Proceedings Paper:
Kazemi, E., Kyritsakas, G., Husband, S. orcid.org/0000-0002-2771-1166 et al. (3 more authors) (2023) Predicting iron exceedance risk in drinking water distribution systems using machine learning. In: IOP Conference Series: Earth and Environmental Science. 14th International Conference on Hydroinformatics, 04-08 Jul 2022, Bucharest, Romania. IOP Publishing , 012047.

https://doi.org/10.1088/1755-1315/1136/1/012047

Predicting iron exceedance risk in drinking water distribution systems using machine learning

Predicting iron exceedance risk in drinking water distribution systems using machine learning

Ehsan Kazemi[1], Grigorios Kyritsakas[1], Stewart Husband[1], Katrina Flavell[2], Vanessa Speight[1] and Joby Boxall[1]

[1] Department of Civil and Structural Engineering, The University of Sheffield, Mappin Street, Sheffield S1 3JD, UK
[2] Yorkshire Water Services Limited, Western House, Halifax Road, Bradford BD6 2SZ, UK

e.kazemi@sheffield.ac.uk

Abstract. A Machine Learning approach has been developed to predict iron threshold exceedances in sub-regions of a drinking water distribution network from data collected the previous year. Models were trained using parameters informed by Self-Organising Map analysis based on ten years of water quality sampling data, pipe data and discolouration customer contacts from a UK network supplying over 2.3 million households. Twenty combinations of input parameters (network conditions) and three learning algorithms (Random Forests, Support Vector Machines and RUSBoost Trees) were tested. The best performing model was found to be Random Forests with input parameters of iron, turbidity, 3-day Heterotrophic Plate Counts, and high priority dead ends per District Metered Area. Different exceedance levels were tested and prediction accuracies of above 70% were achieved for UK regulatory concentration of 200 µg/L. Predicted probabilities per network sub-region were used to provide relative risk ranking to inform proactive management and investment decisions.

1. Introduction

It is recognised that increased iron concentrations in potable water are associated with customer contacts for discoloured water. Past research indicates that high iron concentrations in drinking water distribution systems (DWDS) are related either to insufficient water treatment or to corrosion phenomena in distribution pipes [1]. While there has been progress understanding the role of iron concentrations and discolouration, most of these phenomena are addressed by water utilities in a reactive way. This is mainly due to the cost issues – such as that water utilities have to focus activities where the reward (or more likely penalty avoided) is greatest – as well as the complex interactions and processes occurring within DWDS and the vast scale and individual complexity of networks together with a lack of predictive tools. It is therefore important to research and develop methods that when applied could transform this reactive approach into a proactive one by providing information that could inform decision making regarding strategical interventions and investment within DWDS.

Water utilities collect a vast number of discrete samples from their customer taps to measure various water quality parameters and regulatory compliance. While this is a huge effort and creates large datasets it is still sparse (spatially and temporally) data compared to the size and complexity of DWDS. Prior work on discolouration that applied machine learning (ML) methodologies to this kind

of data indicated that these methods have the potential for both informing the causes of discolouration and estimating the risk of iron exceedance in the DWDS [2, 3].

In this paper, a classification ML methodology for the prediction of iron exceedance in DWDS sub-regions of a water utility in the UK is presented. The methodology uses as inputs the annual averages of various water quality parameters in previous years at the district meter area (DMA) scale for the prediction of iron exceedance risk in DMAs for the following year. In addition, the methodology provides information regarding the probability of iron exceedance per DMA and thus facilitates creating a DMA risk ranking that can direct and prioritise interventions.

2. Data and approach

The datasets used in this study were obtained from a large drinking water distribution network in the UK, owned and operated by Yorkshire Water, supplying over 2.3 million households collected over a 10 year (pre-COVID) period. The dataset includes 134,803 regulatory water quality samples (including iron, manganese, turbidity, aluminium, chlorine and many others, but not always the same parameters for every sample), 62,695 water quality customer contacts, disinfection type per Water Supply Zone (WSZ) and static asset pipe data (material, diameter, length, and age).

Firstly, a qualitative analysis using an unsupervised ML technique called Self-Organising Map (SOM) was used to analyse the raw data and thus identify the major relationships and key parameters influencing iron exceedance. Then, a risk model was developed based on supervised ML algorithms to quantify the risk of iron exceedance in the network. This was performed by mapping input parameters (identified by the SOMs as the major factors influencing iron exceedance) to the output parameter (the occurrence of iron exceedance) and then predicting probability of exceedance.

3. Identification of key parameters

SOM is a type of data clustering technique, suitable for sparse and incomplete data, especially when the relationships within data are complex and highly non-linear [4]. It uses unsupervised learning algorithms to train the model, i.e., all parameters are fed into the model as input, and through producing a two-dimensional representation of the high dimensional data set (due to multiple variables), the linkages between them are qualitatively and visually investigated.

Table 1. Variables used in the SOM analysis example and as predictors in the risk model.

Variable (unit)	Data source	Comments
Iron (μg/l)	Water quality data	Regulatory samples
Manganese (μg/l)	Water quality data	Regulatory samples
Total chlorine (mg/l)	Water quality data	Regulatory samples
Turbidity (NTU)	Water quality data	Regulatory samples
3-day heterotrophic plate count (HPC) (no/ml)	Water quality data	Regulatory samples
Number of customer contacts per DMA per year	Customer contact data	-
DMA-clustered customer contacts per DMA per year	Calculated	Each cluster includes a minimum of 5 customers contacts per DMA over a minimum of a 2-day period. This parameter is the sum of these clusters per DMA per year
High priority dead-ends per DMA	Standalone data	Total number of dead-end pipes that, according to water utility's estimations, are at higher risk of discolouration per DMA

To uncover the important relationships in the present dataset and identify the key parameters, various combinations of the parameters introduced in Section 2 were tested with SOM. It was found that, for example, the level of iron within samples is not linked to properties of the pipe (material, diameter, age) closest to the location of the property where the water quality sample was collected and that the level of iron within samples is linked to iron, manganese, turbidity, 3-day heterotrophic plate count (HPC), low chlorine, residual type (chlorine and chloramine), number of customer contacts per District Metered Area (DMA) per year, the DMA clustered customer contacts per DMA per year, and high priority dead ends per DMA. These parameters were thus suggested to be used in the risk model to predict iron exceedance. Some of the parameters used in the SOM analysis are presented in Table 1 and an example output is shown in figure 1.

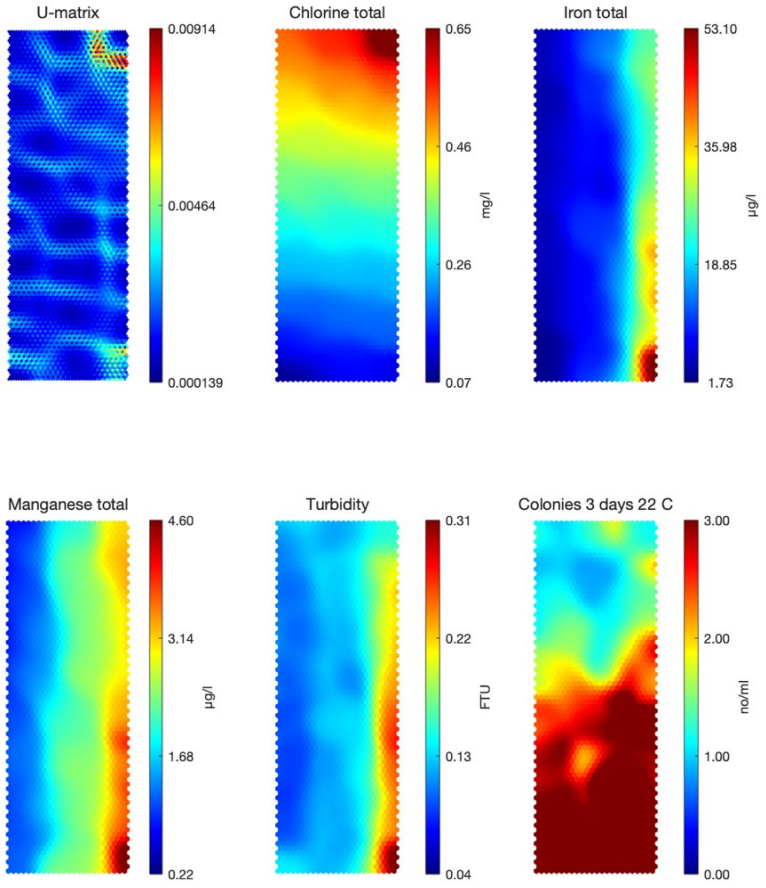

Figure 1. An example SOM developed in this study. Lattices from top left represent the U-matrix, chlorine, iron, manganese, turbidity, and 3-day HPC, respectively.

Figure 1 shows one example of the many SOMs performed in this study. On the lattices, each hexagonal cell (neuron) represents a group of clustered observations; the spatial location of a cell in a topographic map corresponds to a particular domain or feature drawn from the input data; colours show the value of the variables (red: high, blue: low) and each cell in the same position on different maps corresponds to the same cluster of observations/data. In addition to the maps of parameters, an additional lattice called unified distance matrix, or 'U-matrix' is provided (see Figure 1, top left) which shows the strength of the clusters. Blue areas on the U-matrix correspond to the clusters and red lines correspond to where the clusters are separated. This example shows that iron is strongly linked with manganese and turbidity and that high iron is also correlated with high 3-day HPC numbers and low chlorine concentrations in the customers' taps.

4. Risk model

Three supervised ML models were developed from classification Random Forests (CRF) [5], Support Vector Machines (CSVM)[6], and Boosted Trees (CBT) learning algorithm based on Random Undersampling Boosting (RUSBoost) [7]. These models were selected as they provide 'white box' approaches, such that the role of different parameters in the predictions can be understood and appreciated by end users aiding acceptance and uptake. The analysis was performed at a DMA scale, with a one-year time lag between input and output. The parameters identified by the SOMs as important were averaged per DMA per year and used as input parameters; and occurrence of iron exceedance above a certain threshold in DMAs in a year ahead was set as the output parameter. The output parameter consists of two classes of 'exceedance (E)' and 'non-exceedance (N)', referring to whether there was at least one event of iron exceedance above the threshold in the DMA, or not.

The data of years 2009 to 2018 was employed to train the models, and the data of 2019 was used as a validation set to test the performance of the model in predicting unseen data. For assessing the performance of the models and the accuracy of the predictions, True Positive Rate (*TPR*), True Negative Rate (*TNR*), Accuracy (*ACC*), and Matthews Correlation Coefficient (*MCC*) were measured as defined in the following equations.

$$TPR = \frac{TP}{TP + FN} \tag{1}$$

$$TNR = \frac{TN}{TN + FP} \tag{2}$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{4}$$

where *TP*, *TN*, *FP* and *FN* are True Positive, True Negative, False Positive and False Negative, respectively. Positive and Negative denote exceedance and non-exceedance events, respectively. For example, True Negative means actual observational data is a non-exceedance (Negative), and the model predicted it correctly (True). *TPR* represents the probability that the model will correctly predict positive class values (exceedances); *TNR* represents the probability that the model will correctly predict negative class values (non-exceedances); *ACC* is the proportion of samples, positive or negative, predicted correctly; and *MCC* is a measure of summarising performance even when there is a skew in class sizes.

In the following, the ML models are compared (using a threshold of 150 µg/L to define the output parameter, iron exceedance). Then, the model which performs best for the current dataset is used to investigate various effects such as bias in the data (imbalance between exceedance and non-exceedance events), various combinations of input parameters, and exceedance level threshold. Finally, the best configuration is employed to estimate risk of iron exceedance. All the calculations were performed in MATLAB 2019b.

4.1. Comparison of ML models

CRF is an ensemble learning method for classification which combines a collection of decision trees. It improves the prediction by introducing a split on a random subset of features, i.e., it randomly selects observations and variables to build multiple decision trees and then averages all the trees. This makes it robust in dealing with non-linear data with outliers. CSVM uses classification algorithms to find a hyperplane in a multi-dimensional space (due to multiple variables) that distinctly classifies the data points into two groups. It has the advantages of high speed and good performance with datasets where a limited number of observations is available. CBT-RUSBoost is a type of decision trees

suitable for imbalanced datasets. It uses boosting algorithms combined with a random under-sampling method to reduce the size of majority class and thus improve the performance of the model.

The three ML models are compared in terms of training (with the data of 2009 to 2018) and prediction of unseen data (data of 2019), and their performance is summarized in Table 2. CSVM shows very poor performance. It calculated most of the events as non-exceedance (high *TNR* and low *TPR* in training, and *TPR* of zero in validation) due to the highly unbalanced data. Therefore, it is not considered as a suitable method for the present application. RUSBoost and CRF algorithms, although showing better performance than CSVM, still performed poorly, especially in terms of *TPR* and *MCC*. This is also due to the high imbalance in the size of classes of the output parameter (rare exceedance events). The total number of DMA yearly data used in this analysis is 17,507. With an exceedance threshold of 150 µg/L, only 250 of these belong to class 'E'. This indicates that the bias in the data, if defined as M_N/M_E, where M_N = number of non-exceedances, and M_E = number of exceedances, is equal to 69, a situation which cannot be handled by ML algorithms, even with the RUSBoost which is specifically designed for imbalanced datasets.

Table 2. Performance metrics of training of the ML models (with data of 2009 to 2018) and testing it with the validation set (data of 2019). 'NaN' indicates 'Not a Number'.

	Training set			Validation set		
ML model	*TPR*	*TNR*	*MCC*	*TPR*	*TNR*	*MCC*
CSVM	0.127	1.0	0.354	0	1	NaN
RUSBoost	0.697	0.739	0.118	0.440	0.819	0.075
CRF	0.667	0.769	0.123	0.360	0.864	0.073

To overcome the issue, two methods were explored: i) use of synthetic data to improve the balance between exceedance and non-exceedance classes by creating examples of the former; and ii) a random removal of the non-exceedance class events. Due to concerns that the linear extrapolation in the synthetic data generation may be distorting the true underlying patterns in the data it was not considered appropriate; therefore, the random removal of the non-exceedance events was employed for the present analysis.

4.2. Bias in data

The bias was reduced by random down sampling of the majority class (non-exceedances), i.e., reducing M_N. Six tests were performed with M_N/M_E ranging from 69 (original data) to 1 (complete removal of skew).

Tables 3 and 4 present the *TPR* and *MCC* calculated by the CRF and RUSBoost models for training and validation sets, respectively.

Table 3. *TPR* of training and validation of the CRF and RUSBoost models.

		Training set		Validation set	
M_N/M_E	N_t	RUSBoost	CRF	RUSBoost	CRF
69	17507	0.697	0.667	0.440	0.360
10	2783	0.771	0.709	0.423	0.385
6	1771	0.822	0.797	0.412	0.471
4	1265	0.863	0.783	0.630	0.519
2	759	0.667	0.714	0.421	0.421
1	506	0.395	0.649	0.475	0.480

As the imbalance in the data decreases, the size of data (N_t) also decreases, but performance of both models improves. *MCC* constantly increases from M_N/M_E = 69 to 1, especially with CRF, for both

training and validation sets, but *TPR* tends to decrease for $M_N/M_E < 4$. The highest value of *TPR* is at $M_N/M_E = 4$, but *MCC* is poor at this value, particularly for the validation set (i.e., potential for overfitting), while $M_N/M_E = 1$ provides more reliable predictions. In summary, CRF with $M_N/M_E = 1$ performs best and is selected as the best model for the present application.

Table 4. *MCC* of training and validation of the CRF and RUSBoost models.

		Training set		Validation set	
M_N/M_E	N_t	RUSBoost	CRF	RUSBoost	CRF
69	17507	0.118	0.123	0.075	0.073
10	2783	0.359	0.315	0.166	0.137
6	1771	0.408	0.354	0.168	0.132
4	1265	0.453	0.389	0.255	0.191
2	759	0.522	0.455	0.261	0.192
1	506	0.422	0.483	0.359	0.367

4.3. Input parameter combinations

While the number of possible input parameters was reduced by the SOM, there are still a great number of combinations of these parameters that can be used for the risk model. The parameters identified by SOMs were manganese, turbidity, 3-day HPC, low chlorine, residual type, number of customer contacts, DMA clustered customer contacts, and high priority dead ends per DMA. The best model from the last section (CRF with $M_N/M_E = 1$) was employed to test twenty combinations out of these parameters in order to find the best one which gives the highest accuracy of predictions. For brevity, the details of the tests are not presented. It was found that using four of these parameters as input, iron, turbidity, 3-day HPC and number of high priority dead ends, provided the highest accuracy. The performance of the model using these parameters is presented in Table 5.

Table 5. CRF model performance with best combination of input parameters (iron, turbidity, 3-day HPC, and high priority dead ends per DMA).

Metric	Training set	Validation set
TPR	0.612	0.609
TNR	0.874	0.857
ACC	0.739	0.727
MCC	0.502	0.478

4.4. Exceedance level

Using the best input parameter combination, iron exceedance threshold was explored to investigate its effect on the performance of the model. Four levels were tested: 200, 150, 100 and 50 µg/L. The details of the tests and the performance metrics calculated for training and validation sets are presented in Table 6.

For higher iron exceedance thresholds, the size of data is smaller due to rarer exceedance events; however, model fitting to the data is more accurate. For example, when the threshold is set to 200 µg/L, *MCC* of training is 0.704, while for a threshold of 50 µg/L, it is 0.326. For the validation set, the accuracy is highest for the 150 µg/L threshold. However, the misclassification error is also higher. Figure 2 shows the CRF misclassification error calculated for these tests. For exceedance level of 200 µg/L, the error is lower, denoting that more samples contributed to model training, thus the result is more reliable. This analysis shows that the model, including parameters combination, is well suited to identifying the higher levels of iron risk.

Table 6. CRF model performance for different levels of iron exceedance

		Training set			Validation set		
Exceedance level (µg/L)	N_t	*TPR*	*TNR*	*MCC*	*TPR*	*TNR*	*MCC*
200	254	0.866	0.838	0.704	0.500	0.833	0.354
150	485	0.612	0.874	0.502	0.609	0.857	0.478
100	984	0.571	0.797	0.379	0.436	0.906	0.357
50	3202	0.680	0.646	0.326	0.496	0.748	0.252

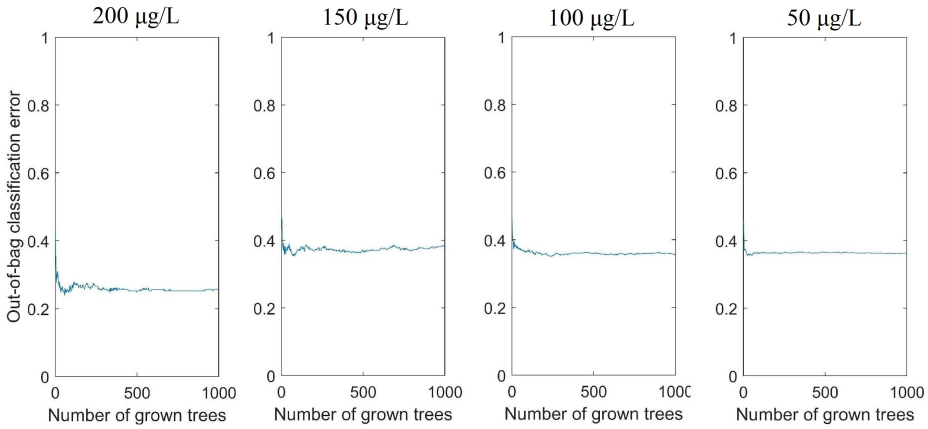

Figure 2. CRF out-of-bag classification error for exceedance levels of 200, 150, 100 and 50 µg/L, from left to right, respectively.

4.5. Performance of the preferred model

Summarising the results of the tests, the CRF model with DMA yearly averaged iron, turbidity, 3-day HPC and number of high priority dead ends as input, and with a balanced data (i.e., $M_N/M_E = 1$) gives the highest accuracy of the prediction of iron exceedance in a year ahead. For an exceedance level of 200 µg/L (the regulatory threshold), the *TPR* and *TNR* of predictions are 0.866, 0.838 for the training set, and 0.5 and 0.833 for the validation set. Therefore, this setup is selected as the preferred configuration for the current dataset and used for the calculation of probability of iron exceedance.

Figure 3 presents the confusion matrices of the model training and validation with the best setup. These show how many of the 'E' and 'N' classes were predicted correctly, or incorrectly. For example, for the validation set, half of the exceedances and most of the non-exceedances (10 out of 12) were predicted correctly.

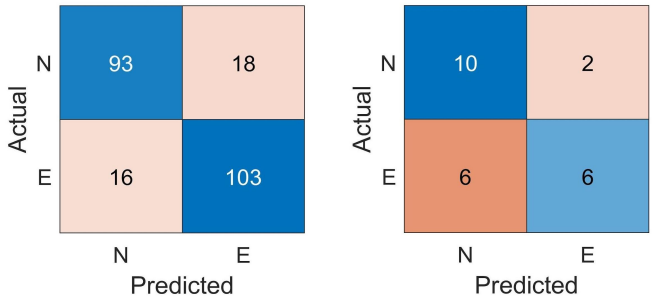

Figure 3. Confusion matrices of model training (left) and validation (right) for the preferred setup.

Figure 4 shows the predictor importance estimates calculated by the CRF model for the best setup. It indicates that the two most important input parameters in the analysis were high priority dead ends and iron, respectively. It means these two had the highest correlation with the output parameter, which is a year ahead iron exceedance, while turbidity and 3-day HPC had smaller contributions.

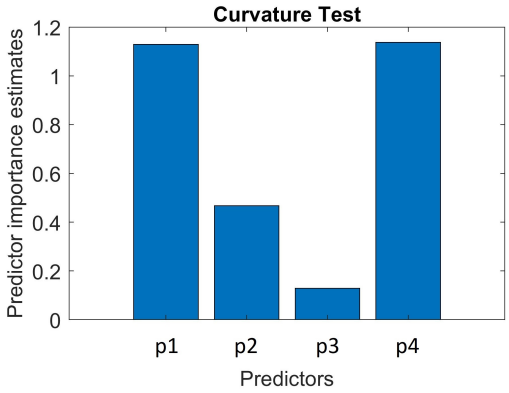

Figure 4. CRF predictor importance estimates for best setup. Input parameters, p1 to p4, are DMA yearly average of iron, turbidity, three-day HPC, and high priority dead ends per DMA, respectively.

4.6. Risk of iron exceedance

CRF grows a collection of individual decision trees. Each tree produces a class prediction ("votes" for a class). Then the forest chooses the class having the most votes over all the individual trees. There are two classes in our application, 'E' and 'N', denoting exceedance and non-exceedance events. Thus, if the fraction of trees that vote for an event to be in the class 'E' is above 0.5, that event is considered as an exceedance event, and if it is below 0.5, the event is a non-exceedance. Probability of a class is then obtained by counting the fraction of trees in the forest that voted for that certain class. The probability calculated for the class 'E' is thus inferred as the probability of iron exceedance, which represents the likelihood an event falls in the class exceedance.

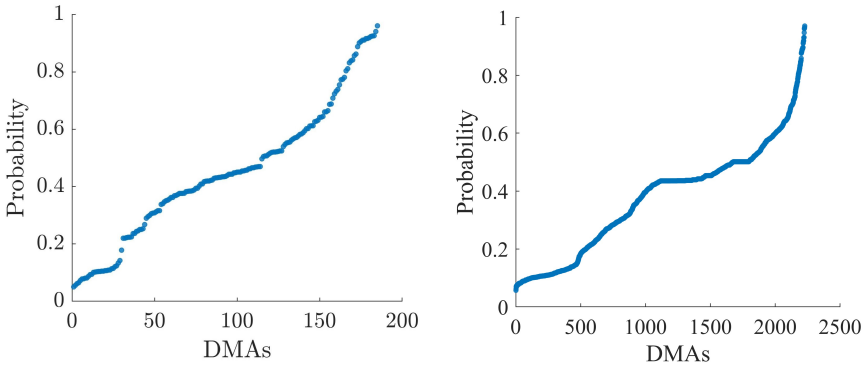

Figure 5. Prediction of relative iron exceedance probability above 200 µg/L of DMAs for the training set (left) and all the DMAs in 2020 (right), sorted by probability values.

The CRF model with the best setup was used to calculate probability of iron exceedance above 200 µg/L for the DMAs in the training set and the result is presented in Figure 5-left, where the DMAs are sorted by the calculated values of probability. The model was then employed to predict exceedance probability for all the DMAs in 2020. For this purpose, data from 2019 was fed into the model and the probability of iron exceedance in 2020 was estimated. The result of this test is presented in Figure 5-right, where DMAs are sorted ascendingly by probability. The graph exhibits a 'kick' upwards at

higher probabilities. The vast majority (90% and over) of DMAs have probabilities of exceedance < 0.5, while < 10% have values above 0.7~0.8. This suggests that targeting interventions and resources to these DMAs should provide the greatest returns for managing iron exceedance risk.

It is noted that the calculated probability is a relative measure of iron exceedance risk in a DMA compared to all the other DMAs in the analysis, rather than being taken as an absolute value of risk. As a result, it is suitable for the purpose of risk ranking and informing proactive management.

5. Conclusions

Machine learning was used to analyse a large dataset of water quality samples, discolouration customer contacts and static asset data from a drinking water distribution network to understand the causes of high iron concentrations at a sub-region (DMA) level, and to thus develop a risk model giving predictions of relative probability of iron exceedance above a certain threshold.

A preferred setup was identified in terms of accuracy and reliability for risk modelling across the network. The setup included a CRF model trained with yearly DMA averaged iron, turbidity, 3-day HPC, and high priority dead ends per DMA as input parameters, with output a year ahead prediction of relative probability of iron exceedance above 200 µg/L.

The trained model was used to estimate and rank likelihood of iron exceedance for all DMAs in year 2020. This ranking showed that less than 10% of DMAs pose a risk compared to the entire system, suggesting that targeting these for interventions should provide the greatest return for managing iron exceedance risk.

The developed model can inform proactive management and is easily applicable to investigate other discoloration parameters such as turbidity, and transferable to other datasets or regions.

Acknowledgments

For the purpose of open access, the author has applied a creative commons attribution (CC BY) license to any author accepted manuscript versions arising.

References

[1] Vreeburg J and Boxall J 2007 Discolouration in potable water distribution systems: A review *Water Research* **41**(3) 519–529 https://doi.org/10.1016/j.watres.2006.09.028

[2] Speight V, Mounce S and Boxall JB 2019 Identification of the causes of drinking water discolouration from machine learning analysis of historical datasets *Environ. Sci. Water Res. Technol*. **5** 747–55 https://doi.org/10.1039/C8EW00733K

[3] Mounce SR, Ellis K, Edwards JM, Speight VL, Jakomis N and Boxall JB 2017 Ensemble decision tree models using RUSBoost for estimating risk of iron failure in drinking water distribution systems *Water Resour. Manag.* **31**(5) 1575–89 https://doi.org/10.1007/s11269-017-1595-8

[4] Kohonen T 1990 The Self-Organizing Map *Proc. IEEE* **78**(9) 1464–80 https://doi.org/10.1109/5.58325

[5] Breiman L 2001 Random forests *Machine Learning* **45** 5–32 https://doi.org/10.1023/A:1010933404324

[6] Hastie T, Tibshirani R and Friedman J 2008 *The Elements of Statistical Learning* 2nd edition (New York: Springer)

[7] Seiffert C, Khoshgoftaar T, Hulse J and Napolitano A 2008 RUSBoost: Improving classification performance when training data is skewed *Proc. 19th Int. Conf. Pattern Recognition* 1–4.