Available online at www.sciencedirect.com

ScienceDirect

Procedia Engineering 00 (2016) 000–000

Procedia Engineering

www.elsevier.com/locate/procedia

12th International Conference on Hydroinformatics, HIC 2016

Developing Decision Tree Models to Create a Predictive Blockage Likelihood Model for Real-World Wastewater Networks

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

[a] *Dŵr Cymru Welsh Water (DCWW), Pentwyn Road, Nelson, Treharris, Mid Glamorgan CF46 6LY, UK*
[b] *University of Exeter, College of Engineering, Mathematics and Physical Sciences, Exeter, UK*

Abstract

To reduce the blockages occurring on wastewater networks, reducing costs, customer and environmental impact, greater levels of proactive maintenance are being conducted by water and sewerage companies. For effective prioritisation of this maintenance, an accurate model of blockage likelihood is required. This paper presents the development of a model, for provision of a blockage likelihood level and verification using unseen data, based on previous decision tree models constructed using the asset and historical incident data from the wastewater network of Dŵr Cymru Welsh Water. The model has been developed here using the geographical grouping of sewers and the application of ensemble techniques, with the results illustrating the potential benefits which can be derived from these techniques.

Peer-review under responsibility of the organizing committee of HIC 2016.

Keywords: Sewer; Wastewater; Blockage; Likelihood; Model; Decision Trees; Ensemble;

1. Introduction

In the continued drive for Water and Sewerage Companies (WaSCs) to improve financial and service performance [1], there is the potential for increased proactive maintenance to prevent incidents before they occur and remove the impact on customers. To maximise the benefit from this approach there is a need for well prioritised proactive maintenance. A number of investigations [2] [3] [4] have been completed with the aim of utilising the available data on the incidents and assets of WaSCs to help predict incidents using data-driven modelling

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

Peer-review under responsibility of the organizing committee of HIC 2016.

techniques. This paper presents the development of decision tree models produced using the asset and incident data from the wastewater network of Dwr Cymru Welsh Water (DCWW) to predict the likelihood of blockage, and inform the prioritisation of proactive maintenance. This paper presents work investigating geographical aggregation, ensemble modelling techniques and the use of a historical input feature for the enhancement of model performance. These developments are aimed at improving the performance of the models, to maximise the effectiveness of DCWW's proactive maintenance.

For WaSCs, there is a desire to gain information at the greatest resolution, ideally in the provision of predicted risk at a pipe level. Given issues with the resolution of data held, the result has previously been limited to an area based output, for example in the work of UKWIR [2] and Savic [3], although pipe level outputs have now been widely produced [5] [6] [7]. Due to the issues present in the data, for example the poor linking of incidents to assets, and the provision of a risk output which can be used by WaSCs to assign proactive maintenance, there is the potential for a geographical aggregate to provide benefits in the performance of data-driven models. Also, given the stochastic nature of blockage events and the potential influence of the surrounding network on the likelihood of blockage on any individual pipe length, geographical aggregation may also provide benefits in terms of the reduction of the noise present in the data and the representation of the surrounding network in the inputs to the model, for each area. These potential benefits are balanced against the desire for the greatest resolution of risk information. Fenner [4] purposefully used a geographical aggregation method as part of initially identifying hotspots, before combining this with the influences of the pipe characteristics to modify the aggregate level risk score, based on the characteristics of the pipe, for those areas showing the highest risk of blockage. This work's use of a geographical approach was motivated by the provision of practical areas for management by WaSCs, the consistency of input features within each area for which data was (e.g. asset age, type) and was not (e.g. soil type and vegetation) available, and reduction in the data preparation which was required. In an analogous situation, Kleiner [8] uses geographical clustering to aid in the explanation of water mains breakage rates, by using this as a surrogate for data which may be missing but geographically related, such as soil data and land development.

Ensemble modelling techniques are widely used within modelling applications, with the predictive power provided by the models within the ensemble maximised, while minimising the correlation between the models, to maximise the overall predictive power of the ensemble. This can be achieved through a variety of methods, including the manipulation of the training dataset, the input features or the output features [9]. Random Forests [10], for example, have been widely used to improve the performance of models, in applications such as ecology [11] and bioinformatics [12]. There is large variety in the cause of blockages, which can at a high level be due to acute or chronic problems, each of which may have a number of factors influencing them. Ensembles may provide a benefit in an output which can represent more of this variation and provide a greater exploration of the search space of solutions than is provided by the single decision trees used in the initial modelling work [13].

2. Methodology

The dataset used consists of the asset and incident data from DCWW's wastewater network, which was previously prepared for modelling [13] and has been adapted for the three areas of investigation. This dataset contained variables including those referring to the assets, the properties connected and adjoining network. The predictor of blockage flag was used, defined by whether a sewer had blocked within the period of historical data sourced. The sections below outline the further use and adaptation of this dataset.

2.1. Data

The data was sourced from DCWW's asset databases, covering the whole of their area of responsibility. The input features to the models included data on the assets (sewer diameter, sewer length and gradient), the proximity of properties and food producers (the number of property and food producer connections per sewer), postcode level ACORN classification and information on the property types and ages present (terraced/detached/semi-detached/basement present). Other input features were derived, including: property density and sewer velocity, calculated using the Manning formula using the normal depth assumption. The predictor feature was a flag of whether a sewer had suffered a blockage during the period of historical data held. Each incident was recorded

against the asset on which the incident had occurred, with any incidents not listed against an asset infilled using a spatial assignment to the nearest sewer, excluding any unlikely to suffer a blockage (such as surface water sewers). The flag field was derived based on whether each length of sewer was present within the dataset of incidents which had occurred, and the asset on which they had occurred.

2.2. Geographical Aggregation

To define the geographical areas to be used for modelling, a number of different areas were investigated, including: postcode, and 100m by 100m grids. The aim was to produce an area which aligned with the typical area issued by DCWW for proactive maintenance, and contained a narrow distribution of total sewer length, so that evenly sized groups were produced. Following these investigations, the final area used was that of postcode, with larger postcodes broken down by 100m by 100m grids, to reduce the frequency of the largest areas skewing the distribution. In addition, to remove areas containing single sewers, sewers were re-assigned to adjacent areas based on the connections within the network.

To define the variables for the aggregates, for categorical variables the length of network within each category was found, while for continuous variables length weighted averages were taken, as well as the length of network within bands defined by discretization, equivalent to the categorical variables. For the discretisation of the continuous variables, if breaks are present in the distribution (e.g. the consideration of sewers of diameter 225mm or less as small bore) then these were used, with remaining distributions discretised based on percentile values to produce evenly sized bands. For the definition of the predictor variable, the public and PST networks have different amounts of historical data available, and therefore show different proportions of sewers which have and have not blocked. To account for this, within each geographical group the proportion of sewers which have blocked relative to the average for that type of network (Public or PST) were found, with a length weighted average of these two relative proportions then calculated to give a continuous measure. Different thresholds within this defined relative blockage proportion were investigated, with a value of one, representing an average proportion of sewers which have blocked, used initially.

This resulted in a dataset of around 120 000 geographical areas used as the input to Classification and Regression trees. The inputted dataset was balanced by the boosting of the minority class, with training and testing datasets defined in the ratio 70:30 using random assignment. Each model was grown to maximise the performance of the testing dataset, as evaluated using a Receiver Operator Characteristic (ROC) curve and the area underneath this curve (AUC).

2.3. Ensemble Modelling

The approach taken to produce ensemble models was to use a random input, or random combination of the inputs to produce individual decision trees, and combine the outputs from each into an ensemble, using the same query definitions as used by Breiman [10]. To produce the random input queries, a random selection of the input variables was taken, with replacement, with categorical variables more likely to be selected by a factor of the number of categories. For the random input variables, the continuous variables were used as in the original, with categories within categorical variables randomly assigned as 0 or 1, for each input query. To produce the random combination queries, a random selection of variables were taken, with replacement, and used to create a new set of variables, by the linear combination of the original variable, each with a random coefficient varying between -1 and 1. For continuous variables, the z-score for each record was found based on the average and standard deviation of the training dataset, and used with the 0 or 1 output from the processing of the categorical variables, as in the random input queries.

The generated queries were used to grow Classification and Regression Trees to a maximum depth of 15, with no pruning used, for the subset of public, combined sewers, for which a model had been developed previously [13]. Testing and training datasets were defined in the ratio 70:30 using random assignment. The model outputs were combined either through a voting or raw propensity weighted voting method. The type of query, combination

method, number of inputs to each model and number of models run were all investigated to assess the impact on model performance, evaluated using the ROC curve and AUC.

3. Results and Discussion

3.1. Geographical Aggregation

The results from the models produced on the geographical areas show performance ranging between 0.65 and 0.69 for the testing AUC, as shown in Figure 1. These results compare to the results obtained on the pipe level models, for which performance ranged between 0.65 and 0.72 [13], but do not show any large improvement. The results show that better performance was obtained from the models predicting the most extreme categories used: those which showed the highest (relative blockage proportion threshold 8) and lowest (relative blockage proportion threshold 0) tendencies to block, although again the difference is not large. From the application of this work, there is a very large network with a small proportion of sewers which can be surveyed in any year, meaning the ideal output from the models would be the highly accurate prediction of a set of very high risk sewers or areas, which could be proactively maintained each year by WaSCs. Figure 2 shows the typical shape of the ROC curve for the

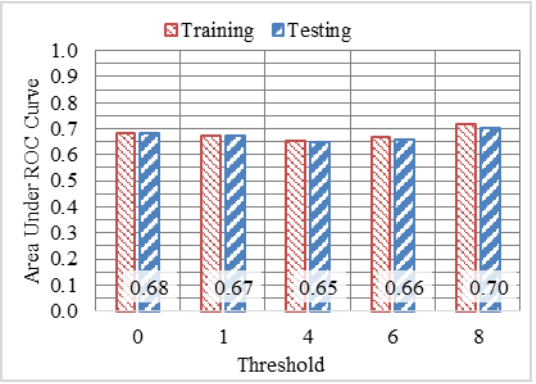


Figure 1: Chart showing the results of the models built using the geographical aggregates. The chart shows the training and testing AUC's for the different thresholds used in the relative proportion of blockages, which defined the predictor flag for the models.

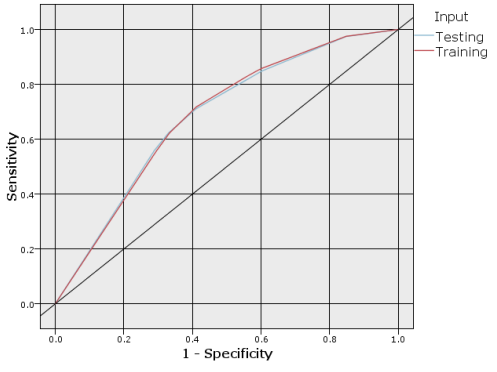


Figure 2: Chart showing the ROC curve for the model built on the geographical aggregate, using a defined threshold of 0.

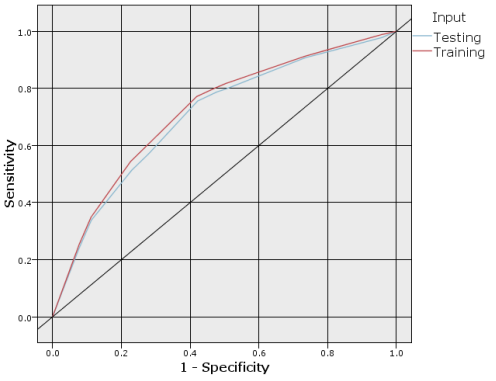


Figure 3: Chart showing the ROC curve for the model built on the geographical aggregate, using a defined threshold of 8.

models for thresholds 0 to 6, showing no distinction in terms of the risk score for around the top 35% of sewers and providing little information on those highest likelihood sewers. Figure 3 for the threshold 8 model, shows more information provided for these highest risk sewers, with the model itself using around the top 4% of highest risk areas to form the positive flag being predicted, and therefore potentially giving more useful information to WaSCs.

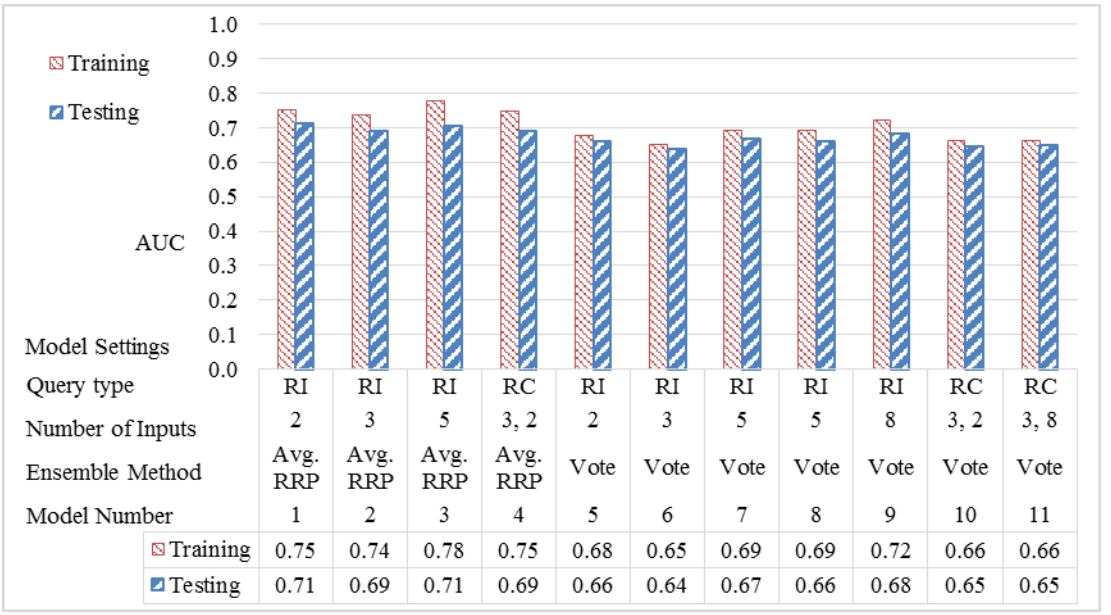
Figure 7 gives the decision tree output for the model produced using a threshold in the relative blockage proportion of 8. The variables appearing within the tree are similar to those which appeared in the other aggregate models, and those which appeared in the single decision trees. These factors include: the number of property connections to each sewer, along with the average length of sewer, the sewer velocity and the length of sewer where the downstream sewer is of the same diameter. These results show the importance of the number of property connections in explaining blockage likelihood, as was found with the single decision trees produced previously [13]. The presence of sewer velocity may show the added influence of this factor when considered within a network, where, within an area, repeated sewers of high or low sewer velocity influence the likelihood of blockage.

The postcode areas used would seem to represent a logical area for aggregation given that they form a useful size for the assignment of proactive maintenance, and that postcodes tend to follow the layout of streets, which also tend to be followed by the sewer network. This should result in adjacent parts of the network being grouped together, and give relative consistency in the input features, although unconnected parts of the network could still be grouped together. In addition, the input features to the models need to represent the characteristics of the network, and the relative effect of the characteristics of the pipes within the group on the likelihood of blockage, which would be aided by consistent input features. Further complications arise from the presence of unmapped sewers within the asset database used and held by DCWW, which may impact the likelihood of blockage within the geographical areas but their presence will not be represented in the input or output features derived for each aggregate. These factors may therefore influence the potential benefit of the geographical aggregation method.

3.2. Ensemble Modelling

Figure 4 gives the results of the ensemble models produced on the testing and training datasets defined. Overall, the results are similar to the best single model results for this subset of sewers, of 0.69 [13]. A few of the models show slightly improved performance, with for example models 1 and 3 showing testing AUCs of 0.71, although the models also show overfitting to the training dataset. Investigating the ROC curves for the models also does not appear to show any benefits from the ensembles. As mentioned above, the desire for the application of the models to WaSCs' networks is that a set of high risk sewers could be accurately predicted and highlighted for monitoring. Comparing one of the best performing models (Model 1)(Figure 5) with the previous best model (Figure 6) shows ROC curves of a similar shape, with the other models showing similar or flatter shapes, and therefore providing no benefits from the modelling using ensembles in this regard.

The different settings used to produce the ensembles show different impacts on the performance of the models. For the number of inputs, models 1 to 3 and 5 to 9 show a comparison between models produced using the same settings, but with the number of inputs to each model changed. The chart shows little variation in the performance of the models as the number of inputs is changed. With the aim to improve the performance of the single models in the ensemble, but limit correlation between them, the number of inputs used leads to a balance between these two measures [10], with this conflict potentially resulting in this lack of change in performance. For the ensemble method, the raw propensity weighted voting method generally seems to give better performance on the testing dataset, comparing models 1 to 3 and 5 to 7 respectively, but also a higher level of overfitting. The improved performance would be influenced by the increased weight given to the better performing models, where for the voting method the weight given to each model within the ensemble is the same. Given the lack of pruning of the models within the ensemble, the best performing models, which are more strongly weighted, may be expected to show overfitting, which may then be seen in the final ensemble output. For the query type, the random input and random combination methods both show similar performance. The random combination method was designed to increase the number of features in situations where there are a low number of input features, to aid an increase in strength, while preventing correlation between the models [10]. This is not shown in the performance of the models,



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. RRP	Avg. RRP	Avg. RRP	Avg. RRP	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 4: Bar chart showing the model settings and results for the models built using ensemble techniques. The chart shows the training and testing AUC's for each of the models, along with a table of results and number used to reference each model.

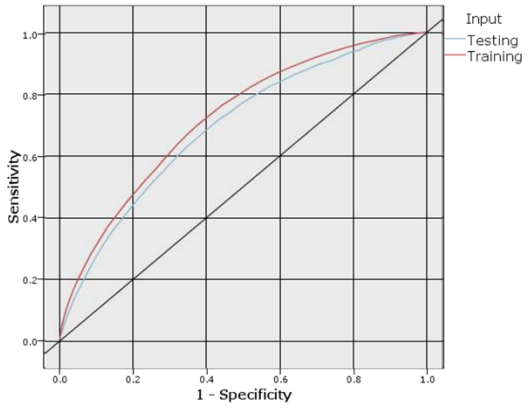


Figure 5: ROC curve showing the performance of model number 1, one of the best performing ensemble models.

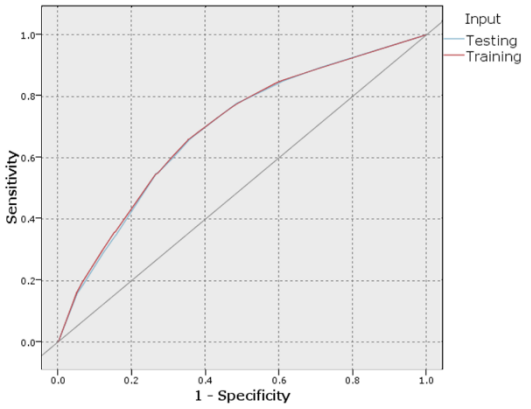


Figure 6: ROC curve showing the performance of the best performing single decision tree [13]

which may be due to the relatively large number of input features already available to the random input models limiting the applicability of the random combination method.

To be of sufficient explanatory capability, a consistent dataset of blockages across a number of years is required for both an increasing number of years of input data and use as an input feature. While no further benefit was derived from the larger number of years of data, this may be due to the variables derived and there may be potential for deriving further features based on the historical data, which aid in identifying sewers which, for example, consistently show a tendency to block. In addition, the achievement of similar results with only four years of input

data (two for the input features and two for the output features) suggests that, following potential future regulatory or operational changes, benefit can relatively quickly be derived from the historical data on incidents.

4. Conclusion

This paper presents the development of decision tree models using the three techniques of: geographical aggregation, ensemble modelling and derivation of a historical blockage input feature. The results show limited benefit from the geographical aggregation and ensemble modelling, while the additional input feature provided small benefits in model performance, and further benefits in the more accurate prediction of the high likelihood sewers, which is of particular benefit for this application. The paper provides further information on the application of these techniques, their potential benefits and the situations in which these benefits may be derived.

5. 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).

References

[1]. OFWAT. 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.

[2]. Hall, M., et al. Deterioration Rate of Sewers. UKWIR. s.l. : UK Water Industry Research Limited2005/6.

[3]. Savic, Dragan A. The use of data-driven methodologies for prediction of water and wastewater asset failures. Risk Management of Water Supply and Sanitation Systems. s.l. : Springer Netherlands, pp. 181-190.

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

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

[6]. Statistical analysis and definition of blockages-prediction formulae for the wastewater network of Oslo by evolutionary computing. Rita Ugarelli, Stig Morten Kristensen, Jon Røstum, Sveinung Sægrov, Vittorio Di Federico. 8s.l. : IWA Publishing, 2009, *Water Science and Technology,* Vol. 59pp. 1457-1470.

[7]. Modelling sewer failure by evolutionary computing. Dragan Savic, Orazio Giustolisi, Luigi Berardi, Will Shepherd, Slobodan Djordjevic, Adrian Saul. 2s.l. : Thomas Telford, 2006, *Proceedings of the ICE-Water Management,* Vol. 159, pp. 111-118.

[8]. Static and dynamic effects in prioritizing individual water mains for renewal. Kleiner, Yehuda and Rajani, Balvant. 2007, *Water Management Challenges in Global Change,* pp. 61-68.

[9]. Ensemble Methods in Machine Learning. Dietterich, Thomas G. s.l. : Springer, 2000, *Multiple Classifier Systems,* pp. 1-15.

[10]. Random Forests. Breiman, Leo. *Machine Learning*, Vol. 45, pp. 5-32.

[11]. Random forests for classification in ecology. Cutler, D.R., Edwards Jr, T.C., Beard, K.H., Cutler, A., Hess, K.T., Gibson, J. and Lawler, J.J., *Ecology*, Vol. 88, pp. 2783-2792.

[12]. Gene selection and classification of microarray data using random forest. Díaz-Uriarte, Ramón, and Sara Alvarez De Andres. *BMC bioinformatics* Vol. 7p. 1.

[13]. Predictive risk modelling of real-world wastewater network incidents. Bailey, James, et al. s.l. : Elsevier, 2015, *Procedia Engineering,* Vol. 119, pp. 1288–1298.

[14]. Estimation of burst rates in water distribution mains. Boxall, J. B., O'Hagan, A., Pooladsaz, S., Saul, A. J., & Unwin, D. M. London : Institution of Civil Engineers by Thomas Telford Ltd, 2007. *Proceedings of the Institution of Civil Engineers-Water Management.* Vol. 160, pp. 73-82.

[15]. Approaches to sewer maintenance: a review. Fenner, RA. 4s.l. : Elsevier, 2000, *Urban water*, Vol. 2, pp. 343-356.

Appendix A.

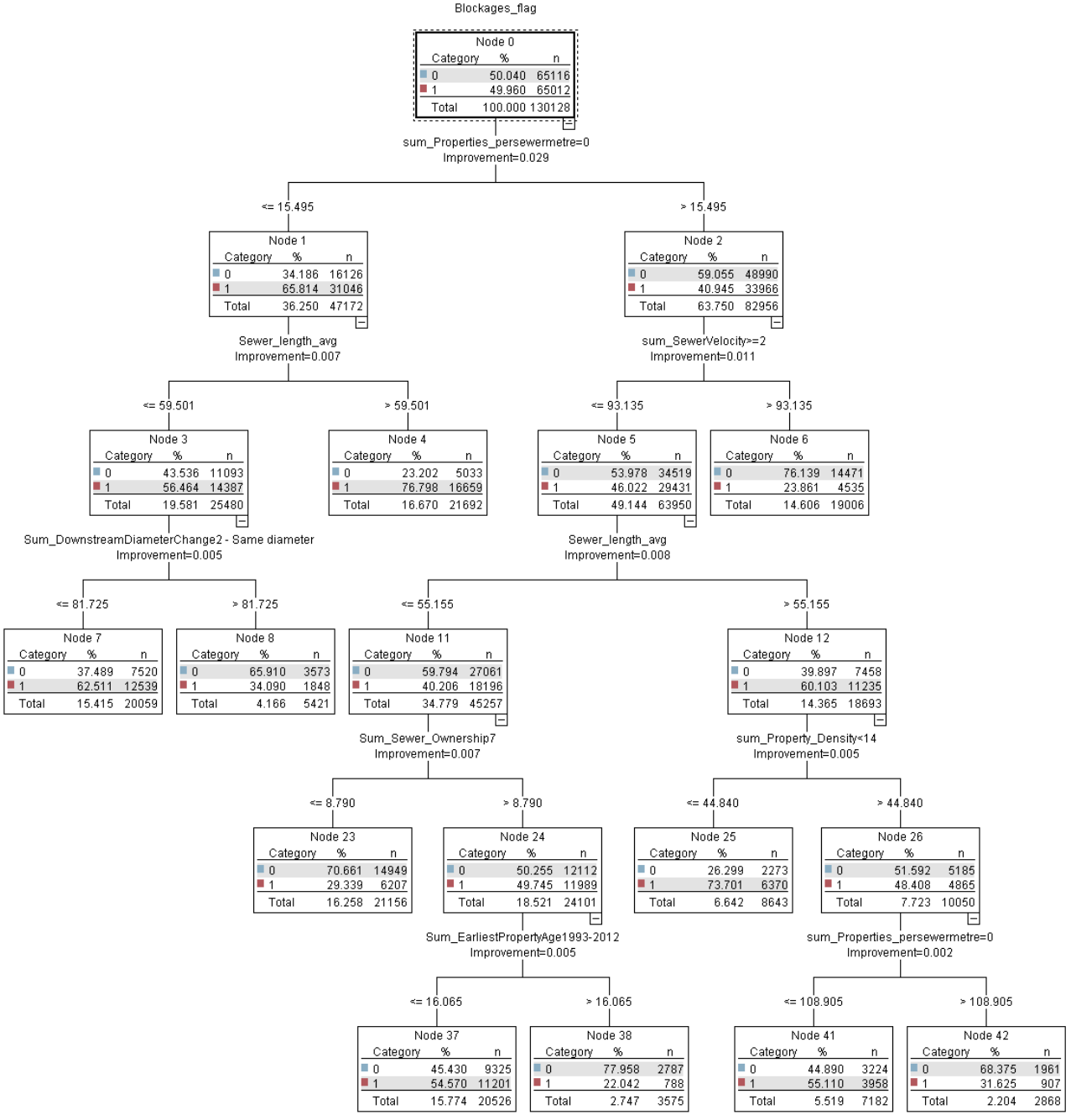


Figure 7: the decision tree output for the model of relative blockage proportion threshold 8. The variables shown in the decision tree represent length weighted averages (suffixed with _avg) or total length (prefixed with sum_) for each of the aggregated areas.