White Rose Research Online

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk
Universities of Leeds, Sheffield and York

Deposited via The University of Sheffield.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/id/eprint/187741/

Version: Accepted Version

Article:
Boxall, J., Speight, V., Kyritsakas, G. et al. (2022) The application of artificial intelligence techniques to better manage iron in drinking water distribution systems. Institute of Water Journal (7). pp. 28-34.

© 2022 The Authors. This is an author-produced version of a paper subsequently published in the Institute of Water Journal. This version is available under the terms of the Creative Commons Attribution Licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Reuse
This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here: https://creativecommons.org/licenses/

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

UNIVERSITY OF LEEDS
University of Sheffield
UNIVERSITY of York

The application of artificial intelligence techniques to better manage iron in drinking water distribution systems

Joby Boxall[1], Katrina Flavell[2], Vanessa Speight[1], Grigorios Kyritsakas[1], Ehsan Kazemi[1], Stewart Husband[1], Samuel Bright[2], Sam Ledgar[2], Luke Montgomery[2]

[1] Department of Civil and Structural Engineering, University of Sheffield, Sheffield. UK

[2] Yorkshire Water Services, Bradford. UK

Abstract

The concentration of iron in drinking water at the customer's tap is a key measure of both water quality and network performance. The reasons and factors that contribute to this are complex and interact. It is important that water companies understand these so that constrained resources can be best targeted to minimise the risks of non-compliance and avoid associated financial and reputational penalties. Understanding of iron behaviour in water distribution networks is reported and discussed here based on the application of artificial intelligence techniques. This enabled the mining of over a decade of companywide, messy and incomplete data, linking across different data types: discrete sample data, asset data and customer contacts. The qualitative understanding derived, such as lack of influence of local pipe material and association with heterotrophic plate count data, were then captured in numerical predictions. A form of ensemble decision trees was found to provide robust results, giving 'S' curve rankings at district metered area (DMA) level that clearly highlighted areas of highest risk. These rankings have been used to target flushing interventions.

Keywords

Drinking water quality, Iron compliance, Discolouration, Big data, Machine learning, Risk ranking

1. Introduction

The UK regulatory standard for iron in potable water is 200 μgL^{-1}. This level is set for aesthetic rather than health reasons, with elevated levels of iron associated with customer contacts for discoloured water. Like many UK water companies, Yorkshire Water Services (YWS) undertakes a targeted programme of mains flushing to address both iron compliance and discoloured water. This is important to deliver good service and ensure that YWS meets the specific Performance Commitments relating to elevated levels of iron: Water Quality Compliance (CRI) and Drinking Water Contacts (Ofwat, 2019). A prioritisation process identifying areas of highest risk is needed to ensure that flushing programmes direct resources efficiently and deliver the best return for investment. The work presented here delivers a development to the current prioritisation process used within YWS.

2. Background

2.1 Iron behaviour within drinking water distribution networks

The source of iron within drinking water distribution systems can be carry over, generally at low concentrations, from water treatment works and/or corrosion of iron fittings and fixtures within drinking water distribution systems. The subsequent behaviour within the network is complex with many factors known to contribute and be influential, in particular where and how iron accumulates and where and when iron may be mobilised or released. The chemical dominated processes of corrosion are important, as highlighted by Sarin et. al. (2004) linking water quality degradation due to iron release, in soluble or particulate form, from corroding pipes to the bulk water. Seth et. al. (2003) showed iron together with manganese to be the dominant metal constituent of discoloration flushing samples, irrespective of pipe material. Sly et. al. (1990) also noted the co-deposition of iron and manganese, observing it to be microbial mediated. The physical conditions of hydraulic pipe operation can also be important. Beckett et. al. (1998) found stagnant zones to influence higher metal concentrations, although Li et. al. (2020) found that no flow conditions limited corrosion processes, associating this with a lack of supply of dissolved oxygen. This brief review highlights that the processes leading to risks of elevated iron concentration are complex and cannot be readily derived by simple analysis or interpretation.

2.2 Digital water

Digital water, the exploitation of digital technologies to provide and exploit data, improve business processes, create markets and entirely new products and services, has been widely recognised as an important opportunity for the water sector, such as IWA (2019). An important subset of this is the application of artificial intelligence (AI) or machine learning (ML) techniques to extract or mine understanding from the data historically and currently collected by water companies, but which is often used for siloed, specific purposes.

2.2.1 Self-Organising Maps

Self-Organising Maps (SOM) are a type of Artificial Neural Network (ANN) that is trained using unsupervised learning; this means that when the inputs are presented to the ANN, it forms its own clustering of the training data thus allowing the potential to derive information from data without any previous knowledge. This feature eliminates the need to specify relationships prior to the data analysis, a critical attribute for the analysis and data explored here. A further important attribute of SOM is their resilience to missing and incomplete data, important when considering the data available for this analysis. The Kohonen self-organising feature map (Kohonen 1990), generally referred to as the SOM, is an ANN model which resembles the way biological brain maps spatially order their responses by modelling the self-organising and adaptive learning features of the brain. SOM enables the visualisation of high-dimensional input data in a low (usually two) dimensional output space. The outputs are qualitative, but provide evidence-based understanding of the interactions and influences of different parameters on a given outcome, in this case elevated iron concentrations. SOM have been applied to a variety of disciplines, including economics, genetics, climatology, engineering, and water applications such as resources (Kaltech et al 2008) and drinking water quality (Speight et al 2019).

2.2.2 Risk ranking

Supervised ML techniques can map input parameters to an output parameter, such as elevated iron concentrations. In a supervised learning method, in contrast to unsupervised learning approaches like SOM, input and output parameters are defined, thus when the

model is trained, using a set of input data, it will be possible to 'predict' the corresponding output parameter. Different ML algorithms were explored here, Classification-based Random Forests (CRF) (Breiman, 2001), Support Vector Machines (Hastie et al., 2008), and Boosted Tree learning algorithms based on Random Undersampling Boosting (RUSBoost) (Seiffert et al., 2008). Mounce et al (2017) previously employed boosted trees effectively to categorise risk of iron failure, finding that *"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"*.

3. Method

The rare event nature of elevated iron concentrations (in 2019, just 0.53% of regulatory iron samples collected by YWS exceeded the compliance standard) is challenging for understanding the causes and for estimating future risk (probability). This rare nature is compounded by the vast size, complexity and age of drinking water distribution system infrastructure. While very significant time and effort are invested in sampling to confirm water quality at the customer's tap, the discrete sample data derived gives only single point values in time and space that underrepresents the entire system.

The rare event nature of elevated iron concentrations means that it is essential that the method adopted takes advantage of all possibly relevant data, and is at significant spatial and temporal scale. The analysis was hence conducted at company wide scale and for data collected over a decade. The approach consisted of two phases as set out in Figure 1. The first is termed knowledge discovery, although it might also be called prior expectation exploration. In this phase SOM were used to explore the complex, messy data set, testing prior assumptions and looking for multi-parameter correlations. The second phase was to provide insight into which district metered areas (DMAs) have the highest risk (relative probability) of elevated iron concentrations. Different methods were assessed for provision of DMA-scale, year-ahead prediction of elevated iron concentrations and ultimately DMA risk ranking based on relative probabilities of this.

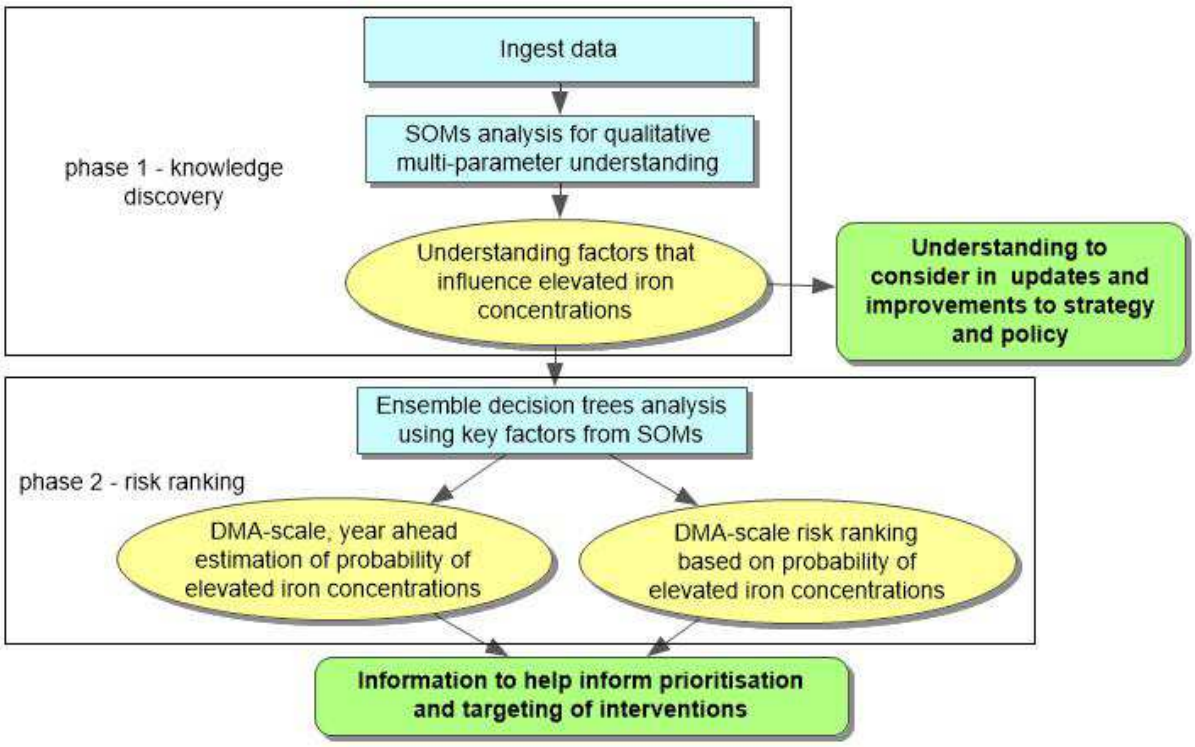

Figure 1. Schematic representation of approach

3.1 Data preparation

Various data that may pertain to elevated iron concentration are collected by water companies, for different purposes and functions. The data is typically stored in different systems and ranges from static asset to discrete water sample to connectivity data. Prior to attempting knowledge discovery, this data must be checked, linked or associated and made accessible. This was done making use of unique YWS-specific reference numbers and geospatial searches, such as associating tap samples to their nearest water distribution main but ignoring pipes not flagged as live, and to associate pipes and samples to DMAs. This required a time-consuming understanding of the data structure and systems used to ensure that the data are correctly associated. Once done and checked it provides a source that can be explored for a range of different questions and challenges, but only elevated iron concentrations are reported here. Following data preparation there were 134,803 records for regulatory samples at customer taps, 816,703 pipe records and 62,695 discolouration customer contacts, representing a decade of pre-covid data.

3.1 Knowledge discovering using self-organising maps

SOM were used for the identification of potential multi-parameter correlations among water quality, pipe and customer contact data. SOM were generated using Matlab2019b and the SOM Toolbox2.1 (Kohonen, 2014). There are a vast number of permutations and combinations of parameters and subdivisions of the data sets that can be explored. The act and process of posing questions, investigating these with the SOM and discussing and interpreting the results is a rewarding and beneficial process. This was undertaken by an expert group across practitioners and academia. Only a snapshot of the SOM produced is included here to capture some of the deeper understanding gained as well as key parameters that were taken forwards to the predictive ranking.

3.1.1 Analogous description of SOM

Fill a football field with people, ask the people to question each other about some selected parameters (hair colour, height, job, whatever is of interest) and to form groups (clusters) of similar attributes (values) across the parameters based on the answers. Give a long time to question and form groups. Now give each group a set of coloured cards and ask them to hold up the colour that represents their group in response to a question about the parameters. This is, sort of, a human SOM. Importantly computers don't get bored like people and will keep asking the same questions to form groups, so computer groupings should be more reliable.

3.2 Risk ranking

The second phase was to provide quantitative insight into which DMAs have the highest risk (relative probability) of elevated iron concentrations. Candidate input parameters were identified by SOM with the output a categorical parameter consisting of two classes of 'elevated (E)' and 'non-elevated (N)' based on a threshold (150 and 200 μgL^{-1} were both tested). If there was at least one iron concentration elevated above the threshold in a DMA in a year, then the DMA-year sample was classified as elevated 'E'. Models were trained with prior data and then employed to 'predict' elevated iron concentration at individual DMAs in a certain year and calculate the probability of the predicted elevated concentration.

The elevated iron concentration data is highly biased; the number of elevated iron concentrations is rare compared to non-elevated. The total number of DMA yearly iron data used in this analysis was 21,307. When a threshold of 200 μgL^{-1} is applied, only 154 of the

samples belong to class ‘E’, a non-elevated to elevated ratio of 137. This ratio is 69 when a threshold of 150 μgL^{-1} is used. The machine learning models, even the RUSBoost algorithm, which is specifically designed for imbalanced data, were not able to handle this imbalance. Therefore, two methods were explored to address this; generation of synthetic elevated concentrations created using the Adaptive Synthetic Model (ADASYN) in MATLAB 2021a, and random down sampling (random removal of data points) of the non-elevated data.

Modelling, using 90% data to train and 10% to test, then comparing to a year ahead to assess accuracy, showed promise. Overall accuracy of 0.9 was achieved with the generation of synthetic data. However, there was concern that the synthetically created data might have distorted the true underlying relationships within the data, as it is based on linear interpolation between existing minority classes. Therefore, random down sampling of the non-elevated data with different levels of bias reduction (‘E’ to ‘N’ ratios of 1/69 to 1/1) were explored with the different machine learning algorithms. It was found that the CRF model showed the best performance (accuracy ~ 0.7); and that the ‘E’ to ‘N’ ratios below 1/5 provided an acceptable balance.

While the number of possible input parameters was reduced by the SOM, the number of combinations and permutations possible was still vast. A range and variety of these were explored, informed by the expert user group. Overall performance assessment is complex, with simple ‘goodness of fit’ insufficient to capture, for example, if a model that predicts all real events but also a high number of false events is better than a model that has low false positives but misses some real events. Similarly, it is informative to know if increasing model complexity (i.e. one that uses more parameter or has more branches – decisions – in the tree) is yielding sufficiently improved predictions for the additional complexity. A number of different measures capturing such aspects where used to enable selection of the overall preferred model. These included confusion matrices (capturing the performance in terms of true and false positives and true and false negatives), observation of the reduction in out-of-bag classification error as a function of number of trees grown (capturing model complexity), and quantification of predictor importance estimation via curvature tests (informing on the value obtained from the number and combinations of parameters used).

The selected model, trained on all prior years, was employed to predict relative probability of elevated iron concentration of individual DMAs in a year using the DMA yearly averaged quantities in the prior year. The predicted probability was used to rank DMAs from worst to best based on their likelihood of elevated iron concentrations.

4. Results

4.1 Knowledge discovery

The SOM produced are heavily dependent on the parameters included and any subdivisions of the data (for example chlorine versus chloramine residual was explored). Credence should be given to correlations and associations that manifest across multiple SOM. The SOM presented here provide examples of correlations that were robustly seen across the many analyses performed. The colour ranges of each component plane of the SOM have been set to cover 5 to 95% of the data range for each parameter in each SOM, they are not consistent between figures. Non-numerical parameters such as pipe material were considered by post labelling the SOM, i.e. had no part in deriving the clusters formed. The U-matrix plane gives a measure of the strength and difference of the clusters formed.

Figure 2 shows an example of what became something of a ‘base-line’ SOM. It shows strong correlation between iron, manganese and turbidity down the right-hand side of the component planes. This confirms the complementary chemical behaviour of iron and manganese, as well as their easier assessment via measurement of turbidity. The figure also shows strong inverse correlation between chlorine and heterotrophic plate counts data (‘Colonies 3 days 22 C’), top to bottom of the component planes. Of note is that the higher concentrations of iron (and manganese and turbidity) are towards the bottom right, suggesting that increased biological activity in some instances increases iron risk. SOM segregated by disinfection type, free chlorine or chloramine, showed this correlation more strongly for the chlorine data (SOM not shown). Two clusters of high iron became visible, one associated with manganese and hence suggesting chemical processes and the other with elevated heterotrophic plate counts, suggesting more biologically mediated processes. Although other clusters of elevated plate counts were also evident such that increased biological activity is not necessarily always an indicator of increased iron concentrations. Contrary to prior expectations, post labelling the SOM by pipe material, for the pipe nearest to the sample, did not yield a strong correlation. While there are perhaps visually more green cells (indicating cast iron) to the right-hand side, this is tentative and more a product of greater red cells (indicating no dominant pipe material) to the centre and left.

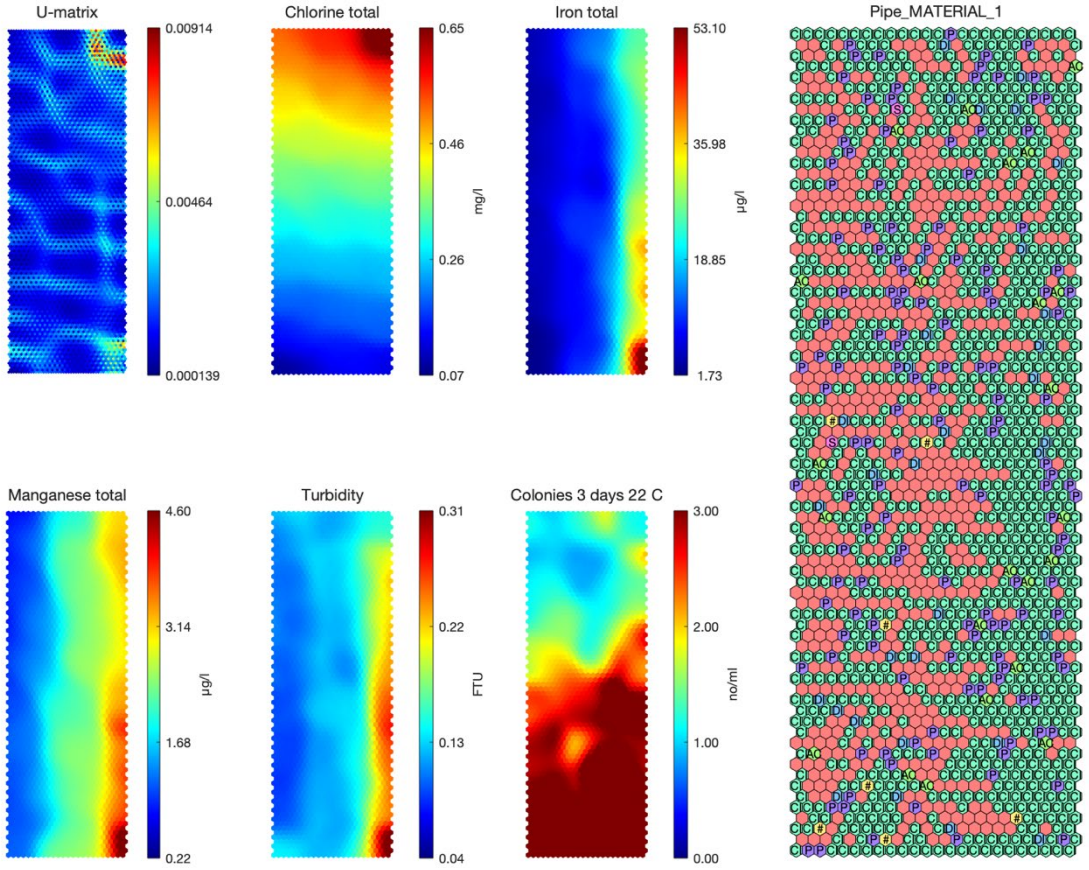

Figure 2. SOM using key regulatory samples, post labelled with water mains material

(Green: Cast Iron. Purple: Plastic. Cyan: Ductile Iron. Dark Green: Asbestos cement. Red: No dominant pipe material. Yellow: no connection)

The role of pipe material was further investigated. Figure 3 shows an example of this, including pipe material as a numerical value such that it does influence the clustering. This was done by assigning the percentage of plastic and iron pipes within each DMA to each

sample. The final two component planes show the expected inverse relationship of these two parameters. Clustering is lost in all other parameters, suggesting that the percentage of pipe material does not correlate with risk of elevated iron concentration, or with chlorine decay or bacterial activity. This lack of importance of pipe material was further explored and confirmed in several other SOMs.

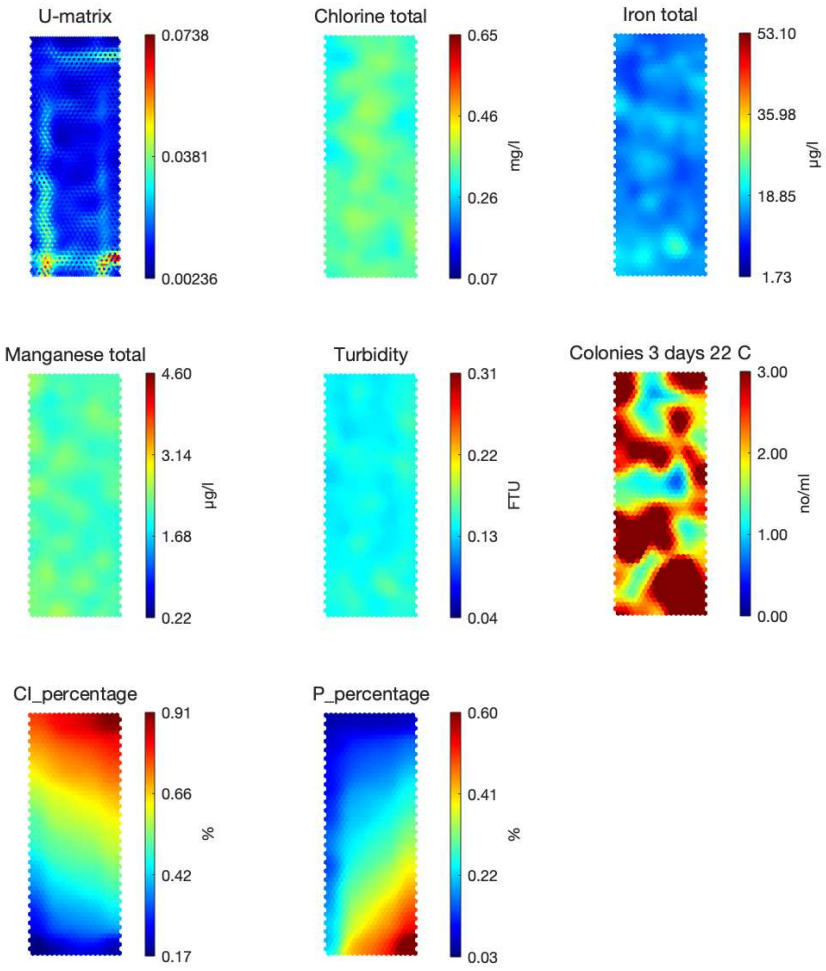

Figure 3. SOM using key regulatory sample data and pipe material as a percentage per DMA (Cl: cast iron. P: plastic)

Figure 4 shows how a company specific measure was usefully incorporated into the analysis. YWS have an internal system based on expert judgement that enables classification and count of the number of 'high priority dead-ends' in a DMA. From Figure 4 it can be seen that these high priority dead-ends cluster at the bottom right of this SOM, correlating with medium to high clusters of iron. While the usual corresponding clusters of turbidity and manganese are present, they are less clearly defined. Surprisingly there is no correlation or even pattern to the chlorine or heterotrophic plate count data, which might have been expected due to the often low turnover and high residence times associated with dead ends. This is perhaps due to the unique company specific measure that define high priority as opposed to dead-ends more generally. From Figure 4 it can be seen that high numbers of dead-ends in general cluster to the bottom left and do not correlate with iron, or anything else.

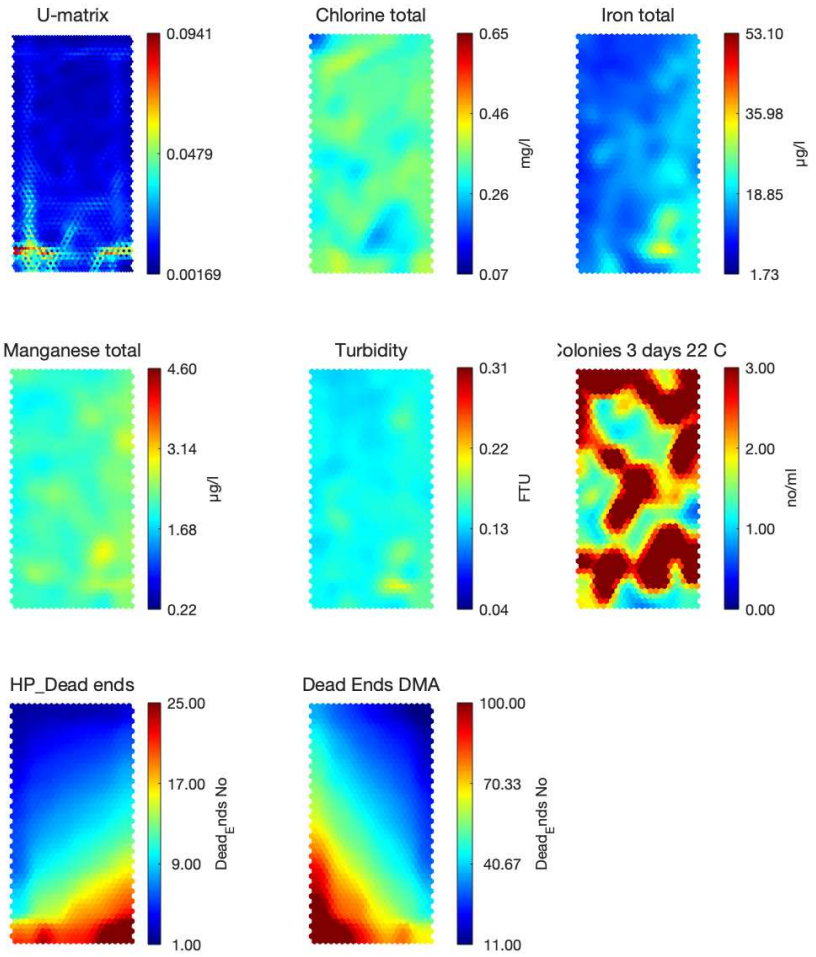

Figure 4. SOM using key regulatory samples, and dead-ends data per DMA

4.2 Ranking

Predictive modelling was conducted with combinations of up to 10 different input variables identified from the SOM analysis. Ultimately a model that used only iron, turbidity, 3-day heterotrophic plate counts and high priority dead-ends was selected as the preferred model as it gave only a small reduction in overall performance compared to more complex models. Four applications of the best model were compared to confirm robustness; predictions for two different latest years and for analysis at elevated iron concentrations of 150 and 200 μgL^{-1}. Results across these, such as 24 out of 30 (top 1%) highest ranked DMAs being common, were deemed to indicate robust results.

Figure 5 shows spatially the results of DMA ranking based on the predicted relative probabilities of elevated iron concentration. It does not show any spatial correlation or clustering, suggesting that there is limited influence of source or treated water quality. It also highlights that risk-ranking at water supply zone level is not possible. It is not possible to predict relative probability of elevated iron concentration for DMAs with no measured data in the prior year, so these are blank (white) in Figure 5. The missing data was always iron; the model would be improved if at least one iron sample is collected from each DMA each year in future.

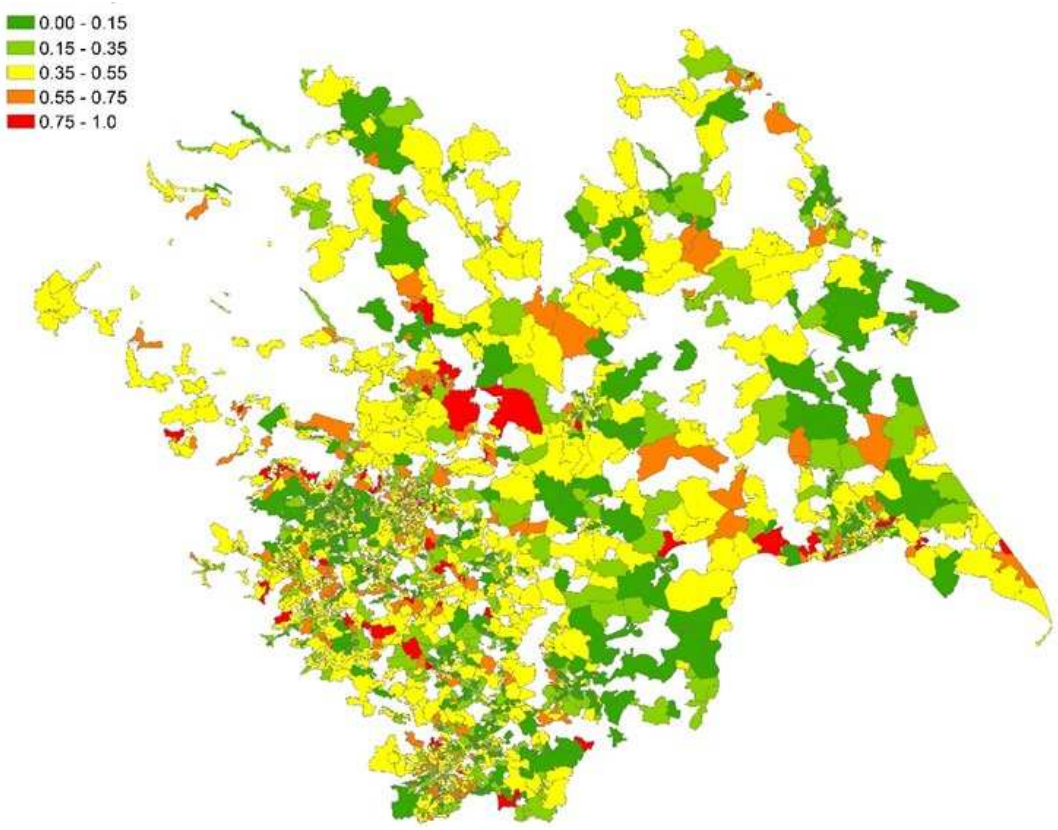

Figure 5. Map of relative probability prediction rankings for elevated iron concentration of 200 μgL^{-1}. Note DMAs are white where there was no iron data for the previous year

Figure 6 shows the same predicted relative probabilities of elevated iron concentration as figure 5, but rank ordered. It shows that over 90% of DMAs were found to have a relative probability of elevated iron concentration of less than 50%. The steep climb in probability of elevated iron concentration within the last few percentage of the DMAs suggests that targeting interventions such as flushing to the top few percent with higher predicted probabilities should be effective at managing risk of elevated iron concentration.

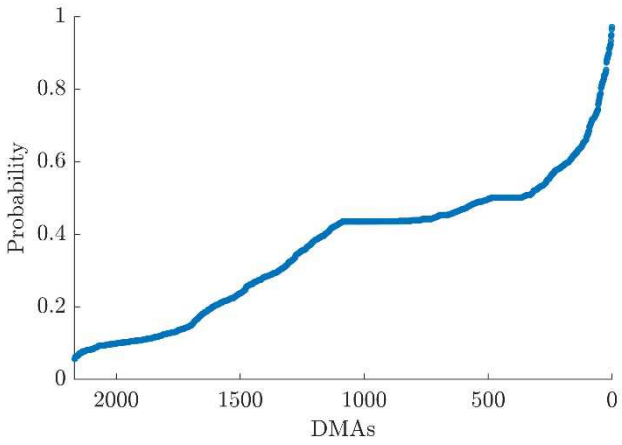

Figure 6. Relative elevated iron concentration probability ranking for 200 μgL^{-1}

5. Discussion

The SOM confirm the importance of chemical processes for iron behaviour within drinking water distribution systems, in particular the association with manganese (seen via the relationship between manganese and turbidity). There is also some evidence of iron correlating with biological activity. This is likely through the role of biofilms, following the growing association of biofilms with discolouration processes of material accumulation and subsequent mobilisation by hydraulic events, Husband et. al. (2016). The partial nature of the correlation with biological activity is consistent with this, the discrete samples that the data are derived from will capture the bulk water organisms, not those attached as biofilm. Additionally, they will only capture that which can be cultured by heterotrophic methods.

The lack of correlation between iron and pipe material found in the SOM analysis is unexpected. The lack of association with the pipe closest to the sample could be explained by the fact that discrete samples under normal operating conditions are more likely to capture cumulative effects between water treatment works to the sampling point. The lack of correlation with DMA percentages of pipe material is not due to lack of variation in mix of pipe materials between DMA, there are DMAs with ~15 to 90% cast iron and ~5 to 60% plastic visible in the SOM (Figure 3). An interpretation of this could be that while corrosion of iron pipes and fittings is widely recognised as a source of material, this is perhaps not the limiting factor in elevated iron concentration when assessed at DMA-level.

The correlation with high-priority dead-ends was a valuable finding from the SOM knowledge discovery, identifying a parameter that was important in the predictive modelling. It is perhaps surprising that it was the YWS-specific and proprietary classification of high priority that was important, the simpler count of all dead-ends in a DMA was not correlated. A SOM that explored diameter and all dead-ends was tried, but this was also insufficient to reveal any correlations. This shows a strength of this type of analysis, readily enabling the incorporation and exploration of different data and ideas.

The analysis conducted had the defined aim of ranking at DMA scale. Even so some of the understanding explored was to pipe level, such as lack of association with the nearest pipe material and association with the high priority dead-ends. It is likely that if further pipe-specific data could be incorporated deeper insight and finer resolution of ranking and hence prioritisation would be possible. A likely valuable source of information would be from hydraulic models, providing information such as peak velocities and water age that impacts water quality behaviour, Machell and Boxall (2014).

Most of the performance of the predictive modelling could be obtained with only four input parameters. It was notable, but perhaps unsurprising, that previous iron data was key amongst these. Comparison to a model that was based on only iron and turbidity data provided significantly reduced performance and different rankings, confirming that the other parameters were providing value. This was also confirmed by the curvature values, which rated high priority dead ends as most important, closely followed by iron, then turbidity and then a drop to 3-day heterotrophic plate counts having a small contribution. Interpreting across the results it appears the predictive contribution of manganese was covered by the turbidity data. This should not be confused as meaning that the co-behaviour of iron and manganese is not important, rather that this effect can be captured from the turbidity data.

The results from the predictive modelling have been incorporated into YWS's prioritisation of DMAs for routine flushing. This will help ensure that the programme is efficient by ensuring that high-risk DMAs are targeted. Of particular value is the finding that less than 10% of DMAs were found to have a relative probability of elevated iron concentration greater than

50%. This is a relatively small number of DMAs, and the manageable scale of this means that the worst-performing DMAs can receive the greatest level of focus for both routine flushing, and potentially other future capital interventions.

6. Conclusions

The research reported here shows the potential to derive value from data collected by water companies. This was achieved by breaking the silos of data storage, creating associations and linkages between data architectures. Once linked, tools such as self-organising maps (SOM) can be applied to explore the multidimensional dataset created. While commonly termed knowledge discovery, this is perhaps better described as knowledge exploration. Testing different ideas and concepts to see if these are supported or not by multi-parameter correlations between the various data. Key observations here included the expected, such as strong association between iron, manganese and turbidity, and the unexpected such as the lack of association with the material of the pipe closest to the sampling point. The understanding gained through SOM is qualitative, and requires expert input to guide the processes. This can then be complemented by tools such as decision trees, a quantitative 'white box' process. Here ensemble classification trees were used to provide year ahead predictions of probability of elevated iron concentration at DMA-scale. These were used to provide a relative ranking. These showed no spatial correlation, but did categorise the very highest probabilities as associated with only the top 1 or 2% of DMAs. This suggests that targeting these for interventions should reduce the probability component of risk of elevated iron concentrations.

Acknowledgements

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

Data used in this research is not available due to its confidential and GDPR sensitive nature.

References

Beckett, M.A., V.L., S., Jim, K., Sarin, P., Kriven, W.M., Lytle, D. and Clement, J.A. A. (1998). Pipe Loop System for Evaluating Iron Uptake in Distribution Systems. Proc. AWWA Water Quality Technology Conference, Paper 5C-4. San Diego, CA.

Breiman, L. (2001). Random Forests. Machine Learning 45, pp. 5–32.

Hastie, T., Tibshirani, R. (2008) and J. Friedman. The Elements of Statistical Learning, second edition. New York: Springer.

Husband, S., Douterelo, I., Fish, K.E. and Boxall J.B. (2016) 'Linking discolouration modelling and biofilm behaviour within drinking water distribution systems' IWA Water Science and Technology: Water Supply Vol. 16 No. 4 pp. 942-950 DOI: 10.2166/ws.2016.045

IWA (2019) Digital Water :Industry leaders chart the transformation journey. [https://iwa-network.org/publications/digital-water/ accessed 7th Macrh 2022]

Kalteh, A. M. Hjorth P. and Berndtsson R. (2008) Review of the self-organizing map (SOM) approach in water resources: Analysis, modelling and application, Environ. Model. Softw., 2008, 23(7), 835–845.

Kohonen, T. 1990. 'The Self-Organizing Map'. *Proceedings of the IEEE* 78 (9): 1464–80. https://doi.org/10.1109/5.58325.

Kohonen, T. 2014. *MATLAB Implementations and Applications of the Self-Organizing Map*. http://docs.unigrafia.fi/publications/kohonen_teuvo/.

Li, M., Wang, Y., Liu, Z., Sha, Y., Korshin, G.V., and Chen, Y. (2020) 'Metal-release potential from iron corrosion scales under stagnant and active flow, and varying water quality conditions' Water Research, Vol. 175, 115675, doi.org/10.1016/j.watres.2020.115675.

Machell, J.M. and Boxall, J.B. (2014). 'Modelling and Field Work to Investigate the Relationship between the Age and the Quality of Drinking Water at Customer's Taps' J. Water Resources. Planning and Management, Vol. 140 (9) pp. 1943-5452. DOI: 10.1061/(ASCE)WR.1943-5452.0000383

Mounce SR, Ellis K, Edwards JM, Speight VL, Jakomis N, Boxall JB. (2017) Ensemble Decision Tree Models Using RUSBoost for Estimating Risk of Iron Failure in Drinking Water Distribution Systems. Water Resour Manag. 2017;31(5):1575–89.

Ofwat (2019) "PR19 Final determinations, Yorkshure Water – Outcome performance commitment appendix" https://www.ofwat.gov.uk/wp-content/uploads/2019/12/PR19-final-determinations-Yorkshire-Water-%E2%80%93-Outcomes-performance-commitment-appendix.pdf [accessed 23rd April 2022]

Sarin, P., Snoeyink, V. L., Lytle, D. A., and Kriven, W. M. (2004). "Iron corrosion scales: Model for scale growth, iron release, and coloured water formation." J. Environ. Eng., 130(4), 364–373.

Seiffert, C., Khoshgoftaar, T., Hulse, J., Napolitano, A. (2008). RUSBoost: Improving classification performance when training data is skewed. 19th International Conference on Pattern Recognition, pp. 1–4.

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

Sly L.I., Hodgkinson M.C. and Arunpairojana V. (1990). Deposition of manganese in a drinking water distribution system. In: Applied and Environmental Microbiology, Vol. 56, No. 3 pp 628-639.

Speight V, Mounce S, Boxall JB. (2019) Identification of the Causes of Drinking Water Discolouration from Machine Learning Analysis of Historical Datasets. Environ Sci Water Res Technol. 2019;5.4:747–55.