Data-Driven Models of Blockage Likelihood in the Wastewater Network

Submitted by James Richard Bailey to the University of Exeter as a thesis for the degree of Master of Philosophy in Engineering, September 2016

I certify that all material in this thesis which is not my own work has been identified and that no material has previously been submitted and approved for the award of a degree by this or any other University.

Signature: .....

Abstract

Blockages are a major problem for Water and Sewerage Companies (WaSCs), impacting on customers and the environment through flooding and pollution incidents. Proactive maintenance aims to reduce this impact by identifying issues and clearing them before there is any impact. Given the large size of the networks, accurate predictions of blockage likelihood are required for this maintenance to be cost-effective. Data mining has the potential to provide these predictions by finding patterns in large datasets. This work presents the novel application of these techniques to the datasets on incidents and assets covering the whole region of a WaSC. The work also contributes an investigation of an input feature formed from a sewer's blockage history and application to real-world data of the techniques decision trees and ensembles methods.

Initially, decision trees were used to produce models at a sewer and area level. General models for the network and for the different causes of blockages were developed. The models are of reasonable accuracy, give a blockage likelihood output and understanding of the important variables relating to blockages. The sewer level models had improved area under the ROC curve (AUC) and gave greater spatial resolution than the area level models. Therefore these were developed using both ensemble techniques and experiments which evaluated the effect of an input feature based on a sewer's blockage history. The historical input feature improved performance, particularly for those sewers most likely to be proactively maintained. Finally, the best performing models were validated using a further dataset of incidents and survey results. The model outputs combined with the historical blockage rate showed good performance for both blockages and flooding incidents on the unseen dataset.

Overall, decision trees gave accurate models on this real-world data and informed which factors influence blockages. Good accuracy was achieved using models including the sewer characteristics, property information and blockage

history. These outputs, validated using the further dataset of incidents, demonstrate the performance of these data mining techniques on real-world data.

Acknowledgements

I would firstly like to thank Innovate UK and Dŵr Cymru Welsh Water for their financial support of the Knowledge Transfer Partnership which enabled me to complete this work.

I would like to thank my supervisory team: Prof. Ed Keedwell, Prof. Zoran Kapelan and Prof. Slobodan Djordjevic. They have helped guide, support and inspire me over the course of the project. I have really enjoyed working with them and the discussions we've had.

To everyone at DCWW who has given up their time to explain things to me, find that useful piece of information or offered support thank you.

My final thanks go to my partner Catherine, without her love, support and the great times we've had together I would not have got this far.

Contents

List of Figures

List of Tables

List of Abbreviations

The following are abbreviations used with the thesis, listed here in alphabetical order:

- ARP - average raw propensity
- AUC - Area Under the Curve
- CARE-S - Computer Aided Rehabilitation of Sewer and Storm Water Networks
- CART - Classification and Regression Tree
- CSO - Combined Sewer Overflow
- DCWW - Dŵr Cymru Welsh Water
- EPR - Evolutionary Polynomial Regression
- FOG - fat, oil and grease
- GIS - Geographical Information System
- OFWAT - Water Services Regulation Authority
- OSAPR - Ordnance Survey Address Point Reference
- PST - Private Sewer Transfer
- ROC - Receiver Operator Characteristic
- SAP - Systems, Applications & Products in Data Processing
- SIM - Service Incentive Mechanism
- SPSS - Statistics Package for the Social Sciences
- UKWIR - United Kingdom Water Industry Research
- VC - vitreous clay
- WaSC - water and sewerage company

Chapter 1

Introduction

This thesis describes the investigation of blockages on the wastewater network and the development of models to predict their likelihood using data mining. The following sections give a background to the problem, the aims and objectives of the work, the publications and contributions from the thesis, and a description of the organisation of the thesis.

1.1 Background

Blockages on the wastewater network occur as a result of the build-up of material which restricts flow. This presents a major cause of flooding and pollution incidents [1] and a large impact on Water and Sewerage Companies (WaSCs). The impacts relate to:

- Direct impact on customers and the environment
- Cost
- Customer service

Flooding and pollution incidents impact on customers and the environment, flooding properties and polluting watercourses. Blockages, along with collapses, are one of the causes of these incidents [1]. WaSCs are also measured on serviceability. When blockages occur the level of service is no longer being provided by that part of the network, affecting this serviceability measure. While individual blockages are relatively cheap to clear, the high frequency of blockages mean

the overall cost is significant. The Service Incentive Mechanism (SIM) [2] score is based on levels of customer service and is also impacted by blockages [3]. Customers contacting the company regarding blockages and their clearance will adversely affect the SIM score of each company. The problem has been compounded by the Private Sewer Transfer (PST) sewers which have recently been adopted by WaSCs in the UK [4]. WaSCs are now responsible for previously private sewers which, for example, are shared sewers from multiple properties. This has caused a large increase in the size of the network. The sewers adopted are those of small diameter close to people's homes, which are believed to represent the highest likelihood of blockages. Given these effects, WaSCs want to understand the likelihood of blockages and the factors influencing this likelihood.

There are a range of causes of blockages, with different mechanisms and different explanatory factors likely to affect their occurrence. Blockages are generally classified into two groups [5]:

- Acute
- Chronic

Acute blockages occur as a result of items like nappies, wipes or rags suddenly blocking the sewer. Chronic blockages occur as result of the gradual build-up of material. For example, silt or fat, oil and grease (FOG) accumulate and reduce the capacity of the sewer until it is insufficient to transport the flow through the sewer. The response of WaSCs has included:

- Public Information
- Proactive Maintenance

Blockages such as those caused by nappies, wipes and rags, and FOG are influenced by behaviour. Public information strategies engage with customers directly following an incident or use campaigns to target specific geographical areas. This aims to change behaviour, reducing the likelihood of this material entering the sewer. Proactive maintenance aims to find blockages as they are building up and

clear them before there is any impact. Given the size of wastewater networks it would be infeasible to proactively survey the whole network. There is a need to identify a small set of sewers which are highly likely to block and target those for maintenance. Hence the need to predict the likelihood of blockage as accurately as possible.

The situation on wastewater networks is therefore a very large network of assets, with a large history of blockages. Data mining techniques allow patterns to be found in data, for example in the occurrence of blockages. This predicts the likelihood of a blockage to prioritise maintenance. This work aims to use real-world data from DCWW to predict the likelihood of a blockage and apply this to the prioritisation of proactive maintenance. The use of this real-world data allows validation and evaluation of these data mining techniques.

The approach taken has been to develop models of blockage likelihood using DCWW's real-world data. Models at both a sewer and area level have been developed using decision trees. The models at a sewer level were also developed using ensemble techniques and the investigation of the use of a historical input feature. The final output from the models has been validated using existing survey results and further incident data.

1.2 Research Questions, Aims and Objectives

The work described here was completed as part of a Knowledge Transfer Partnership between Dŵr Cymru Welsh Water (DCWW) and the University of Exeter. The aim of the work was to apply data mining techniques to calculate a measure of the likelihood of blockages. This likelihood could then be used to prioritise proactive maintenance and reduce the number of blockages occurring.

Research Questions

- How well do existing data mining techniques perform on the prediction of the likelihood of blockages using real-world data from a wastewater network

- Which data mining techniques perform well on the prediction of the likelihood of blockages on the wastewater network
- Which are the most important factors for describing the likelihood of blockages on the wastewater network

Aims

- To develop new, data mining models for predicting the likelihood of blockages

Objectives

- Accurately predict the likelihood of blockages, ideally for a set of high likelihood sewers
- Understand the important explanatory factors describing the likelihood of blockages
- Validate the performance of data mining techniques on real-world data

1.3 Publications arising from this thesis

The following publications have arisen as a result of this thesis:

- Bailey J, Keedwell E, Djordjevic S, Kapelan Z, Burton C, Harris E. Predictive risk modelling of real-world wastewater network incidents, *Computing and Control in the Water Industry*, Leicester, UK, 2nd – 4th September 2015. Procedia Engineering. Volume 119, 2015, Pages 1288 – 1298.
- Bailey J, Harris E, Keedwell E, Djordjevic S, Kapelan Z. Developing decision tree models to create a predictive blockage likelihood model for real-world wastewater networks, *International Conference on Hydroinformatics*, Incheon, South Korea, 21st – 26th August 2016. Procedia Engineering. Volume 154, 2016, Pages 1209 – 1216.

1.4 Contributions

The main contributions from this thesis are:

- The application of data mining techniques to the datasets on incidents and assets covering the whole region of a UK Water and Sewerage Company.
- The validation of data mining techniques on real-world data.
- The investigation of an input feature formed from a sewer's blockage history and its effect on model performance.
- Further information on the important explanatory factors describing blockages.

1.5 Organisation of the thesis

The remaining chapters of this thesis are organised as follows:

- **Chapter 2** contains a review of the literature covering the techniques used to predict the likelihood of blockage, as well as which variables have been investigated and which have been found to be significant.
- **Chapter 3** - describes the preparation of data prior to modelling and the development of sewer and area based models.
- **Chapter 4** - describes the development of the sewer level models using ensemble techniques and the evaluation of the inclusion of a historical input feature.
- **Chapter 5** - gives a summary of the data used and models developed in each stage, the conclusions from this thesis and recommendations for future work.

Chapter 2

Literature review

2.1 Introduction

This review covers the modelling of blockage incidents on the wastewater network of WaSCs. The aim is to investigate both previous work in modelling blockages and the selected explanatory factors used in the models. Existing modelling work includes statistical analyses and a number of different modelling techniques, both of which are reviewed. The review of explanatory variables aims to inform the choice of variables to include in the models to be built, as well as understanding which variables have already been found to be important.

The review is organised to initially cover the analysis and modelling undertaken, before reviewing which variables have been investigated and which found to be important, and finally giving examples of the benefits of ensemble techniques as a method.

2.2 Modelling of Blockages

In the literature, the modelling of blockages is relatively limited. This review considers studies which have used Bayesian modelling, factorial based models, case-based reasoning and Evolutionary Polynomial Regression (EPR) to predict the number of blockages suffered by sewers, or groups of sewers.

2.2.1 Evolutionary Polynomial Regression

Evolutionary Polynomial Regression (EPR) [6] is a hybrid regression analysis and genetic algorithm technique which aims to produce models that include the important factors. The genetic algorithm seeks to optimise the variables which are included in the models. The regression analysis then determines the coefficients and exponents of each of the factors. This provides information on the important variables for modelling and provides an output which predicts the number of blockages.

EPR was applied to two catchments of two different UK WaSCs [1]. EPR was used with five years of incident data to develop models for each of the catchments, predicting the number of blockages at an area level. For each catchment, a model was developed and tested, before being validated on the other catchment. The models performed well in predicting the number of blockages in their own catchments, with coefficients of determination (CoD) of 0.90 and 0.78. However, when the best performing model was used to predict for the other catchment, the CoD dropped to 0.63. As the authors state, this change is most likely due to differences in data recording practices between the two companies. While the technique produces accurate models on the initial data, there is no validation of performance on data from subsequent years.

Savic [7] used data from two UK sewer systems to develop models for blockages using EPR. The dataset represented areas of network, with five years of incident data. The approach derived equations which included the most important variables, with coefficients of determination of 0.86 and 0.76 for the two systems. Savic *et al.* [8] also applied the technique to a dataset of sewer level information, with 10 years of incident data. The coefficient of determination for the developed model was 0.825, showing good predictive performance. Ugarelli *et al.* [9] applied EPR to a dataset from Oslo. The work separated the sewers into cohorts, completing a statistical analysis on the data before developing models for each of the cohorts. The coefficient of determination for these models ranged from around 0.9 to 1, with the exclusion of a cohort which showed a low presence in

the dataset.

These studies demonstrate the benefits of EPR and the accuracy of the models built using it. The technique allows parsimonious models to be built, selecting only the important factors. The most important factors can also be understood from inspection of the derived relationships. However, there is little evidence of how well the technique works on unseen data. There is large variability in the location of blockages and so models must predict well on unseen data, for example from further years of incident data. None of the studies reviewed used a validation dataset which would have demonstrated performance on unseen data.

2.2.2 Bayesian Modelling

Fenner *et al.* [10] used Bayesian modelling to develop sewer level models to extend area level models. At an area level, hotspots of blockage were found and predictions of the number of blockages made using statistical analysis. For the hotspots, sewer level models were then developed. Bayesian modelling was used with the characteristics of the sewers and the historical incidents. This allowed the area level output to be adjusted into a sewer level output.

The area level approach may be of use in identifying hotspots of blockages. By highlighting areas, it may be easier for WaSCs to initially evaluate risk in the network. An area output may also allow consideration of different options for intervention, for example a public information campaign. The probability of failure for each combination of sewer characteristics allows the importance of each of the factors to be understood. The sewer level output then provides the blockage likelihood which can be used to prioritise proactive maintenance at a more detailed level.

The work by Fenner *et al.* provides both an area and a sewer level output. However, the Bayesian models are based on only a few factors. Given the large variation in the occurrence of blockages, ideally as many factors as possible could be included. This would allow more relationships to be derived, which may help explain more of the variation. However, in some methods, such as decision trees,

the inclusion of more variables can interfere with the derivation of relationships. Here the benefits of techniques such as EPR, which allow more variables to be included but select the most important, are demonstrated. Again there is no evidence of how well the outputs perform on subsequent years of data.

2.2.3 Case Based Reasoning

Fenner *et al.* [11] applied a case-based reasoning approach to the modelling of maintenance interventions. This approach develops a set of fully defined cases, including blockage history, maintenance regime and sewer characteristics. New sewers can then be matched to the most similar case, using this case to define the maintenance strategy. For the defined set of cases, each is characterised using the physical attributes of the sewer and the maintenance history, giving indices for condition, performance and management. For new sewers, methods are used to find the most similar sewer within the developed cases. The management strategy of the match is then applied to the new sewer. Expert knowledge can also be incorporated into the approach. A weighting can be given to each attribute that changes the importance of each attribute when a matching case is found. The approach was applied to a case study using data from 20 drainage zones from a water company's region. The cases and indices were developed for this case study area. A set of twenty validation sewers was used to test the performance in finding similar sewers from the cases and the performance in selecting the correct maintenance strategy for that sewer.

Case-based reasoning has advantages in that it gives WaSCs an understanding of the basis for decisions. By reviewing the defined set of cases and the weighting applied to the similarity calculation, it is possible to understand how the method works to add new cases. This will build WaSCs' confidence in the approach. A maintenance strategy can also easily be applied to a newly built sewer without the need for historical data to build models from. However, the approach does not provide any output of the likelihood of a blockage or the number of blockages. The approach also does not use any analysis to define which factors

are important in influencing sewer similarity. This could lead to the sub-optimal matching of sewers to each other.

2.2.4 Factorial Modelling

A factorial model was developed as part of the CARE-S project [5] and used to develop models of blockage risk. Ugarelli *et al.* [12] applied the same technique to a dataset analysed with EPR.

For each sewer a total risk factor is defined based on the characteristics of the sewers. Using a statistical analysis a set of important variables was selected. For each variable, different categories within this variable were defined. For example, for the variable sewer function, categories of combined flow, stormwater and sewage were defined. By dividing the blockage frequency for sewers of that category by the overall average blockage frequency, a risk factor is developed for that category. This risk factor shows the importance of the different variables and categories within them. For any given sewer, the total risk factor can be found by multiplying the individual risk factors together. This is combined with the average blockage rate to give a number of blockages per year. The approach is further refined by minimising the correlation between variables. Variables showing a high correlation will result in a greater weight being given to that risk factor. By selecting variables which show little correlation, an unbiased total risk factor can be found.

The approach was applied in the development of the CARE-S network rehabilitation decision support system. The approach allows the important factors for understanding the likelihood of blockage to be understood from the statistical analysis and the resulting factors. The output is a blockage likelihood which can be used in prioritising proactive maintenance. Ugarelli *et al.* [12] used the technique to evaluate which variables were selected as important factors for predicting the likelihood of blockage, comparing this to those selected using EPR. However, there was no analysis included of how well the approach worked to predict the number of blockages.

2.3 Statistical Analysis

There have also been a number of statistical analyses of blockages by Arthur *et al.* [13], [14]. These studies have investigated the use of an inferred sewer age and CCTV survey data [13], and the detailed consideration of all incidents within a small catchment to produce a statistical analysis [14].

Arthur *et al.* [13] used historic maps of Edinburgh to date developments in the growth of the city. These were used as a surrogate for sewer age when investigating the effect of age on the number of complaints suffered. Sewer condition grades were also used to understand how this varied with sewer age and affected the number of complaints. Their approach was to find the number of complaints per kilometre for each of the categories in sewer age and condition grade. The analysis showed that complaints per kilometre increased with increasing age and increasing condition grade. This demonstrated the benefit of inferring sewer age and the effect of condition grade and age on the likelihood of complaints. However, significance testing would have benefited the conclusions drawn. A scatter graph is plotted to show the average number of complaints for each development period and sewer age. The points on the graph could instead have represented individual sewers or network areas. This would allow an evaluation of the variation in the number of complaints for each category and allowed a trend to be drawn based on all of these points.

Arthur *et al.* also used a dataset of around 30km of sewer from a catchment in south-east England [14]. A small catchment was considered and every incident within that catchment was analysed to produce a statistical analysis of the effect of each of the factors and their statistical significance. The consideration of each incident with a hydraulic model of the catchment allowed inclusion of variables such as modelled surcharge state, lack of capacity and whether flows meet self-cleansing criteria. These factors are more difficult to analyse on a large scale due to the lack of up-to-date models for all catchments, potentially large computation

times and sometimes manual interpretation of the results. The inclusion of statistical significance gives greater robustness to the findings. This work demonstrates the importance of some of the additional factors considered, such as those from the hydraulic model.

2.4 Review of Explanatory Factors

The work covered in sections 2.2 and 2.3 investigated a number of different variables for predicting the likelihood of blockage. The following section reviews the factors which have been investigated and which were found to be important for predicting blockages. The section refers to the papers shown in table 1.

Table 1 shows the work covered in the previous sections, which factors were included within their analysis and which they found to most strongly predict blockage likelihood. A number of different groups of variables have been investigated, including sewer characteristics and the attributes of the surrounding network. The following sections review the appearance and significance of each of the groups of variables investigated.

Table 1: Table showing the papers reviewed and the explanatory factors which they included. A 'Y' indicates the factor was included and found to be significant, a 'N' indicates the factor was included but not found to be significant, while blank squares indicate the factor was not included. The table only includes factors which appeared in two or more of the papers reviewed.

	Sewer Characteristics										Age		Incidents					Network					
Paper No.	**Sewer function**	**Section 24 sewer**	**Depth**	**Diameter**	**Sewer length**	**Gradient**	**Criticality**	**Material**	**Condition Grade**	**Shape**	**Age**	**Era of Construction**	**Previous Incidents**	**Nearby Incidents**	**Property Density**	**Soil Conditions**	**Surface Use**	**Surcharging Sewers**	**Proximity to hydraulic control**	**Flow / Dry Weather Flow**	**Sewer Velocity**	**Position in the Network**	**WaSC interventions**
Fenner *et al.* [10]	Y		Y	Y	Y	Y				Y			Y	N									
Hall *et al.* [1]		Y	N	N		N			Y		Y	N			N	N	N	N					
Savic [7]		Y	N	Y	N	N			N		N				N	Y	N	N					
Savic *et al.* [8]	N		N	Y	Y	Y	N	N		N													
Berardi *et al.* [15]			N	N	N	N		N	N	N	N	N	N	N		N	N		N	N	N		N
Ugarelli *et al.* [9]	Y			Y	N	N		N			Y	N											
Fenner *et al.* [11]	N		N	N	N	N	N	N			N					N	N	N					
Ugarelli *et al.* [12]	Y			Y		N		N	N		Y					N				N	N	N	
Savic *et al.* [16]				Y	Y	N		N	Y												Y		N
Arthur *et al.* [13]																							
Arthur *et al.* [14]	Y			N		Y												N	N		Y		
Hafskjold *et al.* [5]	N			Y		N		Y	N		N											Y	Y

2.4.1 Sewer Characteristics

The sewer characteristics group shows the greatest number of attributes investigated. As there is a very high likelihood of this data being held in WaSCs asset databases, this is widely investigated.

Of the attributes appearing most often, diameter and sewer function appear to be important variables. However, the importance of sewer function depends on whether surface water sewers are included. There is little difference between combined and foul sewers, but there are significant differences between those and surface water sewers. Gradient is regarded as an important variable, with physical implications on the flow through the sewer. However, it was widely found not to be significant. Sewer length also appears quite commonly and does show significance in a number of studies. Condition grade shows significance in two studies from the six which included it. Depth is investigated in a number of studies, possibly because of its influence on collapses, but it seems to be less important for blockages, only being significant in one study.

2.4.2 Age

Both sewer age and era of construction have been investigated. Sewer age is included in a number of studies and found to be important in a number of those. Era of construction was not found to be significant in the studies which included it.

2.4.3 Historical Incidents

The use of blockage history as an input feature to models is limited. However, Fenner *et al.* [10] evaluated a number of features for predicting likelihood of blockage at an area level and found blockage history to be the best predictor of future blockages.

2.4.4 Properties

Property density was investigated in two studies but neither found it to be a significant factor.

2.4.5 Surrounding Material

Soil conditions were investigated in a number of studies but only found to be significant in one. The soil conditions may be of more significance for the physical condition of the sewer, for example for predicting collapses.

2.4.6 Surface Use and Network

A number of attributes in relation to the network have been investigated, including: sewer velocity and proximity to hydraulic control. Surcharging sewers were not linked to an increased blockage risk in any studies which considered this. However, sewer velocity was found to be important in two of the four studies which included this.

2.4.7 Other

A few studies looked at the impact of maintenance interventions, with one finding it to be significant. Other studies aimed to use this data but had problems in processing the data.

2.5 Ensemble Techniques

Ensemble techniques have been widely used in other modelling applications. They function by producing many different models and combining the results from each into a single output. The predictions given by each individual model are maximised, while minimising the correlation between them. This optimises the overall predictions given by the ensemble [17]. This section presents examples of the application of ensemble techniques to complex domains, outside of wastewater

networks. Of the reviewed literature, none used ensemble techniques for the prediction of blockages but they present many benefits in their modelling capability [17].

Cutler *et al.* [18] applied the ensemble technique Random Forests [19] to ecological data and compared the performance with other techniques, such as classification trees alone. The ecological data used was highly dimensional, non-linear and had complex interactions between variables, including many missing values. This therefore provides an interesting case to compare to the data from the wastewater network. Cutler *et al.* found that Random Forests matched or outperformed the other techniques used. It was more accurate and more stable to perturbations in the data. This example provides further evidence as to the benefit of ensemble techniques on complex datasets and their performance in comparison to classification trees alone.

Diaz-Uriarte *et al.* [20] applied Random Forests to the selection of genes in gene expression studies. Here the aim is to find the smallest possible set of genes which give the required predictive performance. The data used has lots of noise and a large number of variables compared to the number of observations. The importance of each variable from this selection must be identified. Random Forests was compared to methods such as k-nearest neighbour and support vector machines. The results showed Random Forests gave a very small set of genes with the same level of prediction as the other techniques. This demonstrates the potential of ensemble techniques for the evaluation of variable importance.

This section has reviewed two examples of ensemble techniques and their potential benefits. The examples showed how the techniques can perform well on complex data and be useful in the evaluation of variable importance.

2.6 Conclusion

A number of studies have used or developed different techniques for predicting blockage likelihood or the number of blockages. However, given the variability in blockage likelihood, it would be of interest to see validation of the models'

performance on additional datasets.

Of the variables investigated none show a clear trend of always being significant but a number are shown to be significant in some cases. Multiple studies show sewer function, diameter, length, gradient, age and velocity to be significant, therefore these are found to be the most important explanatory factors in blockage prediction.

Ensemble techniques as applied in other domains were also reviewed and found to offer improvements in the outputs produced by data mining models.

Chapter 3

Blockage Likelihood Prediction Models using Decision Trees

In this chapter, the developed models of blockage likelihood are presented. The development of sewer level models is described initially, before area level models are developed and evaluated.

3.1 Data Preparation

The following section gives an overview of the datasets which were sourced and investigated for the project, along with an evaluation of the data quality, and overview of the data cleaning and infill required.

To source a list of datasets for use, the corporate systems of DCWW were interrogated, along with datasets which had been used for previous modelling work. The datasets sourced are shown in Table 2.

3.1.1 Dataset of Sewers

The sewer dataset, sourced from DCWW's Geographical Information System (GIS), was used as the basis for the assets to be analysed. This dataset includes fields such as the sewer diameter, sewer length and sewer material which are believed to be of potential for understanding the likelihood of blockage. Each field was evaluated for the amount of missing data and the distribution of data within the categories. Any fields which showed large proportions of missing data

Table 2: Datasets sourced for modelling

Dataset	Source	Basis
Asset and Geographical		
Sewers	DCWW's GIS	Asset
Chambers	DCWW's GIS	Asset
Sub-Catchments	DCWW's GIS	Sub-Catchment
Property locations	DCWW's GIS	Individual property
ACORN	DCWW	Postcode
Property age and type	DCWW	Postcode
Food producer locations	DCWW	Individual property
Maintenance	DCWW	Individual job
Incidents		
Blockages	DCWW Regulatory return	Incident

or records mainly within one category were excluded from the analysis.

Table 12 gives the results of the proportion of missing and invalid records for each field in the sewer dataset. Also shown in Appendix A are the distributions of each of the variables within this dataset. A number of fields are dominated by one category. These include 'lining design type' and 'type of protection' where, of the values, nearly all are completed as 'none', and 'joint type' where nearly all of the values are 'Spigot Socket Mortar'. 'Ground loading', 'ground type' and 'water table' all show very poor levels of completeness. Given these issues, these fields will be excluded from the analysis. The internal and external condition grades would offer explanatory capability but are poorly populated. This information is based on site survey and the coverage across the whole network is low. Also, the older the results, the less likely they are to represent the current condition of the sewers, limiting the confidence in the data. The condition of the sewers could also be highly influenced by local conditions, making it harder to infill this missing data using other characteristics of the sewers. It was therefore decided not to include this field. For sewer gradient, given its implications on flow through the sewer, this variable was infilled. Using the other variables in the sewer dataset,

multiple logarithmic regression was used to infill the missing values of the gradient. The variables used for this were: the categories of sewer ownership, sewer type, sewer function, sewer material, sewer shape and backdrop flag, along with the diameter and length of the sewer. A field denoting the sewer gradient, as well as the source of the gradient were derived.

A number of erroneous values were also found within the datasets. These include the defaulting of sewer construction year to 01/01/1900. Any erroneous values were removed from the datasets in the preparation stage. For the sewer construction year this meant around 35% of the around 700 000 sewers did not have this information. Data verification was also completed by specifying the expected values or range of variables, checking that the values were within these and removing any outside.

3.1.2 Dataset of Blockages

The incident dataset is based on DCWW's regulatory return data for blockages. Eight years of historical data processed and audited in a consistent manner were available. The accuracy of less recent historical data was believed to be affected by inconsistent processing so was not sourced for modelling.

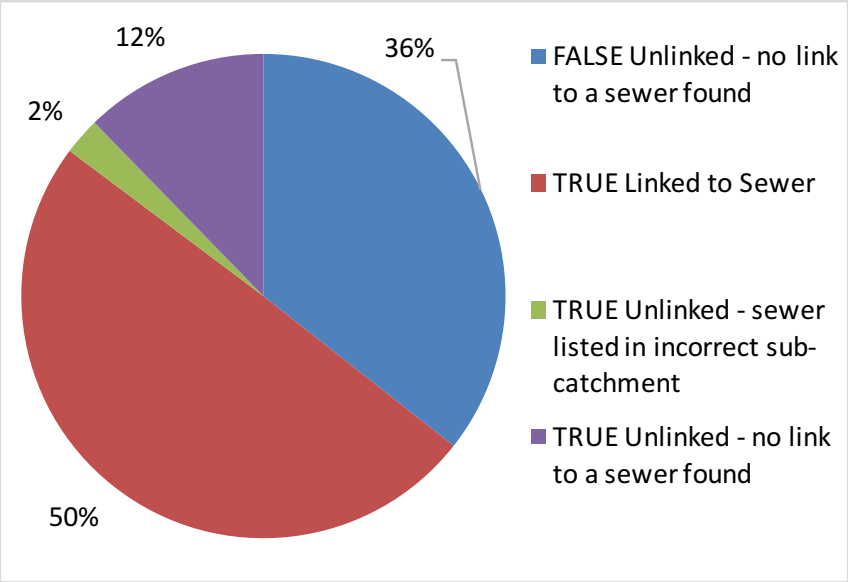
The main fields of interest were the asset to which the incident was assigned and the property information. The asset information allows the predictor variables to be derived for each sewer. The property information is of use for the infilling of the asset information. Figure 1a shows the proportion of incidents which are linked to an asset (True) and the proportions which have no asset information listed (False). Of those with asset information, the incidents are further broken down by whether that asset information matches a sewer in the sewer records. This is denoted by 'Linked to Sewer' and 'no link to a sewer found' respectively. The final category shown is for incidents which are listed against a sewer but where the sub-catchment of the property does not match the sub-catchment of the sewer to which it is assigned. This suggests an erroneous assignment and

so these assignments were removed from the records. This then gives the proportion of incidents which are linked to a sewer and the proportion which are not (shown by Linked or Unlinked in the figure). Figure 1a shows that 64% of incidents are linked to a sewer in the sourced incident dataset, with 50% linked a sewer in the sourced sewer dataset. The graph also show that once erroneous assignments and missing records have been included, 50% of incidents are not linked to a sewer.

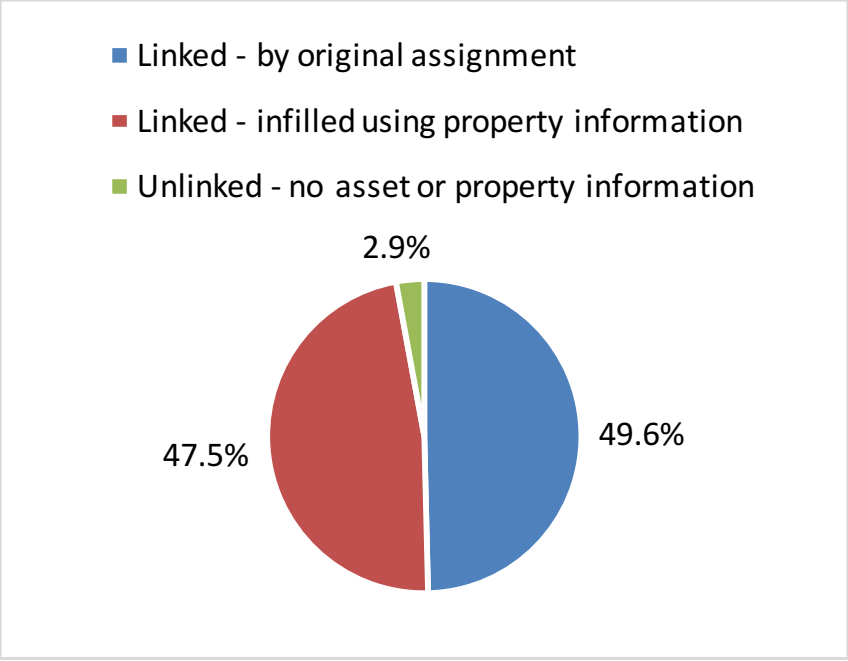
The missing sewer assignment was infilled using a spatial assignment. The location of the property and the assets were used to find the nearest sewer to the property and assign the incident to this asset. The set of sewers used for assignment was limited by excluding those in which it is highly unlikely a blockage would occur, e.g. surface water sewers and rising mains. In addition, the subset of sewers used was limited to those of 225mm diameter or less, which was defined as the set of sewers on which blockages are likely to have occurred. In addition to this, a limit of 25m was placed on the maximum possible distance between the incident and the asset to limit erroneous matching.

Figure 1b shows the proportion of incidents which can be linked to an asset, when the original and spatial assignments are included. The figure shows those incidents linked to an asset using the original asset information and those linked by spatial infilling using the property information. The final section shows those incidents which have no asset or property information in the records and so cannot be linked to an asset. Figure 1b shows that the proportion of incidents which can be linked to an asset is around 97%. The chart also shows that around half of blockages can be linked to a sewer using the original assignment, with another 47.5% that can be linked spatially using the property information. This leaves 2.9% for which no sewer or property information is listed and so no assignment can be made.

The dataset of incidents linked to assets was used to set a flag field representing whether each sewer had suffered a blockage within the years of historical data available. A blockage rate was also calculated using the number of incidents



(a) Pie chart showing the proportion of incidents which have and have not been assigned to an asset, showing True or False for whether a link to an asset was found and linked or unlinked for whether the sewer assigned was found in the dataset.



(b) Pie chart showing the proportion of incidents which can and cannot be assigned to assets. The graph shows those which can be linked using either the asset or property information and those which cannot be linked.

Figure 1: Graphs showing the proportion of incidents which are assigned and could be assigned to assets.

per sewer, normalising the value by the length of the sewer and the number of years of data available. The maximum number of years for the public and PST networks were eight and one respectively. However, if the sewer was found to have been built after the start of this data, then the number of years used was the age of the sewer.

3.1.3 Additional Datasets

The following datasets were sourced from DCWW's corporate systems and included as input features to the models.

Dataset - ACORN Classification ACORN classification [21] is a geo-demographic dataset giving a consumer classification based on "demographic data, social factors, population and consumer behaviour" [22], which is held at a postcode level for the whole of the UK. The dataset gives three classifications: category, group and type. The category and group classifications were used. There are a large number of types, which will limit any understanding gained from its use.

Dataset - Property Age and Type A dataset of property ages and types for the whole of Wales is held by DCWW. This dataset contains, at a postcode level, flags for whether properties within certain age bands, property types and properties containing basements are present. This dataset was used to set the flag fields of whether basements are present, the types of property and for deriving the earliest property age. The earliest property age has been used as a surrogate for sewer age because sewer construction date is poorly completed.

Dataset - Planned and Proactive Maintenance The planned and proactive maintenance data was sourced from DCWW's SAP system. These datasets list the regular planned maintenance, which is completed on sewers and CSOs, and proactive maintenance undertaken following an incident to mitigate risks of a future incident. Issues were found with this data in linking the work done to particular assets and so this dataset was not included in the final dataset for modelling.

3.1.4 Additional Datasets - Derived

The following datasets were sourced from DCWW's corporate systems and used to derive input features to the models.

Dataset - Property Locations The Ordnance Survey Address Point Reference (OSAPR) for properties are held within the GIS systems of DCWW, containing a reference and location for all of the properties in Wales. The dataset was used to derive datasets of the property density and the number of property connections to each sewer. The property density was derived by dividing the area into 100m grids and calculating the number of properties within each grid. For the property connections, the method used to assign blockages to the nearest sewer was used with the properties. The number of property connections to each sewer can then be found. Normalisations of this field were investigated for explanatory capability. These normalisations were sewer length and length multiplied by diameter squared. Each of the different normalisations, and the original field, were evaluated by producing decision trees. It was found that the number of properties per sewer metre provided the best explanatory capability and so this was used. This was believed to make logical sense, with each connection representing a potential site for rags to become caught or protuberances to interrupt the flow through the sewer and increase the likelihood of blockage. The density of these along the length would be related to blockage likelihood in this way.

Dataset - Food Producers A list of food producer locations, linked to an OSAPR reference, were used to derive a number of variables. The number of food establishments within each postcode and 5 digit postcode were derived, along with the number of property connections per sewer. This number of connections was derived in the same way as the property connections, with normalisations again investigated. The normalisation by the length and diameter squared was found to give the best performance. This was again believed to make logical sense. The number of food producers represents a load on the sewers which

would need to be handled by the capacity of the sewer. This capacity will be represented by the length and cross sectional area of the sewer.

Dataset - Sewer Velocity A dataset of sewer velocity was derived to give a measure of the self-cleansing ability of the sewer. The Manning velocity formula under the full pipe assumption was used to derive a sewer velocity. The material of the sewer was used to find the roughness coefficient of the sewer, with the data on gradient and diameter used in the calculation.

Dataset - Upstream Values The network surrounding each sewer is believed to influence the likelihood of blockage. For example, repeated lengths of low sewer velocity may contribute to an increased likelihood of blockage. Two fields were derived, using the upstream sewer velocity and upstream properties per sewer metre. If multiple sewers were found upstream then the average value for the variable was used.

Dataset - Diameter Changes A dataset of diameter changes was derived to investigate the effect of downstream constrictions in flow affecting the likelihood of blockage. The chamber references for the start and end of the sewers were used to find the downstream sewer and its diameter. A nominal field was derived, giving the categories of: smaller diameter downstream, larger diameter downstream, same diameter downstream, no sewers found downstream or multiple sewers downstream found.

Dataset - Chambers and Fittings The datasets for the chambers and fittings were sourced from DCWW's GIS system and used to derive the type of connection to the sewer downstream. This allowed investigation of the effect of manholes in the network.

3.2 Methodology

3.2.1 Statistical Analysis

Using the prepared data, statistical analysis evaluated the explanatory capability of each variable. The blockage rate, normalised by the length of sewer and number of years of data, was used.

For continuous fields, the Pearson correlation coefficient was calculated between each variable and the blockage rate. A T-test was conducted to evaluate the statistical significance of the relationship. For the categorical variables, the average blockage rate was compared for sewers inside and outside of each category. For example, the rate for combined sewers was compared to the rate for all sewers which are not combined. A T-test was performed to evaluate statistical significance, with 5% used to define significance. Any categories which showed a low presence in the dataset were combined into a joint category to reduce the number of categories evaluated.

3.2.2 Data Mining - Sewer Level

The prepared dataset used in the statistical analysis was used to build models predicting the likelihood of blockage. The first stage of modelling used decision trees to build sewer level models.

Decision trees [23] separate the input dataset into groups which in this case show different occurrences of blockages. This forms the nodes within the decision tree. Each separation is based on a single variable and a defined point of separation in that variable. A measure is used to find the best variable for separation and the best point within this to form the separation. Different measures are used for different decision tree algorithms. In Figure 2 this measure is shown as Improvement, with the value of the improvement score shown for each split in the tree. In Figure 2 the first split in the tree is made by the variable 'Properties per sewer metre'. A value of 0.01 is used to separate the records into the two nodes shown in the second layer of the tree. Sewers with a value for 'Properties

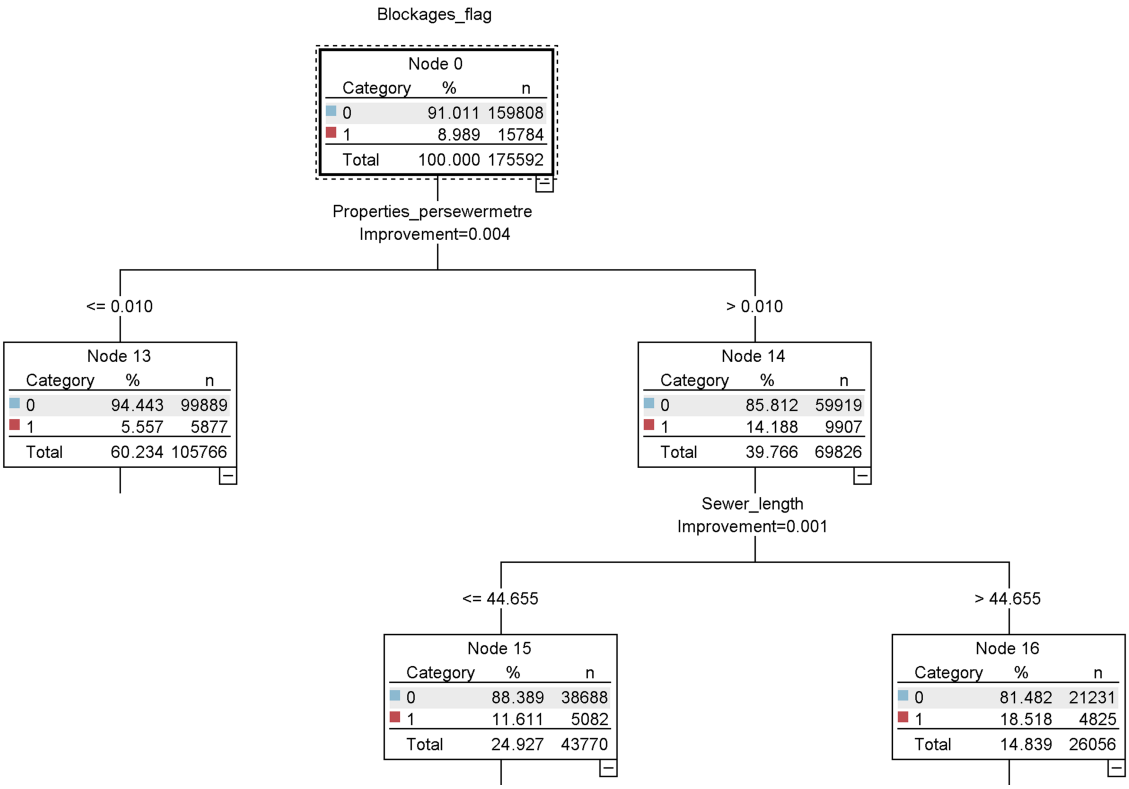


Figure 2: An example decision tree

per sewer metre' greater than 0.01 form one branch of the tree, and those equal to or below form another branch. This variable and value combination were found by the algorithm to give the best improvement score and so were used to form the split. The tree continues to split forming new nodes until a stopping criterion is met. These stopping criteria relate to a minimum change in the measure used for splitting or minimum number of records in the node. When splitting has stopped a decision tree is formed with terminal nodes at the end of each branch. Each inputted record belongs to one of these terminal nodes. The prediction of the output feature is taken from the terminal node. The variables and splits within those variables which form the branches leading to this terminal node are the variables influencing that record's prediction. In Figure 2, for the node labelled Node 16, the likelihood score is formed by the 81% to 19% split between 0 and 1 for the blockage flag, for the records in that node. The variables influencing that score are Properties per sewer metre (greater than 0.01) and sewer length (greater than 44.655). The algorithm selects the best variable for splitting at each point and so the variables appearing in the tree are the important variables. The variables at

the top of the decision tree are therefore the most important. In Figure 2, the best variable and point within the variable for separating the input records was made by properties per sewer metre and the value 0.01. Properties per sewer metre was the best performing variable for separating the records which makes it the most important variables for describing the likelihood of blockage in the dataset used.

As the size of the decision tree becomes larger, overfitting can occur. This is when the model predicts very well on the training dataset but is too specific to that dataset. This means performance on the testing dataset is poor and is likely to be poor on other unseen datasets. Some algorithms therefore include pruning, which is an established method for preventing overfitting [23] and was found to reduce overfitting. Pruning removes nodes from the tree after a stopping criterion has been met by finding and removing the node which causes the smallest increase in error. Pruning continues until a maximum increase in error has been met. Predictions given by the decision tree can be adjusted using balancing and misclassification costs. Balancing alters the input dataset to give a dataset with an equal number of positive and negative cases. This is achieved by taking a sample from the class showing a majority or by repeatedly sampling the class showing a minority. Misclassification costs give a weighting to misclassified records. So a positive record predicted to be negative can be given a greater weight than the vice versa. This adjusts the predictions made by the decision tree.

Decision trees were chosen because they allow a visual understanding of the important explanatory factors as well as producing a blockage likelihood score. The relative importance of the factors can be understood from their appearance in the decision trees and their position in the tree. More important variables will appear closer to the top of the tree than less important ones. Both Classification and Regression (CART) and C5.0 decision trees were used, produced using the software SPSS Modeler [24].

The outputs from the models are evaluated using a Receiver Operator Characteristic (ROC) curve and the area underneath (AUC).

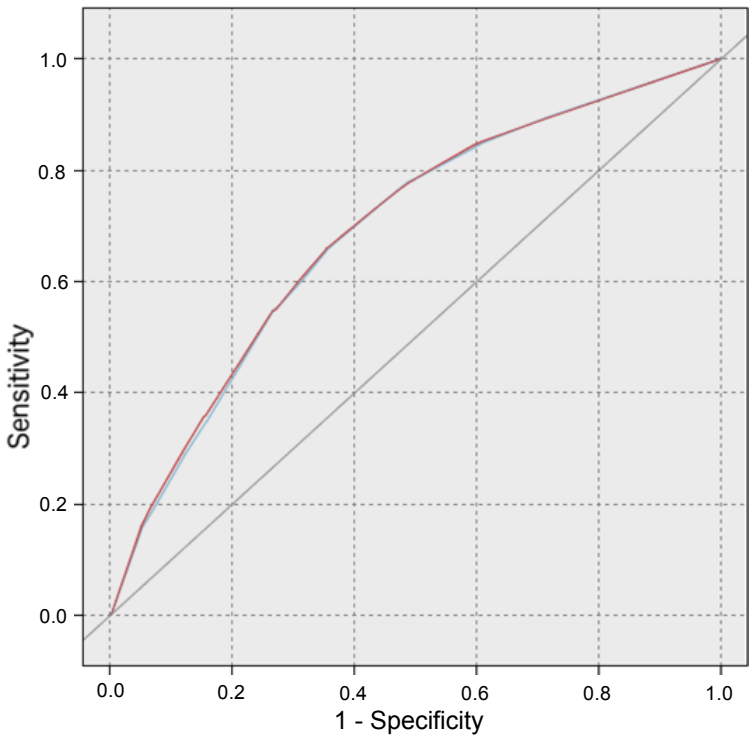


Figure 3: Example Receiver Operator Characteristic Curve (ROC)

A ROC curve plots the sensitivity (or true positive rate) against one minus the specificity (or false positive rate). The y-axis represents prediction performance for the positive records in a binary decision while the x-axis represents negative records. The y-axis represents the proportion of correctly classified positive records, the x-axis the proportion of incorrectly classified negative records. The point (0, 0) means all records are classified as negative, the point (1, 1) all records as positive. A curve describing perfect classification would pass through the point (0, 1). This point means all positive and negative records are correctly classified. A curve passing through (0, 0), (0, 1) and (1, 1) would have an area underneath (AUC) of 1. A curve along the 45° line represents random assignment of records to positive and negative and has an AUC of 0.5. The AUC therefore gives a measure of the classification with the greater the AUC (closer to 1), the better the predictions of the models.

The output feature being predicted was the blockage flag, indicating whether each sewer had suffered a blockage in the period of incident data available. 0 represents no historical blockage, while 1 represents a sewer which has suffered historical blockages. The input features were made up of the other variables in

the prepared dataset, with no joining of categories as was used in the statistical analysis.

Both balancing of the datasets and misclassification costs were investigated to adjust the model predictions. Oversampling of the minority class was used to balance the dataset. The training and testing partitions were formed by random assignment of each sewer in the proportion 70:30. Each model was evaluated using a Receiver Operator Characteristic (ROC) curve and the area underneath the curve (AUC).

Models were produced for the different subsets of the network and for the different causes of blockages. The public and Private Sewer Transfer (PST) parts of the network have different levels of available historical data. This means that the relative proportion of sewers which have blocked is different. To account for this, separate models were built for these parts of the network. The sewer function (surface water, foul, combined) was also used to separate the network. Some sewers, such as surface water sewers, are highly unlikely to suffer a blockage so were excluded from modelling. Foul and combined sewers may show different influencing factors because of the presence of surface water in the combined sewers. Separate models were therefore also produced for these subsets of the network.

Investigations into the different causes of blockage were also made. Given the different mechanisms of blockage formation, there are likely to be different explanatory factors which are important. The different classifications used by DCWW were therefore grouped into those that showed similar mechanisms. For these models, the blockage flag was then formed by whether each sewer had suffered that type of blockage.

3.2.3 Data Mining - Area Level

Models were also produced at an area level, again using the same prepared dataset of incidents and assets. The possible geographical areas for modelling were evaluated and the best selected for use. The input and output features were

then aggregated to this area and a single model produced for all sewers.

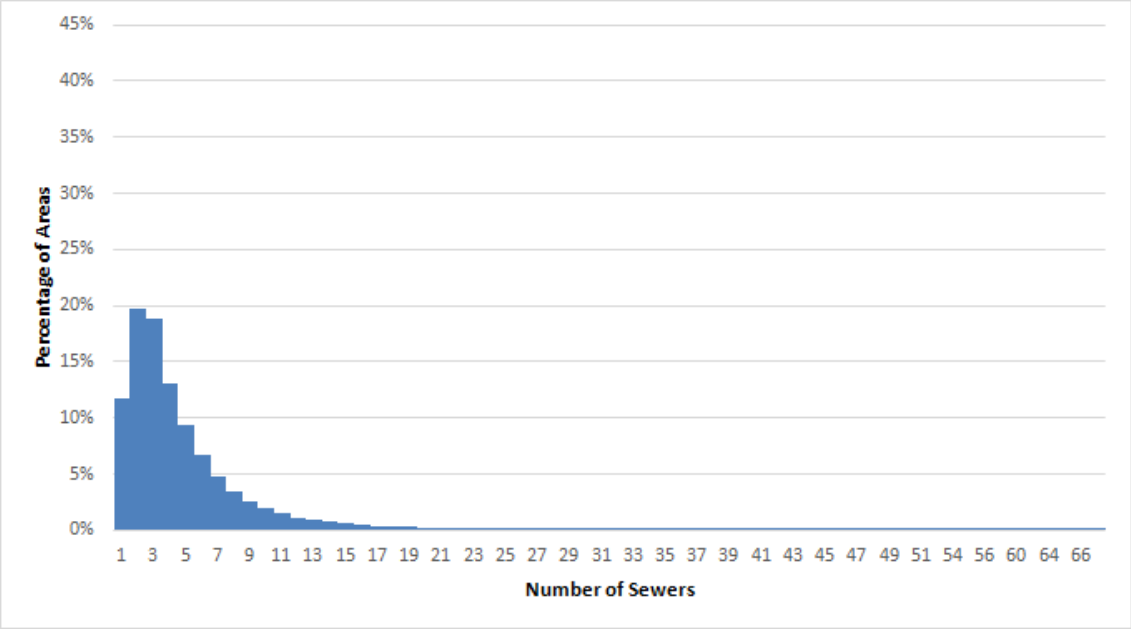
Selection of Area

The selection of area was based on the ideal of groups with similar total lengths of sewer and areas which matched those issued by WaSCs for proactive maintenance. The areas evaluated were postcode, 100m by 100m grids and combinations of the two.

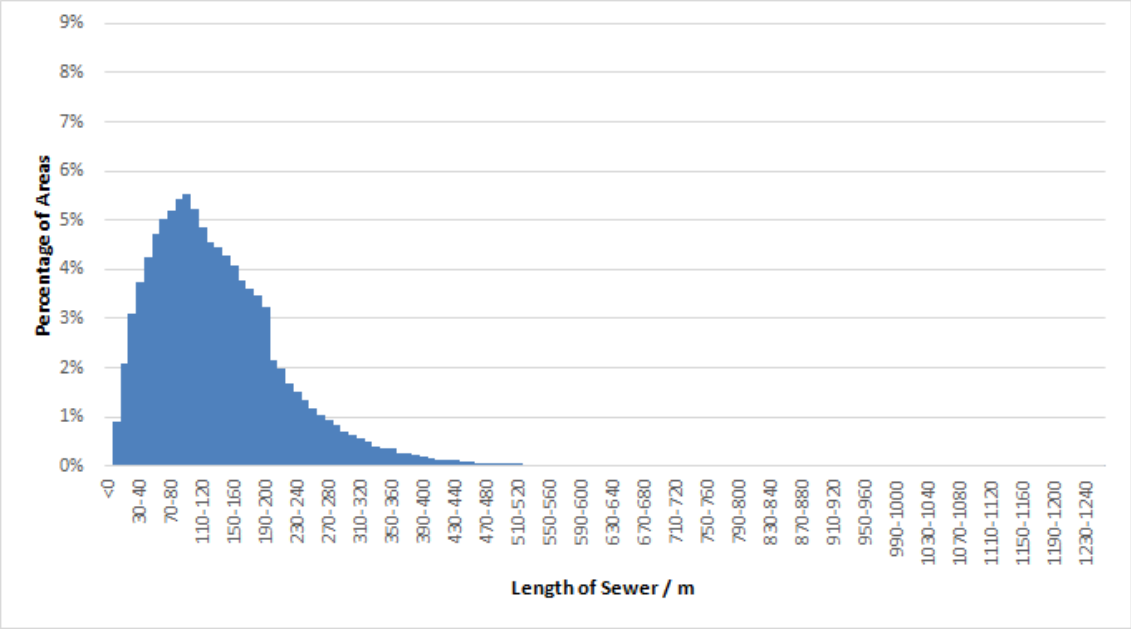
The geographical groups selected gave a distribution (Figure 4) which showed a relatively narrow distribution of sewer lengths and counts, while also reducing the number of groups which only contained one sewer. The groups used postcode to generate the initial groups. Larger postcode areas were then broken down into smaller groups using the intersection of postcode and 100m by 100m grids. This results in a large proportion of groups with only one sewer. Groups with only a single sewer were reassigned to neighbouring groups. If a downstream sewer is present assignment was made to this group, otherwise the upstream sewer. If neither up or downstream sewers were present, the sewer remained in a group of its own.

Derivation of Aggregated Variables

Input Features Different methods were used for continuous and categorical variables. For categorical variables, a new variable was created for each category. For each category within each categorical variable, the total sewer length was found. For example, for sewer function the length of combined sewer and length of foul sewer was found for each area. For continuous variables, the first aggregation used a length weighted average for each area. The second method discretised the variable and treated them as categorical variables. If a break in the definition of the range of data is present, then this was used for discretising. If none of these were present, then quartile values were used to give evenly sized groups.



(a) Distribution of sewer count



(b) Distribution of sewer length

Figure 4: Distributions of sewer count and length for aggregated areas derived using postcode, breaking larger postcodes down using 100m grids and re-assigning groups with a single sewer to up or downstream sewers.

Food Producer and Property Connections For the food producer and property variables, the majority of sewers had a value of so the bands used were 0 and greater than zero.

Diameter For diameter, sewers of 225mm or smaller are regarded as small bore sewers. The majority of sewers are less 225mm so a further threshold of 150mm was used to derive three groups.

Sewer Velocity For sewer velocity, 1 m/s has been used as a definition of self-cleansing velocity so this defined one group. The majority of sewers are above this threshold, so the sewers were further split using 2m/s as a threshold.

Gradient and Property Density For gradient and property density, no particular cut-offs could be defined, so the quartile values were used to define bands.

Output Feature To derive the output feature, the difference in historical data between public and PST had to be accounted for. A relative proportion of blocked sewers was derived. For each group, for public and PST, the proportion of assets which had blocked was found. These values were divided by the average proportion of blocked sewers for all areas, calculated separately for public and PST. The length weighted average of these two relative proportions was then found. This gives a continuous measure, where one represents an average relative proportion of blocked sewers and zero a group showing no blocked sewers. The models were produced to still predict a blockage flag. Different thresholds in the relative proportion were used to derive blockage flags. These were then predicted. Models were produced using thresholds of 0, 1, 2, 4, 6 and 8.

3.3 Results and Discussion

3.3.1 Statistical Analysis

Figures 5 and 6 show the results of the statistical analysis completed to understand the explanatory capability of the variables sourced and derived. The correlation analysis gives each of the variables, also showing whether the result was found to be statistically significant. For the difference in average analysis, only the statistically significant variables with above average blockage rates are shown. The variables are ordered from left to right by the absolute difference in the blockage rates.

Correlation Analysis

The top predictor shown in Figure 5 is that of property density. An increased property density is likely to be linked to a higher number of properties connected to the sewer. This could lead to a greater potential for a large volume of material, from the properties connected, to enter the sewer and quickly form a blockage. The greater the number of properties, the greater the potential for this spike in material. A greater number of properties will also give a generally higher load on the sewer and potentially more material which can form a blockage. The property connections into the sewer also represent potential sites for flow disruption which could cause material to settle out from the flow, which could also increase the likelihood of a blockage.

The highest negative predictor found is sewer diameter, showing that smaller diameters result in a higher rate of blockage. Material from residential properties enters the network, initially into small diameter sewers. A sudden spike in material entering the sewer or flushing of rags or debris has the potential to quickly form an acute blockage, preventing further flow through the network.

Construction date also shows as a negative predictor, showing that older sewers show a higher rate of blockage. Older sewers have the potential to be built

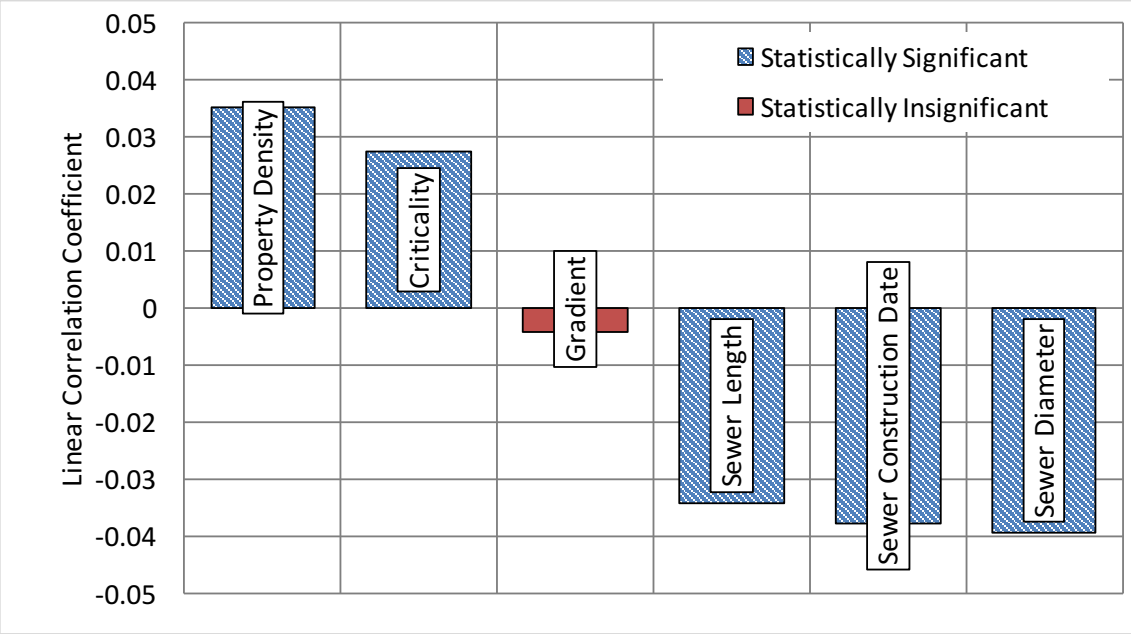


Figure 5: Graph for public sewers showing the Pearson linear correlation coefficient between the factors being investigated and the rate of blockage in blockages per km per year.

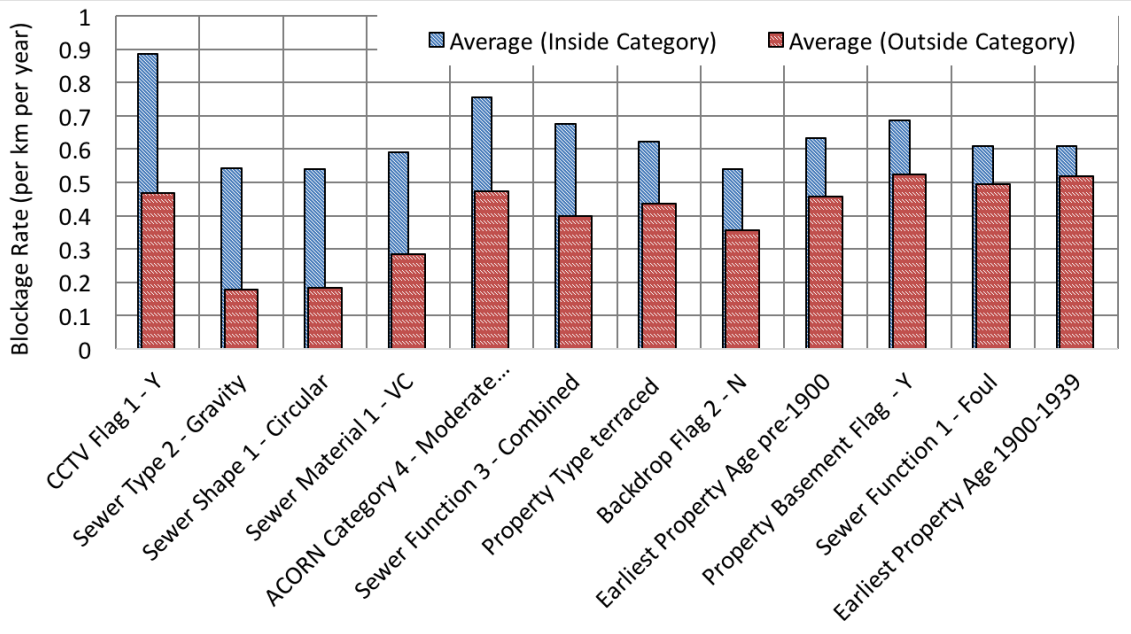


Figure 6: Graph for public sewers showing the effect of the categorical variables investigated. For each category within each variable, the average blockage rate (per km per year) is plotted for sewers in that category with the average blockage rate for sewers which are not in that category. The graph only shows differences which are statistically significant at a 5% significance, with the variables ordered from left to right by the absolute difference in blockage rate.

to different design standards and could be more likely to block than modern sewers. Older sewers also have a greater potential for defects to be present over the time they have been installed, which can result in material becoming caught in the sewer and the build-up over time to a full blockage.

Sewer length shows a weak correlation to blockage rate, meaning that shorter sewer lengths show a higher likelihood of blockage per length of sewer. This could be linked to a number of factors. Short lengths of sewers mean that there are manholes, junctions or chambers in the network, breaking sections up into a greater number of smaller sewers. Within DCWW, interceptors are believed to influence blockages, with work competed previously to alter the network around them and allow easier clearance of blockages formed in them. With interceptors installed in chambers the risk of blockage from interceptors may be helping increase the risk of blockage on short lengths of sewer. The potential for blockages in manholes was also highlighted, with the potential obstructions in the manholes helping form blockages. Arthur *et al.* [14] also found blockages forming near to sewer junctions, where disruptions to the flow are present.

Difference in Average

The categorical field analysis, Figure 6, highlights a number of variables, including: sewer diameter, sewer length and other fields: ACORN category, backdrop flag, property ages, material type, sewer function and sewer shape. Property age was highlighted by the presence of the pre-1900 and 1900-1939 categories from the earliest property age variable. The property age is linked to the age of sewers with the earliest property age often being used as a surrogate for the age of sewer. As discussed from the correlation analysis the age of the sewer is linked to the likelihood of blockage.

ACORN Category 4 and a number of ACORN groups were highlighted as having high incident rates. The ACORN Category highlighted was 4 and, of the groups highlighted, two of the three category 4 groups are present, along with one from category 3. Further investigation was conducted to evaluate the effect of

Table 3: Variation of property density across ACORN categories.

	Average Property Density	No of Records
ACORN Category - 1	13.7	232364
ACORN Category - 2	36.2	22193
ACORN Category - 3	22.5	202191
ACORN Category - 4	29.4	129778
ACORN Category - 5	25.0	139730
ACORN Category - 6	19.5	2796
Overall	21.8	729052

Table 4: Variation of property density across ACORN groups.

	Average Property Density	No of Records
ACORN Group - A	13.5	75538
ACORN Group - B	11.3	82016
ACORN Group - C	16.6	74810
ACORN Group - D	27.4	6044
ACORN Group - E	38.7	5180
ACORN Group - F	39.9	10969
ACORN Group - G	33.1	29375
ACORN Group - H	20.9	105709
ACORN Group - I	18.7	54232
ACORN Group - J	27.6	12875
ACORN Group - K	45.1	1572
ACORN Group - L	28.2	31130
ACORN Group - M	29.5	97076
ACORN Group - N	23.7	109826
ACORN Group - O	28.5	24745
ACORN Group - P	34.6	5119
ACORN Group - Q	46.6	40
ACORN Group - U	19.5	2796
Overall	21.8	729052

these and the effect of property density which was found to be a strong predictor in the correlation analysis and would be expected to vary with the ACORN Category. Tables 3 and 4 give the average property density for each of the ACORN categories and types. This shows that Category 4 has a property density of 29.4 properties per 100m square, compared to the overall average of 21.8. While this is not the highest property density, this value is above average, which could suggest that the effect of ACORN Category is linked to the property density. This will be further informed by the decision tree modelling, which will evaluate the explanatory potential of each of the input features.

For sewer function, there are three possible functions: foul, combined and

surface water, which represent most of the sewer dataset. Foul and combined sewers are both shown as high predictors, but because the other category is surface water, where blockages would not be expected, the rate of blockage for foul and combined sewers would be expected to be higher than the overall average rate of blockage. However, the higher rate for combined sewers does provide further information. The combined network will be affected by the rainfall patterns, with periods of dry weather causing the build-up of material which is then found when the first rainfall occurs and the material which has built-up constricts the flow through the sewer.

Backdrop flag being defined as 'No' is shown as a high predictor, with the possible values being Missing, 'Yes' or 'No'. Backdrops are used in areas of steep gradients where the backdrop is used to drop the sewer deeper, by including a step in the sewer. Backdrops are believed to be linked to blockages because they represent a potential obstruction in the sewer. It would not therefore be expected that the confirmed lack of presence of a backdrop would be linked to higher blockage rates. Again, the apparent explanatory capability of this variable may be due to the link to other variables, which can be further understood by the modelling using decision trees.

Table 5: Table comparing the average sewer length, diameter and construction date for sewers which are (1) and are not (0) of material type 1 (vitreous clay)

Sewer Length

MATERIAL_T=1	Average	No of Records	Standard Deviation
0	46.8	110386	95.0
1	27.1	621971	26.0

Sewer Diameter

MATERIAL_T=1	Average	No of Records	Standard Deviation
0	420.7	110384	493.9
1	166.6	621954	69.3

Construction Date

MATERIAL_T=1	Average	No of Records	Standard Deviation / Days
0	10/12/1956	110386	14951.2
1	08/09/1942	621971	14262.2

Material types of vitreous clay (Material type 1 - VC) were highlighted in the

analysis and their characteristics were investigated (Table 5). Sewers in this category are older, shorter and smaller in diameter, which may influence the higher rate of blockage.

Another field to highlight is gradient, which was expected to be linked to blockages but does not appear in this analysis. A shallower gradient will mean a greater difficulty in moving material through a sewer and potentially a higher rate of blockage. It is possible that the low level of completeness of the gradient field means it is more difficult to find a significant relationship, so that gradient was not highlighted in the analysis.

For PST sewers, the linear correlation coefficients shown in Figure 7 show weaker correlations between blockage rate and the fields investigated. Construction date is the exception to this, showing a similar result to that for public blockages, again indicating the older the sewer, the higher the rate of blockage. The correlations may be weaker because of the similarity of characteristics reducing the spread of values and the possibility of finding a trend in the values. PST sewers are the smaller diameter sewers, close to properties and therefore have similar characteristics. For sewer diameter, for example, PST sewers are on average smaller and show less variation in their values. PST sewers have an average diameter of 125mm, standard deviation of 66mm, while public sewers have an average of 249mm, standard deviation of 264mm.

The categories highlighted in Figure 8 show a mix of results, with a number repeated from the analysis of public blockages and some additional categories: Material type -1 (the combined category of low presence categories) and a larger mix of ACORN categories and groups. Of the repeats, combined sewers and the presence of a basement appear higher up the list for blockages on PST sewers. The corresponding categories found include older, shorter and smaller diameter sewers, along with ACORN Category 4 and the presence of basements. For the ACORN fields, there is a larger mix of categories than was found for public sewers, with both the ACORN Categories 4 and 5 appearing, along with ACORN

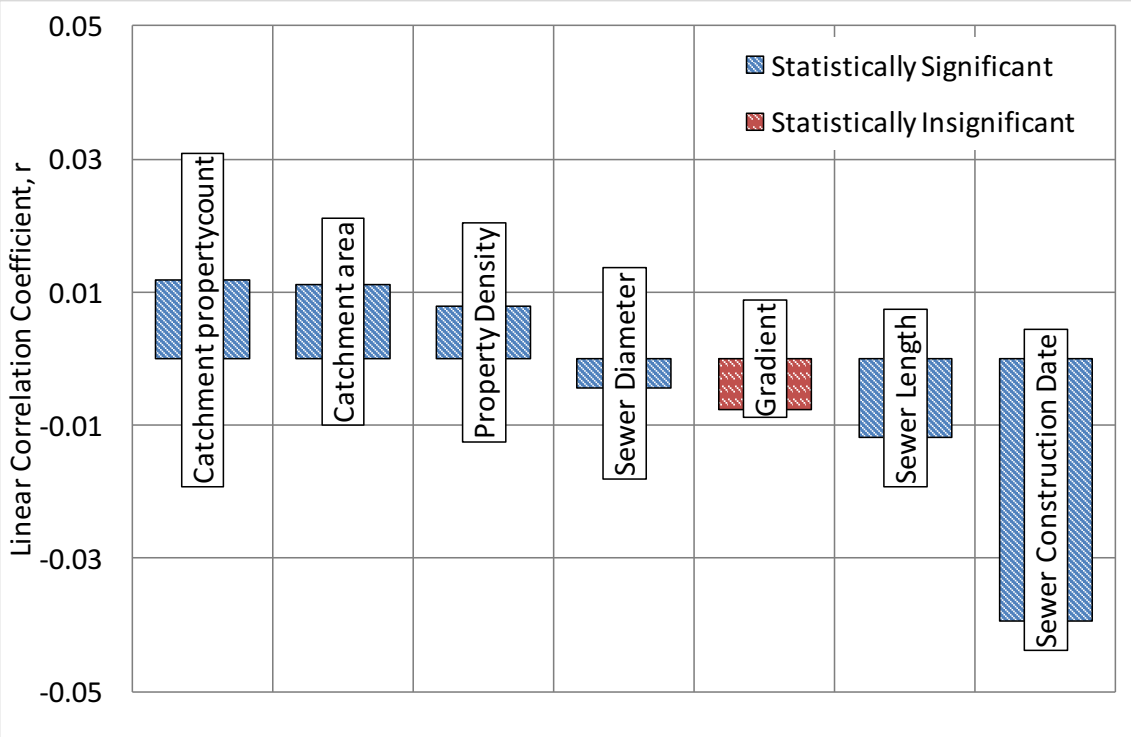


Figure 7: Graph for PST sewers showing the Pearson linear correlation coefficient between the variables being investigated and the rate of blockage in blockages per km per year.

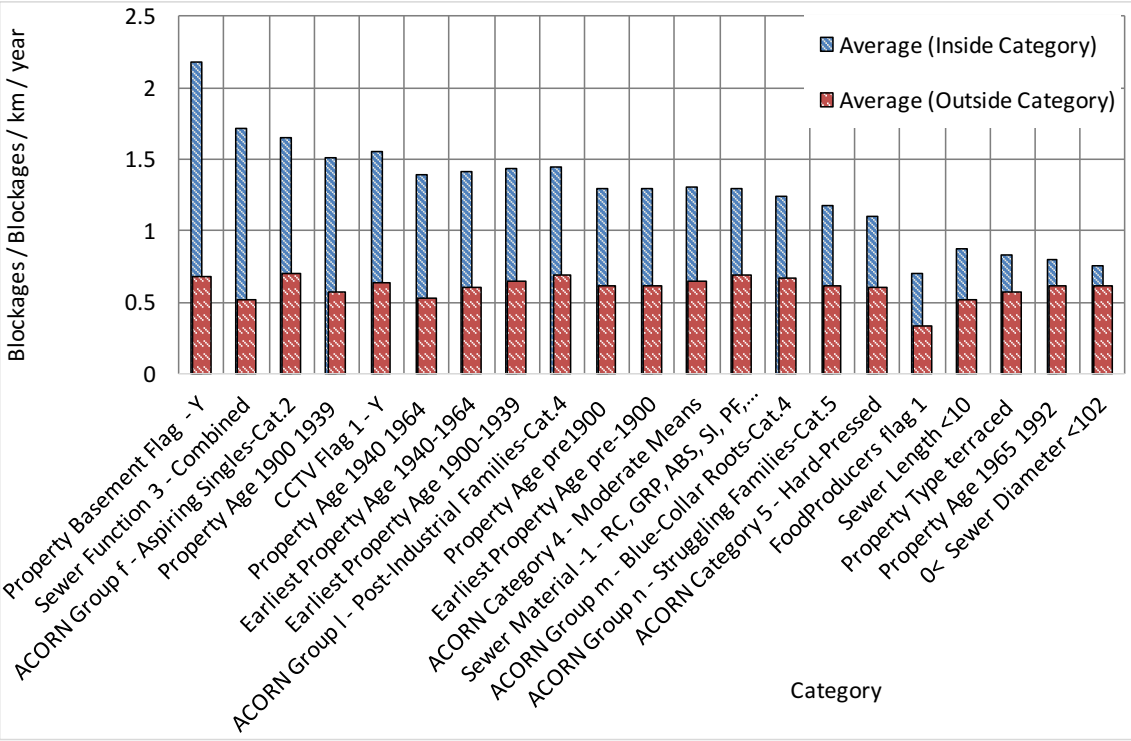


Figure 8: Graph for PST sewers showing the effect of the categorical variables investigated. For each category within each variable, the average blockage rate (per km per year) is plotted for sewers in that category with the average blockage rate for sewers which are not in that category. The graph only shows differences which are statistically significant at a 5% significance, with the variables ordered from left to right by the absolute difference in blockage rate.

groups from categories 2, 4 and 5. All three of these show higher property densities than the other ACORN categories, with Group F also showing a higher property density, as shown in tables 3 and 4. This may mean that it is the higher property density which contributes to the increased risk, although the correlation analysis did not show a strong relationship between blockages and property density, so this may be less applicable for PST incidents.

Material type -1 is significant for PST sewers (Figure 8) where it was not for public sewers. Again links to other fields were investigated but no significant differences were found in comparison to the dataset as a whole. It is difficult to make further conclusions for reasons behind the increased blockage rate, given the large number of material types which make up this combined set.

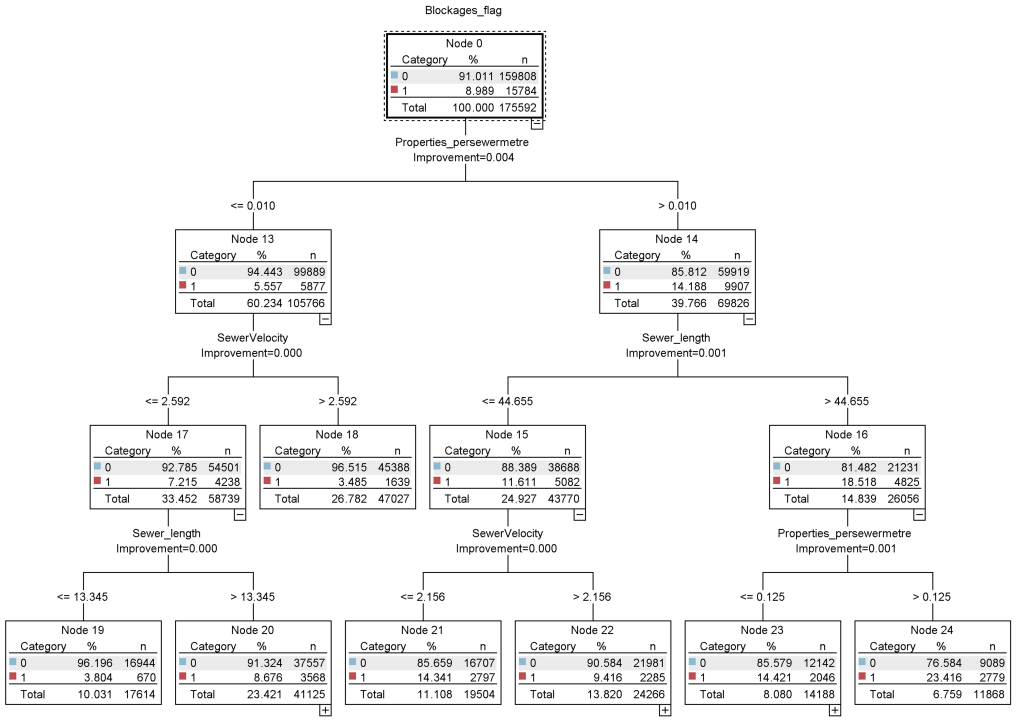
3.3.2 Data Mining - Sewer Level

All Blockages

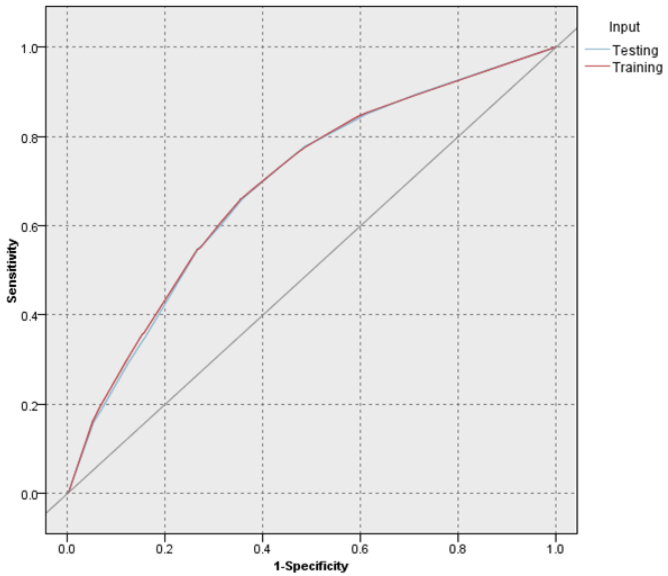
Table 6 gives the results of the decision tree modelling for the sewer level models developed. Overall, the results show reasonable performance in predicting blockages, with the AUCs ranging from 0.65 to 0.72. The following paragraphs review the decision trees of each subset of the network for their performance and variables highlighted.

Table 6: Results of the single decision tree model on the four subsets of the overall sewer network.

Model	Accuracy	AUC
Public - combined	65%	0.69
Public - foul	64%	0.65
PST - combined	62%	0.66
PST - foul	65%	0.72



(a) Decision Tree, showing the top four layers of the tree.



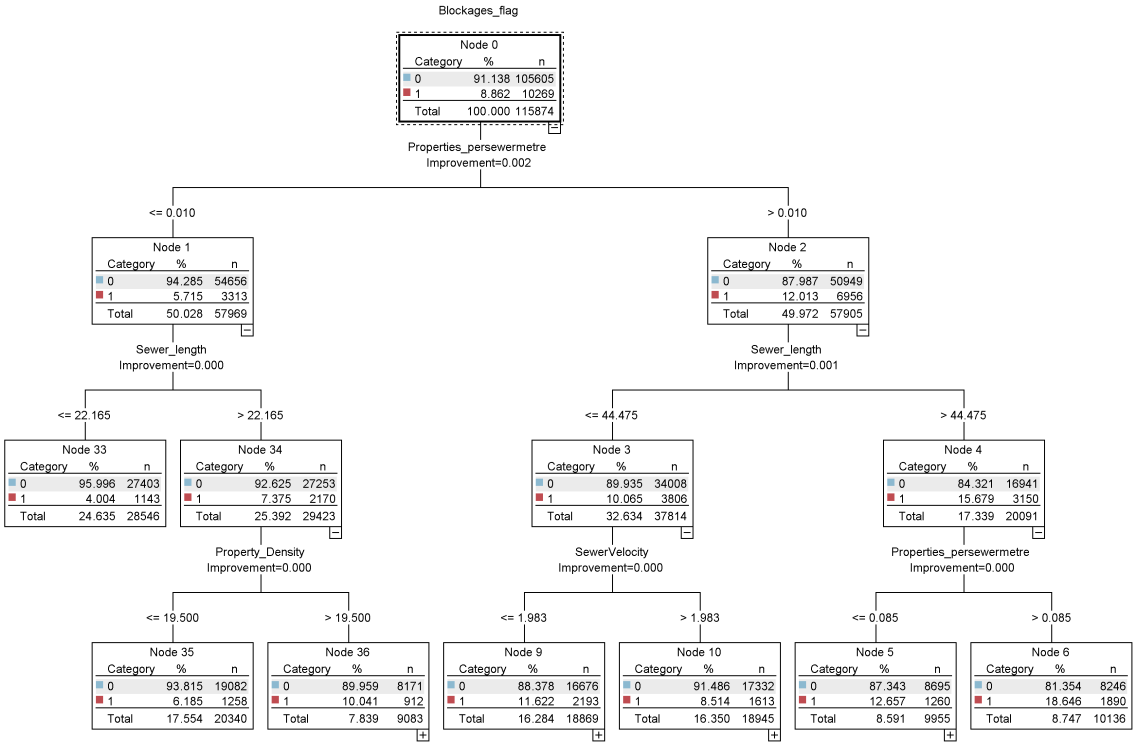
(b) ROC curve

Figure 9: Results obtained from the models for the public, combined subset of the network.

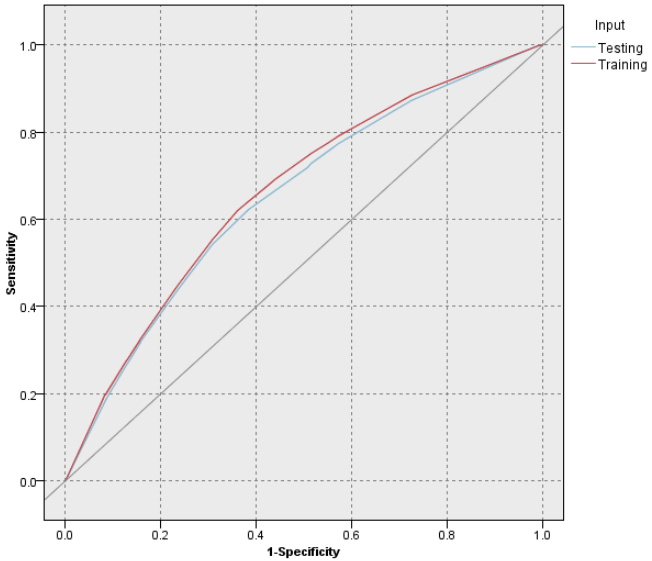
Public Combined Sewers This model (Figure 9) shows a reasonable level of prediction and useful relationships for understanding the blockage likelihood.

The ROC curve (Figure 9b) shows an AUC of 0.69 on the testing dataset, with a fairly even shape to the curve. This AUC is around the average performance for these models. The application of the work is to the prioritisation of proactive maintenance. Given the size of the network, only a very small proportion of the network can be surveyed each year. This means the ideal shape of the ROC curve would give the best performance for the highest likelihood sewers, those which would be prioritised for maintenance. The evenness of the shape of this curve means no particular benefits are seen for the highest likelihood sewers.

The variables at the top of the decision tree (Figure 9a) include properties per sewer metre, sewer velocity and sewer length. These variables make logical sense for representing blockage likelihood. Increasing properties per sewer metre represents an increased load of material on the sewer or increased disruptions to the flow from the sewer connections, both increasing the likelihood of blockage. Sewer velocity represents a self-cleansing ability, linked to the build up of material to form a blockage. Greater sewer length represents an increased length of sewer on which a blockage could have occurred and increased likelihood of blockage. The decision tree also shows sewer velocity rather than gradient providing the greatest explanatory capability.



(a) Decision Tree, showing the top four layers of the tree.



(b) ROC curve

Figure 10: Results obtained from the models for the public, foul subset of the network.

Public Foul Sewers Figure 10 shows results from the model of public, foul sewers. The model again shows reasonable performance and some useful relationships from the decision tree.

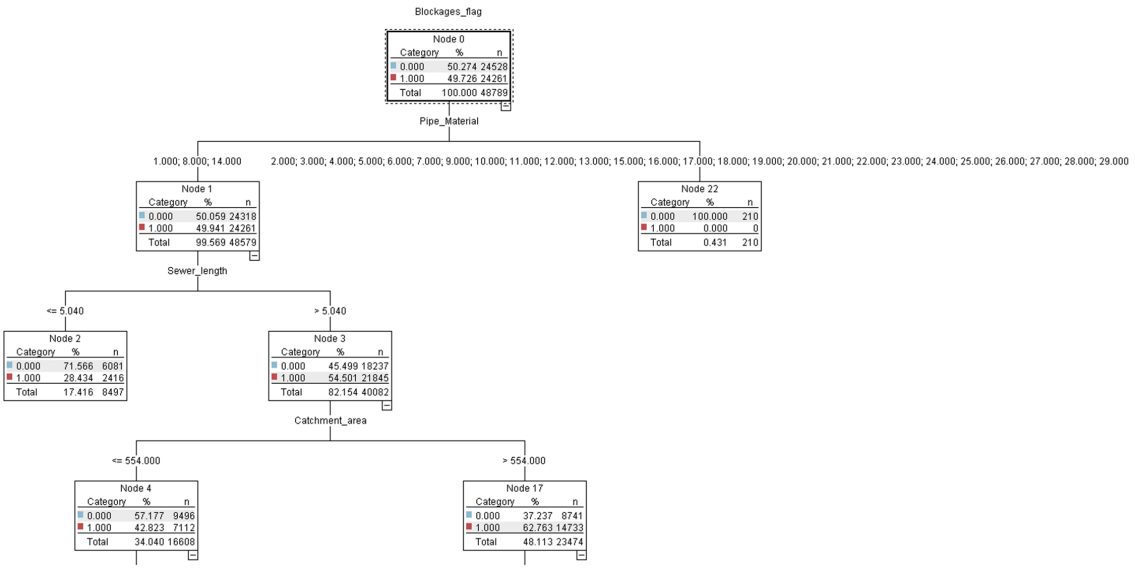
The AUC is 0.65, which is similar to the other single models. Although this model is the worst performing of the four. The ROC curve shows an even shape with little overfitting to the training dataset. As for public, combined sewers this gives no improvement for the highest likelihood sewers.

The important variables in the decision tree are properties per sewer metre and sewer length, with sewer velocity and property density also appearing in the decision tree. The relationships here are slightly less useful because of the increased frequency of sewer length. While this explains the likelihood of blockage - a longer length meaning more sewer on which a blockage could happen, it is less useful for prioritising proactive maintenance. If the sewers maintained were based on the sewer length, then greater weighting would simply be given to the longer lengths of sewer. This wouldn't account for the cumulative risk posed by multiple shorter lengths of sewer. The relationships here, with sewer length more frequent, are therefore less useful in the application of this work.

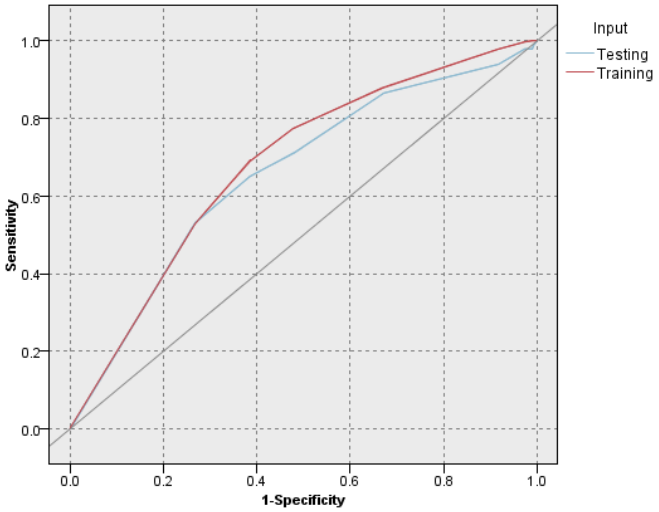
PST Combined Sewers The model of combined PST sewers (Figure 11) shows similar performance to those of the other models. However, the ROC curve does show some overfitting to the training dataset. The variables in the decision tree also show less potential for use in prioritising proactive maintenance.

The ROC curve shows a reasonable area underneath, but an uneven shape. The first part of the curve shows the same gradient. This means that there is no difference in the likelihood scores given for the sewers in this part. This gives poor performance for the highest likelihood sewers. The second part of the curve then shows overfitting to the training dataset. The unevenness to the curve means this model shows poorer performance than those of the public network.

The decision tree shows the variables at the top of the tree are sewer material, sewer length and catchment area. While sewer material could influence the likelihood of blockages, the split in the decision tree simply shows a number



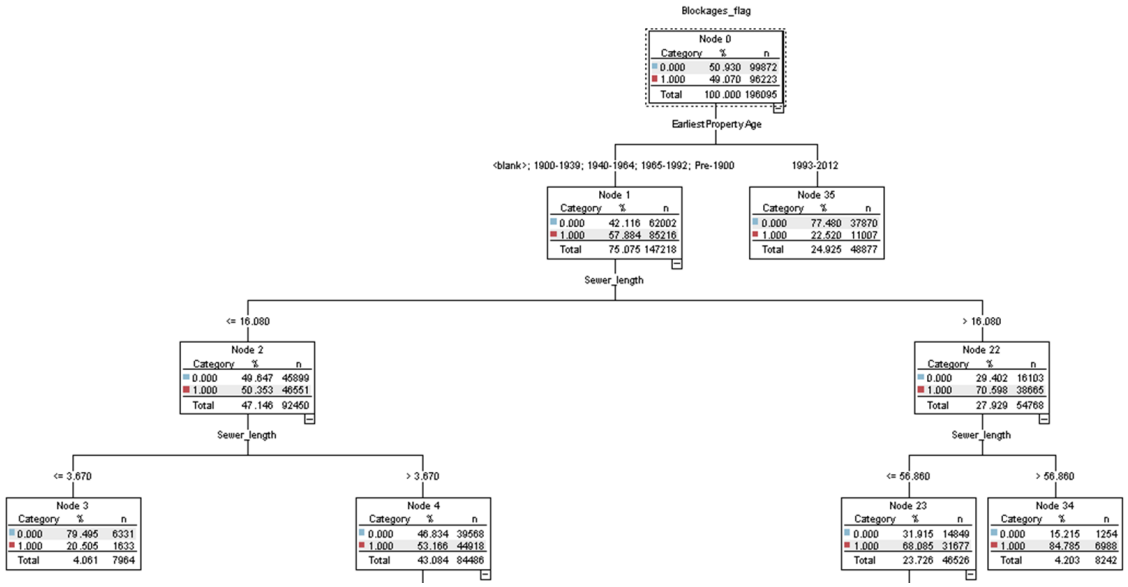
(a) Decision Tree, showing the top four layers of the tree.



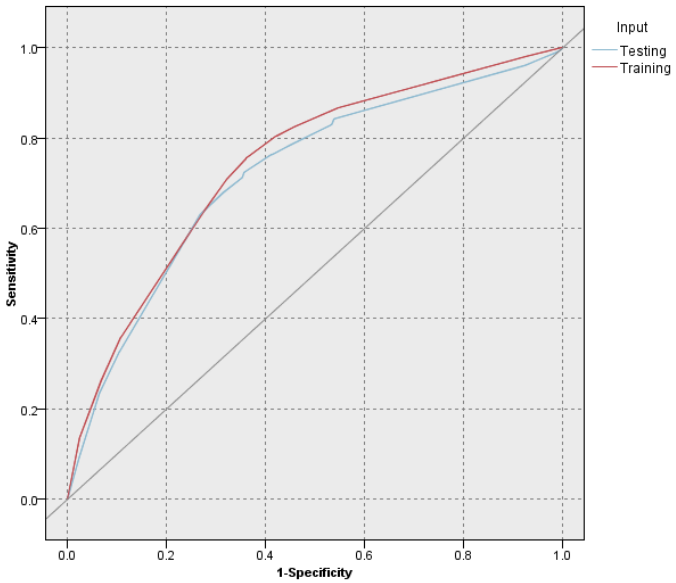
(b) ROC curve

Figure 11: Results obtained from the models for the PST, combined subset of the network.

of material types which have no history of blockage. Given the small number of sewers in this split, this relationship is less useful. Sewer length can be understood to affect the likelihood of blockage but, as explained for public, foul sewers, (page 57) is less useful for this application. The final variable is the catchment area. The sewers in the catchments connect and flow out through a common point, e.g. to a treatment works or pumping station. There wouldn't seem to be a physical explanation which would cause the size of the catchment to influence the blockage likelihood on every sewer in the catchment.



(a) Decision Tree, showing the top four layers of the tree.



(b) ROC curve

Figure 12: Results obtained from the models for the PST, foul subset of the network.

PST Foul Sewers The model of foul, PST sewers (Figure 12) seems to show reasonable overall performance. However, the variables in the decision tree are less useful for prioritising proactive maintenance.

The ROC curve shows an AUC of 0.72, which is similar to the other models. The curve shows a steeper initial part to the curve, giving better performance for the highest likelihood sewers. However, the curve also shows overfitting for the lower likelihood sewers.

The variables at the top of the decision tree are earliest property age and sewer length. The earliest property age is used as a surrogate for sewer age. Older sewers could be linked to a higher likelihood of blockage because of the greater likelihood of defects, or from different design and build standards. Defects in the sewer will act as sites for the flow to be disrupted and increase the likelihood of material settling out or becoming trapped, aiding in blockage formation. Sewer length then forms the remaining splits in the top part of the decision tree. As explained for public, foul sewers (page 57) this variable is less useful for prioritising proactive maintenance.

Overall Discussion The models of the public network show better performance than the models of the PST network. While the overall AUC's for the models are similar, the relationships shown in the decision trees are of less use and the ROC curves show more overfitting.

The models of the public network shows reasonable overall performance, with no overfitting to the training datasets and even shape to the ROC curves. The variables formed from the basic sewer characteristics and property information would seem to give useful relationships and explanatory capability. The decision trees for the public network demonstrate the benefit of some of the derived variables (properties per sewer metre and sewer velocity) from their appearance at the top of the decision trees. Sewer velocity uses the infilled gradient data and sewer diameter. The greater frequency of sewer velocity when compared to gradient demonstrates the greater explanatory capability of the combination of gradient and diameter. This also shows the benefit from the infilled gradient,

despite the originally high levels of missing data.

The increased difficulty in predicting for the PST sewers may be due to the more similar nature of these sewers and the reduced amount of historical data available. The PST sewers are more likely to show similar characteristics, being the smaller diameter sewers closer to homes. This may mean that variables highlighted for the public network do not show the same explanatory capability. For example, the values of properties per sewer metre for the PST network are more likely to be similar and may therefore not give the range of values required to show a relationship. For PST there is only one year of historical data compared to eight for public sewers. This means that there is less information in which to find patterns of where blockages have occurred.

Sewer Level - Blockages by Cause

The following section gives the results from the models built to predict the different causes of blockages.

Table 7: Results of the single decision tree models for the different causes of blockages.

Model	Accuracy	AUC
Blockages due to silt	65%	0.62
Blockages due to debris	60%	0.68
Blockages due to nappies, wipes and rags	54%	0.65
Blockages due to fat, oil and grease (FOG)	65%	0.66
Blockages due to other causes	63%	0.67

Blockages due to Silt Figure 41 shows the model for blockages due to silt. The overall performance is poorer than the overall models of section 3.3.2, with a poorer AUC and greater overfitting to the training dataset.

The ROC curve shows an AUC of 0.62 and a fairly even shape for the training dataset. However, there is significant overfitting, which shows greater difficulty in predicting the likelihood of these blockages. Using only blockages due to a

particular cause limits the size of the incident dataset from which to build models. There is also believed to be inconsistency in the classification of the cause of blockages across DCWW's region. This inconsistency will limit the explanatory capability of the model.

The model for silt shows sewer length, properties per sewer metre and sewer age near the top of the tree. Each of these make logical sense for influencing the likelihood of blockage. Longer lengths of sewer present greater opportunity for silt to settle out along the length, without any additional flows helping transport material. The date of installation will represent the risk of defects interrupting flow, or variation in design and build standards affecting silt deposition. Properties per sewer metre will represent a load on the sewer and the potential for connections being present which interrupt flow. While the relationships make logical sense, the overfitting of the model limits how applicable these relationships will be.

Blockages due to Nappies/Wipes/Rags The model for nappies, wipes and rags (Figure 42) shows reasonable overall performance. There are some useful relationships in the decision tree and no overfitting to the training dataset.

The ROC shows an even shape, with no overfitting to the training dataset and an AUC of 0.65. In addition to the problems of a smaller size of dataset and inconsistency in cause classification, this type of blockage will be acute. Nappies, wipes and rags are likely to suddenly block a sewer without any slow build-up of material. This could lead to a large random element to the likelihood of these blockages and make predicting their likelihood more difficult.

Properties per sewer metre and sewer velocity form the first splits in the tree. Nappies, wipes and rags are likely to enter the sewer from properties, so more property connections will increase the likelihood of these blockages. Greater sewer velocity will increase the likelihood of material being transported through the sewer. This will reduce the likelihood of nappies, wipes and rags becoming trapped in the sewer. These relationships make logical sense and could be useful for understanding the likelihood of these blockages.

Blockages due to Fat, Oil and Grease (FOG) The model for blockages due to FOG (Figure 43) shows reasonable performance, with limited overfitting and some logical relationships.

The AUC is 0.66, which is similar to the models for all blockages. However, there is a small amount of overfitting to the training dataset.

The decision tree shows sewer length, properties per sewer metre and sewer velocity as the most important explanatory variables. Sewer length will represent the likelihood of FOG settling out, as it passes along the length of the sewer. Properties per sewer metre will represent a potential load of FOG and the disruptions caused by sewer connections. Sewer velocity represents a transport capacity and the likelihood of the flow carrying the FOG through the sewer. The food producers per sewer metre area is also present in the decision tree. This will again, particularly for FOG, represent a potential load on the sewer. Its presence indicates its use in understanding blockage likelihood for FOG.

Blockages due to Debris The model for blockages due to debris (Figure 44) shows reasonable overall performance, with little overfitting. However, the relationships in the decision tree are of less use.

The AUC is 0.68, which is similar to that of the overall models for blockages. The ROC curve also shows an even shape, with little overfitting to the training dataset.

The decision tree shows fewer useful relationships. The variables appearing are properties per sewer metre, sewer length and the catchment. Properties per sewer metre is useful for prioritising proactive maintenance and has logical reasons for influencing blockage likelihood. Sewer length, as discussed previously, is of less use for prioritising proactive maintenance. The appearance of catchment name would appear to suggest the inconsistencies in cause classification. For sewers of greater than 0.01 properties per sewer metre, the likelihood of debris blockages varies with the catchment. Given the large size of each catchment, it would seem unlikely that the likelihood of debris blockages would vary between each. This relationship is therefore likely to be due to inconsistent methods for

classification. Based on these points, the relationships in the decision tree are of less use.

Blockages due to 'Other Causes' The model for 'other causes' (Figure 45) shows reasonable performance, with no overfitting. The relationships would appear to be of less use.

The ROC shows no overfitting to the training dataset and an AUC of 0.67. Given the nature of this category, it is more difficult to interpret the relationships in the decision tree. The variables at the top of the tree are properties per sewer metre and sewer length. Properties per sewer metre would seem to reflect the efficacy of this variable shown in the overall models of blockages. Sewer length is of use for predicting blockage likelihood here, but is of less use for prioritising proactive maintenance.

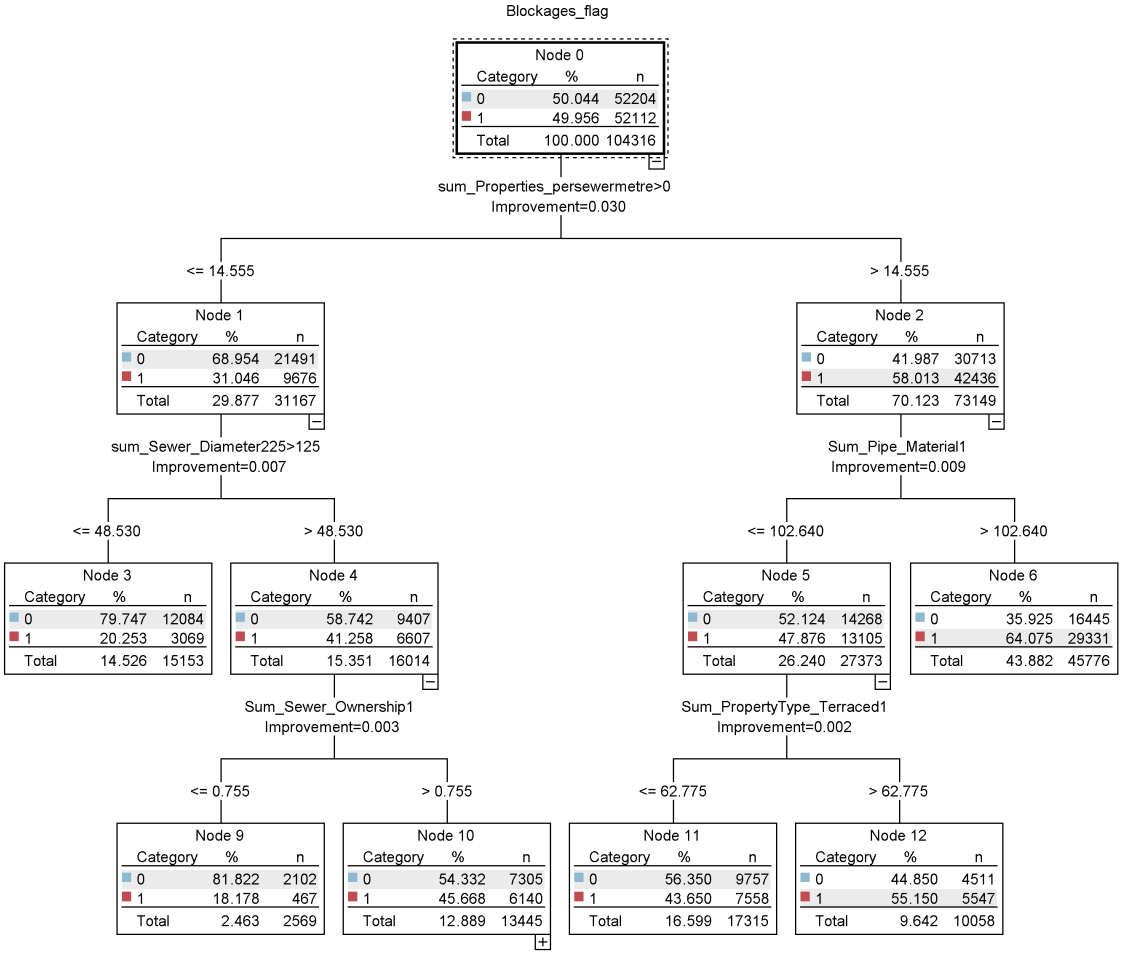
3.3.3 Data Mining - Area Level

The following section gives the results for the area level models. The input features were formed from aggregating the sewer level inputs used for the models in the previous section (3.3.2).

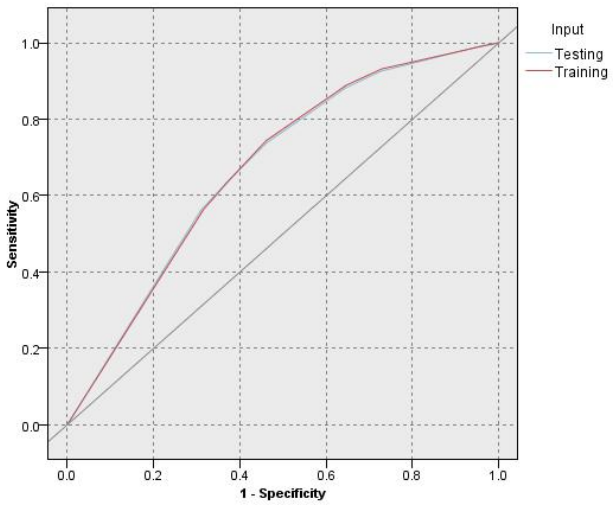
Initial Modelling

The performance of the aggregated model produced is shown in Figure 13 and is similar to that of the sewer level models. The AUC is 0.67, which compares to the AUCs of between 0.65 and 0.72 for the sewer level models. The aggregation was expected to improve performance, but this is not shown in this model. The variables present within the decision trees are also similar to those of the sewer level models.

The variables shown at the top of the decision tree are those of properties per sewer metre greater than zero, sewer diameter between 225mm and 125mm and sewer material type 1 (vitreous clay). These variables are similar to those seen in the sewer level models, where properties per sewer metre and sewer diameter were both common, although sewer material was less common.



(a) Decision Tree, showing the top four layers of the tree.



(b) ROC curve

Figure 13: Results obtained from the area level models, using a threshold in the relative blockage proportion of 1.

The ROC curve shows a flat initial part to the curve and poor performance on the highest likelihood areas. With the lack of overall increase in AUC this shows that there is no performance improvement from this model.

Differing Thresholds

Table 8: Table showing the results of the area level models, using different thresholds in the relative blockage proportion

Threshold	AUC
0	0.68
1	0.67
4	0.65
6	0.65
8	0.69

Table 8 shows the results obtained when the threshold used to define the blockage flag was varied. The results are all within a similar range as the sewer level models, with AUCs ranging between 0.65 and 0.69 for these models compared to 0.65 to 0.72 for the sewer level models. The best results were achieved at the highest (8) and lowest (0) thresholds. At the threshold of 8, the proportion of areas flagged as 1 is only around 3%, while for threshold 0 it is around 25%. The variables shown in the decision trees are again similar to those in the sewer level models. Important variables included properties per sewer metre, sewer velocity and sewer length.

The ROC curves for these models are shown in the Appendices (section D). The curves show a fairly even shape, although a generally flat initial part to the curve, again indicating poorer performance for the higher likelihood areas. The results for threshold 8 (Figure 50) are slightly different. The model shows greater overfitting to the training dataset, but also a steeper initial section to the curve, indicating better performance on the higher likelihood areas.

Discussion

The grouped approach was expected to improve performance, but this was not realised. By grouping the sewers, any noise in the data from the assignment of

blockages to a sewer and in the sewer characteristics should be smoothed out in the groups. For example, there is a stochastic element that material travelling down a sewer may cause a blockage in any of the sewers. On which sewer the blockage occurs may be largely random, rather than represented by any characteristics of the sewer. This noise in the blockage rate on each sewer will be smoothed out in the aggregation, potentially better representing the variation in blockage rate. The effect of any inaccuracies in the assignment of incidents to assets may also be reduced by all of the sewers in the area being represented within the same group.

The grouping of sewers used was based on geographical areas. However, this may not represent the best grouping for modelling. Sewers which aren't in directly connected parts of the network could be in the same postcode and therefore grouped. This may limit some of the expected benefit from reducing noise. The groups may also not represent sewers which show similar characteristics or similar historical blockage rates. The variation in the grouped variables is difficult to represent when represented by a single value in the dataset. This may also affect the value derived from this approach.

3.4 Chapter Summary

In this chapter models at a sewer and area level, and for different causes of blockages have been developed. Data was taken from the corporate systems of DCWW, assessed for quality, prepared and analysed before being used to develop the models. The data included sewer characteristics such as length, diameter and material, property locations, property types and ages, and derived variables such as sewer velocity. At a sewer level, decision trees were used with this data to predict a blockage flag, indicating whether a sewer had blocked during the period of historical data available. Models of the different blockage causes were produced to predict a blockage flag indicating whether a sewer had blocked by that cause. The sewer level data was then aggregated to an area level and models produced. Postcode was used to form the geographical groups modelled,

with larger postcodes broken up and smaller postcodes combined. The sewer and area level models are of reasonable accuracy, providing a likelihood of blockage output and informing the importance of explanatory factors. The models of the different causes of blockages gave some useful relationships but inconsistency in the datasets limited predictive performance. Given the greater resolution of the sewer level models, these are the best performing from this part of the work. In the next chapter these models will be developed with the aim of improving their performance. The two methods investigated in the next chapter will be ensemble techniques and the derivation of an input feature from a sewer's blockage history.

Chapter 4

Blockage Likelihood Prediction Models using Ensembles of Decision Trees and Historical Input Features

In Chapter 3 models were developed at a sewer and area level. Given that WaSCs desire information at the greatest resolution and there was little difference in the performance of the sewer and area level models, it was decided to use the sewer level models for further development. In this chapter two approaches are applied to improve performance. The first uses ensembles, producing multiple models and combining the outputs from each. The second investigates the inclusion of an input feature based on the blockage history of each sewer.

4.1 Methodology

The two extensions to the work described are the building of ensemble of decision trees and the addition of the historical input feature to the models. The methodology used in each case is described below.

4.1.1 Ensembles

Ensemble techniques [17] generate many models, combining the outputs from each into a single model. This was found to be more accurate than the individual models. Dietterich [17] describes how ensembles can improve performance of models, the main reasons being statistical, computational and representational.

For example, in the computational reason, it can be difficult for some algorithms to find the best model. The methods of developing the model can mean the output becomes limited to a local optimum. Producing many models, from different parameters, and combining the result can give a better model than any of the individual models. Cutler *et al.* [18] applied an ensemble method, Random Forests, to ecology data. Random Forests showed slightly or substantially better performance than decision trees or alternative linear methods in the cases investigated. To give the best performance, the performance of the individual models must be maximised while the correlation between them is minimised [17]. There are many ways to generate the ensembles, which include:

- Manipulating training data
- Manipulating the input features
- Manipulating the output features
- Adding randomness

The approach taken here produced the individual models from a selection of the input features. Different methods of selecting and combining the variables were investigated, as well as methods of combining the model outputs.

Derivation and selection of variables

To produce the variables, either random input [19] or random combination [19] methods were used. For the random input, a random selection from the full set of input features were taken and used to produce the individual models. For the random combination, a random selection was taken and combined to produce additional features then used to produce the models. To process categorical variables, a category was chosen at random and a binary flag derived for whether each record was a member of that category. Categorical variables were made more likely to be chosen by a factor of the number of categories. To process continuous variables, the average and standard deviation for the variable in the training dataset were used to derive z-scores for each variable. The variables

were combined by summation, with a weighting applied to each, the coefficient being randomly chosen from the range -1 to 1.

Number of variables used

For each type of query (random input and random combination), a different numbers of variables used for selection and combination were investigated. For random input, this gave the number of variables selected. For random combination, this gave a number of variables selected and then a number of output variables produced from this selection.

Combining Model Outputs

From the individual decision trees produced, the outputs were combined in different methods. The first of these, voting, averaged the 0 or 1 classification output from each of the individual models. The second, average raw propensity, averaged the raw propensity scores from each model to produce the overall model score.

Models Produced

To evaluate the approach ensemble models were produced for the public, combined subset of the network as this showed the best combination of performance and useful relationships from the sewer level models. The models allowed assessment of the overall improvement in performance and the effect of each of the different methods used.

4.1.2 Historical Input Feature

The aim was to investigate the addition of an input feature based on historical incidents. The effect of the number of years of data and use of an input feature was evaluated. Eight years of historical data had been prepared for the modelling and these were used to derive the input and output features in the models. The following sections outline how the input and output features were derived and the

Table 9: Years of historical data used in each model in investigating increasing years of available historical data

Model Number	Years of Data
1	04/2006 – 04/2014
2	04/2007 – 04/2014
3	04/2008 – 04/2014
4	04/2009 – 04/2014
5	04/2010 – 04/2014

three experiments undertaken.

Deriving Input and Output Features

The output feature to be predicted was the blockage flag, as was used in the previous stages of modelling. Previously, this flag defined whether the sewer had blocked at any point in the years of incident data available. At this stage different selections of years were made and these used to define whether the sewer had blocked in this time period. This leaves some of the incident data to form an input feature, again formed from different selections of years. The input features derived were a blockage flag, whether the sewer had blocked, and a blockage rate, defined as the number of blockages per year, per km of sewer length.

Increasing Years of Input Data

This part of the investigation used differing amounts of input data to build the models. The aim was to understand potential future changes in performance, when increasing amounts of data are available.

For the models, the output feature was formed from an increasing number of years of data, as shown in Table 9. To form the training and testing partitions, a random selection of sewers were taken from the full set of sewers. The models were produced for the public, combined subset of sewers.

Varying Input and Aggregated Output

Here we investigated how the size of the input and output features affect performance. The full eight years of data were used to form inputs and outputs using

Table 10: Table showing the models produced to investigate varying the size of the input and output features

Model Number	Years of Historical Incident Data		
	As Input	As Output	For Testing
1	1 — 4 04/2006 — 04/2010	5 — 7 04/2010 — 04/2013	8 04/2013 — 04/2014
2	1 — 5 04/2006 — 04/2011	6 — 7 04/2011 — 04/2013	8 04/2013 — 04/2014
3	1 — 6 04/2006 — 04/2012	7 04/2012 — 04/2013	8 04/2013 — 04/2014

varying amounts of historical data. The models produced are shown in table 10

The models in Table 10 show an increasing number of years used for the input feature and reducing number for the output feature. All of the models are tested on the same year of data. For these models the training and testing datasets both contained all of the sewers. The models were trained using 'As Output' and then tested using 'For Testing'. The models were produced for the public, combined and public, foul subsets of the network.

Windowing

Table 11: Table showing the models produced for the windowing investigation

Model Number	Years of Data (N)	Years of Historical Incident Data		
		As Input	As Output	For Testing
1	2	5 — 6 04/2010 — 04/2012	7 04/2012 — 04/2013	8 04/2013 — 04/2014
2	2	— —	5 — 7 04/2010 — 04/2013	8 04/2013 — 04/2014
3	4	3 — 6 04/2008 — 04/2012	7 04/2012 — 04/2013	8 04/2013 — 04/2014
4	4	— —	3 — 7 04/2008 — 04/2013	8 04/2013 — 04/2014
5	6	1 — 6 04/2006 — 04/2012	7 04/2012 — 04/2013	8 04/2013 — 04/2014
6	6	— —	1 — 7 04/2006 — 04/2013	8 04/2013 — 04/2014

The windowing approach investigates different amounts of available data and the effect of the historical input feature. The approach uses the differing amounts

of incident data and, for each, compares using the data to form the output feature, and to form both an input and an output feature.

The models to be produced are shown in Table 11. For each set of available data two models are produced: with a historical input feature and without a historical input feature. The other inputs to the models remain the same in each. When a historical input feature is included:

- As Input: Additional input features of a blockage rate and blockage flag are derived from *N* years of historical incident data
- As Output: The model is trained by predicting the blockage flag formed for the year of incident data *N+1*
- For Testing: The model is tested on the blockage flag formed for the year *N+2*.

Without an input feature:

- As Input: There are no additional input features derived from historical incident data
- As Output: The model is trained by predicting the blockage flag formed from *N+1* years of incident data
- For Testing: The model is tested on the blockage flag formed for the year *N+2*

For these models the training and testing datasets both contained all of the sewers. The models are trained using 'As Output' and then tested using 'For Testing'. The year of incident data forming the output for testing remains constant in each of the models. The models are produced for the public, combined and public, foul subsets of the network.

4.1.3 Validation

The aim was to evaluate the potential benefits of the models and validate the approach. Two validation datasets, not originally used, were sourced: the blockages

which occurred in the reporting year 2014-5 and results from on-site surveys. Gain curves are used to estimate how the model's likelihood score output could prevent blockages, as a function of the length of sewer surveyed. An estimation of the cost savings for 2014-5 was also made, based on figures provided by DCWW.

Three likelihood scores were used to evaluate and compare performance:

- the output from the decision trees
- the historical blockage rate (per year per km of sewer length)
- a combination of the two scores (adjusted likelihood score)

For the decision tree output, the best performing models were used. For the public sewers this was the models including the historical input feature. For the PST sewers there was insufficient data to derive the historical input feature so the best performing models were from the initial stage of modelling. The historical blockage rate is based on the full eight years of available blockage data (2006 - 2014) and was normalised per year, per km of sewer length. The adjusted likelihood score was derived to combine these two outputs. The scores give a greater weighting to sewers which have suffered a blockage. For sewers which showed a history of blockage, the historical blockage rate was the score, while for sewers which showed no history, the output from the models was used. The historical blockage rate was added to one in the score, so that all of these sewers were given a higher weighting than those which had not suffered a blockage.

Data - Survey Results

The survey results evaluated the model performance using issues which had been found on on-site surveys. The dataset is based on a geographical area in which all of the sewers were surveyed, with records kept of problems found on any sewers. The survey results were used to compare the model output and the historical blockage rate and evaluate overall model performance. For evaluation, an ROC curve was produced using each likelihood score. The positive events

were defined as sewers which had suffered an operational issue which required response, such as jetting or rodding of the sewer.

Data - Incidents 2014-5

The incidents from 2014-5 have been used in the production of gain curves and in estimating cost savings for the year 2014-5. The incidents included within this are blockages, flooding and pollution, based on DCWW's regulatory return data. For flooding, there is a further classification for incidents being due to hydraulic overload or due to other causes (such as blockages or collapses). Only flooding incidents due to other causes have been included in these estimations.

The incidents were used to estimate the prevention of blockages for the range of sewer length maintained. There is no data on the effectiveness of proactive maintenance and how long they prevent or reduce the likelihood of blockages for. It has therefore been assumed that sewers defined as being surveyed have been prevented from suffering blockages for the year. Included in this, is that if a sewer suffers multiple blockages in a year then all of those blockages will have been prevented. This gives a proportion of sewers on which blockages have been prevented, from which the proportion of blockages prevented can be found. These assumptions will give an over-optimistic estimate of the benefit of proactive maintenance, but without the data on proactive maintenance effectiveness it is not possible to develop this further.

Due to the poor level of assignment of incidents to assets, the incidents included and proportions quoted are based on using those incidents which have been assigned to an asset. For the cost savings, the calculation of the number of incidents prevented is based on calculating a proportion of incidents which have been assigned and scaling this up to the proportion of all incidents.

Gain Curves

The gain curve plots the proportion of incidents prevented against the proportion of the length of sewers. The incidents from 2014-5 were used to plot curves

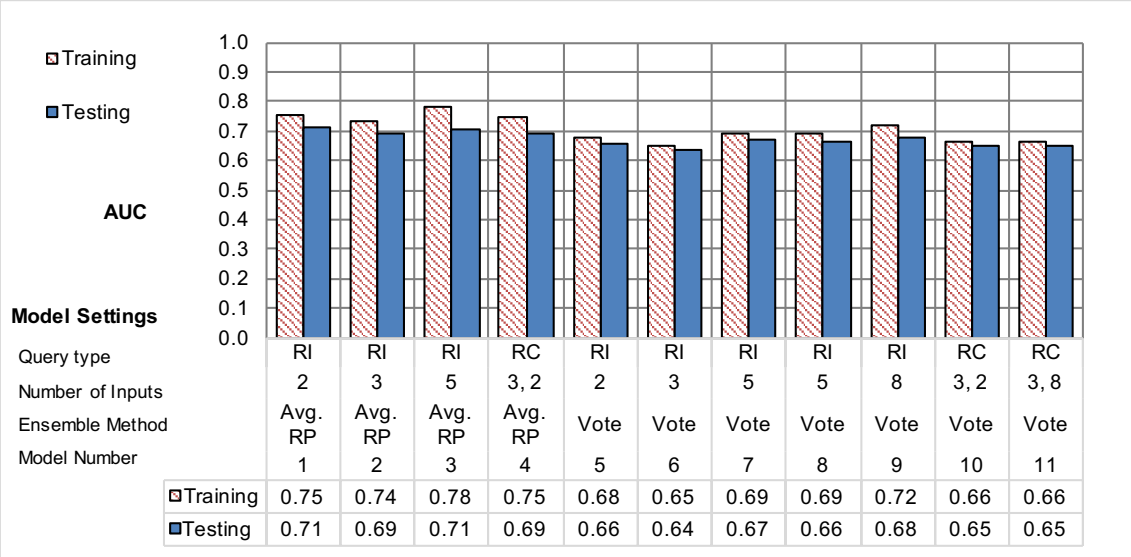
for each type of incident, to compare the three likelihood scores and the overall performance of the models.

Incident Prevention Estimation

Results for incident prevention are based on figures provided by DCWW related to the current levels of sewer surveying and the calculations used to derive the gain curves. Using these, it is possible to estimate the proportion of each incident type which would have been prevented, based on this level of surveying.

4.2 Results and discussion

4.2.1 Ensembles

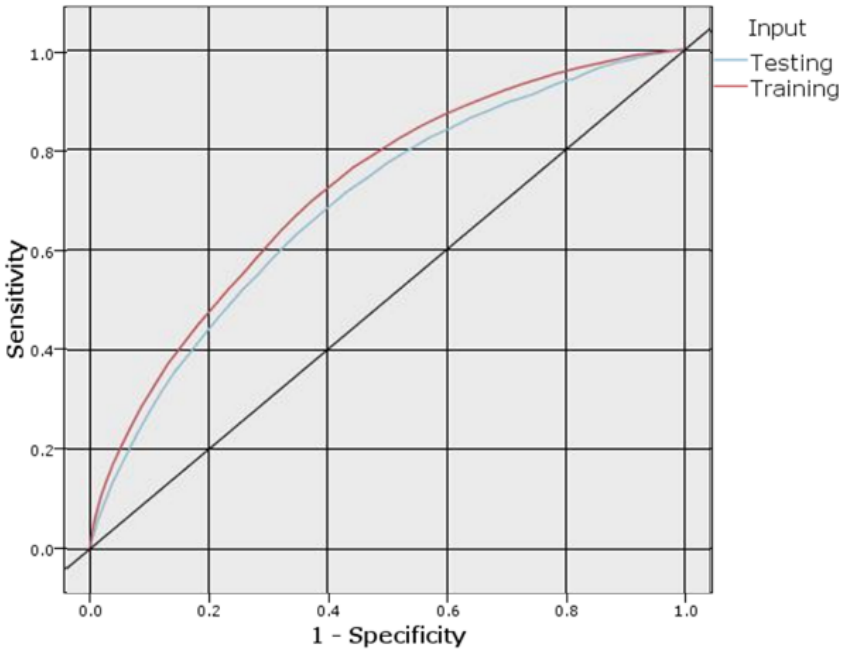


Model Settings											
Query type	RI	RI	RI	RC	RI	RI	RI	RI	RI	RC	RC
Number of Inputs	2	3	5	3, 2	2	3	5	5	8	3, 2	3, 8
Ensemble Method	Avg. RP	Avg. RP	Avg. RP	Avg. RP	Vote	Vote	Vote	Vote	Vote	Vote	Vote
Model Number	1	2	3	4	5	6	7	8	9	10	11
Training	0.75	0.74	0.78	0.75	0.68	0.65	0.69	0.69	0.72	0.66	0.66
Testing	0.71	0.69	0.71	0.69	0.66	0.64	0.67	0.66	0.68	0.65	0.65

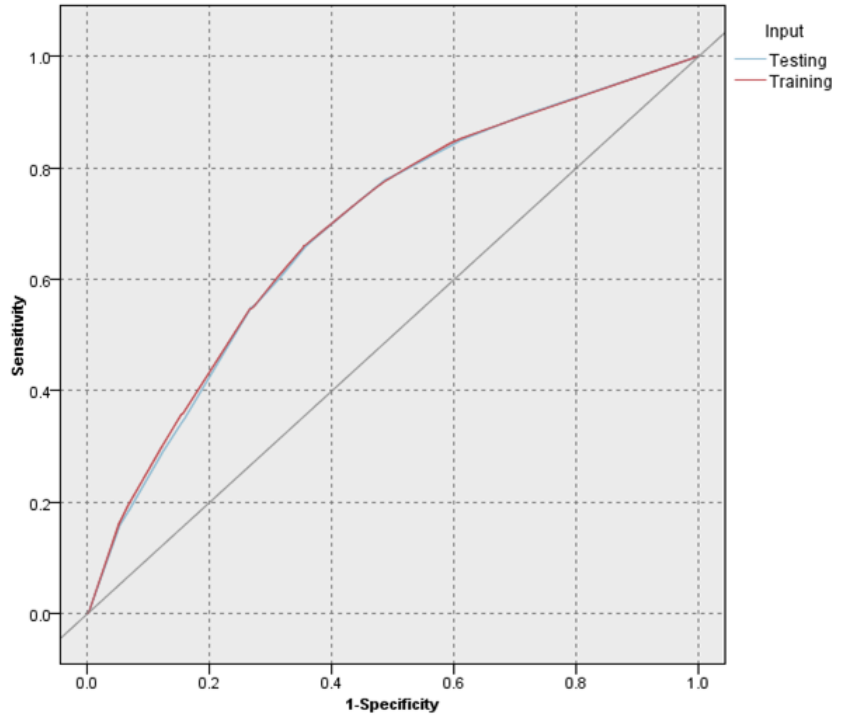
Figure 14: Overall results of the modelling using ensembles. The figure shows the parameters used to produce each model and the AUC for training and testing.

Figure 14 shows the performance of the models produced using the different settings of the variables listed above.

Overall there is no significant improvement in performance over the best single decision tree, which had an AUC of 0.69. A few models show slightly improved performance, but also overfitting to the training dataset. Those showing less overfitting also show poorer performance overall. For example, models one and three



(a) ROC curve of the best performing ensemble model



(b) ROC curve for the best performing single decision tree model

Figure 15: ROC curves comparing the best performing ensemble and single decision trees

show testing AUCs of 0.71 but overfitting. Model 11 shows little overfitting, but poorer performance compared to the previous best performing model. Figure 15 shows the ROC curve for model 1 (15a), the best performing ensemble model, with the ROC curve for the best performing single decision tree (15b). Both show similar, even shapes with little performance improvement for the highest likelihood sewers.

The following paragraphs evaluate the effect of the different model building parameters.

Number of Inputs

Models one to three and five to nine show a comparison between models produced using the same parameters, but with the number of inputs changed. These show only a small variation in overall performance and little effect of the number of inputs. The results do seem to show an increase in overfitting as the number of inputs is increased, as would be expected. In the ensemble the aim is to maximise the performance of the single models, while limiting the correlation between them, to maximise overall performance [19]. The number of inputs will affect the overall performance and correlation between models. However, here there does not seem to be an optimal point, giving the balance between these two effects.

Query Type

The query type, random input or random combination, does not appear to significantly change performance. Models 2 and 4 and 6, 10 and 11 were produced using the same inputs, except that 4, 10 and 11 used the random combination method. Comparing the two sets of results shows little change in the AUC for the testing dataset.

The random combination method was designed to increase number of input features where a low number exist [19]. This aims to increase the strength of the models, while preventing correlation between. There is a relatively large number of input features available to the models. This may mean that the effect of the random combination method is limited.

Model Combination Method

The ensembles produced using the average raw propensity (ARP) show generally better performance than those using voting. However, these models also show greater overfitting to the training dataset. Models 1 to 3 and 5 to 8 allow comparison between models produced using the same parameters. In voting, the same weight is given to the outputs of each model, where the ARP models are weighted by their performance. This may mean that by giving greater weight to the better performing models, the ARP models show better performance. For each individual model there is no pruning applied. The better performing models may show larger decision trees and greater overfitting to the training dataset. This would mean the ARP models show greater overfitting when compared to the voting models.

4.2.2 Historical Input Feature

The results from the three parts of the investigation are shown in the sections below, with the overall discussion of the results given following the results.

Results - Increasing Years of Input Data

Figure 16 shows the results of the models built with increasing amounts of historical data. The results show very similar performance for each model, with only 0.005 separating the best and worst test performances. This shows that increasing the amount of historical data does not necessarily improve performance. Investigating the ROC curves also shows very similar trends for each of the models, with no variation in shape across the decision trees.

Results - Varying Input and Aggregated Output

Figure 17 shows the results from the models produced. Overall, the AUC shows little change as the historical data is varied between forming the input or the output feature. Compared to the best performing models in Chapter 3 there is an increase in performance. For public, combined sewers the AUC increased from

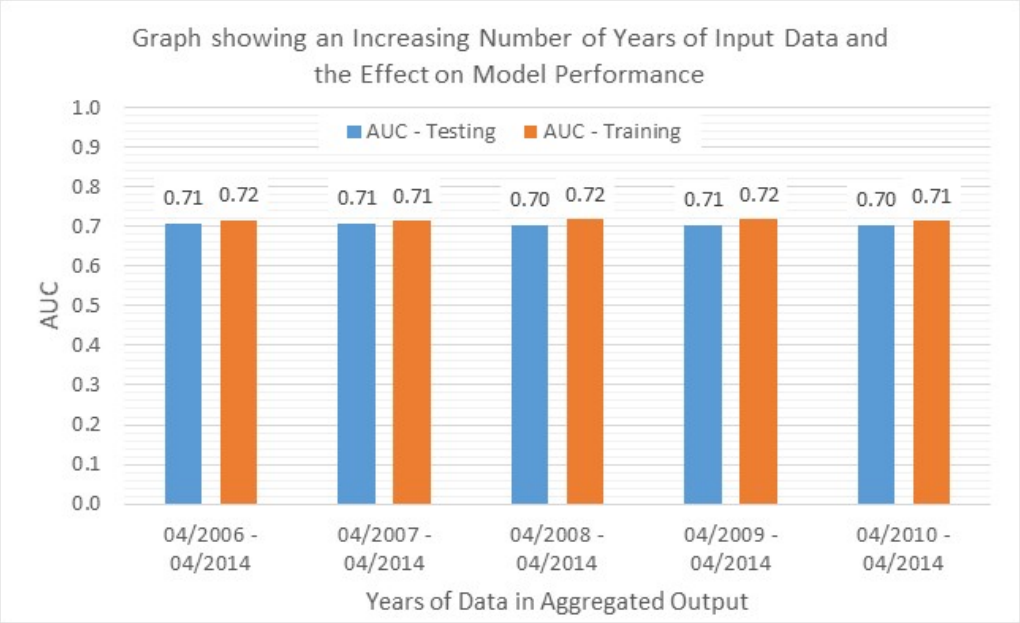
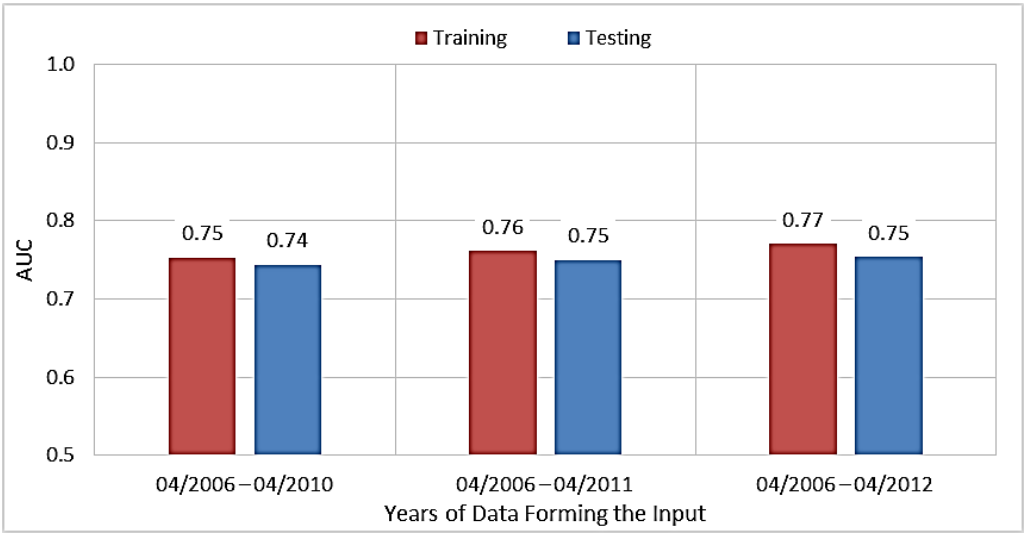


Figure 16: Results of the models built using increasing amounts of historical data in terms of the AUC

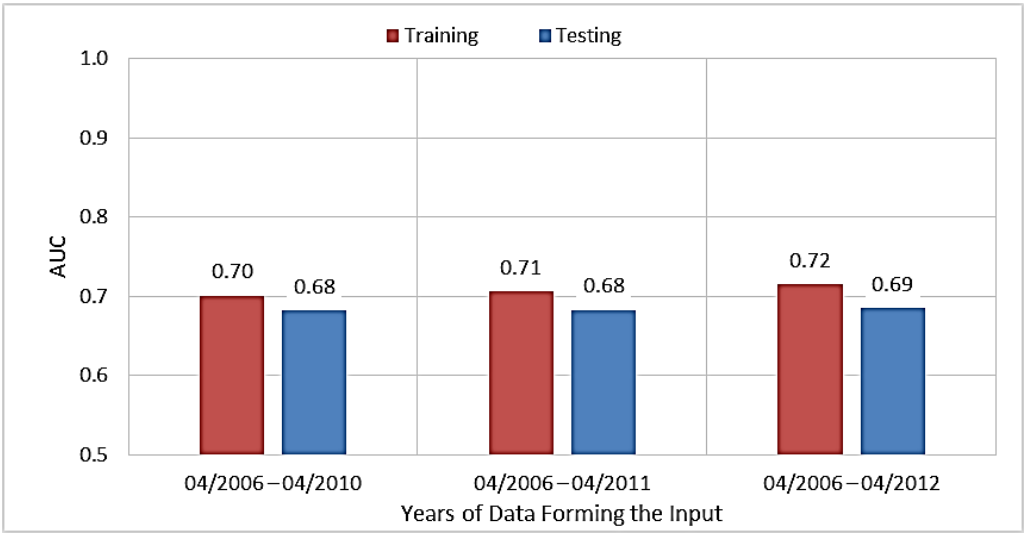
0.69 to 0.75 and for public, foul sewers the AUC increased from 0.66 to 0.69. Figure 18 shows the ROC curves for model 1 and model 3 for the public, combined network. Both ROC curves show an initial steep part to the curve, showing better performance for the highest likelihood sewers. The ROC curve for the public, foul network shows a similar shape.

Results - Windowing

Figure 19 shows the overall results for the public, combined and public, foul parts of the network. The results show little change in model performance as the number of years of incident data is increased. Figure 19b for public, foul sewers shows slightly better training and testing performance for those models including an input feature. While Figure 19a for public, combined sewers shows slightly better performance for the training dataset but not the testing. Figure 20 shows the ROC curves from four of the models, showing models built using two and six years of data, for those with and without the historical input feature. Comparing the models with a historical input feature (Figures 20a and 20c) to those without (Figures 20b and 20d) shows a steeper initial part to the curve. This indicates better performance for the highest likelihood sewers and matches with the results



(a) Results for the subset of sewers on the public, combined network

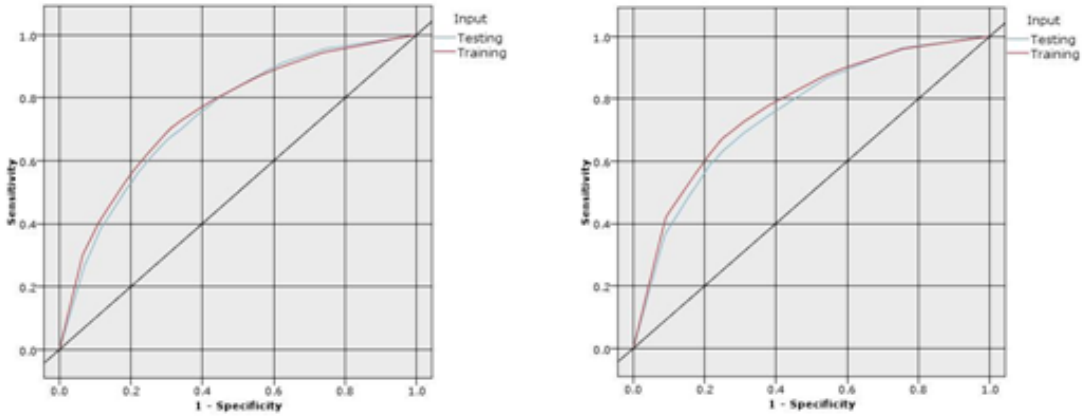


(b) Results for the subset of sewers on the public, foul network

Figure 17: Overall results of the models built in terms of AUC

shown in the investigation varying the input and aggregated output. The ROC curves for the public, foul network match these, with a steeper initial part to the curve, without the slightly improved overall performance. The effect of the number of years of incident data, for those including a historical input feature, can also be seen from the ROC curves by comparing the models built using two years of data (Figure 20a) to those built using six years of data (Figure 20c). The ROC curves show varying length and gradient to the initial part to the curve. The model built with two years of data shows a steeper but shorter section when compared to the model built using six years.

The decision trees (Figure 51) themselves can be investigated to understand



(a) From model 1, produced using 04/2006 to 04/2010 to form an input

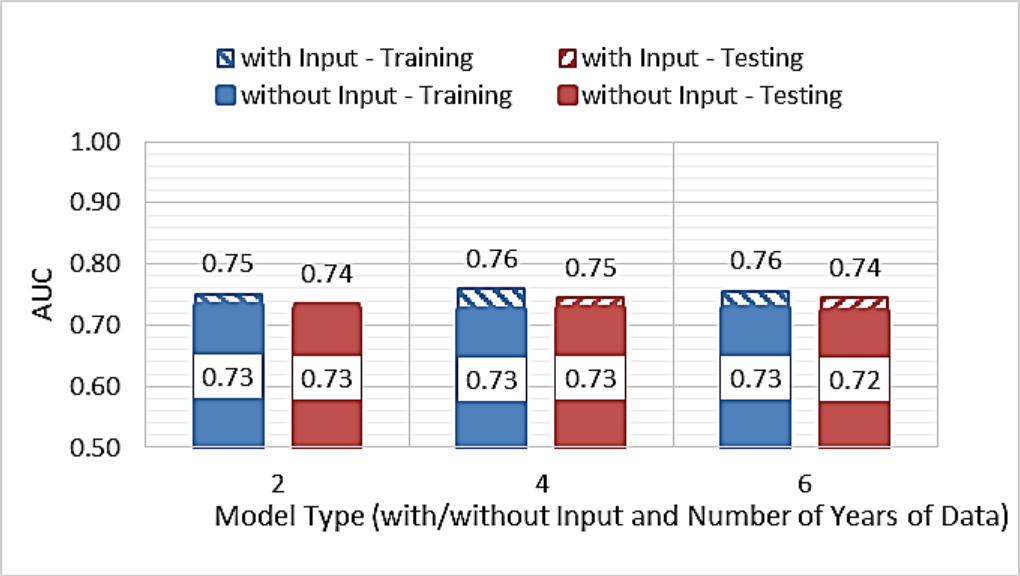
(b) From model 3, produced using 04/2006 to 04/2012 to form an input

Figure 18: ROC curves from two of the models for public, combined sewer

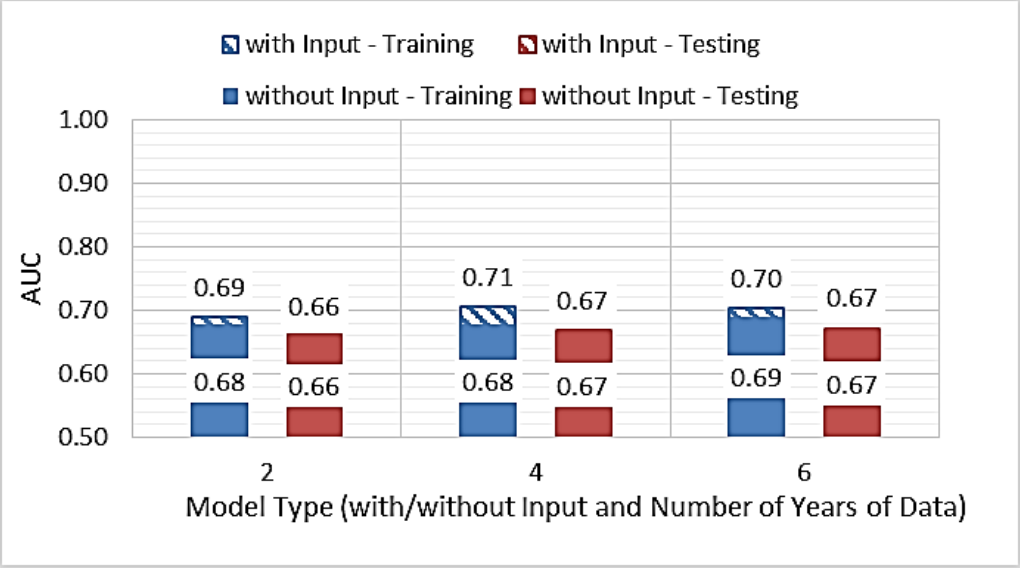
how important the input feature is. The historical input feature forms the first split in the decision tree for each model including it. This represents the most important variable for explaining the likelihood of blockage. The variable used is also the blockage flag rather than the blockage rate.

Discussion

The inclusion of a historical input feature shows improvements over models without this feature. For some models this includes overall improvements in AUC, while all show performance improvements for the highest likelihood sewers. This could suggest either that: there are features of these sewers not represented in the other variables, or that the occurrence of blockages influences the likelihood of another blockage. Repeated blockage locations could be due to the particular load on the sewer, for example from customer behaviour, or characteristics of the sewer, for example defects which are present. Blockages due to nappies, wipes and rags would be expected to occur on smaller diameter sewers as a result of customer behaviour. If the behaviour is repeated, then the likelihood of another blockage will remain and increase the likelihood of a repeated blockage location. Fat, oil and grease (FOG) or silt blockages, where behaviour disposing of FOG or infiltration of silt, will mean a blockage remains likely to occur. Sewers could also have construction defects or damage which make them more likely to suffer



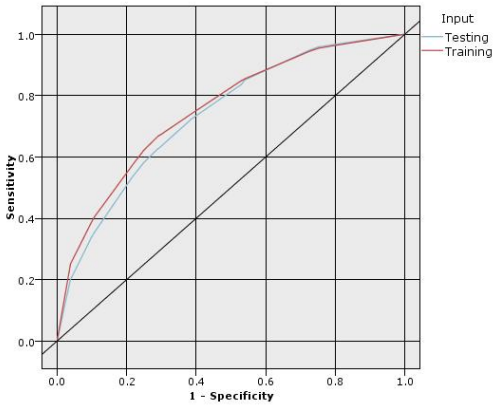
(a) Results for the subset of sewers on the public, combined network



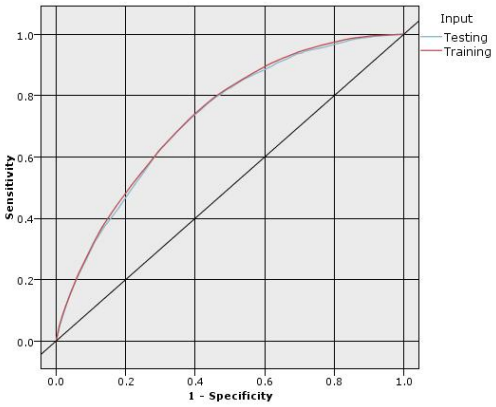
(b) Results for the subset of sewers on the public, foul network

Figure 19: Overall results of the models built in terms of AUC

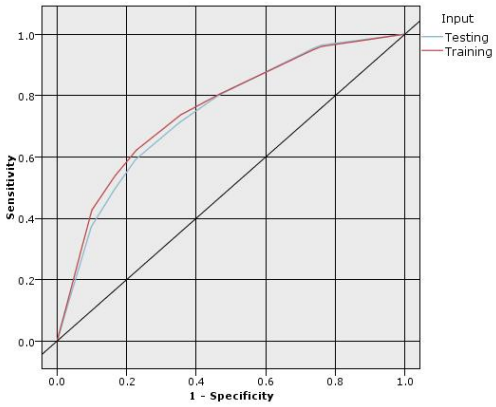
from blockages. Sewer condition is measured using CCTV surveying and could be used to explain blockage likelihood. However, there is a small proportion of the network which has been surveyed and condition changes over time limit the period for which the survey is relevant. This means the data was not included in the models developed, but could be part of the explanatory capability offered by the historical input feature. WaSCs make interventions to prevent blockages and are likely to target repeated location or hotspots. These interventions aim to reduce the likelihood of blockage in that location and potentially reduce the efficacy of the historical input feature. The likelihood of repeated locations could be



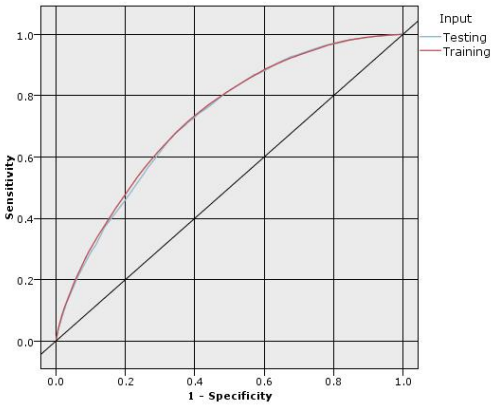
(a) ROC for model 1 built using two years of data, including an input feature



(b) ROC for model 2 built using two years of data, without an input feature



(c) ROC for model 5 built using six years of data, including an input feature



(d) ROC for model 6 built using six years of data, without an input feature

Figure 20: ROC curves from four of the models built on the public, combined part of the network

investigated to understand whether and how blockages affect the likelihood of a repeated location and for how long this effect exists. Attempts to derive an estimated condition could also be made. Using existing condition surveys and the characteristics of the sewers, an infilled condition grade could be developed and tested for efficacy.

Increasing the amount of available data does not seem to improve performance significantly, whether a historical input feature is present or not. There are different explanations for this for those models with and without the input feature.

Without the historical input feature, more years of incident data gives more data in which to find patterns of blockage occurrence. It was therefore expected that more years of incident data would improve performance. The lack of improvement could be due to changes in the data limiting the relationship between the most recent incidents and those more historical. Changes in reporting or operational practices over the period could both affect the data. The lack of improvement could also be from the use of the blockage flag rather than the rate altering the effect of the additional incidents. For the blockage flag, more blockages on the same sewer won't affect the value of the output feature. Whereas the blockage rate will be affected by further, or lack of further, incidents and could mean performance improvements when more years of data are used.

With a historical input feature, the benefit from further data will also depend on the likelihood of repeated blockage locations and for how long the effect lasts. The benefit of models with a historical input feature over those without would suggest the likelihood of repeated blockage locations. The length of the effect was interrogated using the initial part to the ROC curves. All of the decision trees with a historical input feature have their first split as this feature. For the Windowing experiment all of the models are also trained and tested on the same datasets. The only change is the amount of data forming the input feature, allowing a comparison of the effect that this has. As more years of incident data are used, there will be more incidents and a greater proportion of sewers which have blocked in this period. This will increase the length of the initial part to the

curve. The height of this initial part to the curve will indicate the proportion of blockages occurring on sewers which have previously had blockages (in the time period of the input feature). If the historical incident data provided the same explanatory capability, then the gradient would be expected to stay the same. The proportion of repeated blockages would remain the same, even as the proportion of sewers suffering incidents increases. The values of these proportions have been calculated for comparison. As the number of years is increased from 2 to 4 to 6, the proportion of sewers suffering incidents increases from 2.5% to 4.2% to 6.1%. However, the proportion of these sewers suffering more than one blockage decreases from 14% to 12% to 11%. This shows that the sewers suffering blockages more recently have a higher proportion of repeated locations. This indicates that the more recent incidents better explain the likelihood of blockage than those less recent. The better performance of the more recent data could be due to a greater likelihood of the effect remaining. If the likelihood of a repeated blockage is because of customer behaviour or sewer defects then the more recent the incident, the more likely it is that this effect remains.

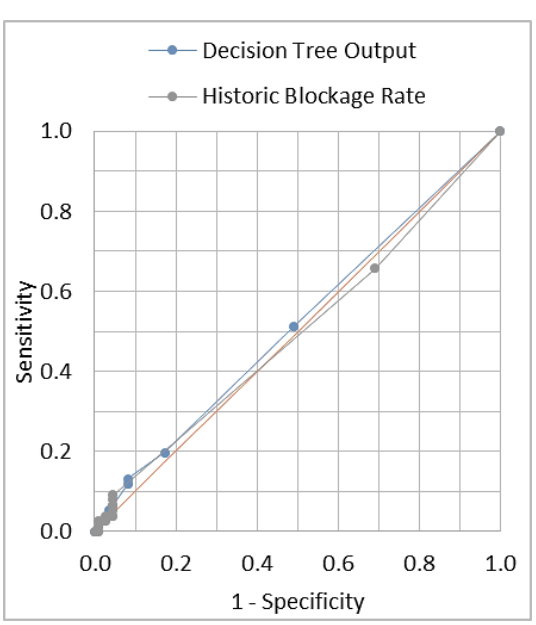
It may also be possible to derive further input features using differing amounts of historical incident data, which would add further explanatory capability. This could, for example, use all available historical data but weight more recent data more highly.

4.2.3 Validation

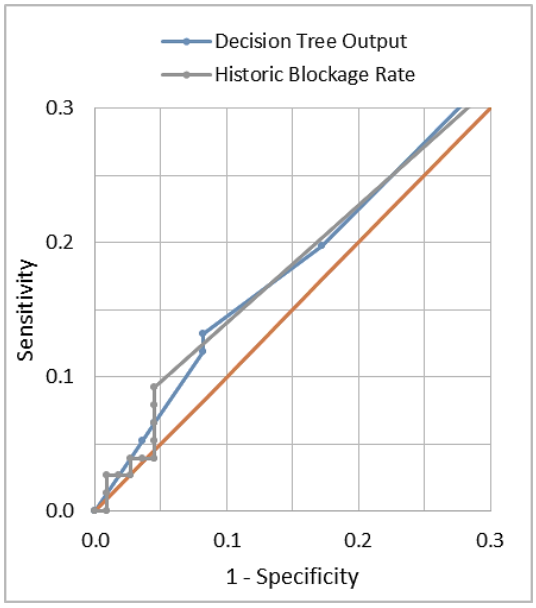
Survey Results

Figure 21 shows the ROC curves from the survey results, showing the output from the decision trees and the historical blockage rate. The results show similar and generally poor performance for both methods. Although both show best performance for the highest likelihood sewers.

The AUC for the decision tree output is 0.52, compared to 0.5 for the historical blockage rate. At the highest likelihood end of the curve, there is some deviation from 45°. Beyond this, the historical blockage rate is slightly worse than random



(a) Full ROC



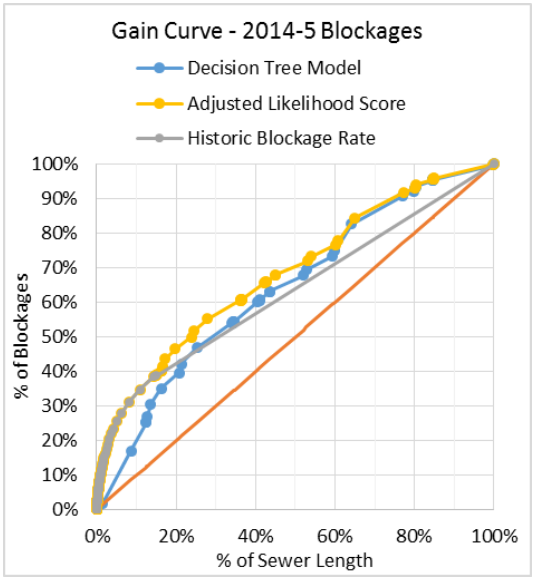
(b) ROC curve for the initial part of the curve

Figure 21: ROC curve constructed using the survey results and comparing the output from the models to the historical blockage rate.

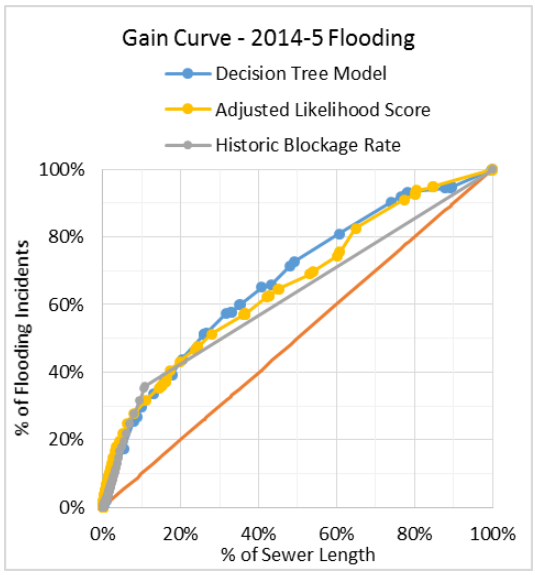
and the decision tree output slightly better than random. The dataset used to derive the results contains a survey of only 109 sewers, giving a small dataset for evaluating performance. The poorer performance may also result from the definition of a blockage found. From the survey data, issues which required an operational response were defined as blockages. This defines the build up of material as being the issue, rather than structural damage to the sewer. These operational responses will include blockages which would have caused issues and been recorded for regulatory return and partial blockages which wouldn't have caused an issue and been cleared without any intervention. This will mean the definition will vary between the incident used to build the models and calculate the historical blockage rate and the incidents to which they are being tested.

Incidents 2014-5

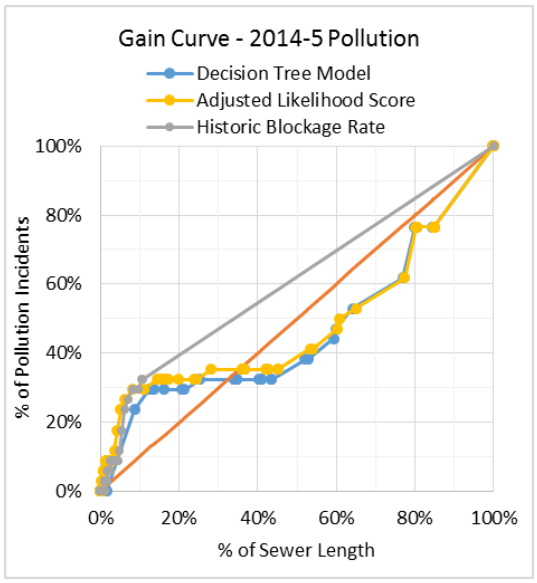
Figure 22 shows the gain curves produced for each incident type using the three likelihood scores derived (described in section 4.1.3). The results show good performance for blockages and flooding, with poorer performance for pollution incidents. In general, the adjusted likelihood score shows the best performance



(a) Gain curve for blockages



(b) Gain curve for flooding incidents



(c) Gain curve for pollution incidents

Figure 22: Gain curves showing the performance of each of the derived likelihood scores on the three types of incidents occurring.

from the three. The adjusted likelihood score is made up of the combination of the historical blockage rate and the outputs from the decision trees.

For blockages and flooding, the likelihood scores show good performance for the prevention of incidents. The graphs show a steep initial part to the curve with the curves maintaining divergence from the 45° for the remainder of the curve. The steep initial part shows good performance for the highest likelihood sewers, those most likely to be surveyed by WaSCs. For pollutions, the curve again shows a steep initial part, suggesting good prevention for the highest blockage

likelihood sewers. For the remainder of the curve the model outputs show poor performance. The historical blockage rate provides no further explanatory information due to the majority of pollutions occurring on sewers with no history of blockage. The random selection of sewers to form the remainder of the curve in the blockage rate output outperforms the other two outputs. While blockages are more likely to occur on smaller diameter sewers near customers' homes, pollution incidents are more likely to occur on sewers away from homes and closer to watercourses. Flooding incidents are also likely to occur near customers' homes, which may help explain this better performance.

Comparing the scores used, the adjusted likelihood score is the best performing. For blockages, the initial part of the curve shows better performance for the historical blockage rate, while the later part shows better performance for the output from the models. The adjusted likelihood score gives the performance benefits of each, matching the historical rate initially and then out-performing the model outputs. For flooding, all outputs show similar performance in the initial part of the curve. In the later part, the decision tree model shows the best performance, followed by the adjusted likelihood score and historical blockage rate. For pollutions, the initial part of the curve shows all three outputs with similar performance. For the later part none of the outputs perform very well. The remainder of the historical blockage rate is based on the random selection of sewers. The other two measures show performance plateauing and then performing worse than random. Using blockage and flooding incident performance, which would be best predicted from blockage likelihood models, the adjusted likelihood score gives the best performance. The score matches the historical blockage rate in the early part of the curves and then shows performance around that of the model outputs for the remainder of the curve. This result also demonstrates the efficacy of the blockage history in predicting blockages, especially for the highest likelihood sewers. It also suggests potential for deriving further explanatory capability from the historical blockage rate.

Incident Prevention Estimation

The current levels of surveying for DCWW are around 30km per year, around 0.1% of the network. Using the gain curves this would give blockage prevention of around 3%. Given the very small proportion of the network being surveyed, this represents very good performance from the models.

4.3 Chapter Summary

This chapter investigated the use of ensemble techniques and the derivation of a historical input feature. The aim was to improve the performance of the sewer level models from Chapter 3. For the ensemble models, individual models were produced by taking a selection of input features. Different methods of selecting the input features, combining the models' outputs and different numbers of input features were investigated. For the historical input feature, the effect of an input feature and the number of years of historical data were investigated. The outputs from the models were then validated using a further dataset of incidents and survey results. This allowed comparison of the model outputs and the historical blockage rate for the different types of incidents, and validation of the model's ability to prevent blockages. The ensemble techniques showed little improvement in performance. Most models showed no improvement while those that did showed greater overfitting to the training dataset. More years of incident data were not found to influence model performance but the inclusion of a historical input feature showed an improvement in performance, especially for the highest likelihood sewers. The validation showed that a combination of the historical blockage rate with the output from the decision trees gave the best predictive performance, performing well for blockage and flooding incidents. The next chapter gives an overall summary and the conclusions from this thesis, and makes recommendations for future work.

Chapter 5

Summary and Conclusions

5.1 Summary

5.1.1 Decision trees - Sewer Level

Decision trees were used to produce models of the likelihood of blockage at a sewer level. Data on sewers formed the inputs to the models while data on blockages formed the output being predicted. The inputs to the models included: sewer characteristics such as diameter, material and length, property locations, the types and ages of properties and derived variables such as sewer velocity. The models predicted a blockage flag indicating whether a sewer had blocked in the period of historical data available. Models for different subsets of the network and for the different causes of blockages were built using Classification and Regression (CART) and C5.0 decision trees. A number of relatively accurate blockage prediction models were produced. These demonstrate the efficacy of using decision trees for finding patterns in large datasets, providing further understanding of the most useful explanatory factors and allowing the prioritisation of proactive maintenance. Some of the basic sewer characteristics such as length, diameter, gradient, combined with property data, provide good explanatory capability in these models. The appearance of sewer velocity in the decision trees also demonstrates the benefits from infilling the gradient and combining this with diameter. The models of the different blockage causes are affected by the inconsistency in the data. The models are inconsistent with some showing poor

performance or overfitting to the training dataset. However, some models do perform better and derive useful relationships for predicting that cause of blockage. This suggests that if a more consistent dataset were available, it may be possible to produce good models of the different causes of blockages. This would help inform the factors influencing these different mechanisms.

5.1.2 Decision trees - Area Level

Models were developed at an area level from the aggregation of the sewer level data. Different geographical areas were investigated for their similarity to areas of proactive maintenance and consistency of sewer length. Postcode formed the initial groups, with larger postcodes broken up and smaller ones combined. The input variables to the sewer level models were aggregated to an area level. The blockage flag, initially predicted, formed a relative blockage proportion for each group. CART trees were produced for different thresholds in this continuous measure. The results showed limited benefit from the geographical aggregation, with similar performance achieved by the sewer level models. Given the greater spatial resolution of the sewer level and similar performance, these models are less beneficial.

5.1.3 Decision trees - Ensembles

Ensemble techniques produce a number of individual models, the outputs from which are combined into a single output. By maximising the performance of the individual models and minimising the correlation between them, the best performing models are produced. To evaluate the benefit from these techniques in this application, sewer level models for the public, combined subset of the network were developed. CART trees predicted the blockage flag, with each tree produced using a selection of the available input variables. Different methods of selecting the input variables, combining the outputs and the number of inputs were investigated. The results showed limited benefit from the ensembles. Most models were poorer than those of the original sewer level models. Of those which

were better most showed greater overfitting to the training dataset.

5.1.4 Decision trees - Historical Input Feature

A historical input feature based on a sewer's history of blockage was investigated. Different experiments evaluated the benefit of a historical input feature and the effect of the number of years of incident data. CART decision trees were used to produce models for the public part of the network. The historical input feature was used in addition to the variables used in the sewer level models, with a blockage flag predicted. For models with the input feature, there was improved performance overall and particularly for the sewers with the highest likelihood of blockage. The number of years of available data had little influence on performance. The more recent blockage history gave greater explanatory capability than the less recent history.

5.1.5 Validation

Performance was validated using an existing dataset of survey results and further datasets of incidents. The outputs from the best decision tree models, a historical blockage rate and combination of the two were compared. Gain curves were produced for blockage, flooding and pollution incidents comparing the three outputs. The outputs did not perform well on the survey results, the outputs deviating little from random. However, the outputs did perform well for predicting incidents, particularly blockage and flooding. The combination of the historical blockage rate and model output predicted these incidents best.

5.2 Conclusions

The following are the conclusions from this thesis:

- Decision trees gave relatively accurate models on this real-world data and informed which factors influence blockages.

- Validation using existing survey results and further datasets of incidents demonstrated the potential of the models to prevent blockages.
- Models including basic sewer characteristics, property information and sewer blockage history showed good accuracy, particularly for the highest likelihood sewers.
- A sewer's blockage history is a strong predictor of future blockage likelihood. This is particularly beneficial for identifying the sewers most likely to block. The use of blockage history has not been widely investigated within the literature.
- A sewer's more recent blockage history better predicts the likelihood of blockage.
- Sewer velocity influences the likelihood of blockage more than gradient alone. This is based on gradient data which has been heavily infilled. However, the influence of sewer velocity also shows the benefit of this infilling.
- Models of the different causes of blockages can be developed but a consistent dataset must be available for sufficient accuracy to be achieved.
- An area level approach could be useful for identifying blockage hotspots and areas for WaSCs to prioritise but the aggregation to an area level does not automatically improve predictions.
- Ensembles have been widely used to improve model performance but no benefit was found in their application in this thesis.

5.3 Future Work Recommendations

The following are the recommendations for future work:

Further investigation of aggregation methods The aggregation could be: alternative geographical areas or a system of grouping joined parts of the network

together. A different geographical area may better group sewers. Grouping based on network connections would ensure that sewers in the same aggregated group are connected. Groups could also be separated at major changes in sewer characteristics to give groups with more consistent characteristics.

Further investigation of historical input features The investigations conducted showed the explanatory capability of blockage history and the greater capability of the more recent history. However, it may be possible to derive further explanatory capability from the inclusion of all of the blockage history, separated into periods of more and less recent history.

Inclusion of sewer condition grades Sewer condition grades give informations which is believed to highly influence the likelihood of blockages. The possibility of its inclusion is worth further investigation. This could be the evaluation of explanatory capability on a smaller area with good coverage of this information. Alternatively, a method of estimating the condition grade could be developed, or taken from existing studies.

Investigation of different blockage causes Models of the different causes of blockages would give greater understanding to the factors influencing these specific mechanisms. A more consistent dataset could be sourced and used to investigate this. These models could also be combined into an ensemble model predicting the overall likelihood of blockage. This would give an overall likelihood and a breakdown of likelihood by cause.

Inclusion of further dynamic factors There are likely to be temporal factors which influence the likelihood of blockage. These could be rainfall, seasonality or previous occurrence of blockages. These factors may influence different sewers to different extents. This may improve the prediction of blockage likelihood and could be used to predict at a more detailed temporal resolution, for example for a given month.

Prediction of blockage rate The models developed in this thesis all predict a blockage flag based on whether a sewer has blocked. However, predicting a blockage rate would further inform the magnitude of the likelihood.

Application to another water company This thesis describes a process which could be applied to the datasets of another water company. The main datasets used are those of the sewer characteristics, property ages, property locations and incidents, which could be sourced for other areas of the country. The datasets could be prepared and missing data infilled for important variables creating a dataset for modelling blockages or incidents like flooding and pollution. A historical input feature would be recommended, derived in the manner described here. Decision trees have been shown to perform well in this application: accurately predicting a likelihood of blockage and informing the important factors related to blockages. The validation approach should then help demonstrate the performance of the models in predicting future blockages and the benefits provided from this.

Bibliography

[1] M Hall, Z Kapelan, R Long, and D Savic, “Deterioration Rates of Sewers,” UKWIR, Technical Report PP/05/051, 2006.

[2] OFWAT. (2010). Putting water consumers first - the service incentive mechanism. Accessed 23/08/2016, [Online]. Available: http://www.ofwat.gov.uk/wp-content/uploads/2015/11/prs_inf_simsup1.pdf.

[3] S. U. Custance-Baker, R. U. Long, and N. U. Muggeridge, “The Practicality of a Planned Preventative Maintenance Approach for Managing Sewer Blockages,” UKWIR, Technical Report 16/SW/01/15, 2015.

[4] OFWAT. (2015). Transfer of private sewers. Accessed 23/08/2016, [Online]. Available: http://www.ofwat.gov.uk/publications/transfer-of-private-sewers/.

[5] L. Hafskjold and A. Kø nig, “Improved assessment of sewer pipe condition,” in *CityNet 19th European Junior Scientist Workshop*, Meaux-la-Montagne, France, 2004, pp. 1–8.

[6] O. Giustolisi and D. Savic, “A symbolic data-driven technique based on evolutionary polynomial regression,” *Journal of Hydroinformatics*, vol. 8, no. 3, pp. 207–222, 2006, International Water Association Publishing.

[7] D. Savic, “The use of data-driven methodologies for prediction of water and wastewater asset failures,” in *Risk Management of Water Supply and Sanitation Systems*, Springer, 2009.

[8] D Savic, O Giustolisi, and W Shepherd, “Modelling sewer failure by evolutionary computing,” *Proceedings of the Institution of Civil Engineers - Water Management*, vol. 159, no. 2, pp. 111–118, 2006.

[9] R. Ugarelli and S. Kristensen, “Statistical analysis and definition of blockages-prediction formulae for the wastewater network of Oslo by evolutionary computing,” *Water Science and Technology*, vol. 59, no. 8, pp. 1457–1470, 2009.

[10] R. Fenner, L Sweeting, and M. Marriott, “A new approach for directing proactive sewer maintenance,” *Proceedings of the Institution of Civil Engineers-Water and Maritime Engineering*, vol. 142, no. 2, pp. 67–77, 2000.

[11] R. Fenner, G McFarland, and O Thorne, “Case-based reasoning approach for managing sewerage assets,” *Proceedings of the Institution of Civil Engineers - Water Management*, vol. 160, no. 1, pp. 15–24, 2007.

[12] R. Ugarelli, G. Venkatesh, H. Brattebø, V. Di Federico, and S. Sæ grov, “Historical analysis of blockages in wastewater pipelines in Oslo and diagnosis of causative pipeline characteristics,” *Urban Water Journal*, vol. 7, no. 6, pp. 335–343, 2010, ISSN: 1573-062X.

[13] S Arthur and R Burkhard, “Prioritising Sewerage Maintenance using Inferred Sewer Age - A Case Study for Edinburgh,” *Water science and technology*, vol. 61, no. 9, pp. 2417 –2424, 2010.

[14] S. Arthur, H. Crow, and L. Pedezert, “Understanding blockage formation in combined sewer networks,” *Proceedings of the Institution of Civil Engineers - Water Management*, vol. 161, no. 4, pp. 215–221, Aug. 2008.

[15] L Berardi, O Giustolisi, D. Savic, and Z. Kapelan, “An effective multi-objective approach to prioritisation of sewer pipe inspection,” *Water science and technology*, vol. 60, no. 4, pp. 841–850, 2009.

[16] D. Savic, S Djordjevic, G Dorini, W Shepherd, A Cashman, and A Saul, “COST-S: a new methodology and tools for sewerage asset management based on whole life costs,” *Water Asset Management International*, vol. 1, no. 4, pp. 20–24, 2005, ICE Publishing.

[17] T. Dietterich, “Ensemble methods in machine learning,” in *Multiple Classifier Systems*, Cagliari, Italy: Springer, 2000.

[18] D. Cutler, T. Edwards, K. Beard, and A. Cutler, “Random forests for classification in ecology,” *Ecology*, vol. 88, no. 11, pp. 2783–2792, 2007.

[19] L. Breiman, “Random forests,” *Machine learning*, vol. 45, no. 5, pp. 5–32, 2001.

[20] R. Díaz-Uriarte and S. Alvarez de Andrés, “Gene selection and classification of microarray data using random forest.,” *BMC bioinformatics*, vol. 7, p. 3, Jan. 2006, ISSN: 1471-2105. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1363357\&tool=pmcentrez\&rendertype=abstract.

[21] CACI. (2014). The ACORN User Guide, [Online]. Available: http://acorn.caci.co.uk/downloads/Acorn-User-guide.pdf.

[22] ——, (2013). What is ACORN, [Online]. Available: http://acorn.caci.co.uk/.

[23] J. Quinlan, *C4.5: programs for machine learning*. 2014.

[24] IBM, *SPSS Modeler*, 15.0. [Online]. Available: http://www-01.ibm.com/support/docview.wss?uid=swg27023172#en.

Appendices

Appendix A

Distribution of Input Variables

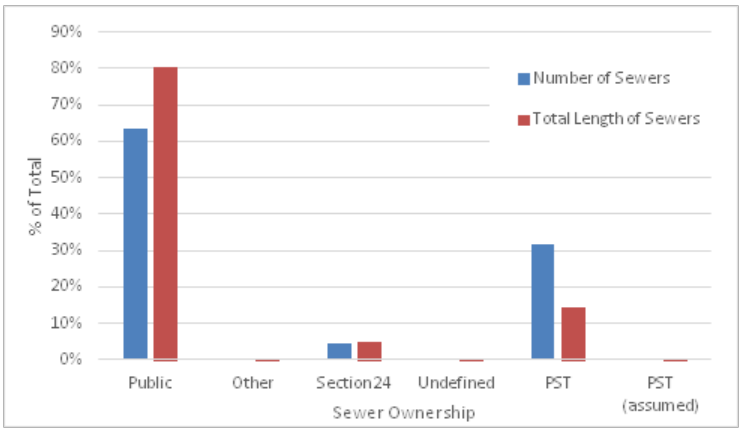


Figure 23: Distribution of input variable for: sewer ownership

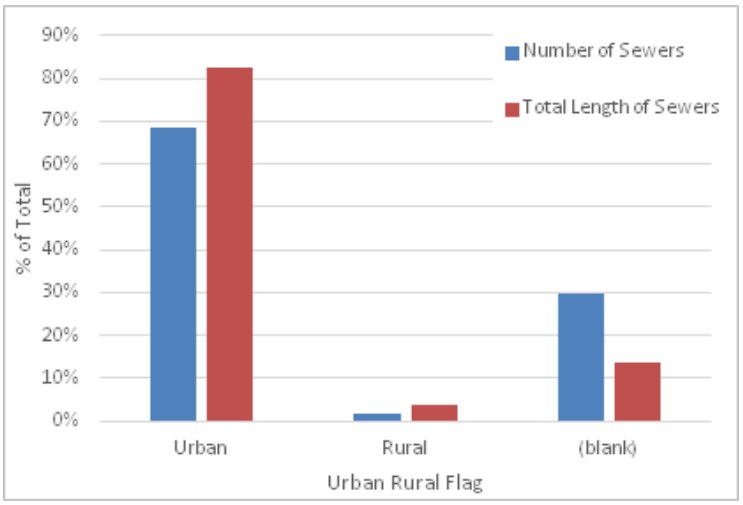


Figure 24: Distribution of input variable for: Urban Rural Flag

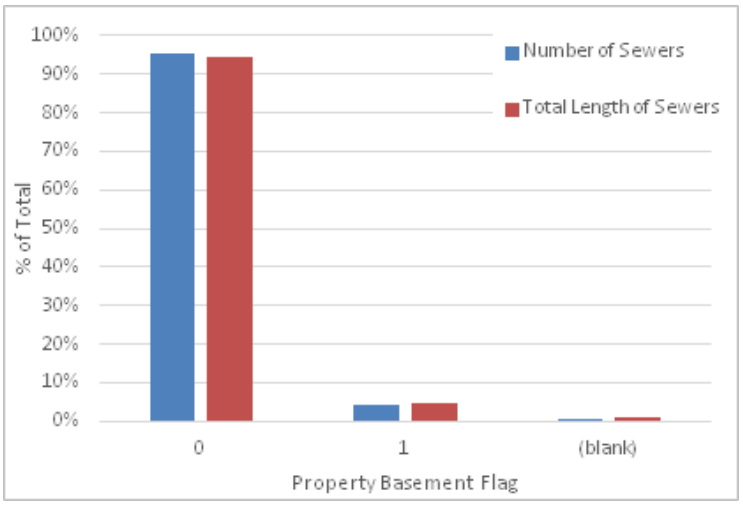


Figure 25: Distribution of input variable for: Property Basement Flag

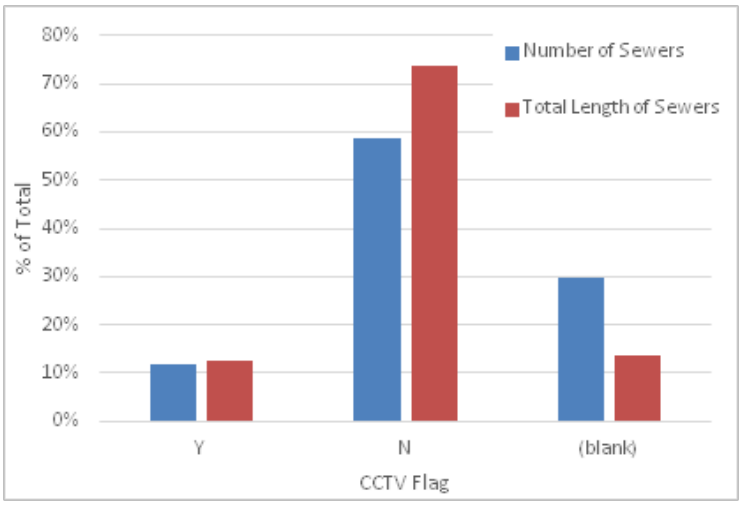


Figure 26: Distribution of input variable for: CCTV Flag

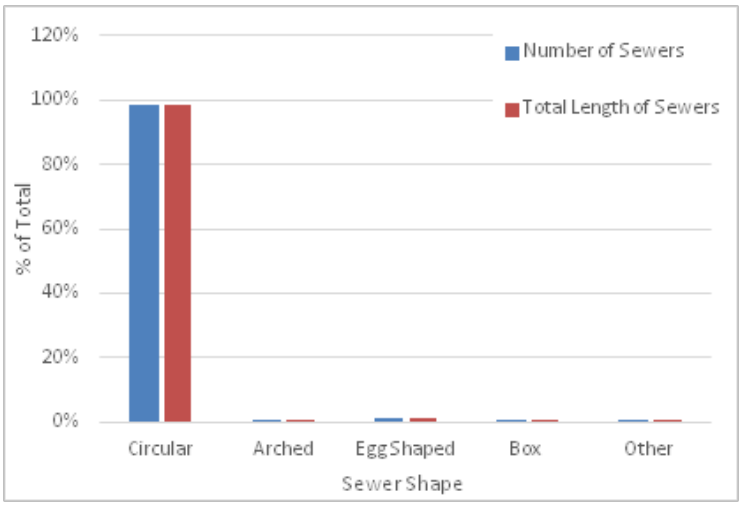


Figure 27: Distribution of input variable for: Sewer Shape

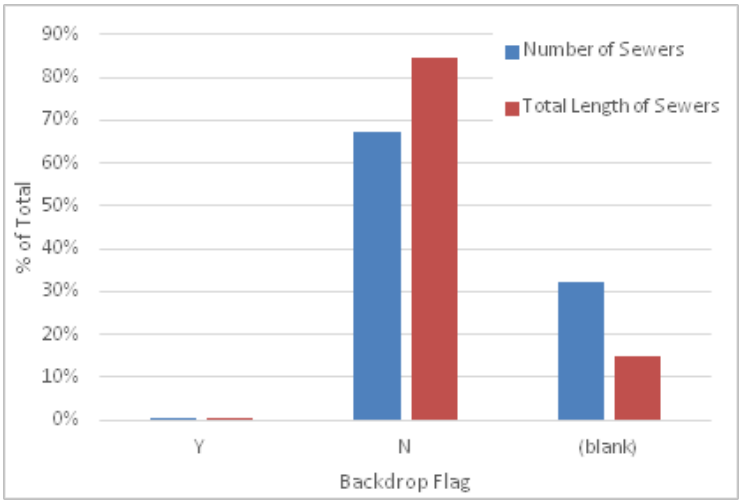


Figure 28: Distribution of input variable for: Backdrop Flag

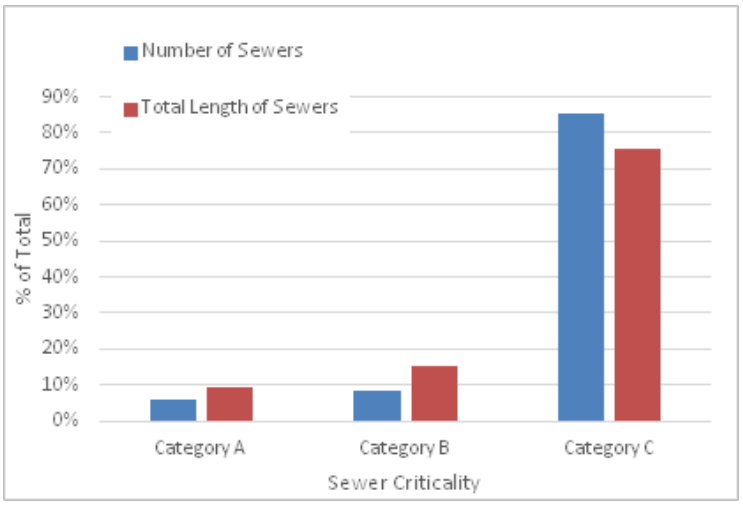


Figure 29: Distribution of input variable for: Sewer Criticality

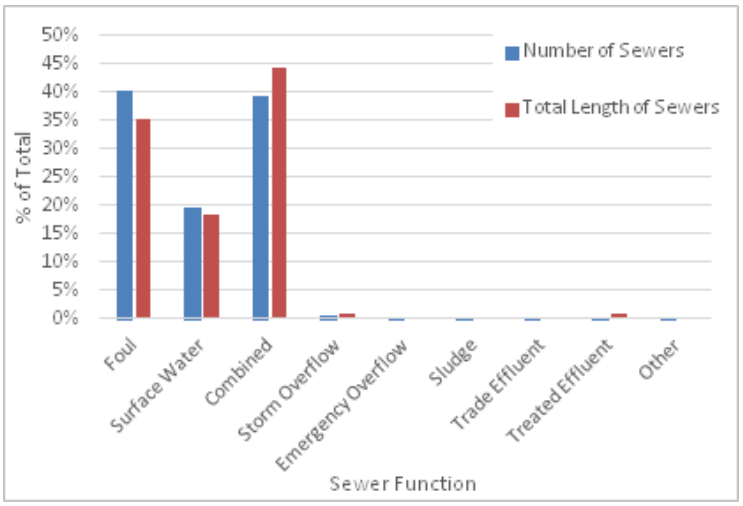


Figure 30: Distribution of input variable for: Sewer Function

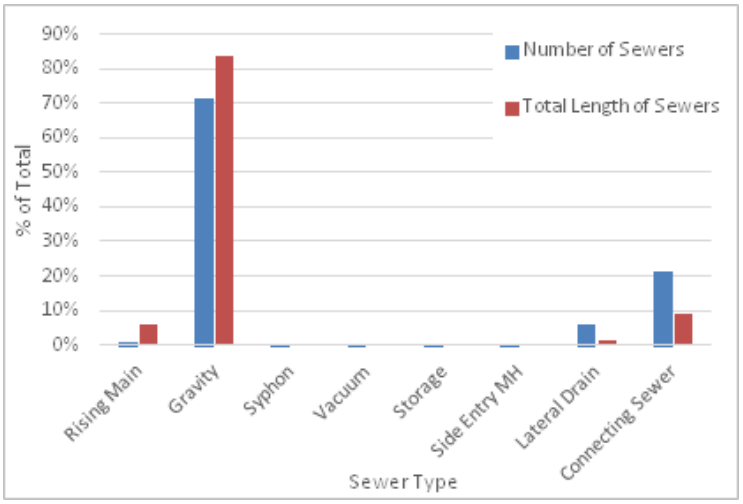


Figure 31: Distribution of input variable for: Sewer Type

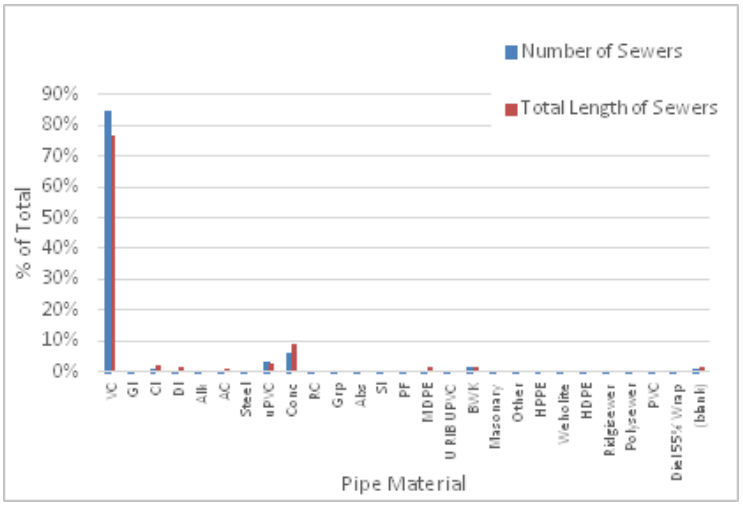


Figure 32: Distribution of input variable for: Pipe Material

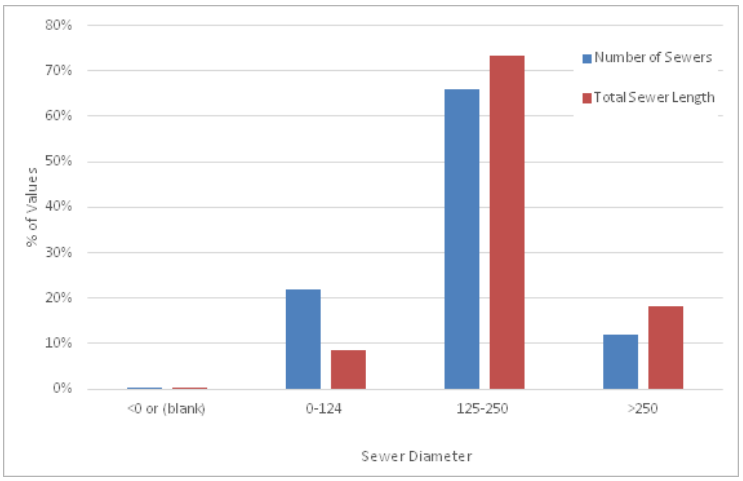


Figure 33: Distribution of input variable for: sewer diameter

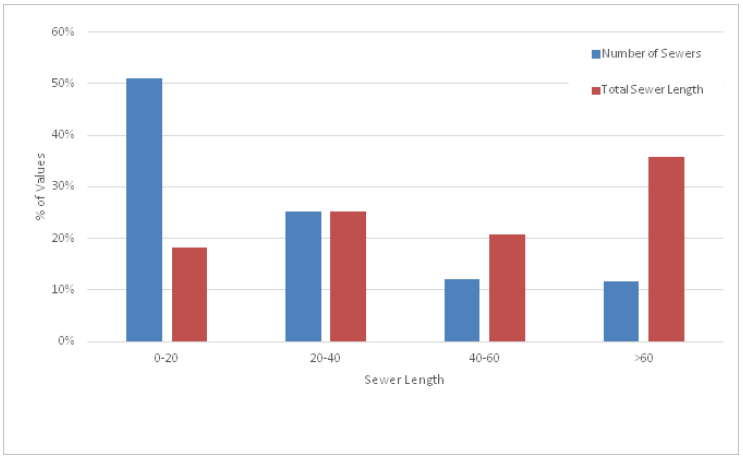


Figure 34: Distribution of input variable for: sewer length

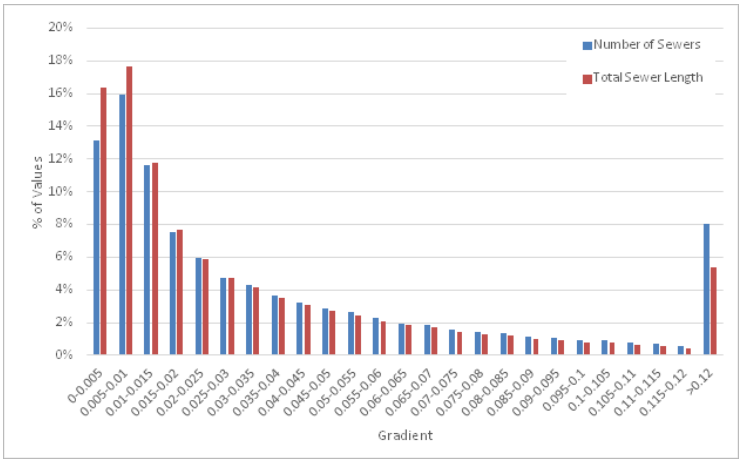


Figure 35: Distribution of input variable for: Gradient

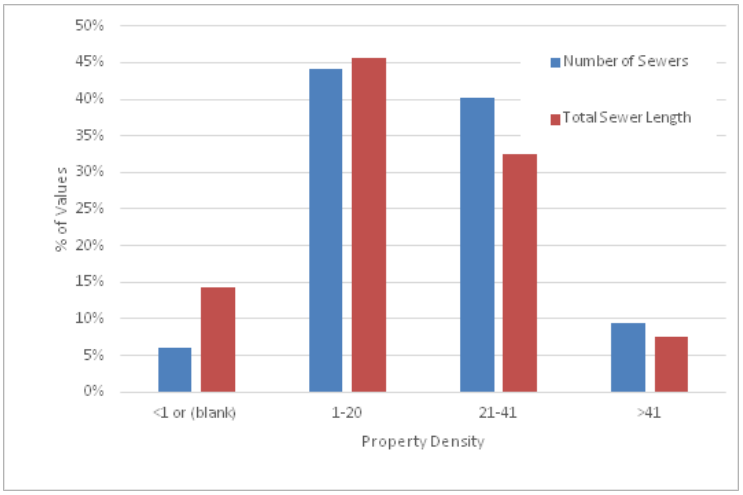


Figure 36: Distribution of input variable for: Property Density

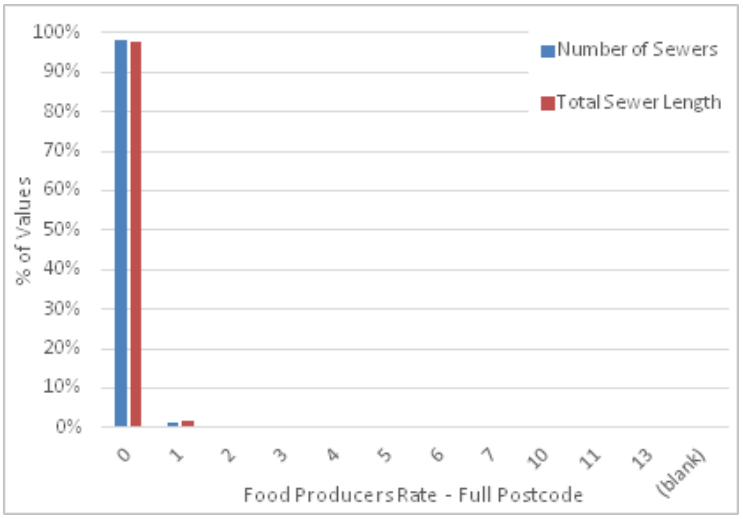


Figure 37: Distribution of input variable for: Food Producers

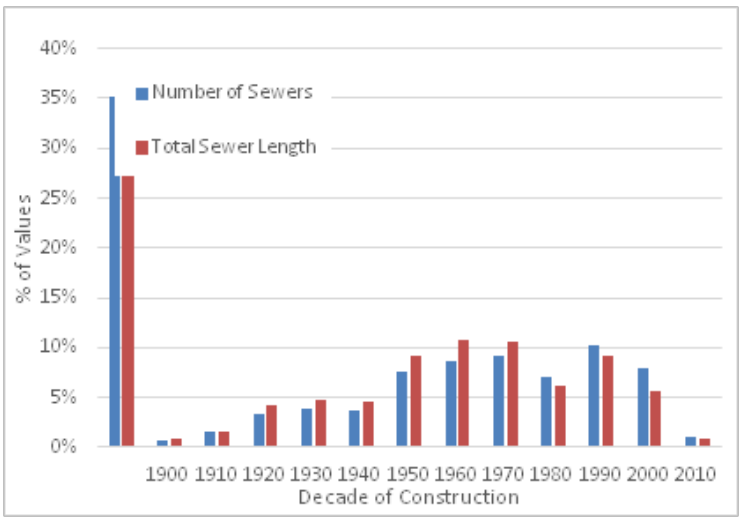


Figure 38: Distribution of input variable for: Construction Decade

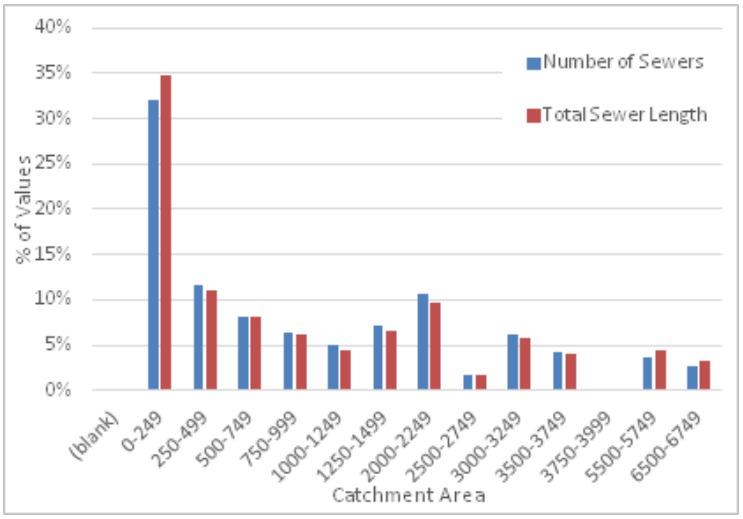


Figure 39: Distribution of input variable for: Catchment Area

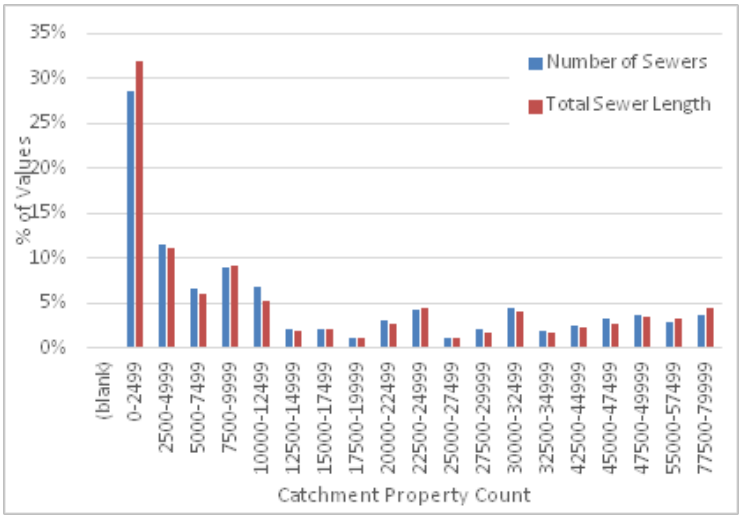


Figure 40: Distribution of input variable for: Catchment Property Count

Appendix B

Sewer Dataset - Data Quality Analysis

Table 12: Table showing the fields present within the sewer dataset, giving the information about the field and the categories present within it and a measure of the amount of missing and invalid data.

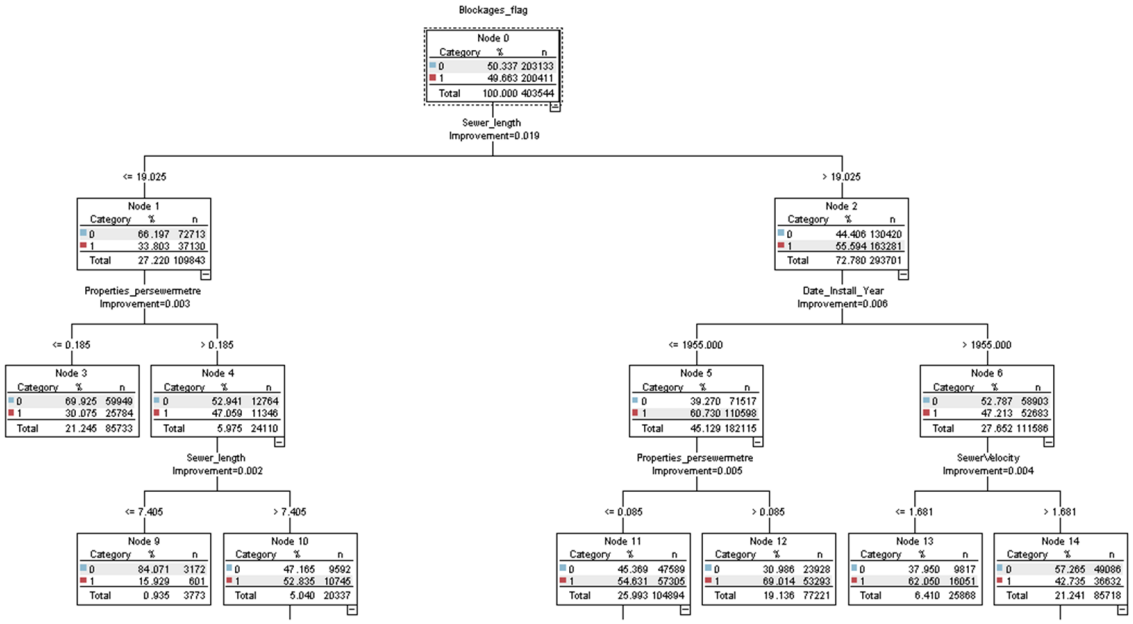
Long Name	Description	Availability	Possible values	Equivalent To	% Missing or Invalid
Asset Datasets					
OBJECTID	GIS generated ID	Available			0%
SAPID	ID used on SAP	Available			0%
IPID	ID	Available			0%
SEWER_TYPE	Sewer type	Available	2 \ 3 \ 4 \ 5 \ 8 \ 9 \ 10 \ 11	Gravity \ Syphon \ Vacuum \ Storage \ Side Entry MH \ Lateral Drain \ Connecting Sewer \ Link Sewer	0%
SEWER_FUNCTION	Sewer function	Available	2 \ 3 \ 4 \ 5 \ 6 \ 7 \ 8 \ 10	SURFACE WATER \ COMBINED \ STORM OVERFLOW \ EMERGENCY OVERFLOW \ SLUDGE \ TRADE EFFLUENT \ TREATED EFFLUENT \ OTHER	0%
OWNERSHIP	Dataset filtered to only include values 1,4,5,6,7,8 (those under the responsibility of DCWW)	Available	4 \ 5 \ 6 \ 7 \ 8	OTHER \ SECTION 24 - DESKTOP \ Undefined \ P.S.T. \ P.S.T. (assumed route)	0%
SEWER_SHAPE	Sewer shape	Available	1 \ 2 \ 3 \ 4 \ 5 \ 6	Circular \ Arched \ Egg Shaped \ Box	0%
DATE_CONSTRUCTED	Date of construction	Available			34%
DATE_ABANDONED	Date of abandonment	Available			100%
DATE_ADOPTED	Date of adoption	Available			91%
CON_CRITICAL_FACTOR	Criticality	Available	1 \ 2 \ 3	Category A \ Category B \ Category C	0%
MATERIAL_TYPE	Material type	Available	1 \ 2 \ 3 \ 4 \ 5 \ 6 \ 7 \ 8 \ 9 \ 10 \ 11 \ 12 \ 13 \ 14 \ 15 \ 16 \ 17 \ 18 \ 19 \ 20 \ 21 \ 23 \ 24 \ 25 \ 26 \ 27 \ 28 \ 29	VC \ GI \ CI \ DI \ ALK \ AC \ STEEL \ uPVC \ CONC \ RC \ GRP \ ABS \ SI \ PF \ MDPE \ U RIB UPVC \ BWK \ MASONARY \ UNKNOWN \ OTHER \ Undefined \ HPPE \ Weholite \ HDPE \ Ridgisewer \ Polysewer \ PVC \ Diel 55% Wrap	1%
INVERT_LEVEL	Invert Level	Available			100%
UPSTREAM_INV_LEV	Upstream invert level	Available			44%
DOWNSTRM_INV_LEV	Downstream invert level	Available			46%
GRADIENT	Sewer gradient (expressed as 1 in x)	Available			81%
Z	Height above sea level	Available			100%
BACKDROP_FLAG	Flag for presence of backdrop	Available	1 \ 2 \ 3	Y \ N \ Unknown	31%
URBAN_RURAL	Flag of urban/rural location	Available	1 \ 2	Urban \ Rural	29%
STUDY	Hydrological study results	Available	1 \ 2 \ 3 \ 4 \ 5 \ 6 \ 7 \ 8 \ 9	Not Modelled \ Ok \ Fail - 1/50 yr storm \ Fail - DG5 \ Fail - DG5 & 1/50 yr storm \ Fail - 1/2 yr storm \ Fail - 1/2 & 1/50 yr storm \ Fail - DG5 & 1/2 yr storm \ Fail - DG5, 1/2 & 1/50 yr	92%

Long Name	Description	Availability	Possible values	Equivalent To	% Missing or Invalid
DATE_STUDY	Hydrological study dates	Available			0%
CONST_TYPE	Type of construction	Available	1 \ 2 \ 3 \ 4 \ 5 \ 6 \ 7	Pipe Jacking \ Thrust Bore \ Elevated \ Micro Tunnel \ Open Excavation \ Tunnel \ Unknown	100%
TYPE_OF_PROTECT	Type of protection	Available	1 \ 2 \ 5 \ 6 \ 7 \ 8 \ 10 \ 11 \ 12	Cathodic Protection \ Thermal Wrapping \ Special Coating / wrap \ Polythene sleeve \ Cement grout \ Epoxy resin \ None \ Unknown \ Other	31%
GROUND_LOAD_TYPE	e.g. A road, motorway, verge	Available	1 \ 2 \ 3 \ 4 \ 5 \ 6 \ 7 \ 8 \ 9 \ 10 \ 11 \ 12 \ 13	Trunk Road \ Motorway \ A Road \ B Road \ Unclassified Highway \ Verge \ Footpath \ Field \ Garden \ Driveway \ Built Over \ Other \ Unknown	99%
GROUND_TYPE	e.g. Ash, rock, gravel, sand	Available	1 \ 5 \ 8 \ 9 \ 15 \ 16	Ash \ Gravel \ Rock \ Sand \ Other \ Unknown	100%
WATER_TABLE	e.g. above/below sewer	Available	1 \ 2 \ 4	Above Sewer \ Below \ Not Known	100%
BED_SURR_TYPE	Sewer bed type	Available	1 \ 2 \ 3 \ 4 \ 5 \ 6	Granular \ Concrete haunch \ Concrete surround \ Soil \ Not Known \ Other	100%
COND_GRADE_INT	Internal condition grade	Available	1 \ 2 \ 3 \ 4 \ 5 \ 6	1 \ 2 \ 3 \ 4 \ 5 \ Unknown	86%
COND_GRADE_EXT	External condition grade	Available	1 \ 2 \ 3 \ 4 \ 5 \ 6	Good \ Fair \ Adequate \ Poor \ Inapplicable \ Awful	31%
CCTV	Flag of CCTV survey completion	Available	1 \ 2 \ 3 \ 4	Y \ N \ Unknown \ Undefined	29%
JOINT_TYPE	Joint Type	Available	5 \ 6 \ 7 \ 14 \ 15 \ 19 \ 20	Sleeve \ Spigot Socket \ Spigot Socket - Rubber Ring \ Solvent Welded \ Spigot Socket - Mortar \ Unknown \ Other	31%
REHAB_TECHNIQUE	Rehabilitation technique	Available	1 \ 2 \ 4 \ 5 \ 6 \ 7 \ 8	Stabilisation \ Pipe Lining \ Insitu Coatings \ Pipe Bursting \ Roll Down Pipelining \ Other \ None	31%
REHAB_MAT_TYPE	Rehabilitation material type	Available	1 \ 2 \ 3 \ 4 \ 5 \ 8 \ 9 \ 10 \ 14 \ 15 \ 16	Cement Mortar \ GRC \ GRP \ VC \ UPVC \ Polyolefins \ Insituform or similar \ Spray Gunite \ Bitumen \ Other \ None	31%
LINING_DES_TYPE	Lining Design type	Available	1 \ 2 \ 3 \ 4	Lining grout and Orig form struc \ Lining structural entity \ Lining has no structural role \ None	31%
CONF_FACTOR_POS	Confidence Factor - Position	Available	1 \ 2 \ 3 \ 4 \ 5 \ 6	A1 \ A2 \ B3 \ C4 \ D5 \ GPS	0%
CONF_FACTOR_DATE	Confidence Factor - Date	Available	1 \ 2 \ 3 \ 4 \ 5 \ 6	A1 \ A2 \ B2 \ B3 \ D5 \ Undefined	0%
CONF_FACTOR_SIZE	Confidence Factor - Size	Available	1 \ 2 \ 3 \ 4 \ 5	A1 \ A3 \ B3 \ C4 \ D4	4%
CONF_FACTOR_MAT	Confidence Factor - Material	Available	1 \ 2 \ 3 \ 4 \ 5	A1 \ A2 \ B3 \ B4 \ D4	4%
CON_AGNT_W_ZONE_IPD	Catchment IPID	Available	References a catchment IPID		0%
CON_LENGTH_	Integer value of length - similar to shape.length	Available			0%
Shape_len	Sewer length	Available			0%
ABS_DIAMETER_MM	Absolute Diameter	Available			0%
CON_NOMINAL_DIAMETER	Nominal diameter	Available			0%
ABS_WIDTH_MM	Absolute width	Available			99%
CON_WIDTH	Sewer width	Available			0%

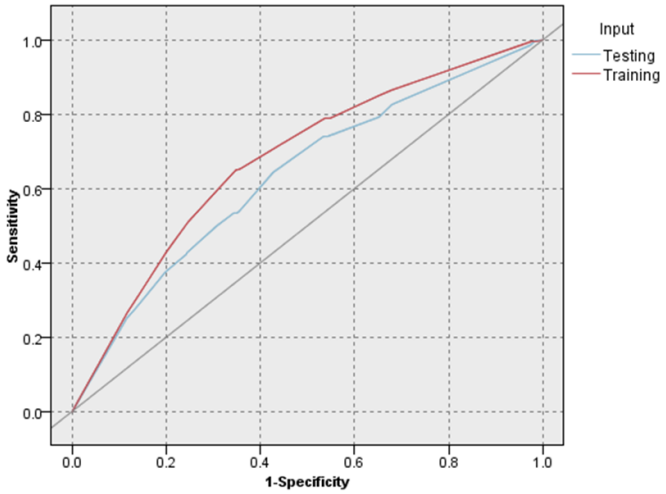
Appendix C

Sewer Level Models - Blockages by Cause -

Decision Trees and ROC curves

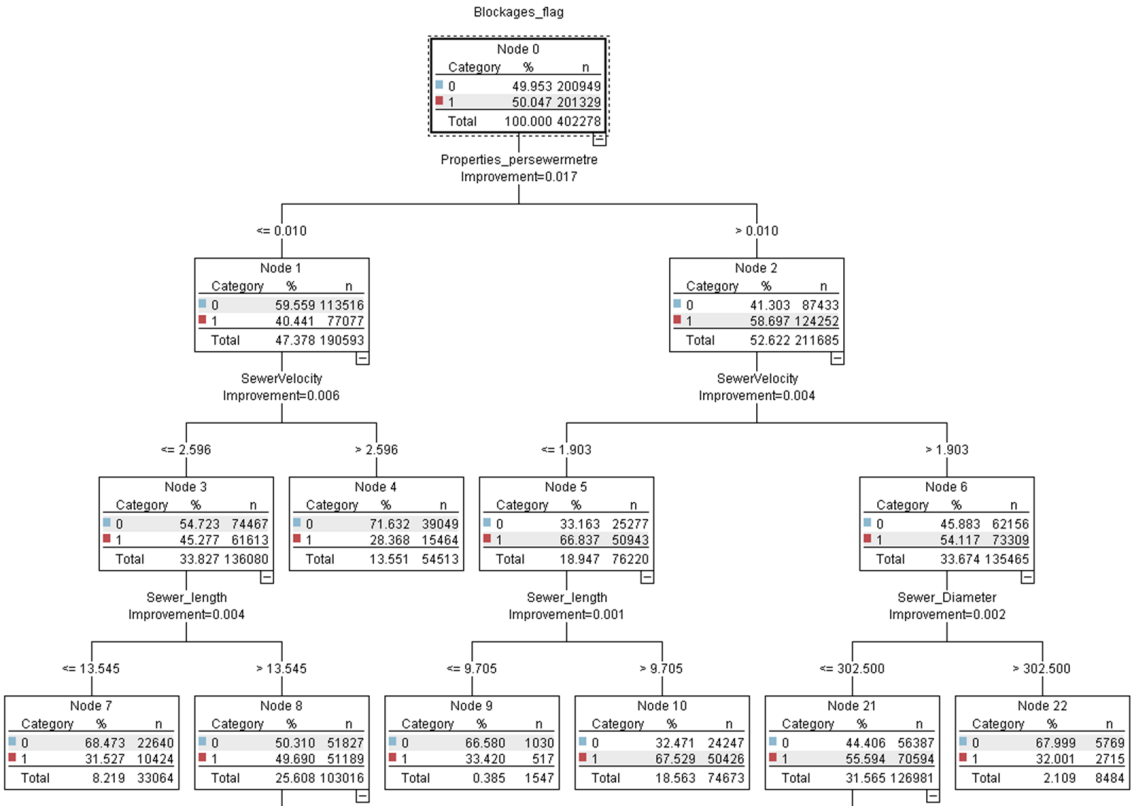


(a) Decision Tree, showing the top four layers of the tree.

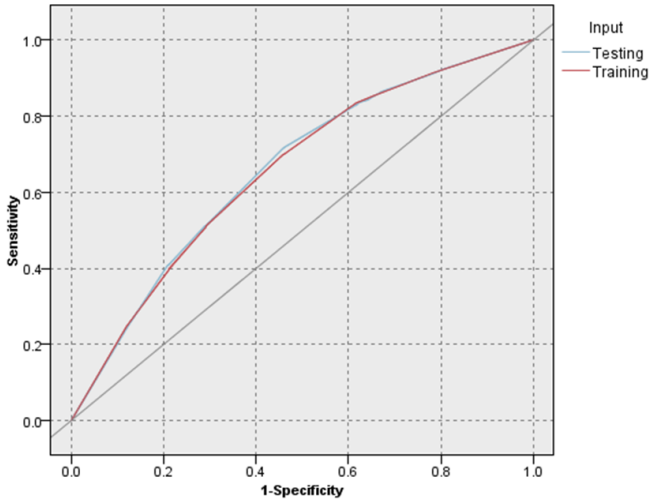


(b) ROC curve

Figure 41: Results obtained from the models for blockages due to silt.

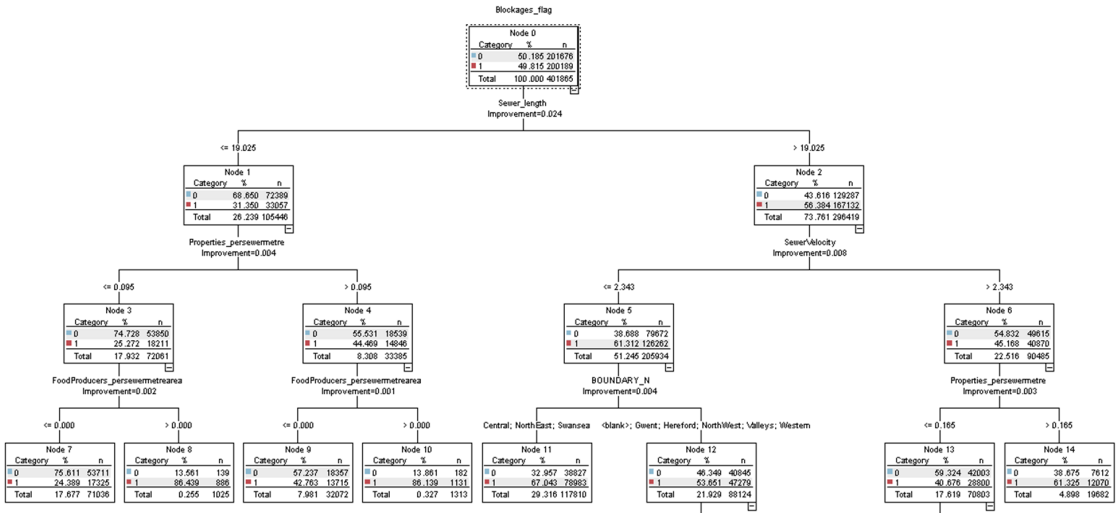


(a) Decision Tree, showing the top four layers of the tree.

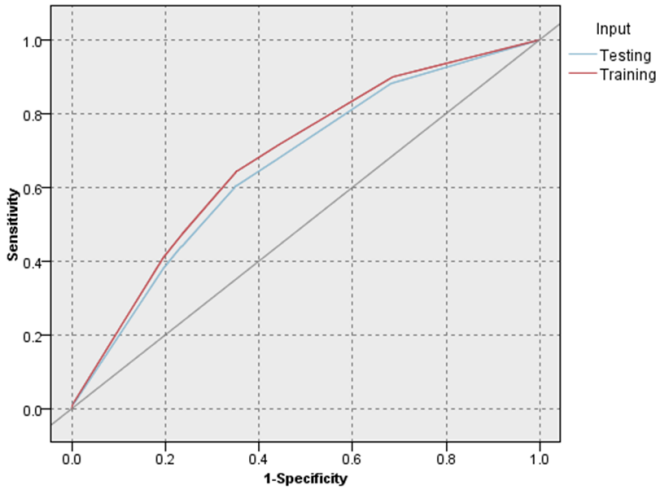


(b) ROC curve

Figure 42: Results obtained from the models for blockages due to nappies, wipes and rags.

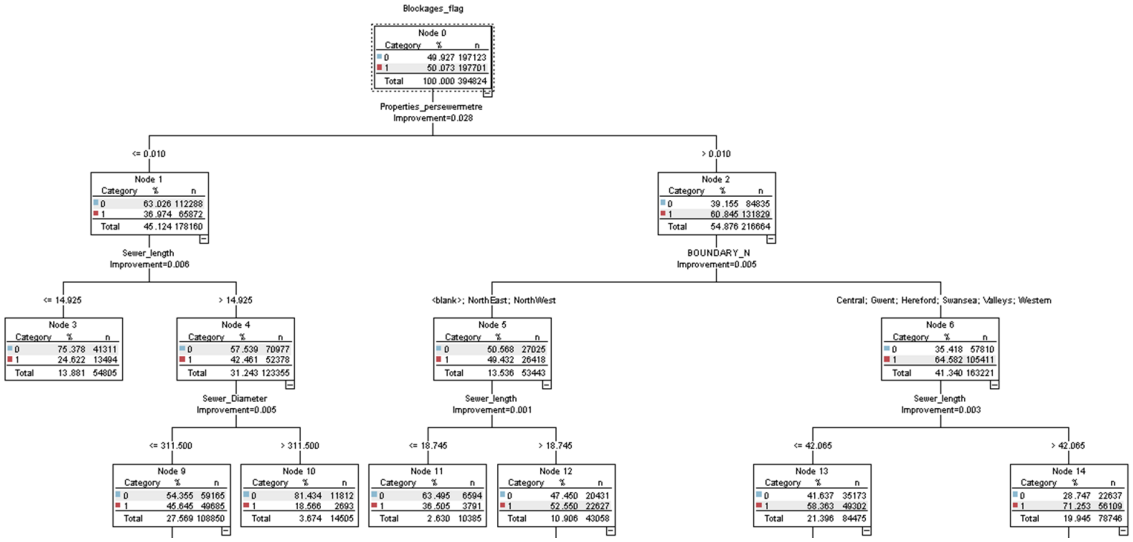


(a) Decision Tree, showing the top four layers of the tree.

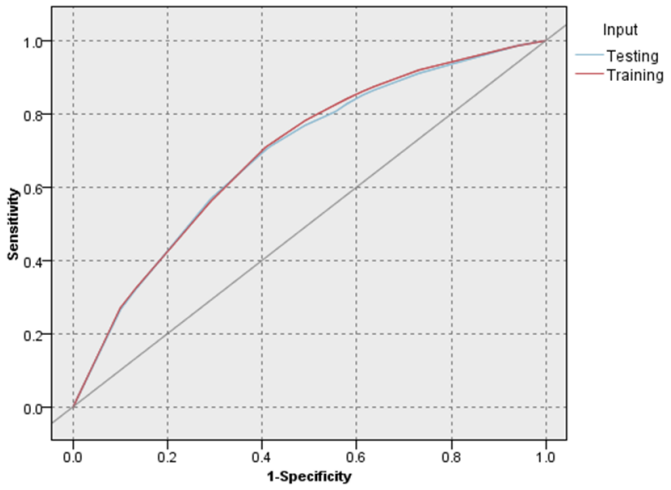


(b) ROC curve

Figure 43: Results obtained from the models for blockages due to fat, oil and grease (FOG).

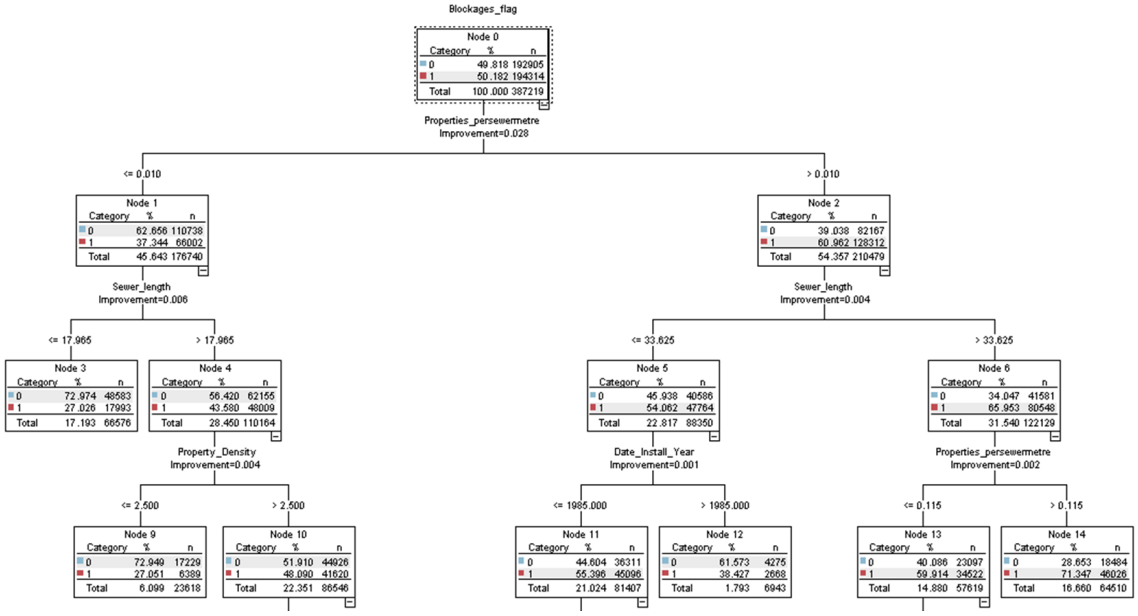


(a) Decision Tree, showing the top four layers of the tree.

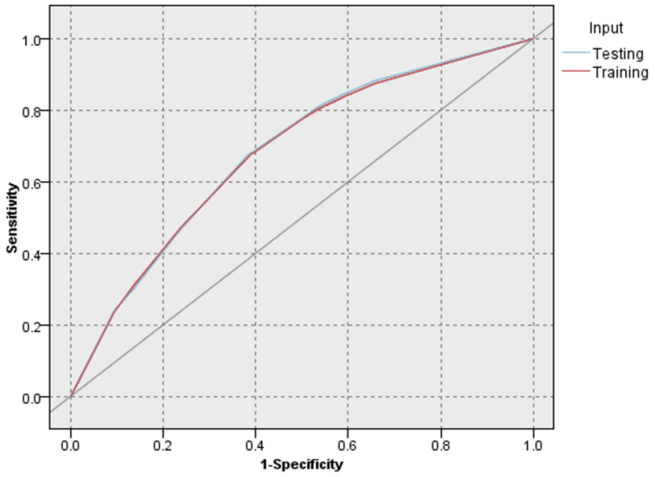


(b) ROC curve

Figure 44: Results obtained from the models for blockages due to debris.



(a) Decision Tree, showing the top four layers of the tree.

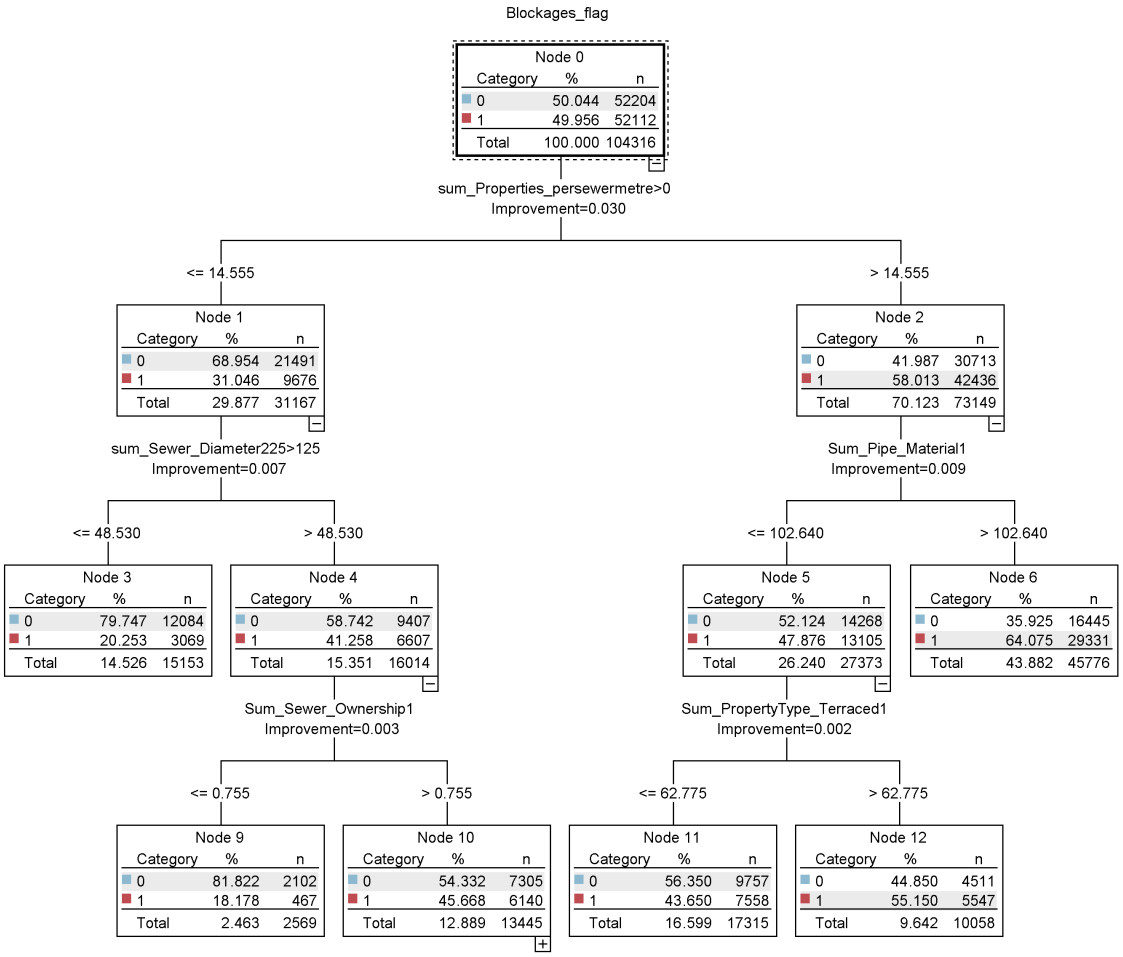


(b) ROC curve

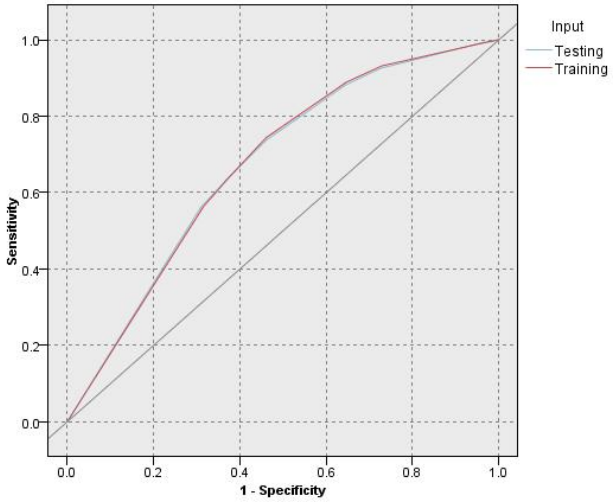
Figure 45: Results obtained from the models for blockages due to 'other causes'.

Appendix D

Area Level Models - Decision Trees and ROC curves

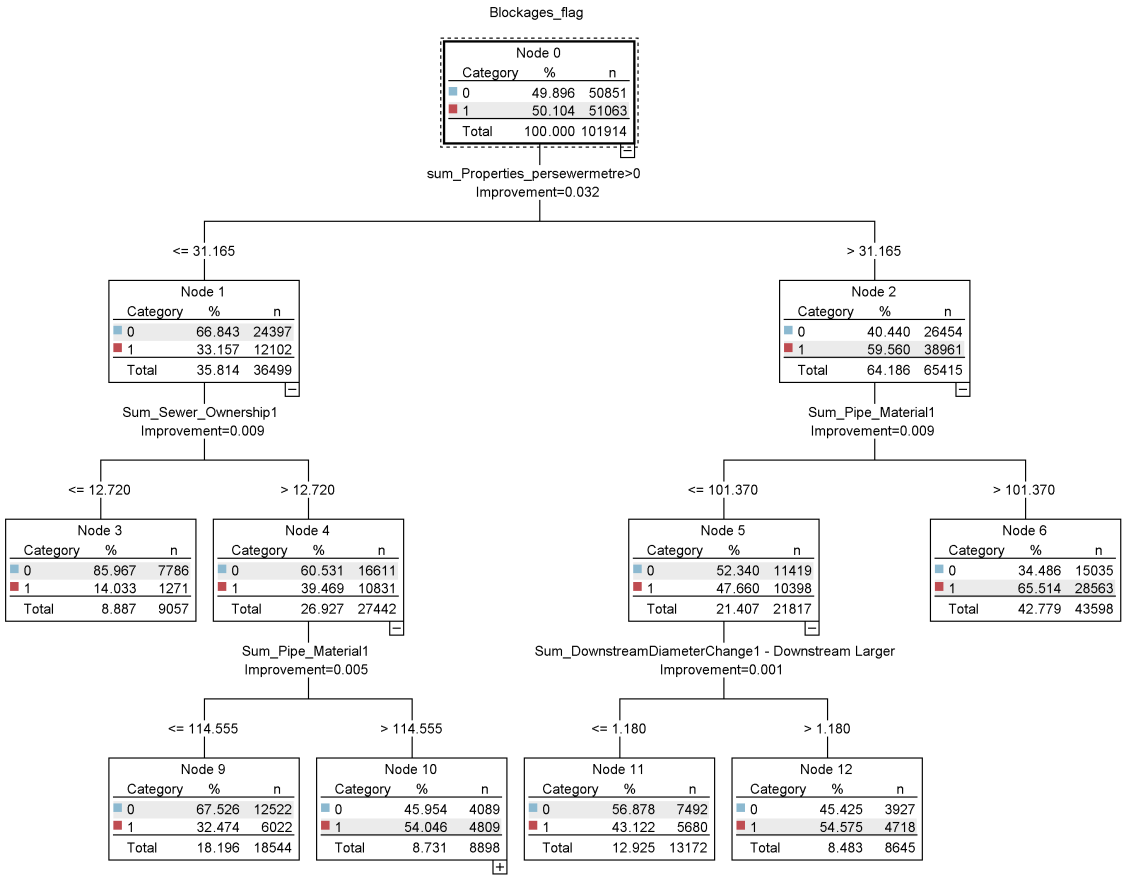


(a) Decision Tree, showing the top four layers of the tree.

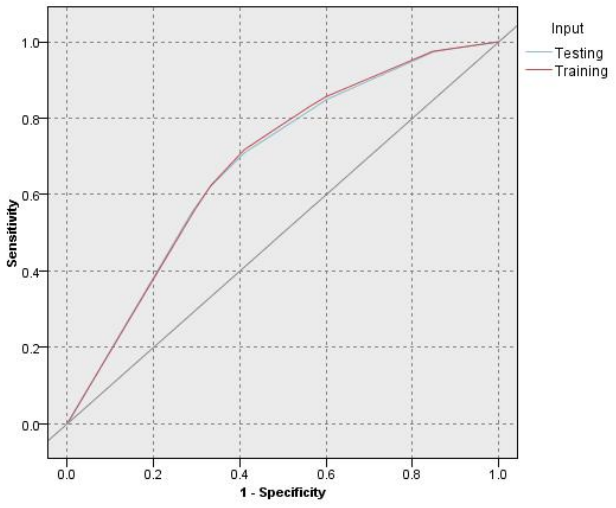


(b) ROC curve

Figure 46: Results obtained from the area level models, using a threshold in the relative blockage proportion of 1.

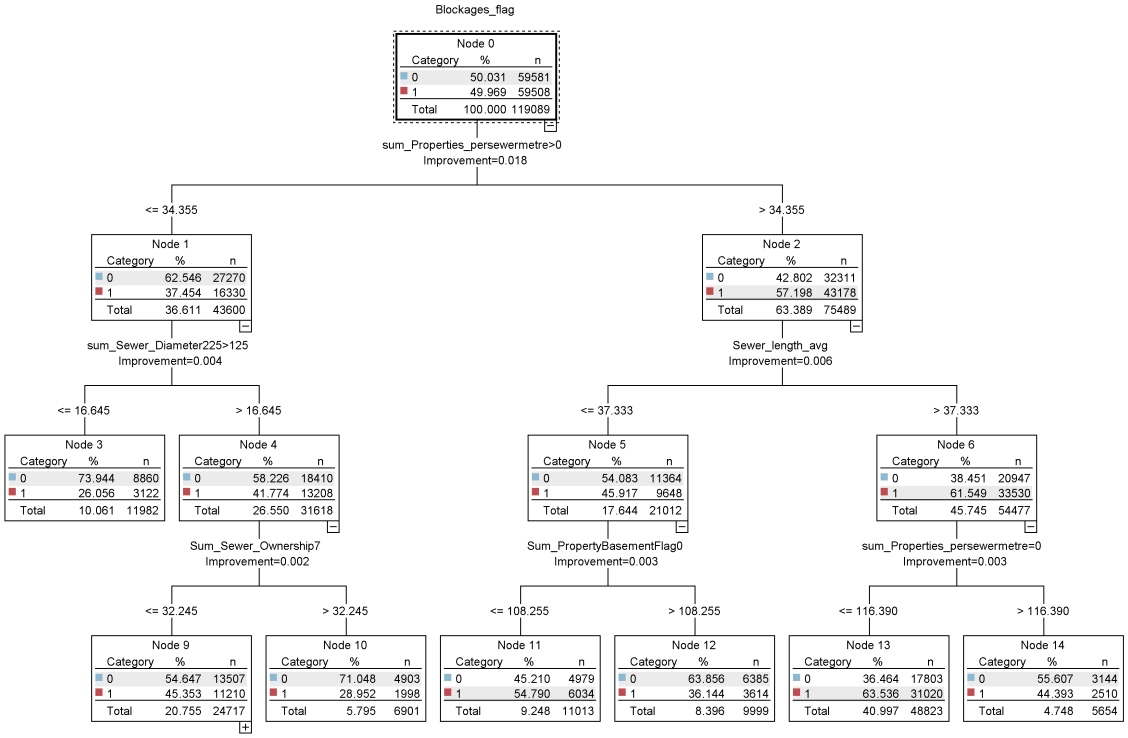


(a) Decision Tree, showing the top four layers of the tree.

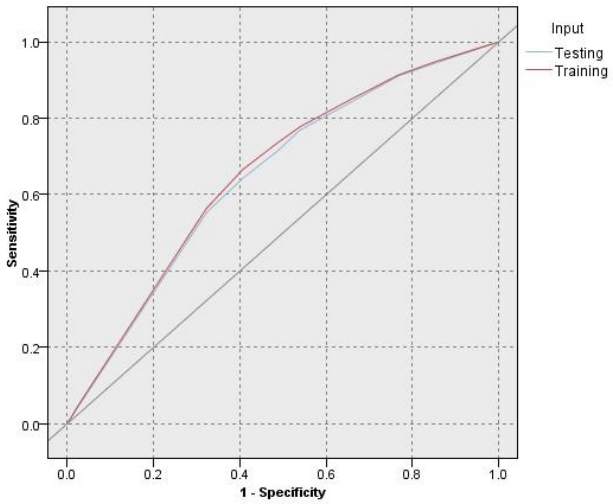


(b) ROC curve

Figure 47: Results obtained from the area level models, using a threshold in the relative blockage proportion of 0.

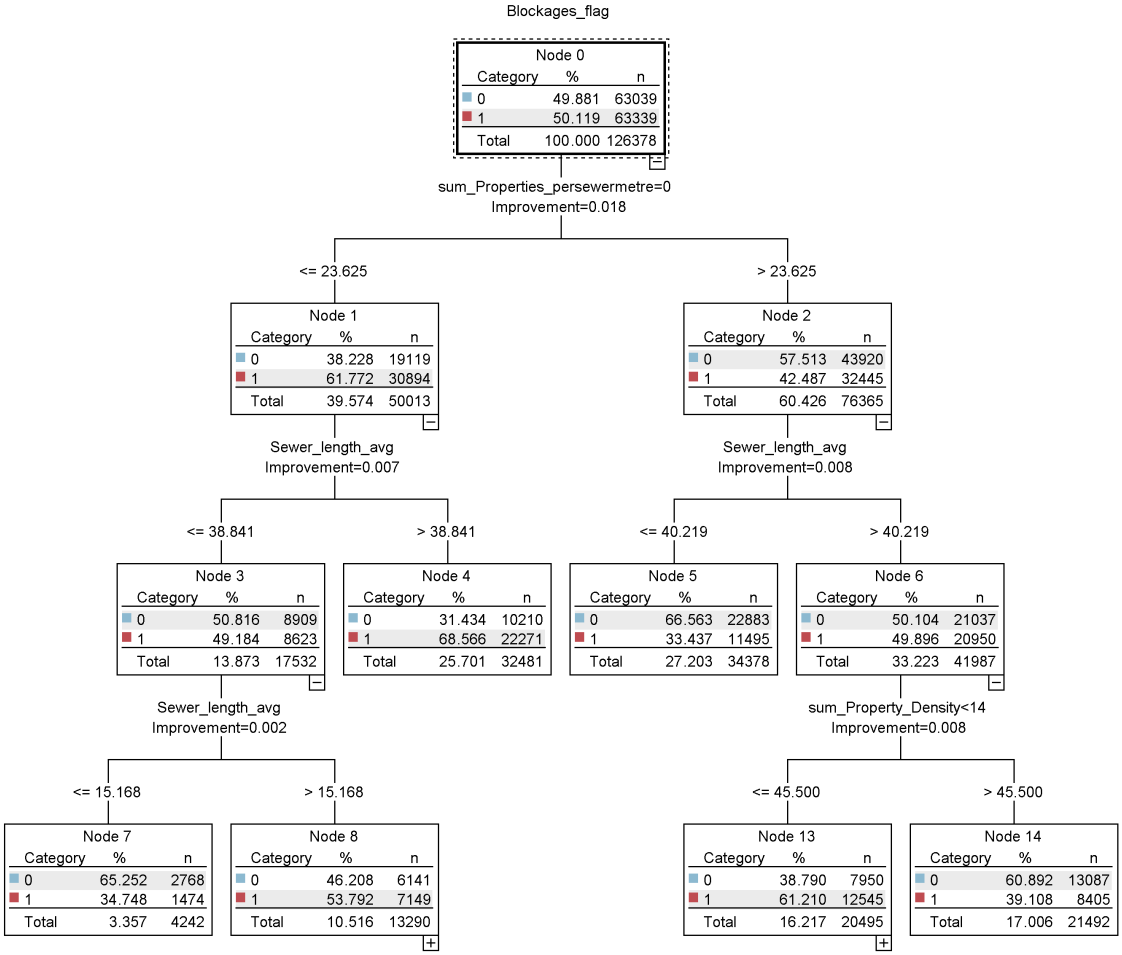


(a) Decision Tree, showing the top four layers of the tree.

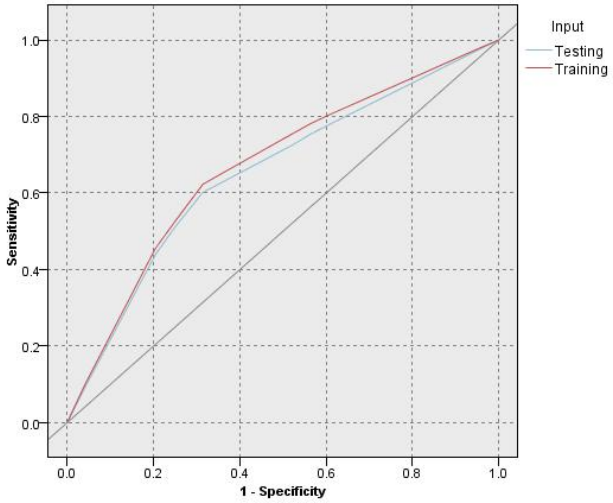


(b) ROC curve

Figure 48: Results obtained from the area level models, using a threshold in the relative blockage proportion of 4.

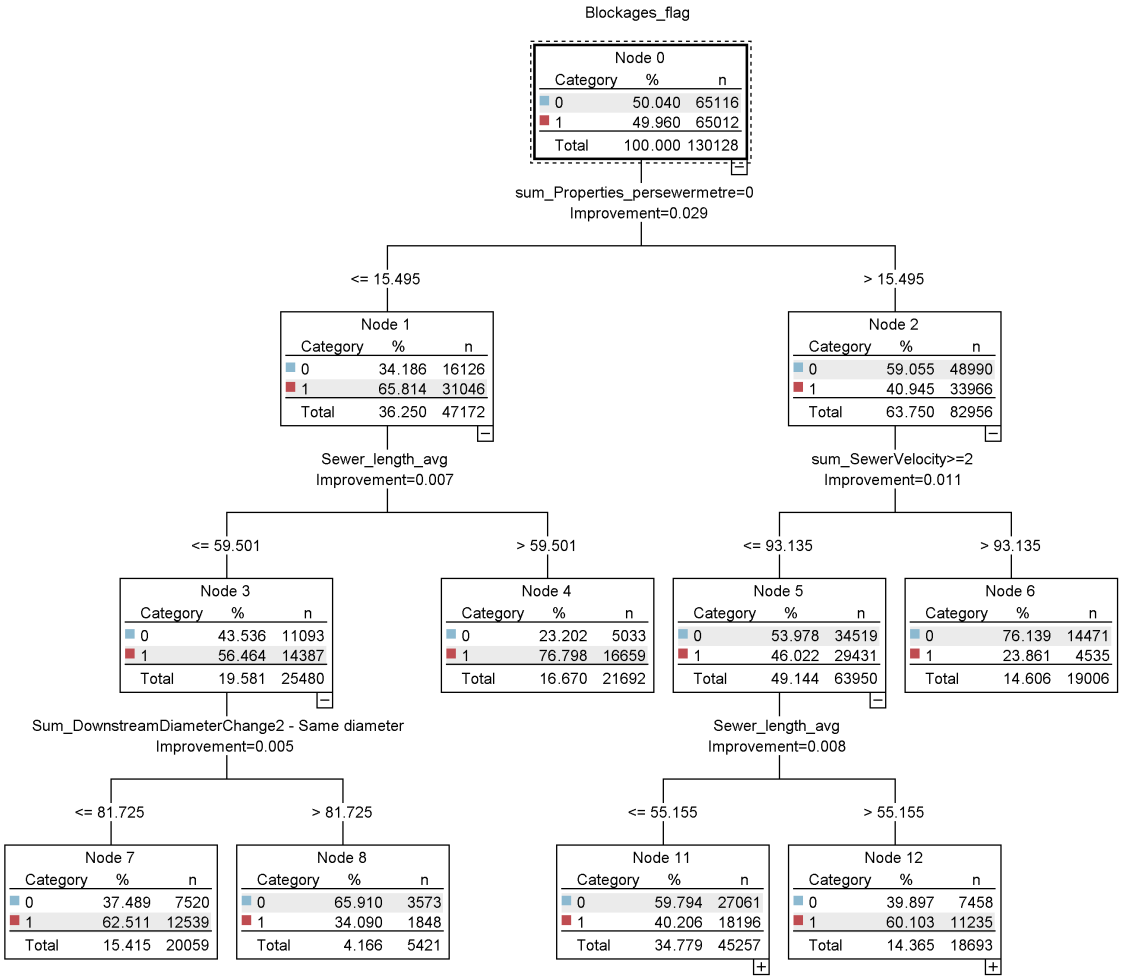


(a) Decision Tree, showing the top four layers of the tree.

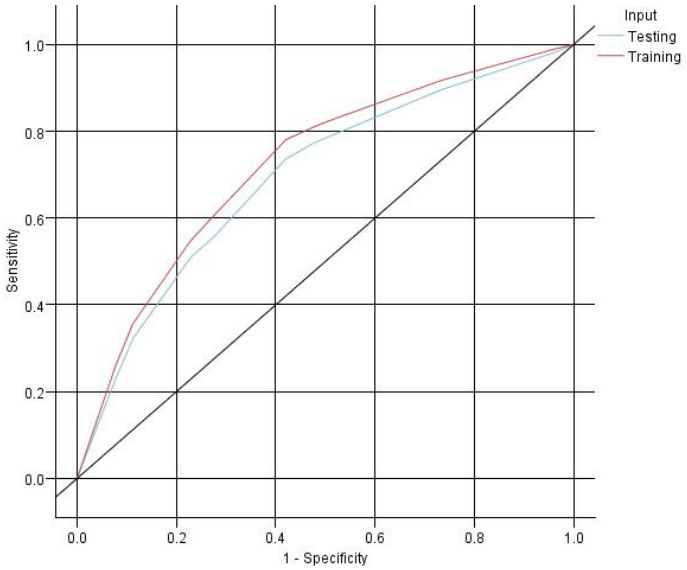


(b) ROC curve

Figure 49: Results obtained from the area level models, using a threshold in the relative blockage proportion of 6.



(a) Decision Tree, showing the top four layers of the tree.

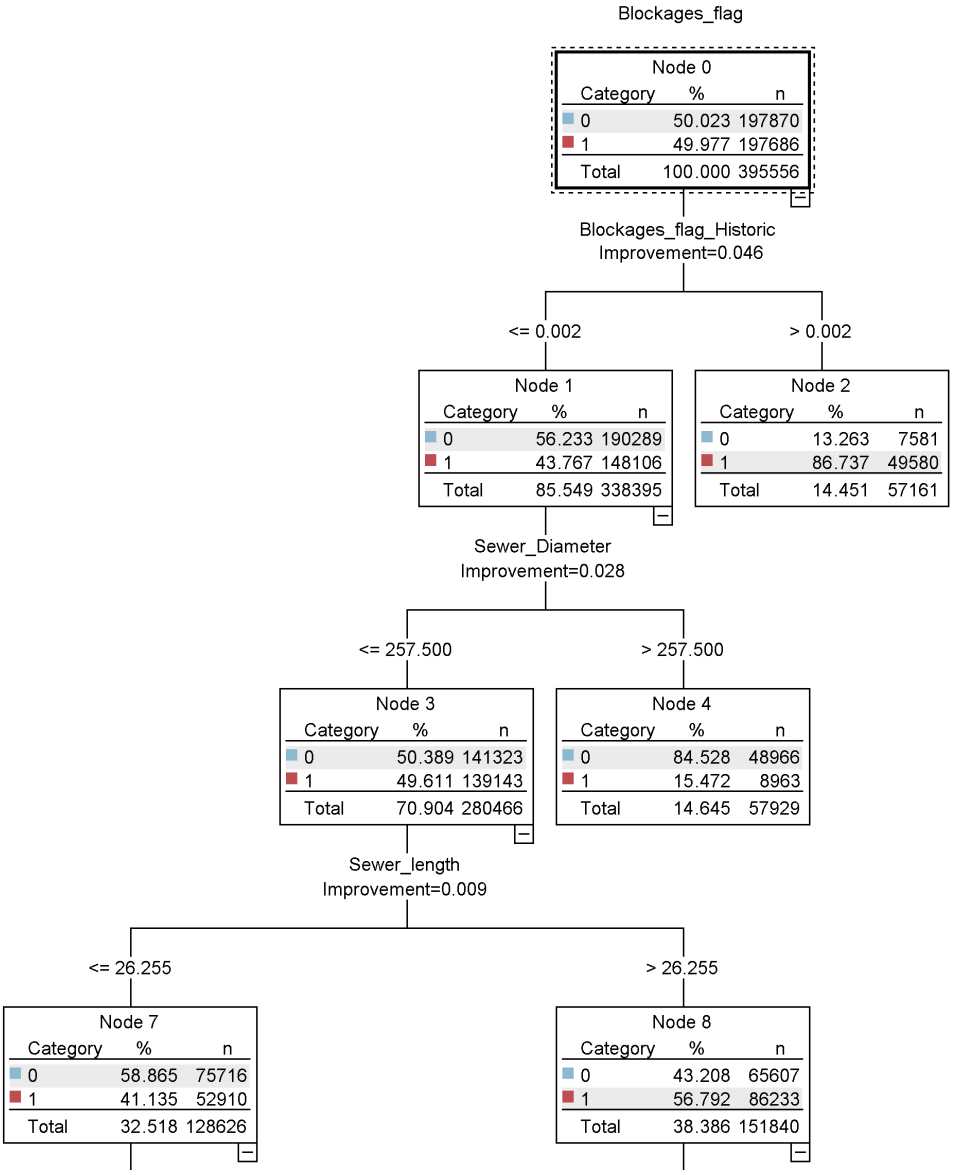


(b) ROC curve

Figure 50: Results obtained from the area level models, using a threshold in the relative blockage proportion of 8.

Appendix E

Historical Input Feature - Decision Trees



(a) Decision tree for model 1, built using two years of data, including an input feature

Figure 51: Decision trees produced using a historical input feature.