Available online at www.sciencedirect.com

ScienceDirect

Procedia Engineering 119 (2015) 1288 – 1298

Procedia Engineering

www.elsevier.com/locate/procedia

13th Computer Control for Water Industry Conference, CCWI 2015

Predictive risk modelling of real-world wastewater network incidents

James Bailey [a, b, *], Edward Keedwell [a], Slobodan Djordjevic [a], Zoran Kapelan [a], Chris Burton [b] and Emma Harris [b]

[a] *University of Exeter, Exeter, UK*

[b] *Dŵr Cymru Welsh Water (DCWW), Pentwyn Road, Nelson, Treharris, Mid Glamorgan CF46 6LY, UK*

Abstract

Due to growing pressure on wastewater network operators to deliver improved serviceability and lower costs to customers, there is a real need for greater understanding of the factors which influence incident rates, enabling effective prioritisation of proactive maintenance. This paper applies decision trees to investigate both static factors, such as sewer material and diameter, and derived factors, such as sewer velocity, for the prediction of blockages on the network of Dŵr Cymru Welsh Water. The results obtained illustrate the effectiveness of the proposed approach when identifying important explanatory factors and predicting sewers that are likely to block.

Peer-review under responsibility of the Scientific Committee of CCWI 2015

Keywords: Data mining; Wastewater; Networks; Blockages; Asset failure; Real-world;

1. Introduction

Blockages form the majority of incidents which occur on wastewater networks, representing a large total cost to water and sewerage companies (WaSC) to reactively clear. Blockages are also responsible for flooding and pollution incidents [1], and contribute to spills from Combined Sewer Overflows (CSOs) [2], resulting in further costs and

* Corresponding author. *E-mail address:* j.bailey@exeter.ac.uk

Peer-review under responsibility of the Scientific Committee of CCWI 2015
doi:10.1016/j.proeng.2015.08.949

impact on the environment and customers. Further impact on WaSC occurs through the Service Incentive Mechanism (SIM) scoring [3] of customer service provided, with calls regarding blockages all contributing to a company's SIM score. This influences the price review and is an area where further performance challenges have been set by OFWAT for Asset Management Plan (AMP) 6 [4]. These combine to drive a continued desire to reduce the number of blockages which occur. The OFWAT Final Determinations for AMP 6 also mean companies must provide an average 5% drop in household water bills over the next five years, adding to the cost pressures faced by WaSC. The result of this has been a move towards proactive maintenance of the wastewater network, to reduce the reactive and total cost to each company. This has increased the desire to understand where and when blockages occur for the most effective prioritisation of work. The relatively small cost to clear a single blockage when compared to the cost of proactive maintenance also contributes to the need for an accurate prediction of risk.

The wastewater network, however, represents a very large network of sewers, with a low coverage of telemetry, and a small number of blockages in comparison to its size. Blockages can also result from many different causes, including: fat, oil and grease (FOG) or sediment (termed chronic blockages [5]), which build up over time and could be proactively cleansed to prevent a blockage occurring, and rags or wipes which can cause a sudden and complete blockage of the sewer (termed acute blockages [5]). Given the size of the network and the amount of data held about each sewer within it, this problem presents an opportunity for data mining techniques to be utilised to find patterns in the occurrence of blockages allowing greater understanding of the important contributory factors and predicting future risk.

The application of these techniques on the real-world data of a WaSC presents a number of challenges related to the selection of the correct datasets for analysis, the handling of imperfect or incomplete data and the fusion of datasets from different sources into a format for modelling. Within the asset database there is also missing data within the variables held and missing information which would be of use for predicting blockages, requiring the infilling of missing data and sourcing of other data. Further issues are caused by the incident datasets, where there is incomplete linking of incidents to assets, and the lack of a consistent methodology for classifying the cause of an incident, limiting the information which can be gained. In recent years the asset base of WaSC's has also been increased by the adoption of previously private sewers through Private Sewer Transfer (PST) [6] and for which there is a lack of historical incident data and for which many are not present within WaSC's geographic information systems (GIS). This paper aims to use data mining technique(s) for analysis of real-world datasets, improving the understanding of the factors which are related to blockages and allowing a prediction of the blockage risk.

A number of studies have been conducted previously with regard to blockages, which have used statistical [7] [8] [9] [10] [11] [2] as well as data mining [7] [12] [13] [1] techniques within different decision support systems (DSS) to understand blockage risk. Arthur et al. [10] [2] [14] aimed to understand the factors, beyond historical incidence, influencing the risk of blockage by using a statistical comparison of the proportion of assets which have blocked. This technique allowed the identification of factors related to blockages, including: smaller diameter, combined sewers of CCTV pipe grade 4 or 5 which do not meet self-cleansing criteria in areas of high population density. Arthur et al. [10] also investigated the inference of age and the use of this for predicting network incidents. Maps from different stages in the development of Edinburgh were used to infer sewer age, which when combined with data on complaints, showed that more recent developments had lower numbers of complaints. The factors related to blockages were also investigated for the development of a model for blockage occurrence using regression analysis as part of the Cost-S Whole Life Cost Modelling project [8] [15]. The analysis compared a normalised blockage rate (blockages per km) with the factors that were believed to be important. Blockage rate was found to decrease with increasing diameter, with gradient and sewer material not showing any clear correlation. These types of statistical analysis allow the relative effect and importance of each of the factors investigated to be evaluated, but do not provide information on any interactions between variables.

A number of studies have also utilised Evolutionary Polynomial Regression (EPR) [12] [16] [17] [18], a hybrid genetic programming and numerical regression tool, to produce relationships for the number of blockages. This has included the use of sewer diameter and mean slope alone to generate a model, showing a good correlation to the rate of blockage [12]. The technique was also used with a larger number of variables, including the number of properties, area of 'hazardous' soil, mean sewer age, surveyed pipe grade and length of Section 24 sewers (those adopted as part of the 1936 Public Health Act). A number of models were generated which showed the length of Section 24 sewers as the strongest predictor of blockage rate. Ugarelli et al. [7] also used EPR with the inputs of sewer age,

length and number of sewers, diameter and slope to predict blockages, for a number of defined classes of pipe. This analysis showed that age, sewer function (foul, combined, surface water) and diameter were strong predictors of the rate of blockage. UKWIR [1] also used EPR to develop models for two case-study catchments, using the data which could be sourced from the water companies' datasets. The model produced was based on the length of sewers of condition grade 4 or 5, sewer age (derived using maps of development) and length of Section 24 sewers. The different case studies using EPR show the ability of this technique to develop accurate performance models, as well as show the explanatory factors on which the models are based. This paper applies decision trees to this problem, with the aim of producing accurate models and gaining an understanding of the important explanatory factors.

2. Methodology

2.1. Data Sources

Data was sourced from Dŵr Cymru Welsh Water's (DCWW) asset databases, with data from the whole of their area of responsibility used to develop models. The data sources used for the analysis were the database of sewers (sewer material, diameter, age), properties (location), historical incidents (blockages, flooding, pollution), location of food producers and postcode level data of ACORN classification, and the property types (terraced/detached/semi-detached) and ages present. The aim was to compile the largest set of variables which would provide explanatory capability, allowing the data mining to be completed to find potential patterns in this data. This resulted in a dataset for analysis of around 700 000 sewers, around 22 000km in length, covering most of Wales and part of England, with around 130 000 blockages, from 8 years of regulatory return data.

2.2. Data Preparation

The initial preparation involved the removal of duplicated values within the sewer and incident datasets, and cleaning of the data to remove any records which had the default value for the field or were outside of the expected range. For the fusion of all of the datasets each sewer also required the addition of spatial references (a postcode and 100m grid reference) for linking to the geographical data sources.

Following an analysis of the level of data completeness within the different sources, a period of data preparation was undertaken to infill missing data and derive some of the variables believed to be relevant. This included using multiple linear regression [19] to infill the sewer gradient, the derivation of property density using a spatial (i.e. GIS type) query of the property dataset and the derivation of the number of food producers connected to each sewer. Further data preparation was conducted to link blockage incidents to a sewer asset, a field poorly completed within the dataset, where all incidents are linked to a property rather than asset location, and where some incidents are incorrectly linked to a particular asset. This linking was achieved using a proximity analysis [8] between the property location listed for each blockage and the sub-set of sewers to which properties may be connected, allowing the majority of incidents to be assigned to an asset.

Further variables were derived as they were believed to offer greater explanatory capability. This included using the Manning Formula to calculate a sewer velocity under the normal depth assumption, based on the gradient and diameter which was then, in turn, compared to critical velocity to indicate a deposition potential; a flag for diameter changes downstream and the normalisation of properties and food producers connected to each sewer by the length of sewer, and by the length and square of the diameter. These variables were derived for the decision tree modelling, following the statistical evaluation.

2.3. Statistical Analysis

To evaluate the significance of the potential explanatory variables analysed, and provide a comparison between the occurrence of the different incident types, a preliminary statistical analysis was completed. This analysed continuous (e.g. property density, sewer diameter, sewer length) and categorical fields (e.g. sewer function, ACORN Category, terraced properties) to calculate a Pearson correlation coefficient, and to find the difference in average between the categories, respectively. The outcomes were checked by the use of statistical significance testing. This

analysis was completed for different classes of sewer – classified by public and Private Sewer Transfer (PST), and for each of the incident types analysed: blockages, flooding and pollution.

2.4. Data Mining

Data mining was conducted using Decision Trees [20] to predict whether a sewer has blocked historically (based on the eight years of incident data for public sewers and one year for PST sewers) or not, treating each sewer individually, without further aggregation. Decision Trees were chosen because they allow the modelling of blockage risk, with a human-understandable output of the contribution of each factor to the model, visually identifying the important explanatory variables. IBM SPSS Modeler [21] was used to conduct the data mining, with the software's Classification and Regression, and C5.0 trees [22] used to produce the outputs. Boosting of the less common categories and misclassification costs were used to increase the proportion of sewers flagged as blocked by the models. The testing: training split used was 70:30, produced using SPSS Modeler's Partition Node (i.e. randomly). Each model was evaluated in SPSS Modeler through the production of a Receiver Operator Characteristic (ROC) curve, calculation of the associated area underneath and a percentage accuracy of classification.

3. Results and Discussion

3.1. Introduction

The analysis completed below gives an understanding of the important explanatory factors for predicting blockages, along with a comparison of the factors which result in a blockage causing flooding or pollution. The analysis also demonstrates the potential of decision trees for this application, which have been used to model the blockage flag and allowed a greater understanding of the contributory factors and their relative importance, as applied to blockages overall, and to the different mechanisms of blockage.

3.2. Statistical Analysis

3.2.1. Explanatory Factors

For public blockages, Figures 1 and 2, the analysis showed smaller, older and shorter sewers in areas of higher property density were linked to a higher rate of blockage. As it can be seen from these figures, smaller diameter sewers are more easily blocked by large items in the sewer, increasing blockage risk. Older sewers are likely to have a greater number of defects, which disrupt the transport of material, causing or building-up to cause a blockage, and may represent sewers built to different design standards.

Of these highlighted variables, diameter has previously been found to be a strong predictor of blockages [7] [8], along with age [7] [1], in those studies which have been able to source this data, affirming the results found here. Hafskjold et al. [5], however, found manufacturing and construction standards to be more important than age itself. In addition to the construction date field shown in Figure 1, the Earliest Property Age, representing the age of the oldest properties within the sewer's postcode, is listed in Figure 2, suggesting its potential as an explanatory factor in the absence of sewer age, as was also found by Arthur et al. [10] and UKWIR [1].

The increased risk from shorter sewers was linked to the potentially increased presence of manholes, which are believed to increase blockage risk [5]. Other investigations have found mixed results for the influence of sewer length. Hafskjold et al. [5] found that 20% of blockages within the Trondheim catchment studied occurred in manholes. Savic et al. [12] investigated sewer length and found that the fragmentation of the sewer (i.e. greater number of shorter lengths) was linked to increased collapse but not blockage risk. Ugarelli et al. [7] found an unclear relationship with sewer length, where different models produced different relationships between length and blockages. Property density is linked to an increased density of sewer connections and material moving through the sewers, and has been found to influence blockage risk [2][23]. From this, there is an increased risk of material entering the sewer which suddenly blocks the sewer and has been linked to potential defects, with poorly fitted connections acting as defects, increasing blockage risk [23]. Arthur et al. also investigated property density and found a link to an increased risk of blockage [2].

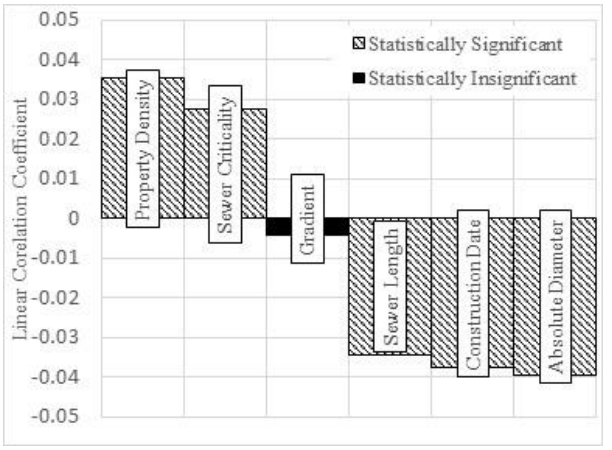

Figure 1 – Chart showing the linear correlation coefficient between the continuous fields and blockage rate per km per year. The bars are shaded differently to highlight whether the result was found to be statistically significant. The fields are ordered on the chart by the size of the correlation coefficient, from most positive on the left to most negative on the right. This chart was produced using the dataset of 465 633 public sewers.

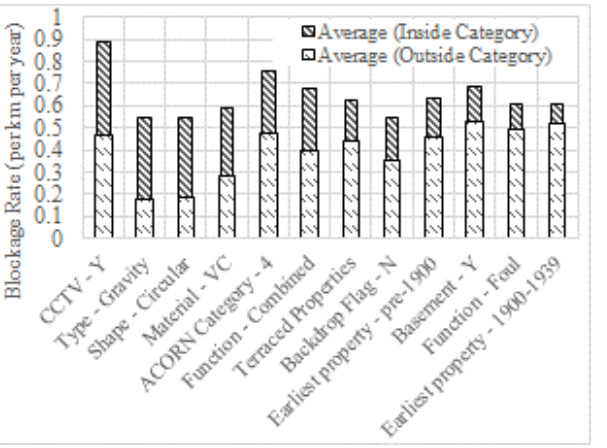

Figure 2 – Chart showing the average incident rate inside each category and for the other results, outside of the category. The categories shown in the chart are those where the average incident rate in the category was higher than the overall average rate, for a result found to be statistically significant. The results are ordered from the largest absolute difference in rate on the left to the smallest absolute difference on the right. This chart was produced using the dataset of 465 633 public sewers.

3.2.2. Comparison of Explanatory Factors for Different Incidents

A comparison was also made between blockages, and pollution and flooding, to investigate explanatory factors which lead to these, and the greater impact they produce. This allows the improved understanding of risk and prioritisation of proactive maintenance. Figures 3 and 4 show that very similar variables have been highlighted for flooding risk, again linked to sewer diameter, length and age, in areas of higher property density.

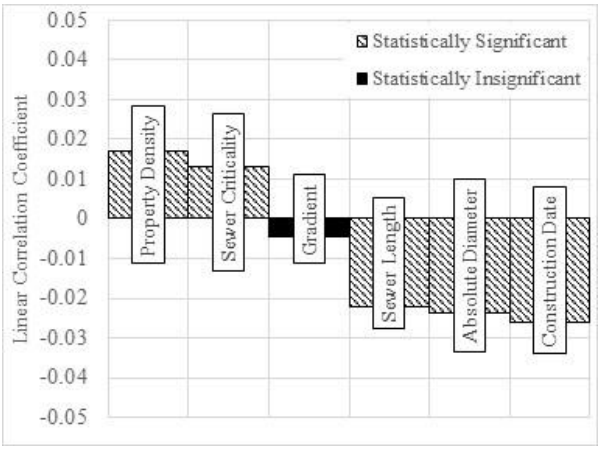

Figure 3 – Chart showing the linear correlation coefficient between the continuous fields and flooding rate per km per year. The bars are shaded differently to highlight whether the result was found to be statistically significant. The fields are ordered on the chart by the size of the correlation coefficient, from most positive on the left to most negative on the right. This chart was produced using the dataset of 465 633 public sewers.

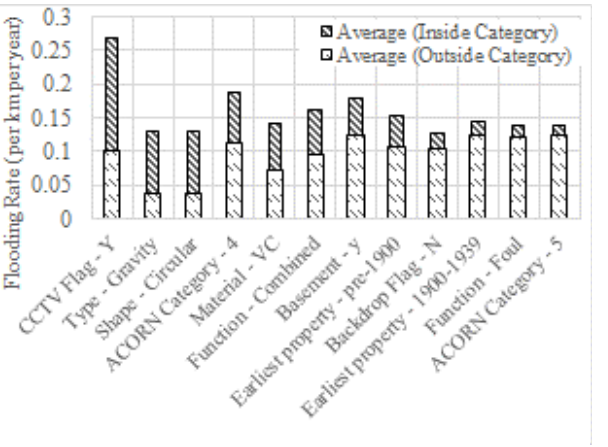

Figure 4 – Chart showing the average incident rate inside each category and for the other results, outside of the category. The categories shown in the chart are those where the average incident rate in the category was higher than the overall average rate, for a result found to be statistically significant. The results are ordered from the largest absolute difference in rate on the left to the smallest absolute difference on the right. This chart was produced using the dataset of 465 633 public sewers.

The comparison to pollutions, Figures 5 and 6, show even weaker correlations, due to the lower number of historical pollution incidents for analysis. This analysis highlights the results for property density and for sewer diameter. The largest correlation found is to property density, where the negative coefficient is in contrast to that found for blockages and flooding, with sewers flagged as rural showing the largest difference in average pollution rate (Figure 6). This suggests it is in sewers away from houses that pollutions occur, agreeing with previous analysis completed for DCWW [24]. This was related by DCWW to the distance to watercourse, potential obstructions preventing sewer escape entering a watercourse or the probability of the escape being noticed before entering a watercourse, all of which would affect the probability of a blockage causing a pollution. The categorical analysis also shows that it is larger diameter sewers which are related to an increased pollution risk, again in contrast to blockages and agreeing with previous analysis [24], which suggested the sewers most likely to block do not coincide with those most likely to pollute.

4. Data Mining

Once the preliminary statistical investigations had been completed, it was decided to use Decision Trees to produce models of blockage, initially predicting the flag of whether the sewer has ever blocked. This was begun by modelling the whole dataset, which showed initial splits on sewer ownership (Public / PST) and sewer function (foul or combined / others), due to the differing amounts of historical data and types of sewers where blockages are most likely to occur, respectively. It was, therefore, decided to produce models for the different sewer functions and for different blockage formation mechanisms, as shown in Table 1.

A number of other potential explanatory factors were also derived at this stage, including: sewer velocity, calculated using the Manning formula [26], and normalised values for property and food producer connections, as outlined in section 2.2. The similar measures derived were evaluated by producing decision trees for public combined sewers, assessing overall model prediction accuracy and the presence of the variables in the trees as a measure of their significance as explanatory variables. This analysis showed that the use of sewer velocity, properties per sewer metre and food producers per sewer metre per diameter squared produced the most accurate models. These models make sense from the engineering point of view. For example, if property connections are potential defects then their density along the length of the sewer would impact the risk of blockage with the diameter of the sewer not impacting the risk of being a defect. Whereas, with food producers representing a load on the sewer of fat, oil and grease (FOG), then

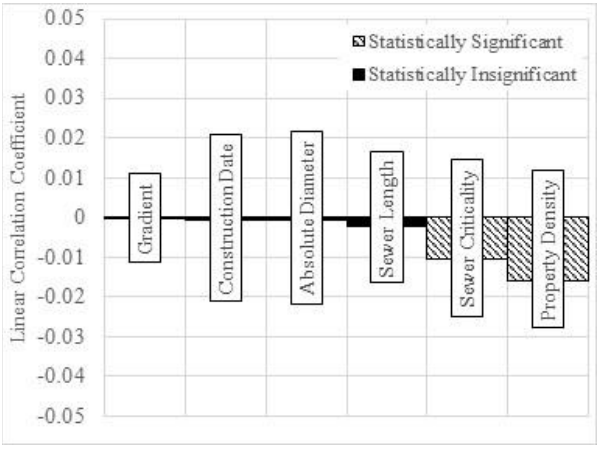

Figure 5 – Chart showing the linear correlation coefficient between the continuous fields and pollution rate per km per year. The bars are shaded differently to highlight whether the result was found to be statistically significant. The fields are ordered on the chart by the size of the correlation coefficient, from most positive on the left to most negative on the right. This chart was produced using the dataset of 465 633 public sewers.

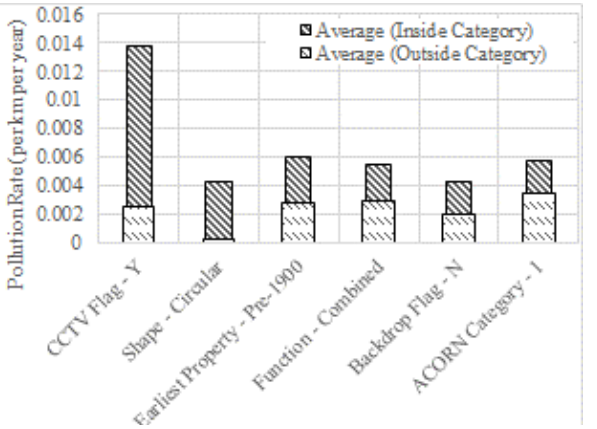

Figure 6 – Chart showing the average incident rate inside each category and for the other results, outside of the category. The categories shown in the chart are those where the average incident rate in the category was higher than the overall average rate, for a result found to be statistically significant. The results are ordered from the largest absolute difference in rate on the left to the smallest absolute difference on the right. This chart was produced using the dataset of 465 633 public sewers.

Table 1 – performance evaluation completed for produced models, showing overall model accuracy and area under the curve (AUC) for the Receiver Operator Characteristics (ROC) curves plotted [25]. The ROC curve and confusion matrices for the public, combined sewer model (Figure 8) are shown in Figure 7 and Tables 2 and 3.

Model	Accuracy	AUC	Model	Accuracy	AUC
Public - foul	64%	0.65	Blockages due to silt	65%	0.62
Public - combined	65%	0.69	Blockages due to debris	60%	0.68
PST - foul	65%	0.72	Blockages due to nappies/wipes/rags	54%	0.65
PST - combined	62%	0.66	Blockages due to fat	65%	0.66

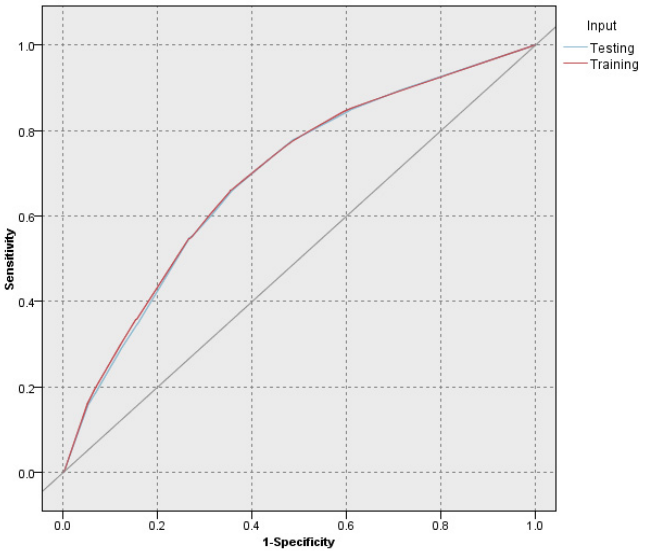

Figure 7 – ROC Curve showing the output using Testing and Training data for the Decision Tree (Figure 8) produced using public, combined sewers.

Table 2 – Confusion matrix for the Decision Tree (Figure 8) produced for public, combined sewers, shown for the training data.

Classified as	0	1
0	91,191	4,765
1	48,224	9,080

Table 3 – Confusion matrix for the Decision Tree (Figure 8) produced for public, combined sewers, shown for the testing data.

Classified as	0	1
0	39,163	2,062
1	20,879	3,889

normalising the load by the overall sewer capacity (length and cross-sectional area) would be expected to be the best method. For a measure of a sewer's self-cleansing ability, sewer velocity was compared to a self-cleansing flag (gradient < 1: pipe diameter) and a flag of whether the sewer velocity meets a self-cleansing velocity of 1 m/s [27] , with the measure of sewer velocity best representing this self-cleansing ability. This analysis allowed the most significant explanatory variables in each group (property connections, food producer connections, self-cleansing ability) to be selected and used in the models. This prevented the similarity of the information provided by the variables from hindering the growth of the decision trees, were all variables from the same group to be included in the modelling.

The first blockage prediction models generated were those for the public sewers, which have more historical data and are a larger dataset than PST sewers, an output of which is shown in Figure 8. Classification and Regression Trees were used to produce the model for combined sewers. This resulted in models with an area under the curve (AUC) of 0.69, respectively, on the training data used. Misclassification costs were used in this case to force the model to increase the prediction of a positive blockage flag.

The C5.0 algorithm, with boosting of the positive blockage flag records, provided the best results for modelling the PST sewers, with an AUC of 0.72 and 0.66 for foul and combined sewers. These levels of accuracy suggest potential for this method of predicting blockage risk and prioritising proactive maintenance, although given the cost constraints faced in clearing blockages, further work may be required to improve prediction accuracy for this application. Overall, a good model accuracy and the provision of blockage prediction models where the contributory factors can easily be understood shows that Decision Trees provide a very useful tool for understanding the important factors related to blockages.

For the public sewer models, the most significant explanatory variables, forming the initial splits in the two trees, are identified as follows: properties per sewer metre, sewer velocity, length, diameter and property density, which are formed from the sewer length, diameter, gradient and data on properties. This affirms previous results [7] [12], which achieved high levels of prediction accuracy using these basic sewer characteristics. Data on the number of properties has been investigated by Arthur et al. [2], who found an increased blockage risk, and by Savic [13] who included property data for modelling, although it did not appear in any of the models produced.

For PST sewer models, the expectation was that the lack of multiple years of historical incident data would limit the explanatory capability of any model produced. Although there is a less consistent set of variables used to form

the models, the performance in terms of accuracy and AUC is similar to that of the public sewer models. The use of sewer length to split, which is used in both models at the top of the tree, provides limited additional information, with a longer sewer more likely to have blocked. Proactive maintenance based on this would focus on the longest lengths of sewers, not accounting for the risk posed by multiple shorter lengths of sewer.

As shown in Table 1, blockage prediction models were also produced for the different mechanisms of blockage formation, using the dataset of public, combined sewers, to provide the largest dataset for modelling. The different causes of incidents listed in the incident dataset were combined to produce the 5 categories used, based on the grouping of similar mechanisms. For FOG type blockages the most significant explanatory factors are sewer length, property connections and sewer velocity. Properties per sewer metre may suggest the increased FOG load on the sewer due to the properties connected, although its presence below the first split, as in the overall model, may suggest a lower importance when compared to predicting any type of blockage. The presence of sewer length may also indicate a role in the formation of FOG blockages with a longer sewer length providing more time for any FOG to settle out and form a blockage. Sewer velocity is linked to a larger velocity reducing the probability that FOG will settle out and increasing the probability that the flow at a given time could dislodge and transport any settled FOG. There is little previous work that investigates the use of historical data to predict blockages through different mechanisms. Modelling the different mechanisms is made more difficult by the potential lack of consistency in the classification of incidents.

The blockage prediction model for nappies, wipes and rags has properties per sewer metre as the first split, representing the increased load on the sewer and link to the source of nappies, wipes and rags. Below this, sewer velocity is present, representing the capacity of the sewer to transport these materials. Sewer diameter is also present, with smaller sewers more likely to suffer an acute blockage, where a larger diameter sewer would require a larger build-up of material to block. However, some sewer transport models have shown that a solid larger in relation to the pipe will receive more of the energy of the flow in the sewer and be transported further through the sewer [23]. Sewer diameter may be expected to appear higher up the hierarchy of factors in the tree, due to the perceived lower propensity of larger diameter sewers to block by this mechanism, although the factor does appear higher than in the overall model, but may be shown to be less important than the load on and transport capacity of the sewer, represented by the property and velocity data.

The blockage prediction model for silt contains sewer length and construction date, with a longer length potentially allowing silt to settle out and construction date being linked to potential defects, impacting the flow and transport capacity of the sewer. The other two models, for debris and other causes, are more difficult to interpret, with the respective physical reasons for blockages less clear.

5. Conclusion

Decision trees were used to produce models of the indicative risk of blockage on the real-world wastewater network of DCWW. A number of relatively accurate blockage prediction models have been produced which demonstrate the efficacy of using Decision Trees to find patterns in the large datasets, provide further understanding of the most useful explanatory factors and allow the prioritisation for proactive maintenance. The modelling has shown that some of the basic characteristics of the sewer (length, diameter, gradient), along with property data, can provide good explanatory capability of the risk of blockage occurrence. In addition, models of different blockage mechanisms has added further understanding of the factors influencing these.

Further work to be conducted could include the prediction of blockage rate, giving a better measure of blockage risk, and the inclusion of CCTV survey data, historical blockage data, and temporal variables, such as rainfall and planned maintenance, to improve model accuracy and allow prioritisation of the timing of proactive maintenance.

6. Acknowledgements

The work has been conducted as part of a Knowledge Transfer Partnership (KTP) with funding provided by Innovate UK and Dŵr Cymru Welsh Water (DCWW), working in collaboration with the University of Exeter's Centre for Water Systems (CWS).

Appendix A.

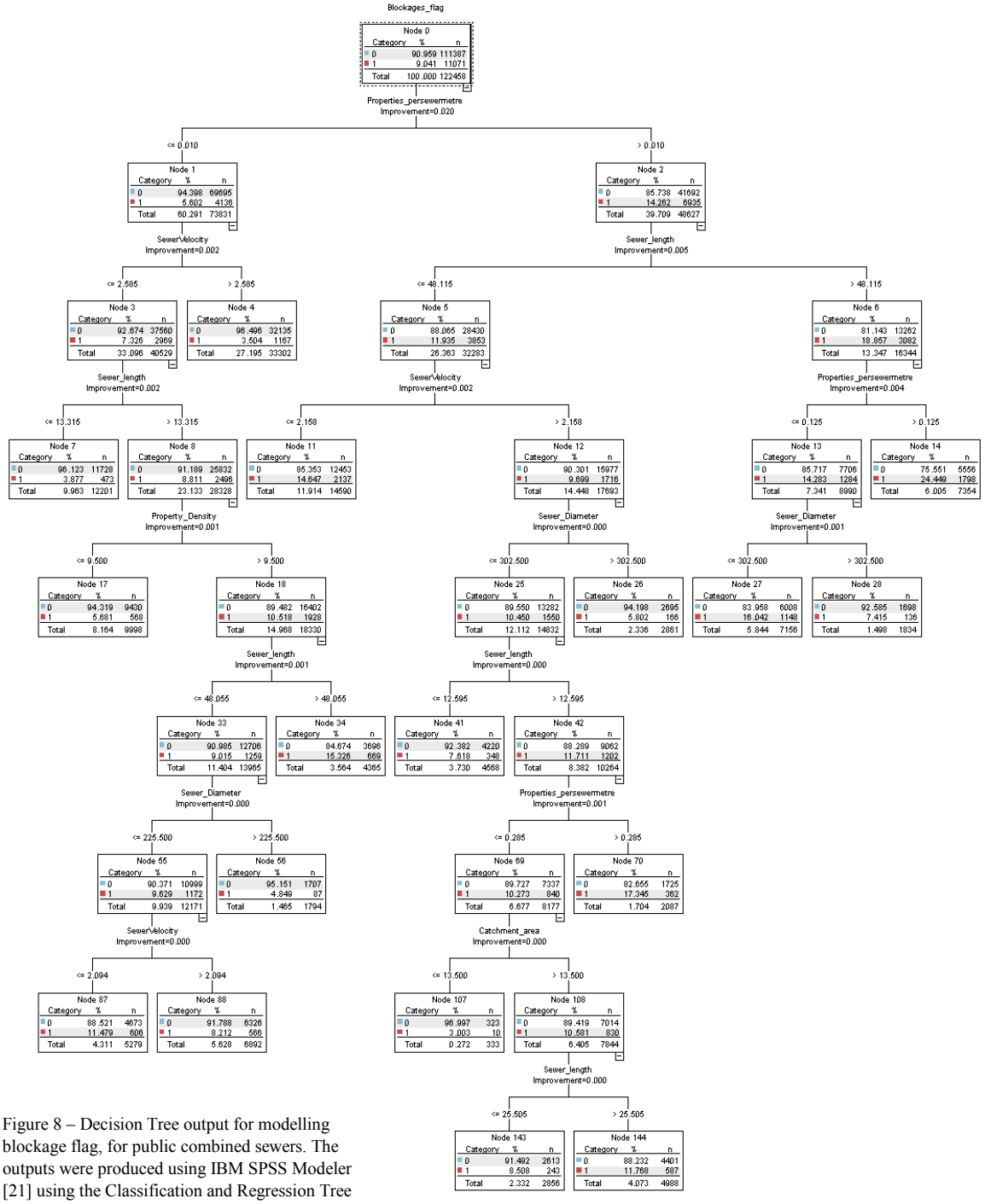

Figure 8 – Decision Tree output for modelling blockage flag, for public combined sewers. The outputs were produced using IBM SPSS Modeler [21] using the Classification and Regression Tree [22].

References

[1] M. Hall, Z. Kapelan, R. Long, D. Savic, Deterioration Rate of Sewers, UKWIR, UK Water Industry Research Limited, 2005/6.

[2] S. Arthur, H. Crow, L. Pedezert, Understanding Blockage Formation in Combined Sewer Networks, *Proceedings of the Institution of Civil Engineers*, 2008

[3] OFWAT, Service incentive mechanism, [Online] [Cited: May 22, 2015.] https://www.ofwat.gov.uk/regulating/aboutconsumers/sim.

[4] — Setting price controls for 2015-20 Final price control determination notice: company-specific appendix – Dŵr Cymru. OFWAT Final determinations, [Online] [Cited: May 28, 2015.] https://www.ofwat.gov.uk/pricereview/pr14/det_pr20141212wsh.pdf.

[5] L.S. Hafskjold, A. Kønig, S. Sægrov, W. Schilling, Improved assessment of sewer pipe condition, *CityNet 19th European Junior Scientist Workshop*, 2004

[6] OFWAT, Transfer of private sewers, [Online] [Cited: May 22, 2015.] https://www.ofwat.gov.uk/consumerissues/rightsresponsibilities/sewers/prs_web_sewertransfer.

[7] R. Ugarelli, S.M. Kristensen, J. Røstum, S. Sægrov, V. Di Federico, Statistical analysis and definition of blockages-prediction formulae for the wastewater network of Oslo by evolutionary computing, *Water Science and Technology*, 2009, Vol. 59, pp. 1457-1470.

[8] W. Shepherd, A. Cashman, S. Djordjevic, G. Dorini, A. Saul, D. Savic, L. Lewis., Investigation of blockage relationships and the cost implications for sewerage network management, *Proceedings of 10th International Conference on Urban Drainage*, 2005, pp. 21-26.

[9] R. Ugarelli, G. Venkatesh, H. Brattebø, V. Di Federico, S. Sægrov., Historical analysis of blockages in wastewater pipelines in Oslo and diagnosis of causative pipeline characteristics, *Urban Water Journal*, 2010, Vol. 7, pp. 335-343.

[10] S. Arthur, R. Burkhard, Prioritising sewerage maintenance using inferred sewer age: a case study for Edinburgh, *Water Science and Technology*, 2010, Vol. 61, p. 2417.

[11] R. A. Fenner, L. Sweeting, M.J. Marriott, A new approach for directing proactive sewer maintenance, *Proceedings of the ICE-Water and Maritime Engineering*, 2000, Vol. 142 (2), pp. 67 - 77.

[12] D. Savic, O. Giustolisi, L. Berardi, W. Shepherd, S. Djordjevic, A. Saul, Modelling sewer failure by evolutionary computing, *Proceedings of the ICE-Water Management*, 2006, Vol. 159, pp. 111-118.

[13] D. Savic, The use of data-driven methodologies for prediction of water and wastewater asset failures, *Risk Management of Water Supply and Sanitation Systems,* pp. 181-190.

[14] S. Arthur, H. Crow, L. Pedezert, Understanding blockage formation in sewer systems a case-by-case approach, *Water Manage. Proc. Inst. Civil Eng*, Vol. 161, pp. 125-221.

[15] D. Savic, S. Djordjevic, G. Dorini, W. Shepherd, A. Cashman, A. Saul, COST-S: a new methodology and tools for sewerage asset management based on whole life costs, *Water Asset Management International*, 2005, Vol. 1, pp. 20-24.

[16] D. Savic, O. Giustolisi, D. Laucelli, Asset deterioration analysis using multi-utility data and multi-objective data mining, *Journal of Hydroinformatics*, 2009, Vol. 11, pp. 221-224.

[17] L. Berardi, Z. Kapelan, Multi-Case EPR strategy for the development of sewer failure performance indicators, *Proc. World Environmental and Water Resources Congress*, 2007, Vol. 10, p. 243.

[18] L. Berardi, O. Giustolisi, D. Savic, Z. Kapelan, An effective multi-objective approach to prioritisation of sewer pipe inspection, *Water science and technology*, 2009, Vol. 60, p. 841.

[19] D. Pyle, *Data Preparation for Data Mining,* San Francisco, Morgan Kaufmann Publishers, 1999.

[20] P. Giudici, *Applied data mining: statistical methods for business and industry,* Chichester, Wiley, 2002.

[21] IBM, SPSS Modeler Version 15, [Software], Used under license held by DCWW.

[22] —. IBM SPSS Modeler 15 Algorithms Guide, SPSS Modeler 15.0 Documentation, [Online] [Cited: May 28, 2015.] ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/15.0/en/ApplicationsGuide.pdf.

[23] T.T. Hillas, Reducing the Occurrence of Flooding through the Effective Management of Sewer Blockages, University of Exeter. 2014. pp. 15-33, MPhil Thesis.

[24] Dŵr Cymru Welsh Water, Prediction Pollution Reduction - Analysis Findings, Recommendations & Programme Progress, 2012. Internal Presentation.

[25] T. Fawcett, An introduction to ROC analysis, 2006, *Pattern Recognition Letters*, Vol. 27, pp. 861 - 874.

[26] The Engineering ToolBox, Manning's Formula for Gravity Flow, [Online] [Cited: May 22, 2015.] http://www.engineeringtoolbox.com/mannings-formula-gravity-flow-d_800.html.

[27] C.H.J. Bong, A Review on the Self-Cleansing Design Criteria for Sewer System. 2014, *UNIMAS e-Journal of Civil Engineering*. Table 1.