engineering proceedings

MDPI

Proceeding Paper

A Data-Driven Analysis for Understanding and Risk Estimation of Discolouration in Drinking Water Distribution Systems [†]

Grigorios Kyritsakas [1,2,*], Stewart Husband [1], Killian Gleeson [1], Katrina Flavell [3] and Joby Boxall [1]

[1] Sheffield Water Centre, University of Sheffield, Sheffield S1 3JD, UK; s.husband@sheffield.ac.uk (S.H.); k.gleeson@sheffield.ac.uk (K.G.); j.b.boxall@sheffield.ac.uk (J.B.)
[2] Department of Water Management, Faculty of Civil Engineering and Geosciences, Delft University of Technology, 2628 CD Delft, The Netherlands
[3] Yorkshire Water Services, Bradford BD6 2SZ, UK; katrina.flavell@yorkshirewater.co.uk
[*] Correspondence: g.kyritsakas@sheffield.ac.uk

[†] Presented at the 3rd International Joint Conference on Water Distribution Systems Analysis & Computing and Control for the Water Industry (WDSA/CCWI 2024), Ferrara, Italy, 1–4 July 2024.

Abstract: This paper presents machine learning analysis to understand the factors impacting iron concentrations and discolouration customer contacts in drinking water distribution systems. Fourteen years of network sampling and additional data from a large UK utility were collated, analysed, and interpreted using self-organising maps (SOMs), which include complex network theory (CNT) centrality metrics for the first time, investigating how possible explanatory variables interact. The outputs are used to inform ensemble decision trees for risk estimation of iron exceedance and customer contacts for each of the utility's DMAs, helping inform proactive maintenance.

Keywords: discolouration; machine learning; complex network theory; big data

check for updates

Citation: Kyritsakas, G.; Husband, S.; Gleeson, K.; Flavell, K.; Boxall, J. A Data-Driven Analysis for Understanding and Risk Estimation of Discolouration in Drinking Water Distribution Systems. *Eng. Proc.* **2024**, *69*, 206. https://doi.org/10.3390/engproc2024069206

Academic Editors: Stefano Alvisi, Marco Franchini, Valentina Marsili and Filippo Mazzoni

Published: 11 November 2024

1. Introduction and Background

Discolouration is the primary water quality issue experienced by consumers in the Western world and comprises 65% of water quality customer contacts (CCs) that UK water utilities receive [1]. The main cause of discolouration is the accumulation and subsequent mobilisation of organic and inorganic material, primarily iron, from within the drinking water distribution system (DWDS) [2]. Due to the range of possible interactions occurring within a DWDS, the use of mechanistic models for predicting water quality impacts is not currently practical. Water utilities, therefore, commonly rely on reacting to customer contacts. The only other widespread source of data to inform proactive maintenance is lab-based analysis of discrete samples collected to satisfy regulatory requirements. These are temporally and spatially sparse due to the vast scale and complexity of DWDSs. This sparsity limits derivation of understanding of the trends between the different water quality variables and restricts the ability to generate actionable information that directs investment decisions towards efficient interventions for proactive water quality management with the best return. This sparse historical data can, however, provide key network information via the application of data-driven methodologies, such as machine learning (ML), with the potential to transform the current decision-making approach.

Data-driven methodologies have become popular in the hydroinformatics domain with multiple research projects that use ML for water quality applications reported [3,4]. Regarding the management of discolouration in DWDSs, Boxall et al. [5] used a combination of ML methodologies in a sampling dataset for the identification of key water quality parameters related to increased iron concentrations and the prediction of district meter area (DMA) probability of iron exceedance [5]. In this paper, additional data and further data-driven analysis extends the work of [5] to explore relationships between iron concentrations in DWDSs, discolouration CCs, and different water quality parameters measured from both the water treatment works (WTWs) and from customer taps in a large UK water utility. For the first time, network characteristics and complexity are explored as quantified through complex network theory (CNT) centrality metrics. With relationships established, this paper then examines the application of different interpretable predictive ensemble decision trees to provide a risk estimation of iron exceedance and CCs for each of the utility's DMAs to inform proactive maintenance for the following year.

2. Materials and Methods

2.1. Case Study, Data Collection, and Data Pre-Proccessing

This data-driven analysis was conducted for Yorkshire Water, a water utility that serves more than 5 million customers in the Yorkshire area of the UK. The data used for this work mainly consist of discrete water quality samples collected from WTWs and customer taps over a period of fourteen years (2009–2022). An additional data source containing CC data over the same period was also added. Further data included static asset data, connectivity, and distribution main characteristics per DMA. The analysis included four main stages: clustering, CNT, self-organising maps (SOMs), and ensemble decision trees. Prior to this, the data extracted from multiple sources required initial pre-processing, which included connecting customer tap samples to their associated distribution main pipes; connecting each DMA to the WTWs that feed it; and calculating WTWs' monthly averages per parameter and associating this with DMAs.

2.2. Clustering of Customer Complaints

For the identification of potential discolouration events, a clustering analysis was conducted in the CC dataset using spatial and temporal criteria. For the clustering of the CCs, density-based spatial clustering of application with noise (DBSCAN) was applied using two different clustering approaches to capture different scales and likely causes of events. DMA clustering was defined as 5 or more contacts within a DMA within a 24 h period, and water supply zone (WSZ) clustering was defined as 8 customer contacts over a period of 48 h in a minimum of 2 DMAs [6].

2.3. Complex Network Theory

Complex network theory (CNT) centrality metrics are used as an innovative parameter for understanding and quantifying the complexity of a utility's DWDS. The main centrality metrics used in this work were edge (pipe) betweenness and edge (pipe) n-degree. Edge betweenness is a metric that calculates the number of times an edge appears in the shortest path between two nodes, with a high number likely to relate to hydraulically active pipes and, hence, with less potential for material accumulation. This parameter was calculated for all the pipes of the utility's DWDS with an assumption that each DMA is a unique DWDS. The DMA 30th percentile of this parameter was calculated for each DMA.

2.4. Self-Organising Maps' Application for the Identification of Factors That Influence Both Increased Iron Concentrations and Customer Complaints

SOMs were applied for the qualitative data-driven identification of correlations between parameters. An SOM is an unsupervised neural network clustering methodology that visualises multidimensional datasets in 2-D plots [7]. The visualisation plots can effectively reveal and communicate hidden non-linear complex correlations between multiple parameters, even when part of the dataset is missing or incomplete.

2.5. Ensemble Decision Trees for Calculating the Iron Exceedence and Customer Complaints Risk in DMAs

The last stage of the analysis included a prediction for each of the DMAs' probability of exceedance of iron, CC, and DMA discolouration events (DMADEs) for the next year using the yearly average data of the present year. For this classification problem, a two-category approach was selected, including a low-risk non-exceedance (N) class

(DMAs: iron < 150 µg/L, CC < 5 and DMADE = 0) and a high-risk exceedance (E) class (DMAs: iron ≧ 150 µg/L, DMAs with CC ≧ 5, DMADE ≧ 1). Two "white-box" ML approaches were used, random forest (RF) and boosting—random under-sampling boosting (RUS-Boost) and adapting boosting (ADA-Boost) [8]. The models' outputs were validated using 4 different metrics, accuracy (ACC), true positive rate (TPR), true negative rate (TNR) and Matthews correlation coefficient (MCC).

3. Results and Discussion

3.1. Identifying Correlations with SOMs

Figure 1 shows an example of the SOM analysis including CCs and iron concentrations with some key potential correlating parameters. The analysis was conducted by selecting different combinations of water quality variables based on the literature, experience, and, most importantly, an interactive process between researchers and utility practitioners. The SOM output presented indicates a positive correlation between CCs in DMAs, high iron, and high manganese and an inverse correlation between these parameters and edge betweenness and DMA edge betweenness.

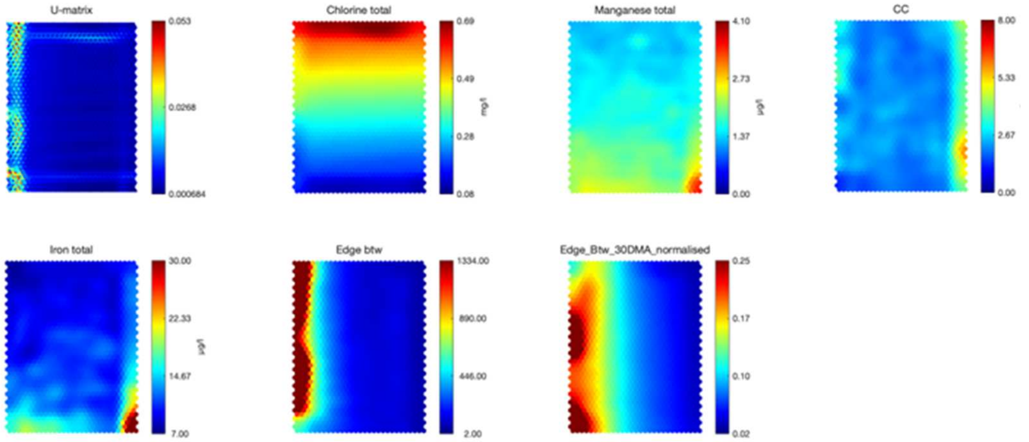

Figure 1. SOM for exploring iron and CC correlations with CNT metrics and chlorine.

The overall SOM analysis indicated that the source of the water and temperature influenced both discolouration CCs and iron concentrations. Discolouration CCs and high iron concentrations are related, but the differences found may be due to uncertainty over individual customer behaviour. The SOM analysis with CNT metrics notably found that edge betweenness and DMA edge betweenness have the potential to be used as a metric for assessing the discolouration risk of both DMAs and pipes.

3.2. Decision Tree Modelling

Predictive modelling was conducted initially using all the available variables and all the available data and then refined using a combination of variables that improved the model's accuracy with reduced complexity. Random under-sampling was applied to address the strong bias of the dataset towards the "N" class (very few network samples had high risk exceedance); the method was applied and iterated to identify the best N to E ratio for training the predictive model. The model was tested using the 2009–2021 data for training and 2022 data for testing.

The performance metric results of the best iron exceedance, CC and DMADE models are presented in Table 1. The best predictive models were then used to predict the probability of exceedance in the DMAs for the year 2023 and their probability outputs were plotted in the utility's DMA map as shown in Figure 2.

Table 1. Performance metrics of the best iron CC and DMADE predictive models.

Predictive Model	ML Method	N E	ACC	TPR	TNR	MCC
Iron	RF	3	0.811	0.714	0.812	0.12
CC	RUSBoost	3	0.781	0.649	0.796	0.306
DMADE	RF	1	0.84	0.545	0.846	0.151

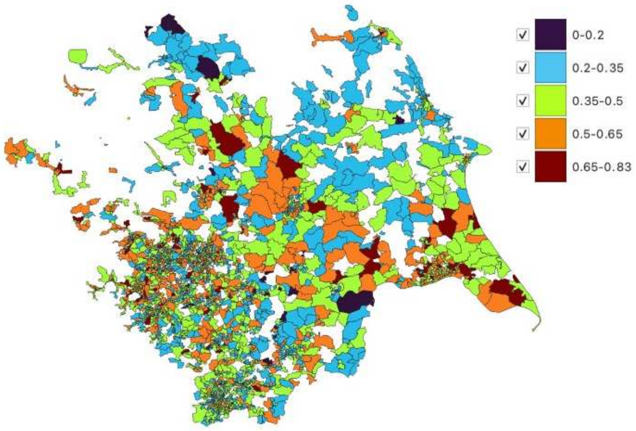

Figure 2. Plot of relative probability prediction for elevated iron (>150 µg/L) in utility's DMAs.

4. Conclusions

Self-organising maps and decision trees are shown to be able to analyse and interpret sparse network data to inform proactive management of iron and discolouration in drinking water distribution systems. Complex network theory metrics were incorporated for the first time for these water quality parameters and found to add value in complex multidimensional space.

Author Contributions: Conceptualization, G.K., S.H., K.G., K.F. and J.B.; methodology, G.K.; data curation, G.K.; writing—original draft preparation, G.K.; writing—review and editing, S.H., K.G., K.F. and J.B.; supervision, S.H. and J.B.; project administration, J.B.; funding acquisition, S.H. and J.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Yorkshire Water.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data belong to a water utility and cannot be shared publicly.

Acknowledgments: The authors are grateful for the support and data provided by Yorkshire Water Services that made this research possible. For the purpose of open access, the authors have applied a creative commons attribution (CC BY) license to any author accepted manuscript versions arising.

Conflicts of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. DWI. *Drinking Water 2020: The Chief Inspector's Report for Drinking Water in England*; DWI: Washington DC, USA, 2020.
2. Boxall, J.; Blokker, M.; Schaap, P.; Speight, V.; Husband, S. Managing discolouration in drinking water distribution systems by integrating understanding of material behaviour. *Water Res.* **2023**, *243*, 120416. [CrossRef] [PubMed]
3. Loucks, D.P. Hydroinformatics: A review and future outlook. In *Cambridge Prisms: Water*; Cambridge University Press: Cambridge, UK, 2023.
4. Li, L.; Rong, S.; Wang, R.; Yu, S. Recent advances in artificial intelligence and machine learning for nonlinear relationship analysis and process control in drinking water treatment: A review. *Chem. Eng. J.* **2021**, *405*, 126673. [CrossRef]
5. Boxall, J.; Speight, V.; Kyritsakas, G.; Kazemi, E.; Husband, S.; Bright, S.; Ledgar, S.; Montgomery, L.; Flavell, K. The application of Artificial Intelligence techniques to better manage iron in drinking water distribution systems. *Inst. Water J.* **2022**, *7*, 28–34.

6. Mounce, S.; Machell, J.; Boxall, J. Water quality event detection and customer complaint clustering analysis in distribution systems. *Water Sci. Technol. Water Supply* **2012**, *12*, 580–587. [CrossRef]
7. Kohonen, T. The self-organizing map. *Proc. IEEE* **1990**, *78*, 1464–1480. [CrossRef]
8. Dietterich, T.G. Ensemble Methods in Machine Learning. In *International Workshop on Multiple Classifier Systems*; Lecture Notes in Computer Science; MCS 2000; Springer-Verlag: Berlin/Heidelberg, Germany, 2000; Volume 1857, pp. 1–15.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.