Water Resour Manage (2017) 31:1575–1589
DOI 10.1007/s11269-017-1595-8

Ensemble Decision Tree Models Using RUSBoost for Estimating Risk of Iron Failure in Drinking Water Distribution Systems

S. R. Mounce[1] · K. Ellis[1] · J. M. Edwards[1] · V. L. Speight[1] · N. Jakomis[2] · J. B. Boxall[1]

Received: 5 January 2016 / Accepted: 15 February 2017 /
Published online: 9 March 2017

Abstract Safe, trusted drinking water is fundamental to society. Discolouration is a key aesthetic indicator visible to customers. Investigations to understand discolouration and iron failures in water supply systems require assessment of large quantities of disparate, inconsistent, multidimensional data from multiple corporate systems. A comprehensive data matrix was assembled for a seven year period across the whole of a UK water company (serving three million people). From this a novel data driven tool for assessment of iron risk was developed based on a yearly update and ranking procedure, for a subset of the best quality data. To avoid a 'black box' output, and provide an element of explanatory (human readable) interpretation, classification decision trees were utilised. Due to the very limited number of iron failures, results from many weak learners were melded into one high-quality ensemble predictor using the RUSBoost algorithm which is designed for class imbalance. Results, exploring simplicity vs predictive power, indicate enough discrimination between variable relationships in the matrix to produce ensemble decision tree classification models with good accuracy for iron failure estimation at District Management Area (DMA) scale. Two model variants were explored: 'Nowcast' (situation at end of calendar year) and 'Futurecast' (predict end of next year situation from this year's data). The Nowcast 2014 model achieved 100% True Positive Rate (TPR) and 95.3% True Negative Rate (TNR), with 3.3% of DMAs classified High Risk for un-sampled instances. The Futurecast 2014 achieved 60.5% TPR and 75.9% TNR, with 25.7% of DMAs classified High Risk for un-sampled instances. The output can be used to focus preventive measures to improve iron compliance.

Electronic supplementary material The online version of this article (doi:10.1007/s11269-017-1595-8) contains supplementary material, which is available to authorized users.

✉ S. R. Mounce
S.R.Mounce@sheffield.ac.uk

[1] Pennine Water Group, Department of Civil and Structural Engineering, University of Sheffield, Sheffield S1 3JD, UK

[2] Dŵr Cymru Welsh Water, Pentwyn Road, Nelson, Treharris, Mid Glamorgan CF46 6LY, UK

Keywords Water distribution systems · Water quality · Iron · Machine learning · Ensemble decision trees · RUSBoost

1 Introduction

Metals of concern in drinking Water Distribution Systems (WDS) include iron, manganese and aluminium. The primary sources of metals in WDS are carryover from water treatment (Vreeburg and Boxall 2007) and corrosion by-products within pipes (Prasad and Danso-Amoako 2014). Once in the pipe network, metals accumulate and then can be mobilised by hydraulic events, which can result in discolouration and exceedance of regulatory standards (Drinking Water Inspectorate 2014). Iron and manganese are the dominant inorganic materials in most UK discolouration samples (Seth et al. 2003). In the UK, the majority of customer complaints about water quality are related to discolouration, comprising 80% of complaints in 2007 (Drinking Water Inspectorate 2008). Due to the association in relative levels (Cook et al. 2015) of these, elevated iron concentrations can effectively be thought of as an early indicator of discolouration, or discolouration contacts considered as extreme iron events.

WDS are subject to a variety of site-specific hydraulic, chemical and microbiological influences that make it difficult to isolate individual reactions and fully understand the mechanisms at work (Husband and Boxall 2011). This coupled with the various uncertainties surrounding complex, ageing, buried pipe infrastructure means that a deterministic modelling approach for estimating iron failures is rarely possible, even for the most advanced Water Service Providers (WSP). Hence alternatively, data driven techniques can potentially be applied, including analytics, modelling and visualisation, to generate new insight and value from complex multidimensional data, often with limited sampling frequency.

This paper presents an ensemble decision tree methodology developed for estimating both current (‘nowcast’) and future (‘futurecast’) risk of iron failure (acting as a surrogate for all metals and turbidity) using an annual scale across the whole of a WSP’s region with a novel predictive model. The objective of this work was to develop a data driven model that would rank the relative risk of iron failure, thereby enabling a WSP to focus its efforts on the District Management Areas (DMAs) with the highest risk of non-compliance for iron, manganese or turbidity. Results from this predictive model are presented for a case study WSP over a number of historical years.

2 Water Quality Data for Drinking Water Distribution Systems

Current knowledge of the WDS is deficient in providing sufficient descriptions of many of the physical, biological and chemical reactions taking place in pipe networks. Together with uncertainty in the exact state of the network (flow rates, degradation/corrosion levels, leakage etc.), the challenge therefore is to make the best possible use of the data available to prioritise interventions across entire networks. Datasets currently maintained by water companies include historic and updated asset records, discrete water quality sampling and associated laboratory analysis, and continuous / online (hydraulic and some water quality) data collection from an increasing telemetry footprint. Companies also keep records of customer contacts; however such data is highly dependent on the vagaries and unreliable nature of individual customer behaviour. The availability and affordability of varying forms of sensing, smart systems, data storage and transmission technology means water utilities are becoming able to collect more data than ever before. Water utility databases are currently growing rapidly, and will continue to do so.

The processes associated with the operation and maintenance of WDS are generally applied over a long timeframe, often assessing water quality data and customer contacts over a period of months and years, to identify trends in network deterioration (Boxall et al. 2011). In order, the key data used in the UK to inform decision making processes are customer contacts, water quality sample data and analysis and network data such as asset records, burst records and pipe samples. The quality of this data is variable from one water utility to the next. Currently interventions are often responsive to customer contacts leading to a reactive management that does not necessarily deal with the underlying issues. Where good quality data is available, accessible and well maintained, the ability of the distribution engineer to monitor, evaluate and make good decisions with regard to the operation and maintenance of the network is greatly enhanced. However this is rare, and there is a need to aid the decision making process, to make the best possible use of the data that is available.

It has traditionally been difficult to justify efforts to improve data quality in the water industry because, although seen as of interest, investment in asset improvements takes priority. It is challenging to put a price on the value of improved data quality and / or increased data collection and even online instrumentation. However, regulation in the UK is encouraging this with companies in the future seeking a basis for their asset planning with the analysis of data from, or directly related to, their operations. However, water quality data is not currently collected in a consistent and automated manner across networks. Often, WDS water quality is monitored through collection and analysis of discrete samples for a variety of aesthetic, chemical and biological parameters. Water quality data analytics are under-developed because there are no significant deployments of real-time water quality monitoring (UKWIR 2013).

The current nature of water utility network data is that it remains sparse in space (e.g. not all locations are sampled) and time and typically is not linked across functions (e.g. water quality data is not linked to hydraulic model data). Maximising the quality of data (its usefulness) requires consideration of a chain of processes and manipulation, e.g. data source, collection, storage, and the anticipated data end-use. The required blend of foresight and experience means a move towards ‘Big Data’ solutions and so called business intelligence (turning an organisation’s data into patterns that help make intelligent business decisions) in WSPs must be somewhat iterative and will require significant development time. Ultimately this will result in data exploration tools for non-ICT specialists, reducing the cost of and high-fidelity data visualisation and thus enabling human cognition and interpretation. An interim position currently exists, where water companies have substantial databases, but lack the connectivity and methods to extract the full potential benefit. This study seeks to tackle this issue by the integration and analysis of heterogeneous data types.

3 Machine Learning Background

Machine learning is a type of artificial intelligence that enables computerised learning from patterns in data without explicit programming, creating a model that can be applied to new data. In classification, a methodology is used for partitioning data instances into a set of categories which are also referred to as classes or labels (such as ‘Iron Risk High’ or ‘Iron Risk Low’). The classification methodology can be based on a data-driven approach (using supervised learning) which learns directly from relationships among variables in available data instances or patterns (usually referred to as the training set). The class labels of the instances in the training set are known and the aim is to build a model in order to label new instances which are presented. An instantiation of a particular algorithm (such as a decision tree) for a specific training set is called a classifier.

3.1 Data Driven Classification Models

There are a number of approaches that can be used to develop a data-driven classification (risk ranking) model in a water quality application. Regression techniques have been used for decades with some success in water distribution problems (Gibbs et al. 2006). Artificial Neural Networks (ANNs) have been applied to a variety of water distribution problems including in the water quality domain (Wu et al. 2014). However, ANNs operate with a black box approach which can be difficult for users to understand and accept. In contrast, a level of transparency is provided by a decision tree paradigm whereby one can follow a tree structure to understand how a classification has been made (Pedrycz and Sosnowski 2001). Furthermore, decision tree models effectively handle a wide variety of continuous data, categorical data, sparse data and skewed data and require very little in the way of data pre-processing and manipulation, unlike traditional statistical techniques like regression.

3.2 Supervised Classification with Decision Trees

Decision tree algorithms are classifier models that learn from an existing dataset so that a flowchart-like tree structure can be obtained that illustrates relationships between input predictors from the dataset and the target class (Quinlan 1987). The topmost node, called the root of the tree, contains all the instances/observations contained in a dataset used for training. Each internal node specifies a test on an input predictor and each branch in the tree represents an outcome of the test. An instance is classified by starting at the root of the tree and then, depending on the values of the attributes, tracing a path down through the branches of the tree. Eventually, a leaf node is reached at the bottom of the tree and a classification can be obtained using the distribution of instances observed in the leaf. A variety of techniques are currently available for constructing decision tree classifiers. One of the oldest and most widely used is the classification and regression tree (CART) methodology developed in the mid-1980s (Breiman et al. 1984). The CART methodology constructs decision tree classifiers using a top-down divide-and-conquer approach. Decision tree classifiers have been used for assessing drinking water quality (Harvey et al. 2015) and for more intelligent management of water supply systems (Rojek 2014).

An example decision tree for this paper's application is shown in Fig. 1.

Due to the complexity of the iron prediction problem, and since there are comparatively few failures upon which to base a predictive model (approximately 4% failure rate across the period of study), a single decision tree is too simplistic. For this reason an ensemble method was used.

3.3 Ensemble Classifier Approach

An active area of research in Artificial Intelligence (AI) is the development of hybrid architectures and so called ensemble methodologies. Ensemble methods use multiple learning algorithms to improve performance (Rokach 2010). These models have been applied in the

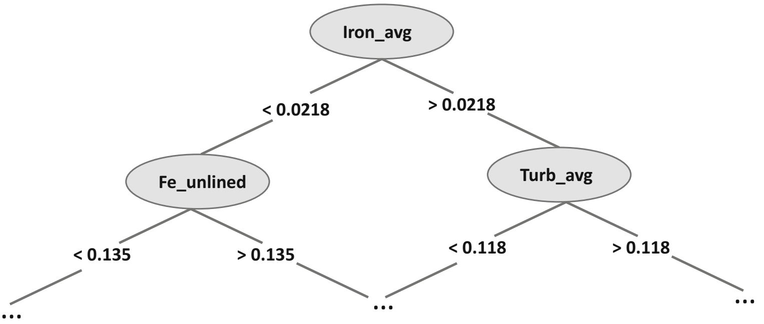

Fig. 1 Example decision tree (shown to depth 2 only) – acronyms from section 4.2

hydroinformatics and water resources domain particularly for regression (Solomatine 2008). For example, Shu and Burn (2004) applied ANN ensembles to the problem of flood frequency analysis and Kim and Seo (2015) applied them for 1-day ahead streamflow forecasting. Model Trees have also been proposed which are tree-structured regression models that associate leaves with multiple linear regression functions used to calculate numerical values (Jung et al. 2010).

Ensemble based classifiers allow the combining of the predictive power of hundreds of individual classifiers (e.g. decision trees or ANN). Decision trees are particularly well-suited for ensembles because they are fast and unstable (Gashler et al. 2008). Decision trees can be unstable because small variations in the data might result in a completely different tree being generated. By using them within an ensemble this problem is mitigated. For ensemble learning, multiple classifiers are used and weighted and then combined in order to obtain a classifier with superior performance to individual classifiers. In one sense, the approach captures the 'wisdom of the crowd' concept (Baker and Ellison 2008) in that humans will seek and weigh several opinions before making an important decision.

Three common methods are utilised for constructing ensembles of decision trees: boosting, bagging (bootstrap aggregation) and random subspace (random forests) as compared in Dietterich (2000). The most popular model-guided instance selection is boosting. Boosting is a general method for improving the performance of a weak learner (such as a decision tree). The method works by repeatedly running a weak learner on various distributed training data. The classifiers produced by the weak learners are then combined into a single composite strong classifier in order to achieve a higher accuracy than the individual trees would be capable of. AdaBoost (Adaptive Boosting) was first introduced in Freund and Schapire (1996), and is a popular ensemble algorithm for binary classification that improves the simple boosting algorithm via an iterative process whose main aim is to give more focus to patterns that are harder to classify. A range of algorithm variants exist as described in section 4.3.

4 Methodology

4.1 Case Study Data Set

A dataset of relevant data from multiple corporate systems including asset characteristics, water treatment works and service reservoir sampling, regulatory water quality sampling and customer contacts was assembled. The WSP's supply area (a European country of size approximately 20,800 km^2 and population three million people) is divided into 86 Water Quality Zones (WQZs) based on their source of supply, with a total of 1312 DMAs within those WQZs consisting of 26,500 km of water mains. Traditional analysis of variables commonly associated with iron compliance (e.g. extent of unlined iron pipe percentage or number of customer contacts) revealed that no single parameter could reliably predict iron failures, thus necessitating a multivariate approach as implemented herein (Ellis et al. 2015).

4.2 Data Analysis and Pre-Processing

The compilation of data from a variety of disparate WSP sources required a significant amount of computational effort but resulted in a comprehensive dataset to support iron risk analysis at a company-wide scale. Whilst it would be desirable to consider iron compliance at the individual pipe level, the data requirements for asset characteristics for the entire network made this goal infeasible. The WSP indicated that a DMA-level spatial analysis would suit their operational needs and summary statistics were available for DMA characteristics, such as pipe material. Therefore a DMA-based scale of analysis was selected.

Iron failures are relatively rare events and therefore the temporal scale of the analysis needs to reflect the failure frequency of occurrence (in addition some DMAs may be un-sampled during a particular year). Use of a monthly scale would have resulted in the majority of DMAs with zero failures. Therefore a yearly time scale was selected to allow for sufficient number of failures to be included in each analysis period while still providing relatively up-to-date information, with the ability to rerun analysis on an annual basis. A complex and repeatable data processing and formatting process involving multiple MS Excel (Microsoft Corporation, New Mexico) and MATLAB® 2014b (The Mathworks Inc., Massachusetts) operations led to the development of year-by-year data matrices for the period January 2008 to September 2014. The resulting data matrix includes 1312 DMAs with 56 fields of associated parameters per year (see Online Resource 1). Many of the data fields were calculated from raw data sources. Some examples include:

- Water quality sampling data linked to DMA
- Water chemistry data for Service Reservoirs (SRVs) and Water Treatment Works (WTW)
- Water supply system connectivity such as which WTWs supply which WQZ (and hence DMA) and which SRVs supply which DMAs
- DMA information: % of certain types of pipe material calculated from overall pipe lengths
- Clusters of customer contacts: individual customer contacts are rarely indicative of a wide-scale water quality problem, but multiple contacts within a short time window could be (Husband et al. 2010). The Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm was used in MATLAB to identify temporal clusters of contacts (Ester et al. 1996). A cluster was defined as having three or more contacts at the DMA-level (minPts: minimum number of points to form a cluster) within a 48 h rolling window (ε: neighbourhood distance). The results were then limited to WQZs with at least two DMAs containing clusters and this field was then added to the matrices for each year: 'cc_clusters'.

Overall, 25% of data were missing, highlighting the need for data analysis techniques that can flexibly handle missing data. Missing data occurred because a) the regulations do not require all DMAs to be sampled for all parameters every year, b) some datasets were only brought on-line later in the study period and c) asset data was not always present due to the age of the network.

The data matrix facilitated an extensive exploration of parameters associated with iron failures using a variety of traditional statistical and machine-learning techniques. Self-organising maps (SOMs) were used to visually explore the interrelationships between various types of water quality, hydraulic and asset parameters (Ellis et al. 2015). The application of SOMs provided a method for visualisation of clusters within different groupings of parameters and insight into predictor variables. For example, iron and turbidity were closely linked at the DMA level. Iron and manganese were sometimes associated with each other but occurrences of elevated manganese also took place without an accompanying rise in iron. The presence of unlined cast iron pipe in a DMA was not necessarily correlated with a higher number of iron failures. Customer contacts in general were correlated with iron failures but to a different degree depending on the predominant pipe material within the DMA. The presence of a cluster of customer contacts within a WQZ that spanned multiple DMAs appeared to be associated with some iron failures, indicating that trunk main operations/bursts or WTW iron sources were playing a role in those failures. Additional statistical and data driven evaluation to inform feature selection was conducted (such as a bivariate correlation coefficients and Treebagger for out of bag feature importance). Online Resource 2 provides some further details. The three most important features are indicated to be the average of median iron measurements, the average of median turbidity measurements and the total number of customer contacts (complaints) about water quality. The final model parameters were selected using data analysis, in the context of data availability and to maintain relative simplicity for decision tree interpretation. Note that SRV iron concentration data was not utilised in the final set of predictors. It had very low data availability 2008–2011 (less than 1%), but for the period 2012–2014 this had improved to 63.5% availability, which is better but still below an acceptable threshold for inclusion. The data analysis indicated that SRV iron could add to accuracy of the model once the data coverage is improved.

Based on the results of the data exploration, the following matrix parameters for the ensemble model were selected:

- Average of median iron concentration from all samples per DMA per year, 'Iron_avg'
- Average of median manganese concentration from all samples per DMA per year, 'Mn_avg'
- Average of median turbidity from all samples per DMA per year, 'Turb_avg'
- Percentage of unlined iron pipe per DMA, 'Fe_unlined'
- Percentage of lined iron pipe per DMA, 'Fe_lined'
- Average of median turbidity from supplying WTWs per year and, where applicable, calculated as a weighted average for DMAs with multiple supplying WTWs, 'WTW_turb_avg'
- Number of customer contacts per DMA per year, 'cc'
- Number of customer contacts in WQZ-level clusters per DMA per year, 'cc_clusters'

Customer contacts (and the derived customer contact clusters) were only available for year 2011 onwards.

4.3 Ensemble Design and Algorithm Implementation

The target for the ensemble was defined to be the classification of risk for each DMA as follows: High (H) which corresponds to one or more iron failures in the DMA or Low (L) which corresponds to zero iron failures in the DMA (for the yearly period under consideration).

Building an effective ensemble classification model can be a challenging task if the data used to train the model are imbalanced. When examples of one class greatly outnumber examples of the other class(es), traditional machine learning algorithms tend to favour classifying examples as belonging to the overrepresented and dominating (majority) class. In this study, due to the very limited number of iron failures within the water quality dataset, this imbalanced situation exists. Typical data sampling approaches are to oversample the minority class (see Chawla et al. 2002 for a description of SMOTEBoost) or undersample the majority class. The RUSBoost (Random Under Sampling) algorithm is designed to classify when one class has many more observations than another and good reference results have been obtained (Seiffert et al. 2010). Blackard and Dean (1999) describe an ANN classification of an imbalanced dataset achieving 70.6% accuracy, whereas RUSBoost obtained over 76% classification accuracy. The majority of class-imbalance learning techniques currently implemented, including RUSBoost, have been designed for two-class problems. Figure 2 provides an outline of the RUSBoost algorithm with *X* the feature space and *D* the weights.

MATLAB implements ensembles with boosting and bagging – which trains each model in the ensemble using random subsets of the training data. Also, rather than just a single classification, by aggregation across the learners the relative weight for a particular class label can be obtained. Due to the very limited number of iron failures (typically around 4–7% across DMAs in a particular year), results from many weak learners (1000 decision trees were utilised

```
RUSBoost Algorithm
Given: Set $S$ of examples $(x_1, y_1),...,(x_m, y_m)$ with minority class $y^r \in |Y| = 2$
Weak learner (decision tree), WeakLearn
Number of iterations, $T$
Desired percentage of total instances to be represented by the minority class, $N$
    1 Initialise $D_1(i) = \frac{1}{m}$ for all $i$.
    2 Do For $t = 1,2,...,T$
        a. Create temporary training dataset $S_t^{'}$ with distribution $D_t^{'}$ using random under-sampling
        b. Call WeakLearn, providing it with examples $S_t^{'}$ and their weights $D_t^{'}$.
        c. Get back a hypothesis $h_t : X \times Y \rightarrow [0,1]$.
        d. Calculate a pseudo-loss (for $S$ and $D_t$):
        $\in_t = \sum_{(i,y):y_i \neq y} D_t(i)(1 - h_t(x_i, y_i) + h_t(x_i, y))$.
        e. Calculate the weight update parameter:
            $\alpha_t = \frac{\in_t}{1 - \in_t}$.
        f. Update $D_t$:
            $D_{t+1}(i) = D_t(i) \alpha_t^{\frac{1}{2}(1 + h_t(x_i, y_i) - h_t(x_i, y : y \neq y_i))}$
        g. Normalise $D_{t+1}$: Let $Z_t = \sum_i D_{t+1}(i)$.
            $D_{t+1}(i) = \frac{D_{t+1}(i)}{Z_t}$
    3 Output the final hypothesis:
        $H(x) = \arg\max_{y \in Y} \sum_{t=1}^{T} h_t(x, y) \log \frac{1}{\alpha_t}$.
```

Fig. 2 RUSBoost (adapted from Seiffert et al. 2010)

in the final models, with deep trees for higher ensemble accuracy and setting minimal leaf size of 1 and learning rate of 0.1) were melded into one high-quality ensemble predictor using RUSBoost in this application. A protocol for equalising classes and randomly removing data points for particular model subsets is used to remove the imbalance. The ensemble results were obtained as predictions of relative likelihood of risk per DMA: High or Low. The model also provides the weighting of High vs Low based on the relative weighting across the set of decision trees. For each observation and each class, the score generated by each decision tree is the probability of this observation originating from this class computed as the fraction of observations of this class in a tree leaf. These scores are averaged over all trees in the ensemble. Finally, a ranking placement of each DMA can be calculated based on the weighting.

Table 1 provides two key metrics for the holdout dataset (described later in section 5.1) when using the ensemble methods AdaBoostM1, LogitBoost, GentleBoost (all described in Friedman et al. 2000), Bag (Breiman 1996), RobustBoost (Freund 2009) and LPBoost (Warmuth et al. 2006). RUSBoost has the highest TPR value by a large margin due to imbalanced classes.

4.4 Nowcast and Futurecast Models

The availability of the historical data provided and the design of a yearly temporal scale model introduced two useful options for using the data. Since it is envisaged the model (s) will be run at the end of each calendar year (when all necessary data is compiled per DMA and the data matrix for that year generated) two models were decided upon: 'Nowcast': the situation at end of calendar year (all data used in training and testing) and 'Futurecast': attempt to predict end of next year fail situation from this year's data (hence test data is completely 'unseen').

4.4.1 Nowcast

For the Nowcast model, all years up to and including the present year were used to train the model with a target output of H/L ranking for each DMA for the present year. Then, the present year's data was presented as input without a target to get the predicted model outputs (predicted class (H/L), HIGH SCORE, LOW SCORE and rank), as shown in Fig. 3a. Whilst the present year's Iron fails will be known at the end of the year and thus a model is not needed to predict them, their inclusion in the analysis allows the Nowcast to rank the current performance of DMAs that have not been sampled for iron during the year (for example, in 2013 this was 33% of DMAs).

4.4.2 Futurecast

For the Futurecast model, all years up to the present year minus 1 are used to train the model, with an output target of H/L DMA ranking for the next year. This permits the prediction of iron risk one year in advance, using data from up to the end of the current year. So for Futurecast

Table 1 Comparison of ensemble methods on imbalanced dataset

Ensemble Method Metric	AdaBoostM1	LogitBoost	GentleBoost	Bag	RobustBoost	LPBoost	RUSBoost
TPR	0.20	0.24	0.27	0.29	0.26	0.28	0.72
TNR	0.98	0.97	0.98	0.98	0.97	0.98	0.87

2014, at the end of 2013 the model outputs including risk ranking are available for the predicted situation at the end of 2014, as shown in Fig 3b.

5 Results and Discussion

5.1 Nowcast Hold-Out Validation

Initial exploration of the data involved a partition for quality assessment of the ensemble classifier approach. The Nowcast approach using RUSboost and all years holding 50% of the data back for completely unseen analysis, resulted in a model that gave 71.5% correct classification of iron fail DMAs (H) and 83.5% of no fail DMAs (L) (note that only DMAs that had been sampled were used). Further refinement of the model using years 2011 to 2013 with full data availability of all identified input attributes resulted in better accuracy than the all (six) years analysis, with the best model obtained after experimentation and analysis giving 71.9% correct classification of iron fail DMAs (H) and 87.3% of no fail DMAs (L) on 50% unseen holdout data. Overall percent accuracy may not provide a reliable indicator of predictor performance for models trained using imbalanced datasets as it may provide a false impression of capabilities for the minority class of interest, which in this case is High Risk DMAs. A better performance metric is a confusion matrix. In binary classification problems, classification can be grouped into four categories as illustrated in Table 2. This confusion matrix summarises the percentage of predictions versus actual values in the Low and High categories and thus gives an indication of the number of false positives and false negatives associated with the output. Ideally, there will be very few false positives and false negatives (i.e. high accuracy). Confusion matrices should be considered along with overall model percentage accuracies to judge the model performance. Two other metrics can be useful which are readily available: sensitivity or True Positive Rate (TPR) = $\frac{TP}{TP+FN}$ and specificity or True Negative

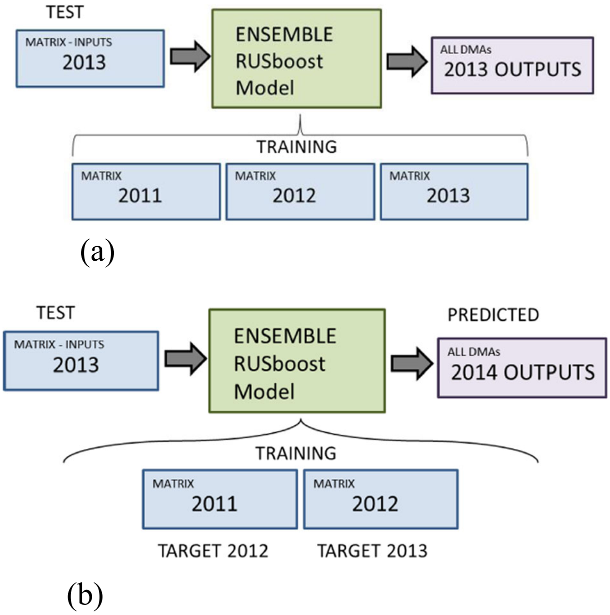

Fig. 3 **a** Nowcast scheme for ensemble model (example for 2013). **b** Futurecast scheme for ensemble model (example for 2014)

Table 2 Confusion matrix for Nowcast 2011–2013 (50% holdout)

Actual	Percentage of DMAs Predicted as:	
	Low	High
Low	85.4 (TN)	14.6 (FP)
High	24.7 (FN)	75.3 (TP)

(TP) True Positive = a high risk DMA correctly predicted to be in the 'H' class by the model

(FP) False positive = a low risk DMA incorrectly predicted to be in the 'H' class by the model

(TN) True Negative = a low risk DMA correctly predicted to be in the 'L' class by the model

(FN) False Negative = a high risk DMA incorrectly predicted to be in the 'L' class by the model

Rate (TNR) = $\frac{TN}{FP+TN}$. For reference, the data set in Table 2 consisted of 6.8% H labels and 93.2% L labels emphasising the imbalanced class categories (for training data TPR = 0.753, TNR = 0.854).

5.2 Summary Nowcast Results

The final models use all available data with no hold out. Three Nowcast models were developed, 2012, 2013 and 2014. The later years' models were trained on more years of historical data (Fig. 3a). Results for Nowcast 2014 are presented in Table 3. For the year 2014, 3.3% of DMAs had class label H, 44% had class label L and 52.7% were unsampled (96.7% predicted as Low, 3.3% as High).

The Nowcast 2014 Model had the ability to predict 100% of the high risk DMAs and a reasonable number (3.3%) of high risk DMAs in the un-sampled category. For earlier models, Nowcast 2012 had performance TPR = 1.0, TNR = 0.99 (with 21.8% predicted High from the un-sampled DMAs) and Nowcast 2013 had performance TPR = 1.0, TNR = 0.97 (10.7% predicted High from the un-sampled DMAs).

The model output also provides a weighting for each of the (potentially) 1000 decision trees within the ensemble. An evaluation of the weighting for the top three trees was performed to understand the degree to which the model was able to be described by a few trees. The top three decision trees contributed 19.1% to the scoring for this model (see Online Resource 3). Exploring these larger weighted decision trees can enable an understanding of what is most significant in determining a class label for a particular DMA.

5.3 Summary Futurecast Results

Three Futurecast models were developed, 2013, 2014 and 2015 according to Fig. 3b. Predicting the DMAs' relative iron risk for the following year is of course a much more challenging problem, because the Futurecast performance is tested on unseen data. Table 4 provides the Futurecast 2014 results. As was the case for the Nowcast models, it was expected that training the Futurecast models on larger datasets would improve the proportion of correctly classified DMAs and this was observed with improved performance in later models (for Futurecast 2013, TPR = 0.541 and TNR = 0.749). The Futurecast 2014 model was able to accurately predict 60.5% of the high risk DMAs (in comparison to 54.1% for Futurecast 2013). However a large number (25.7%) of high risk un-sampled DMAs were predicted, which is significantly worse than the actual observed performance. The contribution to scoring from the

Table 3 Confusion matrix for Nowcast 2014

Actual	Percentage of DMAs predicted as:	
	Low	High
Low	95.3 (TN)	4.7 (FP)
High	0.0 (FN)	100.0 (TP)
TPR = 1.0 TNR = 0.953		

top three trees from the ensemble was only 7.2% for this model, compared with 19.1% for Nowcast.

5.4 Futurecast Predictions for 2015

For Futurecast 2015, a prediction of the system performance for 2015 was performed using data from 2011 through 2013 to train the model, with 2014 as test data (the matrix for 2015 being unavailable) This model yielded a total contribution from the top three trees of 7.2%, which is similar to the other Futurecast models. A large number of DMAs (22.9%) are predicted to be high risk as was the case for previous Futurecast models but the actual system performance is likely to be similar to previous years.

Classifying DMAs as either High or Low risk presents the WSP with a list of DMAs requiring attention and the Futurecast 2015 results in 300 High risk DMAs for 2015 when using only the binary output of the ensemble model (see Fig. 4a plotted using GIS: ArcMap 10.1 (ESRI, California)). In order to produce a more fine grained and focussed assessment, the model output also ranks the DMAs so that the highest risk sites can be targeted first (using the generated weightings across the ensemble). Figure 4b shows the Futurecast 2015 rank predictions colour-coded, with the 20 DMAs having the highest risk in red (darkest shade).

This ranking list provides a potential method for prioritising interventions into high risk DMAs, some of which may experience iron failures during 2015. Early findings from 2015's compliance for the live network showed that of the five iron non-compliances that had been detected (as of May 2015), four were in DMAs predicted as High risk by the Futurecast 2015 model. The fifth failing DMA was one where no past sampling data were available for training or testing the model (Ellis et al. 2015).

5.5 Discussion

The ensemble technique combines many weak learners in an attempt to produce a strong learner. If an ensemble could be reduced to a single decision tree with univariate splits, there

Table 4 Confusion matrix for Futurecast 2014

Actual	Percentage of DMAs predicted as:	
	Low	High
Low	75.9 (TN)	24.1 (FP)
High	39.5 (FN)	60.5 (TP)
TPR = 0.605 TNR = 0.759		

would be no point in growing the ensemble - i.e. it is hard to visualise in multivariate space. However, just a few trees may represent a significant weighting of the ensemble. The resulting image of a tree is interpretable, even by those unfamiliar with the data mining process. For example, in Fig. 1, DMAs are initially separated based on median iron concentration, with a concentration cut-off of approximately 0.02 mg l^{-1}. Those DMAs with iron concentrations of less than this value were then classified by percentage of unlined iron pipeline in the DMA and those greater than the cut-off were alternatively divided by median turbidity. This process continues down branches of the tree until a leaf node is reached (class label of H/L).

The RUSBoost decision tree ensemble methodology has been demonstrated to have potential for producing a DMA-specific prediction of risk of iron failure using multiple data sources assembled in a comprehensive yearly scale matrix. The model presented here was tested on an incomplete dataset (January to September 2014). In basing the predictions for Futurecast 2015 on three years of training data it could be more accurate than those for Futurecast 2014. It would nevertheless have been beneficial to be able to test the model on a full year's data and it is possible that accuracy may have been compromised by having a truncated dataset for 2014. In future years, larger training datasets will be available to improve the accuracy of the model and the WSP should have greater confidence in utilising the model outputs in operational decisions.

The decision tree(s) utilises the water quality and system parameters that are likely to result in iron non-compliance. The decision tree ensemble models generated in this project (in MATLAB), produce an output that can be plotted in GIS software to highlight the DMAs most at risk of iron failures. The results from this work are already enabling the WSP to better target specific areas of their network to prevent iron non-compliance, with a view to extending the impact with future iterations of the model.

6 Conclusions

This paper presents a novel data driven methodology to enable the estimation of iron failure risk from sparse multidimensional data from UK water company corporate systems. Iron failures result in a significant quantity of customer complaints and water company investment, so methods to understand and predict intervention strategies are of great value. The work has shown how the use of ensemble methods with multiple learning algorithms (of the same base learner such as a decision tree) can be used to improve performance and supply sufficient

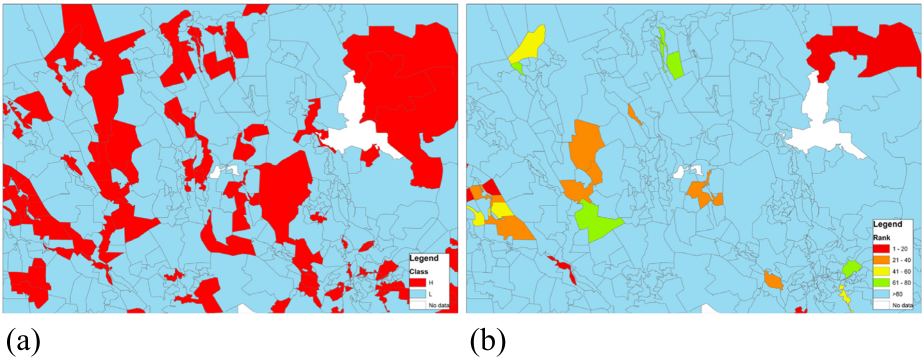

Fig. 4 **a** Binary output for Futurecast 2015 plotted on excerpt of WSP region. **b** Futurecast 2015 ranking results plotted on excerpt of WSP region (scale non-linear to highest risk DMAs)

coverage across a sparse multidimensional problem space. In particular the paucity of target data, iron fails, was overcome with the results from multiple ‘weak’ decision trees melded into one high-quality ensemble predictor using the RUSBoost algorithm, which is designed for situations where one class has many more observations than another (imbalanced class label distribution). The final outputs are ensembles of decision tree classifiers that enable assessment of both current and future iron risk at DMA scale. Results can be obtained as predictions of relative likelihood of risk (by averaging across decision tree weightings over the ensemble) of iron failure per DMA: High or Low, and as a relative ranking of DMAs by risk.

Two model formulations are presented here, Nowcast and Futurecast. The Nowcast model allows comprehensive review of current data and extrapolation to unsampled areas (as much as 33% for the data set explored here). The Nowcast achieved 71.9% TPR and 87.3% TNR on a 50% holdout. The Nowcast 2014 model with all data achieved 100% TPR and 95.3% TNR, with 3.3% of DMAs classified High Risk for the un-sampled instances. The Futurecast 2014 model achieved 60.5% TPR and 75.9% TNR, with 25.7% of DMAs classified High Risk for the un-sampled instances. The real world success and value of this Futurecast has been demonstrated by 80% correct prediction of failures (with the only missed fail being in a previously un-sampled DMA) from January to May 2015.

Overall this work demonstrated how ensemble decision trees can be applied to explore and data mine sparse, multidimensional space for WDS data, producing outputs that show operators and managers the basis of how results have been obtained, making them an accessible tool for decision-making unlike black box approaches. The widespread application of these types of ensemble decision tree models for classification analysis within the water industry could revolutionise the way that investment is justified, prioritised and implemented for proactive pipeline management, making them more efficient and more cost-effective.

Acknowledgements The authors would like to gratefully thank Dŵr Cymru Welsh Water for their support, data provision and permission to publish the details included herein. The work reported here was also supported by the UK Engineering and Physical Sciences Research Council (EPSRC) Challenging Engineering grant EP/G029946/1 and Platform Grant EP/1029346/1.

References

Baker L, Ellison D (2008) The wisdom of crowds - ensembles and modules in environmental modeling. Geoderma 147:1–7

Blackard JA, Dean DJ (1999) Comparative accuracies of artificial neural networks and discriminant analysis in predicting forest cover types from cartographic variables. Comput Electron Agric 24:131–151

Boxall JB, Machell J, Dewis N, Gedman K, Saul A (2011) Operation, maintenance and performance. In: Water Distribution Systems ICE (ed) Dragan A Savic and John K. Banyard

Breiman L (1996) Bagging predictors. Mach Learn 26:123–140

Breiman L, Friedman J, Olshen R, Stone C (1984) Classification and regression trees. CRC Press

Chawla N, Bowyer K, Hall L, Kegelmeyer W (2002) Smote: synthetic minority over-sampling technique. J Artif Intell Res 16:321–357

Cook DM, Husband PS, Boxall JB (2015) Operational management of trunk main discolouration risk. Urban Water J. doi:10.1080/1573062X.2014.993994

Dietterich TG (2000) An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization, Mach. Learning 40(2):139–157

Drinking Water Inspectorate (2014) Drinking water 2013: a report by the chief inspector of the Drinking Water Inspectorate. Drinking Water Inspectorate, London

Ellis K, Mounce SR, Edwards JM, Speight VS, Jakomis N, Boxall JB (2015) Interpreting and estimating the risk of iron failures. Procedia Engineering 119(2015):299–308

Ester M, Kriegel H, Sander J, Xu X (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. In: Simoudis E, Han J, Fayyad U (eds) Proceedings of the second international conference on knowledge discovery and data mining. AAAI Press, California, pp 226–231

Freund Y (2009) A more robust boosting algorithm. arXiv:0905.2138v1, 2009

Freund Y, Schapire RE (1996) Experiments with a new boosting algorithm. In: Machine learning: proceedings of the thirteenth international conference, 325–332

Friedman J, Hastie T, Tibshirani R (2000) Additive logistic regression: a statistical view of boosting. Ann Stat 28(2):337–407

Gashler M, Giraud-Carrier C, Martinez T (2008) Decision tree ensemble: small heterogeneous is better than large homogeneous. The Seventh International Conference on Machine Learning and Applications 2008:900–905. doi:10.1109/ICMLA.2008.154

Gibbs MS, Morgan N, Maier HR, Dandy GC, Holmes M (2006) Investigation into the relationship between chlorine decay and water distribution parameters using data driven methods. Math Comput Model 44(5–6): 485–498

Harvey R, Murphy HM, McBean EA, Gharabaghi B (2015) Using data mining to understand drinking water advisories in small water systems: a case study of Ontario first nations drinking water supplies. Water Resour Manag 29(14):5129–5139

Husband P, Boxall J (2011) Asset deterioration and discolouration in water distribution systems. Water Res 45: 113–124

Husband P, Whitehead J, Boxall J (2010) The role of trunk mains in discolouration. Water Management 163(WM8):397–406

Inspectorate DW (2008) Drinking water 2008; drinking water in England and Wales 2008. A report by the Chief Inspector, Drinking Water Inspectorate, London 83

Jung NC, Popescu I, Kelderman P, Solomatine DP, Price RK (2010) Application of model trees and other machine learning techniques for algal growth prediction in Yongdam reservoir, Republic of Korea. J Hydroinf 12(3):262–274

Kim SE, Seo IW (2015) Artificial neural network ensemble modeling with exploratory factor analysis for streamflow forecasting. J Hydroinf 17(4):614–639

Pedrycz W, Sosnowski ZA (2001) The design of decision trees in the framework of granular data and their application to software quality models. Fuzzy Sets Syst 123:271–290

Prasad T, Danso-Amoako E (2014) Influence of chemical and biological parameters on iron and manganese accumulation in water distribution networks. Procedia Engineering 70:1353–1361

Quinlan JR (1987) Simplifying decision trees. International Journal of Man-Machine Studies 27(3):221. doi:10.1016/S0020-7373(87)80053-6

Rojek I (2014) Models for better environmental intelligent management within water supply systems. Water Resour Manag 28(12):3875–3890

Rokach L (2010) Ensemble-based classifiers. Artif Intell Rev 33(1–2):1–39

Seiffert C, Khoshgoftaar TM, Hulse JV, Napolitano AA (2010) RUSBoost: a hybrid approach to alleviating class imbalance. IEEE Transaction on Systems, Man and Cybernetics-Part A: Systems and Human 40:1

Seth A, Bachmann R, Boxall J, Saul AJ, Edyvean R (2003) Characterisation of materials causing discolouration in potable water systems. Water Sci Technol 49(2):27–32

Shu C, Burn DH (2004) Artificial neural network ensembles and their application in pooled flood frequency analysis. Water Resour Res 40:W09301. doi:10.1029/2003WR002816

Solomatine DP (2008) Committees of models in hydrologic modelling: boosting, mixtures and trees. In: Practical Hydroinformatics: Computational Intelligence and Technological Developments in Water Applications (Abrahart, See, Solomatine, eds), Springer-Verlag

UKWIR (2013) “Cost Benefit Analysis of Ubiquitous Data Collection in Water Distribution - CBA Scenarios”. 13/DW/12/2 - ISBN: 1 84057 692 8

Vreeburg J, Boxall J (2007) Discolouration in potable water distribution systems. Water Res 41:519–529

Warmuth M, Liao J and Ratsch G (2006) Totally corrective boosting algorithms that maximize the margin. Proc. 23rd Int'l. Conf. on Machine Learning, ACM, New York, 1001–1008

Wu W, Dandy GC, Maier HR (2014) Protocol for developing ANN models and its application to the assessment of the quality of the ANN model development process in drinking water quality modelling. Environ Model Softw 54:108–127