

PAPER • OPEN ACCESS

Understanding district metered area level leakage using explainable machine learning

To cite this article: Matthew Hayslep *et al* 2023 *IOP Conf. Ser.: Earth Environ. Sci.* **1136** 012040

View the [article online](#) for updates and enhancements.

You may also like

- [Exploring the impacts of tourism and weather on water consumption at different spatiotemporal scales: evidence from a coastal area on the Adriatic Sea \(northern Italy\)](#)

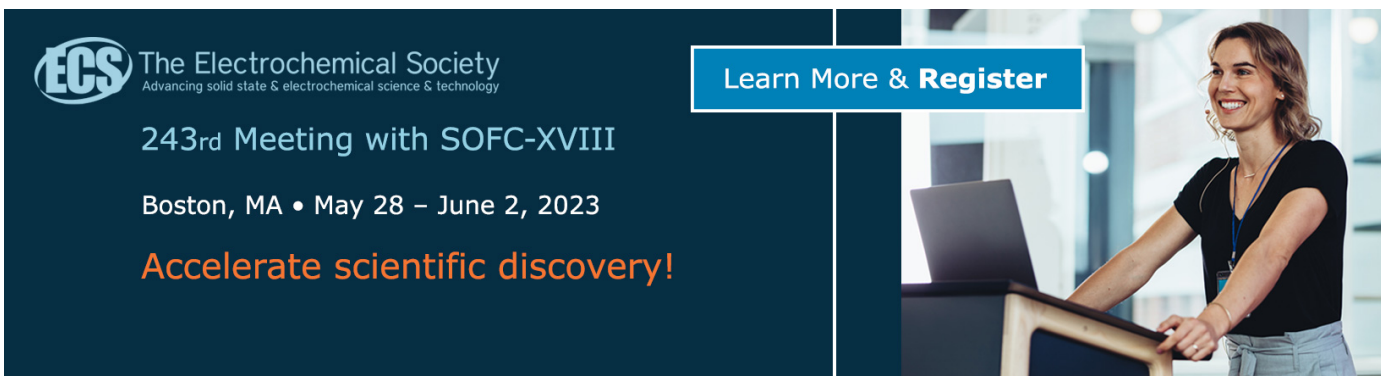
Filippo Mazzoni, Valentina Marsili, Stefano Alvisi et al.

- [CHANDRA OBSERVATIONS OF THE GALAXY GROUP AWM 5: COOL CORE REHEATING AND THERMAL CONDUCTION SUPPRESSION](#)

A. Baldi, W. Forman, C. Jones et al.

- [A COMBINED LOW-RADIO FREQUENCY/X-RAY STUDY OF GALAXY GROUPS. I. GIANT METREWAVE RADIO TELESCOPE OBSERVATIONS AT 235 MHz AND 610 MHz](#)

Simona Giacintucci, Ewan O'Sullivan, Jan Vrtilek et al.



ECS The Electrochemical Society
Advancing solid state & electrochemical science & technology

243rd Meeting with SOFC-XVIII

Boston, MA • May 28 – June 2, 2023

Accelerate scientific discovery!

Learn More & Register

Understanding district metered area level leakage using explainable machine learning

Matthew Hayslep¹, Edward Keedwell¹, Raziye Farmani¹, Joshua Pocock²

¹ College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter, EX4 4PY, United Kingdom

² DWS Water Resources & Water Efficiency, South West Water, Exeter, EX2 7HR, United Kingdom

mh989@exeter.ac.uk

Abstract. Understanding the various interrelated effects that result in leakage is vital to the effort to reduce it. This paper aims to understand, at the district metered area (DMA) level, the relationship between leakage and static characteristics of a DMA, i.e. without considering pressure or flow. The characteristics used include the number of pipes and connections, total DMA volume and network density, as well as pipe diameter, length, age, and material statistics. Leakage, especially background and unreported leakage, can be difficult to accurately quantify. Here, the Average Weekly Minimum Night Flow (AWM) over the last 5 years is used as a proxy for leakage. While this may include some legitimate demand, it is generally assumed that minimum night flow, strongly correlates with leakage. A data-driven case study on over 800 real DMAs from UK networks is conducted. Two regression models, a decision tree model and an elastic net linear regression model, are created to predict the AWM of unseen DMAs. Reasonable accuracy was achieved, considering pressure is not an included feature, and the models are investigated for the most prominent features related to leakage.

1. Introduction

It has been consistently noted throughout the literature that leakage is a pervasive and important challenge for water companies which has an economic, environmental and sustainability impact [1]. To combat this and reduce leakage, water distribution networks in the UK are commonly divided into District Metered Areas (DMAs) [2, 3]. These DMAs have flow sensors at the inlets and outlets, providing a measure of how much water is used in that DMA. There are multiple ways of measuring and analyzing leakage [4]. This paper uses Average Weekly Minimum Night Flow (AWM) as a proxy for leakage which focuses this study on background and unreported leakage (see Section 2). In addition, Minimum Night Flows (MNFs) are part of the reporting required of water companies in the UK by the regulator OFWAT [4, 5], meaning that the data necessary to undertake this study is widely available.

Other research has estimated non-revenue water ratios and leakage rates [6-9]. Many of these works use principal component analysis and artificial neural networks [6-9]. These methods, while more sophisticated, are less explainable than the methods used in this paper. Explainability is important from the perspective of water companies to support decision making about leakage mitigation methodologies to implement and areas on which to focus resources. Explainable models allow us to understand the relationship between input features and outputs (e.g. predicted leakage),



furthering our understanding of the problem itself and what factors impact leakage the most. For water companies, this means additional factors can be considered, or better understood, when deciding on actions, company policies, and management.

2. Methods

This paper investigates two models, a decision tree and an elastic net. Using explainable models is important because it allows the exploration of the *reasons* why a certain prediction is made. This means that the models are useful and informative outside of their direct application.

In this paper, leakage is defined as Minimum Night Flow (MNF), measured in m^3h^{-1} , and is the flow during the hour of least flow between midnight and 4am. MNF is recorded, ideally, every day for each DMA. The Weekly MNF is the lowest MNF recorded during a Monday-Sunday period. The use of Weekly MNF smoothes out what can be an extremely noisy signal and will also mask bursts or other major leakage events that do not last for at least a full week, which focuses this work on background and unreported leakage. Weekly MNF data covers a 5-year period from the beginning of 2017 to the beginning of 2022 for 866 real DMAs in the UK. The availability of Weekly MNFs for any single DMA varies from a few examples to almost total coverage of the 5-year period. To further reduce this problem from a time series forecast problem into a regression problem we use the Average Weekly MNF (AWM) over the last 5 years. Each model takes as input the static characteristics of a DMA and outputs the predicted AWM for that DMA. A prediction of, for example, $10 \text{ m}^3\text{h}^{-1}$ would indicate that the average, over the last 5 years of operation, Weekly MNF for that DMA would be $10 \text{ m}^3\text{h}^{-1}$.

The aim of this paper is to predict the AWM of unseen DMAs based on their static characteristics using explainable, data-driven methods. Emphasis is put on predicting AWM, which is derived from MNF, *without* a hydraulic model, pressure, or flow information. By predicting AWM in this way, we can further our understanding of the causes of leakage. Accurate predictions of the expected AWM may also inform new leak detection methodologies, which often rely on some understanding of ‘normal’ operation.

Predictions are based on the number of customer connections in a DMA and various pipe features. The models learn from a subset of the real DMAs and are tested on the remaining subset of completely unseen DMAs. There is considerable variety amongst each of the 866 real DMAs in terms of total length (less than 1km to over 10km), pipe materials (100% metal to nearly 100% plastic), number of customer connections (100s to 1000s), etc. In total, more than 80 features were created. Most of these features were calculated using the pipe data available for each DMA. Each pipe provides the following information: installation date (used to calculate age), length, diameter, and material. Various materials were grouped together into ‘metal’ (various types of iron and steel), ‘plastic’ (PE, PVC, etc.) and ‘other’ (mainly asbestos cement). From this, various features were created such as ‘total length’, which is the sum of all pipe lengths in a DMA, ‘proportion of metal pipes’, which is the ratio of the number of metal pipes to all pipes, ‘proportion by volume of pipes older than 20’, which is the ratio of the volume of all pipes older than 20 years to the volume of all pipes, and so on. The volume of a pipe is calculated, based on the assumption that a pipe is a cylinder, using its length and diameter. Additional naming conventions are that large/small refers to diameter while long/short refers to length. Therefore, the smallest pipe diameter refers to the *diameter* of the pipe with the smallest diameter, whereas shortest pipe age refers to the *age* of the pipe with the shortest length. In addition, there are several ‘weighted age’ metrics which are calculated as follows:

$$f \text{ weighted age} = \frac{\sum_{i=1} f(p_i) \cdot a(p_i)}{\sum_{i=1} f(p_i)} \quad (1)$$

where p_i is the i^{th} pipe in a DMA, f is the parameter of interest, for example diameter, length, or volume, and a is the age. Additional features of note include the number of customer connections and

network density (customer connections per meter of pipe). A full table of the features investigated can be found in the appendix (table A1).

Each DMA is randomly placed into the training set or test set. The training set, which consists of 60% of the DMAs, is used to train the models. The test set, consisting of the remaining 40%, is used to report the performance of the models, these DMAs are not seen during training. The same test-train split was used consistently across all models, i.e., the same DMAs were always part of the training set and vice-versa for the test set. This is important because our purpose is to find the best model, not the most representative training set.

Many machine learning models make assumptions about the normalization of data, making it important to preprocess the DMA data. Therefore, features which are not proportions (since these are already normalized) are normalized. This is done by removing the median and dividing by the interquartile range, i.e., the range between the 1st and 3rd quartile. The median is used to scale instead of the mean because the data is not normally distributed. Therefore, using the median is more robust to outliers. To ensure that all features lie in the range 0 to 1, the features which are not proportions are then rescaled to fall within that range. The precise values for these transformations are found using the training set, rather than the entire dataset. This is important to avoid ‘leaking’ information about the test set back into the training set. This method of normalization is similar to the ‘Z-score’ in [6, 9] but replaces the mean and standard deviation with the median and interquartile range.

Cross-validated grid search is used to finetune the hyperparameters of the decision tree and elastic net. Cross-validation takes a random subset of the training data away, without replacement, to serve as an intermediate test set, known as the validation set. This process is repeated, 8 times in this paper, until every training sample has been in the validation set exactly once. Each validation set is used to test different model hyperparameters, with the average over the repeated random subsets being used to determine the best hyperparameter values.

Each model is assessed primarily using the R^2 metric. Explained variance and mean absolute error (MAE) have also been reported for completeness. R^2 values range from $-\infty$ to 1, with 1 being the best possible score. A model which simply returns the average value of all samples would get an R^2 of 0.

2.1. Decision Tree

Decision trees [10] are simple models where predictions are made based on a series of if-then-else decisions, similar to a flow chart. Each decision, or node, is based on a threshold on a single feature. At the end of a series of decisions, a leaf node determines what value is outputted. The depth of a tree, i.e., the maximum number of decisions before reaching a leaf node, determines its complexity. At each node, the decision (which feature and what threshold) which minimizes the mean squared error (MSE) in each child node is chosen. This is repeated until the tree is complete. In this paper, the leaf nodes on the decision trees output a constant value. The hyperparameters of the decision tree that were searched using cross-validated grid search were: maximum tree depth, minimum samples required in a leaf node, and minimum samples required to split a node.

2.2. Elastic Net

Elastic Nets [11] are linear regression models that use both l_1 and l_2 regularization. Regularization adds additional components to the loss function, which is the function that is minimized during training to determine the coefficients of the model. While l_1 regularization adds a penalty based on the absolute value of the coefficients, l_2 regularization adds a penalty based on the squared value of the coefficients. Elastic net tries to combine the benefits of lasso linear regression, which prefer to have fewer features with non-zero coefficients by using l_1 regularization, and ridge linear regression, which penalizes the magnitude of coefficients by using l_2 regularization. Elastic net is particularly useful when there are groups of correlated features [11], which is certainly the case in this instance. When given correlated features, other linear regression models would pick one of the correlated features to be the only important one, whereas an elastic net spreads the importance amongst the correlated features. The hyperparameters of the elastic net that were searched using cross-validated grid search

were: alpha, which controls ratio between the l_1 and l_2 regularization terms and lambda, which scales both the l_1 and l_2 regularization terms.

3. Results and Discussion

This section discusses the results of the two methods that were used. Ten different runs, including cross-validation for hyperparameter tuning, with different random seeds were conducted, as these models are not deterministic. The following sections detail the results of the best model for each method from these runs. Table 1 shows the overall results for each of the best models. From this table, we can see that the elastic net model significantly outperforms the decision tree. This is to be expected because of the simplicity of the decision tree and due to its fundamental limitations when used as a regressor, i.e., discrete single value outputs. Section 3.1 details the decision tree results while Section 3.2 details the elastic net.

Other models, such as a neural network, support vector machine and random forest, which are not detailed here, were tested. Overall, the elastic net model gave the best results, but most models achieved an explained variance and test R^2 of at least 0.6. It is possible that the performance of the elastic net is near the best achievable with the current set of features. Better performance might be achievable with more or different explanatory factors.

Table 1. Overall results for predicting AWM.

Method	Explained Variance	Test R^2	Train R^2	Overall R^2	Test MAE	Train MAE	Overall MAE
Decision Tree	0.572	0.571	0.632	0.606	3.031	2.730	2.851
Elastic Net	0.680	0.680	0.698	0.691	2.741	2.465	2.575

3.1. Decision Tree

Decision trees are simple, but explainable, models with the limitation of only being able to output discrete values, which is not ideal in a regression context. Therefore, it is common to use many different decision trees in an ensemble, i.e., a random forest. However, in this case the performance of the lone decision tree is good enough to justify examining it. Figure 1 shows the full decision tree with the thresholds transformed back into their original space. The sample sizes and MSEs are based on the training set. The decision tree model splits the data using ‘total length’ three times, ‘number of customer connections’ and ‘shortest pipe age’ twice, and ‘diameter weighted age’ and ‘total volume of metal pipe’ once.

The root node splits roughly two thirds of the training set away based on total length. The right-hand subtree then splits again on a much larger total length value with DMAs with more than 33km of pipe being assigned the highest AWM. If a DMA does not have more than 33km of pipe, then the number of customer connections is considered. DMAs with fewer customer connections are assigned a lower AWM than those with more. Overall, the right-hand subtree can be summarized as ‘larger DMAs have a higher AWM’ using different metrics.

Meanwhile, the left-hand subtree splits on shortest pipe age. The shortest pipes in DMAs in this particular dataset are often less than 1m long and might be better understood as representing junctions. Therefore, one way of interpreting the significance of this metric is that the age of the shortest junction in DMAs is an important factor. If the DMA has an older shortest pipe, the assigned AWM is dependent on size metrics, where larger DMAs are assigned a higher AWM. If the DMA has a younger shortest pipe, then the diameter weighted age and shortest pipe age are considered. Overall, the left-hand subtree can be summarized as ‘smaller, younger infrastructure leaks less than older, larger infrastructure’, i.e. the importance of infrastructure condition [4] and size.

Splitting on the number of customer connections and total length could be interpreted as the importance of network density, in terms of connections per meter of pipe. However, network density was a metric included as a feature which the decision tree did not consider important. This might

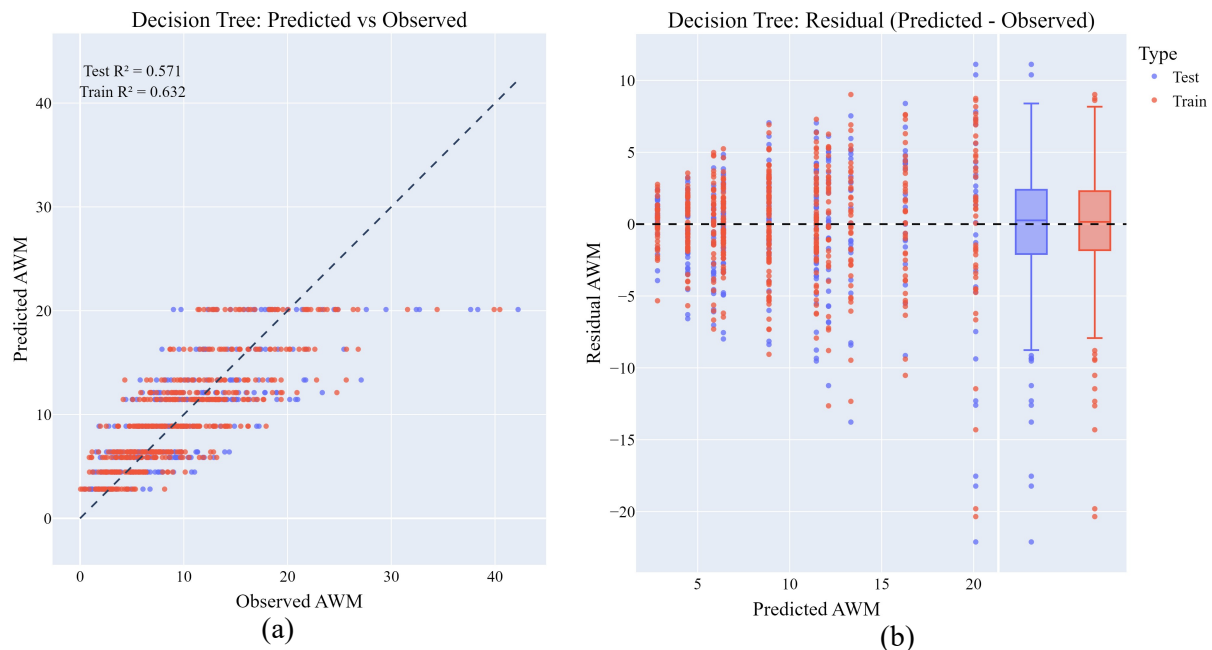


Figure 2. Decision tree comparison between (a) predicted and observed AWM, and (b) predicted and residual AWM (prediction - observed). The dotted line in (a) and (b) shows the zero-error line, i.e., perfect predictions. Figure 2b also shows the distribution of samples along the residual axis as a box plot.

Figure 2 shows the predictions of the decision tree across both the test and training set. The discrete outcomes of the decision tree can be seen clearly in both figure 2a and 2b. While there is some overlap between a DMA's predicted AWM and observed AWM, the box plot in figure 2b shows that most predictions lie within $\sim 2.5 \text{ m}^3\text{h}^{-1}$ of their observed value. The absolute median percentage error is 25.7% and 30.0% for the training and test set respectively. In addition, the decision tree is not significantly less accurate on the test set compared to the training set, which is confirmed by the R^2 of each set, and indicates that the model did not overfit. However, the model does have difficulty predicting the AWM of extreme outlier DMAs, although this is not completely unexpected behavior.

Overall, the performance of the decision tree model is good, especially considering its simplicity and limitations. In addition, the decision tree as shown in figure 1 allows straightforward assessment of more DMAs without using of the model directly.

3.2. Elastic Net

The Elastic Net model results are detailed in this section. Unlike the decision tree, the elastic net model has a continuous output because, fundamentally, it is a linear model. Therefore, we find that this model outperforms the decision tree model. This comes at the cost of additional complexity, though it is still a linear model. Figure 3 shows the predictions of the elastic net model across both the test and training set. Figure 3b again shows that most predictions lie within $\sim 2.5 \text{ m}^3\text{h}^{-1}$ of their observed value. The absolute median percentage error is 23.8% and 25.0% for the training and test set respectively. Furthermore, the accuracy of the model is similar on the test set and the training set, which is confirmed by the R^2 of each set, indicating the model has not overfitted. The elastic net model, as with the decision tree, has more difficulty predicting the AWM of extreme outlier DMAs. Again, this is not surprising as they are outliers.

As a linear regression model, the elastic net model derives a set of coefficients which allow us to understand why the model makes certain predictions. Table 1 lists the coefficients of the elastic net linear model. These coefficients are for the transformed data. They allow us to understand which

features the elastic net has determined to be the most important for predicting the AWM of a DMA. In general, coefficients of less than 1 are insignificant, but have been reported for completeness.

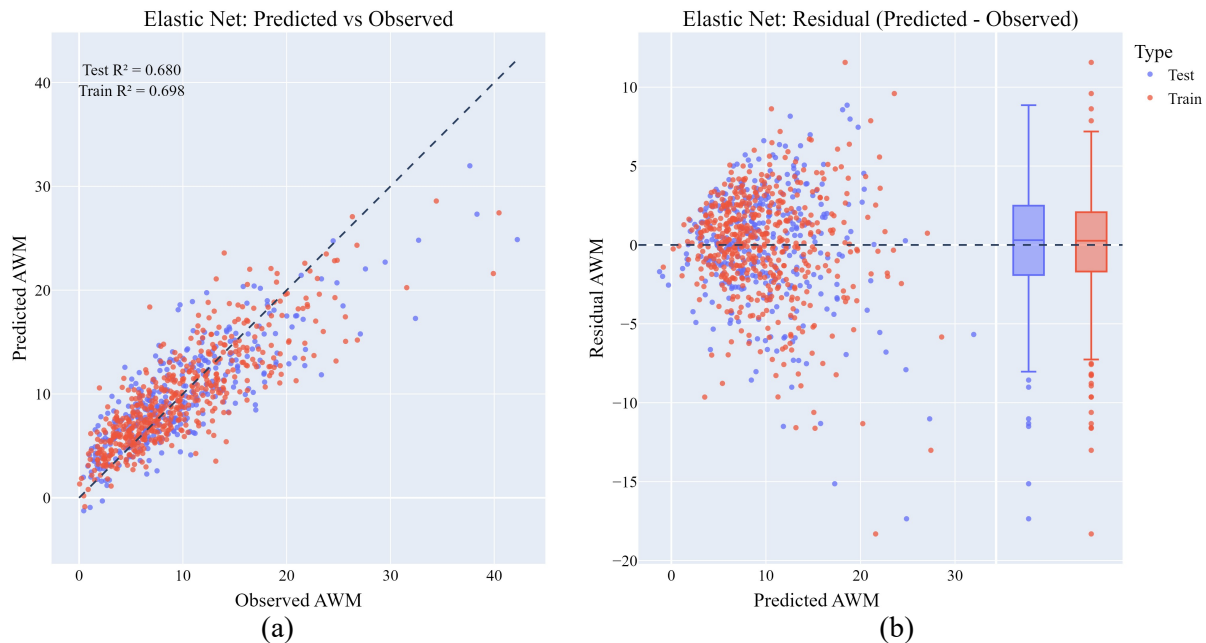


Figure 3. Elastic net comparison between (a) predicted and observed AWM, and (b) predicted and residual AWM (prediction - observed). The dotted line in (a) and (b) shows the zero-error line, i.e., perfect predictions. Figure 3b also shows the distribution of samples along the residual axis as a box plot.

From table 2 we can see that the number of customer connections is by far the most important factor and correlates with an increase in AWM. This is consistent with research on the leakage at service connections and water mains, where service connections, due to the high density of joints and fittings, are responsible for more leakage than their total length might indicate [4]. This is also consistent with the practice of reporting leakage in relation to the number of customer connections in an area [4]. This may also reflect the legitimate night usage of customers. The next most important coefficients are ‘total length of pipes larger than 100mm’ and ‘total length’, both correlating with an increase in AWM. Both of these features are closely correlated with each other because the pipes smaller than 100mm are usually very short. Again, this is consistent with the literature, in that the more pipe there is, the more leakage there is.

Table 2 also shows the factors which correlate with a decrease in AWM. Of these, the most important is ‘smallest pipe diameter’. This suggests that, if the diameter of the smallest pipe in a DMA is larger, then the AWM is smaller. Bernoulli’s principle states that a decrease in pipe diameter will decrease pressure and increase velocity. This increased velocity may result in more wear and tear, leading to more leakage, in smaller pipes compared to larger pipes. This increased velocity may be at its height during the MNF, as this is usually the period with the highest pressure, unless pressure management policies are in place. Furthermore, there is an interplay between the positive coefficients: ‘largest pipe diameter’ and ‘total length of pipes larger than 100mm’, which indicate that larger pipes cause more leakage, and the negative coefficient ‘smallest pipe diameter’ which indicates that larger pipes cause less leakage. One interpretation is that this describes the relationship *between* pipes of different sizes in a DMA. Specifically, those significant changes in pipe diameter within a DMA, such as stepping down from a 200mm pipe to a 20mm pipe, and the resulting pressure changes, are a significant underlying cause of these coefficients.

Table 2. Coefficients for the elastic net model in descending magnitude. Left shows positive coefficients, right negative. Prop. is the abbreviation of proportion.

Feature	Coeff.	Feature	Coeff.
Number of Customer Connections	9.53	Smallest pipe diameter	-3.14
Total length of pipes larger than 100mm (m)	3.54	(mm)	
Total length (m)	3.31	Prop. of Metal pipes	-1.30
Oldest pipe length (m)	2.79	Longest pipe length (m)	-1.22
Prop. by volume of pipes older than 20	2.67	Total length of Plastic	-1.20
Total volume of Metal pipe (m ³)	2.51	pipes older than 20 (m)	
Smallest pipe length (m)	2.17	Prop. by length of Plastic	-0.87
Prop. by volume of Metal pipe	2.09	pipes older than 20	
Shortest pipe age (y)	2.08	Total length of pipes	-0.84
Prop. by length of pipes longer than 400m	2.01	older than 20 (m)	
Largest pipe diameter (mm)	1.85	Prop. by volume of Other	-0.79
Total length of Metal pipe (m)	1.84	pipe	
Number of pipes	1.67	Total volume of pipes	-0.71
Prop. by length of Other pipe	1.56	longer than 400m (m ³)	
Average Metal pipe diameter (mm)	1.39	Total volume of Plastic	-0.59
Largest pipe age (y)	1.05	pipe (m ³)	
Total length of Plastic pipe (m)	1.03	Longest pipe diameter	-0.55
Prop. by length of Metal pipe	0.88	(mm)	
Total volume of pipes larger than 100mm (m ³)	0.88	Prop. by volume of pipes	-0.55
Total length of Metal pipes older than 20 (m)	0.77	longer than 400m	
Prop. of Plastic pipes	0.76	Prop. by volume of pipes	-0.39
Diameter Weighted Age (y)	0.66	larger than 100mm	
Total length of Other pipe (m)	0.65	Number of Metal pipes	-0.34
Prop. by length of pipes larger than 100mm	0.58		
Prop. of Other pipes	0.56	Number of Other pipes	-0.33
Network Density (mc ⁻¹)	0.37		
Average Metal pipe length (m)	0.33	Prop. of pipes longer than	-0.26
Prop. by length of Metal pipes older than 20	0.33	400m	
Length Weighted Age (y)	0.30	Prop. by length of Plastic	-0.22
Number of Plastic pipes	0.27	pipe	
Oldest pipe diameter (mm)	0.12	Number of pipes longer	-0.08
Average pipe age (y)	0.07	than 400m	
Prop. of pipes larger than 100mm	0.07	Total volume of Other	-0.05
Average Other pipe diameter (mm)	0.03	pipe (m ³)	
Total length of pipes longer than 400m (m)	0.03	Number of pipes older	-0.03
Shortest pipe length (m)	0.02	than 20	

Table 2 also shows the impact of age on AWM with ‘oldest pipe length’, ‘proportion by volume of pipes older than 20’ and ‘shortest pipe age’ all correlating with an increase in AWM. In addition, the impact of different materials is also present, though more difficult to decipher. For example, the proportion of metal pipes (by number) in a DMA correlates with a decrease in AWM while the proportion *by volume* of metal pipe correlates with an increase in AWM. This might mean that for smaller/shorter pipes metal is better than other materials, but for larger/longer pipes other materials are better than metal. This may be because of the mechanical joints of metal pipes which allow more minor leaks than, for example, a welded plastic pipe [4]. In addition, there is ‘total length of plastic pipes older than 20’ and ‘proportion by length of plastic pipes older than 20’, both of which correlate with a decrease in AWM. This could either mean that as plastic pipes age they do not degrade as badly

as other pipe materials, or it could mean that the type of plastic used more than 20 years ago is better than the type of plastic used since. All the factors that affected the decision tree are present as coefficients for the elastic net model, though not necessarily as prominently as their importance in the decision tree might suggest.

Many of these features are related to others. For example, ‘smallest pipe diameter’ is related to ‘total length of pipes larger than 100mm’, which is also related to the total length of all pipes. If the smallest pipe diameter is 100mm then ‘total length of pipes larger than 100mm’ is equal to the total length of all pipes. Because of this, the interactions between the different coefficients are more complex in some cases than a cursory glance might suggest, and it becomes important to interpret which features are related to each other. Therefore, it is much harder to answer a question such as ‘how would the AWM change if we replaced the oldest and largest metal pipe with a plastic one’ without calculating the new statistic for the theoretical DMA and running it through the model. This is one of the main downsides of this linear model compared to the decision tree: while the elastic net is more accurate, it is also more difficult to interpret.

4. Conclusion

This paper has presented two models for predicting Average Weekly MNF of DMAs in a real-world data-driven case study. By doing so we hoped to further our understanding of the causes of long-term leakage. The results have illustrated the main factors that contribute to these predictions for each model and attempts have been made to interpret why these factors are important. The models and results are based on a data-driven approach which relies on little to no domain knowledge. The results presented support and demonstrate many of the ideas on what influences leakage presented in the literature. Specifically, the importance of DMA size, the number of customer connections and service pipe sizes, the different material properties for different pipe sizes and the interplay between different materials and their age all have an effect on Average Weekly MNF.

The simplicity and explainability of the methods used should allow these results to be repeated and improved upon. Further work may include the addition of more features. In particular, pressure may be the most obviously missing metric. Adding pressure or other features may help to improve the accuracy of these models even further. In addition, increasing the size of the dataset with more DMAs may also improve the accuracy of the predictions. Finally, further work may also incorporate these predictions as part of a leak detection methodology.

Acknowledgments

The first author would like to acknowledge the financial support from South West Water in collaboration with the University of Exeter Centre for Resilience, Environment, Water and Waste (CREWW). Special thanks are given to James Mercer for his assistance in accessing the data used in this paper. This work was undertaken as part of the first author’s pursuit of a PhD.

Appendix

Table A1. Table of all features, in no specific order, created for each DMA.

Features		
Number of pipes	Diameter Weighted Age (y)	Length Weighted Plastic Age (y)
Total length (m)	Average pipe diameter (mm)	Prop. of pipes larger than 100mm
Oldest pipe age (y)	Largest pipe diameter (mm)	Average Other pipe diameter (mm)
Average pipe age (y)	Longest pipe diameter (mm)	Number of pipes longer than 400m
Longest pipe age (y)	Average Other pipe age (y)	Average Metal pipe diameter (mm)
Largest pipe age (y)	Average Metal pipe age (y)	Total length of Plastic pipe (m)
Prop. of Metal pipes	Smallest pipe diameter (mm)	Number of pipes larger than 100mm
Prop. of Other pipes	Shortest pipe diameter (mm)	Total volume of Plastic pipe (m ³)
Number of Metal pipes	Youngest pipe diameter (mm)	Average Plastic pipe diameter (mm)
Youngest pipe age (y)	Prop. of pipes older than 20	Prop. by length of pipes older than 20

Shortest pipe age (y)	Average Plastic pipe age (y)	Prop. by volume of pipes older than 20
Smallest pipe age (y)	Prop. by length of Other pipe	Total length of pipes older than 20 (m)
Number of Other pipes	Prop. by volume of Other pipe	Total volume of pipes older than 20 (m ³)
Network Density (mp ⁻¹)	Average Other pipe length (m)	Prop. by length of pipes longer than 400m
Oldest pipe length (m)	Length Weighted Other Age (y)	Prop. by volume of pipes longer than 400m
Prop. of Plastic pipes	Number of pipes older than 20	Prop. by volume of pipes larger than 100mm
Number of Demand Points	Prop. by length of Metal pipe	Total length of pipes longer than 400m (m)
Volume Weighted Age (y)	Prop. by volume of Metal pipe	Prop. by length of pipes larger than 100mm
Average pipe length (m)	Length Weighted Metal Age (y)	Total volume of pipes longer than 400m (m ³)
Longest pipe length (m)	Average Metal pipe length (m)	Total length of pipes larger than 100mm (m)
Largest pipe length (m)	Total length of Other pipe (m)	Total volume of pipes larger than 100mm (m ³)
Length Weighted Age (y)	Total length of Metal pipe (m)	Prop. by length of Other pipes older than 20
Number of Plastic pipes	Total volume of Other pipe (m ³)	Prop. by length of Metal pipes older than 20
Total volume of DMA (m ³)	Prop. of pipes longer than 400m	Total length of Other pipes older than 20 (m)
Shortest pipe length (m)	Total volume of Metal pipe (m ³)	Total length of Metal pipes older than 20 (m)
Smallest pipe length (m)	Prop. by length of Plastic pipe	Prop. by length of Plastic pipes older than 20
Youngest pipe length (m)	Prop. by volume of Plastic pipe	Total length of Plastic pipes older than 20 (m)
Oldest pipe diameter (mm)	Average Plastic pipe length (m)	

References

- [1] Puust R, Kapelan Z, Savic D A and Koppel T 2010 A review of methods for leakage management in pipe networks *Urban Water Journal* **7**(1) 25–45 <https://doi.org/10.1080/15730621003610878>
- [2] Romano M 2021 Review of techniques for optimal placement of pressure and flow sensors for leak/burst detection and localisation in water distribution systems *ICT for Smart Water Systems: Measurements and Data Science* **102** ed A Scozzari et al (Cham: Springer International Publishing) chapter 2 pp 27-63 <https://doi.org/10.1007/978-3-319-405>
- [3] Savić D and Ferrari G 2014 Design and performance of District Metering Areas in water distribution systems *Procedia Engineering* **89** 1136–43 <https://doi.org/10.1016/j.proeng.2014.11.236>
- [4] Farley M and Trow S 2005 *Losses in Water Distribution Networks: A Practitioners' Guide to Assessment, Monitoring and Control* vol 4 (London, UK: IWA Publishing) <https://doi.org/10.2166/9781780402642>
- [5] Ofwat 2018 *Targeted review of common performance commitments final report* (Ofwat, UK) pp 281 <https://www.ofwat.gov.uk/publication/targeted-review-common-performance-commitments-final-report/>
- [6] Jang D and Choi G 2017 Estimation of non-revenue water ratio for sustainable management using artificial neural network and Z-score in Incheon, Republic of Korea *Sustainability* **9**(11) 1933 <https://doi.org/10.3390/su9111933>
- [7] Jang D and Choi G 2018 Estimation of non-revenue water ratio using MRA and ANN in water distribution networks *Water* **10**(1) 2 <https://doi.org/10.3390/w10010002>
- [8] Jang D, Park H and Choi G 2018 Estimation of leakage ratio using principal component analysis and artificial neural network in water distribution systems *Sustainability* **10**(3) 750 <https://doi.org/10.3390/su10030750>
- [9] Kizilöz B 2021 Prediction model for the leakage rate in a water distribution system *Water Supply* **21**(8) 4481–92 <https://doi.org/10.2166/ws.2021.194>
- [10] Breiman L, Friedman JH, Olshen RA and Stone CJ 1984 *Classification and Regression Trees* (New York: Routledge) p 368 <https://doi.org/10.1201/9781315139470>
- [11] Zou H and Hastie T 2005 Regularization and variable selection via the elastic net *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**(2) 301–20 <https://doi.org/10.1111/j.1467-9868.2005.00503.x>