Exploring machine learning applications for improving drinking water quality

Grigorios Kyritsakas

Thesis submission for the degree of Doctor of Engineering

Academic Supervisors: Prof. Vanessa Speight, Prof. Joby Boxall

Industrial supervisor: Claire Thom

The University of Sheffield
Department of Civil and Structural Engineering
October 2021

In loving memory of my parents

DECLARATION

I declare that no portion of this work has been submitted in support of an application for another degree or qualification of this or any other university or institute. The work presented is my own except were indicated.

PhD Abstract

Water utilities in the UK collect vast amounts of water quality data during their monitoring programs to assure that the final product that they deliver to consumers is of a high quality. This data, once checked over the compliance with the regulations, is archived and not used for further analysis. However, advanced data analytics tools, such as machine learning (ML), have the potential to uncover hidden information, regarding the complex processes that occur in the drinking water distribution systems (DWDS), from such types of data. This work contributes to the research over the application of these techniques in real world water quality problems when the water quality datasets are used as inputs. More specifically, this research investigates the potential of these techniques, by exploring their ability in analysing real drinking water quality problems and by proposing to the water utilities a new operational approach on the management of the water quality data for creating evidence that will support decision making over proactive interventions in the DWDS.

The main contribution of this work is a Big Data framework that works as a guide for the water utilities to solve water quality related problems in their DWDS by applying ML applications in the data that have already been collected. This framework proposes, in the form of 4 layers, a new holistic approach that demands changes in the way the data storage, integration, analysis and visualisation is made. It also includes a novel process to facilitate the selection of the most appropriate ML technique, based on the water quality related problem and the existing data for analysis.

Moreover, this research investigates the ability of some of the most common ML techniques by developing data-driven methodologies and applying them on water quality case studies for a water utility that supplies 5.5 million people. These methodologies are: a) a methodology that identifies correlations between different parameters and, thus, identifying factors that contribute in water quality deterioration; b) a methodology that predicts the risk of bacteriological deterioration in water exiting service reservoirs; c) a methodology for the short term forecast of free chlorine losses in drinking water trunk mains; d) a methodology that predicts the bacteriological behaviour of the water exiting the WTWs - flow cytometry total cell counts prediction in the WTWs outlet.

The results obtained by the application of these methodologies, reported in this thesis, demonstrate the huge potential of ML techniques in both understanding the factors of deterioration and predicting future water quality behaviour. Overall, the data-driven methodologies and the framework presented in this thesis, open a new discussion to researchers regarding the identification of the appropriate data and methods for creating models that improve drinking water quality, and direct water utilities over a new data management approach to gain beneficial information for their DWDS operation and maintenance.

ACKNOWLEDGMENTS

This research was sponsored and supported by Scottish Water and Sheffield University under the EPSRC platform grant STREAM IDC EP/L015412/1. I'm very thankful to Scottish Water for supporting my thesis and giving me access to their data systems, without which this work would not been achieved.

I would like to extend my thanks to the following individuals:

My two academic supervisors, Professor Vanessa Speight and Professor Joby Boxall for giving me this opportunity, for their continuous support and technical guidance through these 4 years and for believing on my abilities even in periods that I was not believing in myself.

My industrial supervisor, Clair Thom and other members of Scottish Water for the managerial support and for helping me find and understand their data collection and storage systems and thus making this thesis possible.

All the members of the Water Distribution Research Group at the University of Sheffield for the research and social discussions that developed me as a researcher and as a person.

Special thanks to Professor Costas Synolakis and Mr Evangelos Voukouvalas. The former for particularly convincing me that it's never too late to persuade PhD studies and without him I would have not applied for this position. The latter for his assistance and guidance in the use of MATLAB especially in the initial period where I was struggling to understand the basics.

Last, but not least, I would like to personally thank my two closest companions, Dr Katerina Antonopoulou, my other half, and our dog Stathy for their love and support, and for being always there where I need them.

Table of Contents

List of Figures

List of tables

List of Abbreviations

AB	AdaBoost - Adaptive Boosting trees
ADASYN	Adaptive Synthetic Minority Oversampling Technique
AI	Artificial Intelligence
ANN	Artificial Neural Network
ANN-FF	Feed forward Artificial Neural Network
BT	Boosting Trees
CM	Combined model
DBPs	Disinfection by-Products
DM	Data Mining
DMA	Discrete Meter Area
DNA	Deoxyribonucleic Acid
DNN	Deep Neural Network
DWDS	Drinking Water Distribution system
DWI	Drinking Water Inspectorate
DWQR	Drinking Water Quality Regulator
FCM	Flow Cytometry
FN	False Negative
FP	False Positive
HPCs	Heterotrophic Plate Counts
HRT	Hydraulic Retention Time
ICCs	Flow Cytometry Intact Cell Counts
IWA	International Water Association
LSTM	Long Short-term Memory Networks
MAE	Mean Absolute Error

MCC	Matthews Correlation Coefficient
ML	Machine Learning
MSE	Mean Square Error
NARX	Nonlinear Autoregressive Exogeneous Algorithm
NMSE	Normalised Mean Square Error
NOM	Natural Organic Matter
PCA	Principal Component Analysis
R^2	Coefficient of Determination
RB	RusBoost - Random Under Sampling Boosting
RF	Random Forest
RNNs	Recurrent Neural Networks
RMSE	Root Mean Squared Error
RSZ	Regulatory Supply Zone
SCADA	Supervisory Control and Data Acquisition
SMOTE	Synthetic Minority Oversampling Technique
SOMs	Self-Organising Maps
SR	Service Reservoir
SR_RT	Service Reservoir Retention Time
SVM	Support Vector Machine
SW	Scottish Water
TCCs	Flow Cytometry Total Cell Counts
THMs	Trihalomethanes
TN	True Negative
TNR	true Negative Rate
TOC	Total Organic Carbon
TP	True Positive
TPR	True Positive Rate

tSNE	t-distributed Stochastic Neighbor Embedding
USEPA	United States Environmental Protection Agency
WDNs	Water Distribution Networks
WDTMs	Water Distribution Trunk Mains
WHO	World’s Health Organisation
WOA	Water Operation Area
WQ	Water Quality
WSZ	Water Supply Zone
WTW	Water Treatment Work
WUs	Water Utilities
WWTP	Wastewater Treatment Plant
ΔCl	Chlorine losses

1. Introduction

1.1. Background & motivation

Water utilities' (WUs) aim to provide their customers with drinking water of high-quality standards set by stringent national and international regulations. Thus, it is in their duties to guarantee that the water that arrives in their customers' taps is of a high quality. This means that it is WUs' responsibility to guarantee that the travel of the drinking water from the source to the taps through their drinking water distribution systems (DWDS) is safe. This travel includes the water treatment in the water treatment works (WTWs), with innovative treatment technologies, the distribution to service reservoirs (SRs) through the water distribution trunk mains (WDTM), the storage of the water in SRs that are well maintained, and the distribution to the taps through the water distribution mains (WDM). In the UK, the drinking water delivered by the WUs is of a high-quality standard, and microbial contamination of water pollution are very rare phenomena. However, even small incautions by the WUs could produce serious health impacts to humans and huge negative economical and reputational impacts to the companies. An example to give is the *Cryptosporidium* contamination that occurred in Lancashire, U.K. affected 575000 people and forced United Utilities to pay a £300,000 fine (DWI 2017).

For eliminating the risk of water quality deterioration, guaranteeing that these phenomena will not occur, and complying with the regulations, WUs monitor the water throughout its travel from the source to the WTWs, the SRs, and the customer taps. Regarding the WTWs, in the UK, WUs monitor all the processes using sensors that measure the main water quality (WQ) parameters such as, turbidity, pH and Cl_2, in steady frequency (usually 5 minutes). These measurements are stored in the supervisory control and data acquisition (SCADA) system, and, through this, the processes are adapted to properly treat the raw water and produce a final product of high standards. However, monitoring the water during the distribution in the same way as WTWs, is more complicated as the water distribution mains are buried. Therefore, in the UK, for monitoring the water quality inside their DWDS, the WUs take samples from their WTWs' outlets, their service reservoirs (SRs') outlets, and randomly from some of their customer taps. For the DWDS monitoring program there is some minimum number of samples per year that WUs should collect from their assets, which is set by the Drinking Water Inspectorate (DWI) in England and Wales and the Drinking Water Quality Regulator (DWQR) in Scotland (DWI 2016; DWQR 2019a). Routine monitoring schedule is mostly focusing on the identification of indicator bacteriological microorganisms (coliforms, E. coli, enterococci and clostridium perfringens) and the heterotrophic plate counts (HPCs) as parameters that indicate biological degradation in the DWDS (Standing Committee of Analysts 2002). In addition, other parameters that could be indicators of chemical or aesthetic degradation of the drinking water are also measured. Overall, in Scotland, DWQR requires Scottish Water (SW) to undertake regular monitoring samples for 51 parameters such as iron, manganese, lead, turbidity etc. (DWQR 2019a). SW has undertaken more than 300,000 samples in their WTWs, SRs and tap samples only for the year 2016 (DWQR 2017).

The common policy for the WUs, in the UK, is to check the measured parameters over their compliance with the regulations limits and prioritize their interventions in the DWDS if the parameters exceed them. This approach by the WUs, however, is a reactive management approach that includes interventions in the DWDS (flushing, additional disinfection, cleaning of the network, mains replacements etc.) to handle water quality degradation that occurred there. This approach sets the WUs' customers at risk of consuming drinking water with low quality standards for the period between the water quality incident occurring and the intervention to act. Therefore, new methodologies are required that could provide early warning advice to WUs decision makers and transform the reactive management of the DWDS to a proactive one. As WUs collect large amounts of data that increase year after year, the research over these new methodologies should include data-driven techniques that extract valuable information hidden in raw datasets. Positive outputs of this research could give sufficient evidence and support WUs over their proactive management of their networks.

Data analytics or data science is not a new scientific field. It is a field where visualisation methods and other approaches are developed for understanding and interacting with datasets (Gandomi and Haider 2015). With today's continuous development of computing systems and data storage it is possible to store huge amount of data (digital traces) and, therefore, new approaches in data science were required to gain better knowledge from the available "big data". Machine learning (ML) is a computer science domain that develops mathematical and statistical algorithms to give computers the ability to learn from data (Alpaydin 2014). ML methods are applied for prediction of future trends, for a deeper understanding of the relationships between various variables, for clustering unlabelled data and for detecting events. ML offered new approaches and techniques to data scientists for better understanding of the available data. Nowadays, its use is expanded in various scientific fields such as medicine, science, and engineering (Praveena and Jaiganesh 2017; Zekić-Sušac, Mitrović, and Has 2020; Dai and Wang 2019). WUs in the UK are in the process of creating data analytics departments in within their organisations and have already successfully collaborated with Academia in various cases for understanding the roots of discolouration in their DWDS (Speight, Mounce, and Boxall 2019) understanding the roots of bacteriological failures in the WTWs (K. Ellis et al. 2014, 2015), predicting discolouration in water distribution trunk mains (Meyers, Kapelan, and Keedwell 2017; Kazemi et al. 2018), predicting water quality in the DWDS (Garcia, Puig, and Quevedo 2020; S. R. Mounce et al. 2017; Vries et al. 2016; Kühnert et al. 2014) etc. As the scientific domain increases and new ML techniques are created, this field could offer benefits in the drinking water quality management. Unfortunately, though, there is no clear connection between the water quality problems that the WUs require to be solved with data, the available datasets that could be used for that, and the ML techniques that could be applied in these datasets to provide WUs with actionable information. Therefore, further research is required for understanding better the logic behind each ML technique, investigating which one could be more appropriate for a certain water quality investigation using specific datasets, and, mainly, for providing WUs with a holistic and strategic tool that could aid them operate and maintain their DWDS using their own data.

1.2. Research questions, aim and objectives

1.2.1. Research questions

This work explores the ability of data-driven techniques in supporting the management of drinking water quality. Therefore, the research questions that the thesis addresses are:

1. What WUs can learn and how can they benefit from the application of machine learning techniques in their data in terms of better managing their systems and improving the quality of the drinking water that they serve?
2. Which machine learning type is the most appropriate for each specific water quality problem?
3. What is required to facilitate the application of machine learning methods by WUs for deriving new knowledge that could support the management of their drinking water distribution systems?

1.2.2. Aims and Objectives

The overall aim of this research is to provide a new operational approach on the use and the analysis of water quality data that WUs collect on their daily monitoring routine programs, that develops new knowledge regarding drinking water quality behaviour in the DWDS and creates evidence for supporting proactive interventions for their maintenance. The main objectives are as follows:

1. To investigate the existing machine-learning techniques and the ways that these could be applied to drinking water quality problems (addresses research question 1).
2. To develop data-driven models for understanding the roots for drinking water deterioration in DWDS (addresses research question 1 & 2).
3. To develop predictive data-driven models for water quality deterioration events in the DWDS (addresses research question 1 & 2).
4. To compare the various data-driven methods and suggest the most appropriate for a specific water quality problem (addresses research question 2).
5. To present a new strategic approach that includes changes in WUs mode of collecting, integrating, and analysing their own data to create evidence that supports decisions over a proactive management of their DWDS (addresses research question 3).

1.3. Thesis structure

Including the introduction chapter, this thesis is divided in 11 chapters. In this section, a summary of the remaining chapters is presented:

Chapter 2: Literature review

This chapter is divided into 2 parts. The first part focuses on presenting the main aspects of drinking water quality in the DWDS, concentrating mostly in the sources of bacteria in the DWDS, the parameters that influence bacteriological regrowth in the DWDS, the types of bacteria that are used as indicators of bacteriological regrowth. In addition, a further overview of the main causes of discolouration in the DWDS is also provided. In the second part of this review, the ML categories are explained, and a brief presentation of some key ML applications in various scientific fields and in the water sector, is provided.

Chapter 3: Machine learning techniques selection for drinking water quality problems and performance metrics for their evaluation

In the first part of this chapter, a methodology for selecting ML techniques based on the type of water quality (WQ) problems that could be solved using data-driven approaches is presented. This methodology proposes a 6-step approach that should be followed based on Mitchell's definition for machine learning algorithms (Mitchell 1997). In the second part of this chapter, the main ML methods that were investigated for this thesis are presented. Some of these ML methods were applied in real case studies presented in the remaining chapters of the thesis, and some of them were just investigated for their potential but not applied in a real WQ problem. Finally, in the last part of the chapter, the performance metrics used for evaluating the ML methods' performance in the case studies are presented. This chapter addresses objective 1 and partially objective 5.

Chapter 4: Scottish Water's water quality data analysis

This chapter presents the samples water quality monitoring program that SW follows to control the quality of the drinking water in their systems. In addition, the steps that were taken to transform the raw data from various sources into a main samples' water quality dataset are provided.

Chapter 5: Understanding bacteriological activity in service reservoirs by applying data-driven techniques on water quality datasets

In this chapter two ML techniques (self-organising maps and principal component analysis) were applied in SW's SRs and WTWs WQ dataset for identifying the main factors that increase bacteriological activity and bacteriological failures in the SRs. This chapter addresses objectives 2 and 4.

Chapter 6: A SOMs application on water quality datasets for investigating the impact of switching disinfection type on drinking water quality

In this chapter, self-organising maps are applied in SW's tap WQ dataset taken from their systems that during the period between January 2012 and May 2020 have switched their disinfection type from chlorination to chloramination. The main aim is to understand the impact on drinking water quality of switching disinfection by identifying correlations between the various water quality parameters before and after the disinfection switch. This chapter as the previous one addresses objectives 2 and 4.

Chapter 7: A comparison between ensemble decision tree models for the classification of service reservoirs using drinking water quality data

This chapter investigates the potential of ensemble decision trees, a group of ML methods, in predicting low chlorine events and coliform events in SRs. The methods used in this work are random forest, Adaboost and Rusboost and the proposed methodology was developed for the identification of high - risk SRs in the forthcoming month based on the samples WQ monitoring data, taken in the SRs outlet in the previous months. A comparison of the models is made based on their performance and a new model that combines the best single ML model is created. This chapter addresses objectives 3 and 4.

Chapter 8: Predicting short-term chlorine losses in water distribution trunk mains using machine learning applications

In this chapter a data-driven model for the short-term prediction of chlorine losses in water distribution trunk mains is presented. The data-driven model uses three different ML techniques, random forest, feed-forward artificial neural network and non-linear autoregressive artificial neural network, and their performance is compared. This chapter addresses objectives 3 and 4.

Chapter 9: A data-driven investigation on the performance of Balmore water treatment plant

A data-driven investigation in Balmore, one of SW's largest treatment plants, is presented in this chapter. This investigation aims to, initially, understand the factors that contribute the most in the increase of flow cytometry total cell counts in the outlet of Balmore WTW and then to predict the total cell counts in the WTWs' outlet up to certain hours ahead. This chapter addresses objectives 2,3 and 4.

Chapter 10: A Big-Data framework for actionable information to manage drinking water quality

This chapter contains the main contribution that this thesis is providing. Based on the outputs of the previous chapters, this chapter proposes a framework for the holistic approach on how to use ML applications for specific water quality problems from the standardisation of the data storage to the data integration, the data analysis, and the visualisation of the outputs. The main purpose of this framework is to provide evidence to WUs decision makers to direct their investments into certain areas and, thus, guarantee the good operation and maintenance of their DWDS. This chapter is a reproduction of a paper submitted at the Environmental Science Water Research & Technology journal and is under review at the time of the writing and addresses objectives 1 and 5.

Chapter 11: Discussion - conclusions and future work recommendations

This chapter summarises the outputs of the previous chapters, presents the novel contributions of this thesis and makes recommendations for future research.

1.4. Resources and Publications

1.4.1. Resources

The data used in this thesis belong to SW and, therefore, is not publicly available. The data include discrete monitoring samples data, telemetry data, time-series data, water distribution modelling data and assets information. The rainfall data collected in the rainfall stations that belong to the Meteorological Office of the UK (Met Office) were used in chapter 5 and chapter 7. This data is available to purchase under request and for research purposes.

The codes used for this thesis are written in MATLAB 2018b,2019a and 2019b (MathWorks). All the different algorithms are available in the following github repository: https://github.com/goresonic/PhD-Codes.

1.4.2. PhD publications

This research work produced and will produce a number of publications that are presented in this section and summarised in the table 1-1.

1.4.2.1. Journal paper

Grigorios Kyritsakas, Joseph B. Boxall and Vanessa L. Speight (2020). A Big Data Framework for Actionable Information to manage Drinking Water Quality. J. *Environmental Science: Water research & Technology themed issue: Data-intensive water systems management and operation (under review)*
(Chapter 10 is a reproduction of this paper)

1.4.2.2. Conference papers

Grigorios Kyritsakas, Vanessa Speight, Joby Boxall. (2021). "A data-driven model for the prediction of chlorine losses in water distribution trunk mains" Abstract accepted at the Hydroinformatics conference (HIC) 2022, Budapest

Grigorios Kyritsakas, Vanessa Speight, Claire Thom, Joby Boxall. (2020). "A machine learning approach for the prediction of chlorine decay in the water distribution trunk mains" Conference paper accepted at the Hydroinformatics Conference (HIC) 2021, Mexico City, Mexico - conference cancelled due to COVID.
(This conference paper was based on the work presented in chapter 8. The paper was accepted but the conference was cancelled due to COVID restrictions)

Grigorios Kyritsakas, Vanessa Speight, Claire Thom, Joby Boxall. (2019). "Investigating drinking water behavior treated by different disinfection with the use of a machine learning technique on water quality datasets". Computing and control in the Water Industry (CCWI) conference 2019, Exeter, UK
(This conference paper was based on the work presented in chapter 6)

Grigorios Kyritsakas, Vanessa Speight, Claire Thom, Joby Boxall (2019). "A machine learning application on water quality datasets for understanding the factors of bacteriological failures on Service reservoirs" *IWA-ASPIRE 2019 conference, Hong Kong*
(This conference paper was based on the work presented in chapter 5)

Grigorios Kyritsakas, Vanessa Speight, Claire Thom, Joby Boxall (2019). "A machine learning approach to predict low chlorine decay in water distribution trunk mains" *Sensing in Water 2019 conference, Nottingham, UK*
(Part of the work presented in chapter 8 was included in this paper)

An additional paper titled: "Predicting Water Quality Events in Storage Tanks Using Data-Driven Techniques" that included part of the work presented in chapter 7 was submitted for presentation to the Water Quality Technology 2020 conference, but the conference was cancelled due to COVID.

1.4.2.3. Future publications

In the future two more journal publications are planned. The first one will include the research presented in chapter 7 regarding the prediction of SRs that are at risk. The second one will include the investigation made in Balmore WTW that is presented in chapter 9.

1.4.2.4. Linking publications with research questions and objectives

The objectives of the thesis and the way that are linked to the research questions and the publications is presented in the following table

Table 1-1: Overview of publications and their linkage to the objectives and the research questions of this thesis

Publication	Research questions answered	Objective addressed
Grigorios Kyritsakas, Joseph B. Boxall and Vanessa L. Speight (2020). A Big Data Framework for Actionable Information to manage Drinking Water Quality. J. *Environmental Science: Water research & Technology themed issue: Data-intensive water systems management and operation (under review)*	a) What is required to facilitate the application of machine learning methods by WUs for deriving new knowledge that could support the management of their drinking water distribution systems? b) Which machine learning type is the most appropriate for each specific water quality problem?	Objectives 1 and 5 a) Investigate the existing machine-learning techniques and the ways that these could be applied to drinking water quality problems b) Present a new strategic approach that includes changes in WUs mode of collecting, integrating, and analysing their own data to create evidence that supports decisions over a proactive management of their DWDS
Grigorios Kyritsakas, Vanessa Speight, Claire Thom, Joby Boxall. (2019). “Investigating drinking water behaviour treated by different disinfection with the use of a machine learning technique on water quality datasets”. Computing and control in the Water Industry (CCWI) conference 2019, Exeter, UK	a) What WUs can learn regarding from their data regarding the impact of switching to chloramination on the drinking water quality b) Is SOMs a good tool for this type of investigation?	Objective 2 a) Develop data-driven models for understanding the roots for drinking water deterioration in DWDS
Grigorios Kyritsakas, Vanessa Speight, Claire Thom, Joby Boxall. (2020). “A machine learning approach for the prediction of chlorine decay in the water distribution trunk mains” Conference paper accepted at the Hydroinformatics Conference (HIC) 2021, Mexico City, Mexico - conference cancelled due to COVID.	a) How WUs can benefit from the application of machine learning techniques in their data in terms of better manage their systems and improve the quality of the drinking water that they serve? b) Which machine learning type is the most appropriate for predicting chlorine decay?	Objective 3 and 4 a)Develop predictive data-driven models for water quality deterioration events in the DWDS b)Compare the various data-driven methods and suggest the most appropriate for a specific water quality problem

Grigorios Kyritsakas, Vanessa Speight, Claire Thom, Joby Boxall (2019). “A machine learning application on water quality datasets for understanding the factors of bacteriological failures on Service reservoirs” *IWA-ASPIRE 2019 conference, Hong Kong*	a) What WUs can learn and how they can benefit from the application of machine learning techniques in their data in terms of better manage their systems and improve the quality of the drinking water that they serve? b) Is SOMs a good tool for understanding the factors that influence the bacteriologica behaviour in the SRs?	Objective 2 a) Develop data-driven models for understanding the roots for drinking water deterioration in DWDS
Future publication based on chapter 7: A comparison between ensemble decision trees classifiers for the risk classification of drinking water service reservoirs	a) What WUs can learn and how they can benefit from the application of machine learning techniques in their data in terms of better manage their systems and improve the quality of the drinking water that they serve? b) Which machine learning technique is the most appropriate in predicting failures in SRs?	Objective 3 and 4 a)Develop predictive data-driven models for water quality deterioration events in the DWDS b)Compare the various data-driven methods and suggest the most appropriate for a specific water quality problem
Future publication based on chapter 9: Investigation of the bacteriological behavior of the WTWs	a) How WUs can improve the bacteriological performance of their WTWs using their own data? b) Which machine learning type is the most appropriate for predicting TCCs in in WTWs outlet? c)How can PCA and SOMs help WUs identify the factors that increase the bacteriological activity exiting WTWs?	Objective 3 and 4 a)Develop predictive data-driven models for water quality deterioration events in the DWDS b)Compare the various data-driven methods and suggest the most appropriate for a specific water quality problem

2. Literature review

2.1. Introduction

As described in the introduction chapter, this thesis is investigating the potential gains that water utilities (WUs) could obtain by applying machine learning techniques on water quality data obtained either for their monitoring programs or for certain investigations. Therefore, prior to the investigation a general understanding of the water quality and the data analytics is required. The aim of this literature review is, firstly, to highlight the various water quality aspects and concentrate, especially, on bacteriological regrowth and monitoring but also to, briefly, explain the machine learning categories and applications.

More specifically, the aim of this review is:

1. To present some of the tools and the techniques that are used to detect bacteria in the DWDS.
2. To explain the sources of bacteria and metals in the water.
3. To understand the factors that could influence bacteriological regrowth and spread in the DWDS.
4. To present the risks of water discolouration on drinking water quality.
5. To demonstrate various machine learning techniques and their applications.
6. To show examples where machine learning techniques were used in research related to the water sector.

2.2. Method

There were various papers, books, theses, and reports that contributed to this review. Most of the papers were collected from Scopus website, Science direct. The search terms used during the search were “drinking water” “disinfection”, “distribution system”, “bacteria in the drinking water distribution system, “supervised learning”, “unsupervised learning”, “data-driven methods”, “data-driven, water”, “data-driven modelling”. Papers were also found as references of other papers. The books that were used the most in this review were *“Introduction to Machine Learning”* by E. Alpaydin (Alpaydin 2014), “A*pplied predictive modelling*” by M. Kuhn and K. Johnson(Kuhn and Johnson 2013), “*Pattern recognition and machine learning*” by C. Bishop (Bishop 2006), “*The Elements of Statistical Learning*” by T. Hastie et al. (Hastie, Tibshirani, and Friedman 2008), “*Introduction to potable water treatment processes”* by S. Parsons and B. Jefferson (Parsons and Jefferson 2009) and *“Microbial growth in Drinking Water Supplies”* by P. van der Wielen and D. van der Kooij (van der Krooij and van der Wielen 2014). The review includes articles published before September 2021.

2.3. Water quality in drinking water distribution systems

Access to drinking water is a human right recognized by the United Nations (United Nations 2010). However, even if clean drinking water is taken as granted in the Western World, around 2.1 billion people lack access to safe and readily available water at home (WHO, 2017). Treatment and safe distribution of water are highly important as unsafe water is a potential source of various diseases. Water Treatment is obtained in the Water Treatment Works (WTWs) after passing various different processes depending on the water source. Water distribution is achieved through its transportation inside a complicated network of service reservoirs (SRs), pipes and valves known as Drinking Water Distribution Network (DWDS). Thus, the thorough and systematic control of the WTW operation, the maintenance of both the WTW and the DWDS and the water quality monitoring are an obligation for the WUs.

Water quality monitoring is a highly important procedure that guarantees that the water is safe for drinking. WUs are collecting samples from the WTWs, the SRs and customer's taps to analyse them for various physical, chemical, and microbiological parameters (WHO 2011). The European Directive (Council of the European Communities 1998) and the "Public water Supplies in Scotland" regulation (DWQR 2014) indicate the quality standards that the drinking water in Scotland should maintain in order to be considered as safe to consume. In both standards, it is specified that there are mandatory and non-mandatory parameters that WUs should measure in their monitoring program. The results of the samples' analysis should be below the indicating values suggested by these regulations for every parameter measured, otherwise further investigation over a potential water quality accident is required.

In this section, the main methods for detecting bacteria in discrete monitoring water quality samples taken from various parts of the DWDS, the sources of bacteria and the indicator parameters for bacteriological regrowth are presented. Furthermore, the effect of iron and manganese presence on water quality and the discolouration phenomenon are, briefly, described.

2.3.1. Bacteriological monitoring

WUs in the Western World are forced by the regulations (Council of the European Communities 1998; Gorchev and Ozolins 2011) to take samples for bacteriological monitoring. Identification of bacteriological existence in the drinking water is crucial for human health and understanding the potential sources of microbiological failure in the DWDS or WTW and preventing future contamination is very important.

There are various methods for the identification of indicator bacteria and other microorganisms that are either based on bacteria's physical characteristics or their deoxyribonucleic acid (DNA). The first category of tests, known as phenotyping methods, are all the typical culture tests that WUs routinely do. The second category, known as genotyping methods, are methods that are applying molecular analysis for the identification of specific microorganisms. In this literature review, the focus is on presenting the methods that Scottish Water is using for monitoring bacteria. Some of these parameters are mandatory to measure (heterotrophic plate counts, Indicator bacteria) and some are not (flow cytometry total and intact cell counts).

2.3.1.1. Heterotrophic Plate Counts (HPC)

Heterotrophic plate counts (HPC) is the most common monitoring tool for WUs. Heterotrophic microorganisms including bacteria, fungi, protozoa etc. (Berry & Raskin, 2006) need organic carbon as energy source for growth (Bartram et al., 2004). HPC include a variety of culture-based tests that intent to recover microorganisms by grow them in a carbon-based plate environment (Bartram et al. 2004). Typically, the microorganisms are incubated at 22^{o}C and 37^{o}C and, according to Water Research Centre, when the microorganisms that grow in the 37^{o}C environment are more than those growing in 22^{o}C environment the water should be considered as contaminated (Water Research Centre 1976). It is generally accepted that even if higher numbers of HPC microorganisms are found in the test, these should not be necessarily considered as a threat to human health (Allen, Edberg, and Reasoner 2004; M. W. LeChevallier, Seidler, and Evans 1980; Dick van der Kooij and van der Wielen 2014a; Payment P. 2003). It is also proven that there is no direct relationship between *coliform* presence in the water and HPC (K. Ellis et al. 2014; Sam Van Nevel et al. 2016; G. Liu et al. 2013) . The percentage of HPC microorganisms recovered in the plates in the routine monitoring usually represent less than the 1% of the total concentration of bacteria in a DWDS or WTWs and the results always vary widely between locations, seasons of the year etc. (Bartram et al. 2004; van der Kooij and van der Wielen 2014; McCoy and Olson 1987). Thus, HPC is not a method that could show the level of microbial regrowth in the DWDS. However, if HPC microorganisms exceed the 500 CFU/ml threshold, could interfere with coliform and *E.Coli* enumeration (Allen, Edberg, and Reasoner 2004). Furthermore, the outputs of an experts meeting, as presented by Bartram et al. (Bartram et al. 2004), indicate the importance of the HPC measurements in DWDS. Therefore, WUs continue doing the HPC tests as it is a cheap and easy method to investigate a failure of maintaining a disinfectant residual, a possible water contamination and a fouling of the DWDS (van der Kooij & van der Wielen, 2014a).

2.3.1.2. Indicator bacteria monitoring tests

2.3.1.2.1. Coliform Bacteria - E. Coli test

Coliform and *E.Coli* tests are the most important routine tests that the WUs are doing as these tests provide a potential detecting faecal contamination in the raw water or the DWDS but also to check the effectiveness of the WTWs. Coliform bacteria are a genus of bacteria that include *Escherichia, Klebsiella, Enterobacter, Citrobacter, and Serratia*. The main characteristics of this genera are that these bacteria are oxidase-negative and produce acid from lactose at 37 ^{o}C (Standing Committee of Analysts 2009). *Escherichia Coli* also belong to this genera and have the same characteristics but the difference is that *E.Coli* produce acid from lactose at 44 ^{o}C (Standing Committee of Analysts 2009) and they also produce indole from tryptophan. Coliform Bacteria are indicators of water quality deterioration and *E.Coli* are indicators of a potential faecal contamination (van der Kooij and van der Wielen 2014). *E. Coli* are not a pathogenic microorganism, but their detection could indicate the existence of *E. Coli* O157:H7 which presence is related to the development of haemorrhagic colitis and haemolytic uraemic syndrome (Standing Committee of Analysts 2016). LeChevallier (Mark W. LeChevallier 1990) suggested that because coliforms' survival and regrowth in the DWDS is frequent, their presence does not necessarily mean recent contamination in the system.

The main mechanisms for the coliform and *E.Coli* introduction to the drinking water are the defective performance of the WTWs or the intrusion into the DWDS (Besner et al. 2002). Coliforms identification in the WTWs outlets is, usually, related to the poor treatment performance with respect to raw water quality changes due to rainfall or other events that change the organic matter (Besner et al. 2002). The main conditions for coliforms growth in the DWDS are pipe corrosion, pipe burst, temperature above 15^{o}C, low disinfection residual and sediments (Mark W. LeChevallier 1990; Besner et al. 2002; Dick van der Kooij and van der Wielen 2014b).

2.3.1.2.2. Enterococci

Enterococci are genera of bacteria that are used as indicators of faecal contamination of water. Enterococci identification tests are used for assessing the risk of contamination on water that is used in recreational parks, for assessing the effectiveness of disinfection in swimming pools and for detection of microorganisms in wastewater treatment plants (Standing Committee of Analysts 2015b). WUs in the UK, are using Enterococci identification tests as secondary tests for faecal pollution when coliforms are detected and no *E.Coli* test has been made.

2.3.1.2.3. Clostridium perfringens

Clostridium perfringens are Gram-positive, rod-shaped, anaerobic, spore-forming pathogenic bacteria used as subsidiary indicators. They form spores that resist to environmental loading for long periods. Therefore, the presence of *Clostridium perfringens* in clean water is related to faecal contamination and when detected on water samples in absence of other indicator bacteria indicates a remote source of water pollution (Standing Committee of Analysts 2015a). Their main characteristics are that they are able to reduce sulphite to sulphide at 44^{o}C in less than 24h and they produce phosphatase which is the main characteristic that make them unique in the genera of Clostridia bacteria (Standing Committee of Analysts 2015a).

2.3.1.3. Flow Cytometry (FCM)

Genotyping methods for the identification of bacteria are split into two groups PCR-based and luminescence/ fluorescence methods. Flow Cytometry (FCM) is a fluorescent method that facilitates measurement of a cell by fluorescence using laser (Y. Wang et al. 2010) This method was firstly recommended by Hammes et al. (Hammes et al. 2008) and Y. Wang et al. (Y. Wang et al. 2010) for use in the identification of bacteria in drinking water. FCM is an accurate technique that is used for counting the total number of cells counts (TCCs) (Hammes 2008, Wang 2010) when nucleic acid stain is used, or the number of intact cells counts (ICC) (Helmi 2014) when nucleic acid is combined with viability stains (S. Van Nevel et al. 2017). Van Nevel et al (2017) suggest that FCM can soon replace the HPC method in the WUs routine bacteriological monitoring due to its accuracy, its rapid analysis, its reasonable cost, and the high information that it provides. Various studies reinforced this argument (G. Liu et al. 2013; Helmi et al. 2014; Sam Van Nevel et al. 2016). SW is the first company in the UK that used FCM in their WTW outlet and SRs (SR) outlet samples as a routine sampling tool and the first company in Europe that uses online FCM for more accurate and faster results. However, arguments against FCM exist and include: i) the difficulties in identifying the viable and non-viable bacteria (Sam Van Nevel et al. 2016; Gillespie et al. 2014) ii) the identification of indicator bacteria (Kathryn Ellis 2013; Gillespie et al. 2014) and iii) the problem with undercounting the number of bacteria when these are in clusters and attached in inorganic compounds (van der Kooij & van der Wielen, 2014).

2.3.2. Sources of Bacteria

Raw water cleaned from the WTW, travels through the DWDS to arrive in our taps. In this travel, the abundant microorganisms and other contaminants raw water is treated through in various process stages in the WTW and exits as a high-quality drinking water. However, during its transport to our taps deteriorates as it passes through a complicated network of the DWDS and SRs. This is because water during the travel to the taps i) passes through the complicated, usually buried, old and sometimes in poor condition network of the DWDS ii) remains for a certain period in SRs iii) contacts with the biofilm of the DWDS, a complex community of microorganisms and inorganics attached in the pipe surface.

2.3.2.1. Raw Water

Raw water could be impacted by industrial waste, sewage overflow spills, pesticides, and other chemicals from agricultural run-offs (Parsons and Jefferson 2009). Treatment of the raw water is dependent on the type of the source (groundwater, lowland surface water, upland surface water) (Parsons and Jefferson 2009). Upland waters are generally low in minerals and contain high dissolved organic matter and their treatment focuses on the organic matter reduction and disinfection (Parsons and Jefferson 2009). Lowland waters receive water from upland but also effluents from sewage treatment works, industrial wastewater and agricultural chemicals, therefore their treatment should include pesticides removal, chemical removal, and organic matter reduction (Parsons and Jefferson 2009). Groundwater is usually a low bacteriological source so the treatment for this source is focusing more on the chemicals and inorganics removal (Parsons and Jefferson 2009). A typical treatment process includes coagulation – flocculation, clarification, filtration, and disinfection (Parsons and Jefferson 2009). Even if disinfection is the last stage of treatment, it could not be effective for bacteria removal if the previous treatment stages do not perform well (Kathryn Ellis 2013). Various studies prove that changes in organic matter or inorganics concentration of the raw water could have a negative impact on the drinking water quality (Korth et al, 2004; Kulbat & Sokołowska, 2017). Korth et al. (Korth et al. 2004) found that the increase of natural organic matter (NOM) in the German water sources generates a NOM increase in the treated water, requires a higher disinfectant demand, and, finally, could potentially spike the number of bacteria entering in the DWDS. Research on climate change and precipitation effects on microbial pollution show that rainfall increases the microbial load in the raw water (Tornevi, Bergstedt, and Forsberg 2014; Curriero et al. 2001). Tornevi et al. (Tornevi, Bergstedt, and Forsberg 2014) suggest that heavy rainfall and extreme events will lower the water quality of the raw water, and this will be a huge challenge for the WUs in the future. Curriero et al. (Curriero et al. 2001) study the waterborne diseases in the US and found that there is a significant correlation between heavy rainfall events and waterborne diseases.

2.3.2.2. Pipe Bursts

Water pipe networks and service reservoirs (SRs) are designed in a way to maintain high water quality and to prevent any contamination of the water by external factors. However, the pipe network and the SRs are vulnerable infrastructures that corrode through age, damage and bad operational practices and maintenance.

Bursts occur in the DWDS pipes because of the phenomenon of pressure transients (also known as water hammer). Pressure transients are caused by unexpected increase of the water velocity because of power cut or operational activities (e.g., sudden close of a valve). During bursts, the flow rate suddenly increases and thus the system starts having negative pressures that allow intrusion of contamination into the DWDS through leaks. LeChevallier et al (Mark W Lechevallier et al. 2003) investigated the health risks associated with bursts and found that the soil and the groundwater that could potentially enter the system are sources of faecal pathogens and bacteria. Besner et al. (Besner, Prévost, and Regli 2011) created a conceptual model to discover the risks of leakage on human health and understand the causes of negative pressure events. They found that duration of the intrusion is a key factor on contamination of the system. According to Yang et al. (Yang et al. 2011), the most effective tool to limit the contamination intrusion due to bursts, would be an optimized pressure management program with an accurate modelling tool that will control the negative pressures. In the same research, they also found that the last protection border against intrusion in chlorinated systems, is to maintain the chlorine residual in the DWDS above 0.2mg/L but as regards the chloraminated systems, maintaining the chloramine residual does not reduce the risk of contamination.

2.3.2.3. Biofilm

Biofilms are complex communities that consist of bacteria, other microorganisms (fungi, protozoa etc.) and inorganic compounds attached in the surfaces of all the systems that treat and transfer the drinking water to our taps. These communities are embedded in a produced by the community matrix of Extracellular polymeric substances (EPS) (K. Fish, Osborn, and Boxall 2017; K. E. Fish et al. 2015; Flemming et al. 2016; Douterelo, Sharpe, and Boxall 2013; Hallam et al. 2001). The process of biofilms' formation is the following: i) microorganisms, carbohydrates and organic acids that survived disinfection passed through the DWDS and got attached to the systems' surface, ii) Proteins are polymers are adhered and the EPS is formatted iii) microorganisms in the surface are creating individual colonies (cells) in the empty spaces of the surface iv) the colonies are embedded in EPS matrix and inorganic compounds are attached v) biofilm formed vi) release of bacteria due to changes in the hydraulic regime, enabling stabilisation of the biofilm (Dick van der Kooij and van der Wielen 2014a; Camper 2014). Fish et al. (K. E. Fish, Osborn, and Boxall 2016) suggests that biofilm appears in every DWDS pipe, it should be characterized as a proper microbiological community and since it is impossible to completely remove biofilm for pipe surfaces, the research should focus on the management and the control of it.

Bacteria release from biofilm is very common phenomena within biofilms (Douterelo et al., 2013; Fish et al., 2016; Flemming et al., 2016; Husband et al., 2016; Moore et al., 2000; Petrova et al., 2016). According to Petrova et al. (2016) there are three ways of bacteria release from biofilms: i) desorption, the direct transfer from the upper layers of the biofilm to the bulk water, ii) detachment that occurs when hydraulic change, such as shear stress, occur and iii) dispersion which is the passive release of bacteria from biofilms due to some physiological change in the bacterial community. Biofilm detachment is also subcategorized into four different groups depending on the mechanisms that remove it: abrasion, grazing, erosion, and sloughing (Moore et al. 2000). Abrasion occurs when bacteria collide with other particles in the bulk water. Grazing is the result of biofilm bacteria consumption by eukaryotic organisms. Erosion is a low-level detachment that occurs due to a continuous flow increase in a DWDS. Sloughing is the detachment of larger quantities of biofilm due to extreme shear stress forces. Phenomena such as sloughing, or erosion are related with high turbidity but also with bacteriological failure as the detached biofilm enters the bulk water and then ends in the taps (Fish, Osborn, and Boxall 2016).

2.3.3. Disinfection

Disinfection is the last process in WTWs. The purpose of disinfection is to inactivate pathogens and bacteria responsible for the deterioration of the drinking water, the prevention of waterborne diseases. Disinfection efficacy is dependent upon the performance of the upstream treatment processes. There are two groups of disinfection methods:

i) Chemical disinfection with the use of mainly chlorine but also ozone, bromide, heavy metals, and hydrogen peroxide

ii) Physical disinfection with the use of heat, light etc.

Most common disinfection methods are chlorination, Chloramination and Ultraviolet light (UV) which are presented in this literature review (Parsons and Jefferson 2009).

The disinfection process is dependent on two factors, the concentration of the disinfectant and the contact time. Various studies (Fish et al., 2015, 2016; LeChevallier et al., 1985) found that the disinfection process could result in the entrance of small, injured bacteria and microorganisms in the DWDS which might recover under the correct temperature and nutrients conditions.

2.3.3.1. Chlorination

Chlorination is the most common disinfection type used in water treatment. Chlorination disinfection uses either pure chlorine (Cl_2) as liquefied gas or sodium hypochlorite with 13% of Cl_2 or solid calcium with 30% Cl_2. The main advantage of the chlorination is that the formed chlorine residual remains in the DWDS for a long period and thus is protecting it from bacterial regrowth (Parsons and Jefferson 2009). Chlorine gas reacts with water to form hypochlorous acid (HOCl) which further dissociates to hypochlorite anion OCl^- following these equations.

$$Cl_2 + H_2O \rightarrow HCl + HOCl$$
$$HOCl \rightarrow H^+ + OCl^-$$

HOCl is penetrating the bacterial cell walls and destroying the cytoplasm. Thus, HOCl is better disinfectant than OCl^-. The dissociation of HOCl to OCl^- is correlated with high pH values while in pH below 5 HOCl is undissociated. Therefore, the best solution for efficient disinfection would have been to keep the pH in lower levels but, unfortunately, in lower pH values degradation of HOCl to oxygen and hydrochloric acid (HCl) becomes significant when contact times are 30 or more minutes (minimum time required per disinfection). In chlorination, the goal is always to find the balance that promotes efficient disinfection without impacting the taste of the water. Generally, it is recommended a dose between 0.5-1 mg/l of chlorine in the contact tank in order to have a sufficient chlorine residual of 0.2 mg/l in the taps (Parsons and Jefferson 2009).

Chlorine reaction with natural organic matter and inorganic ions inside the DWDS and forms the disinfection by-products (DBPs). Most common DBPs are the trihalomethanes (THMs) and the haloacetic acids (HAAs). DBPs concentrations are related to various health risks (Parsons and Jefferson 2009). Therefore, WHO and USEPA (USEPA 2002a) have regulated the maximum allowed DBPs concentration in the clean water.

The chlorine residual is consumed by biofilm heterotrophic bacteria and chemicals in the water (Fish et al., 2016). Various studies found that biofilm growth strengthens microbial resistance on disinfection (Berry et al., 2006; Fish et al., 2015; Lechevallier et al., 1988). Fish et al. (Fish, Osborn, and Boxall 2016) showed the importance of EPS matrix in the biofilm bacteria resistance to chlorine residual. Other studies observed that chlorine residual is more efficient on the deactivation of bulk bacteria than biofilm bacteria (Besner et al., 2002; Morrow et al., 2008). Low chlorine concentrations have been successful in controlling coliform bacteria (Besner et al. 2002) but LeChevallier et al. (LeChevallier, 1987) found that in some cases even high chlorine doses are ineffective in controlling coliform appearance. In a different study though, LeChevallier (LeChevallier, 1990) found that increasing the chlorine dose has helped to control coliform appearance.

2.3.3.2. Chloramination

Chloramination is the disinfection method where ammonia (as ammonium sulphate NH_4^+) is added either simultaneously with the chlorine dose or in a short time after chlorine dose. Chloramines are not as effective in penetrating microorganisms as chlorine but are very stable and therefore are used in long DWDS with large water age (Parsons and Jefferson 2009). NH_4^+ reacts with HOCl creates 3 chloramine compounds, firstly monochloramine (NH_2Cl), then dichloramine ($NHCl_2$) and finally trichloramine (NCl_3) as shown in the following equations:

$$NH_4^+ + HOCl \rightarrow NH_2Cl + H_2O + H^+$$

$$NH_2Cl + HOCl \rightarrow NHCl_2 + H_2O$$

$$NHCl_2 + HOCl \rightarrow NCl_3 + H_2O$$

Chloramination procedure is pH dependent as in chlorination. Normally, the chlorine ammonia ratio is 5:1 by weight which minimizes the free chlorine residual and the free ammonia concentration (Parsons and Jefferson 2009).

Research works on chloramines reaction with biofilms found that chloramines were more effective on reducing biofilm growth than chlorine (LeChevallier et al. 1988; LeChevallier, 1990). LeChevallier (LeChevallier, 1990) also found that coliforms appearance was more often in chlorinated than in chloraminated systems and van der Kooij (van der Kooij, 2014) found that monochloramines are more effective with Legionella control and reduction. However, the use of chloramination should be carefully controlled as high free ammonia residual in the

DWDS could result in the growth of nitrifying bacteria that could end in nitrification in some parts of the network (Dykstra 2007; Y. Zhang and Edwards 2009). Nitrification is the process of biological oxidation of ammonia to nitrite performed by ammonia-oxidizing bacteria (AOB) leading to loss of chloramine residuals which form an important barrier in the management of drinking water quality and public health (Telfer, 2014). Factors leading to nitrification are low chlorine residual, high temperature, low pH, and high-water age (Telfer,2014). Research on the field shows that approximately two thirds of medium to large chloraminated systems in the US experience nitrification (Dykstra 2007). Thus, Zhang & Edwards (Y. Zhang and Edwards 2009) suggest that with the absence of nitrification, chloramines are more persistent than chlorine, that they reduce corrosion and decrease the amount of HPCs in the bulk water but when nitrification exists or corrosion rates are low, chlorine is more efficient than chloramines.

2.3.3.3. Ultraviolet light (UV)

Ultraviolet light (UV) is used in various countries (Netherlands, Germany, Denmark) as the main disinfection process and replaced the traditional disinfection processes (Medema et al. 2014). The main advantage of the UV systems is that, in these systems, with the absence of disinfection residual no disinfection by- products (DPBs) are produced.

UV is an electromagnetic radiation with higher frequency than visible light but lower than X-Rays. UV can penetrate the cell and react directly with the DNA of the bacteria. Thus, this reaction prevents the replication of these bacteria which means that even if they are not destroyed, they are inactivated (Parsons and Jefferson 2009). In recent years, with the development of computational fluid mechanics, the optimization of UV reactors is achieved and makes UV disinfection a more popular choice for disinfection. However, there are four major factors that influence the UV performance: (i) UV transmission that is related with the transmissivity of the water, (ii) turbidity that scatters the light, (iii) foulants that are using UV radiation to oxidize and (iv) hydraulics that may affect the transmissivity of water (Parsons and Jefferson 2009). USEPA recommends the use of UV disinfection for waters where *Cryptosporidium* is observed as neither chloramination nor chlorination are effective against this species (USEPA 2001).

2.3.4. Indicator parameters for microbial regrowth in DWDS

Bacteria regrowth in the DWDS is dependent on the concentration and nature of biodegradable compounds that serve as a source of energy (van der Kooij and van der Wielen 2014). The use of these compounds from bacteria in the DWDS depends on the growth kinetics of the existing in the DWDS bacteria and various environmental parameters such as the nutrients (nitrogen, carbon, phosphorus), the age of the water, the water temperature, the disinfectant residual, the sediments, and the pipe material (van der Kooij and van der Wielen 2014).

2.3.4.1. Nutrients

Carbon is the nutrient required the most for microorganisms' growth in the DWDS as a source of energy. Natural organic matter (NOM) is a complex mixture of organic compounds that are found in the most water resources. NOM is the main source of carbon in the DWDS even if the majority of NOM is removed in the WTW (Parsons and Jefferson 2009). Total organic carbon (TOC) is the total available organic carbon concentration in the drinking water. From the available TOC in the drinking water only a small fraction is available as a nutrient for the microorganisms known as biodegradable organic carbon (BOC). It is, therefore, comprehensible, that by knowing the BOC concentration, the growth of biofilm and heterotrophic bacteria could be predicted. There are two main methods for the measurement of the BOC: the biodegradable dissolved organic carbon (BDOC) and the assimilable organic carbon (AOC). BDOC is the portion of the biodegradable organic carbon mineralized by heterotrophic microorganism and AOC is the portion of the biodegradable organic carbon that can be converted to cell numbers by a single or by defined microorganisms (Camper 2014). BDOC concentrations are bigger than AOC concentrations due to the larger number of bacteria contributing to the carbon consumption.

Nitrogen is an inorganic nutrient that could be found in water resources that are close to the agricultural areas due to the use of agricultural fertilisers (Parsons and Jefferson 2009). Consumption of high levels of nitrogen concentration in drinking water could cause methaemoglobinaemia and therefore regulations either in the UK but also in the EU require its removal in the WTW (Parsons and Jefferson 2009). A secondary source of nitrogen in the drinking water is through the nitrification process in chloraminated systems (Telfer, 2014). The presence of nitrifying bacteria in the chloraminated systems lead to the formation of nitrite and nitrate in the DWDS. In these systems high chlorine levels or low ammonia levels are required to inactivate the nitrifying bacteria (Berry, Xi, and Raskin 2006).

Phosphorus could be the limiting nutrient in the drinking waters (Lehtola, Miettinen, and Martikainen 2002). In heterotrophic bacteria it is required a ratio of carbon to nitrogen to phosphorus of approximately 100:10:1 (M. W. LeChevallier, Schulz, and Lee 1991). Most of the phosphorus is usually removed in the coagulation/flocculation process in the WTW. The addition of small doses of orthophosphate in the drinking water to prevent the release of lead and suspend corrosion, is the main reason for observing phosphorus in the DWDS (Douterelo, Husband, and Boxall 2014). Research made in the waters in Finland found that the lack of either nitrogen or phosphorus is the principal factor for the limitation of microbial growth and that minor changes in phosphorus concentration may affect the microbial growth potential in the DWDS (Miettinen and Vartiainen 1997). A further study in the same region, showed that phosphorus addition in steady-state biofilms increases the microbial concentrations (Lehtola, Miettinen, and Martikainen 2002). However, Gouider et al. (Gouider et al. 2009) suggested that phosphate does not have any impact on microbial regrowth in the DWDS. Finally, findings by Douterelo et al. (Douterelo, Husband, and Boxall 2014) demonstrated that phosphate increases microbial diversity in the DWDS but their results could not indicate if the presence of phosphorus increases the microbial community as well.

2.3.4.2. Temperature

Temperature is the most important indicator parameter of the bacterial regrowth process as it affects directly or indirectly all the other factors responsible for the microbial growth (Mark W. LeChevallier 1990). More specifically, temperature affects the disinfection efficiency, the disinfectant residual, corrosion rates and the water hydraulics due to increased water consumption (Besner et al. 2002). When the water temperature is equal or lower to 5^{o}C the growth rate of bacteria decreases (van der Kooij and van der Wielen 2014a). LeChevallier et al. found that coliform bacteria are increasing when the water temperature is higher than 15 ^{o}C (LeChevallier et al., 1991; LeChevallier, 1990). Fish et al. (Fish et al., 2016) observed that most bacteriological issues, in the DWDS, are happening in the warmer months, probably because of the increased bacteria growth rate during these months. Temperature is highly correlated with chlorine decay and generally in high temperatures the disinfectant residual is decreased (Parsons and Jefferson 2009). Finally, it is observed in various studies that when the water temperature is higher, the chlorine residual should be increased as the bacteria growth rate increases (Francisque et al., 2009; LeChevallier, 2014).

2.3.4.3. Disinfectant residual

A disinfectant residual limits the bacteria growth rate. The effects of chlorine in the DWDS are explored in detail in a previous section of this chapter. In this section a small summary is presented. High chlorine residual in the DWDS, may result in the formation of disinfection by-product (DBP) as it reacts with NOM and inorganic ions (Parsons and Jefferson 2009). Chloramination is more effective than chlorine in reducing the number of coliform bacteria in the DWDS (LeChevallier, 1990), reducing the biofilm (LeChevallier et al., 1988) and is less reactive with NOM and corrosion by-products (van der Kooij and van der Wielen 2014a). However, in the disinfection with chloramines, the presence of ammonia in the DWDS may contribute in the nitrification phenomena (Y. Zhang and Edwards 2009).

2.3.4.4. Pipe Material

The DWDS consists of pipes composed of different materials such as iron/steel, asbestos cement, copper, plastics and concrete. The influence of pipe material characteristics on biofilm growth, density and composition was investigated in various works (Camper, 2014;

Douterelo et al., 2014; Fish et al., 2017; LeChevallier, 1990; Niquette et al., 2000). Niquette et al. (Niquette, Servais, and Savoir 2000) showed that biofilm on iron mains is more problematic than biofilm on plastic mains. Various other investigations agreed with this view (M. W. LeChevallier, Schulz, and Lee 1991; Camper 2014; K. E. Fish et al. 2015). Finally, Douterelo et al. (Douterelo, Husband, and Boxall 2014) found that there is a significant difference between bacteria communities on plastic and iron pipes, with the density of the biofilm being greater in the iron mains but the bacteriological community in the plastic pipe being more diverse.

Corrosion of pipe material has an impact on the amount of biofilm in the DWDS as corroding iron is realised which is used by iron and manganese oxidizing bacteria (Berry et al., 2006; LeChevallier, 1990). Furthermore, corrosion material could affect the ability of chlorine to inactivate biofilm bacteria (Besner et al. 2002). The importance of iron-corrosion material in the development, the growth and the strength of biofilms in the DWDS was experimentally shown by Camper (Camper 2014). Corrosion control mechanisms, such as the addition of phosphate, are required when a DWDS is at risk, to improve the effectiveness of disinfection and therefore to control the biofilm growth rate in the DWDS (Besner et al., 2002; LeChevallier, 1990).

2.3.4.5. Sediments

Sediments are formatted when particles in the water are accumulated in the bottom of the pipe due to gravity. Studies found that sediments are promoting microbial regrowth (McCoy et al., 1987; Vreeburg et al., 2004; Vreeburg et al., 2008). Sediments are responsible for the discolouration of the drinking water which is the main reason for customers complaints to the WUs (Vreeburg & Boxall, 2007). The origin of particles could be either the WTW or corrosion material from iron mains inside the DWDS (Vreeburg, 2014). The source of particles that enter the network from the WTW could be the incomplete removal from raw water, the precipitation of metals such as manganese and iron passing the coagulation/flocculation process and the bacterial biomass that passed the disinfection process without being completely destroyed (Vreeburg, 2014). However, particles coming as material from the corrosion of iron mains are playing a very important role in the increase of turbidity. An extended study on discolouration by Vreeburg and Boxall (Vreeburg & Boxall, 2007) described the roots of discolouration and the mechanisms of accumulation of the particles in the DWDS and presented methods for predicting discolouration. This study also indicated that particles, related to discolouration, promote biological regrowth. Vreeburg (Vreeburg, 2014) suggests that the percentage of bacteria biomass in particles is between 1-12% which makes the sediments an important factor in water quality degradation.

Unidirectional flushing of the distribution system is the main method that the WUs use to remove sediments from water pipes (Boxall et al., 2003). Vreeburg et al. (Vreeburg et al., 2008) found that flushing in the presence of sediments removes more organic material and thus the regrowth rate is reduced. However, Douterelo et al. (Douterelo, Husband, and Boxall 2014) found that initial changes in hydraulic conditions, due to flushing, remove part of the biofilm but there are layers of sediments that under velocities used in the routine flushing programs, will not be mobilized.

2.3.4.6. Water age

Water's travel to customer taps through the DWDS could last days depending on the distance from the WTWs, the water velocity based on the demand, and the general DWDS' design. In general, an increased age of water increases the bacteriological activity in the water and thus, it also increases the risk of water quality deterioration (USEPA 2002a; Emmanuelle I. Prest et al. 2016).

Over-dimensioning of either the service reservoirs or the pipe network of a DWDS reduces the water mixing and creates areas inside the DWDS that the water could stagnate for weeks before being consumed (USEPA 2002a; Zlatanović, van der Hoek, and Vreeburg 2017). The main impact of water stagnation is the loss of the disinfection residual that directly leads to microbial regrowth and increased numbers of bacteria on customers' taps (Mark W. LeChevallier 1990). Further studies that investigated the factors of bacteriological regrowth in some DWDS, indicated stagnation and water age as the factors that contributed to the decay of the disinfection residual and, consequently, to the bacteriological growth (Kerneïs et al. 1995; H. Wang et al. 2012). In addition, a study on stagnation indicated that it could potentially increase the metal release from pipes, especially in the summer period (Zlatanović, van der Hoek, and Vreeburg 2017). Finally, USEPA includes the disinfection by-products (DBPs) as one of the health impacts that the high water age could cause (USEPA 2002a).

2.3.5. Iron and Manganese

There are various inorganic compounds that can be found in water sources, passed to the DWDS, and influence the drinking water quality including arsenic, nitrate, lead, iron and manganese (Parsons and Jefferson 2009). However, even if water with low and medium iron and manganese concentrations are not considered of high risk to human health (Parsons and Jefferson 2009), these two metals are mostly related to the discolouration phenomenon and thus, the reduction and the control of their concentrations is of a high priority for the WUs (Husband et al., 2011; Husband et al., 2010; Vreeburg & Boxall, 2007; Vreeburg et al., 2008).

Minerals with iron and manganese concentrations could be found in both rocks and soils that water passes through, and therefore both metals are found in any type of source water especially in waters with high organic matter (Parsons and Jefferson 2009). Iron and manganese could, to some extent, be oxidized upon contact with air producing turbid water and stain surfaces and affect water quality (Khadse et al., 2015; Parsons & Jefferson, 2009). Therefore, their concentrations are reduced (or completely removed), in the first stages of the water treatment process, by oxidizing them to insoluble forms, firstly by using oxygen and then by using a strong agent such as chlorine or ozone in combination with pH control for the removal of the most persistent manganese ions (Edzwald, 2011; Ellis et al., 2000; Khadse et al., 2015). However, particles contain iron and manganese appear in the DWDS even after passing the treatment works, because of the ineffectiveness of the treatment processes (Speight et al., 2019; Vreeburg & Boxall, 2007; Vreeburg et al., 2008) or because of the corrosion of pipes (Gerke et al., 2016; Peng et al., 2010; Sarin et al., 2004; Seth et al., 2004).

Iron is the metal that can be found the most in the DWDS, not only in the bulk water but also in the pipe network and equipment which are mostly constructed from iron based material (J. Hu et al. 2018). Therefore, as expected, corrosion of metallic pipes is the main reason for high iron concentrations in the DWDS (Sarin et al., 2003; Sarin et al., 2004; Seth et al., 2004; Vreeburg, 2014). The consequences of pipe corrosion regarding the bacteriological regrowth were explained in the sediments section of this chapter. In addition to that, high concentrations of iron in the drinking water influence the colour and the taste of the water and stain household equipment, such as washing machines, which, subsequently, leads to customer complaints (Seth et al. 2004).

Manganese mostly appears in the DWDS because of its ineffective removal during the oxidation and filtration processes in the treatment works (Sly et al., 1990). However, Peng et al. (Peng et al. 2010) suggested that manganese is also released from PVC and iron pipes. High levels of manganese give a metallic taste to the drinking water that, as in iron, lead to customers' dissatisfaction and, consequently, to damage of the water company's reputation (Seth et al. 2004).

2.3.6. Discolouration

Discolouration occurs when accumulated in the DWDS material, that mainly contain high levels of iron and manganese, is, suddenly, mobilised creating an unpleasant colour in the drinking water (Boxall et al., 2003; Husband et al., 2010, 2011; Vreeburg & Boxall, 2007). Although not an actual measurement of discolouration, turbidity is mainly used as the main parameter for understanding the phenomenon (Husband et al., 2010; Sharpe et al., 2019). The impacts of discolouration on bacteriological regrowth were, briefly, explained in a previous section. In this section, the main causes of discolouration are presented.

Discolouration is observed in various DWDS with different source water, different treatment processes and different pipe material and it is the main cause of customer complaints in England and Wales (Husband et al., 2016). Material contributing in discolouration, contain organic and inorganic compounds and could appear in the DWDS for three main reasons: (i) insufficient removal of the material in the treatment process, (ii) bad maintained coagulation and filtration processes in the WTWs (iii) corrosion of metallic and mainly cast iron pipes (Cook et al., 2011; Husband et al., 2011). Various studies indicated that the dominant material causing discolouration is iron with manganese having an important contribution as well (I. J. H. G. Vreeburg and Boxall 2007; Seth et al. 2004; J. H. G. Vreeburg 2007; Cook and Boxall 2011). The main cause of discolouration is a sudden change in the hydraulic conditions of the DWDS, such as flow increase, that urges sediment layers to mobilise from one part of the network to another (J. H. G. Vreeburg, Schaap, and Van Dijk 2004; P. S. Husband and Boxall 2011). Main events that could, potentially, change the hydraulic conditions of a DWDS are pipe bursts, increases in water demand, hydrant use and misuse of valves (J. H. G. Vreeburg, Schaap, and Van Dijk 2004; P. S. Husband and Boxall 2011).

2.4. Data analytics and machine learning

2.4.1. Data analytics

Data analytics (also known as "Big Data" or Data mining) is the process of collecting, cleaning, transforming, and modelling data to gain useful information, predict future trends and support decision-making. The applications of methods such as machine learning in data, are very common in areas like finance and stock market, in medicine for medical diagnosis, in science and technology. The applications of machine learning, and data analytics are increased with the evolution of computers and computer storage as the datasets year after year are getting larger and they also include, apart from numbers and strings, videos, images, audio, web pages etc. Over the last years Machine Learning and Data Mining techniques are often used in the hydroinformatics and water resources domain for predicting flood frequency, water consumption, water discolouration, calibrate water models, understand the roots and parameters of discolouration etc. Gandomi et al. (Gandomi et al., 2015) summarized in their research paper the definition of big data analytics, its applications, the methodologies that are used and future research needs that are required.
This review briefly explains the field of machine learning, specifies the various machine learning categories, and approaches, and presents some cases where machine learning techniques were used in the water sector.

2.4.2. Introduction to machine Learning

Machine Learning (ML) is a subset of artificial intelligence (AI) that uses collected data in various formats to enable computers to "learn" and improve performance towards a specific task without explicit programming. In other words, ML is the area of AI that studies the algorithms that computers should use to optimize future performances by learning from existing data or past experience (Alpaydin 2014). The programming models that are created include some initial parameters set by the user and during the learning period (also known as training period) the computer program optimizes these parameters with the use of existing data.

The basics of machine learning is math. It uses the theory of statistics to build mathematical models and algorithms that process the data in specific ways and create predictable outputs based on the data patterns. However, the created algorithms determine how the computer interprets the data; therefore, it affects the outcome of the learning process and the final output. Furthermore, the quality and the complexity of the data could affect the learning procedure. Thus, machine learning is part of computer science that includes the creation of algorithms, the cleaning of the data before the use and the training of the computer to provide efficient predictions (Mueller and Massaron 2016).

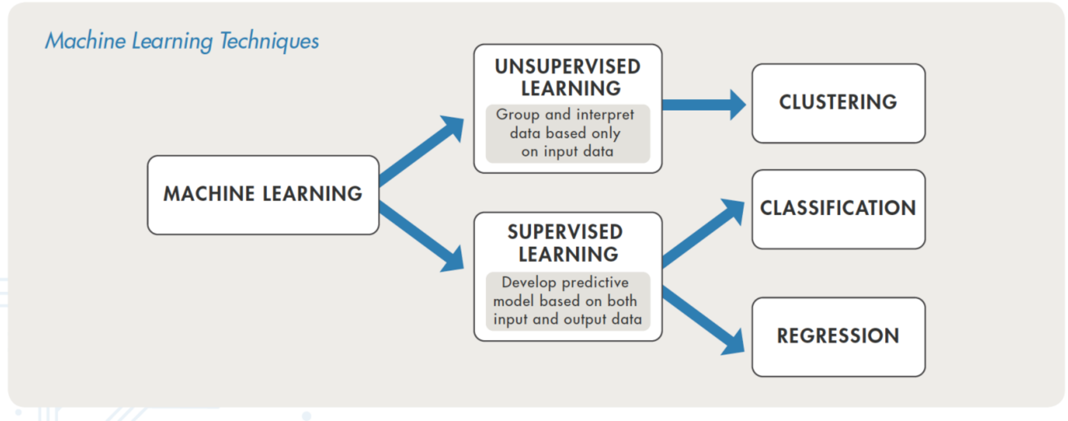

Machine Learning Techniques
MACHINE LEARNING
UNSUPERVISED LEARNING
Group and interpret data based only on input data
SUPERVISED LEARNING
Develop predictive model based on both input and output data
CLUSTERING
CLASSIFICATION
REGRESSION

Figure 2.1:Machine Learning methods (MathWorks 2016a)

There are two main categories in machine learning: supervised machine learning methods and unsupervised machine learning approaches. In supervised learning, the machine learns a function that represents the known responses (output) to the input set of data and then this function is used to predict the response to new input data. In unsupervised learning, there are no specific outputs, and the machine is learning to understand the connections between the data, to group them in clusters with similar characteristics, and to explore hidden patterns in the datasets.

2.4.2.1. Supervised Learning

In Supervised Learning the steps that usually are followed regardless the chosen methodology are the determination of the type of dataset required, the dataset collection, the determination of the input feature in the learned function, selection of the appropriate algorithm – method, running the selected algorithm and evaluation of the accuracy of the learned function (Praveena and Jaiganesh 2017). The Supervised Learning techniques are a form of classification or regression or both.

Classification problems are the problems where the machine is required to identify in which set of categories the new observations (new set of data) belong to, based on a training data set whose category is known. The algorithm that is used for the classification problem is called the classifier. A typical example of a classification problem is the assignment of a new email to inbox or to spam class. Classification predictions can be evaluated by checking the accuracy of the model compared to the actual outputs (more details in the next chapter). In *Regression*, the training data set of inputs and known outputs create a predictive function that gives a continuous output value for any new input value (new observation). The accuracy of regression models is checked by using various performance metrics (more details in the next chapter). There are various supervised learning methods and algorithms that could be used for both classification and regression. Most common algorithms are the Decision Trees and the Artificial Neural Networks (ANN). However, there are methods that could only be used for regression problems such as linear regression and methods that could only be used for classification problems such as logistic regression. In the following paragraphs the main supervised learning methods are presented.

Linear regression is the most common statistical and machine learning method. It is used for solving regression problems. In linear regression, the predictive function is linear whose model parameters are estimated from the input dataset. The main advantages of this statistical method are its simplicity and its ability to predict by solving the function with specific inputs (Alpaydin 2014).

Logistic regression is another statistical method that is used in Machine Learning. Logistic regression is applied to datasets when two possible outcomes are expected. Therefore, Logistic Regression is used for classification problems. The Logistic or sigmoid function calculates the probability of a binary response based on predictor variables. More specifically, it models the probability of an output based on the input values and therefore, by setting a limit to this value the values are divided into two different categories. In cases where more than two output categories are required, multinomial logistic regression is used which combines multiple logistic regression methods (Alpaydin 2014).

Decision trees are predictive Machine Learning algorithms that are trained from existing datasets and create hierarchical tree structures to demonstrate the relationships between the input predictors from the datasets and the target class (Quinlan 1987). Decision trees are used for both classification problems (classification trees) and regression problems (regression trees). Each node of the tree (tree leave) is an attribute that is used in the problem and each branch is the outcome of the attribute test (Pedrycz and Sosnowski 2001). The tree grows from the most meaningful attribute and continues with other attributes at lower nodes. A tree is trained by splitting the input dataset into subsets based on the attribute test and the process continues until the splitting does not add value to the predictions (Praveena and Jaiganesh 2017). Breiman et al. (Breiman et al. , 1984) developed the most widely used decision tree algorithm known as Classification and Regression Tree (CART). CART is a non-parametric decision tree that, after training, could be either a regression or a classification tree depending on the desired outputs.

Ensemble Classifiers or Ensemble methods are supervised methods that use multiple algorithms to create a predictive model with improved performance (Rokach 2010). The main idea behind ensemble classifiers is to apply as many individual predictive classifiers as possible into the training dataset, compare their outputs and combine the most accurate to obtain a "super" classifier that suits the most with the available dataset (Rokach 2010). The mechanisms that are used to build ensemble classifiers are: (i) using various datasets in one learning algorithm (ii) using various parameters in a single algorithm (iii) using different learning algorithms in the same dataset (Kotsiantis et al., 2007). However, they also argued that ensemble classifiers have three main disadvantages: they require increased storage, they increase the computational requirements, and they are difficult to be applied by a non-expert user. The most common ensemble classifier methods are bagging, boosting, and random subspace (random forest); their comparison was presented in a review paper by Dietterich (Dietterich 2000).

Support Vector Machines (Cortes and Vapnik 1995) are supervised learning models for classification and regression analysis. Support Vector Machine (SVM) is an ML method that, after trained with specific datasets, creates an optimal tool to separate values into different groups with the same characteristics. SVM maps the input variables n into a dimensional feature space. A linear decision surface is created based on the training dataset and separates the network into different parts. SVM was successfully applied in various domains including text categorization, classification of images and image segmentation (Alpaydin 2014).

Artificial Neural Networks (ANN) are computing systems inspired by biological neural networks. These systems are trained from 'examples" without being programmed to a specific task (Bishop 2006). A neural network consists of a large number of neurons that are separated in three classes, the input neurons (input layer), the output neurons (output layer) and the in-between neurons known as hidden layers (Kotsiantis, Zaharakis, and Pintelas 2007). There are various ANN types, for example, feed-forward networks where the information, that every layer receives, is taken by the previous layer only or recurrent neural networks (RNNs) where the information passed to one layer contains information from not only its previous layer but from previous layers as well (Bishop 2006). ANNs are widely applied in various domains including medicine, process control, game-playing, face identification, text recognition etc. (Bengio 2009; Aggarwal 2018).

Naïve Bayes Classifiers (NBC) are probabilistic classifiers that are using Bayes' probability theorem (Alpaydin 2014). They take all the attributes of the dataset and analyse them individually as equally important factors. The representation of NBCs is the probability of each class in the training dataset and the probability of each input value given each class. Thus, NBCs learning procedure is fast as only these two probabilities need to be calculated. Based on these two probabilities, the naïve models could perform prediction and classification (Maimon and Rokach 2006).

2.4.2.2. Unsupervised learning

In unsupervised machine learning there are no specific outputs to correspond to the input dataset and therefore the task here is to get some knowledge from the data. In unsupervised learning there are no correct answers, and the algorithms are left to present interesting structures in the data (Mueller and Massaron 2016). Unsupervised learning could be further grouped into clustering, association, dimensionality reduction and anomaly detection techniques (Usama et al. 2017; Hastie, Tibshirani, and Friedman 2008) . Clustering aims to create groups (clusters) that split the input dataset based on common properties and characteristics, association is used for discovering relations between variables in large and complex datasets, the dimensionality reduction algorithms aim to reduce the dimensionality of the dataset and finally anomaly detection algorithms are used for detecting errors in the datasets (Usama et al. 2017; Hastie, Tibshirani, and Friedman 2008). Some of the unsupervised techniques that belong to one of these categories could also be used for problems related to the other unsupervised learning categories. In the following paragraphs, the main unsupervised machine learning categories are presented.

Clustering is the most common reason for using unsupervised machine learning techniques. There are three main clustering approaches based on the way that they follow to separate the datasets in different clusters. The first approach is the *partitional clustering* that separates the data into a strict predefined number of clusters where its sample belongs only to one cluster.

The most common technique that belongs in this category being *k-means,* a method that is briefly explained in the next chapter (Maimon and Rokach 2006) . In this approach belong also the density-based clustering methods, such as *density – based spatial clustering of applications with noise* (DBSCAN), where instead of selecting the number of clusters, the minimum number of samples that belong in a certain neighbourhood (Ester et al. 1996). A special method that also belongs in the density-based family of ML methods is Self-organising maps (SOMs) which is another method that is explored further in the next chapter (T. Kohonen 1990).

The second approach is the *hierarchical clustering* that creates hierarchical clusters either by starting from the “bottom” where each observation is a cluster and then pairs of clusters are merged and move up the hierarchy (agglomerative type), or by putting all the points of the dataset in one cluster and then dividing the cluster in smaller sub-clusters (divisive clustering type) without defining the number of clusters (Usama et al. 2017). The third clustering approach is the *Bayesian clustering* where, once defined, the number of clusters are formed based on probability distributions, such as the Gaussian, and the data are split into different clusters based on the probability of following that distribution. Thus, each datum could belong in more than one clusters. The most common Bayesian method is the Gaussian mixture model (GMM) (Usama et al. 2017).

Association rule learning is a method used for uncovering strong relationships between variables in datasets and it is a very popular tool for mining in commercial databases web mining (Hastie, Tibshirani, and Friedman 2008). In association, the goal is to find the possibility of a variable being present when another variable or variables are present. For example, what is the probability that someone that buys bread and cheese in a supermarket, will buy ham as well. There are various algorithms that generate the association rule with the most known being the Apriori algorithm and the *Eclat* algorithm (Hastie, Tibshirani, and Friedman 2008) .

Dimensionality reduction is the process of reducing the dimensions of a large dataset and representing it with a smaller one that has new features created as functions of all the features of the dataset (Usama et al. 2017). Dimensionality reduction is mainly applied for visualisation of the dataset but could be also used for clustering, for understanding the variables correlation and for feature (variables) extraction. The most common dimensionality reduction algorithm is *principal component analysis (PCA)*, a method that is described in the following chapter (Jolliffe 2002). Another dimensionality reduction method that is presented in the following chapter is *t-distributed Stochastic Neighbor Embedding (t-SNE)* (van der Maaten and Hinton 2008).

Anomaly detection is the process for discovering an outlier or a group of outliers in large datasets. The algorithms that are used for anomaly detection are also used for clustering or classification. Anomaly detection is used for data leakage prevention, fraud detection, consumption anomalies and tumour detection (Usama et al. 2017).

2.4.2.3. Other machine learning approaches

There are some other machine learning methods that are standing between unsupervised and supervised learning. A brief presentation of some of these approaches is presented in the following paragraphs.

Semi-supervised learning (SSL) is settled between unsupervised and supervised learning. It uses unlabelled data in combination with a small amount of labelled data for training to increase learning accuracy (Chapelle, Scholkopf, and Zien 2006). The first approach on SSL was the self-learning classification algorithm in which the algorithm is trained by the labelled data firstly, then labels the unlabelled data according to the existing chosen function and finally is retrained by its own predictions. However, this method's success depends on the accuracy of the selected supervised method. Interest in SSL has increased because there are applications in which there is plenty of unlabelled data that needed to be used such as images, texts, bioinformatics etc. (Chapelle, Scholkopf, and Zien 2006).

Reinforcement learning is the learning method where a computer learns through trial-and-error interactions with the environment. It is used when the desired output is a sequence of correct actions to reach a specific goal (Alpaydin 2014). In the standard reinforcement learning, a computer agent receives an input indication at a specific time and selects a specific action from a set of actions that the user provides, to generate as output. The selected action is sent to the environment and the value of this transition is transmitted through a reinforcement signal. This procedure continues as above, and the agent learns from the trial and error guided by reinforcement algorithms. The main difference between reinforcement learning and supervised learning is that in reinforcement learning there is no input/output training, and the agent is not told if its chosen action is the best available action or which action should have been chosen (Kaelbling and Littman 1996).

Deep learning is a new approach in machine learning (basically is not different to supervised and unsupervised approaches) that learns from data representation and not from algorithms and therefore, requires more data for training and learning(Lecun, Bengio, and Hinton 2015). Generally, a deep learning neural network (DNN) is an ANN with more neurons, more complex ways of connection and multiple layers between the input and output layers (Patterson and Gibson 2017). The deep learning methods were well known since the first years that ANNs became popular, but they were rarely used due to their computational cost (Usama et al. 2017). However, in recent years, due to the technological development of computer hardware and computer storage, DNNs' training ability has improved as they were learning faster and from larger input datasets. DNNs could be applied in supervised, un-supervised, and reinforcement learning tasks and they were implemented in various scientific fields for clustering, data mining, classification etc. (Che et al. 2018; Patterson and Gibson 2017; Z. Y. Wu and Rahman 2017; Ozturk et al. 2020; Yura et al. 2018; Fischer and Krauss 2018; Barzegar, Aalami, and Adamowski 2020).

2.4.3. Machine learning in the water sector

Hydroinformatics is a domain in the water sector that concentrates on applications of high technologies and Artificial Intelligence to address problems regarding water issues. With the ability to store large amounts of historical data, the use of machine learning methods was successfully applied in different domains in the water sector including flood prediction, flood flow forecasting, water resources, water quality, wastewater overflow, water treatment optimisation etc. In this section, some research papers, where machine learning approaches were applied, are presented.

The use of ANN in various water related problems has been very popular since the beginning of the 90's (W. Wu, Dandy, and Maier 2014). In their study Wu et al. (W. Wu, Dandy, and Maier 2014) developed a protocol for comparing various ANNs and, based on this protocol, they reviewed 81 past papers where ANNs were applied in river water quality modelling. A later study by Ahmed et al. (Ahmed et al. 2019) compared 3 different ANN models for the prediction of the water quality deterioration in Johor River, Malaysia. They managed to accurately predict the behaviour of various parameters in the river and thus improve its quality. In flood modelling, an ensemble network with 10 ANNs was used in flood frequency analysis by Shu and Burn (Shu and Burn 2004) and it was proved to be more accurate in flood estimation and less sensitive in the choice of initial parameters than a single ANN. Kim and Seo (Kim and Seo 2015) developed an ensemble ANN model with exploratory factor analysis (EFA) for the 1-day ahead streamflow forecasting. This model was applied in three stations in South Korea and results indicated that this could be a really good tool for 1-day stream prediction for balanced datasets.

ANNs were also applied in various research studies in the water and wastewater systems for the prediction of the behaviour of water quality parameters, understanding and predicting burst events in the DWDS, and monitoring treatment processes. Gibbs et al. (Gibbs et al. 2006) made a comparison between linear regression, multi-layer perceptron (MLP) and general regression ANN (GRANN) to predict chlorine concentrations in two different locations of a DWDS in Adelaide, Australia and found that MPL outperformed the other data-driven techniques. They also indicated the importance of identifying the best parameters for improving the models' performance. Two different studies applied ANNs for the development of an online artificial intelligence system for the prediction of bursts at a DMA level; the first one used past flow timeseries data from 144 DMAs and a hybrid ANN fuzzy inference system (ANFIS) (S. R. Mounce, Boxall, and Machell 2010) and the second one collected the pressure and flow timeseries measurements to use them as inputs to a hybrid event recognition system (ERS) to provide as output the probability of an almost real-time burst occurring in a DMA (Romano, Kapelan, and Savić 2014). ANNs were applied in combined sewer overflows (CSOs) using rainfall data taken from radar devices and CSOs' depth timeseries for the prediction of the CSO's depth up to an hour ahead (S. R. Mounce et al. 2014). As regards the treatment process, extreme learning machine coupled with radial basis function (RBF) ANNs was used for the improvement of the coagulation process in a WTW in Malaysia and predicted the coagulation dosage with a correlation coefficient of at least 0.8 (Jayaweera, Othman, and Aziz 2019). Finally, Kazemi et al. (Kazemi et al. 2018) developed an ANN with the nonlinear autoregressive exogeneous algorithm (NARX) which, in combination with data preparation models, could predict turbidity behaviour in water distribution trunk mains up to 8 hours ahead.

A very popular in the water sector data-drive regression tool named Evolutionary Polynomial Regression (EPR) was developed by Giustolisi and Savic (Giustolisi and Savic 2009). This tool was firstly used for the prediction of groundwater levels in relation to monthly rainfall data (Giustolisi and Savic 2009). Later, this tool was used for other water related problems such as main bursts prediction in relationship with weather conditions (Laucelli et al. 2014) and as a discolouration rate predicting tool in DWDS (S. R. Mounce et al. 2016).

Decision trees and ensemble decision trees are also popular in hydroinformatics for classification and regression problems. Harvey et al. (Harvey et al. 2015) applied a decision classification tree to identify the factors that pose the water quality of small DWDS at risk and predict future water quality deterioration in 158 small DWDS in the province of Ontario, Canada. Alfonso et al. (Alfonso et al. 2018) proposed a combined decision tree model with the prospect theory on decision making under uncertainty, to create a tool that automates the decisions for flood early warning. Mounce et al. (S. R. Mounce et al. 2017) used the RUSBoost boosting ensemble decision tree algorithm to predict iron failures in DMAs of a water company in the UK and achieved 80% of correct iron failure predictions. Random forest classifier algorithm was successfully applied to CCTV footage for the classification of sewer pipe faults (Myrans et al. 2018) and to smart metering datasets for the short-term water demand forecast (Xenochristou, Kapelan, and Hutton 2020; G. Chen et al. 2017).

Unsupervised learning methods were utilised in various research areas in the water sector for evaluating parameters, creating natural groupings, and discovering correlations in complex datasets. DB-SCAN algorithm was used for clustering customer complaints in a DWDS (S. Mounce et al. 2012). K-means was used in combination with a particle swarm optimization algorithm for reducing the number of the parameters that are required for the calibration of hydraulic models (Freitas et al. 2017). A novel clustering approach with the use of t-SNE algorithm was proposed by Mounce (Stephen Mounce 2018) for separating smart water network daily flow data into clusters of residential and commercial customers. PCA was used in plenty of research papers, including modelling the performance of wastewater treatment plants (WWTP) (Abba et al. 2020) and understanding the factors that increase the energy consumption of the water distribution mains (Hashemi, Filion, and Speight 2018).

SOM algorithm applications on water quality datasets are very popular. The application of SOMs in water resources problems were summarized in a review paper by Kalteh et al. (Kalteh et al.,2008). A comparison of the SOM technique against other clustering methods (PCA and cluster analysis) appears in a research paper that examines the river water quality with the use of water quality monitoring data and in which SOM clustering outputs outperformed the outputs of the other two methods (Astel et al. 2007). The first approach in applying SOMs in a drinking water quality related problem was made by Chang et al (Chang et al., 2011). In this work, SOMs were used in combination with K-means and Fuzzy c-means for the clustering of the water mains of a large DWDS in different water quality categories. However, for this study Chang et al. instead of using actual measured WQ data, they used the WQ data as calculated by the hydraulic model. Conversely, there are significant papers in the literature that apply SOMs as their main tool or one of their main tools in measuring drinking water quality data for understanding drinking water quality and identifying correlations between various water quality parameters in DWDS. More specifically, Mounce et al. (S. Mounce et al. 2012) used SOMs in combination with PCA for understanding the relationships between bacteriological species and the chemical and physical characteristics of the drinking water inside a laboratory pipe rig; Blokker et al. (E. J. Blokker et al. 2016) applied SOMs in two water quality datasets, one taken from a UK DWDS ,and one taken from a Dutch DWDS, to identify correlations between water quality deterioration, temperature of water and the age of water on customer taps; Mounce et al. (S. R. Mounce et al. 2016) used SOMs in combination with EPR for the estimation of the rate of discolouration in DWDS in the UK; Ellis et al. (Ellis et.al., 2015; Ellis et al., 2014) utilised SOMs in combination with cross-correlation algorithm to identify the factors that caused coliform failures in two different WTWs; Speight et al. (Speight, Mounce, and Boxall 2019) applied SOMs in monitoring water quality samples datasets of three different UK WUs for understanding the factors that cause discolouration of the drinking water.

Various studies in the water sector used ensemble of machine learning techniques or compared the performance of two or more machine learning techniques. Research by Nourani et al. (Nourani, Elkiran, and Abba 2018) compared three different ensemble techniques that combined support vector machines (SVM), ANNs, fuzzy logic and multilinear regression for the prediction of the performance of a WWTP in Nicosia, Cyprus. Mounce et al. (Mounce, Mounce, and Boxall 2011) compared the performance of a support vector machine (SVM) model with a previously developed ANN model in detecting anomalies in water flow and pressure timeseries data in water distribution networks (WDNs) and found that the SVM model outperformed the ANN model on providing alert over sudden changes in the network. Another study on anomaly identification on WDNs, compared SVMs with clustering methods, and also found that SVM had better performance in predicting anomalies (Vries et al. 2016). In the same subject, Carreno - Alvarado et al. (Carreño-Alvarado et al. 2017) compared the performance of SVM with the performance of Relevance Vector Machines (RVM) - a Bayesian approach on SVM - and suggested that RVM could be also a suitable tool for leakage detection. SVM appears to be the most accurate tool in predicting future water demand followed by random forests in a work held in a city in the south of Spain (Herrera et al. 2010). Finally, in an interesting research work on turbidity forecasting in water distribution trunk mains, Meyers et al. (Meyers, Kapelan, and Keedwell 2017) compared the regression and classification performance of SVM, random forest, and feedforward ANN in predicting turbidity behaviour and turbidity events up to 5 hours ahead and indicated that random forest outperformed the other methods. However, they, also, indicated that as regards the classification model, the longer ahead the prediction is, the higher the number of false positives is.

Lastly, in the recent years that deep learning neural networks and deep learning applications are receiving a lot of attention from the scientific world, studies that apply DNNs in various water related projects can be found in the literature. Most of these research works are related with areas where telemetry systems generate hundreds of thousands of data such as WWTPs monitoring systems. The first study that indicated the effectiveness of deep neural networks compared to other forecasting models, has used deep convolutional neural networks (Deep CNNs) for the prediction of the daily water flow and water level on a catchment of a river in Ireland (Assem et al. 2017). A novel approach that combines hydraulic modelling with DenseNet, a deep learning algorithm, was tested in two different WDNs for the localisation of synthetic pipe bursts (Zhou et al. 2019). Deep reinforcement learning was successfully applied in two water related studies, one for the optimization of the performance of a pumping station inside a WWTP (Filipe et al. 2019) and one for scheduling valves in a DWDS using sensor data and programmable logical systems (PLC) systems for minimizing contamination inside the WDNs by isolating contaminated areas the network (C. Y. Hu et al. 2019). The former decreased the pumping station energy consumption by up to 16.7% and the latter managed to minimise contamination in a WDN, even when multiple contamination source events occurred. Finally, two different studies concentrated on solving problems and improving the process performance of two different WWTPs. More specifically, Dairi et al. (Dairi et al. 2019) used a combined RNN and restricted Boltzmann machine (RNN-RBM) model in combination with classification algorithms for the identification of abnormal influents entering a coastal WWTP in Thuwal, Saudi Arabia, and Mamandipoor et al. (Mamandipoor et al. 2020) applied long short-term memory networks (LSTM) in a WWTP in Treviso, Italy for detecting faults in 12 different sensors during the nitrification and oxidation process and achieved a fault detection rate of up to 92%.

This thesis focuses, as mentioned in the previous chapter, on ML applications for the improvement of drinking water quality. Therefore, to summarise the work that has been done in the field, the research works mentioned in the previous paragraphs that applied ML methods for improving drinking water quality, the problems that addressed and their results are presented in table 2-1.

Table 2-1: Overview table of the research works that apply ML applications on drinking water quality problems

Study title	Publication Reference	Addressed water quality problem	ML method used	Case study Available data	Outputs (pros/cons)
Investigation into the relationship between chlorine decay and water distribution parameters using data driven methods	(Gibbs et al. 2006)	Prediction of chlorine decay in the DWDS and chlorine concentrations in customers taps	- Linear regression model - MPL ANN -GRNN	A DWDS in Hope Valley, South Australia. **Inputs:** Flow, Cl_2 temperature, DOC and UV in the WTW, temperature in the network **Output:** Cl_2 in some points in the network	- ML based models useful when the network hydraulics are not known. -ML models require higher frequency data than the available for more accurate results. -Finding the appropriate inputs increases prediction accuracy
Water quality comprehensive evaluation method for large water distribution network based on clustering analysis	(Chang et al. 2011)	Evaluating water mains in a DWDS based on WQ parameters	-SOMs for correlations identification -K-means and fuzzy c-means for clustering the mains	A large DWDS with available hydraulic model. **Inputs:** Residual chlorine, water age, THMs, TOC and other parameters - all calculated by the hydraulic model	-Clustering of mains in WQ categories based on multiple parameters -ML model in this study uses non measured data (WQ parameters were calculated)
A bio-hydroinformatics application of self-organizing map neural networks for assessing microbial and physico -chemical water quality in distribution systems	(S. Mounce et al. 2012)	Assessing microbiological water quality characteristics in DWDS	-PCA for reducing the T-RFLP profiles dimensionality. -SOMs for correlations identification	A testing loop facility that simulates DWDS. **Inputs:** T-RFLP outputs, WQ parameters	-SOMs provides a visualised output of the microbiological behaviour of the testing loop.
Relating Water Quality and Age in Drinking Water Distribution Systems Using Self-Organising Maps	(E. J. Blokker et al. 2016)	Relating microbiological water quality parameters with water age and temperature in DWDS	SOMs	A Dutch and a UK real DWDS and a testing loop. **Inputs:** drinking water quality parameters and modelled water age	- Temperature and water age are independent parameters -Temperature influences more the microbiological

					activity than water age
Multivariate data mining for estimating the rate of discolouration material accumulation in drinking water distribution systems	(S. R. Mounce et al. 2016)	Estimation of discolouration accumulation rates in DWDS	-SOMs for the data mining -EPR for creating mathematical models' expressions	A UK national dataset collected during flushing operations in 36 different DWDS and two small local DWDS in the Netherlands. **Inputs:** WQ parameters, mains characteristics, flushing outputs **Outputs:** Discolouration accumulation rates	-SOMs indicated that in the UK case study, high accumulation rate is related to high iron in the bulk water, non-plastic pipes, iron coagulation treatment and unlined cast iron mains. -EPR models could be used for assessing iron related parameters to reduce accumulation of materials in the DWDS -No temporal relationship between accumulation rates and WQ parameters could be identified with these two approaches
Improving root cause analysis of bacteriological water quality failures at water treatment works	(K. Ellis et al. 2015)	Understanding the roots of coliform bacteria failure in a WTW	SOMs	A WTW of a water utility in the UK. **Inputs:** WQ parameters taken from various locations in the works and in different frequency.	-Coliform bacteria in the WTW exit related to low turbidity and low disinfectant residual in the works -Heavy rainfall also a potential parameter that influences WTW failure -Study could not provide a definitive cause of bacteriological failure in the works

Identification of the Causes of Drinking Water Discolouration from Machine Learning Analysis of Historical Datasets	(Speight, Mounce, and Boxall 2019)	Understanding the factors that influence discolouration in DWDS	SOMs	A full WQ samples and physical characteristics dataset of a water utility in the UK, a WQ samples physical characteristics dataset for 3 cities that belong in a different water utility and a full WQ samples dataset in a DMA level of third company.	-SOMs is a great tool for visualising corelations between various parameters -This paper opens the discussion over the use of data mining in sparse datasets such as the WQ samples datasets. - SOMs outputs indicate that the risk of discolouration is not necessarily related to the bad condition of iron mains -SOMs gives great visualisation of the mechanisms that could result in iron release and discolouration in DWDS
Using data mining to understand drinking water advisories in small water systems: A case study of Ontario first nations drinking water supplies	(Harvey et al. 2015)	Identifying the factors that cause preventive measures to protect public health from contaminated water in small. DWDS.	Decision tree	A dataset that contains information from the type of the water, the status of the operators, the age of the DWDS, the pipe length the WTW quality etc, from 158 small drinking water systems in Ontario Canada. Outputs: the status of the DWDS (is at risk or not at risk of contamination)	-The model correctly predicted 71% of the systems -Systems that use groundwater source are safer -Systems that are maintained by untrained operators have higher risk to fail -Main aim of this work is to demonstrate that decision trees are user-friendly tools -The model requires further investigation with more data to increase accuracy

Short-term forecasting of turbidity in trunk main networks	(Meyers, Kapelan, and Keedwell 2017)	Forecasting turbidity up to certain hours ahead to aid operational stuff and enabling proactive interventions in DWDS	-Random Forest -Feed-forward ANN -SVM	The case study area is a trunk main that serves 5 DMAs with water. Sensors to measure flow and turbidity are installed in the main. **Inputs:** Flow measurements up to certain hours back, Flow peaks and turbidity peaks **Outputs:** 1.Turbidity values up to certain hours ahead (regression) 2. Turbidity threshold up to certain hours ahead (classification)	-High classification accuracy up to 5 hours ahead -Low regression accuracy achieved by the model (30 minutes ahead) - This model could be useful tool for turbidity predictions in systems with sufficient past turbidity events -If there are not enough turbidity events in the dataset there is a high risk of getting a large number of false positives and low overall accuracy -RF was the best out of the three ML models in this case study
Predicting turbidity in water distribution trunk mains using nonlinear autoregressive exogenous artificial neural networks	(Kazemi et al. 2018)	Predicting turbidity events in water distributions trunk mains for reducing discolouration risk in drinking water	-NARX ANN -Feed-Froward ANN	Model tested in two trunk mains in different DWDS. **Inputs:** Flow measurements in flow events related to turbidity events Risk parameter **Outputs:** Turbidity events	-The model predicted turbidity events up to 10 hours ahead with MAE 0.05NTU error. -NARX ANN was a better method than Feed-forward ANN -The model is a good tool for predicting turbidity and providing proactive information to operators -Model's effectiveness limited when not enough discolouration events available for training

Ensemble Decision Tree Models Using RUSBoost for Estimating Risk of Iron Failure in Drinking Water Distribution Systems	(S. R. Mounce et al. 2017)	Estimating DMAs' risk of iron failure one year ahead to provide water utilities with a DMA risk ranking list that will prioritize their interventions in the DWDS	RUSBoost boosting trees	Case study was a water utility in the UK where their WQ dataset over a period of 7 years was used. **Inputs:** average median iron, manganese, turbidity values from all samples per year per DMA, customer complains per year per DMA, percentage of iron mains per year per DMA **Outputs:** DMA class (failure/non failure) per year.	-Model predicted 60.5% of the high-risk DMAs and 76% of the low-risk DMAs in the upcoming year -RUSBoost is a great tool for classification when the dataset is unbalanced (in this case, very low percentage of iron failures in the dataset) -Model creates large number of false positives due to the imbalance in the dataset
Improved predictive capability of coagulation process by extreme learning machine with radial basis function	(Jayaweera, Othman, and Aziz 2019)	Optimising coagulation process in a WTW to improve drinking water quality	Extreme learning machine with RBF	Segama WTW, Sabah, Malaysia. **Inputs:** pH, turbidity color TDS, alkalinity in raw water and color, TDS, alkalinity, in treated water **Outputs:** coagulant dosage	-Model provided better results and with less computational time than the ANN -Low turbidity water required only 3 parameters for the optimization of the coagulant dosage -High turbidity water required 4 parameters for the optimization of the coagulant dosage -ELM-RBF good tool for coagulation optimization however data quality could effect model's performance

Machine learning approaches to predict coagulant dosage in water treatment plants	(K. Zhang et al. 2013)	Predicting coagulant dosage in a WTW	-K-Nearest Neighbours -SVR	4 WTWs, 2 small ,1 medium and 1 large were used as case studies in this work **Inputs:** pH, temperature and turbidity in the coagulation tank **Output:** Alum dosage	-KNN outperformed SVR in the small WTWs -KNN and SVR had similar performance in the medium and large WTWs -Performance improvement of both methods is highly depended on the quality of the datasets as poor quality data influence both ML methods
Random forest tree for predicting faecal indicator organisms in drinking water supply	(Mohammed, Hameed, and Seidu 2017)	Predicting fecal indicator microorganism in the source water of a WTW	Random forest	Water quality data taken from raw water used as a source for a WTW in Bergen, Norway **Inputs:** conductivity, pH, colour, turbidity and season of the year **Outputs:** Coliform bacteria and E-coli	-RF is vital tool for the prediction of faecal microorganisms -Feature importance outputs indicated that colour and seasonality are the most important parameters that influence the results

2.5. Conclusions

DWDS are complicated systems. Bacteria regrowth in the DWDS could be promoted, as presented above, due to various factors including disinfection residual, pipe material, network condition and maintenance, WTW performance etc. Aging and poorly maintained DWDS could be the main factors of water discolouration but, as various studies indicate, there are many other factors that could cause accumulation and mobilisation of material inside the WDNs. Therefore, it is very difficult to understand the roots that cause water deterioration by just taking water samples periodically for water quality testing.

Methods that have the potential to give a better understanding of the reasons for water deterioration or that could potentially predict future water behaviour and deterioration events are very useful tools for decision makers inside WUs. Machine learning applications, as presented above, were, successfully, applied in various research projects in the water sector, where enough data were available. As WUs are storing water quality data and lots of other types of data from various sources and various points inside their DWDS, large datasets of various parameter measurements are created that could be a very good input for training machine learning models. However, the selection of a machine learning method depends on the data availability and the type of the water quality problem that is required to be solved, which is a gap that the current research works have not answered yet. In addition, there are certain areas in the DWDS where available water quality data exist and no work using data-driven models is made and, therefore, some water quality research questions have not been answered yet. More specifically, the knowledge gaps that this thesis aims to address are described below:

- In the research works where ML applications are applied, the common practice is to create the data-driven model that could solve a certain WQ problem using the available data given by WUs. However, none of the studies aimed to explain which type of ML method is the most appropriate for a certain WQ problem and for the available data. In other words, there is no clear guidance for WUs on applying ML methods for supporting the management of their DWDS. Therefore, to fill this gap, this thesis provides a ML selection process which is presented in chapter 3 and 10.
- WUs are collecting large number of water quality data from their systems. However, there are no papers in the literature that provide a holistic approach regarding the storage of these data, the necessary information that WUs should include in these, and how this information could be connected and integrated to facilitate the application of ML methods for improving drinking water quality. The big data framework, presented in chapter 10, fills this gap, and proposes certain steps that WUs should follow to facilitate data-driven applications using their data.
- There are no research works that aim to identify the factors that increase bacteriological activity in service reservoirs or identifying potential service reservoirs that are at risk of a future failure. This thesis fills this gap by undertaking two data-driven investigations in service reservoirs, one for identifying the factors that increase bacteriological activity in the service reservoirs (chapter 5) and one for predicting service reservoirs that are at risk (chapter 7).
- There are no data-driven research studies that focused on the potential impact of a disinfection switch in drinking water quality. The Self-organising maps investigation in chapter 6 aims to fill this gap.
- The use of ML applications for predicting future turbidity values to prevent discolouration events in trunk mains is explored in at least two different research works. However, it is important to investigate the potential of these models in predicting other parameters that could help WUs reduce the risk of bacteriological risk in their systems. An approach that partially fills this gap is presented in chapter 8

where a predictive data-driven model is used to predict Chlorine losses at the end of 3 distribution trunk mains.

- Over the recent years, bacteriological sensors that measure total cell counts have become a popular tool for recording the bacteriological activity in the WTWs. Potential analysis of these data using ML applications for understanding WTWs' bacteriological activity and predicting future bacteriological behaviour has not been made so far. This area of research is investigated in chapter 9.

3. Machine learning techniques selection for drinking water quality problems and performance metrics for their evaluation

3.1. Introduction

In chapter 2, machine learning (ML) was defined as the computer science domain that constructs algorithms that could learn how to perform a task through past examples (past data). The ML categories and main techniques were also presented. The selection of the appropriate ML technique, though, is fully dependent on the problem that requires to be solved, the type and the amount of the available data and the required output.

In addition, for the water utilities (WUs) the collection of the available data, usually stored in different data warehouses, is a more complicated task. More specifically, the complete collection of all the available data requires the collaboration of different WUs departments that own parts of the data and a huge amount of effort and time for their integration into a unique dataset. As an example, the following chapter presents in detail the effort and all the required steps to create the datasets that were then used as inputs in chapters 5,6 and 7. In Chapter 10, a holistic approach for the application of big-data analytics tools is proposed. This is presented in the form of a framework and includes recommendations regarding the data storage and integration that decreases the amount of time required for the application of ML methods on water quality related problems. This chapter focuses only on the required steps for the selection and application of ML methods without taking into consideration the data storage and integration. More specifically, the aim of this chapter is to propose the ML application steps, from the setting of the water quality problem to the ML output evaluation, that WUs should follow to tackle drinking water quality problems in their DWDS, using only the data that have already been collected in their routine monitoring program. Thus, this chapter addresses the first objective and partially to the fifth objective of this thesis.

3.2. Background

Tom Mitchell in the introduction of his book *Machine Learning* defined ML algorithm as follows: "*a computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E*" (Mitchell 1997). A small example to show how this formalism could be used is the following:

All the employees of a company are receiving a significant amount of spam emails in their professional email accounts including emails that could harm their company's computers. In order to solve this problem with data-driven techniques, it should be redefined as follows: "Could spam emails be identified and sent to the spam folder?". The computer program in this case is the selected ML application and the task T, the performance P and the experience E are defined as follows:

- Task (T): Classify an email into two categories - spam and not spam.
- Experience (E): A large number of emails where some of them are spam and some are not spam to be used as the input data.
- Performance (P): the accuracy of the classification prediction as applied by the machine learning technique.

In addition, the ML model task implementation procedure usually follows these steps (MathWorks 2016):

a. *Collection and understanding of the data*: In the real world, datasets are not perfect. There is messy, noisy, and incomplete data in various formats. The first step before selecting a ML model is to understand the type of the raw available data and to define the hidden information that needs to be uncovered.
b. *Data preparation*: This step includes all the required activities for constructing the final dataset that will be used as the input to a ML model. Cleaning, interpolating, or transforming the data and selecting features are just some of the activities that are included in this step.
c. *Model selection & training*: This is the step where an ML method is selected and trained. The model selection is highly related to the type of desired output and the data format. The training of a model is the procedure that is followed in order to assist the model to understand the data and learn from the data.
d. *Model evaluation*: In this step the performance of the trained model is compared to existing data. The aim is to evaluate the model and the procedure that was followed in the previous steps.
e. *Model improvement*: It may be necessary to change some of the features, steps and either simplify or add complexity to the model. All these decisions are part of the model improvement step. If the model evaluation is satisfactory, this step could be skipped.
f. *Model deployment*: The improved model is now ready to be applied for problem solutions.

For generating the machine learning selection and application steps for water quality problems, Mitchel's formalism in combination with the ML implementation procedure are transformed into steps that should be followed to tackle drinking water quality deterioration problems. In the rest of this chapter, the machine learning selection and application steps, the methodology of the ML techniques that were investigated in the thesis, and the performance metrics used for the evaluation of the techniques are presented.

1. Setting up the machine learning application steps for water quality problems

These proposed steps work as a guide on how to define the water quality problem, to work with the available data and specify the required output. Each step is necessary for the procedure of the machine learning application. The selection of the appropriate ML techniques occurs in step 4 and is facilitated by the use of the *machine learning selection tree* The application steps are as follow:

1. Step 1: Definition of the water quality problem

The spam email example in the previous section indicates the importance of the definition of a problem that needs to be solved. Data-driven techniques could be used to tackle specific problems that are only data related. Therefore, the water quality problem should be a problem that could potentially be solved using historical water quality data or any other available data.

2. Step 2: Clarification of the required output (Task T)

As the spam email indicates, after the water quality problem definition, the required output could be defined. The required output is what Mitchell defines as Task T and it could be a classification output, a prediction of future water behaviour, a correlation between water parameters or clustering of unlabelled and unstructured data.

3. Step 3: Specification of the available dataset (type, amount etc. - Experience E)

What Mitchell defines as experience E in water quality problems are the available water quality data, unstructured data, asset data etc. to train the model. These data could be water quality monitoring data, telemetry data, water quality timeseries data measured for specific investigations. In addition, some data regarding specific characteristics of the assets (pipe age, pipe material, type of disinfection, number of properties etc.) could also be used if they could improve the ML models' performance.

4. Step 4: Selection of an ML technique

The selection of an ML technique is highly dependent on the previous 3 steps. As mentioned above, the appropriate ML technique for solving the problem could be selected from the *machine learning selection tree*. For creating this tree an investigation on some of the available ML techniques is required and which is also one of the objectives of this thesis (objective 1). The techniques that are investigated in this thesis are presented in the following section. Most of these ML techniques were applied in real WQ case studies presented in the following chapters. However, k-means and t-SNE, even if they were investigated over their abilities, were not used in any of this thesis' case studies.

5. Step 5: Preparation and pre-processing of the data

Data preparation and pre-processing refers to all the techniques that are used for changing the data format, adding extra data to fill the missing values or removing the outliers (Kuhn and Johnson 2013). Data preparation is a very important step for the accurate application of the ML technique. The type of data preparation required for each ML technique is also included in the machine learning investigation. Explanation of the data preparation required for each specific technique is included in the machine learning selection tree.

6. Step 6: Machine learning application output

This step refers to the training of the selected ML technique, the types of outputs that it produces, and the testing of its performance. In other words, in this step the ML technique outputs are produced in forms of graphs, tables etc., and then, the performance P of the technique, as defined by Mitchell, is examined.

The machine learning application steps are summarized in the following figure

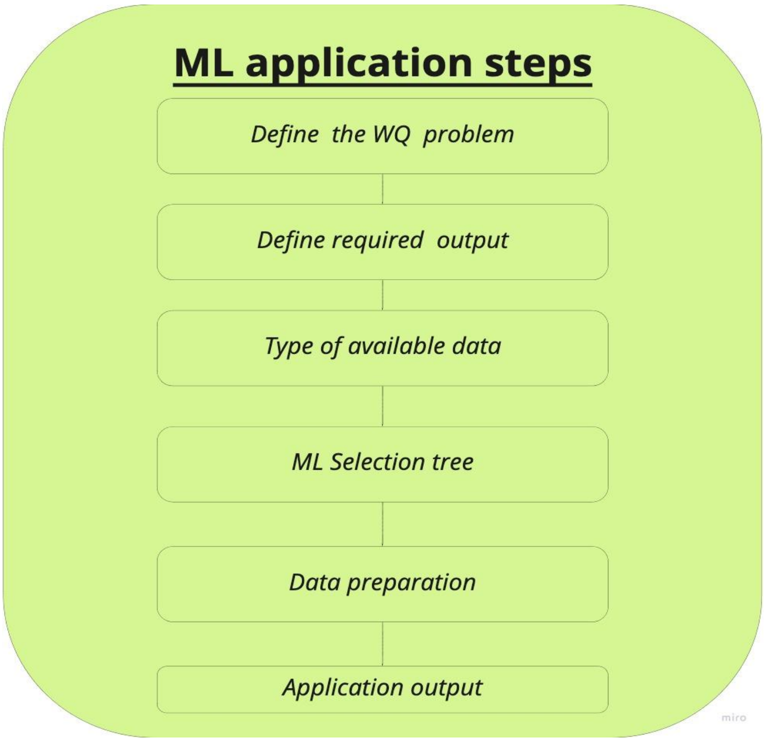


Figure 3.1:Machine learning application steps

3.3. ML techniques and selection tree

The ML techniques that are picked for this thesis are selected based on covering the ML categories (supervised, unsupervised learning etc.), as described in the literature review chapter, and on the basis of covering important water quality problems that Scottish Water (SW) has to tackle. Therefore, the ML techniques selected for this investigation could be divided into three main groups: methods for understanding the water quality behaviour of the DWDS and the factors of water quality deterioration (correlation - association unsupervised learning techniques), methods for clustering of water quality data into groups with similar behaviour (clustering unsupervised learning techniques) and methods for predicting future water quality behaviour (predictive regression and classification supervised Ml techniques). The aim of this research is to both investigate the performance of the most common methods, some of which are already applied in other research projects in the water sector, as described in the previous chapter, and the performance of new ones on water quality problems. The selected methods are presented in the remainder of this section. In the following chapters some of these techniques are applied in real world water quality case studies.

3.3.1. Clustering unsupervised machine learning techniques

Clustering methods in water quality could be used for dividing monitoring water quality samples into different categories with similar behaviour, for example samples with high or low chlorine concentration in certain areas of a DWDS. Clustering could also be used for identifying and visualising the main groups of smart meter datasets, timeseries datasets used for water quality investigations in certain areas. In this thesis two methods are selected, k-means the most common clustering method, and t-distributed stochastic neighbour embedding (tSNE) a new technique mainly used for dimensionality reduction, visualisation of the data and for group clustering.

3.3.1.1. k-means

k-means is the most used clustering method for data mining due to its simplicity. This algorithm separates the input datasets into a k number of clusters where every single input belongs in the cluster with the closest mean value (Maimon and Rokach 2006). Therefore, a cluster is a group of data whose distances are smaller than their distances with data that belong to other clusters (Bishop 2006). The number of clusters k should be defined by the user before the application of the method. Once the number of clusters is defined, the distance between the centre of the cluster (mean value) and the actual observation, using the Euclidean distance, is calculated as follows:

$$d = \sum_{n=1}^{N} \sum_{k=1}^{K} ||x_n - \widehat{\mu_\kappa}||^2$$

Where:

N: Total number of observations

K: total number of clusters

x_n: the nth observation of the dataset

μ_K: the centre of the kth cluster

d: the Euclidean distance of each observation to the centre of each cluster μ_K

The aim is to categorize each datapoint to each cluster so that d is minimized. Therefore, once the initial distance d is calculated, the average of the observations (data points) that belong in the same cluster is computed to obtain the new μ_{κ} of the cluster. Finally, this procedure is repeated until the clusters remain stable or when the predefined maximum number of repetitions is reached (Hastie, Tibshirani, and Friedman 2008).

K-means is easy to implement and separates the data in robust clusters once the k is defined. However, in some cases where complex non-linear relationships exist, it is difficult to select the number of clusters a dataset should be divided, and thus the k-means may not be the appropriate technique for that. In addition, k-means requires an input dataset with no missing values for each selected variable, therefore it could not be applied to water quality monitoring samples (discrete samples) dataset where in each sample (observation) just few of the available variables (water quality parameters) are measured. However, in the water sector k-means was used as a simple tool for grouping pipes to reduce calibration time of the hydraulic models (Freitas et al. 2017) and as a clustering methodology that categorised the mains of a DWDS based on a number of water quality indicator parameters (Chang et al. 2011).

3.3.1.2. t-distributed stochastic neighbour embedding (t-SNE)

t-distributed stochastic neighbour embedding (t-SNE) is a ML technique used for clustering and visualizing high dimensional datasets (van der Maaten and Hinton 2008). The idea of this methodology is to group high dimensional datasets to low dimensions where nearby points are embedded to nearby points in the low dimensional space, and the long-distance points are embedded to distance embedded points in the low dimensional space. The methodology follows the steps of SNE algorithm firstly developed by Hinton and Roweis (Hinton and Roweis 2002) but instead of using a Gaussian distribution for the calculation of similarity between two points in the low dimensional space, it uses the Student t-distribution. The t-SNE algorithm briefly follows these steps:

a. The distance between the various points of the dataset is calculated. The Standard Euclidean distance is the most used distance metric.
b. The standard deviation σ_{ι} for each row i of the dataset is calculated.
c. The conditional probability $p_{j|i}$ is calculated. The conditional probability $p_{j|i}$ is defined by van der Maated and Hinton (2008) as "the probability that x_i would pick x_j as its neighbour in the low dimension if neighbours are selected in proportion to their probability density under Gaussian centred at x_{κ}" (van der Maaten and Hinton 2008). $P_{i|j}$ is calculated as follow:

$$p_{j|i} = \frac{e^{\left(\frac{-\left|\left|x_i - x_j\right|\right|^2}{2\sigma_i^2}\right)}}{\sum_{k \neq i} \quad e^{\left(\frac{-\left|\left|x_i - x_k\right|\right|^2}{2\sigma_i^2}\right)}}$$

d. The perplexity parameter is defined. The perplexity measures the number of the efficient neighbours to point x_i. t-SNE performs a search over the standard variation to fix the perplexity for each point x_i.

e. An initial set of low dimension points is created.

f. The conditional probability (similarity) q_{ij} is calculated. As mentioned above in the t-SNE the similarity is calculated using the t-distribution and therefore, is calculated as follows:

$$q_{i|j} = \frac{(1 + \left|\left|y_i + y_j\right|\right|^2)^{-1}}{\sum_{k \neq l} \quad (1 + \left|\left|y_k + y_l\right|\right|^2)^{-1}}$$

g. The gradient of Kullback-Leibler divergence between the Gaussian distribution in the high-dimensional dataset and the low-dimensional space is calculated.

h. The low dimensional space is updated, and the steps f and g are repeated until the optimum Kullback-Leibler divergence is achieved.

By following the above algorithm t-SNE plots a high dimensional dataset in a low dimensional space, groups the data into robust clusters and allows visualisation of complex datasets without predefining the number of clusters. However, t-SNE requires an input dataset with no missing values, each line with missing values is removed and therefore, as in k-means, this method could not be applied in gram monitoring water quality datasets. In the water sector this technique was successfully applied as a clustering and visualisation method of smart water meter data into commercial and residential customers (Stephen Mounce 2018) and as a tool for clustering the influents of various WWTPs in order to optimize their processes and improve their performance (Xu et al. 2021).

3.3.2. Correlation unsupervised machine learning techniques

The 2 techniques presented in this section are, as the above, unsupervised ML techniques mainly used for visualisation. In addition, though, the methodology followed by these techniques allows an understanding of correlations between the various variables measured for each observation. Therefore, these methods could be applied on water quality data for identifying the correlations between the various water quality parameters and understanding the factors of water quality deterioration. More specifically, the methods that were investigated in this thesis are Self-Organised Maps (SOMs) a clustering technique, also used for visualisation of the data, and Principal Components Analysis (PCA) a method that is mainly used for dimensionality reduction of high dimensional datasets and feature extraction.

3.3.2.1. Self-Organized Maps (SOMs)

Self-organizing Maps (SOMs) are a type of unsupervised ANN proposed firstly by Kohonen (T. Kohonen 1990). It is a clustering methodology for visual data mining and exploration and has the properties of both vector quantisation and vector projection algorithms (S. R. Mounce et al. 2016). As an unsupervised learning method, SOM does not require to know the relationship between the input variables. There are more than 10000 scientific papers that applied SOMs algorithm in various fields including financial applications, biomedical applications, telecommunications, industrial control, engineering, and genetics (Teuvo Kohonen 2014). In recent years, SOMs have been, successfully, applied in the water sector as well (see machine learning in the water sector section of the previous chapter).

SOMs are composed of two layers, the input layer which contains a number of neurons equal to the number of imported by the user variables and the output layer which is a 2D colour-coded rectangular map that contains a certain number of hexagonal cells. Each cell has an associated weight vector that connects it with the input neurons and an associated weight vector that connects it with its neighbours. The steps that SOM algorithm is following are as follows:

a. The user selects the input nodes, and the input vector is created
b. A weight vector of the output neuron with equal nodes as the input vector is created. The weight vector at the beginning of the training process has random values.
c. The weight vector is initialised either randomly or by using initialisation algorithms
d. The distance between each input and each of the weights is calculated - usually the Euclidean distance is chosen.
e. The output neuron with the smaller distance is the selected (winning) neuron.
f. The influence of the winning vectors to its neighbouring neurons is calculated
g. The weight vector is updated to become more similar to the input vector

This algorithm is repeated for a number of times while the learning rate is decreased. Then, the winning neuron and its neighbouring neurons become almost similar to the input nodes and at the end, a well-trained SOM is created (Chang et al. 2011). The missing input values are ignored by SOM during the distance calculation procedure and during the update weight vector stage, these values are replaced by utilising an imputation SOM as described by Vatanen et al.(Vatanen et al. 2015). SOMs' ability to ignore the missing values during the training process, made this algorithm a very popular method for clustering and visualization. The output 2D map consists of hexagon cells, each of which are associated with a weight vector. For each input variable of the dataset a different component plane is created. Each plane is presented as a grid of colour hexagon cells where red cells contain the high values and blue cells contain the low values of this specific input variable. Therefore, by comparing the values of the different component planes, correlations between the variables or between portions of the variables could be identified (S. R. Mounce et al. 2016).

The SOM algorithm sets the size of the hexagonal cells depending on the available data. Any additional input variable will change the reference vectors and as a consequence their position in the component planes will change as well. Finally, SOM's ability to describe the data is controlled through a *U-matrix*, also included in the SOM output. This matrix is generated to reflect the dissimilarity between the weight vectors in the map. High U-matrix values indicate weaker clusters and correlations.

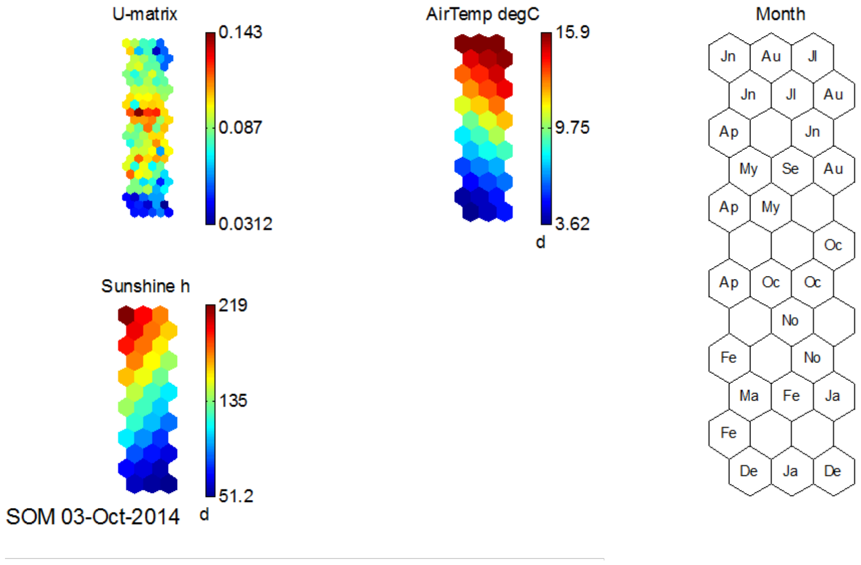


Figure 3.2:Example of SOM air temperature and hours of sunshine per month (Pennine Water Group; ARC Consultancy 2017).

A simple example of SOM output is shown in figure 3.2. This example is taken from a Sheffield University report made on behalf of Scottish Water (Pennine Water Group; ARC Consultancy 2017). In this case, a dataset that contains data for two numerical variables is used. The first variable is the hours of sunshine per day and the second is the air temperature. A third categorical variable is also presented that corresponds to the months of the year. The high air temperature values (red) are clustered in the top right of the component plane, which correspond to the top of the sunshine plane. The top right portion of the planes also corresponds to the months of June, July, and August in the labelled map, indicating a correlation between these variables. Similarly, the cluster of low hours of sunshine (in blue at the bottom of the component planes) corresponds to the months of December and January in the labelled map which are in the same location within each component plane.

SOMs were successfully applied in various scientific works in the water sector as it is a great tool for simple visualisation of multiparameter correlations. More specifically, in the water sector, SOMs were applied for correlation identification of various water quality indicator parameters in the DWDS for categorising water distribution mains (Chang et al. 2011), for understanding the factors of discolouration (Speight, Mounce, and Boxall 2019; S. R. Mounce et al. 2016) and for identifying the relationships between bacteria regrowth, temperature and age of water (E. J. Blokker et al. 2016).

3.3.2.2. Principal Components Analysis (PCA)

Principal Components Analysis (PCA) is a technique that reduces the dimensionality of a large dataset by transforming it to a small number of uncorrelated variables known as principal components (PCs) (Jolliffe 2002). Every PC represents a linear function a_n x that captures part of the variance of the dataset. X represents the vector of the variables of the dataset and a_n is the n^{th} vector which consists of a number of constants equal to the number of the various variables. The a_n vectors are called eigenvectors, representing the linear transformation of the dataset when a scalar vector, also known as eigenvalue, is applied to it. The number of both the eigenvalues and the eigenvectors is equal to the number of the variables in a dataset. The equation that describes the first PC of dataset with n number of variables is as follows:

$$a_1x = a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \cdots . + a_{1n}x_n = \sum_{i=1}^{n} \quad a_{1i}x_i$$

The number of PCs in a dataset could be equal to the number of variables of the dataset but the aim is to capture most of the variance of the dataset in the first 2-3 PCs and therefore, the dimensionality reduction is achieved. Thus, the first PC is a line that captures the maximum possible variance of the dataset, the second PC is a new line calculated in the same way as the first PC and captures the next highest variance with the only condition being uncorrelated to the first PC and so on (Jolliffe 2002).

The five steps of the PCA algorithm are explained by Smith (Smith 2002) in his university lecture notes as follows:

a. Standardization of the dataset: The standardization is required to transform the data into comparable scales and therefore minimise the dominance of larger variables over the smaller variables. For each value x_0 of each variable the standardized value z_0 is calculated as follows:

$$z_0 = \frac{x_0 - mean\ of\ the\ variable}{standard\ deviation\ of\ the\ variable}$$

b. Calculation of the covariance matrix: The covariance matrix is a symmetric matrix where its line and its column number are equal to the number of the variables. For the calculation of the covariance matrix the covariance of each variable with every other variable of the dataset is calculated. The diagonal of the matrix represents the covariance of the variable with itself, therefore it is actually the variance of that variable. The other values of the covariance matrix are symmetric with respect to the matrix diagonal.

c. Calculation of the eigenvalues and eigenvectors of the covariance matrix: Eigenvalues and eigenvectors are calculated from the following formula:

$$Su_1 = \lambda_1 u_1$$

where S is the covariance matrix, u_1 the eigenvector and λ_1 the eigenvalue. There are as many solutions for this formula as the number of the variables of the dataset.

d. Selecting principal components and creating the feature vector: The calculated eigenvectors are sorted from the ones that capture the higher variance to the ones that capture the lower variance. The principal components are the 2-3 (or more) first eigenvectors that capture most of the variables of the dataset. The feature vector is a matrix formed by the selected principal components.

e. Deriving the new dataset: The new dataset with less variables is calculated as follows:
$New\ Data = Feature\ Vector^T x\ Standardized\ Original\ Dataset^T$

Each variable of the original dataset is represented by a vector and the closer the vector to a principal component is, the more the contribution of the vector to the component. By visualising the direction of each variable, the correlations between the various variables are uncovered. Thus, PCA could be applied in datasets when the water quality problem is more related to the identification of the factors of water deterioration. However, as presented above, for the application of the PCA in datasets, the replacement of the missing values is required. The replacement could be achieved by using algorithms such as alternative least squares (ALS), but in datasets with more than 30% of missing data, the accuracy of the ALS is degraded and, consequently, PCA outputs could be unreliable. The SRs water quality investigation, presented in chapter 5, reinforces the previous argument, and indicates the inability of PCA to tackle the missing data issue on discrete monitoring water quality data. However, when the available datasets contain no missing or few missing data, PCA is a great tool for the visualisation of multi-parameters linear correlations and the dimensionality reduction of complex datasets. Thus, for these two reasons PCA was applied in the water sector in different projects. As an example of the former applications, Hashemi et al. (Hashemi, Filion, and Speight 2018) applied a PCA based model for identifying the main factors that increase the energy consumption of the water distribution mains. Finally, as an example of the latter application, Abba et al. (Abba et al. 2020) applied PCA for reducing the dimensions of a WWTP effluents dataset and generating a more simple dataset that was then used as an input in predictive ML methods for modelling the performance of a WWTP.

3.3.3. Predictive machine learning techniques

Predicting future water quality behaviour is very important for the WUs as it allows them to prioritise their interventions and thus prevent water quality deterioration. Depending on the water quality problem, the prediction could be either a prediction of a future water quality event, such as a coliform appearance in a DWDS or the prediction of the behaviour of one water quality parameter up to certain hours ahead. The prediction of the former is a classification problem, and the prediction of the latter is a regression problem. In this thesis, both types of problems are investigated. The predictive techniques presented in this section and in the following chapters are selected either because they are included in the most common ML techniques or because they are new methods, and their potential is promising. More specifically, the methods presented here belong to the ensemble decision trees category (random forest (RF), boosting), the artificial neural networks category (feed forward, NARX) and the deep learning neural networks category (LSTM).

3.3.3.1. Ensemble decision trees

Models that belong in the ensemble decision trees category have the advantage of operating as an "white box" which means that after their application to the dataset, the resulting trees that end up in a specific decision could be explored. Thus, it is possible to find the variables that were more crucial for that decision and remove the features with less or no impact. In addition, ensemble decision trees could handle different types of data, including categorical data and sample data, and require less data pre-processing and manipulation (S. R. Mounce et al. 2017). Furthermore, the produced outputs by these models are calculated as probabilities and so the end users could observe the likelihood of each prediction and make their final decisions. Therefore, for all the above reasons, ensemble decision tree models are significant tools for the WUs. In the thesis, the techniques that were selected for investigation are the random forests (RFs) and some methods that implement the boosting algorithm. A brief explanation of the algorithms that these methods follow is presented in this section, a case study that compares the classification performance of these methods is presented in chapter 7, a case study that compares RFs with an ANN is presented in chapter 8 and finally RF is also used in Balmore WTWs investigation in chapter 9.

3.3.3.1.1. Random Forests (RFs)

Random Forests (RF) is an ensemble decision trees technique that generates a large number of trees whose splitting decision at each node is dependent only to a small randomly selected group of the total number of dataset's variables (Breiman 2001). Random forests could be used for either regression or classification problems. In RF, every generated tree contributes equally to the final decision and therefore, in regression problems the final RF prediction value is equal to the mean of the predicted values of each tree and in classification problems the final class prediction is equal to the class that most of the independent trees chosen.

Hyperparameters in machine learning are the parameters of ML technique that should be defined prior to the learning process. There are three hyperparameters that RFs' performance is dependent on, the number of generated trees, the number of the variables that consist of the splitting group at each node and the tree depth (Scornet 2018). The maximum number of variables that consist of the splitting group is equal to the number of the available variables of the dataset. The hyperparameters should be defined by the user before the application of the method to optimize the model accuracy, minimize the errors and avoid overfitting. Typically, in RFs the number of the variables that consist of the splitting group at each node, is equal to the total number of variables divided by three (Xenochristou 2019).

3.3.3.1.2. Boosting

In this subcategory belong all the algorithms that follow the boosting method. The boosting method initially gives a random set of weights over the dataset and then adjusts these weights depending on the learning experience of the first classifiers. This procedure continues until all the trees of the ensemble have passed the learning procedure. The adjustments increase the weights of the samples that were misclassified and decrease the weights of the correctly classified samples, based on the selected boosting algorithm (Dietterich 2000). Therefore, the main differences between RF and boosting are as follows:

(i) In RF each tree of the ensemble is independent while in boosting each new tree tries to improve the performance of the ensemble in areas of the dataset where the previous trees failed

(ii) In RF all data are independent and used as collected while in boosting weights have been imported in the datasets

(iii) In RF all trees are equally contributing to the final decision while in boosting the high weight trees contribute more than the others in the final decision

(iv) In boosting the aim of the method is to reduce bias toward a certain direction but in RF the aim is to reduce the overfitting problem due to the independence of the data.

The different algorithms that follow the boosting method follow a different procedure to find the weights that could be used in the next step. The first boosting algorithm is the Adaptive boosting or AdaBoost (Freund and Schapire 1997) which is a classification ML technique but there are other boosting algorithms that could be used only in classification problems or in both classification and regression problems. Some examples are Gradient boosting machines (Friedman 2001) and extreme gradient boosting or XGBoost (T. Chen and Guestrin 2016) that are used for both regression and classification, adaptive logistic regression or LogitBoost (Jerome, Trevor, and Tibshirani 2000) a classification technique for data with not perfectly separable classes and random under sampling boosting or RUSBoost (Seiffert et al. 2010) a classification technique for datasets with imbalanced classes. The hyperparameters required in boosting defer depending on the algorithm. In general, the hyperparameters required by all the boosting algorithms are the number of the generated trees, the tree depth, the learning rate and the number of split decisions where learning rate is the factor that each new classifier is allowed to contribute in the change of the weights comparing to the previous classifier and the number of split decisions define the maximum number of subgroups that each node could be split with the maximum value being the number of training samples.

3.3.3.2. Artificial neural networks (ANNs)

As mentioned in the previous chapter, ANNs consist of an input layer, a hidden or a number of hidden layers and the output layer. Due to the hidden layers, the ANNs are considered "black box" methods as it is not possible for the user to check the procedure followed in the hidden units. Therefore, it is not suggested to apply these methods in monitoring water quality samples as it is not possible to check the factors that contribute to the decision and thus, it is not possible to improve the model, besides most of the ANNs require timeseries or continuous data as inputs for their application. However, ANNs have proven their ability in predicting future behaviour in various water related research projects and thus, two ANN techniques are investigated in this thesis. The first one is the feed forward ANN, the first and most common ANN. The second is the Nonlinear autoregressive exogeneous (NARX) ANN which has been proven a really good tool for predicting peak levels in timeseries datasets (Pisoni et al. 2009; Boussaada et al. 2018; Kazemi et al. 2018).

3.3.3.2.1. Feedforward ANN

The feedforward is the first ANN ever invented and is also the simplest and most commonly used ANN. In feedforward, the information travels only in the forward direction from the input layer to the hidden layer/s and then to the output layer with no loops or cycles in the network (Bishop 2006). The minimum number of layers in a feedforward network is two (input and output layers) where in this case the feedforward model is called a single-layer perceptron and the relationship between inputs and outputs is linear. A feedforward network with at least one hidden unit is called a multi-layer preceptor (MLP) as it is composed of more than one perceptron. The structure of a feedforward ANN with 2 hidden layers is shown in figure 3.3. This feedforward ANN, represented in figure 3.3, has 3 input nodes 2 nodes in the first hidden layer, 3 hidden nodes in the second hidden layer and 2 nodes in the output layer.

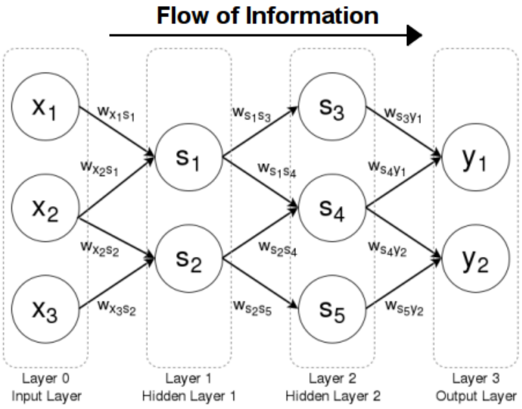


Figure 3.3:Example of a four layer (1 input, 1 output and 2 hidden).

Each node in the hidden layers represents a linear connection between all or some of the input variables. For example, in the feedforward shown in figure 3.3 the hidden node s_1 follows this equation: $s_1=x_1w_{x1s1}+x_2w_{x2s1}+b$ where w_{x1s1} and w_{x2s2} are the weight coefficients that connect x_1 to s_1 and x_2 to s_2 respectively and b is the bias coefficient. However, this linear combination is transformed to a nonlinear relationship via a nonlinear function, known as activation function, and therefore, this procedure gives the ANN the ability to understand and learn non-linear relationships between inputs and outputs (Kuhn and Johnson 2013). More specifically, the training of the feedforward ANN follows the following steps:

a. Initialization: the weights and bias coefficients are given random numbers
b. Hidden layer calculation: The hidden nodes are calculated following the linear relationship described above and transformed with the aid of the selected activation function (for example the Sigmoid function for classification problems).
c. Output layer calculation: The output layer activation values are calculated, once all the hidden units are defined, usually as the sum of the contribution of each hidden unit to the output layer (linear activation function). However, other functions could be also applied (i.e. Softmax function). The mathematical expression of the linear activation function is as follows:

$$f(x) = b_0 + \sum_{k=1}^{H} \quad w_k h_k$$

where b$_0$ is the final bias, h$_k$ is the kth hidden unit and w$_k$ its corresponding weight.

d. Error calculation: The error between the initial output layer and the actual values is calculated using the mean square error (MSE) metric as described in the following section of this chapter
e. Backpropagation: In this step, the aim is to adjust the weights in order to minimise the MSE and thus improve the training of the ANN. The weights are updated using an optimisation algorithm such as the gradient descent optimization algorithm described below:

$$w_{knew} = w_k - a(\frac{\partial Error}{\partial Wx})$$

where ∂Error is the MSE and a is the learning rate which, as in the ensemble decision trees, is the hyperparameter that controls the learning procedure. The backpropagation step is a repeated calculation of the selected optimization algorithm for all the weights and the biases of the network that stops when the error is minimized or after certain repetitions.

The required hyperparameters for the application of the feedforward ANN are the number of hidden layers (the ANN could accept up to 3 hidden layers), the number of units per layer, and the learning rate. It is also important to initially define the activation function for the hidden layer/s, the activation function for the output layer, the optimisation algorithm and the criteria that will stop the ANN optimisation procedure (i.e. number of repetitions, weight decay etc.)

3.3.3.2.2. Nonlinear autoregressive exogenous (NARX) ANN

The NARX model is initially developed as the non-linear approach on the autoregressive exogenous (ARX) model, a model used in time-series analysis, in order to capture the hidden nonlinear relationships between various time-series datasets (Q. Liu et al. 2020). The NARX model equation is as follows:

$$y(t) = F(y(t-1), y(t-2), \dots, y\big(t-n_y\big), u(t-2), \dots, u(t-n_u))$$

This equation means that the one-step ahead prediction y(t) is a function F () of previous outputs y () and of previous independent (exogenous) inputs u(). The n_y and n_u represent the maximum time lags for the y() and u() respectively (MathWorks 2020).

The NARX F() function could be modelled using artificial neural networks (ANNs). There are 2 different NARX ANN architectures, the series parallel (also known as open loop) architecture and the parallel (also known as closed loop) architecture (Boussaada et al. 2018). The open loop NARX is a feedforward ANN that for a future prediction of y(t) uses the past and present values of the u(t) timeseries and the actual past values of the y(t) timeseries. The closed loop NARX is also a feedforward ANN but the difference here is that the future prediction is generated with the use of present and past u(t) time-series values and the past predicted y(t) time-series values of the NARX model. Both models could be used for future predictions of a time-series dataset. The open loop approach has the advantage that during the training period the feedforward network is more accurate compared to the closed loop network, as it uses real time-series dataset, but it cannot be applied for many steps ahead prediction as the closed loop does. Therefore, for timeseries prediction, it is commonly used to combine these two architectures by training the network using the open loop and then using the closed loop for the prediction over many steps ahead.

The hyperparameters required in the NARX ANN model are the same as the feed-forward model described above, but in addition the input delays and the feedback delays should be defined. These two hyperparameters are referred to the maximum past time steps of the input and output time-series that the model should "look" to understand the present output value.

3.3.3.3. Deep neural networks (DNN)

Deep neural networks (DNN) also known as deep learning (DL) methods, are, as briefly mentioned in the previous chapter, ANN with the ability to have multiple hidden layers between the input and the output layer (Lecun, Bengio, and Hinton 2015). Deep learning methods have the advantage, compared to the conventional ANNs, to understand the long-term relationships between different datasets and therefore, to make decisions based on the general seasonal trend of the data. This means, though, that deep learning methods require a large amount of data to capture the seasonality and that the computational and GPU requirements could be expensive due to the complexity of these models. As regards the water quality datasets, it could be understood that it is not possible to apply DNNs in monitoring water quality datasets because both the amount and the seasonality of data are absent in these datasets. However, the success of these methods in various research projects in the water sector, as presented in the previous chapter, implies that these methods could be applied in DWDS where a large number of time-series or telemetry data are available. In this thesis, the long sort-them memory (LSTM) DNN was used in a WTW case study, in chapter 10, using time-series telemetry data.

3.3.3.3.1. Long short-term memory (LSTM)

Long short-term memory (LSTM) is a deep recurrent neural network (RNN) developed by Hochreiter and Schmidhuber (Hochreiter and Schmidhuber 1997). RNNs, in contrast to the feed-forward approach, have feedback connections which means that each new hidden layer is not related only to its previous one but also to other previous layers by using their memory to save and then process the longer sequences of the input datasets. Therefore, RNNs could not only be implemented in time-series datasets but also in speech recognition, video recognition, robotics etc. (Dairi et al. 2019).

The LSTM architecture follows the RNN formulation as follows:

$$h_t = \sigma_h(W_h x_t + U_h h_{t-1} + b_h)$$

Where h_t is the hidden layer at time t, x_t is the input vector at time t, W_h, U_h and b_h are the weights of the input, the hidden and the bias vectors and σ_h is the activation function of the hidden layers. The weights are updated and optimized during the backpropagation through time, a process that takes a very long computational time due to the learning process over long time lags. Furthermore, during backpropagation, the weights error through time could be really small which effects the learning procedure from inputs that are far from the present(Hochreiter and Schmidhuber 1997). To tackle this issue, Hochreiter and Schmidhuber introduced, in their LSTM networks, a memory cell that can learn from the trends of the datasets what to memorize and what to forget. In addition, they introduced three gates that control the information passing through LSTM networks, the input gate that processes the input data and selects which data should contribute to the cell, the forget gate that removes the data that could be ignored, and the output gate that takes the final output from the cell. This complicated procedure requires new activation functions and different weights and bias vectors for each gate. Thus, the hyperparameters required for the application of the LSTM network are the size of the input gate, the number of the hidden layers and units, the learning rate, the gradient threshold, the size of the output gate and to control the feed of the large datasets, 3 other parameters are introduced, the number of epochs, the batch size, and the number of iterations. The first hyperparameter defines the number of the times the neural network passes both forward and backward once, the second hyperparameter is introduced to divide dataset into parts with equal size to avoid entering large datasets all in once and, finally, the iterations are the number of butches that are required in order to finish one epoch. Finally, the activation functions for all the gates and the training algorithm should be defined and if LSTM is used for classification problems, the softmax function should also be included in its architecture.

LSTM has the advantage of learning from a really large time window and understanding long-term relationships between the available data, by using the gates to control the inputs. During the training process the algorithm understands which information is important and which information should be removed. There are various applications that applied LSTM in various sectors including music composition (Eck and Schmidhuber 2002), speech recognition (Graves and Schmidhuber 2005) and time-series prediction (Schmidhuber, Wierstra, and Gomez 2005).

3.3.4. Preliminary Machine learning selection tree

Assuming that the first three steps of the machine learning selection application steps are answered, it is possible to direct our selection towards a specific ML category based on what are our output requirements. As mentioned above, the selection of the appropriate method is facilitated by the machine learning selection tree. The initial idea was to not separate the water quality data into two categories, however it was decided to split the data as the application of some methods in the monitoring discrete samples datasets is impossible due to the significant number of missing values that they have. More specifically, k-means and tSNE were excluded from water quality problems where only this type of data are available. In addition, ANNs and DNNs were also excluded from the analysis of discrete datasets as missing values have a significant impact on their performance (Ennett, Frize, and Walker 2001).

The machine learning selection tree presented in this section is the preliminary one (Figure 3.4 below). In the final one presented in chapter 10, PCA is removed from being a potential method for the analysis of discrete monitoring samples as the investigation over the SRs deterioration, in chapter 5, demonstrates their inability to work with this type of data. In addition, in the final machine learning tree, presented in chapter 10, another factor is introduced and needs to be specified before the final selection of the ML technique. This factor is the “interpretability” of the method which is the ability of the ML technique to make transparent explanations of their outputs.

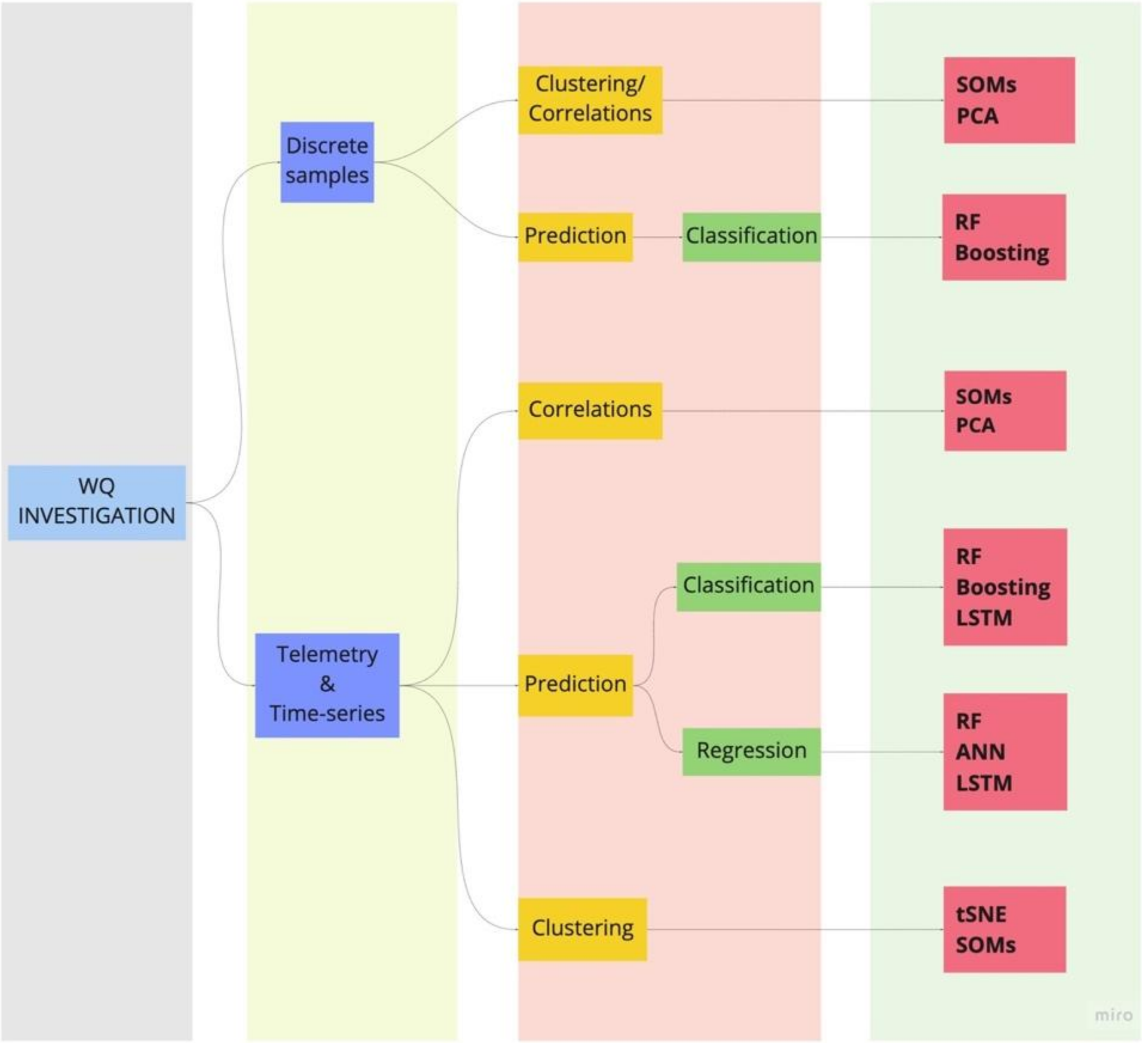


Figure 3.4:Preliminary machine learning selection tree

3.4. Machine learning performance metrics

The investigation over a certain ML technique should, obviously, include the evaluation of the technique's performance to a certain target. The model evaluation includes the ability of the technique to answer to the problem and to analyse the available dataset (i.e. tackle the missing values). In some cases, the model complexity and the computational time required for implementation may be included in the evaluation. It is easy to compare models over their computational time for implementation or their complexity, and to understand if there is a potential with applying a certain method to a certain dataset; however, evaluating their ability to answer to a certain problem (how well they performed) is not that straightforward and differs from predictive techniques to clustering techniques. In clustering, the evaluation is based on how weak or strong are the clusters or the final correlations between the variables. In the predictive models, performance metrics are introduced for the evaluation of the performance of the models with respect to the needs of the case studies. There is a large number of performance metrics for either regression or classification type of problems. This is because each one of them is biased towards a certain aspect of the model and provides outputs over the model's performance towards this aspect. Thus, the American Society of civil engineers (ASCE) proposed, in a scientific report, 7 different metrics to evaluate continuous and single event models in order to cover the bias of the model towards the mean or the extreme values, the quality and the quantity of the data and the research questions that the model aims to answer (ASCE 1993). This work concentrated into the performance of hydrological regression models; however, these criteria stand also for other engineering models. As regards the classification performance metrics, Liu et al. (Y. Liu et al. 2014) proposed a strategy for clustering the various metrics in 3 different groups with similar behaviour to help practitioners selecting metrics that evaluate different aspects of the models. The most important performance metrics for both classification and regression problems are presented in the following sections.

3.4.1. Classification performance metrics

For classification models performance, the simplest performance metric is accuracy which is the percentage of observations that are correctly classified. More specifically the formula that calculates the accuracy is as follows:

$$Accuracy = \frac{Number\ of\ correct\ class\ predictions}{Total\ number\ of\ samples}\%$$

However, in unbalanced datasets, accuracy may not be a good indicator of the model performance. This is because the accuracy of a model that classifies the test dataset to the majority class only, will be high, even if no accurate predictions are made for the minority class. Therefore, the confusion matrix is introduced in order to summarize the outputs of the classification models. The confusion matrix for a binary classification is a 2 by 2 matrix that looks as below:

Table 3-1:Confusion Matrix

	Predicted as events	Predicted non events
Actual events	TP	FN
Actual non events	FP	TN

The samples that are correctly predicted as events are defined as true positives (TP) and their sum is added in the first cell of the first row of the matrix. The samples that were correctly predicted as non-events are defined as true negatives (TN) and their sum is added in the second cell of the second row of the matrix. The samples that were wrongly predicted as non-events are defined as false negatives (FN) and their sum is added in the second cell of the first row of the matrix. Finally, the wrongly predicted events are defined as false positive (FP) and their sum is added in the first cell of the second row of the matrix.

Once the confusion matrix is created some other metrics could be calculated. Firstly, the true positive rate (TPR) and the true negative rate (TNR). TPR is also known as the sensitivity or recall of the model and is the proportion of the correctly positive events over all the actual events. TNR is also known as specificity of the model is the proportion of the correctly predicted non-events over all the non-events.

Recall = Sensitivity=$TPR = \frac{TP}{TP+FN}$ Specificity=$TNR = \frac{TN}{FP+TN}$

Another metric commonly used in classification is precision. Precision is the ratio of the correctly predicted true positives over the total number of predicted positive observations (both true and false positives). This metric is used to check the ability of the ML model not to produce a large number of false positive predictions.

$$Precision = \frac{TP}{TP + FP}$$

In the classification over unbalanced datasets, such as the water quality datasets, the above metrics are good criteria for the performance of the model over both classes. However, both TPR and TNR do not include all the available outputs of the confusion matrix (for example TPR does not include TN and FP) and, thus, do not give an overall performance of the model. Therefore, another metric is introduced that includes all the information of the confusion matrix and its result could be used as a comparison indicator of the overall performance of the network. This metric applied in this thesis is the Matthews Correlation Coefficient (MCC) (Baldi et al. 2000). The MCC formula is as follows:

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

The range of MCC values lie between -1 and 1 with models scoring close to -1 being bad models and models scoring close to +1 being good models. This metric was used as an indicator of the best model to use for event predictions in the SRs.

F1 - Score is a metric that combines precision and recall. Its aim is to investigate how harmonic the relationship between these two metrics is, which is a very important indicator for understanding the performance of the models in imbalanced datasets. The F1-score formula is as follows:

$$F1\ score = \frac{2\ X\ (Recall\ X\ \ Precision)}{(Recall\ +\ Precision)}$$

The F1 score values spread between 0 and 1 with models scoring close to 0 being bad classification models and models scoring close to 1 being good classification models.

Finally, another popular metric for binary classification is the Receiver Operator Characteristic - Area Under the Curve (ROC AUC). This is a very useful plot that shows the relationship between sensitivity and false positive rate (FPR=1-specificity). ROC is a curve that plots the FPR vs TPR and the AUC is showing in the plot how much the model could distinguish between the two classes. As regards the model performance, the best models have an AOC value close to 1, the poor model has an AOC close to 0 (meaning that the models predict the opposite to the actual class) and the models that have a value of 0.5 or near to 0.5 have no class separation capacity. Even though the ROC AUC visualises the performance of the models, this metric was not used for the evaluation of the ML models created in this thesis. This is because in the two classification problems presented in this thesis, in chapter 7 and 9, this metric could not be applied for different reasons. In chapter 7, ROC AUC could not be used as the available dataset is highly imbalanced with very few positive points and, thus, a metric that uses the FPR could be misleading. In chapter 9, ROC AUC could not be applied as the problem is a multi-classification one.

3.4.2. Regression performance metrics

There are various metrics to account for understanding the performance of a regression model. For all the metrics it is important to know the following parameters:

- n=number of samples
- O_i=the i_{th} observed value and $\tilde{O}$ = the observed mean value
- Y_i= the i_{th} predicted value$\tilde{Y}$ = the predicted mean value

The simplest of all the metrics is the mean absolute error (MAE) which is expressed as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |O_i - Y_i|$$

MAE is more used as an indication of the overall agreement between the true and the predictive values and could not highlight the larger errors over the small errors. Thus, 4 other metrics are introduced, the mean squared error (MSE), the root mean squared error (RSME), the normalised mean squared error (NMSE) and the coefficient of determination (R^2).

MSE is the average of the squared errors between the observed and the predicted values and RMSE is the square root of the MAE. The mathematical expressions of these formulas are:

$$MSE = \frac{1}{n}\sum_{i=1}^{n} |O_i - Y_i|^2 \qquad RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} |O_i - Y_i|^2} = \sqrt{MSE}$$

MSE is a quality estimator of the model performance as by squaring the difference between observed and predicted values it is easier, comparing to MAE, to penalize the higher errors. The RSME is also measuring the overall quality of the model and is also sensitive to large errors. The only reason that RMSE could be preferred over MSE is that RMSE has the same units as the observed and the predicted values as the square is removed. Values of MSE or RMSE closer to zero indicate a good model performance with zero being the perfect value (perfect model). NMSE is the MSE divided by the variance of the observed values as expressed below:

$$NMSE = \frac{\frac{1}{n}\sum_{i=1}^{n}|O_i - Y_i|^2}{var(O)} = \frac{MSE}{var(O)}$$

NMSE is often used to facilitate the comparison of a model's performance over datasets with different scales.

The coefficient of determination - R^2 is used as an indicator of the correlation degree between the observed and the predictive values. The R^2 lies between 0 and 1 (or 0% to 100%) with models scoring close to 1 being the models that explain better the variance of the observed data. R^2 is calculated as follows:

$$R^2 = \left[\frac{\sum_{i=1}^{n}(O_i - \tilde{O})(Y_i - \tilde{Y})}{\sqrt{\sum_{i=1}^{n}(O_i - \tilde{O})^2 \sum_{i=1}^{n}(Y_i - \tilde{Y})^2}}\right]^2$$

The Nash-Sutcliffe Model Efficiency Coefficient (NSE) is commonly used in hydrological modelling to investigate how the predicted by the model values fit the observed data. When NSE=1 the model is optimal, when NSE=0, the model has the same predictive skills as the mean of the observed data and when NSE is less than zero the mean of the observed data is a better predictor than the model. NSE is defined as follows:

$$NSE = 1 - \frac{\sum_{i=1}^{n}(Y_i - O_i)^2}{\sum_{i=1}^{n}(\tilde{O} - O_i)^2}$$

NSE and R^2 both belong to the group of metrics that are used for understanding how well the predicted data fit with the observed values. Therefore, in this thesis, for avoiding a repetition of similar model evaluations, only the R^2 was used.

4. Scottish Water's water quality data analysis

4.1. Introduction

The aim of this chapter is to briefly present the water quality monitoring program that Scottish Water (SW) follows and to describe the steps followed by this thesis for setting up a water quality dataset using discrete monitoring water quality data. The dataset setting up procedure includes the downloading and the collection of water quality data from the period between the 1st of January 2012 to the 31st of May 2020, the downloading and the collection of meteorological data for the same period, the extraction of data from SW GIS, the collection of information for various SW assets and for the way that SW organises their WDNs, the collection of technical details for each asset, and the organisation of the data in a format that would, firstly, help us understanding the dataset and then use it as input for data-driven techniques for further water quality investigations. The created water quality datasets were used as inputs to the machine learning techniques in some of the water quality investigations presented in the following chapters.

4.2. Scottish Water's water quality monitoring program

4.2.1. Definition of water for human consumption in Scotland

SW monitors water quality by taking, sporadically, samples from their WTWs exit points, their SRs exit points and from some of their consumers' taps, selected randomly. The monitoring program follows the *"Public Water Supplies (Scotland) Regulations 2014"*. The most updated version of these regulations came into force on the 1st of January 2015 (DWQR 2014).

The regulations define the, supplied by SW, water as water for human consumption only if it follows these requirements:

(i) does not contain microorganisms

(ii) does not contain substances and parasites at concentrations that could be dangerous for human health

(iii) does not contain any water parameters that exceed its upper concentration or value limit prescribed in Table A & B of the regulations at the point of measurement (WTW, SR, customer tap) defined by the same tables.

4.2.2. Monitoring program on consumers' taps

The regulations obligate SW to divide, before the beginning of each year, the areas that supply water into zones with similar water quality and of a maximum 100000 population. These zones are defined in the regulations as Water Supply Zones, however inside SW these are defined as Regulatory Supply Zones (RSZs), as Water Supply Zones (WSZs) are the distribution areas that are fed with water from the same SR. The monitoring program, as defined by the regulations, sets the minimum number of samples per year that should be taken from different consumers' taps inside each RSZ, to measure each one of the specific parameters described in the Table A, B and C of the regulations (DWQR 2014). For bacteriological parameters such as coliform bacteria, *E-Coli*, Colony counts, and disinfection residual, SW should always monitor with a minimum frequency of 12 measurements per 5000 people per year at each RSZ. However, for some chemical parameters monitoring is only required if certain circumstances, defined by Table 1 of the regulations, occur. For example, iron and manganese should be measured only if the water source is surface water. In addition, for each one of the chemical parameters only, the regulations allow SW to reduce the number of samples that should take the following year to those defined in Table 2 of the regulations, only if in the previous year none of the samples of that specific parameter exceeded the upper or lower limit(DWQR 2014).

4.2.3. Monitoring program on water treatment works

The regulations oblige SW to monitor the quality of the water exiting the works with a specific frequency defined at Table 4 of the regulations (DWQR 2014). The number of samples taken per WTW per year are dependent on the volume of water supplied by the WTW per day in m^3/d. In the WTWs, the minimum number of samples taken for each one of all the parameters (chemical, physical and bacteriological), could be reduced to the number defined by Table 4 of the regulations, only if in the previous year none of the samples of that specific parameter exceeded the upper or lower allowed limit.

4.2.4. Monitoring program on service reservoirs

The regulations require that SW takes one sample per week from the exit of every SR that is in use to measure all the main bacteriological parameters (Coliform bacteria, *E-Coli*, colony counts at 22 and 37 ^{o}C and disinfection residual).

4.2.5. Additional sampling

SW is doing additional monitoring sampling for various reasons including:

(i) Measuring other water parameters that are not included in the regulations, such as flow cytometry total cell counts, and intact cell counts in their assets
(ii) investigating over customers complaints in specific areas of the network
(iii) investigating over the continuous over the limit concentration of a certain water parameter
(iv) investigating over a potential water quality deterioration due to certain events that occurred in the DWDS (e.g., pipe burst)

4.2.6. Publication of yearly monitoring water quality results

SW must prepare and maintain a record of all the samples taken including the name of the zone, the WTW and SR where the sample is taken, and the lab analysis output of every parameter measured in that sample and notify the Drinking Water Quality Regulator (DWQR) for any water quality deterioration event. SW must, also, publish by the end of the March in each year the results of their monitoring program in a report including the number of samples that were taken and the number of the compliance and non-compliance samples. Then, DWQR should check SW's report and produce their own annual report which is public and available to everyone. According to the 2019 *"Drinking Water Quality in Scotland 2018"* report, SW undertook an overall of 319124 samples and achieved a 99.9% compliance for the year 2018 (DWQR 2019a).

4.3. Creating water quality datasets

The final format of each dataset is dependent on the desired data-driven analysis and, thus, it cannot combine a mix of data that do not have any physical connection. For example, a dataset that combines monitoring water quality data taken from both the taps and the WTWs together, does not make any practical sense as it combines analysis results from samples taken from different parts of the network. It is also important that the final datasets are organised in a way that is easily readable and could be used by the machine learning techniques directly. Therefore, three different monitoring water quality datasets were created, one that included all the SW WTWs water quality data for the years 2012-May2020, one that included all the SW SRs water quality data for the same years, and one that included all the samples taken from the consumers taps fed by SW's DWDS. In addition, each row at each one of the datasets represents a different sample taken and each column represents a different parameter for each specific sample including the sample ID and the point and the date that it was collected. The parameters that were not measured in a sample are left blank (missing values).

4.3.1. Collection of monitoring water quality data

At the time that the datasets were created, there were three different tools that SW's employees were using for downloading monitoring water quality data. Each one of these tools had its advantages and disadvantages. More specifically:

Tool A was the official tool for downloading and monitoring water quality data**.** The format that the downloaded data had, was the required one, however, it only allowed 1000 data samples per time used. In addition, for each tap sample the information given was the address point and the RSZ that the property belonged to. Moreover, it was a difficult software to use for extracting WTWs water quality data as it was not clear if the samples were taken at the entrance or the exit of the works.

Tool B was a macro excel file that was directly connected to the SW laboratory database. This tool was more user friendly than tool A and there was no limit as regards the number of samples to download per time. However, in the final extract format, each measured parameter was exported in a different line, creating large datasets with lots of unwanted and repeated lines, and as regards the tap samples, a limited information was given (RSZ and postcode of the property that the sample was taken).

Tool C was a data processing tool that allowed users to create reports by selecting the information that he/she wanted to extract. The software was not easy to use and, as in tool B, each parameter was extracted at a different line. However, with this tool it is possible to extract important information regarding the water quality samples. More specifically, for each tap sample information regarding the property eastings and northings and the DMA, the WSZ, the WOA that the property belonged to could be given, and for WTWs samples the volume of the water supplied by the plant and the type of the water source feeding the plant were given.

Knowing the DMA and the exact location that a tap sample is taken, offers the opportunity to do an analysis at a DMA level and at distribution pipe level by linking the tap sample to its distribution pipe. Therefore, the tool used for downloading the data was tool C.

4.3.2. Collection of rainfall data

For understanding the relationships between rainfall and water deterioration the daily and hourly rainfall data were collected from the Met Office stations in Scotland (Met Office 2020) for the period between the 1st of January 2012 to the 31st of May 2020.

4.3.3. Collection of any other information

In the following table 4-1, the files and the tools used to collect any other information regarding SW's DWDS and the type of the collected information are presented.

4.3.4. Data manipulation and water quality datasets production

The steps followed for combining the above collected data, changing their format, removing unnecessary information, and finally creating the monitoring water quality datasets were as follows:

Table 4-1: Data sources used in the datasets

Sources	Extracted information
GIS	-Pipe material -Pipe commission date -Pipe diameter -Pipe location -Pipe ID -Pipe length -No of bursts per pipe -DWDS hydraulic hierarchy
Scottish Water database file 1	-Service reservoir retention time -Water age exiting the SR (as sum of the retention time of all the cascading SRs - pipe travel time not included) - Number of feeding DMAs -Last cleaning date per SR -WTW fed by
Scottish Water database file 2	-DWDS hydraulic hierarchy (including connections between WOAs and the WTWs that fed them etc.) -Number of properties per WOA, WSZ, DMA
Scottish Water database file 3	-SRs' year of construction
Scottish Water database file 4	-WTWs names -WTWs disinfection type

Step 1: Changing the format of the raw water quality data

The raw water quality data that were downloaded using the tool C, required a reshape of their format in order to get a final table where each line represents a different sample and each column a different parameter for this sample, including its location, its area and its measured water parameter. Thus, a code was created in MATLAB R2018a (The MathWorks Inc., Massachusetts). The code was also identifying the samples with two or more measurements per parameter and retaining the first measurement only. The code was run separately for the tap, the SRs and the WTWs data creating three different datasets. The code for this step is in the GitHub repository mentioned in chapter 1.

Step 2: Connecting SRs and WTWs asset information to the water quality datasets.

The next step, after creating the datasets, is to add information regarding the DWDS that each one of the samples belongs, for example, including the water age leaving the SR's in the SRs and the customer taps water quality datasets, and the disinfection type of the system to all three of the datasets. Firstly, the connection between the sampling point and their SR or their WTW was made using searching tools. Then, by using the *"outerjoin"* command in MATLAB and the files from SW's database described in the above table, the asset information for the WTWs and the SRs was included in both the WTWs and the SRs water quality datasets. As regards the tap water quality datasets, two further steps were required, the first was to connect the Water Supply Zone (WSZ) that the sample belongs to the SR serving this area with water and to connect the Water Operation Area (WOA) that the sample belongs to the WTW that is fed by. This step was achieved, again, by using firstly searching tools and then the *"outerjoin"* command in MATLAB.

Step 3: Connecting tap samples to their nearest distribution main
The raw extracted tap monitoring water quality data do not include any information regarding the distribution main that fed the properties, an information that could be useful to include in the research. Therefore, the process followed to reference the tap data to their nearest water main was as follows:

- Import the tap data in the ArcGIS (ESRI., California) using the eastings and northings
- Use of the "Spatial Join" command to connect each sampling point to the nearest main using a straight line to count the distance between them
- Creation of a new layer that contains both the sample's and water main's information
- Export layer in a csv file and merge it with the tap dataset using the "outerjoin" command in MATLAB

Step 4: Relating tap and SR water quality data with the WTW that are fed by
To relate SRs' water quality with the water quality of the WTW that is fed by, and the customers' taps water quality with the water quality in both their WTWs and their SRs, a code in MATLAB was created. This code calculated the monthly average values of all the parameters measured in the samples taken in the WTWs exit points and then merged them to the SR water quality dataset and the tap dataset. Thus, in the final dataset, for every sample taken at a specific month, the average value of each parameter measured at its related WTW at that month, is included. The same code was also used to relate each tap sample to the SR that is served by. As with the previous code, this one is also added in the GitHub repository.

Step 5: Adding meteorological data in the water quality datasets.
The aim in this step was to include in the water quality datasets the average daily rainfall per month per year and the total monthly rainfall per year of the area where each SR and each WTW are located. This requires calculating the values for these two parameters for every Met Office station and then referencing a Met Office Station at each WTW and at each SR. The former was achieved with a MATLAB code and the latter was achieved by following the same process as in step 3 in ArcGIS.

4.3.5. Final water quality datasets

The types of information included at each one of the final datasets are summarized at the following table.

Table 4-2:Type of information included at the water quality dataset

Dataset	Type of information
WTW water quality dataset	- Sample ID & date collected - Water quality analysis results - Average daily precipitation & total monthly precipitation - Disinfection type
SR water quality dataset	- Sample ID & date collected - Water quality analysis results - Average daily precipitation per month & total monthly precipitation in the SRs - Average daily precipitation per month & total monthly precipitation in the WTWs - Monthly average values of the WTWs water parameters - SRs asset information (age of SR, bacteriological failures etc.) - Disinfection type - Age of water exiting the SRs
Tap water quality dataset	- Sample ID & date collected - Water quality analysis results - Average daily precipitation per month & total monthly precipitation in the SRs - Average daily precipitation per month & total monthly precipitation in the WTWs - Monthly average values of the WTWs water parameters - Monthly average values of the SRs water parameters - SRs asset information (age of SR, bacteriological failures etc.) - Disinfection type - Age of water exiting the SRs - DWDS hierarchy information (WOA,WSZ DMA etc.) - Distribution main information

5. Understanding bacteriological activity in service reservoirs by applying data-driven techniques on water quality datasets

5.1. Introduction

Service reservoirs (SR's) are assets inside the DWDS used for balancing water supply variations. Drinking water, during its travel from the WTWs to consumers' taps, could stay for a significant amount of time (from a couple of hours to a few days) inside these assets and, therefore, SRs are crucial components of the systems. As explained in the previous chapter, water utilities (WUs) are taking samples from the SR's outlet to monitor the bacteriological activity and the chemical and physical parameters of the water exiting these assets. However, monitoring sampling is, as previously mentioned, sparse in time and, therefore, it is not possible to clearly understand the factors that could cause bacteriological failures in the SRs by just checking if these samples comply with regulations limits. It is for the WUs benefit, to further investigate methods and techniques that could transform the sparse SRs' water quality datasets into valuable material for understanding water quality behaviour inside them. Speight et al. (Speight, Mounce, and Boxall 2019) demonstrated that Self organising Maps (SOMs), a clustering and visualisation machine learning technique, has the potential of identifying the causes of discolouration when applied in consumers taps' water quality datasets. In this chapter, SOMs and PCA are applied in the SRs water quality dataset with the aim to, firstly, understand the factors that increase bacteriological activity in the SRs and to, secondly, investigate the potential of these two techniques as supporting tools for decision making on interventions in SW's DWDS. SOMs and PCA are selected in this investigation, instead of other clustering techniques, as they are data mining techniques that generate clear and simple visualisation outputs that indicate the correlations between multiple water quality parameters, an ability that other clustering techniques do not have.

The overall aim of this chapter is to do an investigation over the bacteriological activity in the SRs, the factors that influence this activity and to further explore the ability of these two ML methods on data mining of sparse water quality samples dataset. In this investigation, various water quality indicator parameters are used in addition to other information, such as the precipitation in the SRs, and WTWs and the disinfection type used in these. As there is limited research in the area that concentrates in the WQ in service reservoirs (Doronina et al. 2020), this work aims to aid water utilities on understanding what causes bacteriological deterioration in the water that exits their SRs and, therefore, better maintain their DWDS. In addition, by comparing the outputs produced by both SOMs and PCA, this chapter aims to make a proper comparison between these two methods and propose the most appropriate one for these types of WQ problems. Therefore, this chapter seeks to address objectives 2 and 4.

5.2. Methods

5.2.1. Data Collection & analysis

The SR dataset includes a total of 405464 samples taken in all the Scotland's SRs outlets that belong to SW in the period between 2012 and May 2020. For some SRs there are just few data available, or the available data are for certain periods of the year. This is because some of the SRs were either abandoned at some point during the analysis period or these SRs are used in certain periods of the year that the water demand is increased. However, the data from these SRs is also included in the analysis. As mentioned in the previous chapter, in the final dataset, apart from all the water quality parameters measured in the SRs' outlet, the average monthly values of the parameters measured in samples exiting the WTWs, the precipitation in both the WTWs and the SRs and other qualitative parameters were also included.

5.2.2. Self-Organising Maps

Self-organising Map analysis was carried out using the MATLAB® SOM Toolbox version 2.1 (Teuvo Kohonen 2014) in MATLAB® version 2019b. For the analysis only some of the parameters were used depending on the research question that needed to be analysed. Therefore, three different algorithms were created in MATLAB® (codes are stored in GitHub) to, firstly, extract the selected quantitative and qualitative parameters from the main dataset and then to call the Toolbox for the analysis. The Toolbox, normalised the data, conducted a rough training and created the final SOMs plots. In the final output, each selected parameter was represented with a different map including the qualitative parameters that were created after the main SOM training was finished. By default, the Toolbox creates a colour bar scale that changes from deep blue for the low values of each parameter to deep red for the highest values of each parameter. In this analysis, the outliers initially were included, however, to guarantee that the final SOMs were not skewed by extreme values, the maps' colour range was standardized to use all the values that were between the 5th and the 95th percentile of the dataset. Finally, a different algorithm was created to include the number of samples per parameter, the average and the standard deviation in the final analysis.

5.2.3. Principal Components Analysis

PCA analysis was conducted using MATLAB® version 2019b. The PCA function in MATLAB was used in an algorithm produced to apply this method. In addition, for the selection of the parameters selected for the PCA analysis the algorithms that were created for the same purpose in the SOMs analysis were used. The final number of samples per parameter and the average and standard deviation of each parameter used in each analysis were included in the final analysis.

5.3. Machine learning application steps

The machine learning application steps are filled as follows:

a. Define the water quality problem

In this chapter, the aim is to understand the factors that increase bacteriological activity and cause bacteriological failures in the SRs. Therefore, the water quality problem should be defined as follows:

What are the main water quality parameters related to increased bacteriological activity and high numbers of bacteriological failures in the SRs?

b. Type of the available data

The available data for the investigation are the water quality monitoring samples taken from the SRs' outlets and the WTWs' outlets.

c. Define required output

The required output in this investigation is to identify clear correlations between water parameters measured from the water quality samples

d. Machine learning selection

Clustering and identification of the correlations between various parameters requires an unsupervised machine learning technique. By following the machine learning tree presented in chapter 3, the selected techniques could be either the SOMs or PCA.

e. Data preparation

The data for both techniques should be prepared in a way that each row is a different observation (sample) and each column represents a different water quality parameter (coliform bacteria, heterotrophic plate count, flow cytometry data etc.). The SRs water quality dataset created as described in the previous chapter will be used as input to SOMs and PCA.

f. Application output

The required outputs are a few graphs that visualise the correlations between the various parameters and therefore the correlations between them could be explored. The procedure followed to produce these outputs is presented in the following sections of the chapter.

The ML application steps for the SRs' water quality investigation are summarized in the following figure.

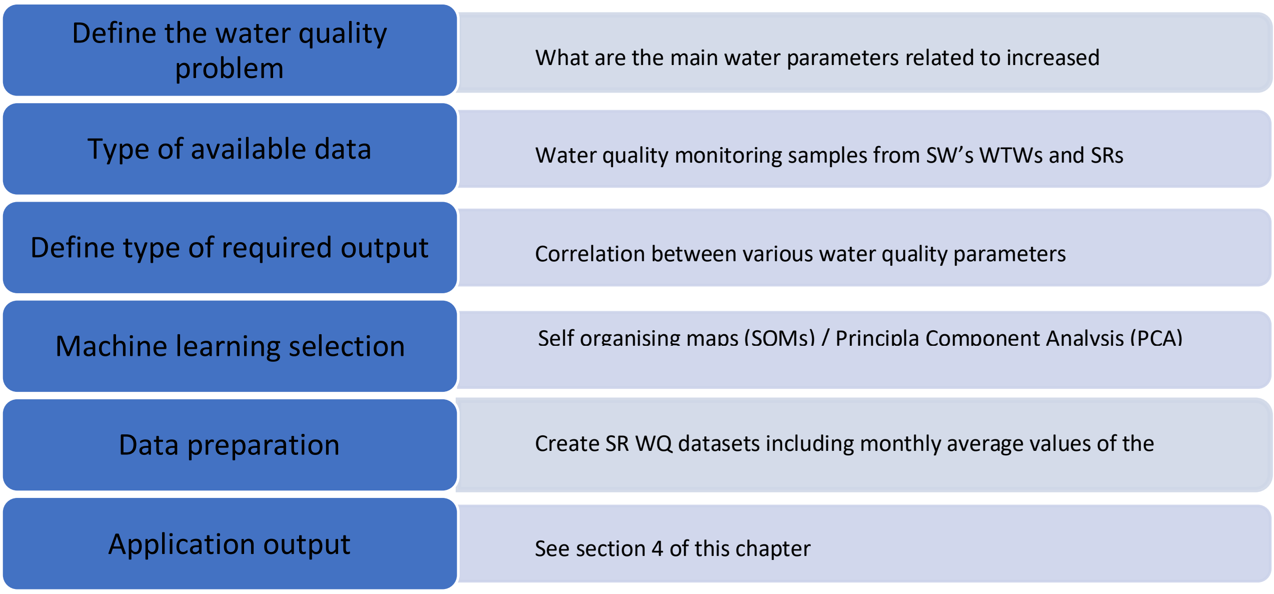

Define the water quality problem
What are the main water parameters related to increased
Type of available data
Water quality monitoring samples from SW’s WTWs and SRs
Define type of required output
Correlation between various water quality parameters
Machine learning selection
Self organising maps (SOMs) / Principla Component Analysis (PCA)
Data preparation
Create SR WQ datasets including monthly average values of the
Application output
See section 4 of this chapter

Figure 5.1: Machine learning application steps for the SR’s water quality case study

5.4. Results

This investigation aims to identify and understand the correlations between the main bacteriological parameters and the other chemical and physical characteristics of the water measured in the SRs’ outlets. The bacteriological parameters that SW measures in the SRs outlets are coliform bacteria and *e-coli,* heterotrophic plate counts (HPCs), flow cytometry total cell counts (TCCs) and intact cell counts (ICCs). In addition, another parameter that SW is using in their investigations is the number of bacteriological failures (coliform or *e-coli* events) per SR per year. As the DWDS are complex systems and the bacteriological activity could be related with multiple parameters a multi-investigation that included the interaction between the various parameters inside the SRs, the impact of the WTWs in SRs biological activity, and the impact of the SR cleaning was generated. Therefore, the main research objective of this chapter - understanding bacteriological activity in the SRs - was subdivided in 4 different research questions as follows:

a) What are the main factors related to bacteriological failures in SRs?
b) Which are the main parameters correlated with increase in the bacteriological activity in the SRs?
c) What is the impact of the WTWs to SRs bacteriological activity?
d) What is the impact of SR cleaning to the bacteriological activity inside them?

5.4.1. Self-organising Maps

For each research question, a different SOM was produced. The results are presented in this section.

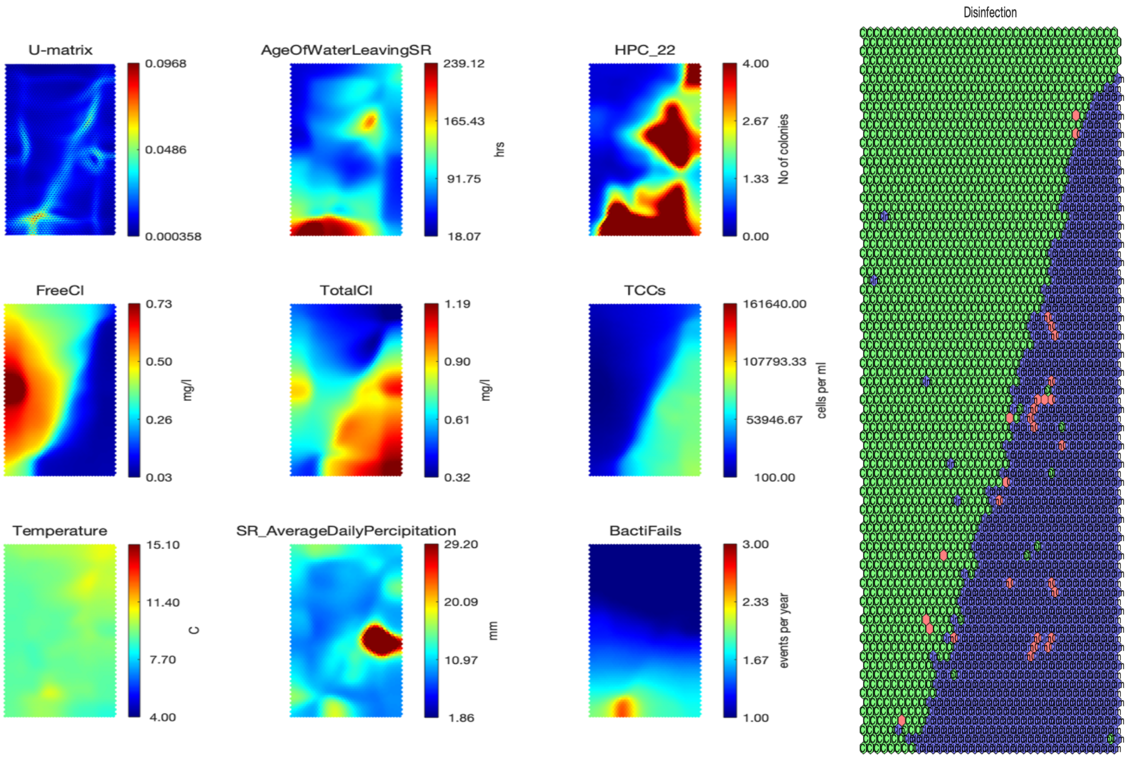
a) Which are the main factors of bacteriological failures on SRs?

A SOM was produced to answer the above question using the following parameters: age of water exiting the SR, HPCs at 22^{o}C, free chlorine, total chlorine, FC_TCCs, temperature, average daily precipitation per month per year in the SRs and the bacteriological failures (figure 5.2).

As Figure 5.2 indicates, the high total chlorine clusters on the right of the plot are correlated with the low free chlorine clusters. This is an indication that these clusters belong to chloraminated SRs which is confirmed in the labelled SOM on the right where most of the high chlorine clusters are correlated to chloraminated SRs (blue cells). Contrariwise, high, and medium free chlorine clusters are correlated to chlorinated systems (blue cells). These two correlations were already known, but their appearance in this specific SOM, is a clear demonstration that Self-organising Maps could be a very useful tool for clustering monitoring water quality data.

By looking the high bacteriological failures cluster located in the bottom of the plane, high bacteriological failures clusters are correlated with medium to high temperature clusters, low total and free chlorine clusters, high age of water, medium to high average precipitation per month in the SRs and could happen in both disinfection type systems. It is also clear that the increase in bacteriological failures is mostly related to the age of water exiting the SRs as the increased bacteriological failures clusters as appeared in the bottom of their plane follow almost the same trend as the age of water clusters in their map.

Other interesting findings include the clear correlation of the medium to high TCCs in the left of their map with the chloraminated systems, a correlation that also appears in most of the high HPCs clusters but as the age of the water cluster indicates is high in these systems. High HPCs also appear in the chlorinated systems with medium or high age of water exiting the SRs (top right, middle, bottom left), low free chlorine and medium to high temperature of the water (top right, bottom middle to right).



Green cells: chlorinated
Blue cells: chloraminated

Figure 5.2:SOM for bacteriological failures, including secondary disinfection labelled map

Table 5-1: Summary of variables for SOM presented in figure 5.2

Variable (units)	Variable short form for SOM	Data Source	Number of samples	Average value	Standard deviation
Age of water leaving the SR (as sum of the retention time of this SRs and the retention time of the previous SRs that the water passed through - hrs)	AgeOfWaterLeavingSR	SW Asset files	401926	86.51	77.35
HPCs @22C (No of colonies)	HPC_22	Water quality	394206	1.95	16.61
Free Chlorine (mg/l)	FreeCl	Water quality	396194	0.32	0.25
Total Chlorine (mg/l)	TotalCl	Water quality	373708	0.70	0.28
Flow cytometry Total Cell Counts (cells per ml)	TCCs	Water quality	51753	34272	143257
Temperature (C)	Temperature	Water quality	240095	9.30	4.87
Average daily precipitaton per month in the SRs (mm)	SR_AverageDailyPercipitation	Met Office	365760	11.63	26.55
Bacteriological failures (events per year)	BactiFails	SW Asset files	61278	1.24	0.63

b) Which are the main parameters correlated with increase in the bacteriological activity in the SRs?

The aim is to investigate the correlations between the bacteriological indicator parameters and the main parameters that, to the best of our knowledge, could be the responsible parameters for the increase of bacteriological activity in the SRs. Thus, the parameters presented in this SOM (figure 5.3) are the age of water exiting the SRs, the HPCs at 22°C, the free and the total chlorine, the temperature of the water, the flow cytometry ICCs and TCCs and finally the time, in days, from the official cleaning date as given by SW asset files. The latter was calculated by finding the difference, in days, between the official cleaning date and the official date that sample was taken. Therefore, negative values indicate samples that were taken before the cleaning date and positive values indicate samples that were taken after the SR was cleaned. A post analysis disinfection SOM was also created with cyan cells for chloraminated systems and red cells for chlorinated systems, as figure 5.3 indicates.

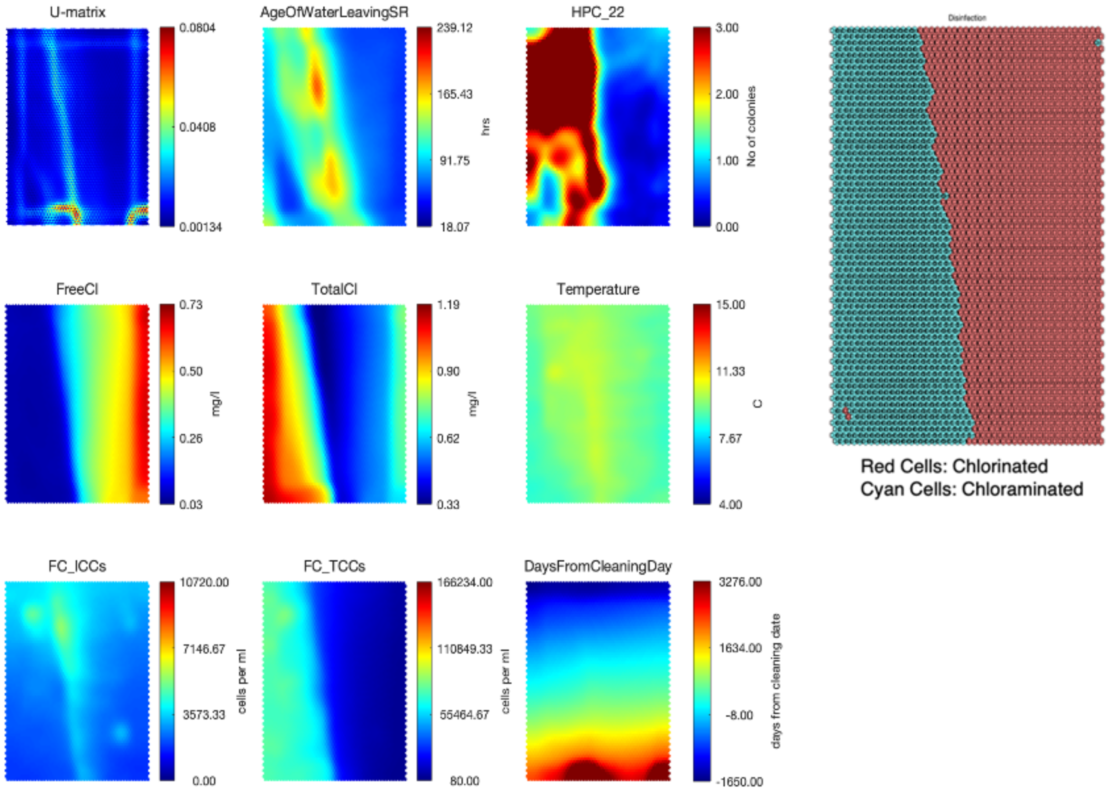


Figure 5.3:SOM for bacteriological activity in the SRs, including secondary disinfection labelled map

This SOM confirms the correlations found in the previous SOM as well. More specifically, this SOM confirms the clear correlation between high and medium TCCs and high HPCs with chloraminated systems, the correlation between high age of water and low total and free chlorine in both systems with HPCs and the correlation between clusters of medium to high temperature with HPCs.

In addition, in this SOM the clusters with increased numbers of ICCs (top and centre of the map) are correlated with clusters of high age of water exiting the SRs, high temperature and low free and total chlorine in both chloraminated and chlorinated SRs. This SOM also indicates that there is some clear correlation between some of the high HPCs clusters and the high ICCs clusters. It also shows that there is no clear correlation between ICCs and TCCs.

Table 5-2: Summary of variables for SOM presented in figure 5.3

Variable (units)	Variable short form for SOM	Data Source	Number of samples	Average value	Standard deviation
Age of water leaving the SR (as sum of the retention time of this SRs and the retention time of the previous SRs that the water passed through - hrs)	AgeOfWaterLeavingSR	SW Asset files	401926	86.51	77.35
HPCs @22C (No of colonies)	HPC_22	Water quality	394206	1.95	16.61
Free Chlorine (mg/l)	FreeCl	Water quality	396194	0.32	0.25
Total Chlorine (mg/l)	TotalCl	Water quality	373708	0.70	0.28
Temperature (C)	Temperature	Water quality	240095	9.30	4.87
Flow cytometry Intact Cell Counts (cells per ml)	FC_ICCs	Water quality	51753	2767	19642
Flow cytometry Total Cell Counts (cells per ml)	FC_TCCs	Water quality	51753	34272	143257
Number of days from the SR cleaning date (days)	DaysFromCleaningDay	Calculated	269795	11.6	26.6

Finally, as regards the impact of the SR cleaning in the bacteriological activity inside the tanks, the SOM analysis indicates a small decrease of the HPCs numbers in the chloraminated systems after the cleaning (bottom left of the map) and a small reduction of the high ICCs clusters after the cleaning in both chloramine and chlorine systems. However, TCCs numbers appear to not be affected by cleaning of the SRs.

c) What is the impact of the WTWs to SRs bacteriological activity?

A different SOM was created (Figure 5.4) to answer the above question. The parameters introduced in this SOM were some of the parameters measured in the WTWs that supplied the SRs to investigate the impact of the WTWs in the bacteriological activity inside the SRs. More specifically, the parameters used in this SOM were the age of water exiting the SRs, the HPCs at 22^{o}C, the free and the total chlorine, the flow cytometry TCCs and ICCs, the monthly average total organic carbon (TOC) in the WTWs, the monthly average Temperature of water exiting the WTWs, the monthly average TCC exiting the WTWs and daily average precipitation per month in the works. As in the previous cases, a post analysis labelled SOM for the disinfection was also produced where the blue cells corresponded to the chlorination systems and the green cells corresponded to the chloramination systems.

The observed correlations in the previous two SOMs were also observed in this SOM (TCCs correlation with chloraminated systems, HPCs with chloraminated systems and high HPCs and ICCs correlation in both systems when the water age is high, and the free chlorine is low). This SOM shows, also, that most of the medium to high WTW TCCs and WTW TOC clusters are correlated with the chloraminated systems, therefore, they are also correlated with the high TCCs clusters and the high HPC clusters. However, the 3 different high TCC clusters - the first in the top left to middle of the plane, the second in the middle and the last one in the bottom of the plane - appear to not have any impact on the SR TCCs in both systems. Out of these three clusters, the top one correlates with the high WTWs water temperature clusters and medium to high WTWs TOC clusters and the bottom clusters appear to have a clear correlation with the medium to high precipitation clusters.

High HPCs clusters are correlated with high and medium WTW temperature of water in the chloraminated systems. Most of their numbers decrease when the WTW temperature of water decreases, however some small clusters of high HPCs appear to correlate with low WTW temperature of water as well. A similar correlation is also observed with the high WTW TOC clusters.

Finally, ICCs appear in both systems and, in addition to the correlations mentioned above and observed in the previous SOMs, high ICCs clusters are correlated with high and medium WTWs TOC clusters and high WTWs water temperature clusters.

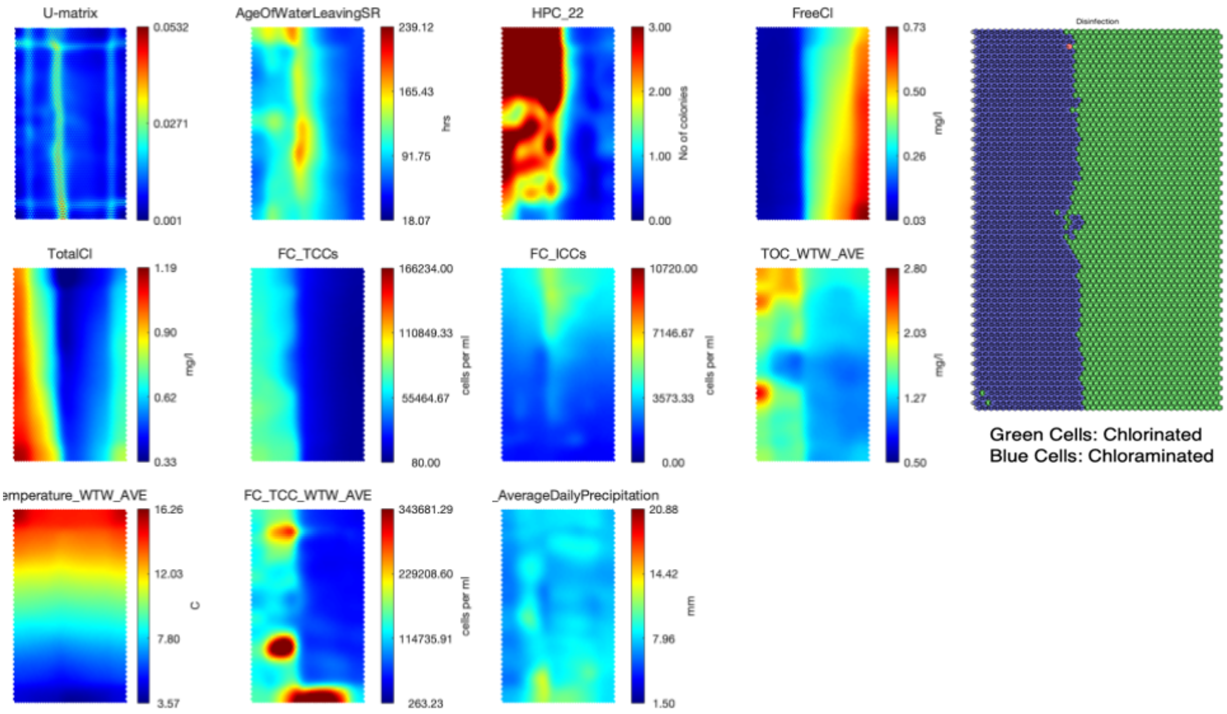


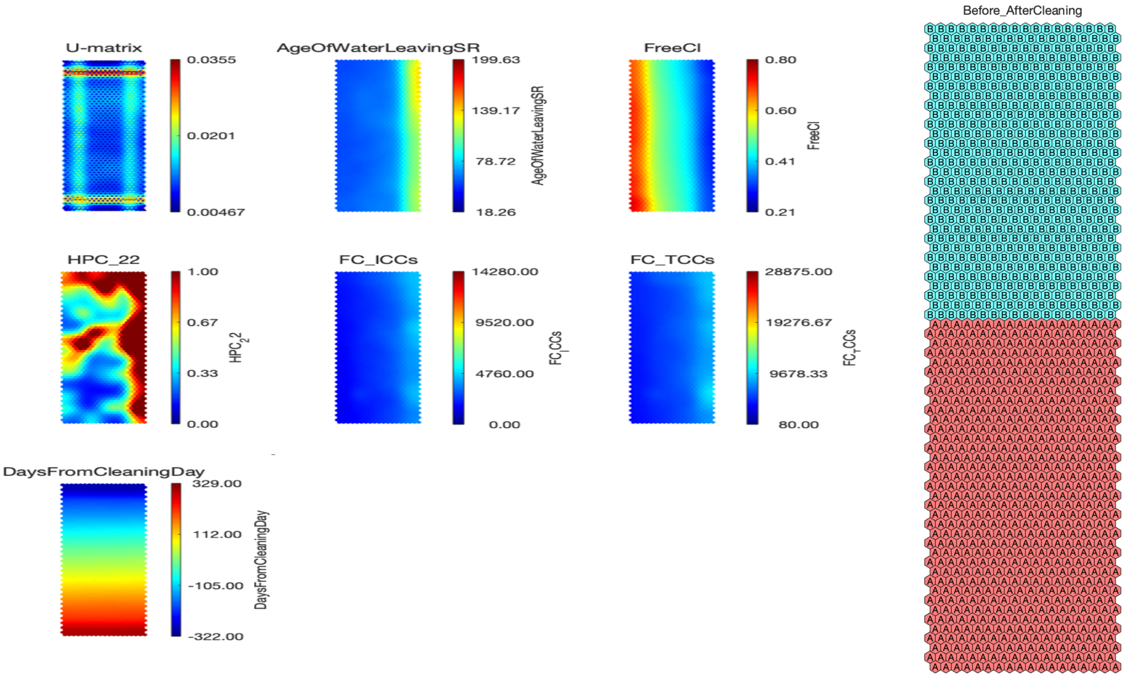
Figure 5.4:SOM for WTWs impact in the SRs, including secondary disinfection labelled map

Table 5-3: Summary of variables for SOM presented in figure 5.4

Variable (units)	Variable short form for SOM	Data Source	Number of samples	Average value	Standard deviation
Age of water leaving the SR (as sum of the retention time of this SRs and the retention time of the previous SRs that the water passed through - hrs)	AgeOfWaterLeavingSR	SW Asset files	401926	86.9	77.9
HPCs @22C (No of colonies)	HPC_22	Water quality	394206	1.95	16.61
Free Chlorine (mg/l)	FreeCl	Water quality	396194	0.32	0.25
Total Chlorine (mg/l)	TotalCl	Water quality	373708	0.70	0.28
Flow cytometry Intact Cell Counts (cells per ml)	FC_ICCs	Water quality	51753	2767	19642
Flow cytometry Total Cell Counts (cells per ml)	FC_TCCs	Water quality	51753	34272	143257
Monthy Average WTW total organic carbon (mg/l)	TOC_WTW_AVE	Water quality	358088	1.46	1.27
Monthy Average WTW water temperature (C)	Temperature_WTW_AVE	Water quality	292130	9.45	4.13
Monthy Average WTW Flow cytometry total cell counts (cells per ml)	FC_TCC_WTW_AVE	Water quality	292130	116099	808274
Average daily precipitaton per month in the WTWs (mm)	WTW_AverageDailyPrecipitation	Met Office	362260	7.97	9.51

d) What is the impact of SR cleaning to the bacteriological activity inside them?

As the SOM in figure 5.3 showed, the cleaning of the SR could have a positive impact regarding the bacteriological activity in the SRs. However, due to the clear correlations between the other parameters the actual impact of the cleaning was not clear. Therefore, two further SOMs were produced for understanding that impact, one for the chloraminated systems and one for the chlorinated systems. To produce these two SOMs, the initial dataset was subdivided into two datasets, one that included all the chloraminated SRs, and one that included all the chlorinated SRs. In addition, to see the actual impact, from these two datasets only the samples that were taken up to a year (365 days) before or after the official cleaning date were included in the final analysis. Overall, the total number of samples included in the chlorinated SOM (Figure 5.5) were 61520 and in the chloraminated SOM (Figure 5.6) were 34111. As in the previous SOMs, flow cytometry TCCs and ICCs, and HPCs were used as bacteriological indicator parameters. The number of days from SR official cleaning date was also used in both SOMs. This time though, free chlorine was used only in the chlorinated SOM analysis and total chlorine was used in the chloraminated SOM analysis. In addition, the age of water was included in the chlorinated SOM analysis, as in the previous SOMs appears to be a crucial factor regarding the increase of bacteriological activity in these systems. Finally, for both SOMs a post analysis labelled SOM was created to indicate the pre post cleaning clusters.



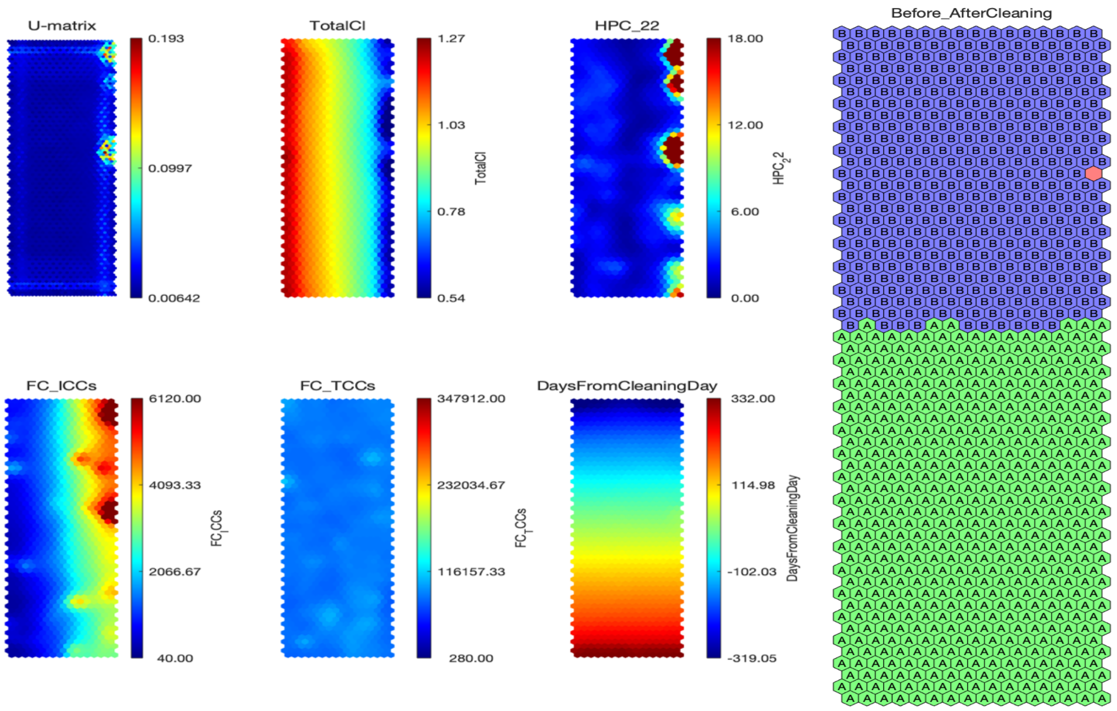
Cyan cells: Before cleaning

Blue cells: Post cleaning

Figure 5.5:SOM for cleaning impact in the chlorinated SRs, including pre/post cleaning labelled map.

Table 5-4: Summary of variables for SOM presented in figure 5.5

Variable (units)	Variable short form for SOM	Data Source	Number of samples	Average value	Standard deviation
Age of water leaving the SR (as sum of the retention time of this SRs and the retention time of the previous SRs that the water passed through - hrs)	AgeOfWaterLeavingSR	SW Asset files	61413	78.75	73.32
Free Chlorine (mg/l)	FreeCl	Water quality	59852	0.48	0.19
HPCs @22C (No of colonies)	HPC_22	Water quality	60139	0.89	11.16
Flow cytometry Intact Cell Counts (cells per ml)	FC_ICCs	Water quality	11436	2925	6397
Flow cytometry Total Cell Counts (cells per ml)	FC_TCCs	Water quality	10425	6397	21768
Number of days from the SR cleaning date (days)	DaysFromCleaningDay	Calculated	61518	18.5	205.8



Blue cells: Before cleaning

Green cells: Post cleaning

Figure 5.6: SOM for cleaning impact in the chloraminated SRs, including pre/post cleaning labelled map

Table 5-5: Summary of variables for SOM presented in figure 5.6

Variable (units)	Variable short form for SOM	Data Source	Number of samples	Average value	Standard deviation
Total Chlorine (mg/l)	TotalCl	Water quality	29992	0.96	0.23
HPCs @22C (No of colonies)	HPC_22	Water quality	33089	4.49	24.15
Flow cytometry Intact Cell Counts (cells per ml)	FC_ICCs	Water quality	4523	2503	15229
Flow cytometry Total Cell Counts (cells per ml)	FC_TCCs	Water quality	4278	84871	250505
Number of days from the SR cleaning date (days)	DaysFromCleaningDay	Calculated	34109	25.4	206.5

As expected, in both SOMs there is a clear correlation between the pre-cleaning days clusters and the pre-cleaning labelled clusters and between the post-cleaning days clusters and the post-cleaning labelled clusters. This is another indication that the SOM clusters' analysis is accurate.

SOM analysis indicates that there is some positive impact regarding the reduction of the HPCs in the chlorinated systems as the clusters with high HPCs are reduced after the cleaning (figure 5.5). However, it is clear that the cleaning has no impact in the ICCs and TCCs as there is no change in the pre and post cleaning clusters. Moreover, high ICCs and TCCs clusters are clearly correlated to each other and to low free chlorine and high age of water clusters. In addition, the high HPCs clusters after the cleaning are also correlated with the above clusters as well. Contrariwise, SOM analysis in the chloraminated SRs (figure 5.6) shows that both ICCs and HPCs are reduced after the cleaning of the SRs. Furthermore, clusters with high values for both of those parameters and in both pre and post cleaning conditions, appear to be correlated with low total chlorine clusters (right part of the planes). TCCs though in the chloraminated systems could appear either before or after the cleaning of the SRs with no clear correlation to any other parameter as well.

5.4.2. Principal Components Analysis

PCA was also applied in the SR water quality dataset to investigate bacteriological activity in SW's SRs. For a better comparison between the two methods, the parameters used at each different SOM analysis to answer each one of the research questions mentioned at the beginning of this section, were also used in the PCA analysis. However, PCA's inability to deal with missing data in the input matrix, reduced the original dataset to a dataset that included only the samples where all the selected for the analysis parameters were measured. The number of samples included in the final dataset for each research question is presented in the corresponding tables.

a) Which are the main factors of bacteriological failures on SRs?

PCA was applied using the 8 parameters used in the first SOM (age of water, temperature, HPCs at 22°C, free and total chlorine, flow cytometry cell counts, the bacteriological failures per year and average daily precipitation in the SRs). The first 2 principal components (PCs) described the 35% of the variance (22.83% and 13.83% respectively), an indication that more components are required to fully describe the dataset. However, the variance described in each one of the other 6 PCs is significantly smaller, therefore the first two components were selected for visualising the relationships between the various parameters of the dataset in a two-dimensional space. As mentioned in chapter 3, each PC is a linear combination of the various parameters, and it is completely uncorrelated to the other PCs.

Figure 5.7 shows the biplot of the PCA applied for understanding the factors of bacteriological failures in the SRs. The x-axis represents the first PC and the y-axis the second PC. All the parameters, apart from the average rainfall precipitation and the TCCs appear to be influential in the PCs. Bacteriological failures and free chlorine appear to be the most influential parameters in the first PC and have less influence in the second. These two parameters also point in reverse directions; therefore, it could be implied that there is a reverse correlation between bacteriological failures and free chlorine. Age of water, total chlorine and TCCs appear to have the same influence in both PCs however only total chlorine and TCCs are pointing in the same direction, an indication of high correlation between those two parameters. Temperature is the main parameter influencing the second PC with rainfall and HPCs being the second and the third parameter respectively most related to the second PC. HPCs at 22°C are in between the age of water and the temperature and point to the part of the plot. This is an indication that HPCs have a small linear correlation with age of water and temperature. This small correlation appears also between bacteriological failures and the age of water.

Table 5-6 indicates that only 12734 out of the 405462 samples represent only 3% of the dataset. Therefore, it is understood that PCA does not represent the whole of SW's SRs dataset.

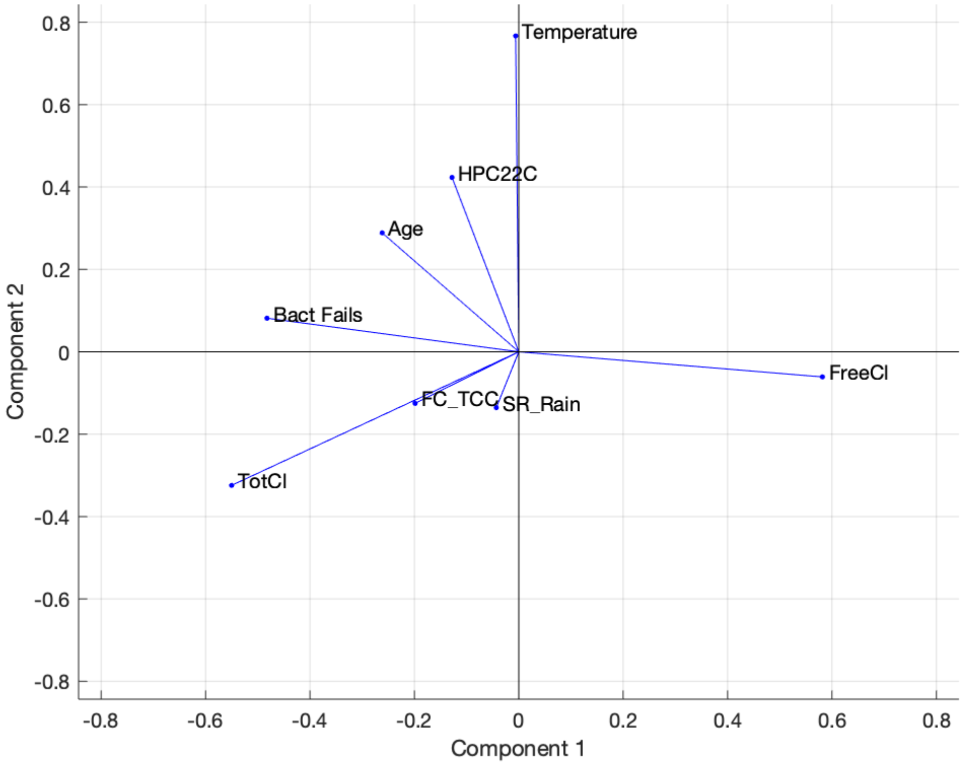


Figure 5.7: PCA biplot for bacteriological failures in the SRs

Table 5-6: Summary of variables of PCA biplot presented in figure 5.7

Variable (units)	Variable short form in PCA biplot	Data Source	Number of samples	Number of samples used	Average value	Standard deviation
Age of water leaving the SR (as sum of the retention time of this SRs and the retention time of the previous SRs that the water passed through - hrs)	Age	SW Asset files	401926	12734	86.99	70.13
HPCs @22C (No of colonies)	HPC_22	Water quality	394206	12734	2.31	18.19
Free Chlorine (mg/l)	FreeCl	Water quality	396194	12734	0.29	0.24
Total Chlorine (mg/l)	TotCl	Water quality	373708	12734	0.75	0.33
Temperature (C)	Temperature	Water quality	240095	12734	9.30	3.84
Flow cytometry Total Cell Counts (cells per ml)	FC_TCCs	Water quality	51753	12734	40280	219483
Bacteriological failures (events per year)	FC_TCCs	SW Asset files	61278	12734	1.28	0.69
Average daily precipitation per month in the SRs (mm)	SR_Rain	Met Office	355403	12734	10.93	21.67

b) Which are the main parameters correlated with increase in the bacteriological activity in the SRs?

PCA was applied using the 8 parameters used in the second SOM (age of water, temperature, HPCs at 22^{o}C, free and total chlorine, flow cytometry intact and total cell counts and the days from the SR cleaning day). The first 2 principal components (PCs) described the 37.5% of the variance (21.21% and 16.44% respectively) which, as in the previous PCA, is an indication that more components are required to fully describe the dataset. As table 5-7 demonstrates this PCA is more representative than the previous one, however even in this one, only 11% of the samples are used.

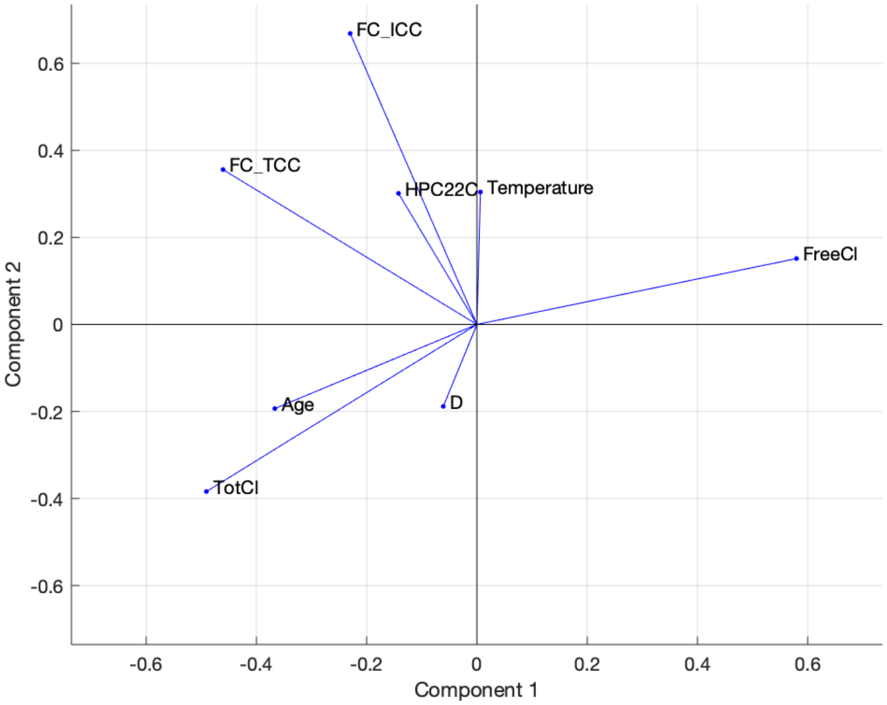


Figure 5.8: PCA biplot for bacteriological activity in the SRs

Table 5-7: Summary of variables of PCA biplot presented in figure 5.8

Variable (units)	Variable short form in PCA biplot	Data Source	Number of samples	Number of samples used	Average value	Standard deviation
Age of water leaving the SR (as sum of the retention time of this SRs and the retention time of the previous SRs that the water passed through - hrs)	Age	SW Asset files	401926	46877	87.32	68.46
HPCs @22C (No of colonies)	HPC_22	Water quality	394206	46877	1.59	15.08
Free Chlorine (mg/l)	FreeCl	Water quality	396194	46877	0.33	0.24
Total Chlorine (mg/l)	TotCl	Water quality	373708	46877	0.71	0.30
Temperature (C)	Temperature	Water quality	240095	46877	9.22	3.92
Flow cytometry Intact Cell Counts (cells per ml)	FC_ICCs	Water quality	51753	46877	2614	18190
Flow cytometry Total Cell Counts (cells per ml)	FC_TCCs	Water quality	51753	46877	32812	140337
Number of days from the SR cleaning date (days)	D	Calculated	401022	46877	904.4	1427.3

Figure 5.8 shows the biplot for this PCA. The parameters with the most influence in the PCs are the ICCs, the TCCs and the total and free chlorine. These parameters are influencing both PCs with ICCs appearing to be the only one out of these parameters more related to the second PC than the first. Temperature appears, again, to influence the second PC, however it is not as influential as in the previous PCA. Days from SR cleaning day is the least influential

parameter in this PCA. As regards the correlations, this PCA indicates that there is high correlation between ICCs and HPCs, and a smaller correlation between these two parameters and TCCs and Temperature. There is also a significantly high correlation between the age of water and the total chlorine, and, finally, a reverse correlation between those two parameters and the free chlorine.

5.5. Discussion

5.5.1. Understanding bacteriological activity in the SRs

The first SOM (Figure 5.2) indicates that the over the investigation period the coliform appearance in the SRs is correlated with high water age exiting the SRs (as sum of the retention time of all the SRs that the water passed), high water temperature, low free chlorine residual in the chlorinated systems and low total chlorine in chloraminated systems. The correlation between high temperature, low chlorine residual and coliform failures was also found in various studies on coliform bacteria regrowth on DWDS (LeChevallier, Welch, and Smith 1996; LeChevallier 1990; Besner et al. 2002). However, these studies were concentrated in the coliform regrowth in the network and thus there were only mentioning stagnation as another factor influencing the phenomenon. The importance of retention time as a factor that influences bacteriological regrowth was mentioned in other research works that were concentrating in the general bacteriological activity in the DWDS. Kerneïs et al. (Kerneïs et al. 1995) found that the higher SR retention time is, the higher are the HPCs inside them. Prest et al. (Emmanuelle I. Prest et al. 2016) in their review paper indicated that all three aforementioned parameters are influencing bacteriological activity in the DWDS. The SOM findings in this work agree with both works as both high ICCs and high HPCs appear in systems with high temperature, low chlorine residual and high age of water (Figures 5.2-5.4).

All the aforementioned research works indicated the importance of temperature as the key factor that influences all the biological processes inside the DWDS, and it is one of the main factors that control the disinfectant decay. In this research this is demonstrated in the SOM analysis (Figures 5.2 - 5.6) where in general low total and free chlorine clusters are correlated with medium to high and high temperature. Controversially, the PCA analysis shows no relationship between temperature and total or free chlorine. This finding demonstrates that this temperature - chlorine decay relationship is more complicated and does not follow the linear relationship that PCA is able to identify.

In contrast, water age (sum of the SR's retention time and the retention time of all the SRs that the water passed after exiting the WTWs and before reaching this SR), has clear linear reverse correlation with free chlorine in the PCA analysis (Figures 5.7-5.8). A simple explanation of the above finding is that the higher the water remains into the SRs the more reactions between the disinfectant residual and the nutrients are happening which means weaker disinfection and thus the bacteriological activity increases. However, as we saw above in such complex systems there is not only one direct influence between one parameter and another but multiple parameters that affect the water stability.

Monochloramine has been proven to be more effective than chlorine in controlling coliform regrowth, even if it is a weaker disinfectant, because it is less-reactive disinfectant (Camper 2014; Mark W. LeChevallier 1990). The SOM analysis in this work indicates this weakness as in every SOM high TCCs and most of the high HPCs are correlated with chloraminated systems. In addition, it also indicates that with high temperatures and high age of water, chloramine disinfectant is also reduced and as a result, coliform bacteria appear in these systems as well.

SOMs analysis in the WTWs influence on the SRs' bacteriological activity (Figure 5.4) indicates that the potential source of high TCCs and high HPCs in the chloraminated SRs are the high numbers of TCCs in combination with the high TOC levels exiting their WTWs. There is no clear explanation regarding the TCC levels exiting the WTWs in the chloraminated systems, however it is an indication that in the WTWs of some of these DWDS the time after adding chlorine and before adding ammonia in the water is not sufficient and thus bacteria cells are exiting the works. The same SOM also indicates high ICCs in chlorinated and chloraminated SRs correlating with high TOC exiting the WTWs and high water temperature. High HPCs that belong in the chlorinated systems also correlate with these two factors. The importance of organic and inorganic nutrients in governing the bacteriological regrowth in the drinking water distribution system has been indicated in various research works (E. I. Prest et al. 2016; K. E. Fish, Osborn, and Boxall 2016). LeChevallier et al. (M. W. LeChevallier, Schulz, and Lee 1991) used a DWDS in the US as a case study for understanding the factors of bacteriological regrowth and found that there is a clear relationship between coliform occurrences and high TOC levels. High carbon levels in the drinking water is utilized by heterotrophic bacteria as an energy resource and, therefore, it contributes to their increase (Mark W. LeChevallier 1990). As mentioned above, the findings in this work agree with these studies and indicate the importance of TOC on bacteriological regrowth in the DWDS. However, it remains unclear the reason that most of the high TOC concentration clusters exiting the WTWs appear in the chloraminated systems.

This work indicates that in chloraminated systems there is no clear correlation between TCCs and ICCs (Figures 5.3, 5.4, 5.6). On the contrary, high ICC and high TCC clusters are fully correlated in the chlorinated systems (Figure 5.5). By checking the actual numbers of cells in both systems, TCC numbers in the chloraminated systems are much higher compared to the ones in chlorinated systems (340000 - 28000) but the opposite occurs with the ICC numbers (6000-12000). This finding, in combination with the appearance of the ICCs in systems with low free chlorine residual and high-water age, indicate, firstly, the importance of maintaining steady free chlorine residual in chlorinated systems and secondly the steadiness of monochloramine as disinfectant even if it is a weaker disinfectant to chlorine. It is interesting that these results agree with the findings of a study by Gillespie et al. (Gillespie et al. 2014) where flow cytometry was applied in three DWDS with different disinfection types (one chlorinated DWDS and two chloraminated DWDS) for understanding the microbiological differences in these two systems.

SOMs' analysis demonstrated the role of heavy rainfall in both the increase of bacteriological activity exiting the WTWs (Figure 5.4) and in the coliform - *E. Coli* occurrence in the SRs (Figure 5.2). In the first case, the high TCC clusters in the WTWs correlation with medium to high intensity average daily rainfall clusters indicates a potential deterioration in the quality of the raw water that was feeding these WTWs. There are various studies that have associated heavy rainfall with increased bacteria and pathogens numbers in the DWDS (M. W. LeChevallier, Schulz, and Lee 1991; Geldreich 1996; Kumpel and Nelson 2013). Potential factors that could relate heavy rainfall with bacteria intrusion into the drinking water are the increase of organic matter in the raw water, the appearance of increased bacteria numbers in the raw water and the slow adaptation response of the WTWs in the new conditions. LeChevallier et al. (M. W. LeChevallier, Schulz, and Lee 1991) showed that there is a time lag between heavy rainfall and coliform appearances in the DWDS. Unfortunately, with the absence of the time element in SOMs analysis, it is not possible to identify this time lag with this analysis. However, a further SOMs analysis using water quality data taken from the raw water may indicate the factors related to this TCCs increase during heavy rainfall periods. In the second case, knowing that all the SRs in SW's network are fully covered and most of them are underground or semi-underground, the correlation between the bacteriological failure clusters (*E.Coli/*coliform events) and heavy rainfall clusters could be an indication of ingress of contaminated water in the SRs through cracks in their structures. Another potential explanation of this correlation could be that the rainfall has transferred nutrients from the soil into the SRs through these cracks. However, there are not enough SR monitoring water quality samples where nutrients (e.g TOC) are measured to further investigate their relationship with rainfall in the SRs using SOMs. Besner et al. (Besner et al. 2002) presented a case study where coliforms appeared in a SR due to issues in the structure that forced the utility company to immediately repair the crack. The findings here cannot directly relate potential structural problems to specific tanks, however, they indicate the importance of adequate maintenance of the SRs.

Finally, the last two SOMs outputs (Figures 5.5, 5.6) show some interesting findings regarding the impact of the SRs cleaning on drinking water quality. Both SOMs indicated that there is a reduction of the high HPC, TCC and ICC clusters after the cleaning in both systems. However, this reduction is more significant in the chloraminated SRs where both HPCs and ICCs are reduced after the cleaning. In these systems, the high ICC and HPC clusters correlate with low total chlorine either before or after the cleaning of the SRs. This finding indicates that in the chloraminated systems, high water quality exiting the WTWs could be maintained in these levels by keeping a steady total chlorine residual and by systematically cleaning the SRs.

Controversy, in the chlorinated systems cleaning is contributing in the improvement of water quality exiting the works, but the main factor that is required to be controlled is the SRs retention time (water age in this study is the sum of the retention time in all SRs that water passed through) that also contributes to the free chlorine decay.

5.5.2. Comparison of PCA and SOMs for water quality samples analysis

SOMs are mainly used for clustering data and PCAs' main use is for dimensionality reduction of large datasets. As this research demonstrated, both methods can also be applied for visualisation of large datasets with multiple parameters and for identification of relationships between those parameters. However, there is a clear difference between those two methods, regarding the path that they follow to identify relationships between variables, the required conditions for their application and the visual outputs that they provide.

PCA follows an algorithm to provide linear relationships between the variables (Smith 2002). This approach allows the identification of only linear correlations between various parameters and thus complex non-linear relationships are not covered. Due to the complex reaction mechanisms that are taking place in the DWDS, most of these hidden relationships between the various water quality parameters could not be identified by PCA when used for this type of analysis. This argument is reinforced by the fact that in both PCA examples presented here, the first two PCs were capturing less than 40% of the dataset's variance, an indication that more PCs were required to fully capture the data and that the relationship between those parameters is more complex.

PCA also requires an input matrix with no, or few missing data as explained in a previous chapter. Monitoring water quality datasets, in general, have a lot of missing or non-measured parameters per sample. Tables 5.6 and 5.7 show the number of samples out of all the available data that were used for the analysis - 3% of the data for the first PCA and 11% for the second). It is, therefore, clear that some trends in the data and some correlations between the various parameters were not captured by PCA analysis.
PCA's visualisation output makes it easy to understand which parameters were the most important for each PCA. It is also easy to see which parameters are clearly correlated, reverse correlated or uncorrelated with other parameters. However, the findings in figures 5.7 and 5.8 indicate that most of the relationships between the water quality parameters are more complex and non-linear apart from some clear exceptions explained in the above section.
SOM technique, in contrast to PCA , follows a non-linear approach for doing the analysis (S. R. Mounce et al. 2016). Therefore, it captures nonlinear more complicated relationships between various parameters and visualising them into two dimensional planes. In this research, SOM's analysis showed more clear correlations than PCA. In addition, the post-analysis labelled maps separated the data into categories (chlorinated vs chloraminated DWDS, pre cleaning and post cleaning data) and gave a better understanding of the

bacteriological activity in the different systems. SOM's ability to ignore missing inputs when finding the corresponding outputs, enabled the use of all the available data in the analysis as the accompanying to SOM tables show. Finally, as regards the actual visualisation outputs, SOM are easy to follow and, thus, it is easier to identify not only the strong correlations, but the weak ones as well. Therefore, it could be said that for an analysis on the water quality samples dataset, SOM is a better technique than PCA.

5.5.3. Operational value of the findings

SOMs have proven to be a very useful tool for discovering relationships between various water quality parameters and trends in the water behaviour. SOMs application in the Scottish water SRs dataset, as presented in this chapter, provided some interesting findings regarding the factors that increase the bacteriological activity and that are potentially related to the SRs bacteriological failures as described in the previous section. Moreover, the application of SOMs for understanding the impact of the WTWs, of the SRs' cleaning, and of the rainfall in the bacteriological activity inside the SRs indicated the need of the following managerial actions by Scottish Water:

- Control and optimisation of the retention time in SRs of both disinfection systems but mainly in the chlorinated SRs where retention time appears to be the main parameter that reduces the free chlorine residual
- Systematic check of the need for a secondary disinfection in the chlorine SRs especially when the water temperature increases
- Systematic cleaning of the chloraminated tanks to reduce the bacteria numbers entering the system
- Systematic check of the SR condition to reduce ingress of contaminated water or nutrients from the soil to the drinking water
- Management and reduction of the TOC concentrations exiting the WTWs
- Improvement of the WTWs' automatization to adapt in sudden changes of the quality of the raw water that feeds them

SOMs are not a method that could be used directly by the WUs' operators to prevent water deterioration in the short-term future. They require data exploration and processing prior to their application. However, by changing the procedures that WUs store and collect water quality data, SOMs could become a really important tool in understanding the general behaviour of the water in the utilities systems. Thus, they could become a supporting tool for the WUs senior managers and water quality experts in their decisions for interventions in their DWDS.

As presented above, PCA is not as powerful as SOMs as regards the analysis of water quality data. However, there are two key findings of the PCA analysis of this work. The first one is the inverse linear correlation between the free chlorine and age of water that indicates the importance of controlling the water circulation. The second one is the linear correlation between temperature, HPCs and ICCs that indicates the importance of a residual disinfection management during the warm periods of the year.

5.6. Conclusions

Two unsupervised data-driven techniques, PCA and SOMs, were applied on Scottish water SRs dataset, created using monitoring water quality samples taken between January 2012 and May 2020, for identifying the correlations between various water quality parameters and understanding the factors that influence the bacteriological activity inside them. The results obtained from this investigation are as follows:

- SOMs is a better technique than PCA for mining monitoring water quality samples datasets and discovering correlations between water quality parameters. It tackles the main issue of the sparse data that these types of datasets have, it is able to identify more complex non-linear relationships between various water quality parameters and its visualisation output could be understood by stakeholders without any ML background knowledge.
- The main factors that contribute to bacteriological failures are high temperature, high water age, low chlorine residual and high rainfall in the SRs.
- Higher bacteriological activity has been found in the SRs that belong to chloraminated systems as there are higher TCC numbers in these SRs than in the chlorinated systems counterparts. However, the large numbers of live bacteria cells (ICCs) in both disinfection systems' SRs when low disinfection residual and high water temperature conditions exist, indicate that disinfection type is not a main factor for the SRs bacteriological failures.
- Bacteriological failures' correlation with high precipitation in the SRs indicates a potential issue with the structures and the condition of some of Scottish Water's SRs.
- The correlation between high rainfall in the WTWs with high TCCs in the WTWs outlet is an indication of a slow adaptation of the WTWs processes in the new raw water quality conditions.
- Systematic SR cleaning reduces the risk of increased bacteriological activity in the chloraminated SRs.
- For the chlorinated SRs, their systematic cleaning will not guarantee a bacteriological activity reduction and controlling and reducing their retention times is also required.
- PCA findings indicate a linear inverse correlation between free chlorine residual and age of water. However, this finding requires further investigation as due to the sparce nature of the dataset and PCA's inability to ignore missing data, only a small part of the data was analysed.

6. A SOMs application on water quality datasets for investigating the impact of switching disinfection type on drinking water quality

6.1. Introduction

Chlorination is the most common type of disinfection that WUs use. It is the process of adding small doses of chlorine in the drinking water before exiting the WTWs to guarantee that the water is free of bacteria and other microorganisms. The main advantage of chlorination is that it forms a residual that could be maintained in the water for long periods and, thus, could limit the bacteriological regrowth during its transportation to the consumers taps (K. E. Fish et al. 2020). However, chlorine, when in high concentration in the water, reacts with natural organic matter (NOM) or inorganic ions to produce disinfection by-products (DBPs) that are related to carcinogenic diseases (Parsons and Jefferson 2009). The common policy for the WUs when they notice an increase of DBPs in a DWDS, is to switch its disinfection type from chlorination to chloramination by adding ammonia at the same time as chlorine is added in the WTWs. This reaction produces chloramines which are weaker disinfectants comparing to chlorine but they also react less with NOM and have longer retention time (Parsons and Jefferson 2009).

The impact in drinking water quality of switching disinfection from chlorination to chloramination has not been largely investigated. However, some of the potential impacts have been reported in American Water Works Association (AWWA) publication by Dyksen et al. (Dyksen et al. 2008). In this report, they investigated 11 DWDS that switched their disinfection from chlorination to chloramination and they found an overall improvement of the quality of the water after the switching in terms of an overall DBPs and HPCs reduction in addition to less odour and taste complains by consumers. However, they also noticed increased levels of metals (lead and copper) and a higher risk of nitrification.

In this chapter, a different investigation over the impact of switching disinfection type from chlorination to chloramination using SOMs is presented. The motivation for this investigation was one of SW's DWDS where iron concentrations in water quality samples continuously exceeded the acceptable limits and an increased number of consumer contacts reporting discolouration was noticed after its secondary disinfection type was changed from chlorination to chloramination. This DWDS is a large system in the southwest of Scotland. Its WTW high capacity is 120 Ml/day and serves a population of more than 210000 people through a complicated network of more than 1400km and more than 60 DMAs. The decision of switching disinfection residual was made in 2014 as a solution for reducing the increased concentrations of trihalomethanes (THMs) in the system. The required for the switch works in the WTW were officially finished by the end of 2017 and in April 2018 chloramination disinfection commenced. Days after the switch, more than 10 consumer contacts reporting taste and odour were noticed and during the following months more than 10 water samples exceeded the iron standard. By the end of the year 2018, the total number of customers complaining about discolouration incidents was way higher compared to the previous two years. SW's investigation for the incidents indicated that a possible cause of the discolouration issue in this DWDS was the destabilisation of the network's biofilm due to the changing in the water chemistry following the change to chloramination (DWQR 2019b).

This event indicated the importance of understanding the impact of changing disinfection residual on the drinking water quality prior to making such a drastic intervention in the DWDS. Thus, this chapter investigates the impact of switching disinfection type from chlorination to chloramination by applying SOMs on monitoring water quality data taken from consumers taps' samples in DWDS that have made this change during the period between 2012 and 2019. The aim of this work is not to examine the decision to switch disinfection type by SW but to understand the potential impacts that this switch could have in drinking water quality, using samples water quality data. Findings of this work could inform WUs about the factors that could influence the deterioration of the drinking water quality after the switch and avoid the same mistakes in the future. As in the previous chapter, this chapter addresses objectives 2 and 4. In this chapter there is no comparison between SOMs and other ML techniques, however the answer to objective 4 is achieved by demonstrating a different way that SOM analysis could be applied for understanding water quality.

6.2. Methods & materials

6.2.1. Data Collection & analysis

The consumers taps' dataset includes a total of 140853 samples taken in all of Scotland's DWDS during the period between 2012 and May 2020. However, for this investigation only the data from the systems that have switched from chlorination to chloramination during this period were required and extracted from the dataset. Overall, during the study period 14 systems have switched their disinfection from chlorination to chloramination. The system described in the previous section is the biggest DWDS out of these 14, serving with water more than 110000 properties. 11 DWDS are small systems with 9 of them serving water to less than 1000 properties. 1 DWDS is a medium to small system serving less than 10000 properties and 1 is medium sized DWDS (approximately 20000 properties). The final dataset for the investigation included a total of 8957 samples taken from consumers taps in these systems. More than half of these samples were taken from taps inside the DWDS A (5324 samples) and the rest of the samples were taken from all the other systems together.

6.2.2. Self-Organising Maps

As in the previous chapter, the MATLAB® SOM Toolbox version 2.1 (Teuvo Kohonen 2014) was used for creating the various SOMs. The analysis was held using MATLAB® version 2019b and the same algorithms as in the previous chapter were used for selecting the parameters and calling the Toolbox. Again here, the maps' colour range was standardized using as reference ranges the 5th and 95 percentiles of the dataset. Finally, a different algorithm was created to include the number of samples per parameter, the average, and the standard deviation in the final analysis.

6.3. Machine learning application steps

The machine learning application steps are filled as follows:

a. Define the water quality problem

In this chapter, the aim is to identify changes in the drinking water quality that may occur by switching from one disinfection type to another, which in this case is a switch from chlorination to chloramination. Therefore, the water quality problem should be defined as follows:

What is the impact of switching disinfection type to the drinking water quality behaviour?

b. Type of the available data

The water quality monitoring samples taken from the customers taps' outlets, the SRs outlets and the WTWs' outlets are the available data for the investigation.

c. Define required output

The required output is understanding the disinfection type change impact on drinking water quality by identifying correlations between various water parameters in the pre and post switch period.

d. Machine learning selection

By following the machine learning tree presented in chapter 3, the selected techniques should have been either the SOMs or PCA. However, the research work on chapter 5 demonstrated that SOMs is the more appropriate technique.

e. Data preparation

The data should be prepared in a way that each row is a different observation (sample) and each column represents a different water quality parameter (coliform bacteria, heterotrophic plate count, flow cytometry data etc.). The consumers taps' water quality dataset created as described in a previous chapter, will be used as input in this investigation.

f. Application output

The required output is a number of graphs that visualise the correlations between the various parameters and therefore the correlations between them could be explored. The procedure followed to produce these outputs is presented in the following sections of the chapter.

The ML application steps for this investigation are summarized in the following figure.

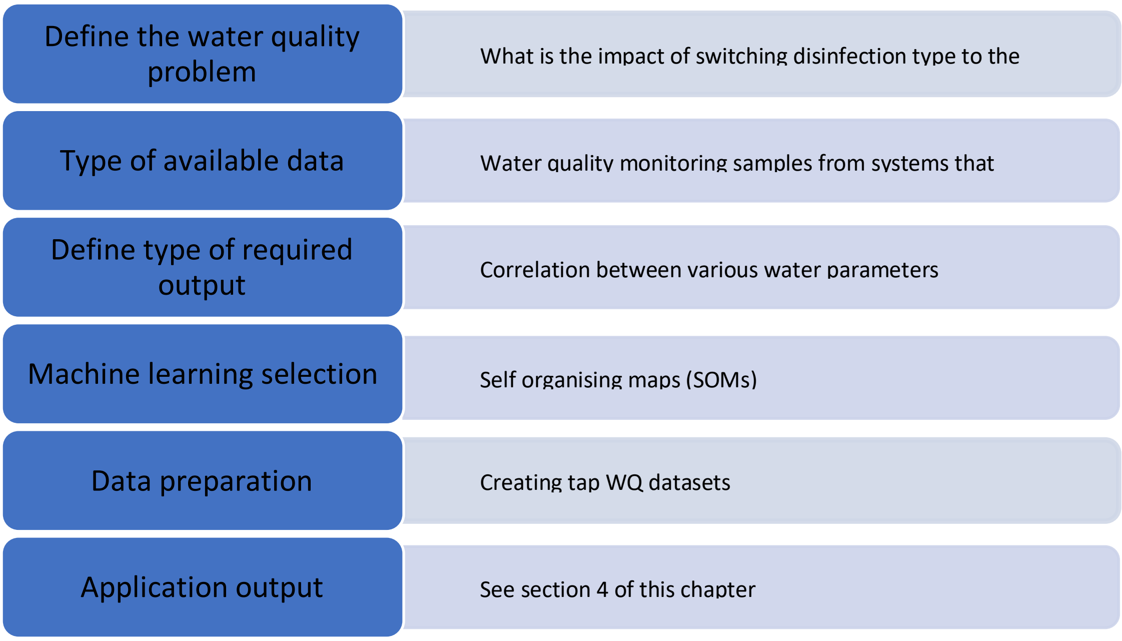

Define the water quality problem
What is the impact of switching disinfection type to the
Type of available data
Water quality monitoring samples from systems that
Define type of required output
Correlation between various water parameters
Machine learning selection
Self organising maps (SOMs)
Data preparation
Creating tap WQ datasets
Application output
See section 4 of this chapter

Figure 6.1: Machine learning application steps for the disinfection type change investigation

6.4. Results and discussion

As shown in the previous chapter, SOM analysis can identify correlations between various parameters, but it has no time element - it is not possible to visualise in the main SOM analysis the parameters correlations and clusters through time. In this investigation, however, the time issue was surpassed by using for each created SOM the post-analysis labelled disinfection map as an indicator of clusters and correlations that are related to chlorination or chloramination. In other words, parameter clusters correlating with chlorination indicate a pre-switch water condition and clusters correlating with chloramination indicate a post-switch condition. Thus, with this way SOMs will be used for investigating the impact of changing disinfectant in drinking water quality. The research questions that this chapter aims to answer are as follows:

a) Has the disinfectant change improved the quality of the drinking water in the customers taps?
b) Which factors contributed to the increased bacteriological activity after the switch?

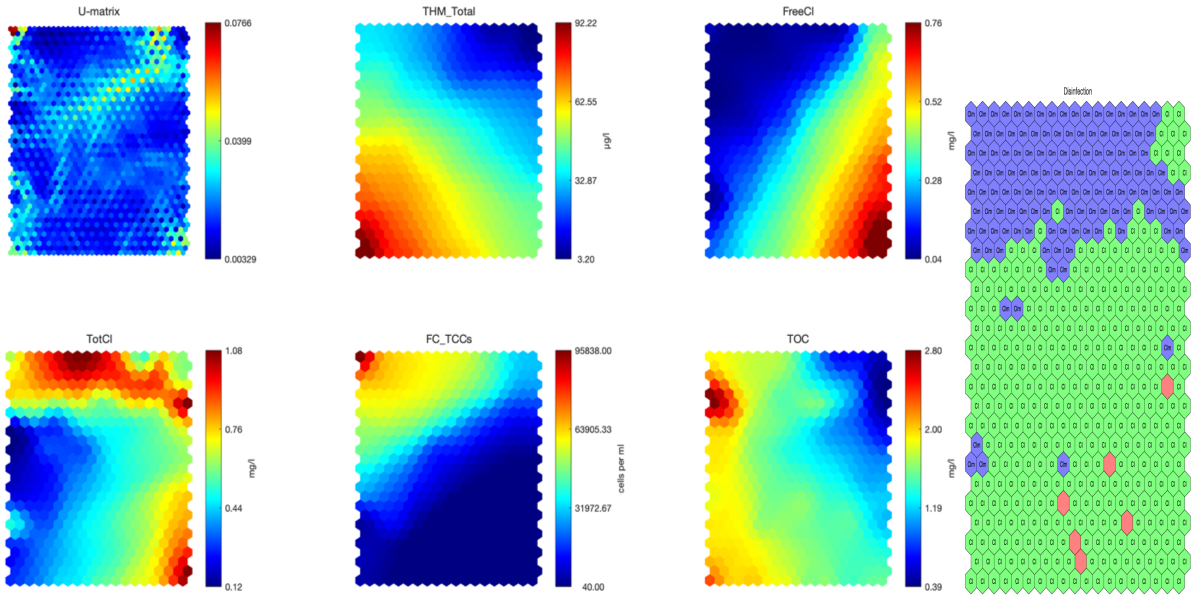
c) Is switching disinfectant the cause of discolouration?

d) Are there any indications of nitrification after changing disinfection residual?

A different SOM was produced for each one of the above research questions.

a) Has the change improved the quality of the drinking water in the customers taps?

Water utilities change disinfection type from chlorination to chloramination in DWDS usually when increased levels of disinfection by-products (most commonly THMs) are observed in the customers taps. Therefore, this SOM (Figure 6.2) investigates the impact of switching to chloramination on THMs in comparison with the impact on bacteriological activity inside the DWDS. The parameters used for this investigation are THMs, flow cytometry TCCs, free and total chlorine and total organic carbon (TOC).



Blue cells: after the switch

Green cells: Before the switch

Figure 6.2: Water quality after changing disinfection, including secondary disinfection labelled map

Table 6-1: Summary of variables for SOM presented in figure 6.2

Variable (units)	Variable short form for SOM	Data Source	Number of samples	Average value	Standard deviation
THMs total (µg/l)	THM_Total	Water quality	2128	41.82	27.94
Free Chlorine (mg/l)	FreeCl	Water quality	7354	0.30	0.26
Total Chlorine (mg/l)	TotCl	Water quality	7206	0.60	0.33
Flow cytometry Total Cell Counts (cells per ml)	FC_TCCs	Water quality	1552	17432	43328
Total Organic Carbon (mg/l)	TOC	Water quality	2051	1.54	0.88

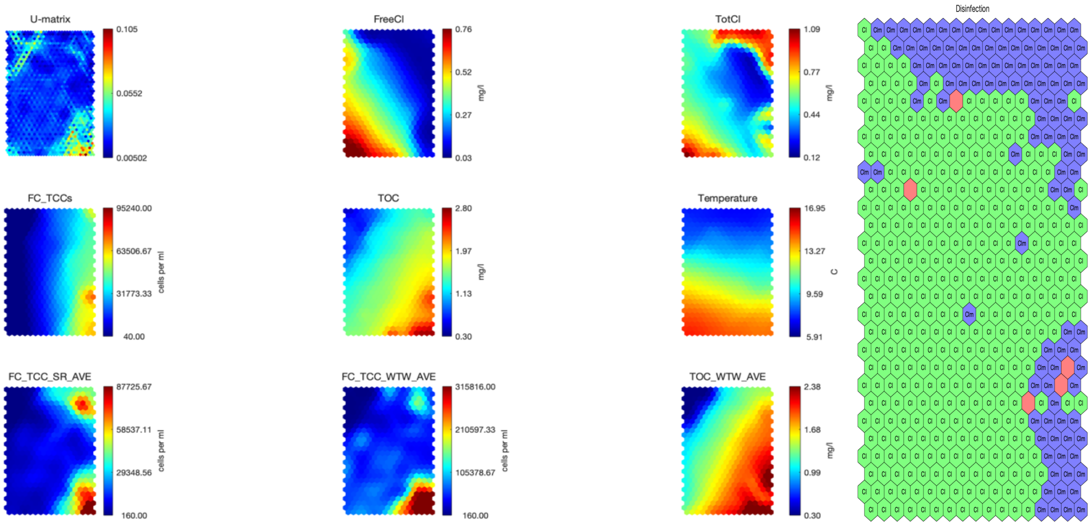
As expected, SOM analysis correlates the high chlorine clusters on the top of their map with low free chlorine clusters and the post-switch clusters (chloramination disinfection) in the labelled map. SOM analysis also indicated a reduction of THMs after the switch as low THM clusters correlate with chloramination (blue cells) in the labelled map, while high THMs are observed with chlorination (green cells) in the labelled map. However, TCCs appear to be increased after the switch, as clusters with high numbers of TCCs correlate with chloramination (blue cells). Finally, as regards the TOC, SOM analysis indicated that there was an increased number of high TOC clusters in the systems before and after the switch but there were noticeably more clusters containing the highest TOC that were related to chloramination than to those related to chlorination.

The increased TCC numbers related to chloramination clusters indicate an increase in the bacteriological activity after the switch which disagrees with the findings by Dyksen et al. (Dyksen et al. 2008). However, this output reinforces the findings presented in the previous chapter regarding the correlation between high TCCs correlating with chloramination systems and supports the consensus that chloramination is a weaker disinfection than chlorine. However, there are other factors that could contribute to this increase, such as changes in the DWDS' biofilm or the age of water arriving at the tap, for which parameters, unfortunately, there were no available data.

THMs are disinfection by products (DBPs) that are produced in chlorinated DWDS when free chlorine residual reacts with natural organic matter (NOM) and especially with organic carbon (Parsons and Jefferson 2009). This SOM shows, high and medium TOC clusters correlate with high THM clusters and medium to low free chlorine clusters. This finding indicates that the high TOC concentrations in consumers' taps are related to THMs appearance in the DWDS before the disinfection switch. The fact that there are more high TOC clusters in the chloramination systems could be related to the absence of free chlorine and the stability of the monochloramine as a disinfectant that reacts less with organic and inorganic material. These increased TOC concentrations after the disinfection switch, could be another factor that contributes to the increased TCCs noticed after the switch.

b) Which factors contributed to the increased bacteriological activity after the switch?

The first SOM indicated that there is a potential increase in the bacteriological activity of the DWDS after the disinfectant change. Therefore, a SOM (Figure 6.3) for identifying correlations between bacteriological activity and other parameters including WTWs and SR parameters was produced. More specifically, the selected for this SOM parameters were the free and the total chlorine, the flow cytometry TCCs, the TOC, the temperature of water, the monthly average flow cytometry TCCs exiting the SRs, the monthly average flow cytometry TCCs exiting the WTWs and the monthly average TOC exiting the WTWs.



Blue cells: after the switch

Green cells: Before the switch

Figure 6.3: Bacteriological activity after changing disinfection, including secondary disinfection labelled map

Table 6-2: Summary of variables for SOM presented in figure 6.3

Variable (units)	Variable short form for SOM	Data Source	Number of samples	Average value	Standard deviation
Free Chlorine (mg/l)	FreeCl	Water quality	7211	0.30	0.26
Total Chlorine (mg/l)	TotCl	Water quality	7206	0.60	0.33
Flow cytometry Total Cell Counts (cells per ml)	FC_TCCs	Water quality	1552	17432	43328
Total Organic Carbon (mg/l)	TOC	Water quality	2051	1.54	0.88
Water Temperature (C)	Temperature	Water quality	1420	11.05	4.31
Monthy Average SRs total cell counts (cells per ml)	FC_TCC_SR_AVE	Water quality	3439	18527	67117
Monthy Average WTW Flow cytometry total cell counts (cells per ml)	FC_TCC_WTW_AVE	Water quality	7566	80522	135215
Monthy Average WTW total organic carbon (mg/l)	TOC_WTW_AVE	Water quality	8820	1.50	0.59

High clusters of TCCs appear in the right of their plane with clusters containing higher values correlating with the chloramination post switch cells (blue cells). The two clusters containing the highest TCC values were a small cluster in the top right of the plane and a bigger one in the bottom right. The first one correlates with high total chlorine, high TCCs exiting the WTWs and the SRs, medium to high TOC and low water temperature. The second one correlates with low total chlorine, high temperature of water, high TCCs in both the SRs and the WTWs, high TOC and high WTWs' TOC.

As regards the correlations between the other parameters, there is a perfect correlation between TCCs exiting the WTWs and TCCs exiting the SRs and between the tap TOC and the WTWs TOC. The high clusters of TCCs in the SRs and WTWs on the top of their plane correlate with chloramination and the high clusters of TCCs in the bottom of their plane are mostly correlated with chloramination, however there is a part that also correlates with chlorination, low free chlorine, and high TOC. High TOC appear in the systems pre and post switching disinfection, however, as in the previous SOM, the clusters containing the highest TOC values are related to chloramination indicating a potential increase after the disinfectant switch.

This SOM indicates that the source of the high TOC concentration in the taps is the high TOC concentration exiting WTWs and agrees with the findings regarding the TOC concentrations of the previous chapter. In addition, this SOM agrees with the finding of the previous SOM regarding the increased TOC concentrations after switching disinfection. A further investigation over the changes that SW did in the WTWs' processes of these DWDS before switching disinfection to allow the drinking water to adapt in the new disinfection environment is required. This investigation could potentially explain further the causes of this increase in the TOC concentrations. However, the findings of this study suggest that a management and a reduction of the TOC concentrations in these DWDS is required to control the bacteriological activity.

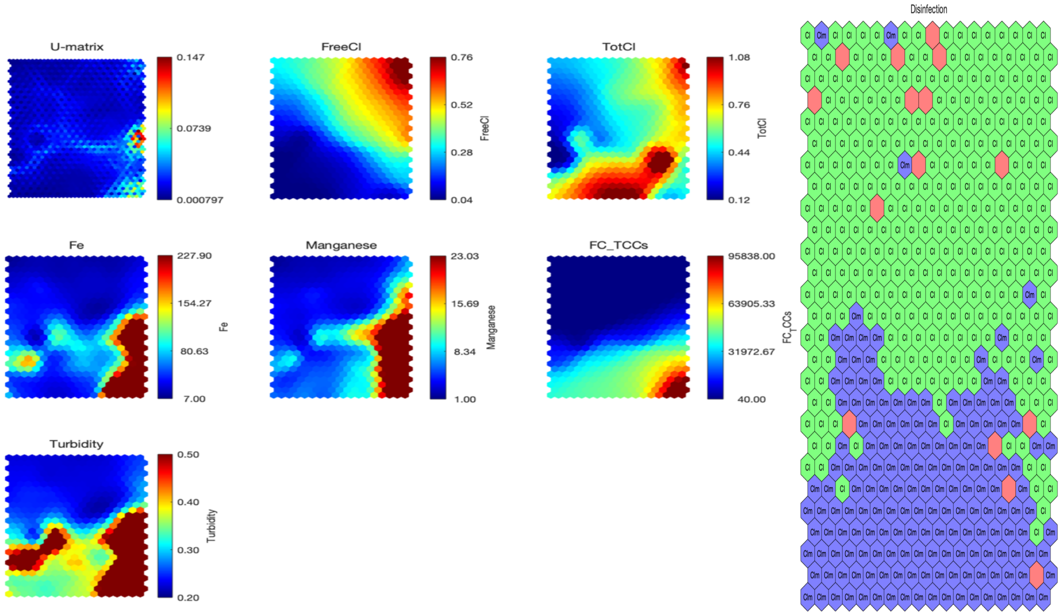
The perfect correlation between high SRs TCC clusters and high WTWs TCC clusters also agrees with the findings of the previous chapter. This finding again shows the weakness of total chlorine as a disinfection residual but, in addition, it also reinforces the hypothesis stated in the previous chapter regarding the insufficient chlorine contact time prior to ammonia addition in the WTWs. However, this is just a hypothesis and, again, a further investigation in the WTWs processes is required.

The other parameters related to high TCC clusters after the switch are low total chlorine residual and high temperature. This finding indicates again as in the previous chapter the importance of temperature in the various reactions inside the DWDS as, usually, low total chlorine residual is a result of high disinfection demand due to high water temperature conditions.

c) Is switching disinfectant the cause of discolouration?

Discolouration is related to organic and inorganic compounds (I. J. H. G. Vreeburg and Boxall 2007). The discolouration that appeared in the main DWDS described at the beginning of the chapter, was related to high iron concentration and high turbidity in the customers taps. The SOM presented here, (Figure 6.4) was produced to investigate if in general switching disinfection residual increases the risk of discolouration. The selected parameters in this SOM were iron, manganese and turbidity, as the main parameters related to discolouration, and also free and total chlorine, and flow cytometry TCCs.

There are two high iron clusters, one in the bottom right of the iron plane and a smaller one in the middle left of the plane. Both clusters correlate with high turbidity clusters and high manganese clusters. The labelled SOM indicates that this correlation appears in both systems, however the part of the clusters that belongs to chloramination cells is bigger than the one that belongs to chlorination cells. The medium and high TCCs, in general, correlate with chloramination systems, with the highest TCC clusters (bottom right of their map) correlating with low total and free chlorine and high manganese and iron clusters. Clusters with low levels of turbidity appear in the systems before the disinfectant change. All the medium to high and high turbidity clusters are located in the bottom of the plane and mostly follow the shape of the blue - post-switch chloramination cells with some exceptions where high clusters of turbidity are related to pre-switch chlorination (middle to bottom left). The chlorination high turbidity clusters are mostly correlated with high iron and high manganese clusters.



Blue cells: after the switch
Green cells: Before the switch

Figure 6.4: Discolouration after changing disinfection, including secondary disinfection labelled map

Table 6-3: Summary of variables for SOM presented in figure 6.4

Variable (units)	Variable short form for SOM	Data Source	Number of samples	Average value	Standard deviation
Free Chlorine (mg/l)	FreeCl	Water quality	7211	0.30	0.26
Total Chlorine (mg/l)	TotCl	Water quality	7206	0.60	0.33
Iron (mg/l)	Fe	Water quality	3668	82.22	550.26
Manganese (mg/l)	Mngs	Water quality	3588	8.83	67.47
Flow cytometry Total Cell Counts (cells per ml)	FC_TCCs	Water quality	1552	17432	43328
Turbidity (NTU)	Turbidity	Water quality	2945	0.48	3.51

Transition from chlorination to chloramination as a secondary disinfectant changes the water chemistry of the DWDS, which could affect the chemical balance of the network and increase the corrosion levels of metallic pipes (The U.S. Department of the Interior 2013). A couple of studies investigated an extreme case in Washington D.C., where increased levels of lead appeared in the tap water due to extreme corrosion of lead pipes and service pipes after switching the secondary disinfection of the DWDS from chlorine to chloramine (Edwards and

Dudi 2004; Edwards, Triantafyllidou, and Best 2009). The findings in this investigation show that there are high iron and manganese clusters before and after the transition to chloramination. By looking at the actual manganese and, especially, to the iron concentrations, it could be assumed that in these systems there is, in general, a big discolouration problem related to inorganic particles. With the available data and with SOMs limitations though, it is not possible to investigate if there were increased corrosion levels in the distribution pipes after the disinfection switch. However, this SOM indicates that, in these systems, switching disinfection increased the organic cell numbers and, consequently, increased the turbidity levels. Thus, it suggests that the increased microbial growth noticed after the disinfection change, in these DWDS, had a potential contribution to increased discolouration levels after the switch. This correlation in combination with the aforementioned high WTWs TCCs correlation with high consumers' tap TCCs, as presented in the previous SOM (Figure 6.3), indicate the importance of producing water of high quality in the WTWs to manage discolouration risk.

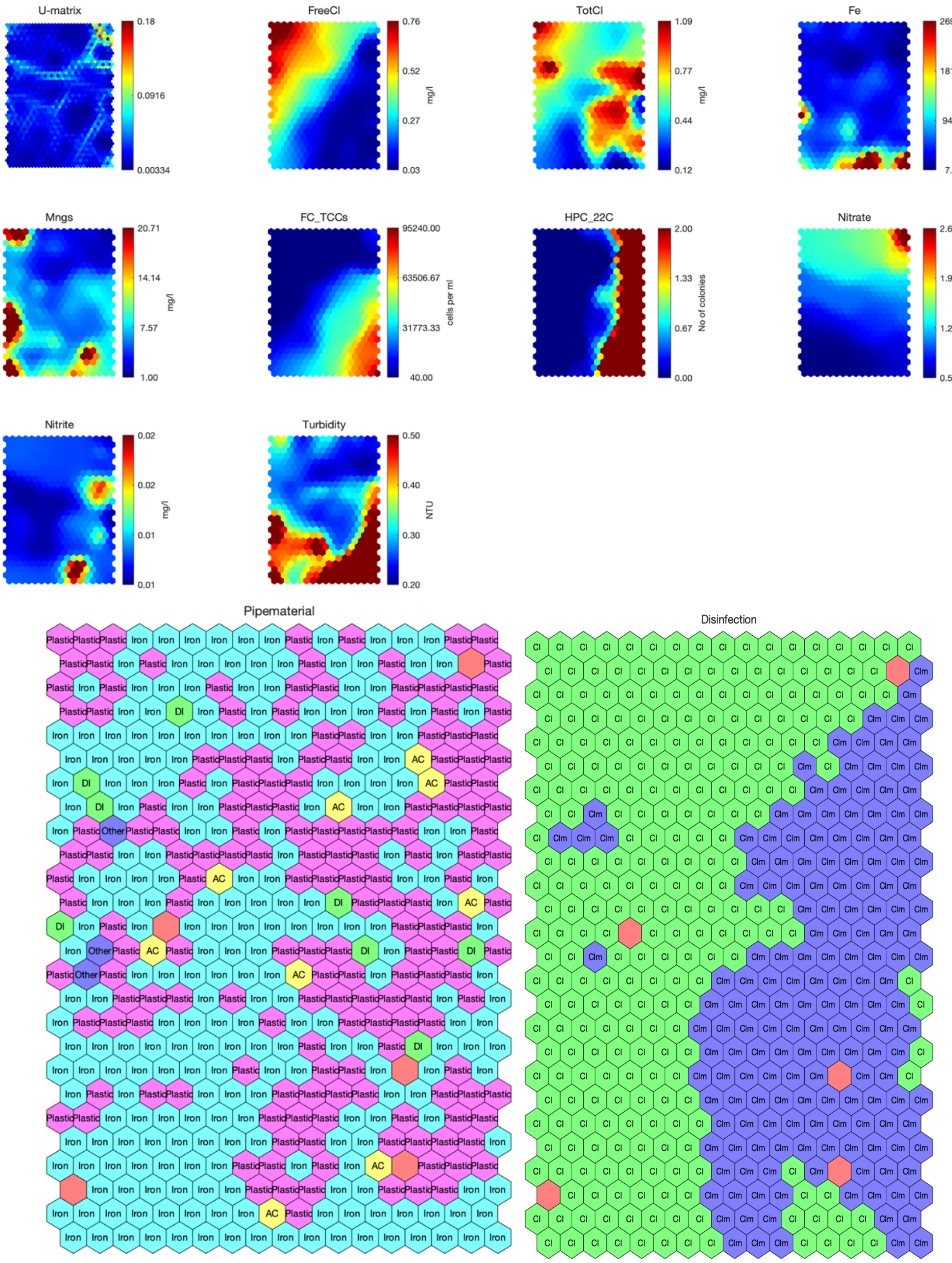
The correlation between high manganese, high iron, high TCCs and high turbidity as shown in this SOM, is a finding that agrees with various research works that study discolouration (I. J. H. G. Vreeburg and Boxall 2007; J. H. G. Vreeburg, Schaap, and Van Dijk 2004; S. Husband et al. 2016; Speight, Mounce, and Boxall 2019). The additional correlation between the clusters containing the highest levels of these parameters and low total chlorine clusters, as shown in the bottom part of their planes, indicates the importance of maintaining a stable disinfection residual to control biological regrowth and manage the DWDS' biofilm.

d) Are there any indications of nitrification after changing disinfection residual?

Nitrification is the phenomenon of oxidation of nitrogen compounds (mainly ammonia) to nitrate and then nitrite by nitrifying bacteria (USEPA 2002b) . The factors that contribute to the increase of these bacteria and, thus, to nitrification are high free ammonia concentration, high temperature, high water age and insufficient chlorine residual (Telfer A. 2014). Nitrification leads to complete loss of the chlorine residual, bacteriological growth (HPC growth in particular), iron release, and decrease of pH (American Water Works Association (AWWA) 2013).

Findings presented in the previous SOM related bacteriological activity with high iron concentrations. In addition, nitrification mostly occurs in chloraminated systems where ammonia is introduced to chlorine to form chloramines. Therefore, a new SOM (figure 6.5) was produced to discover the relationships between the various parameters related to nitrification to examine the scale of the phenomenon after switching disinfection residual. The selected parameters in this SOM were free and total chlorine, iron, manganese, turbidity, flow cytometry TCCs, HPCs at 22°C, nitrate, and nitrite. Two post-analysis labelled SOMs were also produced: one that labels the pipe material and another that labels the type of disinfection.

In the produced SOM, a strong correlation between high nitrite, high TCC clusters, high HPCs, high iron, high manganese, low total chlorine, and high turbidity is shown (bottom centre to right of the planes) and, according to the labelled output, these clusters align with cast iron pipes (cyan clusters in the pipe material labelled SOM) and are associated with the post switch chloramination disinfection. Another high nitrite cluster (middle right of the plane) is correlated with low total chlorine and high TCC and HPC clusters and medium to high turbidity. These clusters are again associated with post switch chloramination, but, in this SOM, apart from iron pipes, they are aligned with plastic pipes as well (pink cells in the pipe material labelled SOM). These findings show that there are signs of nitrification in these systems after the switch as agree with some of the nitrification parameters. In particular, the bottom right to centre cluster indicates the link between nitrification, iron, increased bacteriological activity and iron mains. The other high nitrite cluster, however, indicates that nitrification could be caused due to low or insufficient total chlorine residual or due to the increased bacteriological activity related to low concentrations of the chloramination disinfectant (low total chlorine).



Cyan cells: Iron pipes

Pink cells: Plastic pipes

Yellow cells: Asbestos pipes

Green cells: Ductile Iron pipes

Blue cells: after the switch

Green cells: Before the switch

Figure 6.5: Nitrification after changing disinfection, including pipe material and secondary disinfection labelled maps

Table 6-4: Summary of variables for SOM presented in figure 6.5

Variable (units)	Variable short form for SOM	Data Source	Number of samples	Average value	Standard deviation
Free Chlorine (mg/l)	FreeCl	Water quality	7211	0.30	0.26
Total Chlorine (mg/l)	TotCl	Water quality	7206	0.60	0.33
Iron (mg/l)	Fe	Water quality	3668	82.22	550.26
Manganese (mg/l)	Mngs	Water quality	3588	8.83	67.47
Flow cytometry Total Cell Counts (cells per ml)	FC_TCCs	Water quality	1552	17432	43328
HPCs @22C (No of colonies)	HPC_22	Water quality	2859	2.17	19.74
Nitrate(mg/l)	Nitrate	Water quality	1293	1.03	0.57
Nitrite(mg/l)	Nitrite	Water quality	1338	0.01	0.02
Turbidity (NTU)	Turbidity	Water quality	2945	0.48	3.51

6.5. Operational value of the findings

In this chapter, SOM was used to understand the general change of the drinking water behaviour after switching disinfection. SOMs analysis indicated that there is an increased bacteriological activity after switching disinfection related to high bacteriological activity exiting the works. This finding in combination with the increased TOC concentrations found in the customers taps after the switch indicate the importance of managing and controlling the WTWs to adapt to the changes in the water chemistry after the disinfection switch. Potential improvement of the WTWs performance may also reduce the increased turbidity levels noticed in the DWDS after the switch. Finally, this research emphasises the importance of controlling the free ammonia concentration in the chloraminated systems to reduce the risk of nitrification in the DWDS, a phenomenon that once in its full length, could lead to the complete degradation of the monochloramine residual and to the increased iron levels in the DWDS.

6.6. Conclusions

SOMs were applied on a tap monitoring water quality samples dataset that included all the DWDS that have switched their disinfection type from chlorination to chloramination for understanding the impact of that switch on drinking water quality. The disinfection type, used as qualitative parameter in the SOMs analysis, was also used as temporal indicator to cluster the various quantitative SOM map cells into the pre-switch (chlorination group) and post-switch group (chloramination group). Therefore, the cells in the SOM output map of each water quality parameter that are correlated with the chlorination group, belong in the pre-switch period and the opposite for those that correlate with the chloramination group. The key findings of this work are:

- Switching to chloramination reduces the concentrations of the DPBs in the drinking water

- The bacteriological activity, both in terms of high HPCs and high TCCs, is increased after the disinfection switch
- The increased TOC concentrations in both the customers taps and in the WTWs after switching to chloramination, explain the increased bacteriological activity after the disinfection switch.
- Turbidity levels are increased after the switch which indicates a change in the water chemistry related both to increased metals' concentrations but also to increased bacteriological levels.
- There are indications of nitrification in the systems after the transition to chloramination.
- Water utilities should concentrate in improving the performance of their WTWs, in terms of TOC and metals reduction, of the DWDS that are planning to switch disinfection

7. A comparison between ensemble decision tree models for the classification of service reservoirs using drinking water quality data

7.1. Introduction

In the previous 2 chapters a new direction for understanding relationships between various parameters using an unsupervised machine learning method on monitoring water quality samples was presented. This approach, as it was demonstrated, could be very useful to water utilities (WUs) for understanding the factors of water quality deterioration and the impact of interventions in the DWDS. However, WUs require prediction of future deterioration events to increase the proactive management of the DWDS. Therefore, further investigation over the ability of supervised methods in predicting future water quality deterioration events using discrete monitoring water quality datasets. Successful prediction of such events could be beneficial for WUs as they could direct their actions into the high-risk areas, improve the general water quality in their networks, guarantee a better water quality for their consumers and thus improve their reputation.

Ensemble decision trees, as explained in the literature review chapter, have the advantage that they do not operate like “black boxes” as the ANNs do. The split approach followed by each one of decision trees and the factors that contributed the most to the final decision for each tree that form the ensemble, could be presented to the model user once the prediction is made. Therefore, data-driven models that apply ensemble decision trees methodologies could be used not only as prediction tools but also as supportive tools for decision-making regarding pro-active interventions in the DWDS. To the author’s knowledge, Mounce et al. (S. R. Mounce et al. 2017) were the first to apply a data-driven model based on an ensemble decision tree methodology on tap discrete monitoring water quality data for the prediction of iron failures at a DMA level of a water company in the UK and for the iron risk classification of the company’s DMAs.

In this chapter, a methodology, that uses ML ensemble decision trees methods, is developed for the prediction of future failures in SW’s SRs. More specifically, this model predicts which SRs will fail in terms of having at least one sample with low Cl concentration (<0.3mg/l) the upcoming month or in terms of having at least one sample with coliform bacteria in the upcoming two months. Therefore, as this methodology identifies which SRs will fail and which will not, the binary classification approach is used. If the model predicts that a certain SR will fail in the upcoming month, then this SR will be grouped in the class named as “High risk” class, otherwise it will be grouped in the class named “Low Risk” class.

The ensemble decision trees that the methodology uses are random forest and boosting trees (AdaBoost and RusBoost) and a comparison of their outputs for both the prediction of low chlorine events and coliform bacteria events in SRs is made. This methodology aims to be a predictive tool that could be used by WUs for the identification of the SRs that are more likely to fail in the upcoming month. In addition, as it uses “white-box” ML techniques, it can also provide the water quality parameters that influenced the model’s outputs, an information that could be used for improving the models’ prediction performance further. Therefore, this chapter seeks to address the objectives 3 and 4 of this thesis.

7.2. Data processing and final input dataset production

The SRs’ monitoring water quality dataset for the years between 2012 and 2019 created as described in chapter 4 was used for this investigation. As it was mentioned many times before in this thesis, the monitoring water quality data are sparse both geographically and temporally and therefore the SR investigation was preferred to the tap investigation for two reasons: a) in the SRs each one of the samples was taken from the same point (SR outlet) and b) the regulations require to take 4 samples per month for every active SR to measure chlorine concentration (total and free), coliform bacteria and HPCs. However, by selecting the SR investigation, important parameters such as turbidity and total organic carbon (TOC) could not be included in the investigation as these parameters are not frequently measured in the SRs. For this work, we concentrated on low chlorine events and coliform bacteria events in SRs where at least 3 samples per month for every month of the investigation period were taken. In addition, the low chlorine events investigation, obviously, was made only to the SRs that belonged to the DWDS where chlorine was used for disinfection.

As regards the temporal scale of the analysis, this was selected based on the frequency of low free chlorine and coliform bacteria events. A low free chlorine event was defined as the sample where the chlorine concentration was measured below 0.3 mg/l as the minimum free chlorine concentration on customers taps is 0.2 mg/l (WHO 2000) and part of the chlorine is consumed during the water travel through the pipe network before reaching the taps. A coliform bacteria event was defined as the sample that at least one coliform bacterium is counted as set by the regulations (DWQR 2019b). The counted events were also included in the water quality dataset. Low free chlorine events are rare events and consist of 5-10% of the monthly events. Therefore, based on this frequency the monthly temporal scale was selected for this investigation. On the other hand, coliform events are even more rare events and appear in less than 1% of the yearly samples. However, the coliform appearance is disproportional during the year with most of the events appearing during the summer period and very few appearing in the winter period. For example, in the year 2019, SW had in total 71 coliform events in their SRs, 48 of which appeared during the summer months (from June to September) and for the months of January, February and March there were only 5 coliform events in total. It is, therefore, impossible to predict a coliform event during the winter months using data-driven techniques as there are not enough past events to be used for training the models. In addition, even during the summer period the maximum number of events per month is not bigger than 8. Thus, this work focused on the prediction of coliform events during the summer period only, using a monthly scaled dataset for the months of May, June, July and August for the prediction of the events for the following two months respectively (e.g May input - June, July outputs). Both the monthly (low chlorine events) and the seasonal model (coliform events) are further described in the following section.

Two final datasets were created using MATLAB® version 2019b. The first one was a monthly scaled dataset that included only the chlorine disinfection SRs for the period between 2012 and 2019 and was used as an input in the monthly model. The second one was a summer monthly scaled dataset that included both the chlorine and the chloramine SRs and was used as input for the seasonal model. In each one of the datasets the following parameters were included:

- Mean average values per SR per month for the following parameters:
    1.1.1.1. Free chlorine (Cl_AVE) - used only in the low chlorine events models
    1.1.1.2. Total chlorine (TotCl_AVE) - used only in the coliform events models only
    1.1.1.3. Heterotrophic Plate counts at 22 and 37 °C (HPC22 - HPC37)
    1.1.1.4. Flow cytometry intact and total cell counts (ICCs - TCCs)
    1.1.1.5. Water temperature (SR temperature)

- Mean average values from supplying WTWs per month for the following parameters:
    (i) Free chlorine (Cl_WTW) - used only in the low chlorine events models
    (ii) Total chlorine (TotCl_WTW) - used only in the coliform events models only
    (iii) Flow cytometry intact and total cell counts (ICCs_WTW - TCCs_WTW)
    (iv) Water temperature (TEMP_WTW)
    (v) pH (pH_WTW)
    (vi) TOC (TOC_WTW)

- Standard deviation per month per SR for the following parameters:
    (i) Free chlorine (Cl_std) - used only in the low chlorine events models
    (ii) Total chlorine (TotCl_std) - used only in the coliform events models only

- Age of water exiting the SRs as given by SW (Water age)

- Average daily precipitation per month per SR (SR precipitation) and per WTW (WTW precipitation)

- Three nominal (categorical) parameters:
    (i) SR Name (TWS)

(ii) Month of the year (Month) - used only in the low chlorine events models

(iii) Disinfection type - used only in the coliform events models only

In the final datasets there were no temperature, ICC and TCC data, both in the SRs and in the WTWs, for the years 2012-2014, as SW started measuring these parameters in the year 2015. The model inputs and outputs for both the low chlorine event prediction and the coliform event prediction are shown in figure 7.1

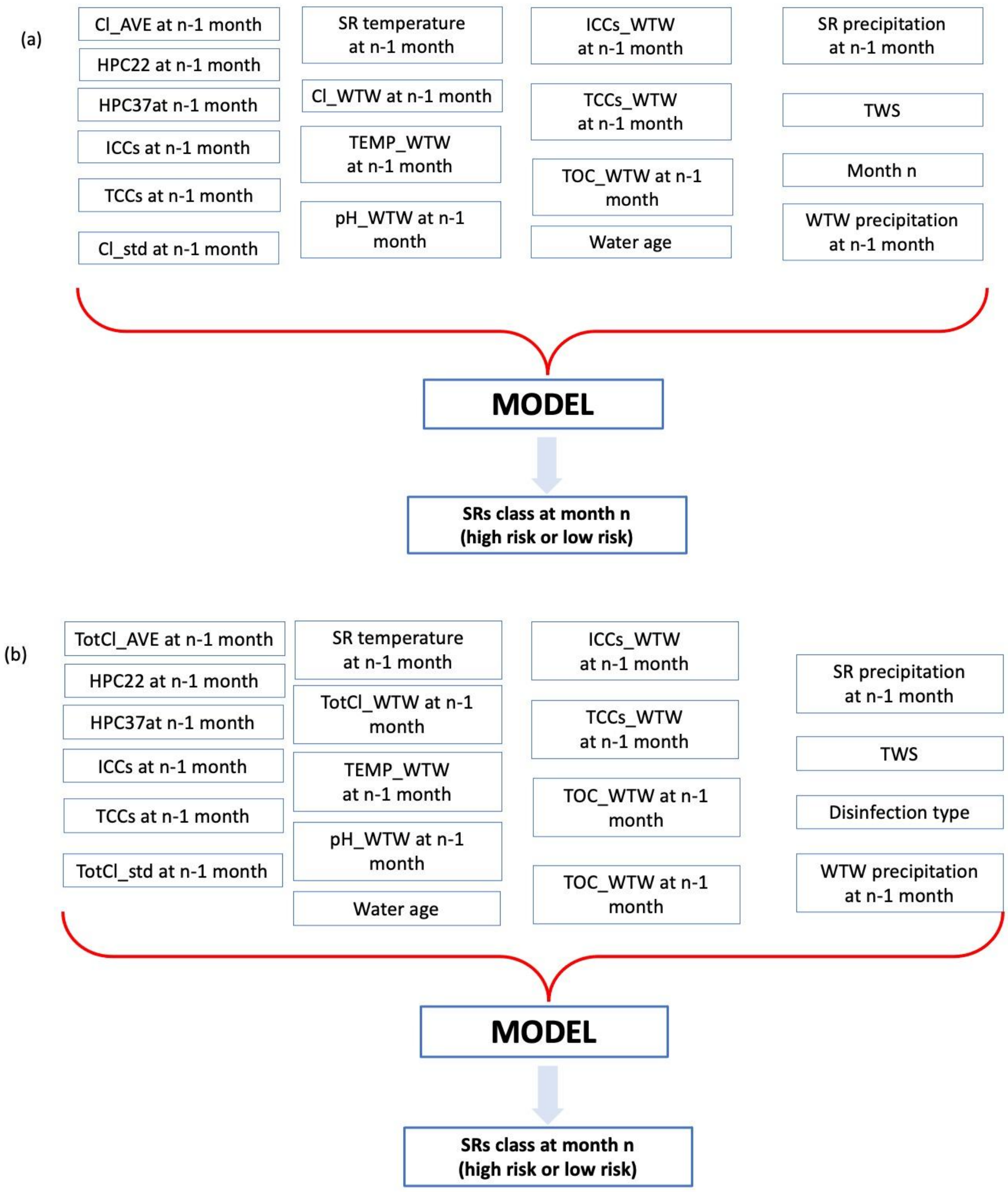


Figure 7.1: a. Simplified diagram of inputs and outputs of the low chlorine events model. b. Simplified diagram of inputs and outputs of the coliform bacteria events model

7.3. Machine learning application steps

The machine learning application steps for this investigation is as follows:

a. Define the water quality problem

In this chapter, the aim is to accurately classify SRs into High or Low Risk class by predicting future low chlorine and coliform events. Therefore, the water quality problem should be defined as follows:

Is it possible to correctly classify SRs into High and Low risk classes by accurately predicting future low chlorine and coliform events in them?

b. Type of the available data

The discrete water quality monitoring samples taken from the SRs outlets and the WTWs' outlets are the available data for the investigation. In addition, the precipitation data, taken from the gauging stations located close to the WTWs and the SRs are also included.

c. Define required output

The required output is the classification of the under-investigation SRs into two different classes (event or non-event class / high and low risk class)

d. Machine learning selection

By following the machine learning tree presented in chapter 3 the most appropriate methods for this investigation are random forest and boosting trees. The final machine learning methods for this investigation were random forest, AdaBoost and RusBoost. The former two ML techniques belong to the broader family of boosting trees.

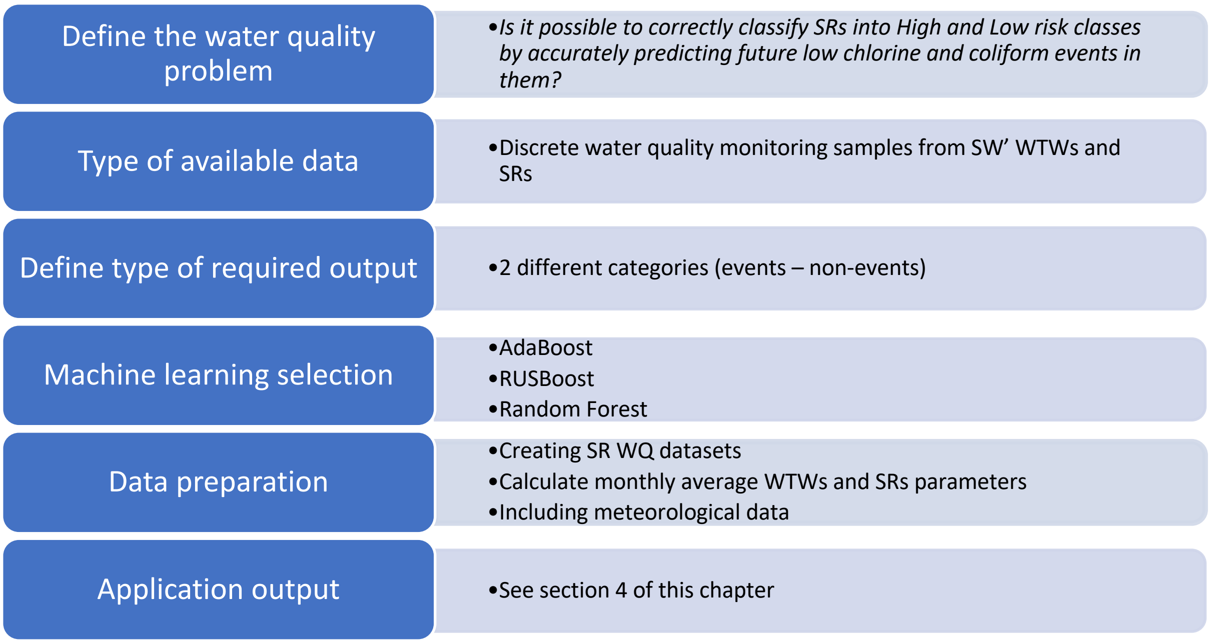
e. Data preparation

The SRs monitoring water quality dataset for the years 2012 to 2019 created as described in chapter 4 was used for this investigation. In addition, to create a steady temporal scale dataset - one monthly and one yearly scaled dataset - the average monthly and yearly values of each parameter of the dataset were calculated, including the WTWs parameters and the precipitation in both the SRs and the WTWs.

f. Application output

The required output is a vector that indicates in which class each one of the under-investigation service reservoirs belongs. The application outputs are presented in the rest of this chapter.

The ML application steps for this investigation are summarized in the following figure.



Define the water quality problem

- *Is it possible to correctly classify SRs into High and Low risk classes by accurately predicting future low chlorine and coliform events in them?*

Type of available data

- Discrete water quality monitoring samples from SW' WTWs and SRs

Define type of required output

- 2 different categories (events – non-events)

Machine learning selection

- AdaBoost
- RUSBoost
- Random Forest

Data preparation

- Creating SR WQ datasets
- Calculate monthly average WTWs and SRs parameters
- Including meteorological data

Application output

- See section 4 of this chapter

Figure 7.2: Machine learning application steps for the prediction of future deterioration events in SRs

7.4. SR classification models methodology

7.4.1. Ensemble algorithms and imbalanced datasets

The main aim of the predictive models was to classify the SRs into either the no-event (Low Risk) class or the event class (High Risk), for the month or the period under consideration, based on their predictions regarding the appearance of low chlorine events (month under consideration) and coliform events (period - season - under consideration).

The initial concept was to investigate traditional ensemble machine learning techniques such as random forest and AdaBoost (the main boosting algorithm) as the main algorithms for the predictive models. These two models were used successfully in various scientific projects including research projects in the water domain (Rojek 2014; Meyers, Kapelan, and Keedwell 2017; Xenochristou et al. 2018) and their algorithms are explained in chapter 3. However, their main weakness is that in heavily unbalanced datasets, like the under-investigation SR dataset, they tend to overclassify towards the majority, in this case low risk, class. Therefore, in this work some other methods that aim to tackle the imbalance issue are also investigated.

The most common type of methods applied to solve the imbalanced datasets problem are the sampling methods (He and Ma 2013). The general idea of these methods is to change the balance of the dataset with a specific mechanism and create a new dataset with more balanced classes. Oversampling of the minority class, undersampling of the majority class and the combination of both are some of the mechanisms used in sampling methods. In this chapter, 3 different techniques were investigated. More specifically, 2 oversampling techniques, SMOTE (synthetic minority oversampling technique) and ADASYN (adaptive synthetic sampling approach) that balanced the final input dataset by creating synthetic data for the minority class (event/high risk class). In addition, an ensemble decision tree method, RUSBoost (random under sampling boosting), that combines the undersampling approach with the boosting algorithm was also investigated.

SMOTE (Chawla et al. 2002) creates synthetic (artificial) data for the minority class following these steps:

1. For each sample, find randomly k nearest minority class neighbours (where k an integer number)
2. Select randomly one of the k nearest neighbours
3. For all the numerical features (parameters)
    a. calculate the Euclidean distance between the sample vector and the selected neighbour vector
    b. Multiply the result with a random number between 0 and 1
    c. Add the result to the sample vector and create a new synthetic vector
4. For all the nominal features (parameters)
    a. Find the nominal feature with the maximum number of appearances over the k minority class neighbours and the sample and add it to the new synthetic sample. If these are more than 1, select randomly

ADASYN (He et al. 2008) creates synthetic data for the minority class by using a weight for each minority class sample that corresponds to its difficulty in learning. Thus, the minority samples that are surrounded by samples of the majority class contribute more in the synthetic data than the minority samples that are surrounded by similar samples. The steps that ADASYN follows are:

1. Find k neighbours (where k an integer number) of each minority sample based on the Euclidean distance and calculate the ratio of the sample by dividing the number of these neighbours that belong to the majority class and k.
2. Normalise the ratio for each sample by dividing the ratio of each sample with the sum of the ratios of all the samples
3. Calculate the number of synthetic samples that should be generated for each sample by multiplying its normalised ratio with the total number of synthetic samples that we want to generate
4. For each minority sample, create the defined in step 3 number of synthetic data by following the SMOTE algorithm.

RUSBoost (Seiffert et al. 2008) is a machine learning technique that combines the random undersampling method for balancing the datasets with the boosting ensemble decision tree algorithm. The random undersampling method simply removes samples that belong in the majority class randomly, so that the final dataset that is used for training is balanced or more balanced than the initial one. Once the final dataset is created each weak learner (each decision tree of the ensemble) is trained based on the new more balanced dataset.

7.4.2. Monthly (low chlorine events) and seasonal (coliform bacteria events) predictive models

The problem, as described above, was a binary classification problem where SR should have been classified in either the Low-Risk no-event class or the High-Risk event class. Therefore, the seasonal and the monthly models were required to predict in which of the two classes each SR will belong in the following two months or the next month respectively.

7.4.2.1. Monthly low-chlorine events predictive models

For the monthly predictive models all the data up to and including the month prior to the present month were included, requiring a target output which in this case was the prediction of low chlorine events in SRs for the present month. This practically means that for predicting the chlorine events in August 2019, the input data from January 2012 with the SR classification for February 2012 up to input data for June 2019 with SR classification for July 2019 were given for training the model and the input data for July 2019 were given to the model to predict the class that each SR belongs in August 2019. Figure 7.3a shows a simple schematic of the monthly model.

7.4.2.2. Seasonal coliform events predictive models

For the seasonal predictive models all the data up to and including the month prior to the 2 investigation months were included, requiring a target output which in this case was the prediction of coliform events in SRs for these two months. So, in the seasonal model for predicting the coliform events in July and August 2019, the input data from May 2012 with the SR classification for Jun - July 2012 up to the input data for May 2019 with SR classification for June 2019 were given for training the model and the input data for June 2019 were given to the model to predict the class that each SR belongs in June-July 2019. Figure 7.3b shows a simple schematic of the seasonal model.

7.4.2.3. Models' parameters and implementation

For the training of each model there was the option of including or excluding a sampling method (SMOTE, ADASYN or none of them) and selecting one of the 3 machine learning algorithms (random forest, AdaBoost and RUSBoost). Therefore, there were 9 possible options for the training of the monthly and the seasonal model. In addition, for each model there is also the option of selecting all the available water parameters, or part of them. For the implementation, an algorithm was created in MATLAB® version 2019b (saved in the Github repository) where the user was able to select the type of model (season or month), the sampling method (SMOTE or ADASYN or none), the machine learning algorithm (random forest, AdaBoost and RUSBoost) and the water parameters to be included in the investigation.

There were a number of parameters that were required to be defined regarding the machine learning algorithm and the sampling methods. For the machine learning algorithms, the number of weak learners was defined as 1000 trees, the minimum leaf size was defined equal to 1, the learning rate for the boosting algorithms was set as 0.1 and the number of randomly selected variables set for each tree split in random forest was set to 3. As regards the sampling methods, the k number of neighbours was set equal to 5 and the number of synthetic data to be created was defined by the user for each different model run.

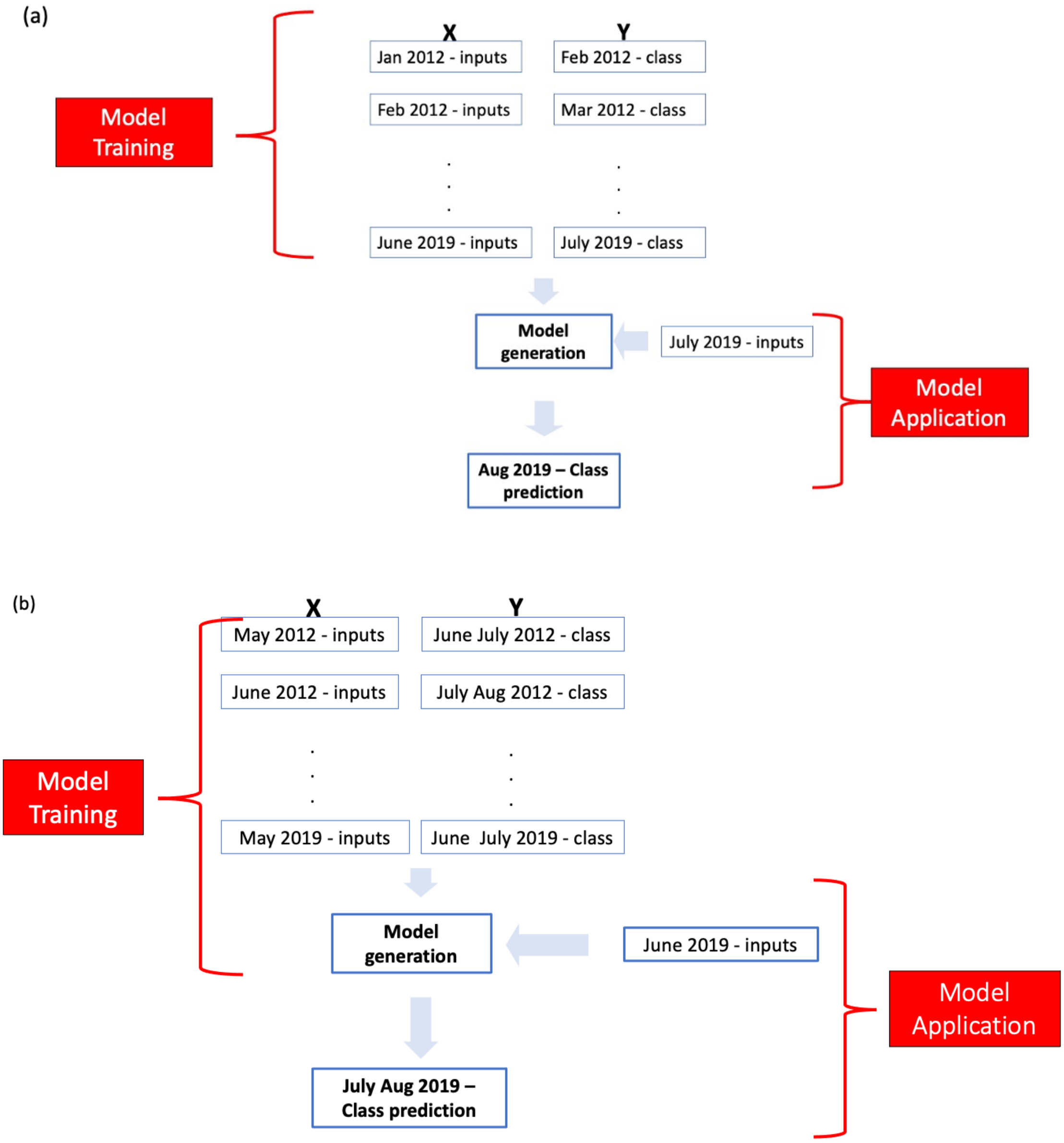


Figure 7.3: a. Monthly predictive model scheme for SR class prediction for August 2019. b. Seasonal predictive model scheme for SR class prediction for the period July-August 2019Performance metrics

Three performance metrics were used in this work for evaluating the models' results in comparison with the real - test data, the true positive rate (TPR) the true negative rate (TNR) and the Matthews correlation coefficient (MCC). The formulas and the meaning of each one of these metrics is explained in chapter 3.

7.4.3. Combining the ensemble models

A further combination of the best predictive models' outputs was made to investigate if the combined model could, potentially, enhance the final performance. In this study, 3 different combined ensembles models were created based on the way that the models contributed to the final decision. More specifically:

a. Simple Average Combined Ensembles Model (SACEM)

In SACEM for each SR each one of the top ensemble models contributed equally in the classification decision with a single vote in the final decision. In case of a tie in the votes, SACEM classified the SR in the High - risk (event) class.

b. TPR Weighted Average Combined Ensembles Model (TPR WACEM)

The TPR WACEM model made its classification decision for each SR by assigning different weights on each one of the best ensemble models based on their TPR performance. The weights on each model were assigned as follows:

$$TPR\_w_i = \frac{TPR_i}{\sum_{i=1}^{N} \quad TPR_i}$$

where TPR_w$_i$ is the assigned weight at the ith model and TPR$_i$ its TPR performance.

c. MCC Weighted Average Combined Ensembles Model (MCC WACEM)

The MCC WACEM model made its classification decision for each SR by assigning different weights on each one of the best ensemble models based on their MCC index. The weights on each model were assigned following the same equation as in the TPR WACEM model but instead of using the TPR performance of each model, the MCC index was used.

7.5. Results

7.5.1. Monthly low chlorine predictive models

7.5.1.1. Checking past events method

For this investigation the month selected as the test month was August 2019. Before training the machine learning models, a simple check in the original dataset for the SRs that had repeated low chlorine events during the month of August in the previous three years (2016-2018) was made. In this approach, it was assumed that an SR that had 2 or more low chlorine events in the past 3 years should be classified as a high-risk SR and the opposite for those with 1 or zero events during the same period. This is an approach that WUs commonly follow for prioritising their interventions in their SRs and was also followed in this chapter to compare its outputs with the ones of the data-driven methodology. The outputs indicated that the checking past events approach had a TPR performance of 0.61 (with correctly predicting 53 SRs in the high-risk class), a TNR performance of 0.75 (243 SRs were correctly classified in the low-risk class) and an MCC index of 0.31.

7.5.1.2. Summary of monthly ensemble models results

Overall, 7 algorithms were used in this work for the monthly model investigation as follows:

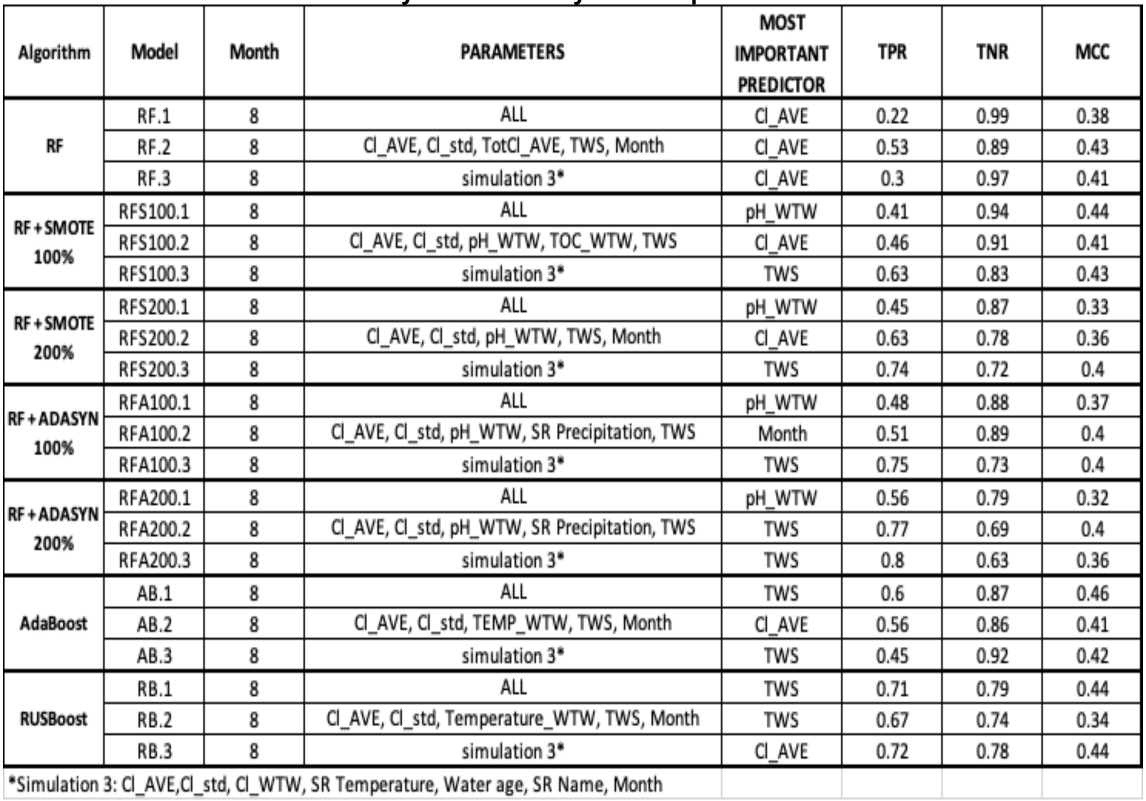
a. Random forest (RF)
b. Random forest with SMOTE with synthetic sampling rate equal to 100% of the minority class so that the minority class equals to around 35% of the final training dataset (RFS100)
c. Random forest with SMOTE with synthetic sampling rate equal to 200% of the minority class so that the final dataset is balanced (RFS200)
d. Random forest with ADASYN with synthetic sampling rate equal to 100% of the minority class (RFA100)
e. Random forest with ADASYN with synthetic sampling rate equal to 200% of the minority class (RFA100)
f. AdaBoost (AB)
g. RUSBoost (RB)

For each one of these algorithms 3 model simulations were made using different groups of parameters. More specifically, for all the algorithms:

- **In simulation 1,** all the numerical and categorical parameters were included
- **In simulation 2,** only the 5 parameters that contributed the most in the final decision of the first simulation, as defined from the post-training graphs (see appendix A), were included
- **In simulation 3,** the three free chlorine parameters (free chlorine average, free chlorine standard deviation, WTWs' free chlorine average), the categorical parameters (SR name and month of the year) plus the water age and the temperature of water in the SRs are included.

SR water age and temperature are included in the third simulation because these were the two main parameters correlated with low free chlorine concentrations in the SRs, as the SOMs investigation demonstrated in chapter 5. The total number of models produced were 21 (7 ML techniques with 3 different sets of parameters per technique). Each model was named based on the algorithm that followed and a number between 1-3 that represented the specific simulation as described above. So, for example, the random forest plus SMOTE 100% model that included all the parameters in the training process was named RFS100.1. The models were trained to predict the low chlorine events for August 2019. The results were compared with the real data for the month of August 2019, and the performance metrics of these models are presented in Table 7-1.

Table 7-1: Summary of the monthly models' performance metrics



Algorithm	Model	Month	PARAMETERS	MOST IMPORTANT PREDICTOR	TPR	TNR	MCC
RF	RF.1	8	ALL	Cl_AVE	0.22	0.99	0.38
	RF.2	8	Cl_AVE, Cl_std, TotCl_AVE, TWS, Month	Cl_AVE	0.53	0.89	0.43
	RF.3	8	simulation 3*	Cl_AVE	0.3	0.97	0.41
RF + SMOTE 100%	RFS100.1	8	ALL	pH_WTW	0.41	0.94	0.44
	RFS100.2	8	Cl_AVE, Cl_std, pH_WTW, TOC_WTW, TWS	Cl_AVE	0.46	0.91	0.41
	RFS100.3	8	simulation 3*	TWS	0.63	0.83	0.43
RF + SMOTE 200%	RFS200.1	8	ALL	pH_WTW	0.45	0.87	0.33
	RFS200.2	8	Cl_AVE, Cl_std, pH_WTW, TWS, Month	Cl_AVE	0.63	0.78	0.36
	RFS200.3	8	simulation 3*	TWS	0.74	0.72	0.4
RF + ADASYN 100%	RFA100.1	8	ALL	pH_WTW	0.48	0.88	0.37
	RFA100.2	8	Cl_AVE, Cl_std, pH_WTW, SR Precipitation, TWS	Month	0.51	0.89	0.4
	RFA100.3	8	simulation 3*	TWS	0.75	0.73	0.4
RF + ADASYN 200%	RFA200.1	8	ALL	pH_WTW	0.56	0.79	0.32
	RFA200.2	8	Cl_AVE, Cl_std, pH_WTW, SR Precipitation, TWS	TWS	0.77	0.69	0.4
	RFA200.3	8	simulation 3*	TWS	0.8	0.63	0.36
AdaBoost	AB.1	8	ALL	TWS	0.6	0.87	0.46
	AB.2	8	Cl_AVE, Cl_std, TEMP_WTW, TWS, Month	Cl_AVE	0.56	0.86	0.41
	AB.3	8	simulation 3*	TWS	0.45	0.92	0.42
RUSBoost	RB.1	8	ALL	TWS	0.71	0.79	0.44
	RB.2	8	Cl_AVE, Cl_std, Temperature_WTW, TWS, Month	TWS	0.67	0.74	0.34
	RB.3	8	simulation 3*	Cl_AVE	0.72	0.78	0.44

*Simulation 3: Cl_AVE,Cl_std, Cl_WTW, SR Temperature, Water age, SR Name, Month

The MCC of these models has a range between 0.33 and 0.46 which is an indication that all the models perform relatively well. However, the TPR variation indicates that some of the algorithms and models outperform others. More specifically, the random forest models (RF.1-RF.3) and the AdaBoost models 2 and 3 (AD.2-AD.3) could not predict sufficient low chlorine events. The use of ADASYN and SMOTE sampling methods improved the positive predictive performance of the random forest algorithm as the TPR metric indicates but with the cost of increased false positive events as the decrease of TNR shows. The RUSBoost models, which also use a sampling method in their algorithms, correctly predicted more positive events compared to the simple boosting model (AdaBoost) but again with the cost of increased false positive events. The MCC metric indicates that overall, the most balanced model is AB.1 as it has the higher MCC number (0.46). However, the TPR of this model is lower compared to other models which, in our case, is a disadvantage as the main goal is to correctly predict as many high-risk SRs as possible. By contrast, RFA200.3, the model that correctly classified the most SRs in the high class (TPR 0.8), had the lowest TNR performance (TNR=0.63) which indicates that this model had also increased false positive predictions which also affects its MCC performance. Thus, a further comparison over the actual number of SRs that were correctly classified in either the low-risk or the high-risk class was also made using the models that had an MCC value of 0.4 or higher but also had a TPR value of 0.6 or higher. In this comparison, the number of the correctly classified SRs using the repeated events approach over the last 3 years, as described at the beginning of the chapter, was also included. The results are presented in table 7-2.

Table 7-2: Comparison of the numbers of the correctly classified SRs by the most accurate ensemble decision tree monthly models and the checking the past years events method

	Checking past years events method	RFS200.3	RFA100.3	RFA200.2	AB.1	RB.1	RB.3	Cl events for August 2019
Correctly predicted high-risk SRs	53	64	65	67	52	62	63	**87**
Correctly predicted low-risk SRs	243	236	237	226	285	257	255	**326**
MCC	0.31	0.4	0.4	0.4	0.46	0.44	0.44	

As the table indicates, without the use of a machine-learning algorithm, only 53 out of 87 SRs would have been predicted correctly as high-risk SRs. The AB.1 model predicted correctly one less SR in the high-risk class compared to repeated events approach, but also generated 41 less false positives, a result that demonstrates the superiority of the data-driven approach over the simple check of past years results approach. The RF models using SMOTE and ADASYN (RFS200.3, RFA100.3, RFA200.2) were the models that predicted most of the low-chlorine events, however, with the cost of creating a lot of false positives, as around 100 SRs were incorrectly classified in the high-risk class. Overall, the RB models (RB.1 and RB.3) appear to be the most accurate models in classifying the high-risk SRs without over producing false positives as the numbers in table 7-2 indicate.

7.5.1.3. Monthly low chlorine combined ensemble models’ results

For the combined ensemble models the 4 models that produced the better results were selected. More specifically, the models selected for the investigation were RFA100.3, AB.1, RB.1, RB.3. The models’ outputs for August 2019 are presented in table 7-3. The results indicate that all three models (SACEM, TPR WACEM) produced good results and correctly predicted at least 60 SRs in the high-risk class. In addition, the MCC indexes of all the combined models were equal or higher to the single ensemble decision trees. SACEM and TPR WACEM models were able to identify more positive SRs than both the RB.1 and RB.3 models and one less compared to the RFA.100. The MCC WACEM model predicted fewer positive events compared to the other two models, however its higher TNR performance indicates that it created less false positive events. Finally, according to the MCC indexes, TPR WACEM is the best out of these three models (0.45 MCC index and 0.44 the other two models). In addition, TPR WACEM performs better to both RB.1 and RB.3 models.

Table 7-3: Comparison of the combined low chlorine events ensembles models outputs for August 2019

Combined Ensemble Model	No of models combined	Models combined	TPR	TNR	MCC	Correctly predicted high-risk SRs	Correctly predicted low-risk SRs
SACEM	4	RFA100.3,AB.1,RB.1,RB.3	0.74	0.77	0.44	64	251
TPR WACEM	4	RFA100.3,AB.1,RB.1,RB.3	0.74	0.78	0.45	64	255
MCC WACEM	4	RFA100.3,AB.1,RB.1,RB.3	0.7	0.8	0.44	60	262

7.5.2. Seasonal coliform events predictive models

7.5.2.1. Checking past years events

The months of July and August 2019 were selected as the test period for our investigation. As in the monthly models, a simple check in the data for past coliform events was also made prior to the models' training. This is again an approach that commonly WUs use for prioritising their interventions in the SRs prior to the summer period and thus avoiding bacteriological failures. In this case, as the coliform events were even more rare, an SR was classified in the high-risk SR class if there was at least 1 coliform bacterium appearance in its water quality samples in the months of July and August for the years 2016-2018. This approach had a TPR performance of 0.25 (with correctly predicting 5 SRs in the high-risk class), a TNR performance of 0.88 (571 SRs were correctly classified in the low-risk class) and an MCC index of 0.07.

7.5.2.2. Summary of seasonal ensemble models' results

The coliform bacteria dataset is even more unbalanced compared to the low-chlorine one and therefore, in the coliform bacteria investigation, only the algorithms that had better performance in classifying the SRs by predicting low chlorine events were selected. More specifically, the RF models and the RF models using SMOTE to create a less unbalanced dataset (RFS100 models) were excluded from this investigation. In addition, the AdaBoost algorithm was used only in combination with SMOTE and ADASYN. Overall, the algorithms that were selected in this investigation were as follows:

a. Random forest with SMOTE with synthetic sampling rate equal to 1400% of the minority class so that the minority class equals the majority class in the final dataset (RFPS1400)
b. Random forest with ADASYN with synthetic sampling rate equal to 1000% of the minority class so that the minority class equals to the 35% of the final dataset (RFPA1000)
c. Random forest with ADASYN with synthetic sampling rate equal to 1400% of the minority class (RFPA1400)

d. AdaBoost with SMOTE with synthetic sampling rate equal to 1400% of the minority class (ADPS1400)
e. AdaBoost with ADASYN with synthetic sampling rate equal to 1400% of the minority class (ADPA1400)
f. RUSBoost (RBP)

For each one of these techniques 3 model simulations were made using different groups of parameters. More specifically, for all the algorithms:

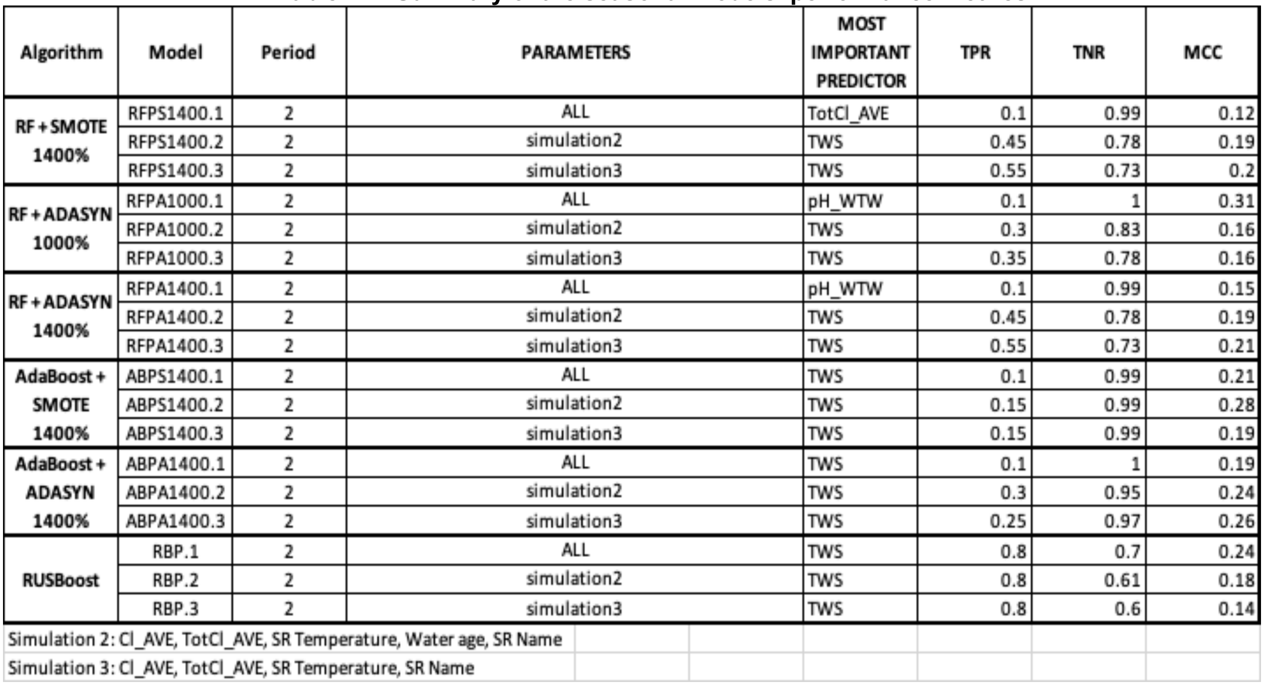
- **In simulation 1**, all the numerical and categorical parameters were included
- **In simulation 2**, the used parameters were the free and the total chlorine, the age of water, the temperature of water in the SRs as these are the parameters that are mostly related to bacteriological failures during the summer period (an argument that agrees with the SOMs analysis findings in chapter 5 as well), plus the SR name as a categorical parameter
- **In simulation 3**, the total chlorine, the free chlorine, and the temperature of water in the SRs plus the SR names as categorical parameters were used. The reason that the water age parameter was excluded in the third simulation was for investigating the performance of each algorithm without the use of a repeated numerical parameter (the age of water was a steady number for each SR).

The total number of the produced models was 18 (6 techniques with 3 simulations per technique). As above, for this investigation, each model was named based on the algorithm that followed and a number between 1-3 that represented the specific parameters used as described above. In addition, the letter P (for period) after the algorithm initials for each model was also included. So, for example, the random forest plus ADASYN 1000% model that included all the parameters in the training process was named RFPA1000.1. The models were simulated to classify the SRs based on their coliform events for July-August 2019. The outputs were compared with the real data for July and August 2019 and their performance metrics are presented in Table 7-4.

The MCC of these models has a range between 0.12 and 0.31 which is an indication that the seasonal coliform prediction models had, in general, a worse performance compared to the monthly low-chlorine prediction models. However, these findings were expected since only 20 out of the total 670 SRs under investigation, had a coliform event during July-August 2019. That also explains the fact that all the models, apart from the RB ones, had a low TPR performance.

The MCC values indicate that the best model was RFPA1000.1 (MCC=0.31) for all the reasons mentioned in the previous paragraph. This is because this model classified correctly all the low-risk SRs and two SRs in the high-risk class but mainly because it did not create any false positives. This model, though, had the lowest TPR performance (TPR=0.1) of all the models (together with the models RFPS1400.1, ABPS1400.1, RFPA1400.1, ABPS1000.1) which is a big disadvantage as its main goal is to predict the most coliform events possible. Nevertheless, when the RFPA1000.1 model is compared with the other aforementioned models with equal TPR, it is clear that this one performs better, as all the others, not only they did not correctly classify a sufficient number of high-risk SRs in the correct class, but they also produced a significant amount of false positive SRs.

Table 7-4: Summary of the seasonal models' performance metrics



Algorithm	Model	Period	PARAMETERS	MOST IMPORTANT PREDICTOR	TPR	TNR	MCC
RF + SMOTE 1400%	RFPS1400.1	2	ALL	TotCl_AVE	0.1	0.99	0.12
	RFPS1400.2	2	simulation2	TWS	0.45	0.78	0.19
	RFPS1400.3	2	simulation3	TWS	0.55	0.73	0.2
RF + ADASYN 1000%	RFPA1000.1	2	ALL	pH_WTW	0.1	1	0.31
	RFPA1000.2	2	simulation2	TWS	0.3	0.83	0.16
	RFPA1000.3	2	simulation3	TWS	0.35	0.78	0.16
RF + ADASYN 1400%	RFPA1400.1	2	ALL	pH_WTW	0.1	0.99	0.15
	RFPA1400.2	2	simulation2	TWS	0.45	0.78	0.19
	RFPA1400.3	2	simulation3	TWS	0.55	0.73	0.21
AdaBoost + SMOTE 1400%	ABPS1400.1	2	ALL	TWS	0.1	0.99	0.21
	ABPS1400.2	2	simulation2	TWS	0.15	0.99	0.28
	ABPS1400.3	2	simulation3	TWS	0.15	0.99	0.19
AdaBoost + ADASYN 1400%	ABPA1400.1	2	ALL	TWS	0.1	1	0.19
	ABPA1400.2	2	simulation2	TWS	0.3	0.95	0.24
	ABPA1400.3	2	simulation3	TWS	0.25	0.97	0.26
RUSBoost	RBP.1	2	ALL	TWS	0.8	0.7	0.24
	RBP.2	2	simulation2	TWS	0.8	0.61	0.18
	RBP.3	2	simulation3	TWS	0.8	0.6	0.14
Simulation 2: Cl_AVE, TotCl_AVE, SR Temperature, Water age, SR Name							
Simulation 3: Cl_AVE, TotCl_AVE, SR Temperature, SR Name							

For all the RF algorithms with 1400% extra synthetic minority data (RFPS1400 & RFPA1400), the 2nd and the 3rd model had an increased TPR performance and a decreased TNR performance compared to the 1st model. Their MCC index though is increased, a factor that indicates that the cost of generating more false positives (smaller TNR performance) did not affect their overall performance. The AdaBoost models (ABPS1400 & ABPA1400 models) had the lowest TPR performance when compared to the respective models of the other techniques. However, their TNR performance which is high (0.95-1) and their MCC index indicate that the Adaboost algorithm produced more balanced models and, more importantly, less false positive SRs. Finally, the RUSBoost models had by far the highest TPR performance (TPR=0.8 for all the RB models). However, the TNR performance and the decrease of the MCC value in the 2nd and 3rd model of the algorithm indicate that to achieve this positive rate, a high number of false positives SRs were also created.

A further comparison over the actual number of the correctly classified SRs of the most successful models (those SRs with MCC 0.2 or higher) and the repeated events is presented in table 7-5.

The repeated events approach had the worst performance as its MCC shows. This is another indication of the importance of the use of data-driven techniques over the simple checking of past events approach. The AdaBoost models (ABPS1400.3, ABPA1400.2 and ABPA1400.3) and RFPA1000.1 model correctly classified less SRs than the other models but have predicted correctly a very high number of low-risk SRs. By contrast, the other RF models (RFPS1400.3 and RFPA1400.3) and the RB model predicted more events but with the cost of a high number of false positives. Overall, the table 7-5 indicates that there is no clear "winning" model.

Table 7-5: Comparison of the numbers of the correctly classified SRs by the most accurate ensemble decision tree seasonal models and the checking the past years events method

	Checking past years events method	RFPS1400.3	RFPA1000.1	RFPA1400.3	ABPS1400.2	ABPA1400.2	ABPA1400.3	RBP.1	Coliform Events for July-August 2019
Events	5	11	2	11	3	6	5	16	**20**
Non Events	571	474	650	474	646	619	633	456	**650**
MCC	0.14	0.2	0.31	0.21	0.28	0.24	0.26	0.24	

7.5.2.3. Seasonal coliform events combined ensemble models results

For the combined models, the 5 models with the most predicted coliform events were selected (RFPS1400.3, RFPA1400.3, ABPA1400.2, ABPA1400.3 and RBP.1). The combined models' outputs for July-August 2019 are presented in table 7-6. As expected, none of the combined models could not predict more positive events than the RBP.1 model. However, both the SACEM and TPR WACEM models had a better performance compared to both the RFPS1400.3 and the RFPA1400.3, as the former predicted the same positive SRs and 38 more negative SRs and the latter predicted 2 more positive SRs and 26 more negative SRs. The MCC WACEM had similar predictions with the AB models which was an expected finding as the ABP models had the best MCC index and, therefore, they were the models that influenced it the most. Overall, by looking at both the performance metrics and the actual numbers of the correctly predicted SRs, it could be said that the combined models performed better than most of the single ensemble decision trees.

Table 7-6: Comparison of the seasonal combined ensembles models outputs for July & August 2019

Combined Ensemble Model	No of models combined	Models combined	TPR	TNR	MCC	Events	Non-Events
SACEM	5	RFPS1400.3, RFPA1400.3,ABPA1400.2,ABPA1400.3, RBP.1	0.55	0.78	0.23	11	508
TPR WACEM	5	RFPS1400.3, RFPA1400.3,ABPA1400.2,ABPA1400.3, RBP.1	0.65	0.75	0.24	13	490
MCC WACEM	5	RFPS1400.3, RFPA1400.3,ABPA1400.2,ABPA1400.3, RBP.1	0.3	0.96	0.24	6	623

7.6. Discussion

7.6.1. Results analysis

The RF classification models without the use of sampling methods were ineffective to correctly classify a high number of high-risk SRs in their correct class, based on their low-chlorine event prediction, as their low TPR indicates. However, this is not a surprising output due to the nature of the RF algorithm (Breiman 2001). RF makes its final classification decision based on a vote between its independent and individual decision trees, each of which was trained in a highly unbalanced dataset. Therefore, it was easier for the RF model to predict the low-risk SRs, as their TNR performance indicates, than identifying the high-risk SRs. The use of the sampling models had improved RF's TPR performance in both the monthly and the seasonal approach. In fact, the more balanced the final dataset was, the more high-risk SRs were predicted. The cost of generating more false positive results when the use of the sampling methods was included, was also expected as the test dataset (August 2019 for the low chlorine events and July and August 2019 for the coliform events) had a small number of positive events while the RFs were overfed with positive events during their training.

ADASYN has been proven to be a better sampling method for creating synthetic data to improve the monthly RF models' TPR performance than SMOTE in this case study. This finding indicates that the ADASYN approach of creating data from the most hidden minority class data and not from all the original dataset, generates more "difficult" minority examples for the RFs' training and, thus, improves their ability to predict more positive events. However, for the seasonal RF models, ADASYN and SMOTE produce almost similar synthetic data as the RFPS1400 and RFPA1400 models' performance demonstrate. This could be potentially explained by the fact that the coliform events constitute only the 4.4% of the seasonal dataset, thus many the synthetic data (1400% of the minority class dataset) was required to balance it and, consequently, for both methods, the minority data has contributed multiple times to accomplish that.

The boosting algorithm (AdaBoost) produced better results regarding the low Cl events prediction compared to the RF. This is because, in contrast to the RF algorithm, each decision tree in boosting is not independent and does not contribute the same in the final decision(Rokach 2010). Each new decision tree is influenced by the performance of the previous trees during the training and, thus, the AdaBoost algorithm could understand the training dataset. As the monthly models' MCCs indicate, AdaBoost are steady models that do not create a high number of false positives or false negatives. However, they were unable to predict enough positive events which is the main scope for WUs.

AdaBoost algorithm remains steady and unbiased towards the minority class even when it is trained with a balanced dataset generated by either SMOTE or ADASYN as the seasonal models' results indicate. The fact that the TPR performance is low may indicate that, when there is no clear difference between the minority and majority class data, the AdaBoost algorithm classifies the SRs into the safer class (low-risk class). This "safe" approach predicts a small number of high-risk SRs, but at least most of them are correctly classified. This is in contrast with the RF algorithm behaviour when the sampling methods are used to balance the dataset as described in the previous paragraphs.

The RUSBoost model was introduced as a model to tackle the problem with the unbalanced datasets by combining the simple removal of unnecessary data from the majority class and the boosting algorithm. The seasonal and monthly RB findings indicate that with this approach the boosting algorithm increases the number of positive classified events but with the cost of losing the stability of the AdaBoost and of classifying incorrectly many SRs in the high-risk class.

The combined models' performance is dependent on the performance of each individual ensemble decision tree. Thus, it is impossible for these models to outperform the individual ones in all the performance metrics. However, the SACEM and TPR WACEM combined models' approach, introduced in this chapter, indicated that the combination of more stable models - models with higher MCC with models with higher TPR - creates models that have sufficient positive rate and, at the same time, reduces the false positives that the high TPR models generate. As regards the MCC WACEM models, they may had predicted less positive SRs, but they also had high TNR performances, which indicates that this is the recommended combined model that someone could use when their main goal is to prioritise the interventions to a limited number of SRs.

The findings demonstrated that selecting the important water quality parameters for training the models, is important for the RF seasonal and monthly models. In general, RF required less parameters to increase their TPR performance. Selecting the top-5 parameters was a good approach but the findings also demonstrated that parameters such as the water age, the temperature the free and the total chlorine contributed the most into creating more accurate results. Therefore, it would have been very interesting to investigate the accuracy of the models if further information was available, regarding the seasonality change of the water age exiting the SRs and the water temperature data for the whole study period, and not only for the last 4 years.

It is worth mentioning that some of the seasonal and monthly boosting models (both AdaBoost and RUSBoost) performed better when all the available parameters were included in their training. This could, probably, be explained by the difference in the way that boosting and random forest make the final decision. As mentioned before, in boosting, the final decision is made using a weighted average and the final trees in the training sequence had learnt to use the information from each parameter proportionally to their contribution in the classification process. In random forest, though, each tree is completely independent and each split decision at each tree is made based on 3 parameters selected randomly each time. Therefore, in random forest the more the available parameters the merrier are the chances that a parameter that is not related to the deterioration event is selected for the splitting decision.

Both models, in this chapter, have been trained using the maximum available data for training and tested only in one month (two months for the coliform bacteria events model). This is because it was aimed to maximise the available data for the training period and get better model outputs. In practice, WUs will follow the same approach by adding the new available data at the end of the month in the model for training and predicting the SRs class in the following month. In the future where data for more years will be available, further work is required to investigate if the existing models' performance could be improved when, in addition to the previous month inputs, the data of the previous year are included.

7.6.2. Operational value of models' application

Checking past years events approach indicated the importance of recording and maintaining a monitoring water quality dataset. This study showed that, even without the use of any data-driven technique, many water quality deterioration events could have been prevented merely by checking repeated past events and the values of their related water quality parameters.

The monthly and the seasonal prediction could be used for different approaches by the WUs. The former could be used as an alarm of low chlorine concentrations and, subsequently, as a bacteriological risk indication in the water supply zones (WSZs). The later could be used for early prediction of coliform events and prioritisation of proactive interventions in the SRs.

The ensemble decision trees make their class decision based on a likelihood of risk per SR that is provided to the user as an output of scores for each SR and for each class. Therefore, this score could be used for ranking the SRs from the highest to the lowest risk and, based on these, WUs can concentrate their interventions into the ones that have higher chances to fail. In addition, each model could export as an image every decision tree of the ensemble. An example of one of the decision trees of the RB.3 model is presented in figure 7.4. This example shows the split criteria of the data in the first 2 leaves of the decision tree. The model user could check the splitting criteria of a certain number of the model's decision trees, and, therefore, understand the reasons that it made a certain decision.

There were some models that outperformed in this work and others that show that they are not appropriate for the available data in the SRs. However, this research could not recommend only one model to use for either the low chlorine events prediction or the coliform bacteria prediction. This is because the models' outputs indicated that there will always be a trade between increasing the TPR performance and decreasing the MCC index by generating a large number of false positives. The optimal trade between these two metrics criteria should be decided on a managerial level, based on the balance between proactive intervention in the predicted high-risk SRs and the available financial sources. The combined ensemble models presented here, demonstrated that could be the solution that could balance the TPR vs MCC. Therefore, it is recommended that the final decision on the proactive management strategy should be taken after checking the outputs of these models.

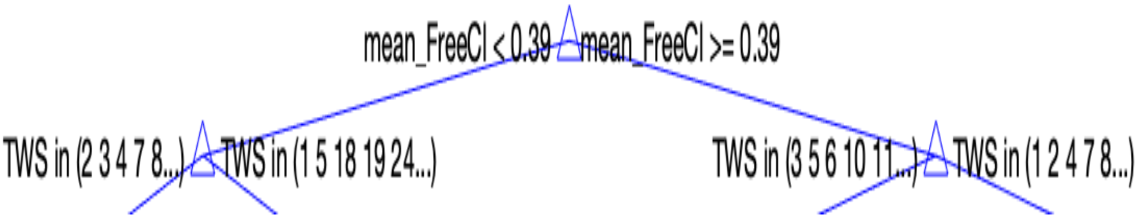


Figure 7.4: Part of one of the 1000 decision trees of the RB.3 model

Finally, it is worth pointing out that this is machine learning approach that uses only available water quality monitoring data in the SRs. This methodology can be applied for any other deterioration event prediction in the DWDS if there are sufficient discrete water quality monitoring sample data for a specific temporal scale and as long as, in the under-investigation data, a range of past deterioration events are included.

7.7. Conclusions

In this chapter, a methodology for classifying SW's SRs into two different classes (high-risk or low-risk class) was presented. This methodology used three different ensemble decision tree algorithms (random forest, AdaBoost and RUSboost) on the monitoring water quality samples dataset taken from the SR outlets. Two different types of models were created, the monthly models and the seasonal models. The monthly models used past water quality data on a monthly temporal scale for the classification of the SRs based on the model's prediction of at least one low chlorine event (samples with free chlorine concentration below the 0.3 mg/l threshold) in the upcoming month. The seasonal models used past water quality data on a monthly temporal scale for the summer months only, for the classification of the SRs based on the prediction of coliform events in the following two months during the summer period

(June and July, July and August, August and September). The SRs that the methodology predicts will have at least one event were grouped in the High-risk class and the others in the Low-Risk class. Due to the unbalanced nature of the monitoring water quality dataset - with the increased majority of the dataset being in the Low-Risk class, two different sampling methods, SMOTE and ADASYN, were also used to generate synthetic data for the minority, High - Risk class, for balancing the training dataset and, thus, assisting the algorithms' training. Overall, 21 different ML - input combinations of the monthly model were tested for their predictions in the month of August of 2019 and 18 different ML-input combinations of the seasonal model were tested for their predictions in the months of July & August 2019. In addition, 3 different combination of the most accurate monthly and seasonal models was also examined. Finally, the checking past events approach that WUs commonly used was also tested for comparison. The results indicated that:

- Checking past years events approach indicated the importance of recording, maintaining and systematically check the past data for preventing some of the future events. However, machine learning based models outperformed this approach.
- The monthly seasonal model produces more accurate results because has less imbalanced dataset than the seasonal model.
- The monthly model that uses RF without additional synthetic data was the worst monthly model as managed to predict a maximum of 53% of the SRs that belonged in the High- Risk class.
- Both SMOTE ad ADASYN improved RF models' performance in predicting more SRs in the High-Risk class but with the cost of producing many false positives (SRs that were incorrectly classified in the High-Risk class).
- The RusBoost based models (both monthly and seasonal) had the best performance as they managed to have a balance between correctly classifying SRs in the High-risk class and not producing many false positives (TPR=0.72 - TNR=0.78 - MCC 0.44).
- Combining the best outputs in both the monthly and the seasonal approach improved the predictions outputs in terms of generating outputs that have a balance between true positives and false positives
- The main advantage of this methodology is of being an "open box" approach. It produces outputs that the user could check for understanding the classification decision criteria and for ranking the SRs from the most likely to fail to lowest likely one. Thus, this methodology could provide evidence to WUs that could support the proactive interventions in the SRs that are more likely to deteriorate.
- As a data-driven approach, this methodology is generic, and it could be also applied in other parts of the DWDS.
- Further work when more data are available, should examine if the models' performance could be improved when, in addition to the previous month inputs, the data of the previous year are included as inputs.

8. Predicting short-term chlorine losses in water distribution trunk mains using machine learning applications

8.1. Introduction

Water Utilities (WUs) usually collect time-series water quality data for specific investigations into their DWDS. The time-series data usually consist of flow, chlorine, and turbidity data in trunk mains. Most commonly, these data are collected from sensors in specific DWDS with known water quality problems, for monitoring the changes in the parameters. Typically, these datasets once checked for a certain investigation or research work are stored without further use. Thus, this chapter aims to explore the potential additional use of these types of time-series datasets as inputs in data-driven models for predicting future deterioration events.

Traditionally, predictions of the future behaviour of a water quality parameter, such as turbidity and chlorine decay, in the DWDS are made using process-based numerical models. These models attempt to mathematically describe the physical and chemical processes that occur inside the DWDS and require time-series data as input for their calibration. In general, process-based models describe the processes inside the DWDS accurately. However, for this hydraulic simulation, they require extended information regarding the DWDS characteristics (e.g., pipe material), accurate time-series data of flow, water temperature and all the water quality parameters whose behaviour is under investigation and a good understanding of the DWDS by the user. In addition, these models are only site-specific and require a lot of computational time for their simulation.

Data-driven techniques could be an alternative approach for the prediction of the water quality behaviour. This is because, as shown in the previous chapters, these are not site-specific and thus, once created, could be applied in any site with similar characteristics and where sufficient data are available. In addition, for their training, no hydraulic model is required and therefore the computational time is minimal. In the past, machine-learning methods were used in various works for the prediction of turbidity in water distribution trunk mains (WDTM) (Meyers, Kapelan, and Keedwell 2017; Kazemi et al. 2018) and chlorine concentrations on customers taps (Gibbs et al. 2006) that are also mentioned in the Literature review chapter (Chapter 2).

In this chapter, the aforementioned methodology that Kazemi et al. (Kazemi et al. 2018) developed, is adapted and redeveloped for the prediction of chlorine losses (ΔCl) at the WDTM end point. More specifically, the model, presented here, is a regression-based model that predicts future chlorine loss events, at the end of the WDTM up to certain hours ahead, using either past temperature and chlorine loss data or past flow and chlorine loss data. This methodology uses three different ML techniques for the future chlorine loss predictions, the

Artificial Neural Network (ANN) based on the nonlinear autoregressive exogenous network (NARX), the Feed-Forward (FF) ANN and Random Forest (RF). The methodology is tested in three WDTM with different hydraulic characteristics located in the same SW DWDS. This work investigates the general performance of the methodology and compares its performance using different parameters (temperature or flow) and different ML techniques. The overall aim of this chapter is to produce a predictive model that could be used by WUs' operational staff to proactive intervene in their DWDS and reduce the risk of distributing water with low disinfectant concentration to their customers. This chapter addresses objectives 3 & 4.

8.2. Machine learning application steps

The machine learning application steps for this investigation is as follows:

a. Define the water quality problem

In this chapter, the aim is to predict a future ΔCl event at the end of the water distribution trunk mains. The water quality problem is defined as follows:

Could the chlorine loss events at the end of the WDTM be predicted up to certain hours ahead using sensor time-series data?

b. Type of the available data

15 minute timestep chlorine, flow and temperature time-series data

c. Define required output

Chlorine losses predictions up to *n* hours ahead

d. Machine learning selection

By following the machine learning tree presented in chapter 3, the most appropriate methods for this investigation are the ANNs (FF and NARX) and the random forest algorithm.

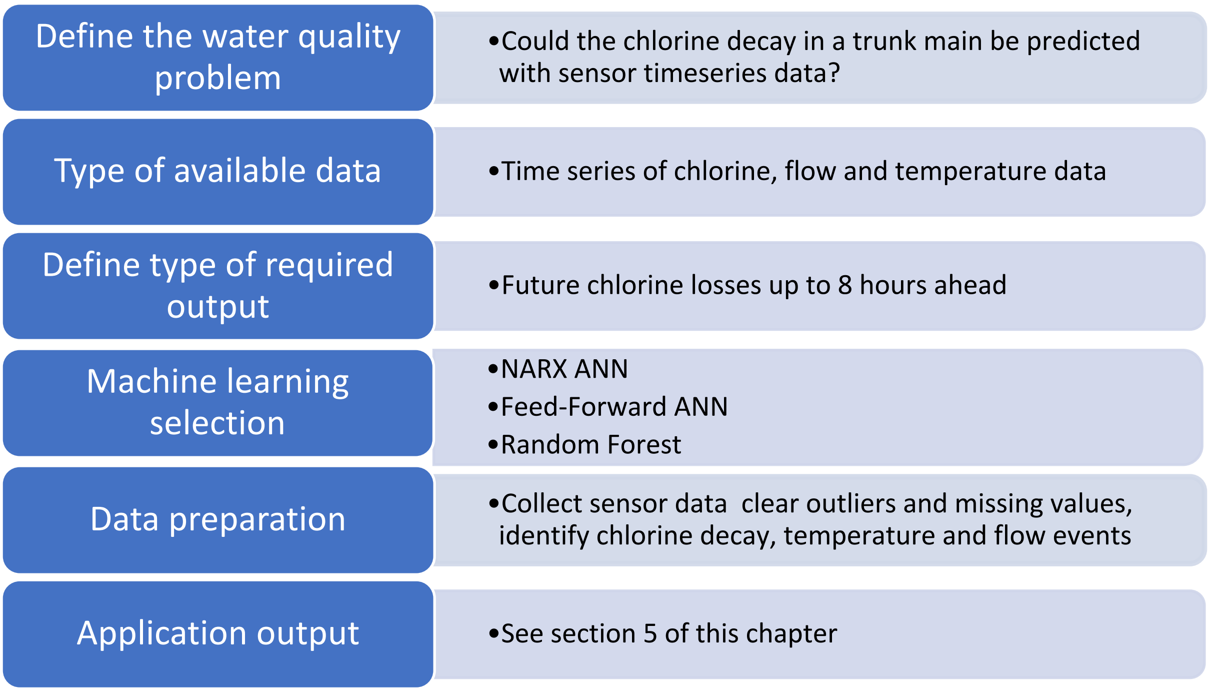
e. Data preparation

The available data for this work were created for the purposes of a hydraulic discolouration research study. Once collected, the dataset should be cleaned from outliers, missing values and prepared for the analysis. In addition, the chlorine losses at the end of the trunk mains should be calculated and the chlorine loss events should be identified.

f. Application output

The application outputs would be a file that contains the model's chlorine losses prediction per timestep. In addition, the model produces graphs where the predicted chlorine loss events are compared with the real chlorine loss events for comparison. The application output is presented in the following sections of this chapter.

The machine learning application steps for this investigation are summarized in the following figure.



Define the water quality problem
- Could the chlorine decay in a trunk main be predicted with sensor timeseries data?

Type of available data
- Time series of chlorine, flow and temperature data

Define type of required output
- Future chlorine losses up to 8 hours ahead

Machine learning selection
- NARX ANN
- Feed-Forward ANN
- Random Forest

Data preparation
- Collect sensor data clear outliers and missing values, identify chlorine decay, temperature and flow events

Application output
- See section 5 of this chapter

Figure 8.1: Machine learning application steps for the prediction of short-term chlorine loss events in water distributions trunk mains

8.3. Site description, data collection and data preparation

8.3.1. Site details and available dataset

The data-sets selected for this investigation were created and collected for a research work on discolouration management in a SW DWDS in Scotland, UK (Sunny et al. 2017). As figure 8.2 shows, the study area consists of 3 different trunk mains (TM-1, TM-2, TM-3) fed with water from the same water treatment work (WTW). Chlorine concentrations were monitored with a frequency of 15 minutes at the WTW outlet pipe and at the end of each trunk main prior to reaching the DMAs. Water temperature was also measured at the end of each trunk main, and flow was measured at the start of each trunk main with the same frequency. The main pipe characteristics for all the trunk mains are similar (table 8-1), however, during the study period, different flow conditions were applied at each one of the mains for the investigation of the impact of flow conditioning on chlorine decay, and, therefore, their hydraulic characteristics differed (Sunny et al. 2017). More specifically, in TM-1 a flow conditioning that had a 40% shear stress increase in addition to the peak shear (normal conditioning) was applied. In TM-2 a flow condition of 15% increase to the peak (passive conditioning) was applied. Finally, in TM-3 had only two flow conditioning interventions, one at the biggening of the investigation period and one at end of the investigation period, 12 months later.

The chlorine loss ΔCl during the travel from the WTW till the end of each trunk main at each timestep is calculated after assuming that during the study period there was no leakage in the trunk mains which means that there were no changes in the monitoring flow during the travel of the water through the trunk mains. Therefore, ΔCl at each timestep and for each trunk main is calculated as the difference between the measured chlorine concentration in the WTWs and the chlorine concentration at the end of each trunk main at a time step Δt which is equal to the required time for the water to travel the pipe's length.

Table 8-1: Water distribution trunk mains characteristics

Trunk main	Mean internal diameter [mm]	Pipe material	Velocity [m/s]	Length from WTW outlet to downstream logger [km]
TM-1	304.8	Unlined CI (25% lined)	0.6	6.4
TM-2	406.8	Unlined CI	0.3	5.6
TM-3	304.8	Unlined CI	0.4	5.9

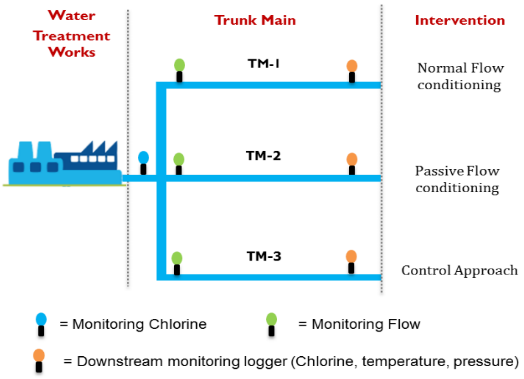


Figure 8.2: Schematic of the DWDS trunk mains (Sunny et al.2017)

Overall, three different datasets, one for each water distribution trunk main, are collected. Each dataset consists of 15 minutes time lag ΔCl, flow and temperature data for a period of 7 months. During this period, there are some months with no available chlorine or temperature data that are excluded from the analysis.

8.3.2. Data preparation

8.3.2.1. Removing outliers and missing values

As in every time-series dataset, several outliers, either spikes or zero and negative values, have been found in these datasets. The spikes in the datasets could be identified as they occur in very short time (usually one timestep) and their values are much higher compared to the values of their neighbour measurements. As the datasets are large, a gradient algorithm is created in MATLAB® version 2019b to identify and remove these outliers (see github repository. The algorithm, firstly, identifies and removes the zero and negative values and replaces them with missing values. Then it computes the gradient between each point and its previous one and if it is greater than a certain threshold, the data point is replaced with a missing value as well.

Once the outliers are removed, the next step is to remove and replace the missing values. In this work, when 4 or more consecutive timesteps with missing values are identified (1 hour without any measurement), are ignored in the final dataset. In all the other timesteps with missing measurements, the missing values are filled using the spline interpolation in MATLAB.

8.3.3. Smoothing the data

The last step of the data preparation is to smooth the data to remove noise that could affect the training of the predictive model. An algorithm that uses the cubic spline function for smoothing the dataset is created in MATLAB® version 2019b. The smoothing length is set equal to 2Δt i.e. 30 minutes as there is a lot of noise in the ΔCl data. A small part of TM-3's ΔCl, flow and temperature original and smoothed data are presented in figure 8.3.

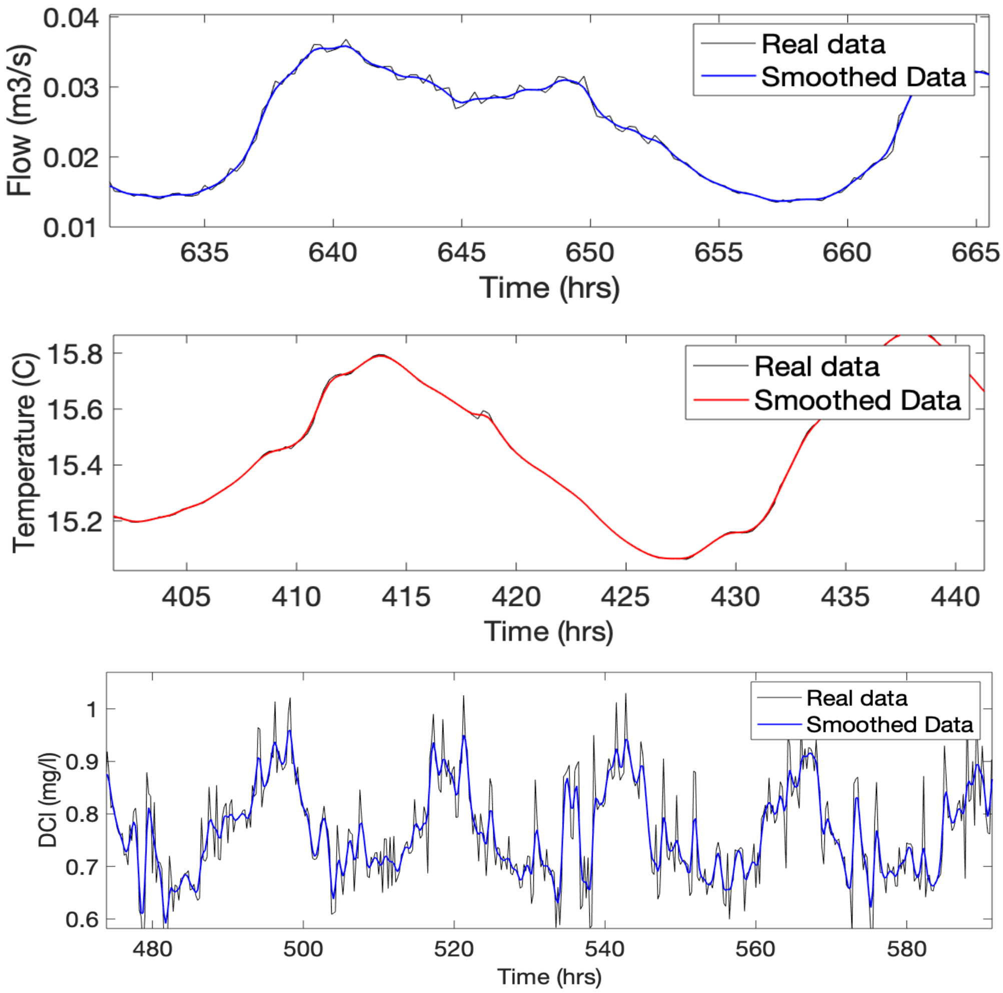


Figure 8.3:Original and smoothed flow (top), temperature (middle) and ΔCl (bottom) data for TM-3

8.4. Data-driven methodology

8.4.1. Detecting high Cl losses events

Commonly, the regression-based models are trained using all the available time-series data (Nourani, Elkiran, and Abba 2018; Jayaweera, Othman, and Aziz 2019; Filipe et al. 2019). However, in this case, the aim is to predict potential future ΔCl events and instead of training the model using all the available data, only the past ΔCl events are used in this process. Therefore, prior to the model training, the ΔCl events and the temperature or flow events related to these events are identified as follows:

a. Cl loss events detention

A model is created in MATLAB® version 2019b to find the Cl loss peaks and extract the events period that start up to a certain time before and continue up to a certain time after each one of these peaks. As chlorine consumption is highly dependent on the seasons of the year, to reduce the seasonality effect the events were selected based on the magnitude of the events instead of using the absolute values. The model extracts the Cl loss events following these 4 steps:

1. The event peaks are detected as the local maximum values.
2. For each peak, the event starting and finishing points are identified based on the gradient of the time-series around the event peak. An event starts when a change in the gradient is noticed, up to a certain period before the peak, and ends when the change in the gradient ends, up to a certain period after the peak. In this analysis, the gradient threshold was set equal to 0.02 mg/l per 15 min, in other words 0.02 mg/l per time step. The period prior to the peak and after the peak are equal to each other and also are equal to half of the required forecasting time. For example, if the required forecasting period is 4 hours ahead, the model searches for a gradient of 0.02mg/l up to 2 hours before and up to 2 hours after the peak.
3. The third step is to find the "base value" of the event which is equal to the pre-event ΔCl value and then the "base line" that connects all the base values is drawn.
4. The magnitude of the event is calculated by extracting the base value from the peak ΔCl value and if this is above a certain threshold (in our case was set up to 0.15mg/l), the event is selected for further analysis.

b. Flow and Temperature event detection

The flow and temperature events are detected once all the Cl loss events are detected. These are defined as sudden changes and peaks that could be associated with each one of the ΔCl events. Flow or temperature events associated with Cl loss events were selected by tracing the datasets over a period of up to n hours before the ΔCl event. A flow and temperature detention model created in MATLAB® version 2019b was used for the identification of the events. The model follows 5 steps with the first 4 being similar to the steps followed in the Cl loss detection model described above. In the fourth step the threshold used for defining a flow event was equal to 10l/s and the one used for defining a temperature event was equal to 0.3 °C. The fifth and final step is to remove all the Cl loss events where no associated flow or temperature peaks are found. Figure 8.4 shows 2 Cl loss event detections in trunk main TM-3 with their associated flow events.

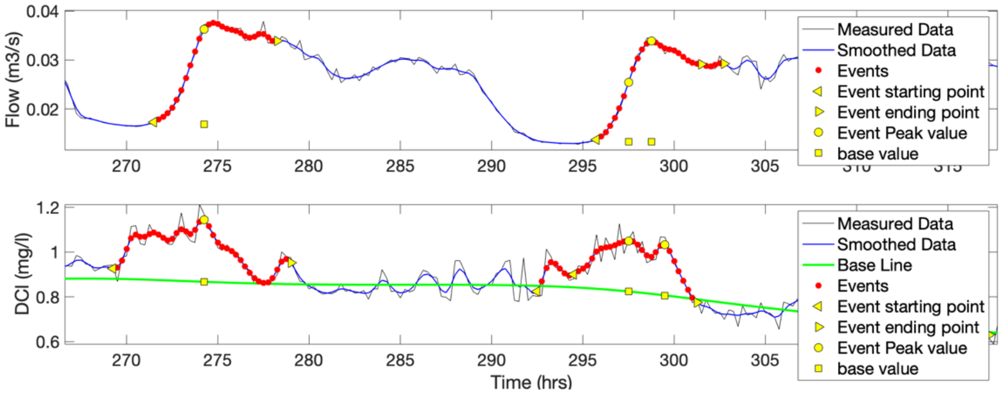


Figure 8.4: Two ΔCl detected events with their associated flow events in TM-3

8.4.2. Predictive model

A multistep prediction model was created using the NARX ANN, the Feed-Forward ANN and the RF algorithms in MATLAB® version 2019b. The model user should define which one of these algorithms and which of the temperature or flow parameters would like to use prior to the model's application. The model is trained using as inputs a number of past flow or past temperature events (10 to 12 past events) and as targets their associated past Cl loss events without considering the temporal distance between these events. Once trained, the model predicts the Cl loss values up to certain hours ahead based on a current flow or temperature event.

The characteristics of each machine learning algorithm are as follows:

- NARX algorithm: 10 hidden layers, 3 input delays and 3 feedback delays
- FF algorithm: 1 hidden layer with a size of 14 neurons
- RF algorithm: 1000 weak trees set the ensemble, minimum number of observations per tree set to 5

8.4.3. Use of the event risk parameter

Kazemi et al. (Kazemi et al. 2018) in their predictive turbidity model suggested an extra input parameter (E_p) to quantify the risk of new turbidity event occurrence based on the temporal distance between this event and a previous captured turbidity event. E_p at the current time t is defined locally as the temporal distance (L_{ev}) between an imaginary event - a weighted average of n past events- and t, divided by the imaginary event magnitude (H_{ev}) following this equation:

$$E_{pt} = \frac{L_{ev}}{H_{ev}} \ where \ L_{ev} = \frac{\sum_{i}^{n} \ (L_i H_i)}{\sum_{i=1}^{n} \ H_i} \ and \ H_{ev} = \frac{1}{n}\sum_{i=1}^{n} \ H_i$$

t: current time

E_{pt}: Event risk at current time t

L_{ev}: Temporal distance between the imaginary event and t

H_{ev}: Magnitude of the imaginary event

i: a past event

n: total number of past events included for the calculation of the imaginary event

H_i: Magnitude of the past event i

L_i: temporal distance between event i and current time t

This equation determines that the longer the temporal distance to past turbidity events or the lower the magnitudes of these events, the higher the risk of a new event to occur at current time. E_p describes in the best the risk of a frequency of a turbidity event, however, the factors that increase the Cl consumption inside the DWDS are usually not related to the occurrence and the magnitude of past events. Nevertheless, in this chapter the E_p parameter was also used as an extra input in the NARX predictive model, and the performance of the NARX + E_p model was compared with the performance of the NARX, FF and RF models where the E_p was not included as input.

8.4.4. Summarized model's inputs and outputs

Once the Cl losses events and their related flow and the temperature events are detected, the data-driven predictive model used 4 possible combinations of input parameter/output parameters for training:

1. Flow events for inputs / Cl losses events for outputs
2. Temperature events for inputs / Cl losses events for outputs
3. Flow events and event risk for inputs / Cl losses events for outputs
4. Temperature events and event risk for inputs / Cl losses events for outputs

This predictive model uses 3 different ML methods, RF, NARX ANN and FF ANN. The event risk parameter was only used in the NARX model. Therefore, the total number of inputs - ML combinations investigated in this chapter was 8 (NARX - flow input, NARX - temperature, NARX - flow and event parameter, NARX - temperature and event parameter, RF - flow input, RF - temperature, FF - flow input, FF - temperature).

The model predicts Cl losses values every 15-minute and up to a predictive period set by the user. So, for example for a predictive period of 4 hours the model predicts 16 different values. Once the model is trained, it predicts one step ahead and reruns up to the set period - for a 4 hour period reruns for another 15 times. This approach was followed as building a new model for each new time-step approach is practically impossible when the temporal distance between the timesteps is that short.

8.4.5. Performance metrics

The metrics used for the evaluation of the models and for the comparison of their performance were the Mean Average Error (MAE), the Mean Squared Error (MSE), and the Normalised Mean Squared Error (NMSE).

8.5. Results and discussion

The aim of this chapter is to suggest the best algorithm with the most appropriate input parameter for the prediction of chlorine losses at the end of the WDTM. To accomplish that, the process followed was, firstly, to train the model considering different combinations of algorithms and input parameters for each one of the trunk mains, secondly, to use the trained model for the prediction of a new Cl loss event, and, finally, to compare the outputs using the performance metrics. This process is also presented in figure 8.5.

Overall, 8 different combinations were tested (NARX - flow, NARX - temperature, NARX - flow & E_p, NARX - temperature & E_p, FF - flow, FF - temperature, RF - flow, RF - temperature) for each trunk main and for different forecasting horizons (2-10 hours). Table 8-2 shows the performance results for each model simulation at each one of the WDTM and for a specific forecasting horizon. Each model has taken its name using the initials of the trunk main that was applied, the ML technique that was used, its input parameter and its predictive horizon hours. Therefore, for example a model that was applied in TM-1, trained using the NARX technique and temperature as input data, and made predictions for up to 4 hours ahead, was named TM1-NARX -T-4.

As table 8-2 indicates, the maximum prediction horizon varies from trunk main to trunk main and is also dependent on the selected input parameter. In general, the smallest predictive horizon was achieved in TM-2 where the ΔCl was predicted up to maximum 4 hours ahead using flow as input parameter, and the largest was achieved in TM-3 where the model managed to predict Cl losses up to 10 hours ahead when flow was used as input parameter. This could be, potentially, explained by the fact that there was a period with many chlorine spikes and negative chlorine measurements in the TM-2 dataset. These noisy data points, once removed or replaced with the process described in the previous section, created a smaller and weak final dataset that affected the training of the model. In addition, the different flow conditions in the three trunk mains during the study period, indicate different hydraulic characteristics that also influenced the performance of the model. The outputs in this work indicate that the model could adapt the normal conditions (TM-3) and the high flow conditioning changes (TM-1) but cannot adapt in small ones (TM-2). However, the flow condition impact in the model's performance should further investigated either in this DWDS or in a different one with more sensor data and, most importantly, with better quality data.

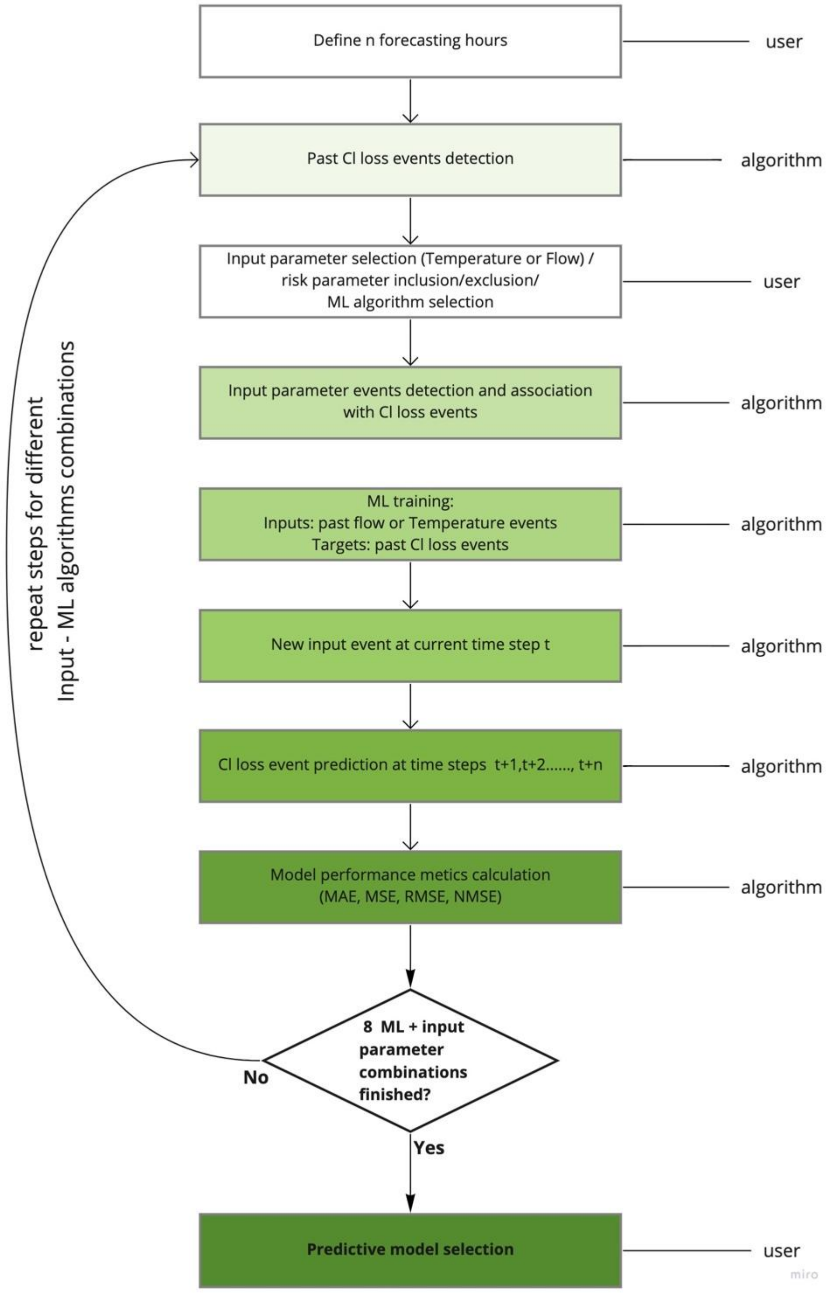

Define n forecasting hours
user
Past Cl loss events detection
algorithm
Input parameter selection (Temperature or Flow) / risk parameter inclusion/exclusion/ ML algorithm selection
user
Input parameter events detection and association with Cl loss events
algorithm
ML training: Inputs: past flow or Temperature events Targets: past Cl loss events
algorithm
New input event at current time step t
algorithm
Cl loss event prediction at time steps t+1,t+2......, t+n
algorithm
Model performance metics calculation (MAE, MSE, RMSE, NMSE)
algorithm
8 ML + input parameter combinations finished?
No
Yes
Predictive model selection
user
repeat steps for different Input - ML algorithms combinations
miro

Figure 8.5: A simple schematic that describes the process for selecting the best predictive ΔCl model

By comparing the performance metrics of all the models' outputs, flow is a better input parameter to temperature for predicting Cl losses in this DWDS. The only temperature model that performed better than its counterparts flow models was TM-1 - NARX - T - 8 (NMSE 0.54

vs NMSE 0.74 for the TM-1 - RF - F - 8). All the other temperature models have unacceptable NMSE values (all of them above 1) and underperformed in all the other performance metrics compared to the flow models. In addition, the smaller number of detected temperature events related to ΔCl events in all three trunk mains compared to the flow events (e.g., in TM-2 there were 12 captured temperature events and 51 captured flow events) implies that the temperature model misses the majority of Cl loss events. As regards the ML algorithms, Table 8-1 justifies that the best among the applied algorithms is NARX ANN as in almost all the cases outperformed all the others. In an overall hierarchical comparison, NARX was the best model with RF and FF being, with a very close distance to each other, second and third respectively, and, finally, by far the worst algorithm was NARX with E_p. These findings confirm, also, the initial hypothesis that the E_p risk parameter is not a parameter that could mathematically describe the frequency occurrence of the Cl loss events. Figure 8.6 shows three timeseries comparisons between monitored and predicted values by the four flow input ML algorithms.

Table 8-2: Performance of NARX, NARX+Ep, RF and FF using different inputs and predictive horizons

Model NAME	Trunk Main	ML algorithm	Input parameter	Predictive horizon (hrs)	MAE (mg/l)	MSE (mg/l)^2	RMSE (mg/l)	NMSE
TM-1 - NARX - F - 8	TM-1	NARX	Flow	**8**	0.06	0.007	0.084	2.12
TM-1 - NARX - FEp - 8	TM-1	NARX	Flow - Ep	8	0.06	0.005	0.071	1.49
TM-1 - RF - F - 8	**TM-1**	**RF**	**Flow**	**8**	**0.04**	**0.002**	**0.045**	**0.74**
TM-1 - FF - F - 8	TM-1	FF	Flow	8	0.05	0.003	0.055	0.94
TM-1 - NARX - F - 6	**TM-1**	**NARX**	**Flow**	**6**	**0.04**	**0.0007**	**0.027**	**0.21**
TM-1 - NARX - FEp - 6	TM-1	NARX	Flow - Ep	6	0.06	0.005	0.071	1.42
TM-1 - RF - F - 6	TM-1	RF	Flow	6	0.05	0.005	0.071	1.29
TM-1 - FF - F - 6	TM-1	FF	Flow	6	0.04	0.005	0.071	0.89
TM-1 - NARX - T - 8	**TM-1**	**NARX**	**Temperature**	**8**	**0.04**	**0.003**	**0.055**	**0.54**
TM-1 - NARX - TEp - 8	TM-1	NARX	Temperature -Ep	8	0.07	0.009	0.095	1.76
TM-1 - RF - T - 8	TM-1	RF	Temperature	8	0.06	0.006	0.077	1.13
TM-1 - FF - T - 8	TM-1	FF	Temperature	8	0.05	0.005	0.071	1
TM-2 - NARX - F - 2	**TM-2**	**NARX**	**Flow**	**2**	**0.03**	**0.0009**	**0.030**	**0.38**
TM-2 - NARX - FEp -2	TM-2	NARX	Flow - Ep	2	0.04	0.003	0.055	1.11
TM-2 - RF - F - 2	TM-2	RF	Flow	2	0.04	0.002	0.045	0.6
TM-2 - FF - F - 2	TM-2	FF	Flow	2	0.04	0.002	0.045	0.77
TM-2 - NARX - F - 2	**TM-2**	**NARX**	**Flow**	**4**	**0.02**	**0.0003**	**0.018**	**0.15**
TM-2 - NARX - FEp -2	TM-2	NARX	Flow - Ep	4	0.06	0.005	0.08	2.44
TM-2 - RF - F - 2	TM-2	RF	Flow	4	0.03	0.001	0.034	0.54
TM-2 - FF - F - 2	TM-2	FF	Flow	4	0.03	0.001	0.026	0.34
TM-2 - NARX - T - 2	TM-2	NARX	Temperature	2	0.05	0.003	0.055	3.44
TM-2 - NARX - TEp - 2	TM-2	NARX	Temperature -Ep	2	0.06	0.006	0.077	8.53
TM-2 - RF - T - 2	TM-2	RF	Temperature	2	0.06	0.006	0.077	8.84
TM-2 - FF - T - 2	TM-2	FF	Temperature	2	0.05	0.005	0.071	3.61
TM-3 - NARX - F - 10	**TM-3**	**NARX**	**Flow**	**10**	**0.04**	**0.002**	**0.045**	**0.61**
TM-3 - NARX - FEp - 10	TM-3	NARX	Flow - Ep	10	0.07	0.007	0.084	1.94
TM-3 - RF - F - 10	TM-3	RF	Flow	10	0.05	0.003	0.055	0.94
TM-3 - FF - F - 10	TM-3	FF	Flow	10	0.05	0.003	0.055	0.89
TM-3 - NARX - T - 4	TM-3	NARX	Temperature	4	0.03	0.001	0.032	1.93
TM-3 - NARX - TEp - 4	TM-3	NARX	Temperature -Ep	4	0.03	0.001	0.032	1.86
TM-3 - RF - T - 4	TM-3	RF	Temperature	4	0.04	0.003	0.055	4.29
TM-3 - FF - T - 4	TM-3	FF	Temperature	4	0.04	0.003	0.055	3.48

Figure 8.6a shows a 5-hour event in TM-1; Figure 8.6b shows a 4-hour event in TM-2; and Figure 8.6c shows a 10-hour event in TM-3. These plots indicate that NARX captures the event better compared to the other models and agree with the performance metrics presented in table 8-2. Therefore, for the available dataset, the model that uses flow data as inputs and the NARX algorithm for training is the best model for the prediction of Cl losses at the end of the WDTM.

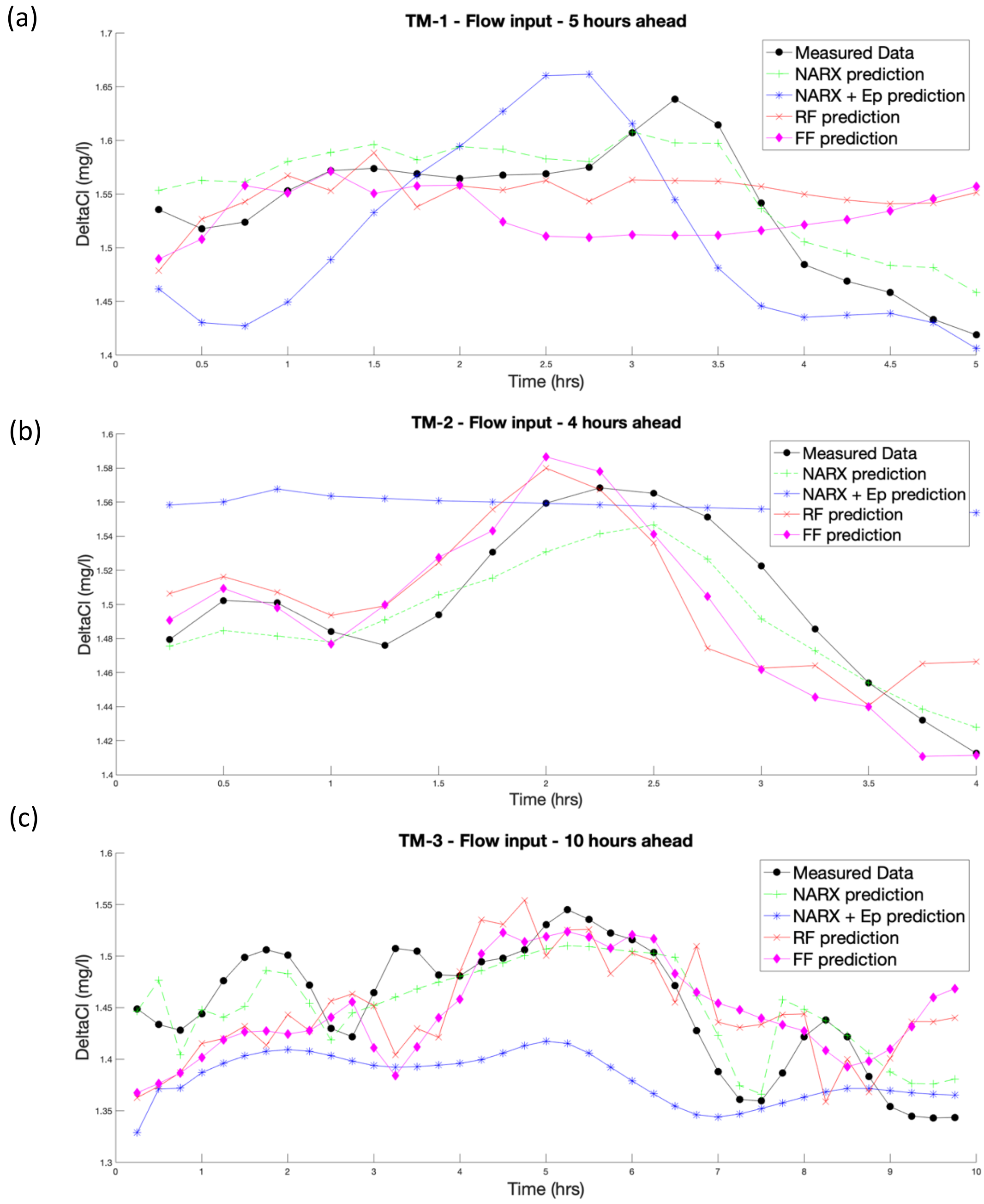


Figure 8.6: Predicted vs Measured data of ΔCl event in the three trunk mains: a) 5-hour event in TM-1, b)4 hour event in TM-2,c)10 hour event in TM-3

8.6. Operational benefits of models' application

The predictive model was tested in a small dataset with many missing data and could not capture the seasonality changes in chlorine consumption. Therefore, a further investigation should be implemented in a DWDS where at least a full year of chlorine and flow time-series data are available to test the model's ability to adapt the chlorine seasonal variations. Once this is accomplished and the prediction outputs are acceptable, the model could be used as a predictive tool to support the proactive strategies of the WUs. More specifically, it could be used in combination with an alarm system connected to the Supervisory control and data acquisition (SCADA) system of the DWDS. The SCADA system could inform the predictive model for a sudden change in the flow and then the model, once trained using the past 10 to 12 previous events, could forecast a potential future ΔCl event. If the prediction exceeds a defined threshold the alarming system is activated and informs the water operators for this potential event. This process, in a DWDS with a modern SCADA, will not require more than 5 minutes from start to finish as the computational time required to train the model is minimal. In addition, this model could be applied for testing the impact of a potential maintenance intervention in the WDTM, which will increase its flow (e.g., flushing or conditioning of the WDTM). More specifically, the model could be asked to predict chlorine losses caused by an artificial flow event that simulates the maintenance intervention to investigate the impact of this intervention in the chlorine concentrations in the drinking water exiting the WDTM.

8.7. Conclusions

In this chapter, a data-driven methodology for the forecasting of a future ΔCl event at the end of a WDTM is presented. This methodology, firstly, identifies past ΔCl events and their related past temperature or flow events. Then, it imports these events as inputs for the training of the predictive model that uses one of the following ML techniques: NARX ANN, FF and RF. An extra input risk parameter $E_{p,}$ used in another work to capture the temporal distance between turbidity events (Kazemi et al. 2018), was introduced as an optional input for the predictive model as well. The methodology is tested in a dataset taken from 3 WDTM mains (TM-1, TM-2, TM-3) of a DWDS that belongs to SW. The aim is to, firstly, investigate the predictive ability of the model and then identify which input (flow or temperature) - ML algorithm (NARX ANN, FF, RF) combination performs better to provide an accurate predictive model that could be used by the WUs for supporting their proactive strategies. Overall, 8 input - ML algorithm combinations of the model were applied at each WDTM (flow - NARX, temperature - NARX, flow + E_p - NARX, temperature + E_p - NARX, flow - RF, temperature - RF, flow - FF and temperature - NARX). The main conclusions are as follows:

- The model managed to predict accurately a future event with a period of 6 hours ahead in TM-1, 4 hours ahead in TM-2 and 10 hours ahead in TM-3.

- The performance metrics (MAE, MSE, RMSE, NMSE) indicated that flow is a better input parameter than temperature for the application of the model in these WDTM.
- NARX has been found to be the best ML algorithm with RF following in second place and FF being third.
- As expected, the use of the E_p did not improve the model's performance, contrariwise the models that included this parameter had the worst performance in all three WDTM.
- This predictive model has the potential of becoming an accurate supporting tool in the WUs' decision making for proactive intervention. However, as the available data for this work were taken for a small period (overall less than 5 months of data were available if we include the spikes and the missing data), further research is required using larger datasets - with at least 1 year of data - to investigate the ability of the model to adapt in the seasonality changes of the Cl concentrations.

9. A data-driven investigation on the performance of Balmore water treatment work

9.1. Introduction

Water treatment works (WTWs) are complicated systems consisting of various treatment stages used by WUs to produce drinking water of high-quality standards for their customers. Inside the WTWs various parameters including the operational function of the electromechanical equipment, the water flow, the water level in the treatment tanks and water quality parameters such as pH, turbidity, and chlorine, are monitored with sensors connected to the supervisory control and data acquisition (SCADA) system. SCADA systems allow WUs' operators to control in real-time the various treatment processes, to interact with the various devices, to adapt the treatment stages when changes in either the water flow or the water quality occur and, finally, to record the data and the various quality events into log files. Most of the parameters are measured with a 5 min to 10 min frequency and, therefore, large datasets are created that, when explored properly, could give significant knowledge regarding the optimisation of WTWs' performance, the parameters that influence the quality of the drinking water entering the DWDS, and the prediction of future water quality deterioration.

Due to this large data availability in the WTWs and the wastewater treatment plants (WWTPs), there is a plethora of projects in the literature where data-driven methods, such as machine learning (ML), were applied. More specifically, data-driven applications were utilised for the optimisation of a process in the works (Asadi et al. 2017; K. Zhang et al. 2013; Jayaweera, Othman, and Aziz 2019), minimise the treatment works energy consumption (Filipe et al. 2019) and improving the overall performance of the treatment works (Nourani, Elkiran, and Abba 2018; Dairi et al. 2019; Mohammed, Hameed, and Seidu 2017). Most of the recent research works that used artificial intelligence (AI) and ML in WTWs are presented in a review paper by Li et al. (Li et al. 2021). In this paper, they grouped these works based on the treatment process that the projects focused on, then they indicated the advantages and the weaknesses of the ML techniques over other approaches in the analysis and control of the WTWs and, finally, they discuss the potential that AI technologies could offer as intelligent models for supporting the management of the DWDS.

This chapter aims to examine further the ability of ML techniques on analysing the available data and supporting WTWs management. More specifically, in this chapter, a data-driven investigation on the bacteriological activity of Balmore WTW, one of the largest Scottish Water (SW) treatment sites, is presented. SW recently installed two online flow cytometers (FCM) to measure total cell counts (TCCs) in the inlet and the outlet of the treated water storage tank in Balmore to investigate the bacteriological activity of the water that reaches

Balmore water operation area (WOA). Thus, in addition to the existing time-series dataset with all the monitored parameters in the WTW, TCCs time-series with a frequency of 1 to 2 hours were also available. A data-driven investigation was made using the available data with the aim to understand the factors that could influence TCCs increase in the treated water and to forecast the TCCs behaviour in the water exiting Balmore WTW up to certain hours ahead. This investigation used both supervised ML techniques for predicting the TCCs behaviour, and unsupervised ML techniques for understanding the factors that increase bacteriological activity in the WTW. This chapter addresses objectives 2, 3 & 4.

9.2. Balmore WTW description & data preparation

Balmore is one of the largest SW's WTWs, located in the north part of Glasgow between Kirkintilloch and Bearsden. It was opened in 2000 and supplies the areas of North Lanarkshire, Falkirk, Grangemouth, West Lothian, and parts of the Glasgow area. Overall, it serves around 600000 people with water - WTW's capacity of 200000 m^3/day. Balmore treats the water coming from Loch Katrine and Loch Lomond via a pre-treatment stage and three main treatment stages - coagulation - flocculation using aluminium as a coagulant, filtering (with 6 double-staged RGFs) and disinfection. The disinfection type in this site is chlorination which is achieved with hypochlorite dosing in the water before reaching the disinfection contact tank. The hydraulic retention time (HRT) of the treatment process is estimated to be roughly 8 hours and then the treated water is stored in the treated water service reservoir for roughly 11 hours (SR_RT=11 hours). Finally, the water reaches the distribution networks via 2 pumps (1 main +1 backup). A Balmore WTW's flow schematic is presented in figure 9.1.

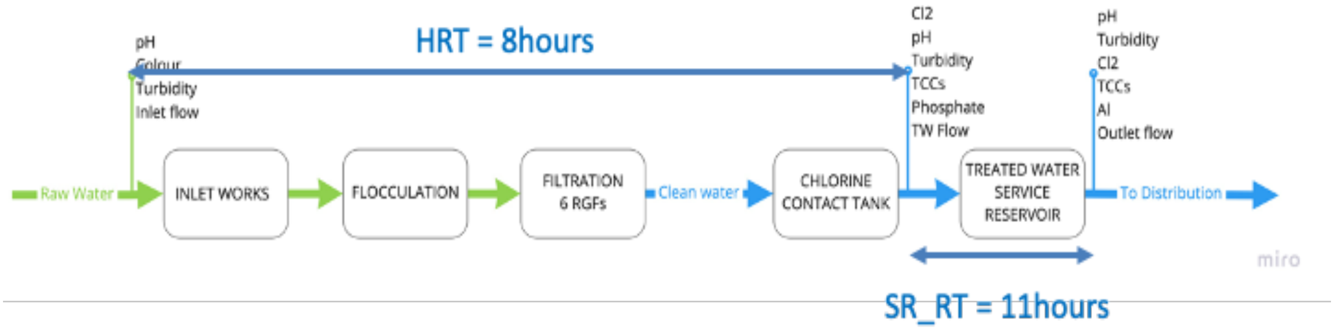


Figure 9.1: Flow schematic in Balmore WTW including the points where water quality parameters are measured

The SCADA system in this site collects water quality and flow data from the inlet of the works, the outlet of the works and from the outlet of each treatment process tank with 5 min frequency. The water quality parameters measured in this site include turbidity and pH at the inlet, the outlet and the treatment stages, colour in the inlet, Cl_2 at the disinfection contact tank outlet and the SR outlet. Due to bacteriological increase noticed in the water exiting the works in the last couple of years, SW installed, in September 2020, 2 online FCMs in the site to measure TCCs at the exit of the disinfection tank and at the outlet of the treated water service reservoir with a frequency between 1 and 2 hours.

From all the available data collected from the SCADA, in this study, the water quality and flow data from the inlet, the disinfection tank outlet, and the SR's outlet were used as the figure 9.1 shows. More specifically, the data used for this investigation were:

- Flow at the inlet (Inlet Flow - m^3/s)
- Turbidity at the inlet (Turb Inlet - NTU)
- Colour at the inlet (Colour Inlet - DegH)
- pH at the inlet (pH Inlet)
- Flow at the outlet of the disinfection tank (TW Flow - m^3/s)
- Cl_2 at the outlet of the disinfection tank (TW Cl - mg/l)
- pH at the outlet of the disinfection tank (TW pH)
- Phosphate at the outlet of the disinfection tank (TW Phosphate - ppb)
- Turbidity at the outlet of the disinfection tank (TW Turb - NTU)
- TCCs at the outlet of the disinfection tank (TW TCCs - cell counts per ml)
- Flow at the works' outlet (Final Flow - m^3/s)
- Cl_2 at works' outlet (Final Cl - mg/l)
- Turbidity at the works' outlet (Final Turb - NTU)
- Aluminium at the works' outlet (Final Alum - mg/l)
- pH at the works' outlet (Final pH)
- TCCs at the works' outlet (Final TCCs - cell counts per ml)

As the TCCs data availability starts from September the 1st 2020 the study period for this investigation starts from 31st of August 2020 up to the 12th of July 2021. Descriptive statistics of these parameters are presented in table 9-1.

Due to the difference in the measurement frequency between the WTW's flow and water quality parameters and the online FCM measurements, prior to the investigation, all the data were transformed to hourly time step data. The WTW parameters were transformed from 5min time-series data to hourly data using the mean of each parameter in the hourly time bin. However, as the FCM data had an hourly or bihourly frequency, the missing TCC data were filled using the cubic spline interpolation. In table 9-2 the descriptive statistics of the actual and the interpolated TCC data are presented.

Table 9-1: Descriptive statistics of the used Balmore WTW parameters

WTW stage	Parameter	Units	Min	Max	Mean	Std
INLET	Inlet flow	m^3/s	0	2.9446	2	0.2067
	Turb inlet	NTU	0.024	6.7655	0.7052	0.3871
	pH inlet	-	5.46	10.1281	6.71	0.8522
	Colour inlet	DegH	-3.36	112.4786	26.404	6.1446
DISINFECTION OUTLET	TW Flow	m^3/s	0	4.5224	1.988	0.2601
	TW pH	-	6.1	10.9049	8.5252	0.2132
	TW Turb	NTU	0.044	1	0.0826	0.0259
	TW Cl	mg/l	0	2	0.977	0.0723
	TW TCC	cells/ml	0	984992	296632	94589
	TW Phosphate	ppb	0	2000	327.407	41.7897
OUTLET	Final Flow	m^3/s	0	5.8258	1.933	0.2526
	Final pH	-	5.37	10.1307	8.1805	0.1595
	Final Turb	NTU	0.00024	1	0.07013	0.0219
	Final Cl	mg/l	0	1.498	0.8047	0.0549
	Final TCC	cells/ml	911	394011	61820	75504
	Final Alum	mg/l	0	0.1797	0.0072	0.0047

Table 9-2: Descriptive statistics of the actual and the hourly interpolated TCCs exiting Balmore WTW

Total Cell Counts

TCCs	Samples	Mean	Std	Median
Actual Data	4525	61805	75504	26183
Hourly Interpolated	7595	59727	73738	25778

The hourly timeseries dataset requires a reorganisation to capture the time lags in the WTW. The overall retention time in Balmore WTW is estimated to be around 19 hours (HRT+SR_RT), which means that the water enters the WTW on the 31st of August at 00:00h, passes the disinfection tank at 08:00h on the same day, and exits the WTW at 19:00h on the same day again. In addition, the TCCs predictive horizon was set to be equal to 12 hours ahead, which means that for the ML models training, the TCC measurement that corresponds to the water exiting the works at 19:00h should be the measurement taken at the WTW's outlet on the 1st of September at 07:00h. The dataset was reorganised as presented in figure 9.2.

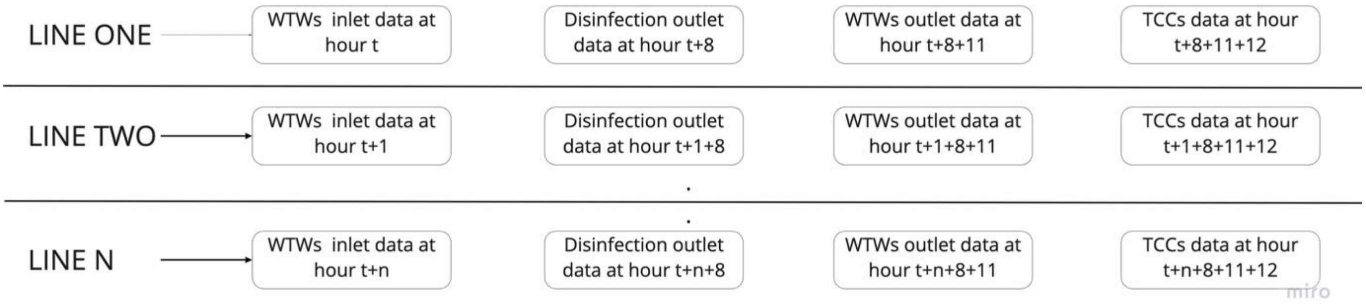

LINE ONE
WTWs inlet data at hour t
Disinfection outlet data at hour t+8
WTWs outlet data at hour t+8+11
TCCs data at hour t+8+11+12
LINE TWO
WTWs inlet data at hour t+1
Disinfection outlet data at hour t+1+8
WTWs outlet data at hour t+1+8+11
TCCs data at hour t+1+8+11+12
LINE N
WTWs inlet data at hour t+n
Disinfection outlet data at hour t+n+8
WTWs outlet data at hour t+n+8+11
TCCs data at hour t+n+8+11+12
miro

Figure 9.2: Example of the reorganisation of the dataset for filling the different time measurements

The final dataset was utilised in this format as input in SOMs and PCA methods for the identification of the factors that increase TCCs. However, for the TCCs prediction, this dataset was separated into input and output data, with input data being all the water quality (including the TW TCCs) and the flow data from all three plant points, and output data being the TCCs exiting the works.

9.3. Machine learning application steps

There are two different water quality problems to investigate. The first one is the identification of the factors that increase bacteriological activity in the water exiting Balmore WTW and, therefore, it is a correlation type of problem. The second water quality problem is the prediction of the future bacteriological behaviour in the water exiting the treatment plant and, therefore, it is a prediction problem. The machine learning application steps for these two problems are presented in the following sections.

9.3.1. Understanding the factors that increase TCCs at Balmore WTW

a. Define the water quality problem

What are the main water quality parameters related to increased bacteriological activity in the water exiting Balmore WTW?

b. Type of the available data

Water quality, flow and online FCM time-series data from the WTW inlet and the outlets of every treatment stage.

c. Define required output

The required output in this investigation is to identify the correlations between high TCCs in the WTW outlet and some of the other available water quality parameters.

d. Machine learning selection

By following the machine learning tree presented in chapter 3, the selected techniques could be either the SOMs or PCA.

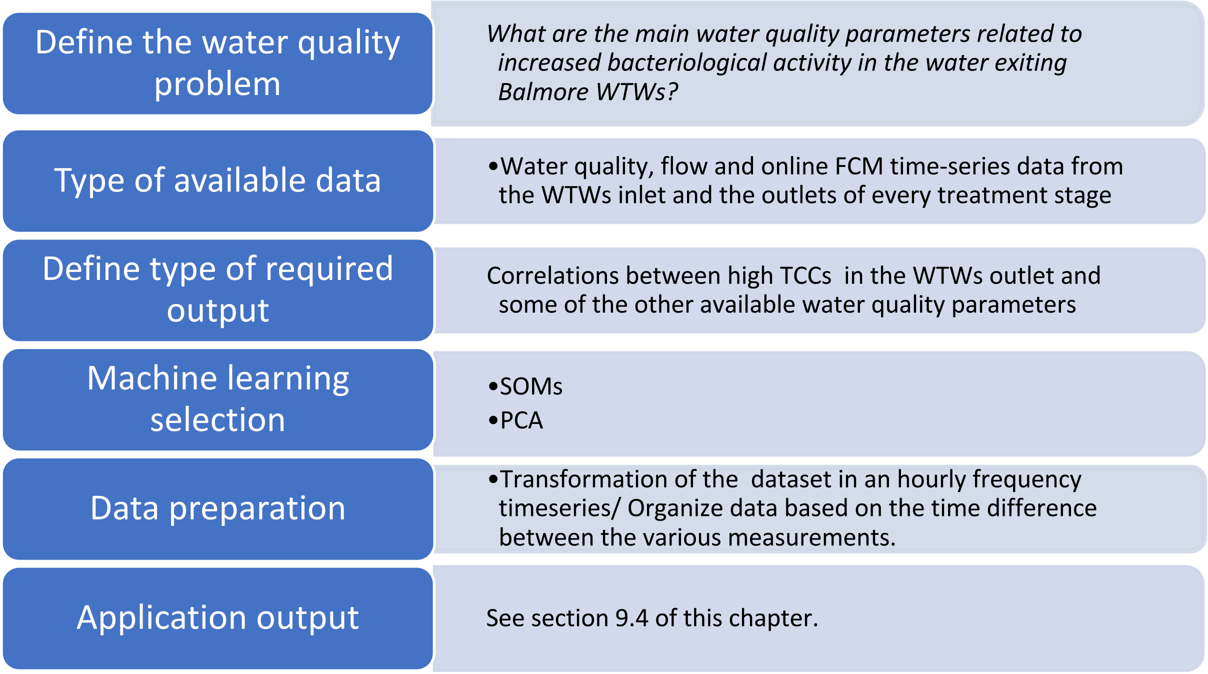
e. Data preparation

The data transformation into a similar temporal scale is required. All data were transformed in hourly time-step frequency.

f. Application output

Graphs that visualise the correlations between the various parameters and therefore the correlations between them could be explored. The procedure and the outputs are presented in this chapter.

The ML application steps for this investigation are summarized in the following figure.



Define the water quality problem	*What are the main water quality parameters related to increased bacteriological activity in the water exiting Balmore WTWs?*
Type of available data	•Water quality, flow and online FCM time-series data from the WTWs inlet and the outlets of every treatment stage
Define type of required output	Correlations between high TCCs in the WTWs outlet and some of the other available water quality parameters
Machine learning selection	•SOMs •PCA
Data preparation	•Transformation of the dataset in an hourly frequency timeseries/ Organize data based on the time difference between the various measurements.
Application output	See section 9.4 of this chapter.

Figure 9.3: Machine learning application steps for relating TCCs with other water quality parameters

9.3.2. Predicting bacteriological behaviour at Balmore WTW outlet

a. Define the water quality problem

Is it possible to predict the bacteriological activity of the water exiting Balmore WTW up to 12 hours ahead?

b. Type of the available data

Water quality, flow and online FCM time-series data from the WTW's inlet and outlets of every treatment stage

c. Define required output

TCCs prediction up to 12 hours ahead or TCCs threshold up to 12 hours ahead

d. Machine learning selection

By following the machine learning tree presented in the chapter 3, the selected techniques could be random forest (RF), feed-forward artificial neural network (FF - ANN or simply ANN), and long short-term memory networks (LSTM).

e. Data preparation

The data transformation into a similar temporal scale is required. The dataset that was already pre-processed for the correlation problem above was used. In addition, before training the model, the data were standardized.

f. Application output

Bacteriological risk ranking prediction for the water exiting Balmore WTW up to 12 hours ahead and graphs that show TCCs behaviour for the same prediction horizon. The procedure and the outputs are presented in section 5 of this chapter.

The ML application steps for this investigation are summarized in the following figure.

Define the water quality problem	*Is it possible to predict the bacteriological activity of the water exiting Balmore WTW up to 12 hours ahead?*
Type of available data	•Water quality, flow and online FCM time-series data from the WTWs inlet and the outlets of every treatment stage
Define type of required output	1.TCCs prediction up to 12 hours ahead 2.Bacteriological risk prediction up to 12 hours ahead
Machine learning selection	•ANN •Random forest •Boosting trees - RuSBoost •Deep learning - LSTM
Data preparation	•Hourly transformation as in the first water quality problem / Standardisation of dataset
Application output	See section 9.5 of this chapter

Figure 9.4: Machine learning application steps for TCCs prediction Methods

9.3.3. Self-organising maps (SOMs)

As in both 3 and 4 chapters, SOMs analysis was carried out using the MATLAB→ SOM Toolbox version 2.1 (Teuvo Kohonen 2014) in MATLAB→ version 2019b. For the analysis all the parameters of the dataset were used. The same algorithms, as the ones used in chapter 5, were used for this case study as well (see GitHub for code details). The SOMs' colour range was standardized to use all the data that were between the 5th and the 95th percentile of the dataset to avoid the skewness of the final outputs.

9.3.4. Principal component analysis (PCA)

PCA analysis was conducted using MATLAB→ version 2019b. The PCA function in MATLAB was used in an algorithm produced to apply this method.

9.3.5. Results

The output graphs from both models are presented in figures 9.5 (SOMs output) and 9.6 (PCA output). Both figures indicate that there is a perfect correlation, as expected, between the flows in all three parts of the plant and high correlation between high TCCs exiting the disinfection tank (TW_TCCs) and high TCCs in the WTW's outlet.

By checking the SOMs output, high TCCs, and, therefore high bacteriological activity in the plant's outlet, is expected when there is (i) high flow in the plant (high inlet, TW and outlet flow) (ii)high turbidity in the inlet (iii) high colour in the inlet (iv) high TCCs exiting the disinfection tank (vi) low turbidity in the water exiting the disinfection tank(vi) low free chlorine in the water exiting the disinfection tank. Moreover, some correlation between final TCCs and pH in the treated water (TW pH) is also present. Finally, there is a clear reverse correlation between aluminium in the final water and final TCCs.

As regards the other correlations, there is an interesting correlation between colour in the inlet and turbidity in the inlet. The increased colour numbers (red and yellow cells) follow the trend of the medium (cyan cells, 0.66-0.80 NTU), medium high (yellow cells, 0.90-1.1 NTU) and high (red cells, 1.1-1.2 NTU) turbidity. This trend is also followed by the medium to high TCCs exiting the disinfection tank (cyan to reads cells) which indicates its correlation with both these parameters and perhaps indicates that the higher the turbidity is in the raw water, the higher the bacteriological activity in these waters is. This SOM also shows a correlation between medium turbidity in the inlet with high turbidity exiting the disinfection tank (right top of the plane). However, this correlation should be ignored as the turbidity exiting the disinfection tank is low with a range between 0.05 to -0.12 NTU.

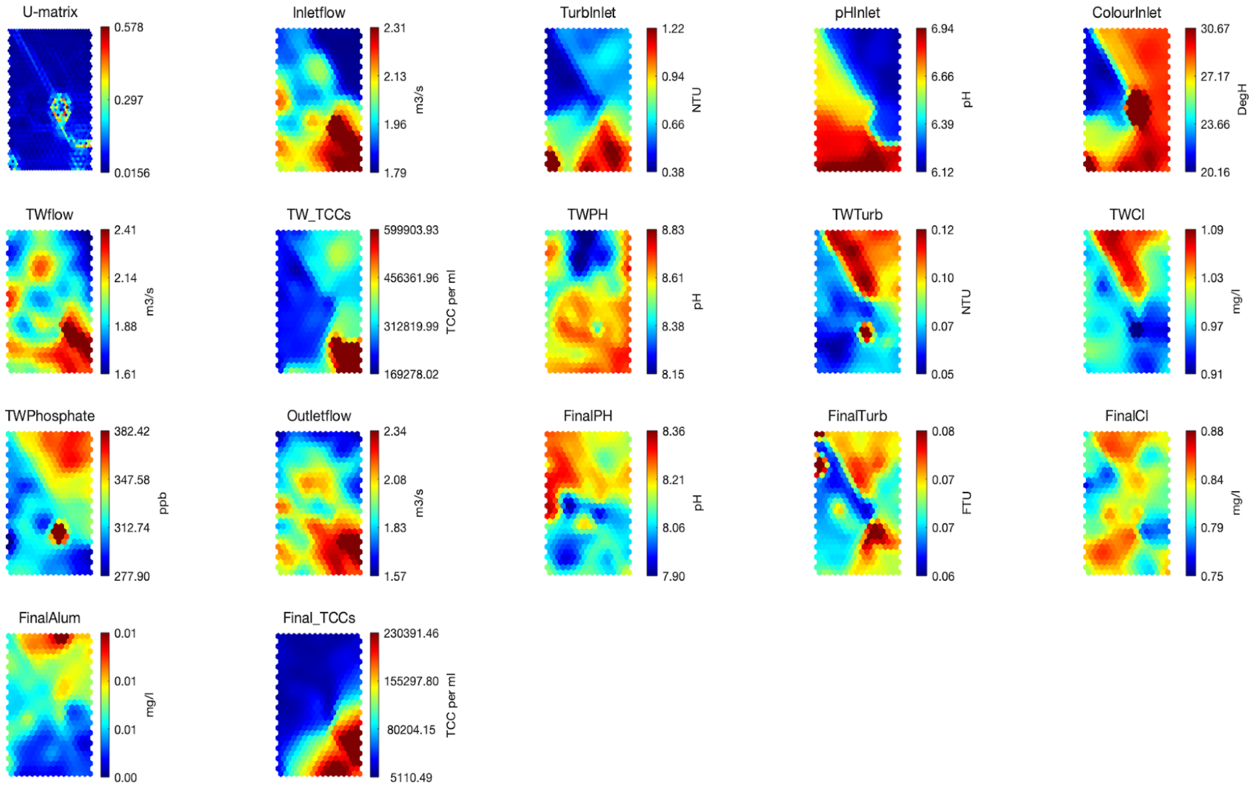


Figure 9.5: SOMs output for Balmore WTW reorganised dataset

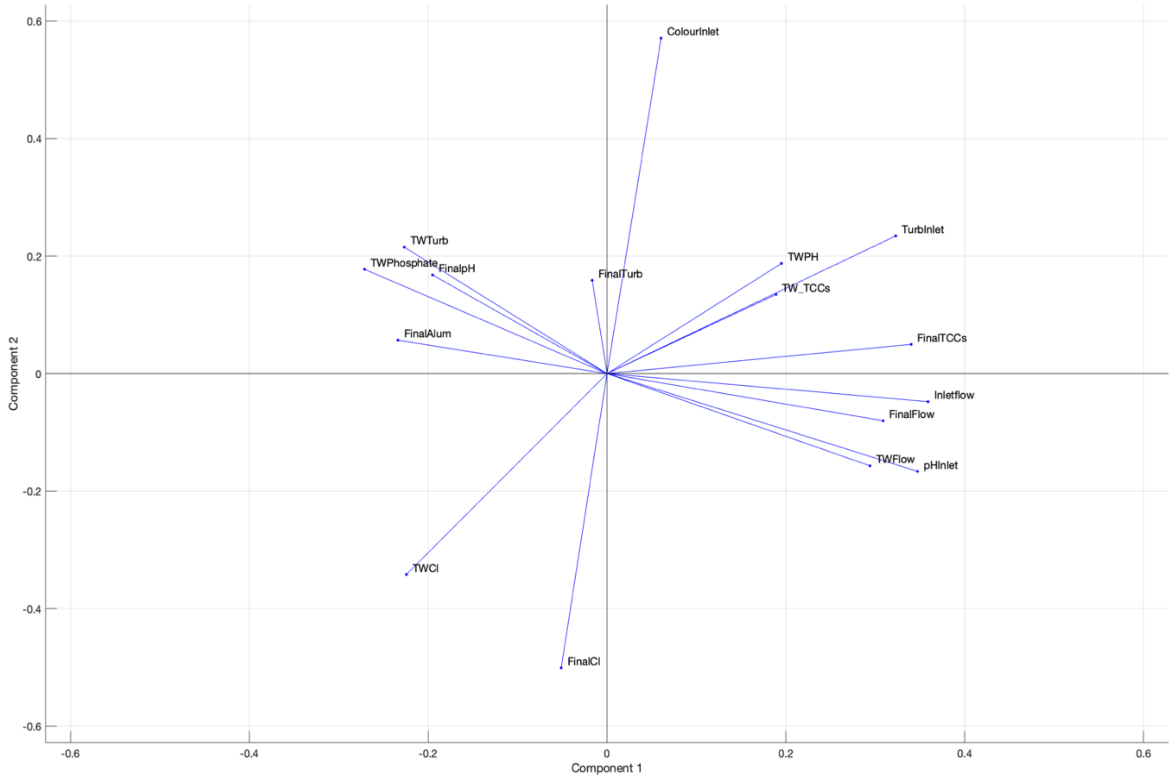


Figure 9.6: PCA output for Balmore WTW reorganised dataset

The first two PCA components explain more than half of the dataset variance (46% and 15% respectively with all the other 14 components explaining less than 8% each. Therefore, the linear correlations represented in figure 9.6 are indicative to the relationships between the various parameters in the data set.

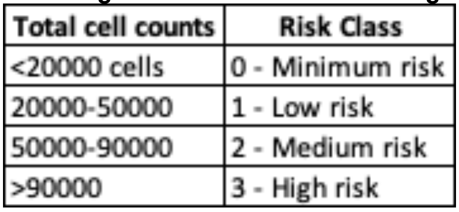
Figure 9.6 indicates that final TCCs have a clear linear correlation with all the flows in the WTW and the turbidity in the inlet, a finding that also appears in SOMs. In addition, a relative linear correlation appears between final TCCs and pH in the inlet which is not clear in the SOMs analysis. Smaller but also significant linear correlations appear between final TCCs and TCCs exiting the disinfection tank and pH in the water exiting the disinfection tank. Also, there is a reverse correlation between final TCCs and chlorine in the disinfection tank, a finding also found in the SOMs analysis. Finally, PCA did not capture any relationship between final TCCs and colour and small reverse correlation between final Aluminium and Final TCCs, correlations that appear clearly in the SOMs output.

9.4. Predicting total cell counts’ behaviour at the WTW’s outlet

9.4.1. Data preparation and model inputs

The model aims to predict the TCCs in the water exiting the WTW 12 hours ahead but also to categorise, for the 12 hours ahead, the water exiting the WTW into different bacteriological risk ranking classes. The TCCs prediction is a regression problem and, thus, for testing the model, the dataset was split into input and output data as described in section 3. The bacteriological risk ranking categorisation is a classification problem with the water exiting Balmore plant being classified into the different bacteriological risk classes based on their TCC numbers. More specifically, 4 classes were defined, based on SW’s criteria for TCCs exiting their works, the minimum-risk class when the TCCs are below 20000, the low-risk class when the TCCs are between 20000 and 50000, the medium-risk class when the TCCs are between 50000 and 90000 and the high-risk class when the TCCs are above 90000 (see table 9-3).

Table 9-3: Risk ranking classes for the water exiting Balmore WTW



Total cell counts	Risk Class
<20000 cells	0 - Minimum risk
20000-50000	1 - Low risk
50000-90000	2 - Medium risk
>90000	3 - High risk

The dataset was divided into training set and testing set using the k-fold cross validation approach (Kohavi 1995). In k-fold, the dataset is broken randomly into k parts, where each time the kth part is used for testing and the rest of the dataset is used for training. Consequently, this means that by using the k-fold approach the model could be tested in k different testing datasets. In our case, the dataset was split into 20 folds. This is because the available dataset is not a large one, covers just a 9-month period and, therefore, if a larger testing set had been used, the training set would have been very small. However, the model was tested only randomly only in 4 of the of the overall 20 folds, the 5th, the 8th, the 12th and the 15th folds.

Once the training and test dataset were created and before training the model, both input and output data (only the regression problem) were standardized (scaled to have mean 0 and standard deviation 1) using the following equation:

$$X_n = \frac{X_o - X_{mean}}{S}$$

where X_n is the normalized data, X_o is the measured (observed) by the sensor data, X_{mean} is the mean of the training set and S is the standard deviation of the training dataset. The variables that follow the time lag presented in section 9.2 are given in the model as inputs and outputs as shown in the following figure.

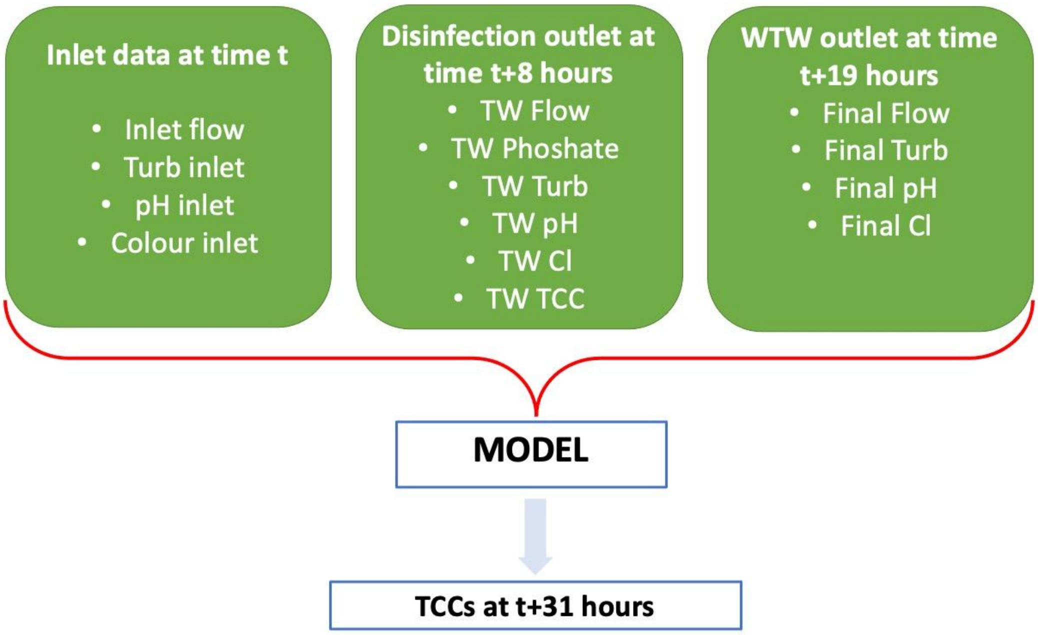


Figure 9.7: Simplified diagram of inputs and outputs of the TCCs predictive model

9.4.2. Machine learning models

The predictive model was created in MATLAB version 2019b using the Statistics and Machine Learning toolbox and the Deep learning toolbox (all codes are stored in the GitHub). The ML algorithms used in the model were as follows:

1. Random forest (RF): The random forest algorithm (Breiman 2001) was used for both the classification and the regression approaches. The number of the weak trees in this algorithm was set equal to 1000, the minimum of the randomly selected parameters

used in each randomly selected subset was set equal to 4 and the minimum tree leaf was set equal to 2.

2. Feed-forward ANN (ANN): The ANN algorithm (Bishop 2006) was used for the regression approach only. The hidden layer's size was set equal to 10 units and the Bayesian regularization backpropagation was used as a training function to update the weights and the bias values.

3. RusBoost boosting trees (RB): The RB algorithm (Seiffert et al. 2008) was used for the classification approach only. As in RF, the weak trees were set equal to 1000, the maximum number of splits per tree was set equal to the number of the training samples and the learning rate was set equal to 0.1.

4. Long short-term memory (LSTM): The LSTM algorithm (Hochreiter and Schmidhuber 1997) was used in both the classification and the regression approaches. For the LSTM algorithm there were several hyperparameters that required to be set, such as the total number of hidden LSTM layers, the number of units per LSTM layer, the initial learning rate, the learning drop factor, the number of epochs and the minibatch size. The number of epochs, the learning drop factor and the minibatch size were set equal to 150, 0.1 and 20 respectively. As regards the other 3 hyperparameters, a Bayesian optimisation was applied to find the best value that reduces the prediction error setting the following ranges:
LSTM layers: 1 to 5 layers / LSTM units: 15 to 150 units / Initial learning rate: 0.01 to 1
The overall architecture of the LSTM algorithm is shown in figure 9.8

5. Combined model (CM): The combined model was simply averaging the prediction outputs in both regression and classification approaches. In regression, for each prediction output, CM was calculating the mean average of the three models used (RF, ANN, LSTM). In classification, each one of the classification models (RF, RB and LSTM) contributed with one vote in the final decision and the CM model produced the final output based on the most popular class selection. In case where each one of the models produced a different output, CM was selecting as output the highest (more risky) of the available 3 classes. So, for example, if RF had predicted that the water belonged in the low-risk class, RB had predicted that the water belonged in the minimum risk class and LSTM had predicted that the water belonged in the high - risk class, the CM classified the water in the high-risk class as well.

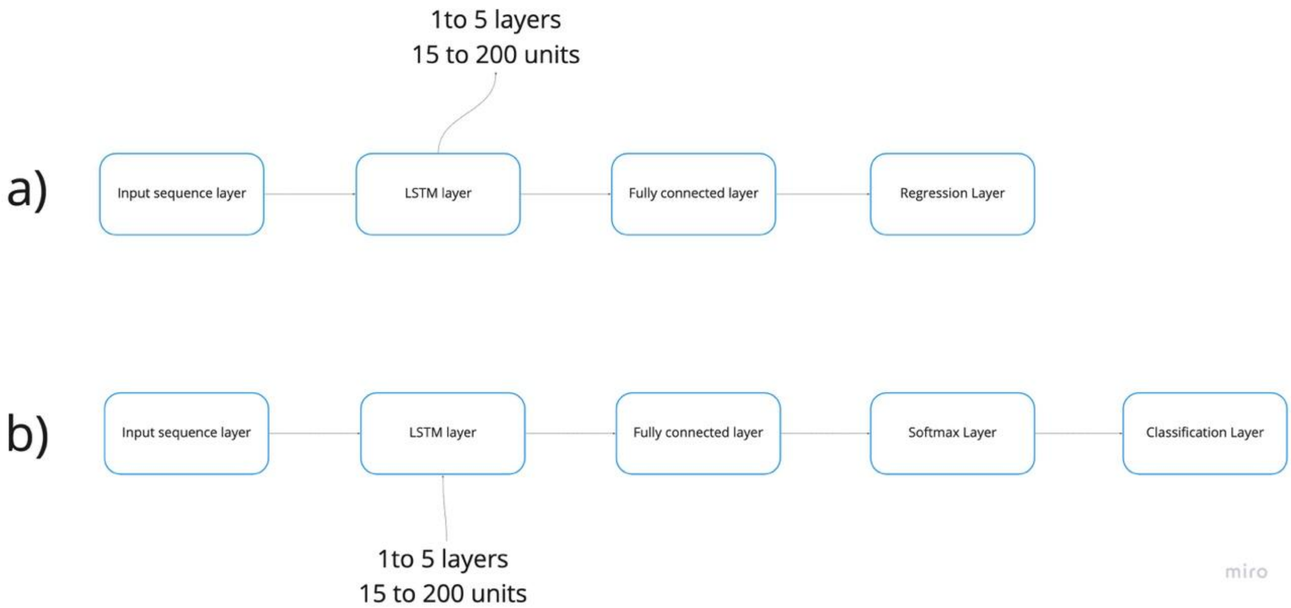


Figure 9.8: LSTM architecture for a) regression approach and b) classification approach

9.4.3. Performance metrics

In regression, the performance metrics used to evaluate the predictive models were the mean squared error (MSE), the root mean squared error (RMSE) of the normalised outputs, meaning that both had no dimensions, the normalised mean squared error (NMSE) of the actual predicted outputs, and the coefficient of determination (R^2).

In classification, 5 metrics were used, the overall accuracy, the TPR of the high-risk class, the macro-recall, the macro-precision, and the macro-F1 score. The macro-recall, the macro-precision, and the macro-F1 score are the arithmetic mean per class of the recall, precision and F1 score respectively.

9.4.4. Results

The predictive model, initially, was trained using one of the ML algorithms and all the available parameters. Then it was trained using the 8 most important parameters of the RF model - 8 parameters, Inlet flow, TurbInlet, Colourinlet, TW_TCCs, TW Cl, Final Turb, FinalCl and Final Alum as shown in figure 9.9. Finally, the predictive model was trained using the 5 parameters that in SOMs analysis appear to be the ones that correlate the most with high TCCs in the WTW's outlet. More specifically, the selected parameters were Inletflow, TurbInlet, Colourinlet, TW_TCCs, TW Cl. Outlet flow and TWFlow were excluded from the latter group of parameters, despite being highly correlated with final TCC, because of their similarities with the Inletflow. In the last case, as no outlet parameters were used the prediction horizon was increased from 12 hours to 23 hours (SR_RT= 11 hours + the 12 hours predictive horizon).

Overall, 48 regression and 48 classification models were produced. Each different model was named based on the algorithm that was used for the predictions (RF, ANN, LSTM or CM for regression and RF, RB, LSTM or CM for classification), the number of parameters used (All, RF8 for the 8 RF parameters, and SOMs for the 5 SOMs analysis parameters) and the predictive horizon (12 or 23 hours). So, for example when the RF with all the parameters was used, the model was named as RF-All-12.

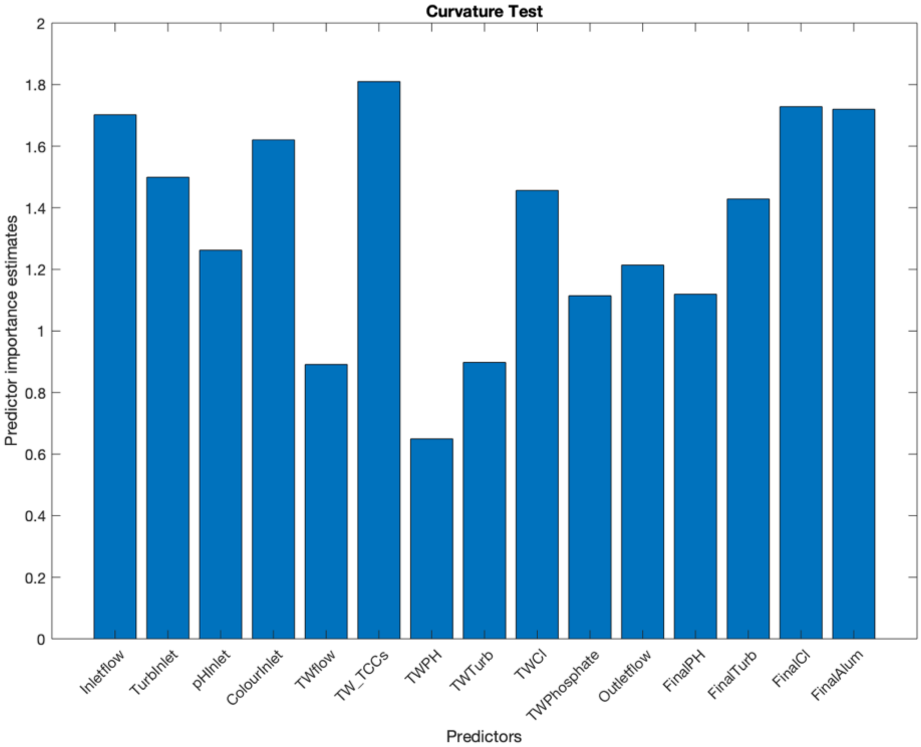


Figure 9.9: Performance importance of each parameter of the RF-All-12 model in the 5th fold (see appendix B for performance importance in folds 8, 12 and 15)

9.4.4.1. Regression results

The performance metrics for each model are presented in table 9-4. It is notable that some of the ML models have produced satisfactory results in predicting the TCCs exiting the WTW.

RF appears to be the best ML method for this WQ problem as the RF models' performance metrics were ranged between 0.08-0.14 and 0.3-0.37 for MSE and RMSE respectively. In addition, the RF models' performance was not influenced by both the parameters reduction and the increase in the predictive horizon (RF - RF8 - 12 & RF - SOMs - 12 models). Moreover, the RF models explained up to 91% of the variance of the model (R^2=87-91%).

From the other two ML methods, LSTM performed better when all the parameters are used comparing to ANN (R^2=0.84-0.86 /R^2=0.82-0.85 respectively, MSE=0.14-0.17/ MSE=0.15-0.18 respectively, RMSE=0.37-0.41/ RMSE=0.38-0.42 respectively). However, ANN performed slightly better than LSTM when the 8 RF parameters or the SOMs are used as input data (R^2=0.79-0.84 /R^2=0.78-0.84 respectively).

Table 9-4: Summary of the regression models' performance metrics

Model NAME	R2	MSEn	RMSEn	NMSE
RF - All - 12	0.89-0.91	0.09 - 0.12	0.3 - 0.35	0.09 - 0.12
ANN - All - 12	0.82 - 0.85	0.15 - 0.18	0.38 - 0.42	0.15 - 0.18
LSTM - All - 12 *	0.84 - 0.86	0.14 - 0.17	0.37 - 0.41	0.14 - 0.17
CM - All - 12 *	0.88 - 0.9	0.1 - 0.13	0.32 - 0.36	0.14 - 0.17
RF - RF8 - 12	0.90 - 0.91	0.08 - 0.11	0.29 - 0.34	0.08 - 0.11
ANN - RF8 - 12	0.79 - 0.84	0.17 - 0.21	0.41 - 0.46	0.16 - 0.21
LSTM - RF8 - 12 **	0.78 - 0.84	0.16 - 0.22	0.4 - 0.45	0.16 - 0.23
CM - RF8 - 12	0.87 - 0.88	0.11 - 0.13	0.34 - 0.36	0.14 - 0.19
RF - SOMs - 23	0.87 - 0.89	0.11 - 0.14	0.34 - 0.37	0.11 - 0.14
ANN - SOMs - 23	0.76 - 0.82	0.17 - 0.21	0.41 - 0.46	0.16 - 0.21
LSTM - SOMs - 23^	0.8 - 0.82	0.19 - 0.2	0.43 - 0.45	0.19 - 0.2
CM - SOMs - 23	0.86 - 0.88	0.12 - 0.15	0.35 - 0.39	0.15 - 0.22

* LSTM - All - 12: 1 LSTM layers,25 units per layer,0.001 Initial learning rate
** LSTM - RF8 - 12: 1 LSTM layers,16 units per layer,0.001 Initial learning rate

^ LSTM - SOMs - 23: 1 LSTM layer,25 units per layer,0.0012 Initial learning rate

MSE and RMSE are calculated for the normalized outputs and therefore they are unitless

The CB model's results indicated, as expected for the regression model, that CBs will not increase the overall accuracy as it is dependent on the accuracy of the single models. However, according to table 9-4, the CB models' performance reduced the bias and the variance of each unique ML model and produced the second most accurate results explaining up to 90% of the model's variance. Finally, table 9-4 justifies that the models' prediction was better when all the parameters were used compared to the 8 RF parameters or the 5 SOMs parameters. The SOMs models' performance, however, should also be considered as satisfactory since their predictive horizon was extended to 11 hours more than the other two model groups - overall predictive horizon of 23 hours.

In figure 9.10, three plots that show the 15th fold test TCCs vs the predicted by the models TCCs using the different groups of input parameters are presented. From this figure it is clear that all the models were able to understand the TCCs increase and decrease trends. However, this graph also shows that none of these models , in all three cases, was able to predict the extreme TCCs values in the time-series. The worst of all the models, as the graphs indicate, is LSTM as it cannot capture any of the extreme events. This is a finding that the performance metrics did not capture and indicates the importance of the graphical representation of the results for evaluating the models' performance.

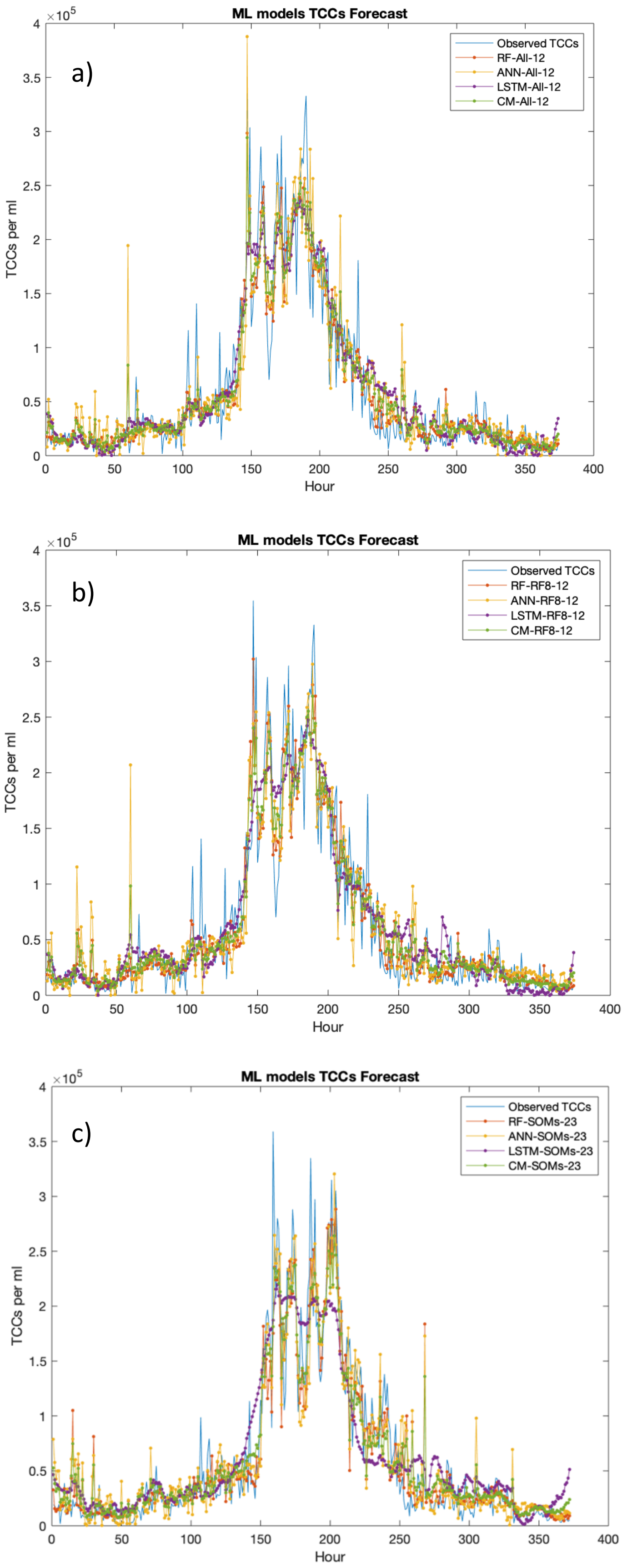


Figure 9.10: 15th fold Observed TCCs vs predicted time series of all the models when a) all the water parameters were used b) the RF 8 parameters are used and c) the SOMs 5 parameters are used (see appendix B for performance importance in folds 5, 8 and Classification results

The classification results are presented in table 9-5. As in the regression approach, the RF models performed better than the other two models according to the performance indices reaching up to 82% total accuracy and 97% recall of the high-risk class when all the parameters were used. The RB models also produced accurate results when all or the 8 RF parameters were used (Accuracy 75-78%, high risk recall (87%-92%). However, RB model performance was decreased when the SOMs parameters were used and, thus, the predictive horizon was increased. LSTM models had the worst performance comparing to the other two models as all the macro - metrics, in all three different cases, were below 70%. The SOMs LSTM model managed to reach a 99% high risk category recall, but both macro-recall, micro precision and macro-F1 metrics for this model were poor (57%-61%, 54%-64%, 0.54 - 0.57 respectively). This finding indicates that LSTM-SOMs-23 has created a high number of false high-risk class positives and therefore, this model should be considered as unreliable. The CB models for the classification approach produced worse results comparing to the regression approach. This performance could be explained by the fact that CB models was not able to reduce the bias and the variance of the bad LSTM models' outputs. Finally, the performance metrics indicate that the best performance is achieved when all the parameters are used, however, as in the regression approach, when only the SOMs' parameters are used, the results are also reliable (except for the LSTM - SOM - 23 model). This finding indicates that it is possible to have an accurate prediction of the bacteriological quality of the water exiting the WTWs 23 hour ahead, which is a sufficient time for the WTWs' operators to act if it is necessary.

Table 9-5: Summary of the classification models' performance metrics

Model NAME	Accuracy	High risk Recall	Macro - recall	Macro - Precision	Macro - F1
RF - All - 12	80 - 82%	90 -97%	68 -72%	72 - 78%	0.7 - 0.74
RB - All - 12	75 - 78%	87 - 92%	68 - 75%	68 - 73%	0.68 - 0.74
LSTM - All - 12 *	71 - 73%	76 - 90%	56 - 65%	61 - 68%	0.61 - 0.69
CM - All - 12 *	77 - 80%	78 - 91%	68 - 75%	69 - 74%	0.68 - 0.76
RF - RF8 - 12	78 - 81%	92 - 93%	69 - 72%	72 - 75%	0.71 - 0.72
RB - RF8 - 12	75 - 78%	88 - 92%	69 - 75%	68 - 73%	0.68 - 0.72
LSTM - RF8 - 12 **	66 - 71%	80 - 93%	57 - 62%	60 - 70%	0.61 - 0.67
CM - RF8 - 12	72 - 80%	79 - 91%	69 - 77%	69 - 75%	0.69 - 0.75
RF - SOMs - 23	75 -77%	90 - 94%	66 - 91%	68 - 75%	0.66 - 0.74
RB - SOMs - 23	70 - 75%	87 - 94%	66 - 72%	64 - 70%	0.64 - 0.72
LSTM - SOMs - 23^	68 - 71%	85 - 99%	57 - 61%	54 - 64%	0.54 - 0.57
CM - SOMs - 23	74 - 76%	87 - 91%	67 - 72%	66 - 70%	0.66 - 0.71

* LSTM - All - 12: 1 LSTM layers,25 units per layer,0.001 Initial learning rate

** LSTM - RF8 - 12: 1 LSTM layers,16 units per layer,0.001 Initial learning rate

^ LSTM - SOMs - 23: 1 LSTM layer,23 units per layer,0.0012 Initial learning rate

9.5. Discussion and operational value of this work

Understanding the factors that increase the bacteriological activity in the water exiting the WTW could be beneficial for the WUs as it could help them adapt the treatment processes to control these factors and decrease the risk of bacteriological contamination of the water that they serve to their customers. This work demonstrates that with the simple use of methods, such as SOMs or PCA, WUs could identify correlations between bacteriological parameters and other WQ parameters using their own data collected from their SCADAs. Once these factors are identified, the plant operators could organise actions to control them and, thus, reduce the number of bacteriological cells in the WTW's outlet. For example, in this work, both PCA and SOMs show that in Balmore WTW there is a clear correlation between high turbidity and flow in the inlet. This means that the operators could be notified when high turbidity and high flow in the inlet is measured, and, thus, prepare the treatment stages to adjust to these increases and reduce the deterioration risk in the outlet.

There were a lot of similarities in PCA and SOMs findings as mentioned in the results. However, SOMs indicated some other correlations between high TCCs and certain parameters such as colour in the inlet, low aluminium in the outlet, low pH in the outlet and turbidity exiting the disinfection tank. This is because SOMs, as an ANN, is able to uncover both the linear and the non-linear correlations in comparison to PCA that can describe only the linear similarities (Speight, Mounce, and Boxall 2019). This finding demonstrates, one more time in this thesis, the complexity of the relationships between the various parameters that influence the water quality in the DWDS. This work, though, recommends the use of both methods by the WUs as by doing so, it helps to separate the linear relationships from the non-linear ones and, thus, understand better the water quality in their systems.

The predictive model's results demonstrated that WUs could be benefited by using the data that they already collect in their own WTWs as inputs in easy to train models and improve their final product. Their decision in which of the two different prediction approaches is more appropriate for their operations, is up to the WUs and their requirements. The regression model could be used as a tool for understanding the direction of the TCCs over a certain predictive period. The classification model could be used as a tool by the WUs if they are only interested in predicting the bacteriological risk of the water exiting their plants.

RF is a well-known method in the water sector and has been applied in various projects (Parkhurst et al. 2005; Mohammed, Hameed, and Seidu 2017; Meyers, Kapelan, and Keedwell 2017). The results of this work indicate that for this dataset, it was the best model for predicting TCCs in the outlet. This is the first case in this thesis where this model was proven to be the best one to select. This is probably because the available dataset here is more balanced compared to the previous two case studies. In agreement with other works where RF was applied (Meyers, Kapelan, and Keedwell 2017; Mohammed, Hameed, and Seidu 2017), it is recommended the application of RF in most of the WQ problems as one of the methods for investigation, especially if the datasets are well distributed. RF is a simple ML model to apply, and its results could be justified and explained to decision makers due to its “white-box” approach.

The fact that LSTM was the worst performing model was not an unexpected output. In the recent year and in various works in the hydroinformatics field, deep learning approaches have been proven to be the models that could produce the most accurate results (Hitokoto and Sakuraba 2018; Dairi et al. 2019; Zhou et al. 2019; Mamandipoor et al. 2020). However, the available datasets in these studies were extremely large in comparison to the available dataset in our study, which is probably the reason that LSTM had that bad performance. In addition, LSTM required the most computational time during the training period. This finding though does not, necessarily, indicate that LSTM is not a good method for WQ problems. Deep learning approaches, in contrast to the traditional ML approaches, are learning directly from the examples and their multiple levels of representations over their consecutive layers (Lecun, Bengio, and Hinton 2015). Hence, a further investigation over the LSTM prediction ability is required in case studies with larger datasets or in the following years when sufficient amount of online TCC data will be available.

The RusBoost algorithm that the RB models applied in the classification approach (also used in chapter 7 is a method that is commonly used in classification problems with unbalanced datasets (Seiffert et al. 2008, 2010; S. R. Mounce et al. 2017). As mentioned above, in this case study, the dataset was more balanced with only the medium risk class (class 2) having fewer samples compared to the other three classes. These models increased the recall accuracy of this class that, consequently, increased the overall macro-recall accuracy as table 9-5 shows. However, their lower performance in both macro-precision and macro-F1 score compared to the RF models, indicates that RBs also produced more false positives than the RF ones. This finding clearly demonstrates the importance of taking an overall consideration of 2-3 performance metrics to decide which model is the best. WUs decision makers should always consider what is the main purpose of the modelling prediction and how much compromise in false positives they can tolerate in their proactive water quality interventions.

The CB models produced good regression results but bad classification results. This is because, as mentioned in the results section, the CB models are dependent on the accuracy of each one of the ML models. Overall, CB models reduce bias and variance of the single ML techniques and therefore, as we also saw in chapter 7, there is a reason to apply them in this type of problems. However, the classification results in this chapter clearly demonstrate that when the single ML models produce inaccurate results, the CB model cannot sufficiently improve them.

This chapter highlights the importance of online microbiological monitoring. The traditional bacteriological monitoring approaches that follow the regulations (DWQR 2014) require a daily sample to measure the 4 bacteriological parameters. However, with this approach the variations of bacteria during the day cannot be captured. In addition, by measuring once per day randomly for coliform bacteria, it is pure luck to find coliforms in the outlet. The online monitoring system provided valuable information regarding the TCC numbers and the parameters that influence the bacteriological increase in the water entering the DWDS. Moreover, online monitoring can capture sudden changes in the water quality and, finally as we demonstrated in this chapter, they could provide data for data-driven models for the prediction of potential future deterioration events.

Figure 9.10 and the figures in the appendix indicate that in the regression approach, all the used ML methods struggle to predict the extreme values. This is probably happening because the available dataset is not sufficient and does not capture all seasonal changes. In addition, there are only a few measurements in the dataset where extreme TCC values have been found. Therefore, the model was not trained properly with enough extreme data to be able to predict some as well. Future work, when more data will be available, should focus on extreme events by using them as an extra input variable for the training period. The classification approach, on the other side, managed to predict the water being in the high-risk class (TCCs>90000) with an accuracy between 85% and 97%. This finding indicates that this model is able to understand when extreme TCC numbers will occur. By using this approach, WUs could get a 12-hour ahead indication of a high bacteriological risk water entering their networks and act promptly.

The predictive model is a data-driven approach that does not require any hydraulic and process model for its implementation. It only uses the data measurements as captured by the sensors and stored in the SCADA of the WTW. Thus, the time that is required for its training is minimal in comparison to process based models that require lots of hours for simulation and high computational power. This is because the data-driven models learn the trends and the patterns of the dataset, in contrast to the process-based models that are using complex hydraulic and process equations to describe the water circulation and treatment process in the WTW. Moreover, the process-based models demand many process and hydraulic parameters that use a large amount of data for calibration. Furthermore, these parameters need recalibration with new data to fit the spatial and temporal changes of the system into the model. Finally, the data-driven model presented here, once built, could be used in any other WTW that has a SCADA system that contains enough water quality parameters data. The process-based models, though, cannot be transferred in other systems as each system has each unique process and hydraulic characteristics.

This work contributes to the general discussion over the use of ML methods in the WTWs. It provides a new methodology that combines unsupervised and supervised ML methods and uses only water quality and flow data for both understanding the factors that influence the bacteriological increase in the WTWs and predicts the future bacteriological behaviour which is the first time that has ever done. The successful implementation of the methodology in the Balmore WTW indicates that these methods should contribute to the existing WTWs managing approach that is manly reactive and transforming it into a proactive one.

9.6. Conclusions

In this chapter a data-driven investigation over the bacteriological activity of the water exiting Balmore WTW is presented. More specifically, this investigation aimed, firstly, to identify the factors that are related with high TCC numbers exiting the Balmore's outlet, using SOMs and PCA ML methods, and, secondly, to explore the potential of a model based on ML methods to predict the TCCs behaviour up to 23 hours ahead. For this investigation, water quality, flow and TCCs data were taken from the SCADA system of the plant. From all the available monitored parameters in the WTW, the parameters measured in the inlet, the disinfection tank outlet and the WTW's outlet were used. The key findings of this investigation are:

- Both PCA and SOMs outputs indicated that the main factors that influence the high TCC numbers in WTW's outlet are the inlet flow, the inlet turbidity and the low chlorine residual exiting the disinfection tank
- SOM's also indicated that there is a high correlation between inlet colour and TCCs, an inverse correlation between TCCs, aluminium in the outlet and pH in the outlet and a weak correlation between the disinfection tank pH and the TCCs
- The additional correlations, captured by the more complex SOM model, indicate the complexity of the processes inside the WTWs that are not always clear
- The regression predictive model managed to capture the TCC trends and the general bacteriological behaviour 12 hour ahead. However, it was not able to predict the highest observed peaks probably because the available dataset was not sufficient
- The classification model captured the extreme events and classified the water that belong in the high bacteriological risk class with an accuracy of up to 97%
- RF has been proven to be the best ML method for both the regression and the classification model
- LSTM had the worst performance out of all the ML methodologies. However, this finding was expected as these type of deep learning approaches require way larger datasets than the available in this investigation

Overall, the outputs of this work demonstrated the benefits that WUs could gain by using the WTWs' data, they already collect, into these data-driven models. SOMs and PCA could be a great tool for decision makers and process engineers to understand the general plant behaviour over a certain period and control specific parameters that increase the bacteriological activity. The predictive model could be developed in an online tool, connected to the SCADA system, and give an early warning to the operators for a potential bacteriological risk in the water exiting the plant few hours ahead. Thus, it will give the necessary time to adjust the processes and prevent deterioration.

10. A Big-Data framework for actionable information to manage drinking water quality

Reproduced from Kyritsakas G., Boxall J.B., Speight V.L (2021). A Big-Data Framework for Actionable Information to Manage Drinking Water quality. Environmental Science Water Research & Technology. Themed issue: Data-intensive water systems management and operation (under review)

Declaration

Chapter 10 is an under peer-review journal paper submitted for publishing in a special edition of the Environmental Science Water Research & Technology journal. The contribution of the first author and the co-authors are the followings:

1. **Grigorios Kyritsakas** is the PhD candidate and first author of the chapter. He developed the main idea of the framework, collected the data, created the datasets, developed the methodologies, and analysed the results in the case studies presented in this chapter. In addition, he wrote the first draft of the chapter.

2. **Prof. Joby Boxall & Prof. Vanessa Speight** are Grigorios Kyritsakas' academic supervisors and co-authors of this chapter. They supervised the development of the framework and provided critical input into the research methodology. In addition, they provided important feedback in the formulation of the abstract, the discussion points and the conclusion. Finally, they contributed to the refinement of the chapter by changing the structure of the draft and by proposing changes in the figures.

		Vanessa Speight.
Mr Grigorios Kyritsakas	Prof. Joby Boxall	Prof. Vanessa Speight
1st author	2nd author	3rd author

Note

The thought for a big data framework for WUs occurred during the period that datasets were generated. Initially, it was considered that the machine learning application steps presented in chapter 3 would be sufficient for the data analysis with ML techniques. However, the increased amount of time that was consumed for collecting the raw data from various sources in different formats and the effort for connecting this and creating a complete dataset indicated that for the fast application of the data-driven techniques, the need of facilitating the data storage and data integration is necessary. This paper reiterates some sections that are developed in the previous chapters of this thesis. More specifically, the big-data framework utilises the ML application steps as developed in chapter 3 as part of the data analysis layer. In addition, a small part of the research conducted for the case studies, presented in chapters 5 and 7, is used in form of two examples that demonstrate the successful application of the framework.

10.1. Abstract

Water companies collect vast amounts of data, but it is stored and utilised in silos. Machine learning techniques offer the potential to gain deeper insight from such data. We set out a Big Data framework that for the first time enables a structured approach to systematically progress through data storage, integration, analysis, and visualisation, with applications shown for drinking water quality. A novel process for selection of the appropriate method, driven by the insight required and the available data, is presented. Case studies for a water utility supplying 5.5 million people validate the framework and provide examples of the actionable information that can be obtained to help ensure the delivery of safe drinking water.

10.2. Introduction

Water utilities have a duty to provide drinking water that complies with the high quality standards set by national and international regulations. This effort requires a multi-step approach that includes the application of efficient and innovative treatment technologies in the water treatment works (WTWs) as well as the prevention of water deterioration during its travel through the drinking water distribution system (DWDS), through the proper maintenance of DWDS and monitoring of the treated water quality from source to tap.

In the UK, water utilities monitor treatment processes using sensors to measure various water quality parameters in a regular frequency (typically every 5-15 minutes), with resulting data, also known as telemetry data, stored in the supervisory control and data acquisition (SCADA) or similar system. In addition, utilities take samples from different points across their DWDS, including exit points from the WTW and service reservoirs (SRs) and randomly selected consumers’ taps. The water quality parameters measured in a typical DWDS monitoring programme are microbial indicators, disinfectant residual, iron, manganese and turbidity as defined by the regulators (DWI 2016; DWQR 2019a).

The typical procedure for DWDS monitoring results at UK water utilities is to archive all the data, once checked for compliance, thereby creating a large store of data in various formats. This archived data is generally not used or analysed further, with its volume increasing year after year. Analysis of these datasets, if done correctly, and when considered with wider asset, operational or even third-party data, can provide a better understanding of the complex processes that occur inside the ageing DWDS and can be used as evidence to direct capital, operational and maintenance activities within a water utility. Advanced data analytics, including tools that broadly fall under the umbrella of artificial intelligence (AI), offer the opportunity to unlock the potential value of otherwise ignored DWDS water quality data.

A few research studies have applied AI technologies, such as data mining (DM) or machine learning (ML), to drinking water quality problems for understanding factors contributing to water quality deterioration (Speight, Mounce, and Boxall 2019; E. J. Blokker et al. 2016) predicting future deterioration events (Kazemi et al. 2018; Meyers, Kapelan, and Keedwell 2017; S. R. Mounce et al. 2017) and optimising treatment processes at WTWs (Li et al. 2021). These studies demonstrate the potential that individual ML techniques could have for analysis of historical water quality data. While the focus of research is often the specific ML technique, the individual ML techniques are just one component of the 'Big Data' analytics approach required to support decision making and inform investment choices. Water utilities who want to benefit from Big Data analytics will also need to transform the ways that they collect, store, process, and visualise data and results. This transformation, sometimes referred to as the pathway to "Digital Water" (IWA 2019) requires holistic consideration of data issues to facilitate the big data applications for improving the delivery of safe drinking water.

This paper proposes a Big Data framework for water utilities, using examples drawn from DWDS water quality applications to demonstrate its application (Speight, Mounce, and Boxall 2019; M. Blokker, Vreeburg, and Speight 2014). This framework fills the gap between different, individual data-driven applications and the integration and processing of the various types of raw data collected by water utilities. This framework is meant to be a guiding approach for water utilities with specific examples related to solving water quality problems in DWDS. By presenting this framework, this work aims to contribute to the "Digital Water" transformation and to lay out the steps for water utilities to undertake on the journey to this digital revolution.

10.2.1. Big Data analytics

Big Data refers to the collection of massive amounts of data that modern digital technologies generate and/or store. The volume of data, however, is just one of the characteristics of Big Data that also includes the velocity, the variety, the veracity, the variability and the value (Gandomi and Haider 2015). Briefly, these six characteristics respectively refer to generation and collections of extreme amounts of data, the speed that the data are generated and analysed, the complexity of the datasets as these could be composed with data in various types of formats, the reliability (in terms of quality) of the available data, the variation of data sources and data flows and the important information that could provide once analysis. Big Data analytics is the science that includes all the processes and the tools required to uncover valuable information hidden in these massive datasets, from the data collection to the mining and the predictive methods (usually DM and ML techniques and algorithms) used for providing outputs to decision makers. This science is evolving rapidly, with new methods and applications contributing to the derivation of new knowledge from data in various scientific domains including medicine (Okada 2021)[11], agriculture (Geetha, Deepalakshmi, and Pande 2019) , environmental protection (Dimokas et al. 2020; Lu 2020), energy efficiency (Zekić-Sušac, Mitrović, and Has 2020), and finance (West and Bhattacharya 2016). Comparing across these works, the different applications pose unique challenges and while there are common features, there are critical differences. Undeniably, successful application of big data science requires collaboration and integration with domain expertise.

10.2.2. Machine Learning

Machine learning is the area of AI that develops the algorithms used to optimize future performance or understand patterns by learning from existing data or past experiences (Alpaydin 2014). ML algorithms are the most common tools that Big Data analytics use for identification of patterns in data and predictions of future trends. There are two main categories of ML algorithms: supervised and unsupervised. Supervised learning algorithms are trained on data that has been labelled as input or output, therefore requiring some specification by the user. Once the training is finished, these algorithms then predict future outputs based on new unseen inputs. Depending on the application, supervised ML predictive modelling can produce outputs of a numerical value or a classification. For classification, during the training period for the given inputs, a specific category or class is specified as the output (e.g. above or below a threshold). Once trained, classification ML algorithms then predict the output category for the new unseen input data. For prediction of numerical values, the ML algorithms typically use regression techniques to develop numerical relationships that can predict future output values when given new unseen inputs. Unsupervised learning algorithms do not have specified inputs and outputs but rather use unlabelled data as inputs to generate clusters of different groups, uncovering hidden structures in the datasets and identifying correlations between the various parameters of the analysis. These types of algorithms are typically used for data exploration rather than prediction.

Machine learning is gaining traction in water related applications, with recent studies having developed and applied algorithms for topics including leak detection in pipes (S. R. Mounce, Boxall, and Machell 2010; Romano, Kapelan, and Savić 2014; Carreño-Alvarado et al. 2017), water demand forecasting (Herrera et al. 2010; Xenochristou et al. 2021), wastewater treatment plant operations (Dairi et al. 2019; Mamandipoor et al. 2020), sewer overflow predictions (S. R. Mounce et al. 2014; Rosin et al. 2018), prediction of chlorine decay at consumers taps (Gibbs et al. 2006), prediction of indicator microorganisms in drinking water supply(Mohammed, Hameed, and Seidu 2017), and prediction of water quality events in DWDS using sensors (Vries et al. 2016; Fellini et al. 2018; Garcia, Puig, and Quevedo 2020). The aforementioned studies generally cover a single application but collectively demonstrate the potential for ML techniques to provide value to water utility operation. However, Speight et al. (2019) and Mounce et al. (2017) report challenges in applying ML techniques in the collection of the data and the processes required to construct a dataset suitable for analysis. The need for a well-founded question, or a more exact articulation of the insight sought, to ensure the Big Data exercise is well directed and leads to consequential new understanding is evident when exploring and differentiating past research. These observations reinforce the need to include consideration of the insight sought along with data collection, storage, and organisation; a Big Data framework that encompasses these and guides the selection of the ML algorithm is essential to create lasting value for water utility applications.

10.2.3. Big Data Analytics Frameworks

The complexities in Big Data applications vary from one organisation to another, depending on the type of collected data and the knowledge that needs to be derived from the datasets. In many scientific domains, discussions over holistic structures, also known as Big Data analytics frameworks, have begun to emerge. Chandarana and Vijayalakshmi (Chandarana and Vijayalakshmi 2014) documented the challenges that organisations face using the different types of data that they collect and the requirements for development of frameworks to organise and analyse this data. As part of this work, the authors emphasised the importance of big-data analysis for deriving valuable information and making better decisions, and gave example areas such as healthcare and intelligence where big-data frameworks could be beneficial.

Most frameworks proposed in the literature comprise a series of rules, in the form of layers, to 1) address the specific data storage and data integration complexities; 2) apply the proper ML, DM, or other data analysis methods depending on the desired outputs; and 3) visualise the outputs (Zekić-Sušac, Mitrović, and Has 2020; Abdullah et al. 2018; Ahmed et al. 2021). For example, Osman (2019) proposed a 3-layer framework for smart cities applications that includes the platform layer which specifies the operating systems and communication protocols for collecting the various types of data, the security layer which specifies the protocols for controlling access to the data and the protocols for data integration, and the data processing layer which specifies the data pre-processing, data analytics and management of the analytics model. The author also included a discussion of principles required for successful implementation, including the integration of static and real time data as well as standardisation of data acquisition. Villanueva Zacarias et al. (2018) suggested a 4-layer framework for the manufacturing sector that includes data storage and integration, consideration of the available IT resources for data pre-processing and analysis, selection of the appropriate ML algorithms for the data analytics, and a dashboard for the visualisation of the different solutions.

Within the water sector, there were two published studies found that discuss Big Data analytics frameworks. One examined the benefits that water utilities could gain in reduction of chemicals in their wastewater using Big Data tools and ML techniques within their datasets (Romero, Hallett, and Jude 2017). This study described the current situation in the water sector, referred to applicable ML tools, and provided two different examples where the incorporation of Big Data tools could strengthen the existing approaches. However, the authors did not develop a specific Big Data framework or ML technique selection process. The second study proposed a 5-layer framework for improving urban domestic wastewater treatment and reducing environmental pollution, consisting of a data perception layer, data transmission layer, data storage layer, data analysis and application layer, and user interface layer (Du, Kuang, and Yang 2019). The authors analysed the volume and type of information that would be required for the application of such data-driven approaches and emphasised the importance of collecting all the necessary data from the wastewater treatment works and networks to support the Big Data framework implementation. The degree of data proposed would require a significant transformation of monitoring practices in wastewater networks compared to the typical level today and the application of the application-specific framework with a smaller dataset or with non-sensor data was not demonstrated. Neither of these two studies refer to selection of specific ML techniques and the criteria required for the application of those techniques.

This paper proposes a comprehensive Big Data framework that is driven from the insight sought, addresses data collection, storage, and management aspects, integrates ML technique selection, and includes visualisation and communication of outputs. Importantly, the criteria for ML selection are based upon the desired drinking water quality investigation and the existing data that is available for the analysis, and the integration of data science and water engineering is essential for this.

10.3. Proposed Big Data Framework

The data that water utilities collect cannot be compared with the amounts of data that IT companies collect every day. In addition, their collection and storage systems are obsolete, and the value of these data is not really explored. Thus, at present, water utilities data do not comply with the big-data definition. However, water utilities’ aim is to follow the digital revolution of the other sectors, a process that requires a complete change of the current setup. Recognising the need of a holistic approach for the management of water quality data in DWDS for data-driven applications, we propose a Big Data framework consisting of 4 layers: 1) data storage; 2) data connection and integration; 3) data analysis; and 4) presentation and communication of data analyses outcomes (Figure 10.1). Importantly, the involvement of different types of expertise for each layer is noted to emphasise that the development of a Big Data framework is not solely within the domain of computer scientists and IT specialists but rather requires collaboration across a number of water utility teams. In this section, the purpose and main principles required for the implementation of each layer is described based upon its application to water quality in DWDS.

10.3.1. The Layers of the Proposed Framework

10.3.1.1. Data Storage Layer

The data storage layer includes the storage of various types of databases through the use of data warehouse or cloud storage technology. The data storage software and system specifications will differ from one water utility to another, but the key capability for all such systems is the ability to store all available types of data including structured, unstructured and asset information. In addition, it is very important that the format of the stored datasets makes them easily accessible for current and future use and supports linkages across databases with the use of unique IDs for every asset or sample. Therefore, the main principle required for this layer is the standardization of data deposition, which sounds straightforward but is not trivial. An examination of the sufficiency of the data collected by water utilities to answer questions of relevance and the quality of the available data sources (accounting for errors in measurement, missing values, etc.) is not the focus of this study.

The data that water utilities collect may be grouped into two categories: the data regarding their assets that is static or changing infrequently over time, and time series data such as water quality and system operations or control data. Within the time series data, 3 further subcategories of data that is collected may be usefully defined: 1) telemetry data from permanent sensor installations at the WTW and key DWDS assets which is transmitted to a central data repository (typically SCADA), 2) discrete water quality (grab) samples from the WTWs and SRs outlets and sparsely from consumers' taps, and 3) other time series monitoring data from temporary sensors in the DWDS, installed for water quality investigations, research, and similar purposes, that is often not stored in the central data repository but rather in a separate application. In addition to internal data collected by the water utilities, Big Data applications may also require data from external sources, including parameters such as rainfall and air temperature.

Table 10-1 summarises the key parameters required to support analytics for different types of water utility data, with a focus on water quality. The specifics of the parameters will differ by water utility, but baseline information will be required for assets and samples.

To gain the greatest value from Big Data techniques, linkages between datasets including aspects of physical connectivity are important to be included. For example, for a given water quality sample from the DWDS, it is ideal to be able to identify the pipes, water treatment works, and other relevant infrastructure supplying the sample location.

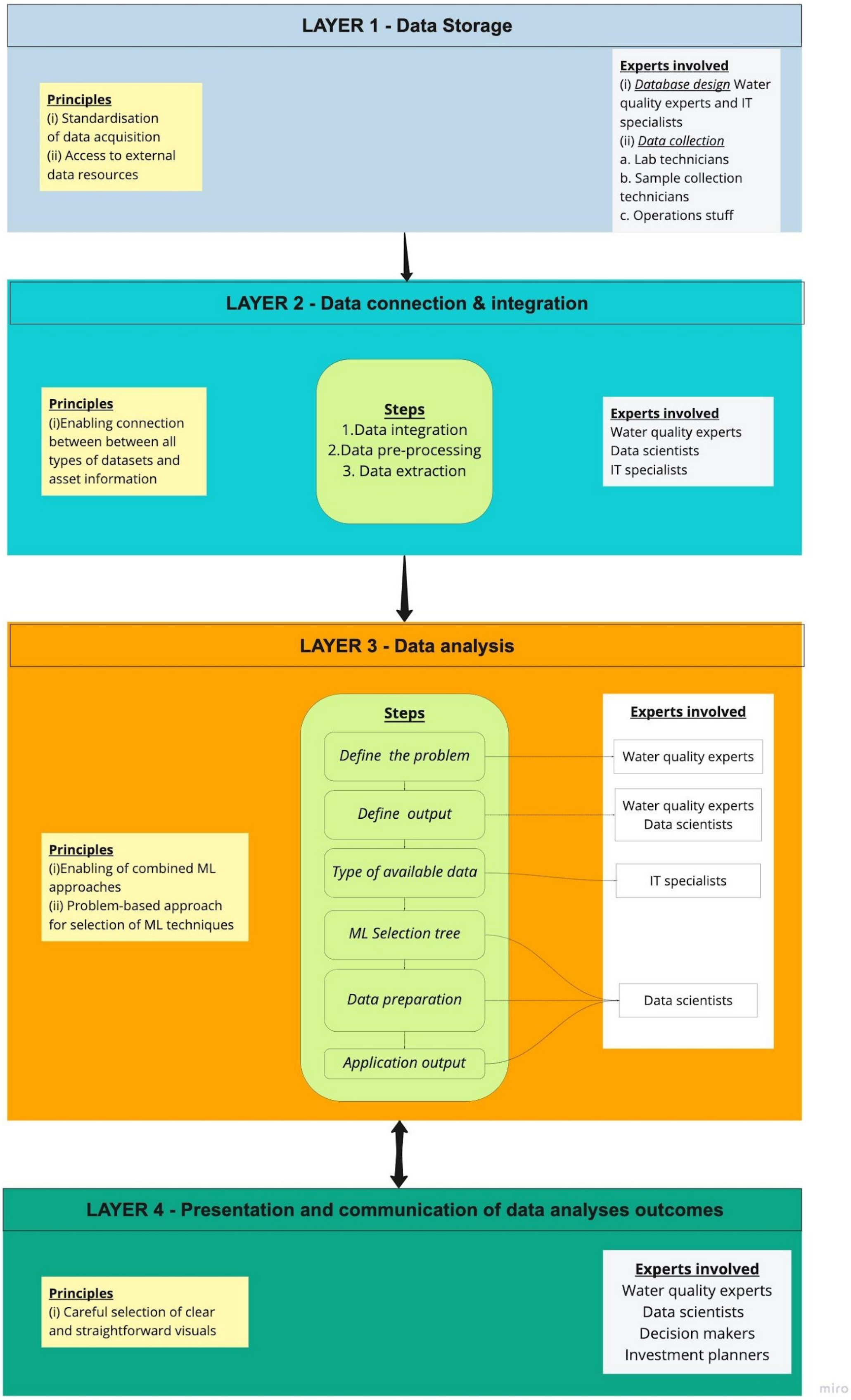


Figure 10.1: The proposed Big Data framework, labelled for data analysis applications in DWDS water quality

Table 10-1: Summary of typical data requirements by asset type and parameter category

Asset Type	Static Data Parameters	Time Series Parameters		
	Physical	Telemetry	Discrete Samples	Other Time Series
Water Treatment Works (WTW)	- Name/ID number, location - Volumes, dimensions, etc. - Source of raw water - Water operation areas (WOAs) served - Type of secondary disinfection - Treatment processes - Locations of monitoring equipment and discrete samples, for permanent and temporary monitoring - Date of any treatment process or secondary disinfection change - Date of other significant events, maintenance activities, etc. - Location of external monitoring (e.g. weather stations)	- All water quality parameters measured with online sensors in each treatment process and the final treated drinking water - Dosing data for key processes - Flows and levels for key process locations	- All water quality parameters measured by discrete sample in each treatment process and the final treated drinking water - Historical data on failures including dates of failures and investigation results - Reason for taking sample (e.g., regulatory, failure investigation)	- All water quality parameters measured with online sensors with dates and times of measurement - Reason for the investigation and results - External parameters e.g. temperature, rainfall
Service Reservoir (SR)	- Name/ID number, location - Volumes, dimensions, etc. volume, dimensions - Type of additional disinfection or chemical addition (if any) - Date of disinfection or other physical changes - Dates of cleaning and other maintenance activities - WTW supplying the SR -Water supply zones (WSZs) supplied by the SR - Location of external monitoring (e.g. weather stations)	- All water quality parameters measured with online sensors (e.g. turbidity, chlorine) - Flows and levels - Chemical dosing data, if applicable	- All water quality parameters measured by discrete sample - Historical data on failures including dates of failures and investigation results - Reason for taking sample (e.g., regulatory, failure investigation)	- All water quality parameters measured with online sensors with dates and times of measurement - Reason for the investigation and results - External parameters e.g. temperature, rainfall
Service Zone	- WOA, WSZ and district meter areas (DMAs) names/ID numbers and locations - Historical data including dates on configuration, operational changes, and maintenance activities, ideally by DMA (e.g.	- Flows, levels, pressure from any monitored locations (e.g. pump stations, control valves, DMA entry points) - Chemical dosing data, if applicable	- All water quality parameters measured by discrete sample from customer taps and other system facilities - Historical data on failures including dates of failures and investigation results	- All water quality parameters measured with online sensors with dates and times of measurement - Reason for the investigation and results

	boundary changes, flushing) - Zone hierarchy from WTW to WSZ and DMA including key facilities like pump stations, valve vaults, and SRs - Type of additional disinfection or chemical addition (if any) - Location of monitoring equipment and discrete sample collection - Location of external monitoring (e.g. weather stations)		- Reason for taking sample (e.g., regulatory, failure investigation)	- External parameters e.g. temperature, rainfall
Pipe	- Location, pipe ID number - Diameter, material, lining, etc. - Date of installation and repairs - DMA, WSZ and WOA for each pipe - Hydraulic model outputs (e.g. velocity, pressure, water age) - Burst history - Properties supplied by the pipe -Number of discolouration or other water deterioration events per pipe	- Flows, levels, pressure from any monitored locations	- All water quality parameters measured by discrete sample from pipe sampling locations	- All water quality parameters measured with online sensors at pipe sampling locations with dates and times of measurement - Reason for the investigation and results

10.3.1.2. Data Connection and Integration Layer

This layer addresses the challenge of combining the various types of data and extracting the necessary parameters from storage in the previous layer data. The spatial connectivity between data elements is an important feature of DWDS data and harnessing this information in the Big Data analysis yields much more significant insight than considering water quality parameters alone.

This layer includes the production of a dataset fit for analysis by integrating across and between the different types of data. The integration component links the various types of data to each other and with their associated DWDS assets. For example, linking each water quality sample to its local pipe and service area asset hierarchy (district metered area, water operations zone, pressure zone, WTW, water source, etc.) is required to fully understand the route that the water follows to enable identification of the causes of deterioration. Data integration can be a complicated task. However, if good standardisation principles are followed in the data storage layer, the process can be facilitated using references and indices like water operations zone names or geospatial coordinates.

In the data pre-processing component, the integrated raw data are further cleaned to remove outliers, bad quality data, missing data, or chronological periods that should not be included in the given analysis. The pre-processing may also require certain calculations to be made using the raw data, such as assigning

classes (for classification analyses) or determining averages. Filling in missing values and removing unwanted parameters from the dataset are also included in this pre-processing.

Once pre-processing has been completed, the final component of this layer is the extraction of the clean dataset in the appropriate, accessible format and ready for further analysis in the next layer.

10.3.1.3. Data Analysis Layer

This data analysis layer is where analysis is performed using ML techniques for the creation of new knowledge from the available data. A critical component of this layer is the selection of the appropriate ML technique, which is dependent on the water quality question to be addressed, the type of output desired, and the type and quantity of the available data. We propose a six-step ML selection and implementation process as follows:

1. **Define the water quality problem:** The first step requires water quality experts to specify a task or a water quality question that has the potential to be addressed using data analytics solutions.
2. **Define the type of required output:** Once the problem is defined, the water quality experts should specify the goal of the investigation and desired type of output. For this framework, ML learning outputs have been categorised into four types: 1) prediction of future class (classification output - e.g., prediction of a water quality failure); 2) prediction of future behaviour (regression output - e.g., predicting the future values of certain parameters); 3) grouping of unlabelled data (clustering - e.g., splitting large datasets in groups based on various criteria); and 4) identifying relationships between parameters (correlation - e.g., identifying parameters that influence water quality deterioration). This definition should not be constrained by the available data initially, although it may need to be adapted through an iterative process with Step 3 to reflect the practicalities of the actual data.
3. **Type of available data:** This step connects this data analysis layer with the previous data integration layer. Here, the final extracted dataset is reviewed to determine the type, format, and most importantly quantity of data available. For example, continuous water quality monitoring data from the DWDS typically results in a dataset that is spatially and chronologically sparse and covers only a few water quality parameters. Given that the quantity of available data and number of included parameters influences the performance of some ML techniques, this review of available data is important for ML technique selection. For example, artificial neural networks (ANNs) are generally not applicable to

small and sparse datasets with a significant number of missing values (Ennett, Frize, and Walker 2001). In some cases, some initial exploratory analyses will be required to determine if the available data is sufficient to address the given water quality question and iterative reconsideration of Step 2 will be required.

4. **Machine learning technique selection:** Building upon the previous three steps, the appropriate ML technique is selected in this step, facilitated by use of *a machine learning technique selection tree* (Figure 10.2, further detail below). Some ML techniques cannot handle missing data and it is not always possible to infill the missing values, , or the target of interest may be a rare event and the technique must be carefully selected to address this. Monitoring samples datasets are temporally and spatially sparse and, moreover, from each sample taken, not all the water quality parameters are measured. The ML methods selected for this framework are presented in detail in the next section.
5. **Data preparation:** Once the ML technique has been selected, this step includes any final changes to the data format required for the selected ML technique.
6. **Application output:** In this step, the selected ML technique/techniques are trained using the available data and tested to check their performance on unseen data. Once the simulations are finished, the outputs are specified. These could include images, values, or tables. The outputs are then reviewed to ensure that the ML technique has produced effective results.

10.3.1.4. Machine learning selection tree

For a defined water quality problem (Figure 10.2, Box A), a path in the tree is followed considering the available data (Figure 10.2, Box B) and desired output (Figure 10.2, Box C), all of which are considered in Steps 1 through 3 of the ML selection and implementation process. The ML technique selection then proceeds by considering an additional factor that needs to be specified (Figure 10.2, Box D). This factor, termed 'interpretability', has been defined as the ability of the techniques to offer a transparent explanation of how outputs were calculated. Techniques with high interpretability, also known as "white boxes", offer a way to clearly demonstrate the logic behind outputs and indicate the contributions of various parameters to decisions. Techniques with low interpretability, also known as "black boxes", provide outputs without any explanatory elements. Different water quality questions might necessitate selection of techniques with higher or lower interpretability. For example, while seeking a prediction of future iron failures in a DWDS, deriving an understanding of the parameters that influence the prediction could be as useful as the prediction itself.

Following through the selection tree process, the final part identifies the appropriate ML technique (Figure 10.2, Box E) based upon the requirements for that specific water quality problem. Table 10-2 summarises such methods, covering those with examples of successfully tested ML techniques. New ML methods are continually emerging from research, hence the likes of table 10-2 require regular updating. The methods investigated for this framework are those with demonstrated applications in the water sector.

Neural networks and deep learning applications are applied to discover knowledge from large and complicated datasets (Lecun, Bengio, and Hinton 2015). Therefore, ANNs and LSTM techniques were both excluded from applications where discrete sample data is used because these datasets are spatially and temporally sparse with many missing values. Similarly, predictions based on regression techniques are not recommended with discrete samples due to their spatially and temporally sparse nature.

10.3.1.5. Presentation and communication of data analyses outcomes

This layer overlaps with the application output step of the data analysis layer, taking the ML model outputs and presenting them using graphs, tables, and images to facilitate understanding and interpretation of the results by decision makers. While the visual formatting is important, it is also critical that the most important and relevant results be carefully selected to clearly explain the ML outputs. Presentation of all outputs created in the ML analysis may create confusion for non-technical stakeholders but editing outputs for clarity must be balanced with providing sufficient evidence for interventions. Well-presented results provide sufficient and correct information to utility staff to make informed decisions for proactive interventions in the DWDS.

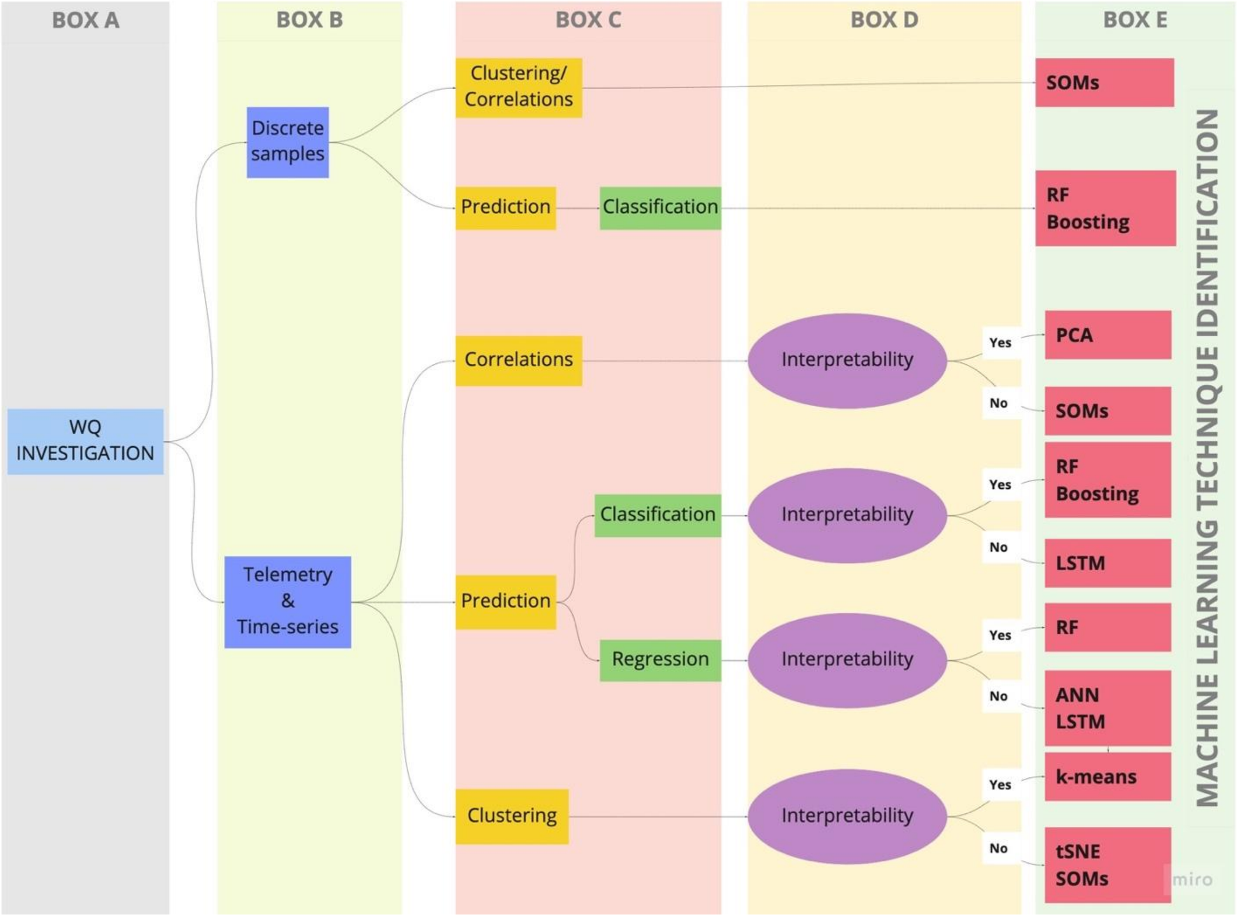


Figure 10.2: Machine learning method selection tree

Table 10-2: Summary of ML methods investigated

Method	Type	Output	Interpretable	Notes/Comments	Example References
k-means	Unsupervised	Clustering	Yes	Not suitable for datasets with missing values	(Maimon and Rokach 2006)
Principal Component Analysis (PCA)	Unsupervised	Clustering/ Linear correlations/ Dimensionality reduction	Yes	Not suitable for datasets with missing values	(Jolliffe 2002)
Self- Organising Maps (SOMs)	Unsupervised	Clustering/ Non-linear correlations	No	Good for datasets with missing values	(T. Kohonen 1990)
Random Forest (RF)	Supervised	Classification/ Regression	Yes	Ensemble decision trees with equal contribution to the final decision	(Breiman 2001)
Boosting Trees (Boosting)	Supervised	Classification	Yes	Ensemble decision trees with weighted contribution to the final decision	(Dietterich 2000)
Artificial Neural Networks (ANNs)	Supervised	Regression	No	Not suitable for datasets with missing values	(S. R. Mounce et al. 2014)
t-Distributed Stochastic Neighbour Embedding (tSNE)	Unsupervised	Dimensionality reduction / clustering	No	Ignores input rows with missing values so not suitable for many discrete sample datasets	(van der Maaten and Hinton 2008)
Long Short-Term Memory (LSTM)	Supervised	Regression / Classification	No	Deep learning method, requires a large amount of data	(Hochreiter and Schmidhuber 1997)

10.4. Application of the Big Data Framework

To validate and evidence the value of the proposed framework, its application with a water utility located in the north of the UK is presented. This water utility serves more than 5.5 million people via 250 WTWs and greater than 50000 kilometres of pipes, with approximately 1100 SRs. Two case studies are presented covering different aspects of water quality performance in SRs. Specifically, the insights sought were: evaluating the factors related to bacteriological activity in SRs, and the prediction of low chlorine concentration events in the SR outlets.

For the case study examples, the data storage layer of the Big Data framework comprised the water utility's in-house data management system with manual integration of external sources like weather data. The following sections present the application of layers 2 through 4 of the Big Data framework.

10.4.1. Example 1: Factors related to increased bacteriological activity in SRs

Layer 2 - Data connection and integration

Discrete water quality samples taken from the outlets of SRs and WTWs were collected for a period between January 2012 and May 2020. In these samples, various parameters were measured including bacteriological indicator parameters such as heterotrophic plate counts (HPCs) at 22^{o}C as well as flow cytometry total cell counts (FC_TCCs) and intact cell counts (FC_ICCs), disinfectant residual parameters, and other physical and chemical parameters. In addition, databases that included asset information such as estimated retention time (water age) within each SR, type of secondary disinfection at each WTW (chlorine/chloramine), and connections between SRs and their source WTWs were also collected. Finally, daily and hourly precipitation data for the same period were retrieved from the relevant Met Office weather stations [41] closest to each asset location.

All the raw data was cleaned and collated, and links between water quality data and the corresponding physical asset were created based on spatial, naming, and/or connectivity data. Precipitation was included as the average daily rainfall (mm) at each WTW. Monthly average values for water quality parameters measured at the WTW were calculated and linked to the SRs within each WTW service zone. Additional parameters were created and integrated with the main SRs' discrete samples dataset as follows: 1) the age of water exiting an SR (hours) as sum of the retention time of the given SR plus the retention time of the SRs that the water passed through upstream of the given SR (AgeofWaterLeavingSR); 2) the time (days) between the last reported SR cleaning date and the sample date (DaysFromCleaningDay), with negative values indicating samples that were taken before the last cleaning; 3) the monthly average total organic carbon in the WTWs (TOC_WTW_AVE); 4) the monthly average temperature of water exiting the WTWs (Temperature_WTW_AVE); 5) the monthly average flow cytometry total cell counts exiting the WTWs (FC_TCC_WTW_AVE); and 6) the daily average precipitation per month near the WTW (WTW_AverageDailyPrecipitation).

Layer 3 - Data analysis

The 6 steps within the ML selection and implementation process were performed for this water quality investigation as follows:

1. Define the water quality problem

The aim of this investigation is to understand the factors related to increased bacteriological activity in the SRs.

2. Define the type of required output

The required output in this investigation is correlations between parameters, with bacteriological parameters as the outcome parameters of interest. Numerical predictions are not required to understand these correlations.

3. Type of available data

As described above, the available data stems from the data integration layer. The outcome parameters of interest (bacteriological measurements HPC, FC_TCC, and FC_ICC) are only available as discrete samples taken from the outlet of the SRs. Telemetry data on water quality as well as calculated daily average water quality from the WTW outlet is also available. Weather data is available as time series data and calculated daily average data. Physical data on the WTWs and SRs is also available.

4. Machine learning technique selection

Given that this problem has an outcome characterised by discrete sample data and that the required output is correlation/clusters, the machine learning technique selection tree (Figure 10.2) directs towards SOMs as the ML method.

5. Data preparation

For the SOM application, the SR water quality data were prepared so that each row represents a discrete sample and each column is a different measured parameter from that sample, including the average monthly values of the water quality parameters at the WTW outlets feeding the sample location and the average daily precipitation per month in the given SR.

6. Application output

Two SOMs were selected from the analysis (Figures 10.3 and 10.4). These outputs consider many of the factors that have been shown to influence bacteriological water quality in the literature. The SOM analysis produces output planes for each parameter that visualise clusters of similar data by colour (low is blue, high is red) based on the range within the dataset, which in this investigation was set to colour-code based on the 5th and 95th percentile for each parameter without excluding any data. The SOM Toolbox 2.1 for MATLAB was used for all analyses[42].

The first SOM (Figure 10.3) investigates the effect of disinfectant residual type and concentration, along with retention time in the SR and temperature, on the bacteriological indicator parameters of HPC at 22oC (HPC_22), flow cytometry total cell counts (FC_TCCs) and intact cell counts (FC_ICCs). Both free (FreeCl) and total (TotalCl) chlorine are plotted, with the type of disinfectant (chlorine or chloramine) used for a post-clustering labelled plot (right hand side of Figure 10.3).

This SOM shows a large cluster of high HPC values (left half of plane) with correlations to high and medium TCCs and high HPCs. The high HPC cluster also has a tendency to correlate with higher age of water exiting the SRs, correlates strongly with low free chlorine, and somewhat with elevated temperature. The labelled map, which is developed post-clustering analysis to assign categorical parameters that best match the members of each cell, shows a very strong correlation between high HPC values and chloraminated systems. High HPCs corresponded to the entire range of total chlorine concentrations and therefore indicates that bacteriological activity is less strongly associated with loss of disinfectant residual than with type of disinfectant.

A cluster with increased numbers of ICCs (top centre of the plane) is correlated with high age of water exiting the SRs, high temperature and low free and total chlorine in both chloraminated and chlorinated SRs. Interestingly, this analysis shows no clear correlation between ICCs and TCCs.

The impact of SR cleaning on bacteriological activity is somewhat less clear from this analysis. There is a cluster of low HPC values (lower left of the plane) within chloraminated systems that correlates with a higher number of days after cleaning. Considering that recently cleaned SRs show as light blue (horizontal band in the middle of the plane), there seems to be slightly lower HPC values corresponding to this cleaning, but not exclusively so.

The second SOM in this analysis (Figure 10.4) investigated the impact of key WTW water quality parameters (TOC and FC_TCCs) in addition to disinfectant type and concentration, water age in the SRs, and precipitation.

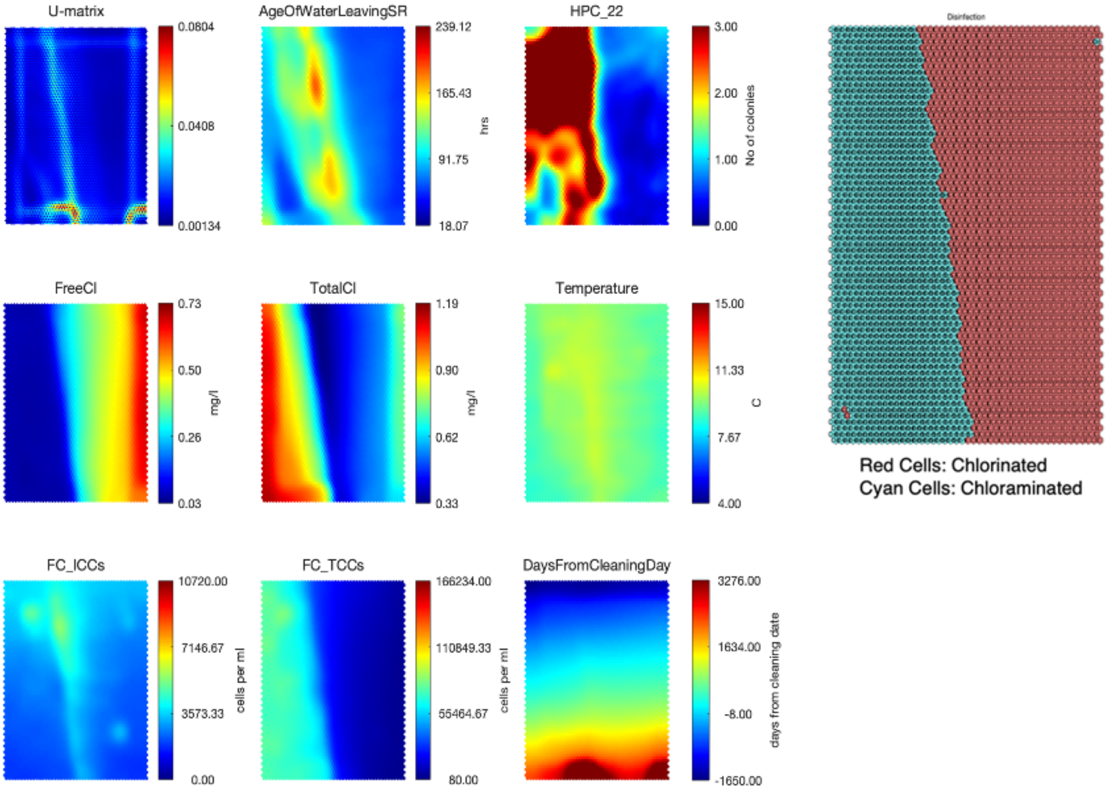


Figure 10.3: SOM showing impact of disinfectant residual type and concentration on bacteriological activity in SRs, including secondary disinfection labelled map

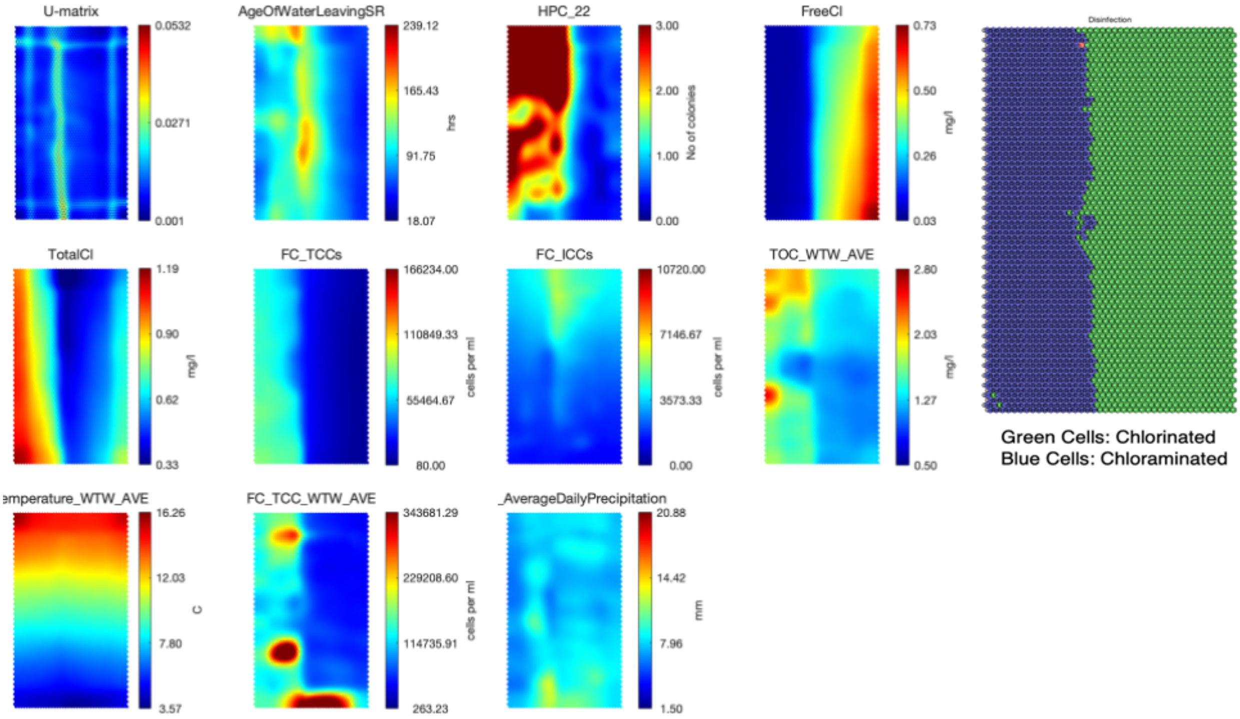


Figure 10.4: SOM showing additional impact of key WTW parameters on bacteriological activity SRs, including secondary disinfection labelled map

Layer 4 - Presentation and communication of data analyses outcomes

In this layer, the application outputs were presented in the form of colour-coded output planes (Figures 10.3 and 10.4). These outputs allowed for visual observation of correlations across multiple parameters and could be understood by a variety of water utility stakeholders. The outputs provided a good indication of which factors influence bacteriological activity in the SRs, although evidence in this format does not provide numerical values or estimates. As such, correlation analyses like these can answer general questions like the one posed in step 1 of the ML selection and implementation process for this case study but cannot address questions that require a numerical output like a score or ranking.

10.4.2. Example 2: Predicting low chlorine concentration events in the SRs

In this example application of the framework, the aim was to classify the SRs into either high-risk or low-risk categories based on a prediction of monthly low chlorine events by the ML models. The monthly temporal scale was selected as a prediction horizon as, on average, 2 to 4 monitoring samples per SR per month are collected for regulatory compliance (DWQR 2019a). A low chlorine event was defined as a sample where the chlorine concentration was measured below 0.3 mg/l to allow a small margin above the allowable minimum free chlorine concentration at customers taps of 0.2 mg/l. The dataset used for this example was the same as the one used for Example 1 and therefore layers 1 and 2 were already complete.

Layer 3 - Data analysis

The 6 steps within the ML selection and implementation process were performed for this water quality investigation as follows:

1. Define the water quality problem

The aim of this investigation is to classify SRs as high or low risk by predicting their low free chlorine events in the upcoming month.

2. Define the type of required output

The required output is the classification of each SRs into low and high-risk categories for the upcoming month based on low chlorine events.

3. Type of available data

The available data for the investigation are the water quality parameters from SR and WTW outlets during the period between January 2012 and December 2019 (complete years of data required, final 5 months from example 1 not utilised). The outcome parameter of interest (FreeCl) is only available as discrete samples taken from the outlet of the SRs. Telemetry data on water quality as well as calculated daily average water quality from the WTW outlet is also available. Weather data is available as time series data and calculated daily average data. Physical data on the WTWs and SRs is also available.

4. Machine learning technique selection

Following the machine learning method selection tree in Figure 10.2**,** two types of ML technique are suitable for this investigation, random forest and boosting trees. The SRs dataset is heavily unbalanced, meaning that most of the available data for training the ML model belong to the non-event, low-risk class. Therefore, RusBoost, a technique that combines random under sampling (of the non-event data) with the boosting tree algorithm was selected for this analysis (Seiffert et al. 2008).

5. Data preparation

The initial dataset contained water quality data taken from all SRs and WTWs on different days. Therefore, this dataset required a final transformation to a monthly scale for analysis. Monthly averaged values per parameter per SR and the chlorine standard deviation per month per SR were calculated. Given that the results of the previous investigation revealed different behaviour in chlorinated and chloraminated SRs, all chloraminated SRs were excluded from this analysis. The historical classification of low-risk or high-risk was calculated for each SR for every month of the dataset. Within a given month, high-risk SRs were those that had one or more low chlorine events for that month (chlorine measured value below 0.3mg/l in at least one discrete sample in that month). Low-risk SRs had no low chlorine events in the given month.

6. Application output

a. ML Model training

Two options for the ML model were tested during training. The first option used the water age exiting the SRs, the average daily precipitation per SR and the average monthly values of 15 water quality parameters per SR per month (Model RB.1 in Table 10-3) and the second focused on the free chlorine parameters: average monthly free chlorine , monthly free chlorine standard deviation, and average monthly WTW free chlorine along with the two parameters from the previous case study that had the highest correlation with low chlorine concentrations: water age exiting the SRs , and average water temperature together comprising Model RB.2 (Table 10-3).

During the training period, the ML model was used to predict the class (high-risk/low-risk) for each SR in the following month and this prediction was paired with the historical class for model training. For example, using the January 2012 water quality data as inputs, the historical SRs classification for February 2012 was produced as output (Figure 10.5). The ML models were developed in MATLAB 2019b utilizing 1000 weak learners for each model and tested for their performance based on their predictions for August 2019. A simple schematic of the model training and testing is presented in Figure 10.5.

The accuracy of the models was evaluated by using true positive rate as calculated with Equation (1) and the Matthews correlation coefficient (Baldi et al. 2000) as calculated with Equation (2), as follows:

$$TPR = \frac{TP}{TP+FN} \qquad (1)$$

$$MCC = \frac{TP\ X\ TN - FP\ X\ FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \qquad (2)$$

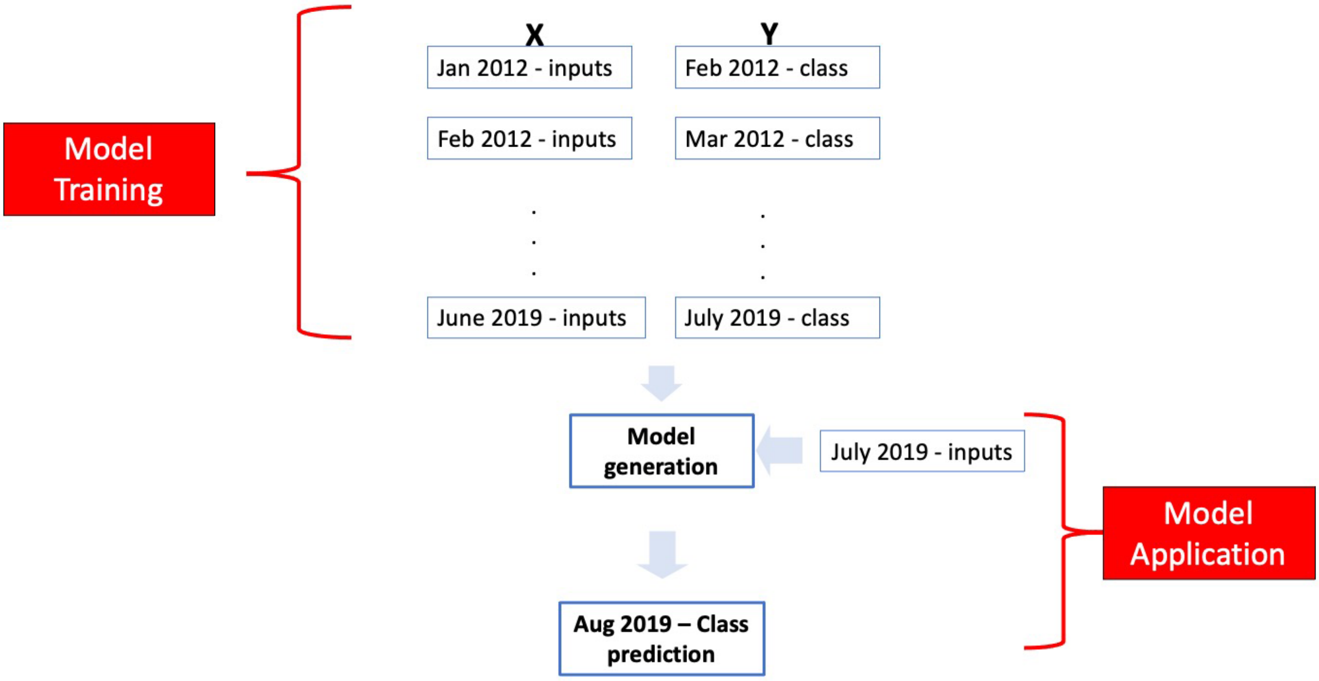


Figure 10.5: Monthly predictive model schematic for SR class prediction for August 2019

where TP is the number of true positive predictions, FN is the number of false negative predictions, TN is the number of true negative predictions, and FP is the number of false positive predictions. True positive rate is used for quantifying the proportion of correctly predicted positives (events) over all the actual events. The Matthews correlation coefficient is used for evaluating the overall performance of the model and has a range between -1 and +1, where a result of -1 indicates a poor predictive capability for the model and values close to +1 indicate good predictive capability.

b. ML Model results

The MCC results for the two ML model options (Table 10-3) show that both RB.1 and RB.2 performed well, especially in keeping a balance between correctly predicting more high-risk SRs (TPR=0.71 and TPR=0.72, respectively) and creating less false positives (MCC=0.44 for both.). Overall, RB.2 can be considered the best model given its slightly higher true positive rate.

In this example, the outputs presented to the utility decision makers included a list of SRs and their predicted risk category for a given month. Given that the boosting tree algorithm is a white-box model, part or all the decision trees that contributed to the final predictions were exported to illustrate which water quality parameters contributed the most to the model predictions (e.g., the most important predictor in Table 10-3). This valuable additional information, associated with the Interpretability factor for the selected ML method, allows the utility to understand not only the risk classification but also the factors that lead to this risk.

Table 10-3: Summary of the performance metrics for the two ML model options tested in Example 2

Algorithm	Model	PARAMETERS	MOST IMPORTANT PREDICTOR	TPR	MCC
RUSBoost	RB.1	Monthly free chlorine average, monthly free chlorine standard deviation, average monthly WTW total chlorine, monthly average HPCs @23C, Monthly average HPCs @37C, monthly average ICCs, monthly average TCCs, average water temperature, water age exiting the SRs, average daily precipitation per month per SR, average monthly WTW free chlorine, average monthly WTW total chlorine, average monthly WTW TCCs, , average monthly ICCs, , average monthly WTW water temperature, average monthly WTW TOC, , average monthly WTW pH	Monthly free chlorine average	0.71	0.44
	RB.2	Monthly free chlorine average, monthly free chlorine standard deviation, average monthly WTW free chlorine, water age exiting the SRs, average water temperature	Monthly free chlorine average	0.72	0.44

Layer 4 - Presentation and communication of data analyses outcomes

In this example, the outputs presented to the utility decision makers included a list of SRs and their predicted risk category for a given month. Given that the boosting tree algorithm is a white-box model, part or all the decision trees that contributed to the final predictions were exported to illustrate which water quality parameters contributed the most to the model predictions (e.g., the most important predictor in Table 10-3). This valuable additional information, associated with the Interpretability factor for the selected ML method, allows

the utility to understand not only the risk classification but also the factors that lead to this risk.

10.5. Discussion

10.5.1. Actionable information

The two example case studies delivered a variety of actionable information. In the first example, the water quality question was to identify factors that influence the bacteriological activity inside the SRs. The SOM outputs provided the evidence to support actions including: 1) closer investigation of chloraminated SRs and maintenance of disinfectant residual, especially when the temperature increases; 2) control and optimisation of the retention time in both the chlorinated and chloraminated SRs; 3) management and reduction of TOC exiting the WTWs; 4) improvements in the WTWs operation to respond to sudden changes in the quality of the raw water due to increased precipitation. Additional studies may be required to more fully explore each of these actions, including collection of data that was not available for the SOM analysis, but the Big Data approach has provided focus and clarity for such additional studies that would not otherwise have been possible.

In the second case study example, a prediction model for high-risk SRs based on their low chlorine events each month was created. The model was able to predict up to 72% of the low chlorine events for the investigation month, which is a high degree of accuracy that allows targeted interventions to take place. Furthermore, the fact that low chlorine events can be accurately predicted using mainly chlorine-related parameters is important, demonstrating that monitoring of supplemental parameters (e.g. other bacteriological indicators) would not be warranted to address this particular water quality question.

For the water quality examples used, changes occur due to complex physical, chemical and biological reactions and interactions occurring inside a pipe network. It was, therefore, important that the Big Data analytics investigation was directed by domain knowledge, the posing of the question and understanding sought that drives the third layer. This is crucial to ensure actionable information results. This finding is an underlying principle of the framework, whatever the application. It requires collaboration between experts in different water utility departments and with data scientists, in all the layers of the framework.

Part of the appeal of Big Data analytics is the ability to answer increasingly complex questions and make predictions for the future. Drawing upon the case study examples, basic data analysis could map the SRs with low chlorine measurements each month, perhaps identifying geographical areas with clusters of high-risk SRs. But such an analysis cannot identify underlying factors that contribute to the low chlorine events and cannot predict which SRs may have problems next month. It is this deeper understanding through iteration of the third layer that was key in obtaining the actionable information derived.

10.5.2. Framework

The Big Data framework, presented in this paper with applications to drinking water quality, emphasises the necessary steps to unlock the power of machine learning and advanced data analytics for water utilities. The framework systemises a process to ensure that actionable information is derived by unlocking the potential of previously siloed data. Importantly a selection tree process to identify the best ML techniques, driven from the insight required and the data available is central. This is based on the knowledge, and illustrates, that there is much more effort required for successful Big Data applications than coding a given ML algorithm.

Standardising data acquisition and storage, organising the data to facilitate analysis including generating links between different datasets, understanding the difference between available data types, and selecting the most appropriate data-driven techniques are all necessary steps to deliver actionable information and supporting evidence to inform operation and management decisions. Implementation of standards that guarantee the collection of good quality data and the organisation of the stored data are often under-resourced tasks at water utilities, yet have been shown to be core elements of this framework. The data-specific nature of current systems within the water industry as well as the lack of historical collaboration between the relevant areas of expertise perhaps explain why there have been so few Big Data frameworks proposed for the water sector before now.

One of the most challenging but vital aspects of layer 2 is the generation and association of links between different data. Linkages between different data sources, such as between asset information and measured water quality parameters, are critical for ML analyses yet are not often performed. For example, analysis of water quality sample data without consideration of the water treatment works supplying a given point in a network often falls short in answering the questions of interest. This is a key area where data scientists and water domain experts must work closely to understand what is possible and to ensure that appropriate associations are made. Unique ID and geocoding data are often useful here, but these should be supplemented by checking secondary data. For example, linking a pipe repair record to GIS data can be done based on location data, with a check made using pipe material data that is also frequently contained in both datasets.

The initial investment of time and effort for data collection, storage, and integration (layers 1 and 2) is often greater than what is needed for data analysis (layer 3). However, once created, layers 1 and 2 can then be used to support a multitude of analyses repeating and iterating layers 3 and 4 without the need to revisit layers 1 and 2. This was shown for the two case study examples presented. The return on initial investment in layers 1 and 2 can be further multiplied many times if automatic updating of datasets can be incorporated. The need for investment in layers 1 and 2 is great but the benefits will be felt across the water industry when the analytical power and decision-making support of layers 3 and 4 is unlocked.

Once layers 1 and 2 are complete, the question, and opportunity, becomes to consider what new and actionable information is needed and how to extract it. Studies in the literature which explored specific algorithm(s) with application(s) to a few water quality parameters have paved the way to greater understanding of the potential in Big Data analytics but the water industry lacked an understanding of how best to apply these techniques and which ones work best in which situations. The machine learning method selection tree proposes a novel, problem-driven approach to this, enabling a wide range of investigations and opening up the possibilities for taking Big Data analytics to the next level of application across the water industry.

10.6. Conclusions

A Big Data framework to enable water utilities to robustly and efficiently apply data-driven methods to derive new understanding from complex, traditionally siloed data is presented. The proposed framework is based on a four-layer approach:

1. the data storage layer, including a system to categorise and sort data
2. the data integration layer, where the importance of associating across and between data is emphasised, irrespective of types, formats and sources of data
3. the data analysis layer, is systemised as a six-stage process that is driven from precise articulation of the decisions that are to be informed. These steps are: 1) definition of the problem; 2) definition of the type of required output; 3) type of the available data; 4) selection of the ML learning technique; 5) data preparation; and 6) application output. A selection tree is used to inform the selection of ML technique based on 3 criteria: the available data (discrete water quality samples or time-series), the required output, and the need for interpretability of the process for producing the outputs.
4. the presentation and communication of data analyses outcomes, where careful selection from the huge number of outputs generated is essential to present a logic effective and evidence-based narrative.

The need for and integration of roles across water engineering and data science are set out across the framework. Layers 1 and 2 are often complex and time consuming. However, once comprehensively accomplished they readily enable a multitude of different explorations of the different data to derive different and deeper understanding by repeating and iterating layers 3 and 4.

Case study examples evidence the application of the framework for drinking water quality. The examples demonstrate the derivation of new understanding, such as the association between disinfection residual type and concentration and age of water exiting a service reservoir combining to correlate with increased two-day plate counts, and for the prediction of low chlorine events at the outlet of service reservoirs. This understanding readily informs operational decisions, such as managing disinfection residual dose and prioritisation of maintenance activities. Overall, the framework is demonstrated to provide robust data-driven understanding and evidence to inform vital water utility operational and maintenance decisions.

11. Discussion

This thesis has sought to aid WUs obtaining new information regarding drinking water quality from the data that they have stored in their systems or collect daily. The ultimate aim of this work is to provide a holistic approach on the data collection, storage, management and analysis and to offer solutions on real-world drinking water quality problems that could improve the current engineering practice. For that reason, a number of data-driven models, based on ML techniques, are generated to tackle real-world water quality investigations using WQ data, already collected by SW. The results of these investigations provide actionable information to SW decision makers regarding both the understanding of some of their DWDS and the prediction of future water quality behaviour in them. These investigations indicate several advantages and weaknesses of the generated data-driven models that are presented in the following section. Moreover, to accomplish its main aim, this thesis main contribution is a 4-layer framework that develops a big-data environment for providing actionable evidence for supporting decisions over the proactive operation and maintenance of the DWDS and guaranteeing high - quality drinking water to consumers. In this chapter, a general discussion over the ML methodologies and the necessity of the big-data framework is presented based on the knowledge gained in the previous chapters of this thesis

11.1. Advantages and disadvantages of the models

11.1.1. Self-Organising Maps and Principal Components Analysis

SOMs and PCA are unsupervised ML techniques. The former one is mainly used for clustering and the latter one is mainly used for dimensionality reduction. However, both could also identify correlations and connections between the various features of the dataset. This thesis indicates that both these techniques could be valuable tools for the water industry. The main advantage of both techniques is that they provide a simple visual representation of complex datasets as it was presented in chapter 5 and 10. In addition, both techniques can provide multi-parameter correlation that helps visualise the factors that cause deterioration events. Chapter 5 and 10 demonstrates, SOMs have four main advantages when used for understanding the correlations between the various parameters compared to PCA. More specifically:

- SOMs could handle missing values while PCA requires either to remove lines with missing values or to replace the missing data with synthetic data. In chapter 5 the analysis made using SOMs has used the entire dataset (tables 5-1 & 5-2) while the analysis using PCA has only a very small percentage on the dataset (tables 5-6 & 5-7).
- PCA could only provide information regarding linear relationships between the various parameters compared to SOMs that, as an ANN, is capable of uncovering non-

linear correlations as well and therefore it is a better tool for understanding WQ inside DWDS where complex reactions occur. For example, PCA was able to identify the relationship between low free chlorine and bacteriological failures (figure 5.7) but not with other parameters such as age of water, temperature, and precipitation in the SRs as the SOMs analysis did (figure 5.2). In addition, in chapter 10, PCA could not indicate as clear correlation between the inlet turbidity and the TCCs in the outlet as the SOMs analysis did (figures 9.6 and 9.5 respectively).

- SOMs is a better tool in visualising the correlations as it could provide general clustering information per parameter and specific correlations between clusters of one parameter and clusters of the other - for example mid to high numbers of HPCs correlating with mid to low numbers of free chlorine in the network (figures 5.2) or TCCs in the disinfection tank outlet following the inlet turbidity and inlet colour trend (figure 9.5).
- Finally, SOMs analysis can produce clusters of qualitative parameters and, thus, identify the correlations between them and the numerical ones. Using this qualitative parameter in chapter 5 and 6, the separation between chlorinated and chloraminated systems was visualised and a further understanding of the bacteriological behaviour in these two systems and of the impact of switching disinfection was achieved.

For the above reasons, chapter 5 and chapter 9 recommend the use of SOMs when the available dataset is a water quality monitoring samples' dataset. However, PCA performed well with telemetry data as chapter 9 demonstrates. The Balmore WTW investigation in this chapter indicates that PCA is a good tool for uncovering the linear correlations in telemetry and timeseries datasets and, in combination with SOMs, understand which correlations are linear and easy to uncover, and which parameters are not linear. In addition, has one more advantage compared to SOMs, it quantifies the correlations between the parameters. This gives an information of how much linearly related these parameters are and how one parameter is dependent on the other.

Both methods have two main disadvantages. While both can provide correlations between parameters, they could not provide information about the chronological starting point that the correlation between the parameters had appeared. In addition, both methods cannot provide any information regarding the temporal distance between these correlations, in other words, which parameter influences the other and after how much time this influence will appear.

11.1.2. SRs' decision trees risk ranking methodology

The SRs' risk ranking methodology, presented in chapter 7, that uses decision trees ML methods has been proven to be a good tool for the prediction of low chlorine and coliform events. The unique models managed to predict up to 72% of the low chlorine events and the combined model up to 74% of these events. Its accuracy is not the methodology's only advantage as its interpretability aids WUs operational and managerial stuff in understanding the main parameters contributing to the model's outputs and the way that the model's final outputs were produced. In addition, the ability of these models to provide the score of each SR based on its likelihood of failure, could help decision makers create a risk ranking list of SRs per month that they could then use to prioritize their interventions only to the tanks that are in a continuous risk. However, their main disadvantage, which is also a general disadvantage for some ML applications, is the ability to predict the extremely rare event. Some of the techniques, presented in chapter 7, may tackle this problem but with the cost of generating higher numbers of false positives. Therefore, decision makers should also consider this factor in their intervention approaches.

11.1.3. Short-term prediction models

The chlorine losses predictive model in chapter 8 and the TCCs predictive model in chapter 9, demonstrate that they can be important tools for WUs' operations. Both models do not require any parameter to be calibrated as the deterministic models do, and once created, they require less time to train and simulate. The models can easily identify patterns in the data that cannot be identified by humans or by simple linear regression models. In addition, once produced they could be trained even without human interaction by just including them in the SCADA system of a DWDS. Another advantage of these models is their simplicity comparing to the traditional numerical models that require a high number of parameters for their simulation. ML models, in general, are improved when more data that include all seasonal changes are available. In these two cases the available data covered a maximum of 9 months period which means that a period of the year was not covered. As mentioned in the discussion sections in both chapters, there are high chances that both models could produce better results once more data are available. Finally, one main advantage of both models is that they are data driven. Both models are trained using just data without taking under consideration the hydraulic characteristics of the DWDS. This means that they could be set and used in every similar DWDS that has enough amount of timeseries data.

As regards their limitations, their main one is their requirements of a large amount of data for their training. In addition, these data should include outliers, extreme values, and rare events to allow the algorithms to learn from them during the training period and make more accurate forecasts. Potential absence or overrepresentation of these type of data can skew the training of the models and make the algorithms biased. Another issue that could generate bad predictions is the absence of relationships in the available dataset between WQ parameters that are known that are related. For example, in this thesis the chlorine loss model, in chapter 8, could not predict the chlorine losses when temperature data was used as inputs, because there were not enough chlorine loss events related to temperature changes in the dataset. Finally, the quality of the data collected by sensors, is at the least equal if not more important parameter than the quantity of the data. Bad quality of data makes these supervised predictive models useless as they are trained in a false environment. As WUs are working for the transformation of their systems towards the digital era, it is very important to identify the protocols required for the correct installation and maintenance of sensors in their DWDS and produce unsupervised ML models that could accurately and promptly detect anomalies in these datasets.

11.2. The necessity of a big-data framework for managing drinking water quality

The drinking water quality big data framework, presented in chapter 10, is a holistic approach on data management for the enablement of data-driven application by the water utilities. This framework has emerged while investigating the various case studies of this thesis. More specifically, the necessity of a proper data storage and a proper data integration layer was emphasised during the period where the grab monitoring samples datasets were generated. This process, presented in chapter 4, consumed a significant amount of time for assembling the grab monitoring samples, the physical assets information, and the external data and integrating them into one dataset. However, once created, the datasets were used in the case studies presented in chapters 5,6 and 7 and provided a better understanding of the processes in SW’s DWDS. These findings highlighted the need for WUs to standardize the data acquisition and storage processes for speeding up the data collection and facilitating the linkage between the different types of data. This process requires an initial time and investment spent by the WUs, but, once it is made, it will enable a deep exploration of the available data, through the data analysis and presentation and communication of data analyses outcomes.

Regarding the data analysis layer, this thesis demonstrated, in these 5 different case studies, that the actionable information for improving drinking water quality could be gained by following its proposed ML application. Considering the type of questions to be answered and the type of evidence required are the first and main steps in the ML application process. Once these are defined and the available data are identified, the machine learning method selection tree enables the application of the most appropriate ML technique for further data investigation that could unlock information for directing WUs investment into a proactive drinking water quality management. This novel approach that starts with the water quality question, and not with the ML technique, is critical in showing what data analytics do and, thus, aid WUs in integrating them in their business.

To summarise the above, the proposed in this thesis, big-data framework, directs the conversation over the application of data-driven techniques in the water industry from the one case study solution to the holistic data-driven approach of the operation and maintenance of the DWDS.

12. Summary, conclusions, and future work recommendations

12.1. Summary

In this section a summary of the 5 research chapters is presented. This summary includes the aim, the approach, and the key outputs of each one of the chapters.

Chapter 5. Understanding the factors that contribute to the bacteriological increase and cause bacteriological failures in the SRs will assist WUs with new knowledge regarding their DWDS and with the ways that future deterioration events could be avoided. This chapter investigates these factors by using SW's SRs water quality samples dataset as inputs in two different correlation clustering ML techniques, SOMs and PCA. The aim is mainly to identify which water quality parameters influence the bacteriological activity in the SRs but also to compare the ability of these two ML methods in analysing water quality samples dataset. The ML methods outputs demonstrated that SOMs is the only, out of these two techniques, that should be applied in water quality samples datasets for correlations. This is because the SOM method overpasses the issue with the large number of missing data that these types of datasets have, but also because it could identify more complex non-linear relationships between the water quality parameters. SOMs outputs indicated 4 main factors that are related with the bacteriological failure in the SRs, the high water temperature, the high age of water, the low chlorine residual and the high precipitation in the area surrounding the SRs. In addition, this research found that there is a higher bacteriological activity in chloraminated systems in comparison to chlorinated systems, however, the highest numbers of live bacteriological cells (ICCs) appear in both systems with high temperature, high age of water and low disinfection residual. Finally, another interesting finding is that in chloraminated systems, cleaning the SRs noticeably reduces the bacteriological activity in the SRs in contrast to the chlorinated systems where in addition to cleaning the SRs, a reduction of the circulation time of the water (reducing the water age) is also required.

Chapter 6. In this chapter, SOMs are used as a main correlation tool for investigating the impact of switching disinfection from chlorination to chloramination on drinking water quality. The data, used here, are drinking water quality samples data taken from various customer taps in 11 DWDS that switched disinfection from chlorination to chloramination over the investigation period (2012-May 2020). The disinfection type qualitative parameter was used as an indication for grouping the SOMs clusters in either chlorination or chloramination and therefore in clusters prior and post the disinfection switch. The aim of this chapter is not to criticize SW's decision for switching disinfection but to highlight the parameters that WUs should monitor before and after switching disinfection and avoid deterioration of the drinking water quality. Results indicated that switching disinfection reduces the DBPs' concentrations in the water, however, increases the bacteriological activity in these systems. A potential reason for the bacteriological increase after the switch is potentially related to the increased TOC concentrations found in both the WTWs and the taps of the DWDS after the switch. In addition, the increased turbidity related to chloraminated systems indicates a change in the water chemistry after the switch. Therefore, SOMs investigation indicated that in order to minimize the impact on the quality of the drinking water, WUs should concentrate in reducing the TOC and metals concentration in the WTWs.

Chapter 7. In this chapter a data-driven methodology that classifies the SRs in two different groups, the ones that will fail (High-risk SRs) and those that will not fail (Low-risk SRs) using 3 different ML ensemble decision tree methods (RF, AdaBoost, RusBoost) is presented. Ensemble decision trees are selected, in this case, as they are "white-box" approaches and both the classification decisions made and the factors that influence these decisions could be identified once the models are trained. Two type of failure criteria were investigated: low chlorine events - a failed SR is the SR with at least one sample of Chlorine less than 0.3mg/l in one month - and coliform bacteria events - a failed SR is the SR with at least one sample that contains coliform bacteria colonies. The methodology uses as inputs the monthly average values of various water quality parameters sampled in either the SRs or the WTWs that feed them and as outputs the class that each SR belongs to. For the low chlorine events prediction, the predictive horizon is set equal to one month ahead using the WQ parameters of the previous month, while for the coliform events a 2-month horizon was set using the WQ parameters of the month before. The aim of this work is to investigate the potential of this data-driven methodology in correctly classifying the SRs and, thus, becoming an important supporting tool for WUs' proactive interventions in their SRs. Thus, the methodology is tested for its performance in both predicting low chlorine events and coliform bacteria events, and the comparison of the various ML methods using different combinations of inputs is made. Initial results indicated that the best ML method for both the low chlorine events and the coliform events was RUSboost (low chlorine events: TPR=0.72, TNR=0.78, MCC=0.44, coliforms:TPR=0.8, TNR=0.7, MCC=0.24). However, a further investigation using a combination of the best performing models indicated that the combined approach increased the performance of the methodology (low chlorine events: TPR=0.74, TNR=0.78, MCC=0.45, coliforms: TPR=0.65, TNR=0.75, MCC=0.24).

Chapter 8. Predicting chlorine losses at the end of water distribution trunk mains up to certain hours ahead could be beneficial for WUs. This is because it will minimise the risk of water with insufficient chlorine residual and, thus, with higher risk of bacteriological deterioration, reaching the water mains and the consumers taps. The aim of this chapter is to build a data-driven methodology that identifies chlorine loss events in the trunk mains and predicts future ones using sensor temperature, chlorine, and flow data. This methodology initially detects the chlorine loss events by identifying the local peaks and the starting and ending points of the events. Subsequently, for each identified event, the methodology traces each associated flow or temperature event. These flow or temperature events are used for training the predictive model. Three different ML methods are used for the prediction model, FF-ANN, ANN-NARX, RF. The methodology is tested in three different trunk mains of the same DWDS with similar pipe characteristics (similar diameters, material, and length) but different hydraulics characteristics. Results indicate that the model could accurately predict chlorine loss events with a period of up to 10 hours ahead for one of those mains (shorter time predictions for the other 2 trunk mains). In addition, the investigation demonstrated that the best input parameter for this analysis is the flow of the water and the best out of the three ML methods is the NARX-ANN.

Chapter 9. In September 2020, two bacteriological sensors that measure TCCs were installed in Balmore WTWs, one at the disinfection tank outlet and one at the WTW outlet. This sensor provides SW data for additional WQ parameters that shows the bacteriological activity in the plant. Thus, in this chapter a data-driven investigation over the general bacteriological performance of the Balmore WTWs is made, using the TCCs data and some of the other available flow and WQ data captured in various locations in the WTW. The aim of this investigation is to, firstly, identify the main parameters that influence the TCCs and then to investigate if a data-driven model can, accurately, predict the TCCs exiting the works. For the former analysis SOMs and PCA methods were used. Their outputs indicated that the main factors that are related to increased bacteriological activity (high TCC numbers) in the water exiting the plant, were the high inlet flow, the high inlet turbidity, the low chlorine residual and high TCCs numbers in the water exiting the disinfection tank. Moreover, SOMs' analysis indicated a further correlation between high inlet colour and high TCCs as well as a weaker correlation between high TCCs in the outlet and low pH in the water exiting the disinfection tank. As regards the predictive model, this was built using past WQ data from the various stages of the plant (inlet, disinfection outlet, outlet) and TCCs data as outputs. The TCCs data were either the actual TCC numbers (regression approach) or some TCCs thresholds that characterised the water exiting the plant (TCCs<20000,20000<TCCs<50000, 50000<TCCs<90000 and TCCs>90000 - classification approach). The predictive model used 4 different ML methods, RF, LSTM and FF-ANN for the regression approach and RF, LSTM and RusBoost for the classification approach. The predictive model was tested in data not seen by the model during the training period. The results indicated that the model is able to predict TCCs with an accuracy of up to 82% for a forecasting horizon of 12 hours ahead and with a prediction of up to 77% for a forecasting horizon of 23 hours ahead.

Chapter 10. WUs collect a vast amount of water quality data that are stored in silos. As ML approaches can provide further knowledge from data such as those that WUs collect, it is worth exploring a holistic approach that will facilitate these applications in drinking water quality problems. Therefore, this chapter proposes a big-data framework that enables the application of these ML methods through a systematic approach consisting of 4 layers. More specifically, the proposed layers of the framework are the storage layer, the data connection and integration layer, the data analysis layer, and the presentation and communication of data analysis outcomes layer. In addition, in this chapter a novel ML selection tree is proposed that selects the most appropriate ML method for the data analysis, based on the type of the WQ problem and the available data for its solution.

12.2. Conclusions

This work attempted to provide a better understanding of drinking water quality and DWDS by investigating the ability of the methods that could use data already collected by WUs as inputs and produce outputs that could be used for supporting decisions for proactive interventions in the DWDS. In addition, this thesis proposed a new operational approach on the use of the WQ data that WUs collect in their daily monitoring routing programs. Overall, this work had 5 main objectives, as presented in the introduction chapter, and manages to answer to these objectives as follows:

1. Investigate the existing machine-learning techniques and the ways that these could be applied to drinking water quality problems

This study investigated 8 different ML models (SOMs, PCA, RF, AdaBoost, RusBoost, Feed-Forward ANNs, NARX ANNs, LSTM) in chapters 5 to 9 and demonstrated in real world drinking water problems their abilities and weaknesses. In addition, the thesis proposed another two ML techniques (k-mean and t-SNE) as potential techniques that could be used by WUs for their benefit.

2. Develop data-driven models for understanding the roots for drinking water deterioration in DWDS

A model based on SOMs and PCA was presented in chapter 5 for understanding the factors that increase bacteriological activity in SRs. In addition, SOMs model was applied for understanding the impact in drinking water quality of switching disinfection from chlorination to chloramination (chapter 6). Finally, both SOMs and PCA were used in a model for the identification of the factors that increase TCC numbers in the water exiting the WTWs and consequently decrease the quality of the water entering the DWDS (chapter 9).

3. Develop predictive data-driven models for water quality deterioration events in the DWDS

A model based on ML decision trees was produced for the prediction of low chlorine events and coliform events on SRs (chapter 8). This model was used for creating an SR risk ranking list for WUs to use as a supporting tool for the management of their SRs. A model was produced for the prediction of chlorine loss events at the end of the water distribution trunk mains(chapter 8). This model managed to predict accurately the chlorine losses up to 10 hours ahead. Finally, a model for the prediction of TCCs exiting the works up to 12 hours ahead was created (chapter 9). This model managed to accurately predict TCCs exiting the works and classify these waters into 4 different groups (minimum, low, medium, high risk) up to 23 hours ahead.

4. Compare the various data-driven methods and suggest the most appropriate for a specific water quality problem

In chapter 5, SOMs technique has been compared with PCA technique on their ability to identify correlations between the various WQ and other parameters when WQ monitoring datasets are used. The work in this chapter indicated that SOMs is a better technique than PCA as it tackles the problem of sparse data and missing values better than PCA.

In chapter 7, RF, RF with SMOTE sampling method, RF with ADASYN pling method, AdaBoost and RusBoost were compared in their ability to classify the SRs into low-risk or high-risk classes . This work indicated that the best methods to use for this type of prediction are RusBoost and RF with ADASYN. Moreover, this work showed that the model that combines the best single models, using a weighted average of the results, could increase the accuracy compared to the single models.

In chapter 8, RF was compared to FF - ANN and to NARX - ANN in their ability of predicting chlorine losses at the end of the water distribution trunk mains up to certain hours ahead. This work indicated that the best out of these three ML techniques was NARX - ANN that managed to accurately predict chlorine loss events up to 10 hours ahead.

In chapter 9, RF, FF - ANN and LSTM were compared in their ability of predicting TCCs exiting Balmore WTW (regression approach) and, for this type of data, RF was found to be the most accurate method, followed by FF-ANN. In addition, RF, LSTM and RUSboost were compared in their ability of classifying the water exiting Balmore WTWs into 4 different classes and, again, RF was proven to be the best out of these three models for this dataset. Finally, in the same chapter a new comparison between SOMs and PCA was made for the identification of the factors that increase TCCs in the Balmore WTW outlet that demonstrated that for this type of data both techniques are instrumental.

5. Present a new strategic approach that includes changes in WUs mode of collecting, integrating, and analysing their own data to create evidence that supports decisions over a proactive management of their DWDS.

Chapter 10 and partly chapter 3 presented a big-data framework that, once applied, could facilitate the ways the data storage and acquisition processes and, consequently, the data analysis with the use of ML techniques is made. The final aim of this framework is to aid WUs extracting valuable information for their DWDS from the data that they collect daily, that could support their decisions regarding the maintenance and the operations of their DWDS.

12.3. Thesis contribution

This thesis has the following key contributions:

1. **The main contribution of this thesis is the big-data framework for providing actionable information for the water quality management in the DWDS (objective number 5).** This framework consists of 4-layers and is developed for creating a data oriented supporting tool that could assist WUs in the proactive operations of their DWDS to manage drinking water quality. The 4-layer framework proposes a new holistic and systematic approach regarding the data storage, integration and analysis using the appropriate ML methods for certain water quality problems. With this approach, WUs gain a greater understanding of the potential of the data analytics and their ability to provide actionable information that could be used to completely transform the management of drinking water quality in their DWDS (chapter 10). Therefore, this framework answers the objective number 5 of this thesis.

2. **The second contribution is a methodology for understanding correlations between various water parameters including WQ parameters, other water characteristics and asset information (objective number 1 & 2).** This thesis demonstrated that this methodology could be used for the identification of the correlations between the various WQ parameters in different water related problems and could aid WUs in identifying factors that contribute to drinking water deterioration (Chapters 5, 6, & 9). Therefore, this methodology addresses objective number 2 of the thesis.

3. **The third contribution is a white box data-driven methodology for predicting the SRs that are at risk of bacteriological compliance (objective number 1, 3 & 4).** This methodology applied various decision trees methods for the prediction of low chlorine and coliform events in the SRs one month ahead and therefore classify the SRs into either the high-risk or the low-risk class based on these predictions. In addition, as a white-box approach, this methodology could help the users understand both the reasons behind the methodology's final decision and the parameters that contributed the most in that. This method can be used for creating an SRs risk ranking that could be used to direct the cleaning interventions in the high-risk SRs (Chapter 7). For this methodology various ML techniques were compared and, therefore, it contributes to both objectives 1, 3 and 4 of the thesis.

4. **The fourth contribution is a new ML based model for the prediction of the chlorine losses in water distribution trunk mains (objectives 1, 3 & 4).** This methodology was used for the prediction of chlorine loss events in the water distribution trunk mains, related to either sudden changes in the temperature of the water or sudden changes in the water flow (Chapter 8). This method could be used as an alarming tool for the proactive intervention to avoid bacteriological failures in the DWDS and, therefore, contributes to objectives 1,3 and 4 of the thesis.

5. **The fifth contribution is a methodology for investigating the bacteriological performance of WTWs (objectives 1,3 & 4).** This methodology combines both supervised and unsupervised ML methods for understanding the factors that increase bacteriological activity in the water exiting the WTWs and for short-term forecasting the bacteriological behaviour of the water exiting the WTWs. This method could be integrated in every WTW SCADA system and assist WTWs' operators in adapting the treatment stages to unexpected changes and improving the quality of the water exiting the plant (Chapter 9). As the chlorine losses model, this model contributes to objectives 1,3 and 4 of the thesis.

6. **The sixth contribution is an improved understanding of the bacteriological activity in the SRs.** The methodology in Chapter 5 indicates that the main factors that increase the bacteriological activity in the SRs are the high-water temperature, the high water age, the low chlorine residual, and the high precipitation in the SRs' area. In addition, it points out that the cleaning of the chloraminated SRs has a huge positive impact as regards the reduction of the bacteriological activity in these DWDS. However, the water age is a dominant parameter in the chlorinated SRs, and, therefore, in addition to cleaning, other operational interventions, such as the control of the water recirculation, are required to reduce the bacteriological activity in these tanks.

7. **The seventh contribution is the attempt for a better understanding on the impact on drinking water quality of switching from chlorination to chloramination.** The findings of the SOMs methodology in Chapter 6 indicated that the disinfection switch could change the chemical balance inside the DWDS and, therefore, could increase the bacteriological activity there. As a consequence of that, the turbidity levels in these DWDS could be increased and discoloured water could be noticed in the customers' taps.

8. **The eighth and final contribution is a general technical guidance for the WUs.** The case studies presented in this thesis, in combination with the proposed framework, could be used as guidance by WUs for improving their data storage and management system, selecting the appropriate models for specific water quality related problems and, therefore, producing high quality water and increasing their reputation.

12.4. Future work

This thesis addresses a few topics regarding the management and the analysis of the data that WUs collect for the quality control of the water that they serve to their customers. The methods that are generated in this thesis could be directly used by WUs to tackle some of their water quality issues in the DWDS. However, this work establishes a point of departure for further research in the field and raises many opportunities for future work in the area.

A starting point for further research is the continuous investigation of other ML methods that could be applied for other water quality investigations. This work used different ML methods and indicated their advantages and disadvantages. However, machine learning as a scientific field is continuously evolving especially in the age of digitisation that we live in. Therefore, new methods are generated, and new approaches are already investigated in other scientific fields that could be useful for water quality investigations. Further work that applies some new methods in real world water quality problems could indicate in which type of water quality problems they could be used and what are their opportunities and limitations. The successful methods could then fill the machine learning decision tree, presented in chapter 10, and provide WUs with more methods for their investigations.

To go one step further, the work in the field should concentrate on extreme deterioration events. As the methodology presented in chapter 7 indicates the limitation of some ML methods in predicting extreme events, future research should focus on methodologies that aim to understand the factors that cause these events better. However, working with extreme events requires a better approach by the WUs. More specifically, extreme events (e.g. a pipe burst, an error in the WTWs or a coliform appearance in a discrete monitoring sample) should be better captured by the WUs. Therefore, future work should concentrate on the historical events captured by the WUs, the information that should have been included in these events, and the potential future accurate detection using mostly unsupervised or semi-supervised ML methods.

Another interesting topic that this thesis recommends focusing on, is the research into the potential of deep learning (DL) applications in the water quality sector. As mentioned before, DL is the state of the art in ML applications with various applications in speech recognition, visual recognition but also event detection and predictions (Lecun, Bengio, and Hinton 2015). In this thesis, due to data unavailability, DL applications were used only in Chapter 9 where the long short-term memory (LSTM) model was applied for the prediction of total cell counts exiting Balmore WTW and, unfortunately, performed poorly. However, their potential in timeseries data has been proven in different engineering works (Assem et al. 2017)(Z. Y. Wu and Rahman 2017; Assem et al. 2017; Wei, Yue, and Rao 2017; Ronao and Cho 2016; Kuremoto et al. 2014; P. Liu et al. 2019; Barzegar, Aalami, and Adamowski 2020).Therefore, future research in the field should focus on the applications of these methods in drinking water quality investigations on DWDS when large amount of data is available.

With this work, the insufficiency of the amount of the discrete data that WUs collect has been prompted. It was indicated, in the thesis, that there are certain limitations on the use of sample datasets for data analysis as these are spatially and temporally scarce. For example, in some case studies, presented here, there were some DWDS where only a sample per year is collected for monitoring purposes. This finding shows that these areas are not fully controlled by the WUs. Therefore, further discussion and research work is required regarding the changes in regulations for monitoring the DWDS. The work should mainly focus on the "how much data is enough to answer our water quality problems" topic. In addition, the research should contribute to the discussion regarding the "digitalisation" of the DWDS, as this is defined by the Institute of Water Association (IWA 2018), by exploring the benefits and the disadvantages of installing sensors in the networks in comparison to the grab monitoring samples collection.

Finally, the WUs direction towards the complete digital transformation of their systems will create many labelled data that could be used in in-line and automated supervised ML methodologies for predicting future water quality events (like the methodology in chapter 9). However, as mentioned in section 11.1.3, this direction requires not only large datasets but also good quality data. Future work on the digitalisation of the DWDS and the improvement of the drinking water quality requires to concentrate in producing protocols for validating the sensor measurements but also focusing on unsupervised methods, such as PCA or tSNE, for the accurate and fast detection of anomalies in the sensors' outputs. This work should concentrate into creating final digital datasets that WUs would feel confident to use as inputs in their data analysis.

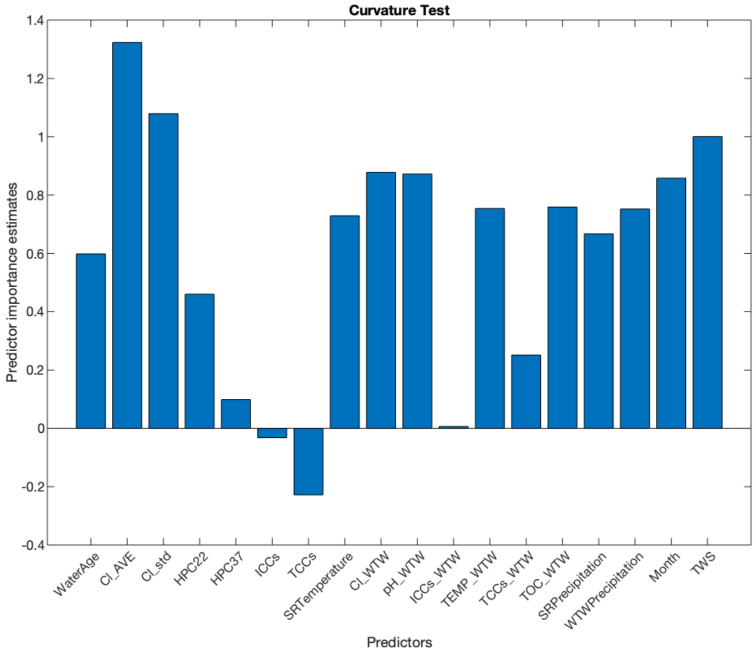
APPENDIX A: Predictor importance graphs for the ensemble classifiers in chapter 7

In this appendix the predictors' performance graphs for the 7 low chlorine events classifiers, when all the available parameters are used, are presented.

Low Chlorine events predictors' performance graphs

RF.1 model

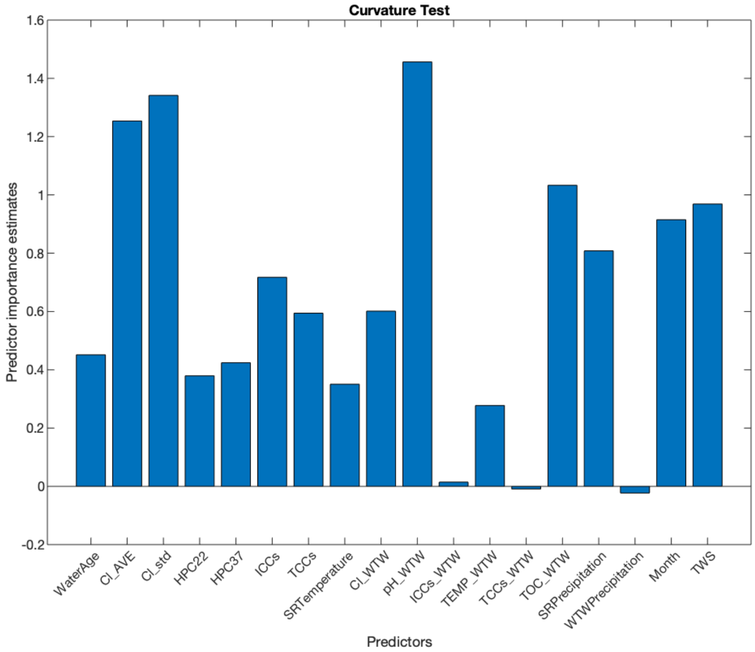
Most important parameters are: average free chlorine per month, free chlorine standard deviation per month, SR Name, Month of the year, average free chlorine in the WTWs per month



Performance importance of each parameter of the RF.1 model

RFS100.1 Model

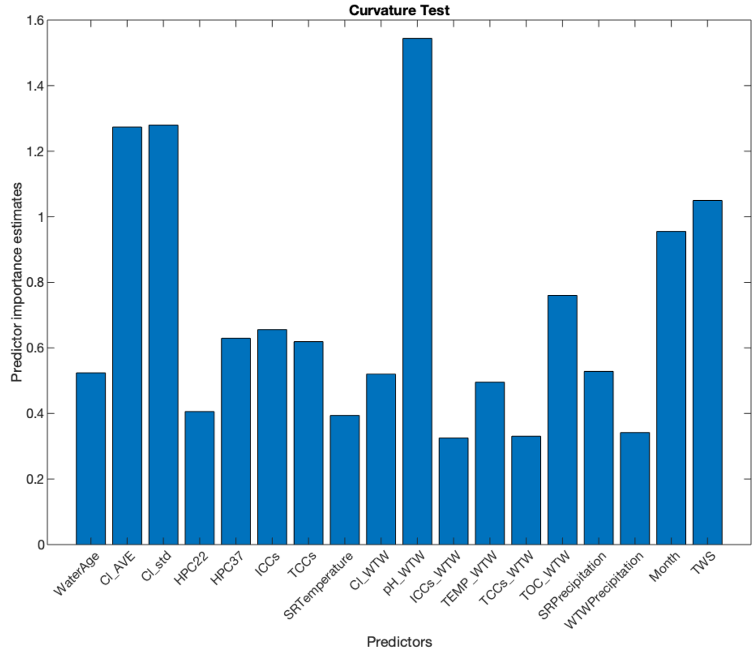
The 5 most important parameters are: average free chlorine per month, SR Name, average TOC per month in the WTWs, average pH in the WTWs per month, free chlorine standard deviation per month



Performance importance of each parameter of the RFS100.1 model

RFS200.1 Model

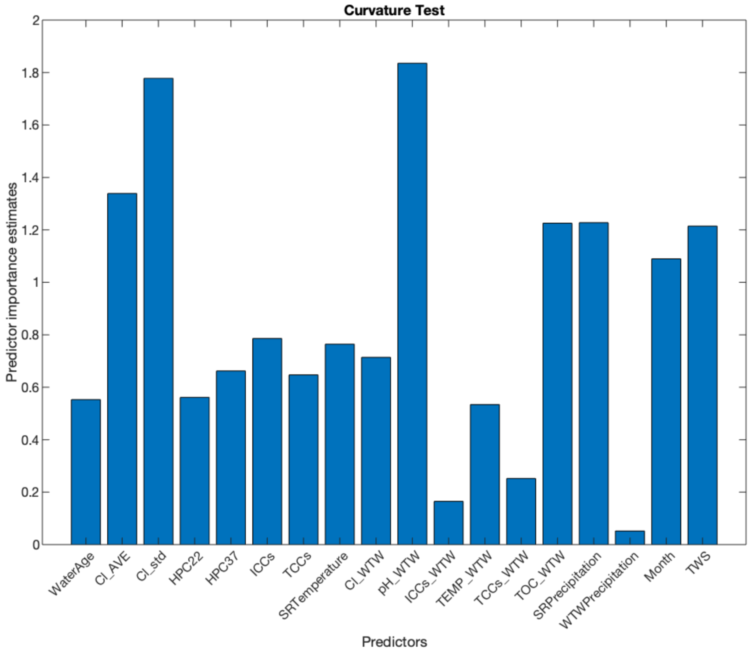
The 5 most important parameters are: average free chlorine per month, SR Name, Month of the year, average pH in the WTWs per month, free chlorine standard deviation per month



Performance importance of each parameter of the RFS200.1 model

RFA100.1 Model

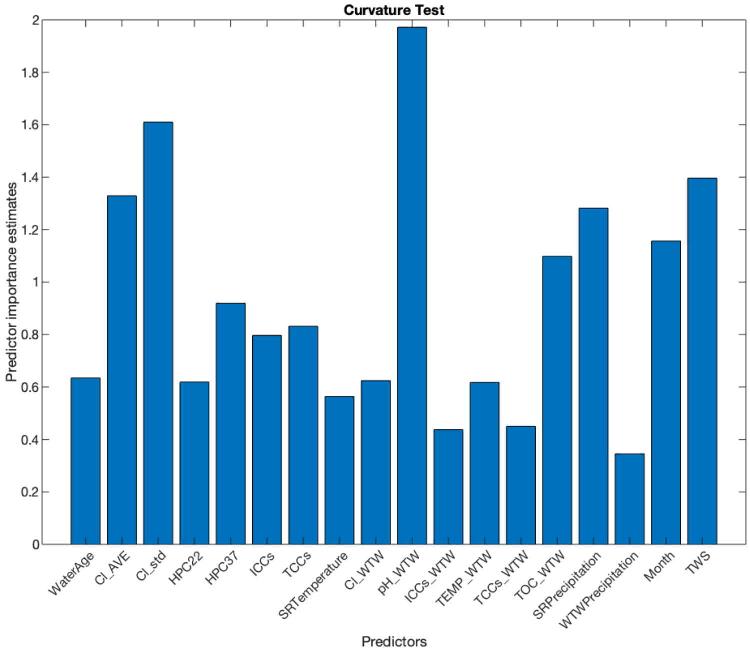
The 5 most important parameters are: average free chlorine per month, SR Name, average pH in the WTWs per month, free chlorine standard deviation per month, average daily precipitation in the SRs per month



Performance importance of each parameter of the RFA100.1 model

RFA200.1 Model

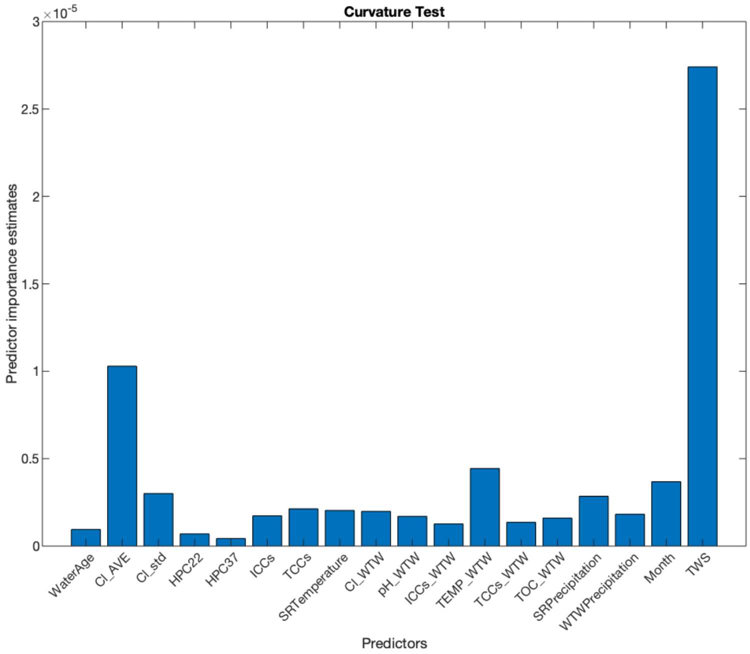
The 5 most important parameters are: average free chlorine per month, SR Name, average pH in the WTWs per month, free chlorine standard deviation per month, average daily precipitation in the SRs per month



Performance importance of each parameter of the RFA200.1 model

AB.1 model

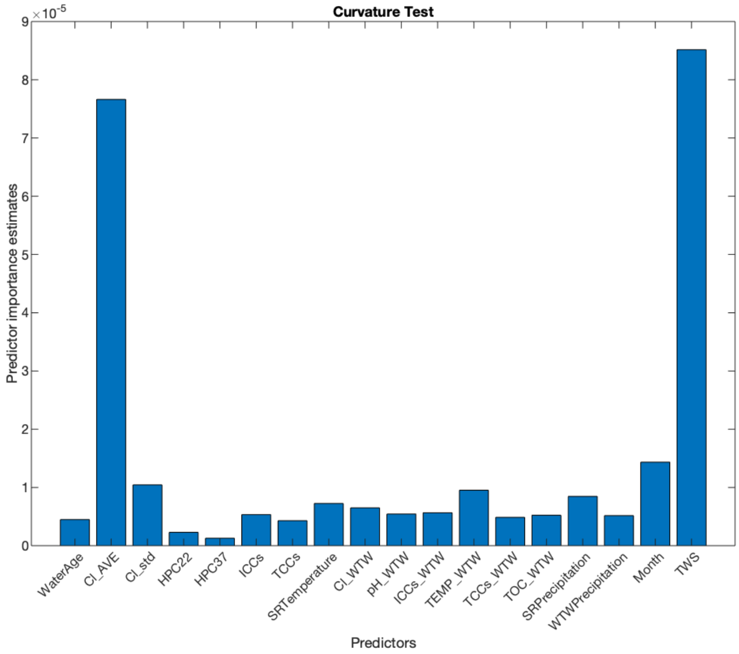
The 5 most important parameters are: average free chlorine per month, SR Name, Month of the year, average temperature in the WTWs per month, free chlorine standard deviation per month



Performance importance of each parameter of the AB.1 model

RB.1 model

The 5 most important parameters are: average free chlorine per month, SR Name, Month of the year, average temperature in the WTWs per month, free chlorine standard deviation per month



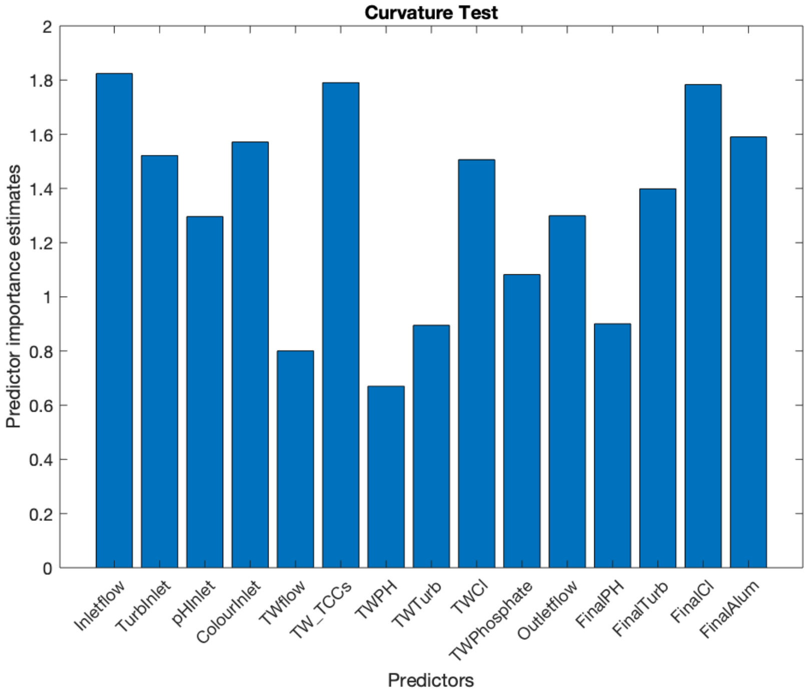
Performance importance of each parameter of the RB.1 model

APPENDIX B: Additional tables and graphs for chapter 9

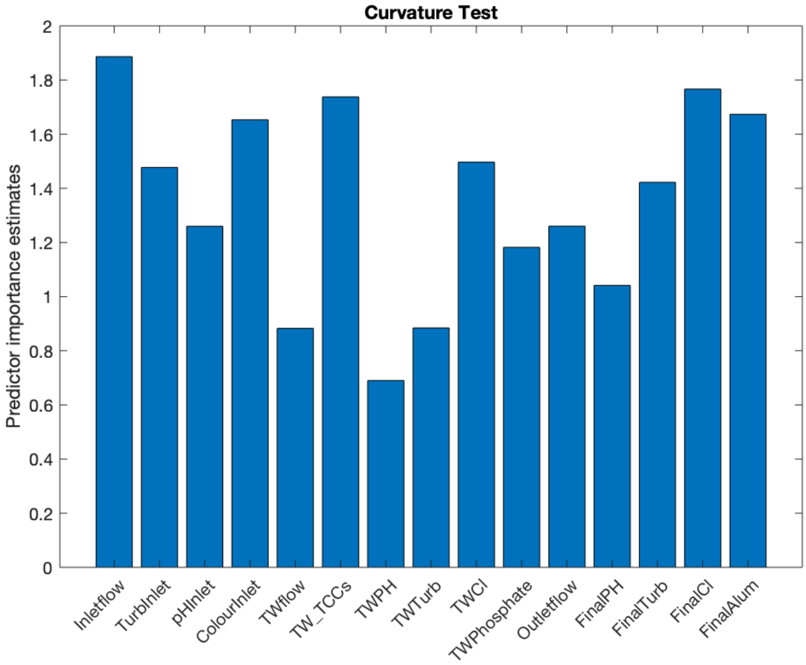
The TCCs predictive model presented in chapter 9 was tested in four out of the available 20 folds that were selected randomly. These folds were the 5th fold, the 8th fold, the 12th fold, and the 15th fold. This appendix contains output graphs and performance metrics tables that are not presented in the main chapter. More specifically, in this appendix, the followings are included:

- the predictors' performance graphs of the RF-All-12 model in the 8th,the the 12th and the 15th folds
- the performance metrics tables for both the regression and the classification approaches for each one of the 4 folds separately
- the predicted by the regression models outputs vs the observed TCCs for the 5th, the 8th and the 12th fold

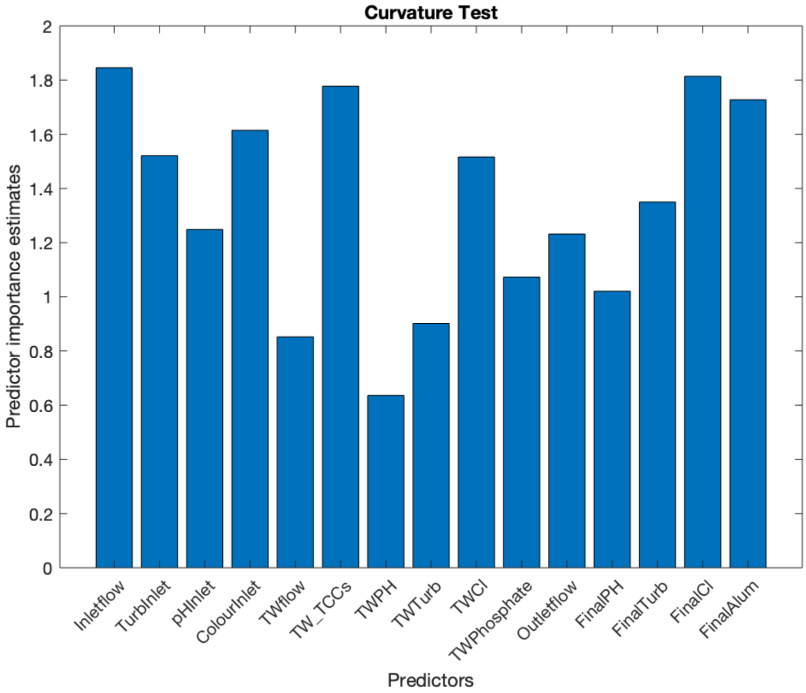
Predictors' performance graphs for the RF-All-12 model



Performance importance of each parameter of the RF-All-12 model in the 8h fold



Performance importance of each parameter of the RF-All-12 model in the 12th fold



Performance importance of each parameter of the RF-All-12 model in the 15th fold

Regression models performance metrics

Summary of the regression models' performance metrics in the 5th fold

Model NAME	R2	MSEn	RMSEn	NMSE
RF - All - 12	0.89	0.12	0.35	0.12
ANN - All - 12	0.84	0.17	0.41	0.17
LSTM - All - 12 *	0.84	0.17	0.41	0.17
CM - All - 12 *	0.88	0.13	0.36	0.17
RF - RF8 - 12	0.9	0.11	0.34	0.11
ANN - RF8 - 12	0.79	0.21	0.46	0.21
LSTM - RF8 - 12 **	0.84	0.17	0.41	0.16
CM - RF8 - 12	0.88	0.13	0.36	0.18
RF - SOMs - 23	0.87	0.14	0.37	0.14
ANN - SOMs - 23	0.76	0.21	0.46	0.21
LSTM - SOMs - 23^	0.8	0.2	0.45	0.2
CM - SOMs - 23	0.86	0.15	0.39	0.22

Summary of the regression models' performance metrics in the 8th fold

Model NAME	R2	MSEn	RMSEn	NMSE
RF - All - 12	0.9	0.1	0.32	0.1
ANN - All - 12	0.85	0.16	0.4	0.16
LSTM - All - 12 *	0.84	0.17	0.41	0.16
CM - All - 12 *	0.88	0.12	0.35	0.14
RF - RF8 - 12	0.91	0.09	0.3	0.09
ANN - RF8 - 12	0.84	0.17	0.41	0.16
LSTM - RF8 - 12 **	0.8	0.2	0.45	0.2
CM - RF8 - 12	0.88	0.12	0.35	0.14
RF - SOMs - 23	0.89	0.11	0.34	0.11
ANN - SOMs - 23	0.82	0.19	0.43	0.18
LSTM - SOMs - 23^	0.81	0.19	0.44	0.19
CM - SOMs - 23	0.88	0.13	0.36	0.15

Summary of the regression models' performance metrics in the 12th fold

Model NAME	R2	MSEn	RMSEn	NMSE
RF - All - 12	0.91	0.09	0.3	0.09
ANN - All - 12	0.82	0.18	0.42	0.18
LSTM - All - 12 *	0.84	0.15	0.39	0.16
CM - All - 12 *	0.88	0.12	0.34	0.15
RF - RF8 - 12	0.91	0.088	0.29	0.088
ANN - RF8 - 12	0.82	0.18	0.43	0.18
LSTM - RF8 - 12 **	0.84	0.16	0.4	0.16
CM - RF8 - 12	0.89	0.11	0.34	0.14
RF - SOMs - 23	0.88	0.11	0.34	0.12
ANN - SOMs - 23	0.81	0.2	0.44	0.21
LSTM - SOMs - 23^	0.82	0.19	0.43	0.19
CM - SOMs - 23	0.87	0.13	0.37	0.18

Summary of the regression models' performance metrics in the 15th fold

Model NAME	R2	MSEn	RMSEn	NMSE
RF - All - 12	0.91	0.09	0.32	0.1
ANN - All - 12	0.85	0.15	0.38	0.15
LSTM - All - 12 *	0.86	0.14	0.37	0.14
CM - All - 12 *	0.9	0.1	0.32	0.14
RF - RF8 - 12	0.91	0.09	0.31	0.1
ANN - RF8 - 12	0.82	0.18	0.43	0.19
LSTM - RF8 - 12 **	0.78	0.22	0.47	0.23
CM - RF8 - 12	0.88	0.13	0.36	0.19
RF - SOMs - 23	0.88	0.12	0.35	0.13
ANN - SOMs - 23	0.84	0.17	0.41	0.16
LSTM - SOMs - 23^	0.82	0.19	0.43	0.19
CM - SOMs - 23	0.89	0.12	0.35	0.17

* LSTM - All - 12: 1 LSTM layers,25 units per layer,0.001 Initial learning rate

** LSTM - RF8 - 12: 1 LSTM layers,16 units per layer,0.001 Initial learning rate

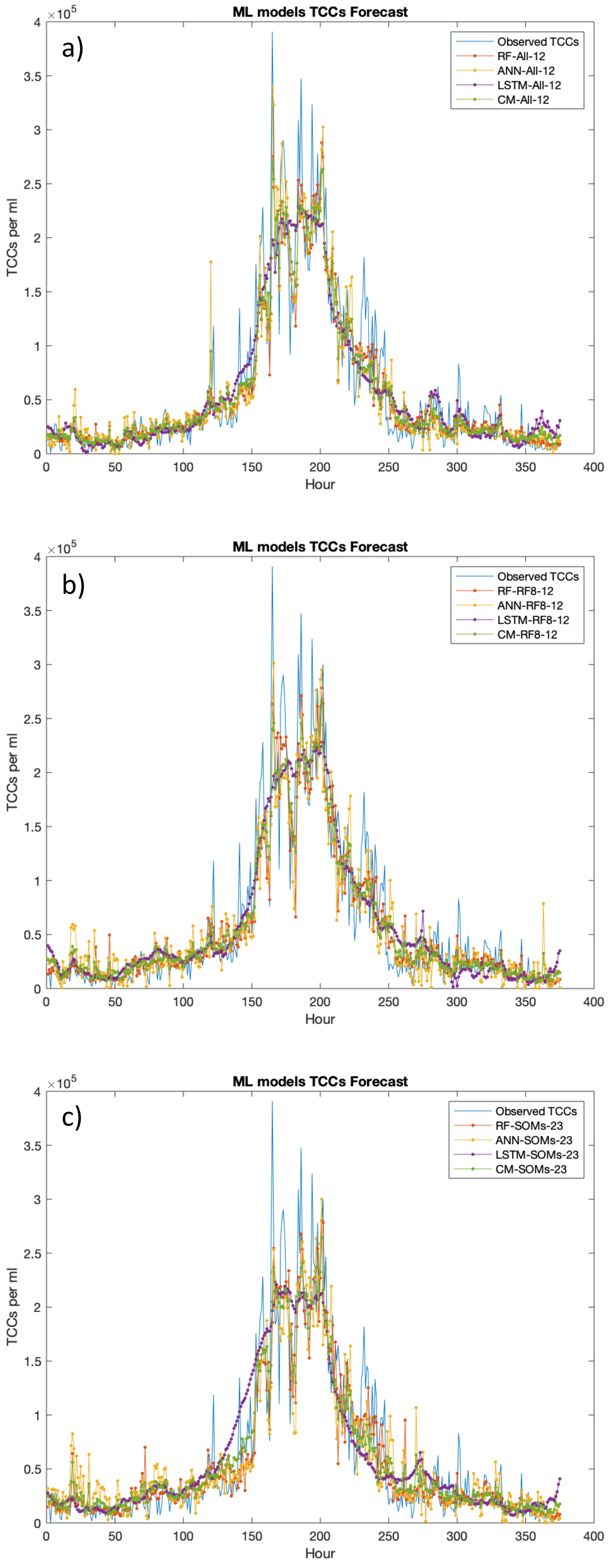
^ LSTM - SOMs - 23: 1 LSTM layer,25 units per layer,0.0012 Initial learning rate

MSE and RMSE are calculated for the normalized outputs and therefore they are unitless

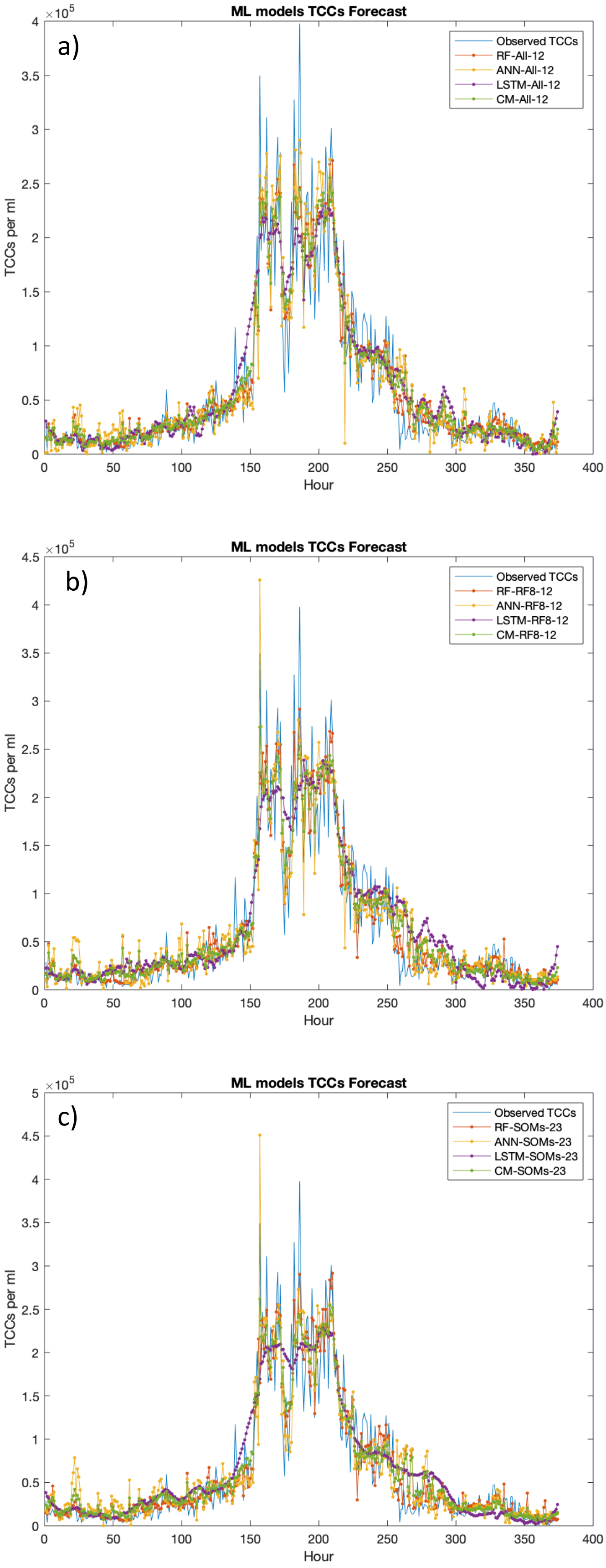
RF8: inlet flow, inlet turbidity, inlet colour, treated water TCCs,treated water turbidity, final water turbidity, final water chlorine, fina water aluminium

SOMs: inlet flow, inlet turbidity, inlet colour, treated water TCCs, treated water free chlorine

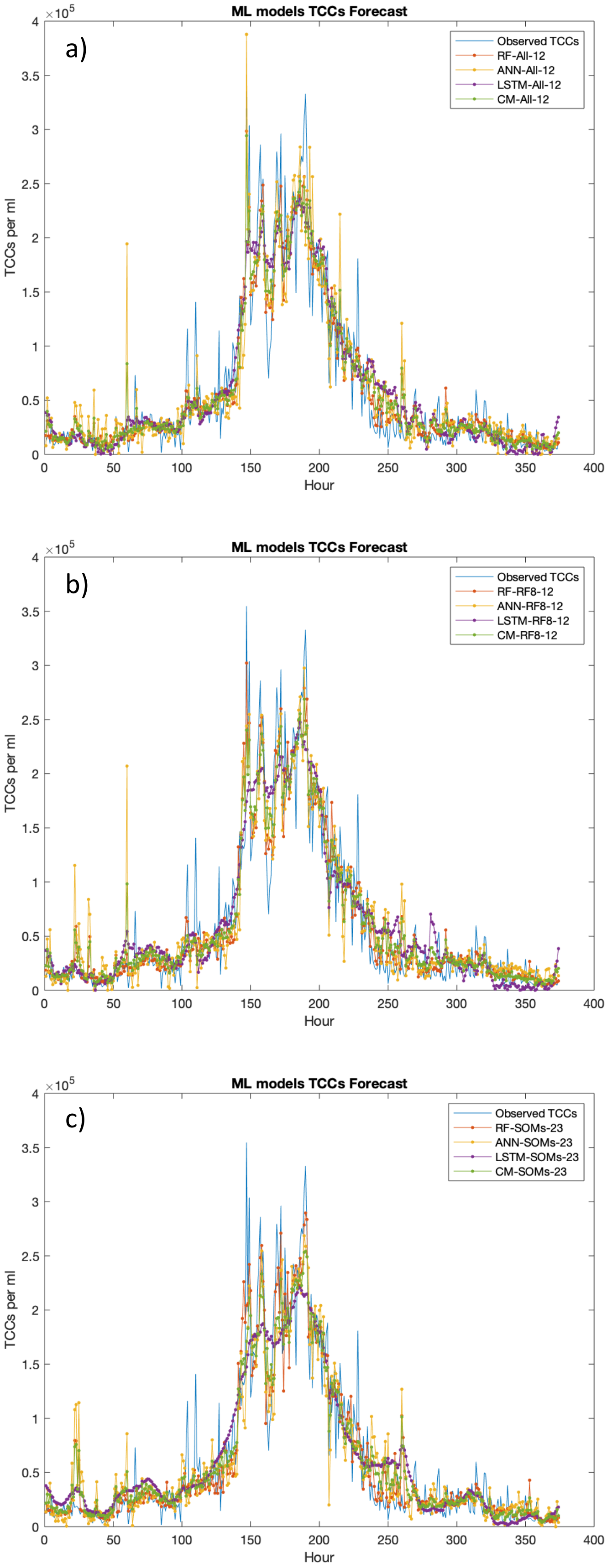
Predicted by the regression models outputs vs the observed TCCs graphs



5th fold Observed TCCs vs predicted time series of all the models when a) all the water parameters were used b) the RF 8 parameters are used and c) the SOMs 5 parameters are used



8th fold Observed TCCs vs predicted time series of all the models when a) all the water parameters were used b) the RF 8 parameters are used and c) the SOMs 5 parameters are used



12th fold Observed TCCs vs predicted time series of all the models when a) all the water parameters were used b) the RF 8 parameters are used and c) the SOMs 5 parameters are used

Classification models performance metrics

Summary of the classification models' performance metrics in the 5th fold

Model NAME	Accuracy	High risk Recall	Macro - recall	Macro - Precision	Macro - F1
RF - All - 12	78.3%	90%	68%	72%	0.7
RB - All - 12	76.2%	88%	70%	71%	0.71
LSTM - All - 12 *	64.50%	76%	56%	61%	0.61
CM - All - 12 *	77.60%	78%	72%	74%	0.744
RF - RF8 - 12	78.40%	92%	69%	72%	0.71
RB - RF8 - 12	78.1%	92%	73%	72%	0.73
LSTM - RF8 - 12 **	71.20%	83%	63%	69%	0.673
CM - RF8 - 12	79.47%	86%	77%	75%	0.754
RF - SOMs - 23	76.80%	90%	91%	75%	0.74
RB - SOMs - 23	75%	92%	72%	70%	0.72
LSTM - SOMs - 23^	68%	99%	57%	54%	0.535
CM - SOMs - 23	74.40%	90%	70%	69%	0.7

Summary of the classification models' performance metrics in the 8th fold

Model NAME	Accuracy	High risk Recall	Macro - recall	Macro - Precision	Macro - F1
RF - All - 12	79.6%	97%	70%	74%	0.7
RB - All - 12	74.6%	87%	68%	68%	0.68
LSTM - All - 12 *	70.6%	90%	64%	67%	0.6722
CM - All - 12 *	79.70%	91%	75%	75%	0.7564
RF - RF8 - 12	79.1%	93%	72%	74%	0.72
RB - RF8 - 12	74.5%	90%	69%	68%	0.68
LSTM - RF8 - 12 **	65.50%	80%	57%	60%	0.616
CM - RF8 - 12	72.10%	79%	70%	69%	0.708
RF - SOMs - 23	74.80%	93%	66%	68%	0.66
RB - SOMs - 23	73%	94%	68%	67%	0.67
LSTM - SOMs - 23^	70%	90%	61%	64%	0.5667
CM - SOMs - 12	74.86%	87%	72%	70%	0.7115

Summary of the classification models' performance metrics in the 12th fold

Model NAME	Accuracy	High risk Recall	Macro - recall	Macro - Precision	Macro - F1
RF - All - 12	79.9%	94%	71%	81%	0.74
RB - All - 12	76.7%	91%	73%	72%	0.71
LSTM - All - 12 *	72.5%	84%	65%	68%	0.69
CM - All - 12 *	79.10%	91%	72%	73%	0.737
RF - RF8 - 12	78.3%	92%	69%	72%	0.71
RB - RF8 - 12	77.2%	91%	73%	72%	0.717
LSTM - RF8 - 12 **	70.90%	93%	62%	70%	0.633
CM - RF8 - 12	77.01%	91%	70%	70%	0.7119
RF - SOMs - 23	76.50%	94%	68%	73%	0.69
RB - SOMs - 23	75%	91%	70%	70%	0.7
LSTM - SOMs - 23^	69%	85%	56%	54%	0.55
CM - SOMs - 12	75.70%	91%	67%	70%	0.6965

Summary of the classification models' performance metrics in the 15th fold

Model NAME	Accuracy	High risk Recall	Macro - recall	Macro - Precision	Macro - F1
RF - All - 12	81.5%	92%	72%	78%	0.733
RB - All - 12	78.0%	92%	75%	73%	0.7404
LSTM - All - 12 *	72.0%	81%	62%	66%	0.632
CM - All - 12 *	76.61%	84%	68%	69%	0.6866
RF - RF8 - 12	80.9%	92%	72%	75%	0.719
RB - RF8 - 12	78.0%	88%	75%	73%	0.729
LSTM - RF8 - 12 **	71.20%	87%	61%	63%	0.61
CM - RF8 - 12	77.69%	90%	69%	69%	0.6945
RF - SOMs - 23	75.00%	91%	66%	68%	0.66
RB - SOMs - 23	70%	87%	66%	64%	0.64
LSTM - SOMs - 23^	71%	99%	60%	63%	0.5708
CM - SOMs - 12	73.66%	91%	67%	66%	0.6681

* LSTM - All - 12: 1 LSTM layers,25 units per layer,0.001 Initial learning rate

** LSTM - RF8 - 12: 1 LSTM layers,16 units per layer,0.001 Initial learning rate

^ LSTM - SOMs - 23: 1 LSTM layer,23 units per layer,0.0012 Initial learning rate

REFERENCES

Abba, S. I., Quoc Bao Pham, A. G. Usman, Nguyen Thi Thuy Linh, D. S. Aliyu, Quyen Nguyen, and Quang Vu Bach. 2020. 'Emerging Evolutionary Algorithm Integrated with Kernel Principal Component Analysis for Modeling the Performance of a Water Treatment Plant'. *Journal of Water Process Engineering* 33 (October 2019). https://doi.org/10.1016/j.jwpe.2019.101081.

Abdullah, Mohammad Fikry, Mohd Zaki, Mat Amin, Mohd Fauzi Mohamad, and M Marini. 2018. 'N-HyDAA - Big Data Analytics for Malaysia Climate Change Knowledge Management'. Proceedings *of 13th International Conference on Hydroinformatics, Palermo, Italy*. Palermo.

Aggarwal, Charu. 2018. *Neural Networks and Deep Learning: A Text Book*. *Artificial Intelligence*. New York: Springer US. https://doi.org/10.1007/978-3-319-94463-0.

Ahmed, Imran, Misbah Ahmad, Gwanggil Jeon, and Francesco Piccialli. 2021. 'A Framework for Pandemic Prediction Using Big Data Analytics'. *Big Data Research* 25. https://doi.org/10.1016/j.bdr.2021.100190.

Alfonso, Leonardo, Han Wang, and Schalk Jan Van Andel. 2018. 'Machine Learning and Behavioral Economics to Simulate Flood Early Warning Decisions'. Proceedings *of 13th International Conference on Hydroinformatics, Palermo, Italy*.

Allen, Martin J., Stephen C. Edberg, and Donald J. Reasoner. 2004. 'Heterotrophic Plate Count Bacteria - What Is Their Significance in Drinking Water?' *International Journal of Food Microbiology* 92 (3): 265–74. https://doi.org/10.1016/j.ijfoodmicro.2003.08.017.

Alpaydin, Ethem. 2014. *Introduction to Machine Learning*. Third Edit. Cambridge,Massachusetts: The MIT Press, Massachusetts Institute of Technology.

American Water Works Association (AWWA). 2013. 'Nitrification Prevention and Control in Drinking Water'. Denver, Colorado.

Asadi, Ali, Anoop Verma, Kai Yang, and Ben Mejabi. 2017. 'Wastewater Treatment Aeration Process Optimization: A Data Mining Approach'. *Journal of Environmental Management* 203: 630–39. https://doi.org/10.1016/j.jenvman.2016.07.047.

ASCE. 1993. 'CRITERIA FOR EVALUATION OF WATERSHED MODELS' 119 (3): 429–42. https://doi.org/https://doi.org/10.1061/(ASCE)0733-9437(1993)119:3(429).

Assem, H., S. Ghariba, G. Makrai, P. Johnston, L. Gill, and F. Pilla. 2017. 'Urban Water Flow and Water Level Prediction Based on Deep Learning'. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 10536 LNAI. https://doi.org/10.1007/978-3-319-71273-4_26.

Astel, A., S. Tsakovski, P. Barbieri, and V. Simeonov. 2007. 'Comparison of Self-Organizing Maps Classification Approach with Cluster and Principal Components Analysis for Large Environmental Data Sets'. *Water Research* 41 (19): 4566–78. https://doi.org/10.1016/j.watres.2007.06.030.

Baldi, Pierre, Søren Brunak, Yves Chauvin, Claus A.F. Andersen, and Henrik Nielsen. 2000. 'Assessing the Accuracy of Prediction Algorithms for Classification: An Overview'. *Bioinformatics* 16 (5): 412–24. https://doi.org/10.1093/bioinformatics/16.5.412.

Bartram, J., J. Cotruvo, M. Exner, C. Fricker, and Axel Glasmacher. 2004. 'Heterotrophic Plate Count Measurement in Drinking Water Safety Management: Report of an Expert Meeting Geneva, 24-25 April 2002'. *International Journal of Food Microbiology* 92 (3): 241–47. https://doi.org/10.1016/j.ijfoodmicro.2003.08.005.

Barzegar, Rahim, Mohammad Taghi Aalami, and Jan Adamowski. 2020. 'Short-Term Water Quality Variable Prediction Using a Hybrid CNN–LSTM Deep Learning Model'. *Stochastic Environmental Research and Risk Assessment* 34 (2): 415–33. https://doi.org/10.1007/s00477-020-01776-2.

Bengio, Yoshua. 2009. *Learning Deep Architectures for AI*. *Foundations and Trends in Machine Learning*. Vol. 2. https://doi.org/10.1561/2200000006.

Berry, David, Chuanwu Xi, and Lutgarde Raskin. 2006. 'Microbial Ecology of Drinking Water Distribution Systems'. *Current Opinion in Biotechnology* 17 (3): 297–302. https://doi.org/10.1016/j.copbio.2006.05.007.

Besner, Marie Claude, Vincent Gauthier, Pierre Servais, and Anne Camper. 2002. 'Explaining the Occurrence of Coliforms in Distribution Systems'. *Journal / American Water Works Association* 94 (8): 95–109. https://doi.org/10.1002/j.1551-8833.2002.tb09529.x.

Besner, Marie Claude, Michèle Prévost, and Stig Regli. 2011. 'Assessing the Public Health Risk of Microbial Intrusion Events in Distribution Systems: Conceptual Model, Available Data, and Challenges'. *Water Research* 45 (3): 961–79. https://doi.org/10.1016/j.watres.2010.10.035.

Bishop, Christopher. 2006. *Pattern Recognition and Machine Learning*. First. Cambridge, UK: Springer US. https://doi.org/10.1021/jo01026a014.

Blokker, E.J., William Furnass, John Machell, Stephen Mounce, Peter Schaap, and Joby Boxall. 2016. 'Relating Water Quality and Age in Drinking Water Distribution Systems Using Self-Organising Maps'. *Environments* 3 (2): 10. https://doi.org/10.3390/environments3020010.

Blokker, M., J. Vreeburg, and V. Speight. 2014. 'Residual Chlorine in the Extremities of the Drinking Water Distribution System: The Influence of Stochastic Water Demands'. *Procedia Engineering* 70: 172–80. https://doi.org/10.1016/j.proeng.2014.02.020.

Boussaada, Zina, Octavian Curea, Ahmed Remaci, Haritza Camblong, and Najiba Mrabet Bellaaj. 2018. 'A Nonlinear Autoregressive Exogenous (NARX) Neural Network Model for the Prediction of the Daily Direct Solar Radiation'. *Energies* 11 (3). https://doi.org/10.3390/en11030620.

Boxall, J. B., P. J. Skipworth, and A. J. Saul. 2003. 'Aggressive Flushing for Discolouration Event Mitigation in Water Distribution Networks'. *Water Science and Technology: Water Supply* 3 (1–2): 179–86. https://doi.org/10.2166/ws.2003.0101.

Breiman, L. 2001. 'Random Forests'. *Machine Learning* 45: 5–32. https://doi.org/10.3390/rs10060911.

Breiman, L., J. Friedman, R. Olshen, and C. Stone. 1984. *Classification and Regression Trees*. Boca Raton: Taylor & Francis.

Camper, Anne. 2014. 'Organic Matter, Pipe Materials, Disinfectants and Biofilms in Distribution Systems'. In *Microbial Growth in Drinking-Water Supplies: Problems, Causes, Control and Research Needs*, edited by Paul W J J; van der Wielen and Dick; van der Kooij, 73–94. London: IWA Publications.

Carreño-Alvarado, Elizabeth Pauline, Gilberto Reynoso-meza, Idel Montalvo, and Joaquin Izquierdo. 2017. 'A Comparison of Machine Learning Classifiers for Leak Detection and Isolation in Urban Networks'. In *Congress on Numerical Methods in Engineering, 3-5 July 2017, Valencia,Spain*. Valencia. https://www.researchgate.net/publication/318275002_A_comparison_of_machine_learning_classifiers_for_leak_detection_and_isolation_in_urban_networks.

Chandarana, Parth, and M. Vijayalakshmi. 2014. 'Big Data Analytics Frameworks'. *2014*

International Conference on Circuits, Systems, Communication and Information Technology Applications, CSCITA 2014, 430–34. https://doi.org/10.1109/CSCITA.2014.6839299.

Chang, Kui, Jin Liang Gao, Wen Yan Wu, and Yi Xing Yuan. 2011. ‘Water Quality Comprehensive Evaluation Method for Large Water Distribution Network Based on Clustering Analysis’. *Journal of Hydroinformatics* 13 (3): 390. https://doi.org/10.2166/hydro.2011.021.

Chapelle, Olivier, Bernhard Scholkopf, and Alexander Zien. 2006. ‘Semi-Supervised Learning’. In *Semi-Supervised Learning*, edited by Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien, 1–13. Cambridge,Massachusetts: Massachusetts Institute of Technology.

Chawla, Nitesh, Kevin Bowyer, Lawrence Hall, and W. Philip Kegelmeyer. 2002. ‘SMOTE: Synthetic Minority Over-Sampling Technique’. *Journal of Artificial Intelligence Research* 16: 321–57. https://doi.org/10.1613/jair.953.

Che, Zhengping, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. 2018. ‘Recurrent Neural Networks for Multivariate Time Series with Missing Values’. *Scientific Reports* 8 (1): 1–12. https://doi.org/10.1038/s41598-018-24271-9.

Chen, Guoqiang, Tianyu Long, Jiangong Xiong, and Yun Bai. 2017. ‘Multiple Random Forests Modelling for Urban Water Consumption Forecasting’. *Water Resources Management* 31 (15): 4715–29. https://doi.org/10.1007/s11269-017-1774-7.

Chen, Tianqi, and Carlos Guestrin. 2016. ‘XGBoost: A Scalable Tree Boosting System’. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,13-17-August-2016*, 785–94. Association for Computing Machinery. https://doi.org/10.1145/2939672.2939785.

Cook, D. M., and J. B. Boxall. 2011. ‘Discoloration Material Accumulation in Water Distribution Systems’. *Journal of Pipeline Systems Engineering and Practice* 2 (4): 113–22. https://doi.org/10.1061/(ASCE)PS.1949-1204.0000083.

Cortes, Corinna, and Vladimir Vapnik. 1995. ‘Support-Vector Networks’. *Machine Learning* 20 (3): 273–97. https://doi.org/10.1023/A:1022627411411.

Council of the European Communities. 1998. ‘Council Directive 98/83/EC on the Quality of Water Intended for Human Consumption’.

Curriero, Frank C, Jonathan A Patz, Joan B Rose, and Subhash Lele. 2001. ‘The Association Between Extreme Precipitation and Waterborne Disease Outbreaks in the United States , 1948 – 1994’. *American Journal of Public Health* 91 (8): 1194–99. https://doi.org/10.2105/AJPH.91.8.1194.

Dai, Hong Jie, and Chen Kai Wang. 2019. ‘Classifying Adverse Drug Reactions from Imbalanced Twitter Data’. *International Journal of Medical Informatics* 129 (April): 122–32. https://doi.org/10.1016/j.ijmedinf.2019.05.017.

Dairi, Abdelkader, Tuoyuan Cheng, Fouzi Harrou, Ying Sun, and Tor Ove Leiknes. 2019. ‘Deep Learning Approach for Sustainable WWTP Operation: A Case Study on Data-Driven Influent Conditions Monitoring’. *Sustainable Cities and Society* 50 (June): 101670. https://doi.org/10.1016/j.scs.2019.101670.

Dietterich, Thomas G. 2000. ‘An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees’. *Machine Learning* 40: 139–57. https://doi.org/10.1023/A:1007607513941.

Dimokas, Nikos, Dimitris Margaritis, Manuel Gaetani, Kerem Koprubasi, and Evangelos Bekiaris. 2020. ‘A Big Data Application for Low Emission Heavy Duty Vehicles’.

Transport and Telecommunication 21 (4): 265–74. https://doi.org/10.2478/ttj-2020-0021.

Doronina, A. V., S. P. Husband, J. B. Boxall, and V. L. Speight. 2020. 'The Operational Value of Inlet Monitoring at Service Reservoirs'. *Urban Water Journal* 17 (8): 735–44. https://doi.org/10.1080/1573062X.2020.1787471.

Douterelo, I., S. Husband, and J. B. Boxall. 2014. 'The Bacteriological Composition of Biomass Recovered by Flushing an Operational Drinking Water Distribution System'. *Water Research* 54: 100–114. https://doi.org/10.1016/j.watres.2014.01.049.

Douterelo, I., R. L. Sharpe, and J. B. Boxall. 2013. 'Influence of Hydraulic Regimes on Bacterial Community Structure and Composition in an Experimental Drinking Water Distribution System'. *Water Research* 47 (2): 503–16. https://doi.org/10.1016/j.watres.2012.09.053.

Du, Jing, Biao Kuang, and Yifan Yang. 2019. 'A Data-Driven Framework for Smart Urban Domestic Wastewater: A Sustainability Perspective'. *Advances in Civil Engineering* 2019. https://doi.org/10.1155/2019/6530626.

DWI. 2016. 'The Water Supply (Water Quality) Regulations 2016,Statutory Instruments (England and Wales) No.614'. http://www.legislation.gov.uk/uksi/2016/614/made/data.pdf.

———. 2017. 'Report of the Drinking Water Inspectorate's Investigation into the Cryptosporidium Contamination of Franklaw Treatment Works in August 2015'. http://www.dwi.gov.uk/press-media/press-releases/Franklaw_Final_Report.pdf.

DWQR. 2014. 'The Public Water Supplies (Scotland) Regulations 2014,Scottish Statutory Instruments No. 364'.

———. 2017. 'Drinking Water Quality in Scotland 2016', 30.

———. 2019a. 'Drinking Water Quality in Scotland 2018: Public Water Supply'.

———. 2019b. 'Incident Summary Bradan Regulatory Supply Zones Iron Failures April 2018 to Present'. https://dwqr.scot/media/42736/dwqr-incident-assessment-bradan-zones-9501-april-2018-to-present.pdf.

Dyksen, John E., Catherine Spencer, Robert Hoehn, Jonathan Clement, Jessica Brandt-Edwards, Melinda Friedman, Amie Hanson, et al. 2008. 'Long-Term Effects of Disinfection Changes on Distribution System Water Quality'. *AWWA Research Foundation Environmental Protection Agency*.

Dykstra, Trevor et al. 2007. 'Impact of Secondary Disinfection on Corrosion in a Model Water Distribution System'. *Journal of Water Supply: Research and Technology - AQUA* 6: 147–55. https://doi.org/10.1139/s06-046.

Eck, Douglas, and Jürgen Schmidhuber. 2002. 'Learning the Long-Term Structure of the Blues'. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 2415 LNCS: 284–89. https://doi.org/10.1007/3-540-46084-5_47.

Edwards, Marc, and Abhijeet Dudi. 2004. 'Role of Chlorine and Chloramine in Corrosion of Lead-Bearing Plumbing Materials'. *Journal - American Water Works Association* 96 (OCTOBER): 69–81.

Edwards, Marc, Simoni Triantafyllidou, and Dana Best. 2009. 'Elevated Blood Lead in Young Children Due to Lead-Contaminated Drinking Water: Washington, DC, 2001-2004'. *Environmental Science and Technology* 43 (5): 1618–23. https://doi.org/10.1021/es802789w.

Edzwald, J. 2011. *WATER QUALITY AND TREATMENT: A HANDBOOK ON DRINKING WATER*.

6th ed. New York: American Water Works Association, American Society of Civil Engineers, McGraw-Hill.

Ellis, Donald, Christian Bouchard, and Gaetan Lantagne. 2000. ‘Removal of Iron and Manganese from Groundwater by Oxidation and Microfiltration’. *Desalination* 130 (3): 255–64. https://doi.org/10.1016/S0011-9164(00)00090-4.

Ellis, K., S. R. Mounce, B. Ryan, C. A. Biggs, and M. R. Templeton. 2015. ‘Improving Root Cause Analysis of Bacteriological Water Quality Failures at Water Treatment Works’. *Procedia Engineering* 119 (1): 309–18. https://doi.org/10.1016/j.proeng.2015.08.890.

Ellis, K., S. R. Mounce, B. Ryan, M. R. Templeton, and C. A. Biggs. 2014. ‘Use of On-Line Water Quality Monitoring Data to Predict Bacteriological Failures’. *Procedia Engineering* 70 (0): 612–21. https://doi.org/10.1016/j.proeng.2014.02.067.

Ellis, Kathryn. 2013. ‘IMPROVING ROOT CAUSE ANALYSIS OF BACTERIOLOGICAL WATER QUALITY FAILURES’. Sheffield.

Ennett, Colleen M., Monique Frize, and C. Robin Walker. 2001. ‘Influence of Missing Values on Artificial Neural Network Performance’. *Studies in Health Technology and Informatics* 84: 449–53. https://doi.org/10.3233/978-1-60750-928-8-449.

Ester, M, H-P Kriegel, J Sander, and X X. 1996. ‘A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise Martin’. *AAAI Press*, 226–31. https://doi.org/10.1016/B978-044452701-1.00067-3.

Fellini, Sofia, Riccardo Vesipa, Fulvio Boano, and Luca Ridolfi. 2018. ‘Real-Time Measurement Fault Detection and Remote-Control in a Mountain Water Supply System’. Proceedings *of 13th International Conference on Hydroinformatics, Palermo, Italy*. Palermo.

Filipe, Jorge, Ricardo J. Bessa, Marisa Reis, Rita Alves, and Pedro Póvoa. 2019. ‘Data-Driven Predictive Energy Optimization in a Wastewater Pumping Station’. *Applied Energy* 252 (February): 113423. https://doi.org/10.1016/j.apenergy.2019.113423.

Fischer, Thomas, and Christopher Krauss. 2018. ‘Deep Learning with Long Short-Term Memory Networks for Financial Market Predictions’. *European Journal of Operational Research* 270 (2): 654–69. https://doi.org/10.1016/j.ejor.2017.11.054.

Fish, K., A. M. Osborn, and J. B. Boxall. 2017. ‘Biofilm Structures (EPS and Bacterial Communities) in Drinking Water Distribution Systems Are Conditioned by Hydraulics and Influence Discolouration’. *Science of the Total Environment* 593–594: 571–80. https://doi.org/10.1016/j.scitotenv.2017.03.176.

Fish, Katherine E., Richard Collins, Nicola H. Green, Rebecca L. Sharpe, Isabel Douterelo, A. Mark Osborn, and Joby B. Boxall. 2015. ‘Characterisation of the Physical Composition and Microbial Community Structure of Biofilms within a Model Full-Scale Drinking Water Distribution System’. *PLoS ONE* 10 (2): 1–22. https://doi.org/10.1371/journal.pone.0115824.

Fish, Katherine E., A. Mark Osborn, and Joby Boxall. 2016. ‘Characterising and Understanding the Impact of Microbial Biofilms and the Extracellular Polymeric Substance (EPS) Matrix in Drinking Water Distribution Systems’. *Environ. Sci.: Water Res. Technol.* 2 (4): 614–30. https://doi.org/10.1039/C6EW00039H.

Fish, Katherine E., Nik Reeves-McLaren, Stewart Husband, and Joby Boxall. 2020. ‘Unchartered Waters: The Unintended Impacts of Residual Chlorine on Water Quality and Biofilms’. *Npj Biofilms and Microbiomes* 6 (34). https://doi.org/10.1038/s41522-020-00144-w.

Flemming, Hans Curt, Jost Wingender, Ulrich Szewzyk, Peter Steinberg, Scott A. Rice, and

Staffan Kjelleberg. 2016. ‘Biofilms: An Emergent Form of Bacterial Life’. *Nature Reviews Microbiology* 14 (9): 563–75. https://doi.org/10.1038/nrmicro.2016.94.

Francisque, Alex, Manuel J. Rodriguez, Luis F. Miranda-Moreno, Rehan Sadiq, and François Proulx. 2009. ‘Modeling of Heterotrophic Bacteria Counts in a Water Distribution System’. *Water Research* 43 (4): 1075–87. https://doi.org/10.1016/j.watres.2008.11.030.

Freitas, Rodrigo, Bruno Melo Brentan, Gustavo Meirelles Lima, and Edevar Luvizotto Junior. 2017. ‘WDNs Calibration Using K-Means Algorithm for Pipes Clustering and a Hybrid Model for Optimization’. In *Proceedings of the Computing and Control for the Water Industry 2017 Conference, Sheffield, UK*.

Freund, Yoav, and Robert E. Schapire. 1997. ‘A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting’. *Journal of Computer and System Sciences* 55 (1): 119–39. https://doi.org/10.1006/jcss.1997.1504.

Friedman, Jerome H. 2001. ‘Greedy Function Approximation: A Gradient Boosting Machine’. *Annals of Statistics* 29 (5): 1189–1232. https://doi.org/10.2307/2699986.

Gandomi, Amir, and Murtaza Haider. 2015. ‘Beyond the Hype: Big Data Concepts, Methods, and Analytics’. *International Journal of Information Management* 35 (2): 137–44. https://doi.org/10.1016/j.ijinfomgt.2014.10.007.

Garcia, Diego, Vicenç Puig, and Joseba Quevedo. 2020. ‘Prognosis of Water Quality Sensors Using Advanced Data Analytics: Application to the Barcelona Drinking Water Network’. *Sensors (Switzerland)* 20 (5). https://doi.org/10.3390/s20051342.

Geetha, S., P. Deepalakshmi, and Shilpa Pande. 2019. ‘Managing Crop for Indian Farming Using IOT’. *2019 International Conference on Clean Energy and Energy Efficient Electronics Circuit for Sustainable Development, INCCES 2019*. https://doi.org/10.1109/INCCES47820.2019.9167699.

Geldreich, Edwin E. 1996. ‘Pathogenic Agents in Freshwater Resources’. *Hydrological Processes* 10 (2): 315–33. https://doi.org/10.1002/(SICI)1099-1085(199602)10:2<315::AID-HYP361>3.0.CO;2-H.

Gerke, Tammie L., Brenda J. Little, and J. Barry Maynard. 2016. ‘Manganese Deposition in Drinking Water Distribution Systems’. *Science of the Total Environment* 541: 184–93. https://doi.org/10.1016/j.scitotenv.2015.09.054.

Gibbs, M. S., N. Morgan, H. R. Maier, G. C. Dandy, J. B. Nixon, and M. Holmes. 2006. ‘Investigation into the Relationship between Chlorine Decay and Water Distribution Parameters Using Data Driven Methods’. *Mathematical and Computer Modelling* 44 (5–6): 485–98. https://doi.org/10.1016/j.mcm.2006.01.007.

Gillespie, Simon, Patrick Lipphaus, James Green, Simon Parsons, Paul Weir, Kes Juskowiak, Bruce Jefferson, Peter Jarvis, and Andreas Nocker. 2014. ‘Assessing Microbiological Water Quality in Drinking Water Distribution Systems with Disinfectant Residual Using Flow Cytometry’. *Water Research* 65: 224–34. https://doi.org/10.1016/j.watres.2014.07.029.

Giustolisi, O., and D. A. Savic. 2009. ‘Advances in Data-Driven Analyses and Modelling Using EPR-MOGA’. *Journal of Hydroinformatics* 11 (3–4): 225. https://doi.org/10.2166/hydro.2009.017.

Gorchev, H G, and G Ozolins. 2011. ‘WHO Guidelines for Drinking-Water Quality.’ *WHO Chronicle* 38 (3): 104–8. https://doi.org/10.1016/S1462-0758(00)00006-6.

Gouider, Mbarka, Jalel Bouzid, Sami Sayadi, and Antoine Montiel. 2009. ‘Impact of Orthophosphate Addition on Biofilm Development in Drinking Water Distribution

Systems'. *Journal of Hazardous Materials* 167 (1–3): 1198–1202. https://doi.org/10.1016/j.jhazmat.2009.01.128.

Graves, Alex, and Jürgen Schmidhuber. 2005. 'Framewise Phoneme Classification with Bidirectional LSTM and Other Neural Network Architectures'. *Neural Networks* 18: 602–10.

Hallam, N.B, J.R West, C.F Forster, and J Simms. 2001. 'The Potential for Biofilm Growth in Water Distribution Systems'. *Water Research* 35 (17): 4063–71. https://doi.org/10.1016/S0043-1354(01)00248-2.

Hammes, Frederik, Michael Berney, Yingying Wang, Marius Vital, Oliver Köster, and Thomas Egli. 2008. 'Flow-Cytometric Total Bacterial Cell Counts as a Descriptive Microbiological Parameter for Drinking Water Treatment Processes'. *Water Research* 42 (1–2): 269–77. https://doi.org/10.1016/j.watres.2007.07.009.

Harvey, Richard, Heather M. Murphy, Edward A. McBean, and Bahram Gharabaghi. 2015. 'Using Data Mining to Understand Drinking Water Advisories in Small Water Systems: A Case Study of Ontario First Nations Drinking Water Supplies'. *Water Resources Management* 29 (14): 5129–39. https://doi.org/10.1007/s11269-015-1108-6.

Hashemi, Saeed, Yves Filion, and Vanessa Speight. 2018. 'Identification of Factors That Influence Energy Performance in Water Distribution System Mains'. *Water (Switzerland)* 10 (4): 1–16. https://doi.org/10.3390/w10040428.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2008. *The Elements of Statistical Learning*. Second. Stanford: Springer US. https://doi.org/10.1007/b94608.

He, Haibo, Yang Bai, Edwardo A. Garcia, and Shutao Li. 2008. 'ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning'. *Proceedings of the International Joint Conference on Neural Networks*, no. 3: 1322–28. https://doi.org/10.1109/IJCNN.2008.4633969.

He, Haibo, and Yunqian Ma. 2013. *Imbalanced Learning Foundations, Algorithms and Applications*. New Jersey: IEEE.

Helmi, K., A. Watt, P. Jacob, I. Ben-Hadj-Salah, A. Henry, G. Méheut, and N. Charni-Ben-Tabassi. 2014. 'Monitoring of Three Drinking Water Treatment Plants Using Flow Cytometry'. *Water Science & Technology: Water Supply* 14 (5): 850. https://doi.org/10.2166/ws.2014.044.

Helsinki University of Technology. 2015. 'SOM Toolbox (for MATLAB)'. 2015. https://github.com/ilarinieminen/SOM-Toolbox.

Herrera, Manuel, Luís Torgo, Joaquín Izquierdo, and Rafael Pérez-García. 2010. 'Predictive Models for Forecasting Hourly Urban Water Demand'. *Journal of Hydrology* 387 (1–2): 141–50. https://doi.org/10.1016/j.jhydrol.2010.04.005.

Hinton, Geoffrey, and Sam Roweis. 2002. 'Stochastic Neighbor Embedding'. *Advances in Neural Information Processing Systems* 15: 833–40.

Hitokoto, Masayuki, and Masaaki Sakuraba. 2018. 'Applicability of the Deep Learning Flood Forecast Model against the Flood Exceeding the Training Events'. *Proceeding of 13th International Conference on Hydroinformatics, Palermo, Italy*. Palermo.

Hochreiter, Sepp, and Jürgen Schmidhuber. 1997. 'Long Short-Term Memory'. *Neural Computation* 9 (8): 1735–80. https://doi.org/10.1162/neco.1997.9.8.1735.

Hu, Cheng Yu, Jun Yi Cai, De Ze Zeng, Xue Song Yan, Wen Yin Gong, and Ling Wang. 2019. 'Deep Reinforcement Learning Based Valve Scheduling for Pollution Isolation in Water Distribution Network'. *Mathematical Biosciences and Engineering : MBE* 17 (1): 105–21. https://doi.org/10.3934/mbe.2020006.

Hu, Jun, Huiyu Dong, Qiang Xu, Wencui Ling, Jiuhui Qu, and Zhimin Qiang. 2018. ‘Impacts of Water Quality on the Corrosion of Cast Iron Pipes for Water Distribution and Proposed Source Water Switch Strategy’. *Water Research* 129: 428–35. https://doi.org/10.1016/j.watres.2017.10.065.

Husband, P. S., and J. B. Boxall. 2011. ‘Asset Deterioration and Discolouration in Water Distribution Systems’. *Water Research* 45 (1): 113–24. https://doi.org/10.1016/j.watres.2010.08.021.

Husband, P. S., J. Whitehead, and J. B. Boxall. 2010. ‘The Role of Trunk Mains in Discolouration’. *Proceedings of the Institution of Civil Engineers - Water Management* 163 (8): 397–406. https://doi.org/10.1680/wama.900063.

Husband, P S, and J B Boxall. 2010. ‘Field Studies of Discoloration in Water Distribution Systems: Model Verification and Practical Implications’. *Journal of Environmental Engineering* 136 (1): 86–94. https://doi.org/10.1061/ ASCE EE.1943-7870.0000115.

Husband, S., K. E. Fish, I. Douterelo, and J. Boxall. 2016. ‘Linking Discolouration Modelling and Biofilm Behaviour within Drinking Water Distribution Systems’. *Water Science and Technology: Water Supply* 16 (4): 942–50. https://doi.org/10.2166/ws.2016.045.

IWA. 2018. ‘Digital Water :Industry Leaders Chart the Transformation Journey’. *IWA*.

———. 2019. ‘Digital_Water: Industry Leaders Chart the Transformation Journey’. *IWA Publications*. London.

Jayaweera, C. D., M. R. Othman, and N. Aziz. 2019. ‘Improved Predictive Capability of Coagulation Process by Extreme Learning Machine with Radial Basis Function’. *Journal of Water Process Engineering* 32 (September): 100977. https://doi.org/10.1016/j.jwpe.2019.100977.

Jerome, Friedman, Hastie Trevor, and Robert Tibshirani. 2000. ‘Additive Logistic Regression : A Statistical View of Boosting’. *Annals of Statistics* 28 (2): 337–74.

Jolliffe, I.T. 2002. *Principal Component Analysis*. 2nd ed. New York: Springer US. https://doi.org/10.1007/BF01884351.

Kaelbling, L.P., and M. L. Littman. 1996. ‘Reinforcement Learning: A Survey’. *Journal of Artificial Intelligence Research* 4: 1475–79. https://doi.org/10.1613/jair.301.

Kalteh, A. M., P. Hjorth, and R. Berndtsson. 2008. ‘Review of the Self-Organizing Map (SOM) Approach in Water Resources: Analysis, Modelling and Application’. *Environmental Modelling and Software* 23 (7): 835–45. https://doi.org/10.1016/j.envsoft.2007.10.001.

Kazemi, Ehsan, Stephen Mounce, Stewart Husband, and Joby Boxall. 2018. ‘Predicting Turbidity in Water Distribution Trunk Mains Using Nonlinear Autoregressive Exogenous Artificial Neural Networks’. Proceedings *of 13th International Conference on Hydroinformatics, Palermo, Italy*. Palermo.

Kerneïs, Alain, Frederique Nakache, Alain Deguin, and Max Feinberg. 1995. ‘The Effects of Water Residence Time on the Biological Quality in a Distribution Network’. *Water Research* 29 (7): 1719–27. https://doi.org/10.1016/0043-1354(94)00323-Y.

Khadse, G. K., P. M. Patni, and P. K. Labhasetwar. 2015. ‘Removal of Iron and Manganese from Drinking Water Supply’. *Sustainable Water Resources Management* 1 (2): 157–65. https://doi.org/10.1007/s40899-015-0017-4.

Kim, Sung Eun S.E., and Il Won I.W. Seo. 2015. ‘Artificial Neural Network Ensemble Modeling with Exploratory Factor Analysis for Streamflow Forecasting’. *Journal of Hydroinformatics* 17 (4): 614. https://doi.org/10.2166/hydro.2015.033.

Kohavi, Ron. 1995. ‘A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection’. In *Proceedings of the 14th International Joint Conference on Artificial*

Intelligence, 1137–1143.

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

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

Kooij, D. van der. 2014. 'Legionella in Drinking-Water Supplies'. In *Microbial Growth in Drinking-Water Supplies: Problems, Causes, Control and Research Needs*, edited by Paul W J J van der Wielen and Dick; van der Kooij, 127–75.

Kooij, Dick van der, and Paul W J J van der Wielen. 2014a. 'Microbial Growth in Drinking-Water Supplies: General Introduction'. In *Microbial Growth in Drinking-Water Supplies: Problems, Causes, Control and Research Needs*, edited by Dick; van der Kooij and Paul W J J van der Wielen, 1–32. London: IWA Publications.

———. 2014b. 'Microbial Growth in Drinking-Water Supplies: Research Needs'. In *Microbial Growth in Drinking-Water Supplies: Problems, Causes, Control and Research Needs*, edited by Dick; van der Kooij and Paul W J J van der Wielen, 423–43. London: IWA Publications.

Korth, A., C. Fiebiger, K. Bornmann, and W. Schmidt. 2004. 'NOM Increase in Drinking Water Reservoirs - Relevance for Drinking Water Production'. *Water Science and Technology: Water Supply* 4 (4): 55–60.

Kotsiantis, S.B., I.D. Zaharakis, and P.E. Pintelas. 2007. 'Machine Learning : A Review of Classification and Combining Techniques'. *Artificial Intelligence Review* 26 (2006): 159–90. https://doi.org/10.1007/s10462-007-9052-3.

Krooij, D. van der, and Paul W J J van der Wielen. 2014. *Microbial Growth in Drinking-Water Supplies: Problems, Causes, Control and Research Needs*. Edited by Dick; van der Kooij and Paul W J J van der Wielen. London: IWA Publications.

Kuhn, Max, and Kjell Johnson. 2013. *Applied Predictive Modeling*. 5th ed. New York: Springer US. https://doi.org/10.1007/978-1-4614-6849-3.

Kühnert, C., T. Bernard, I. Montalvo Arango, and R. Nitsche. 2014. 'Water Quality Supervision of Distribution Networks Based on Machine Learning Algorithms and Operator Feedback'. *Procedia Engineering* 89: 189–96. https://doi.org/10.1016/j.proeng.2014.11.176.

Kulbat, E., and A. Sokołowska. 2017. 'The Occurrence of Heavy Metals and Metal-Resistant Bacteria in Water and Bottom Sediments of the Straszyn Reservoir (Poland)'. *E3S Web of Conferences* 22. https://doi.org/10.1051/e3sconf/20172200093.

Kumpel, Emily, and Kara L. Nelson. 2013. 'Comparing Microbial Water Quality in an Intermittent and Continuous Piped Water Supply'. *Water Research* 47 (14): 5176–88. https://doi.org/10.1016/j.watres.2013.05.058.

Kuremoto, Takashi, Shinsuke Kimura, Kunikazu Kobayashi, and Masanao Obayashi. 2014. 'Time Series Forecasting Using a Deep Belief Network with Restricted Boltzmann Machines'. *Neurocomputing* 137: 47–56. https://doi.org/10.1016/j.neucom.2013.03.047.

Laucelli, Daniele, Balvant Rajani, Yehuda Kleiner, and Orazio Giustolisi. 2014. 'Study on Relationships between Climate-Related Covariates and Pipe Bursts Using Evolutionary-Based Modelling'. *Journal of Hydroinformatics* 16 (4): 743. https://doi.org/10.2166/hydro.2013.082.

LeChevallier, M. W., C. D. Cawthon, and R. G. Lee. 1988. 'Inactivation of Biofilm Bacteria'. *Applied and Environmental Microbiology* 54 (10): 2492–99.

LeChevallier, M. W., W. Schulz, and R. G. Lee. 1991. ‘Bacterial Nutrients in Drinking Water’. *Applied and Environmental Microbiology* 57 (3): 857–62.

LeChevallier, M. W., R. J. Seidler, and T. M. Evans. 1980. ‘Enumeration and Characterization of Standard Plate Count Bacteria in Chlorinated and Raw Water Supplies’. *Applied and Environmental Microbiology* 40 (5): 922–30.

LeChevallier, M. W., A. Singh, D. A. Schiemann, and G. A. McFeters. 1985. ‘Changes in Virulence of Waterborne Enteropathogens with Chlorine Injury’. *Applied and Environmental Microbiology* 50 (2): 412–19.

LeChevallier, M W, T M Babcock, and R G Lee. 1987. ‘Examination and Characterization of Distribution-System Biofilms’. *Applied and Environmental Microbiology* 53 (12): 2714–24.

Lechevallier, M W, C D Cawthon, R G Lee, Mark W Lechevallier, Cheryl D Cawthon, and Ramon G Lee. 1988. ‘Factors Promoting Survival of Bacteria in Chlorinated Water Supplies . Factors Promoting Survival of Bacteria in Chlorinated Water Supplies’ 54 (3): 649–54.

LeChevallier, Mark. 2014. ‘Measurement of Biostability and Impacts on Water Treatment in the US’. In *Microbial Growth in Drinking-Water Supplies: Problems, Causes, Control and Research Needs*, edited by Paul W J J van der Wielen and Dick; van der Kooij, 33–56. London: IWA Publications.

LeChevallier, Mark W. 1990. ‘Coliform Regrowth in Drinking Water. A Review’. *Journal / American Water Works Association* 82 (11): 74–86.

Lechevallier, Mark W., Nancy J. Welch, and Darrell B. Smith. 1996. ‘Full-Scale Studies of Factors Related to Coliform Regrowth in Drinking Water’. *Applied and Environmental Microbiology* 62 (7): 2201–11. https://doi.org/10.1128/aem.62.7.2201-2211.1996.

Lechevallier, Mark W, Richard W Gullick, Mohammad R Karim, Melinda Friedman, and James E Funk. 2003. ‘The Potential for Health Risks from Intrusion of Contmainants into the Distribution System from Pressure Transients’. *Journal of Water and Health*, 3–14.

Lecun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. ‘Deep Learning’. *Nature* 521 (7553): 436–44. https://doi.org/10.1038/nature14539.

Lehtola, Markku J, Ilkka T Miettinen, and Pertti J Martikainen. 2002. ‘Biofilm Formation in Drinking Water Affected by Low Concentrations of Phosphorus’. *Canadian Journal of Microbiology* 48 (6): 494–99. https://doi.org/10.1139/w02-048.

Li, Lei, Shuming Rong, Rui Wang, and Shuili Yu. 2021. ‘Recent Advances in Artificial Intelligence and Machine Learning for Nonlinear Relationship Analysis and Process Control in Drinking Water Treatment: A Review’. *Chemical Engineering Journal* 405 (June 2020). https://doi.org/10.1016/j.cej.2020.126673.

Liu, G., E. J. Van der Mark, J. Q. J. C. Verberk, and J. C. Van Dijk. 2013. ‘Flow Cytometry Total Cell Counts: A Field Study Assessing Microbiological Water Quality and Growth in Unchlorinated Drinking Water Distribution Systems’. *BioMed Research International* 2013: 1–10. https://doi.org/10.1155/2013/595872.

Liu, Ping, Jin Wang, Arun Sangaiah, Yang Xie, and Xinchun Yin. 2019. ‘Analysis and Prediction of Water Quality Using LSTM Deep Neural Networks in IoT Environment’. *Sustainability* 11 (7): 2058. https://doi.org/10.3390/su11072058.

Liu, Qianjie, Wei Chen, Huosheng Hu, Qingyuan Zhu, and Zhixiang Xie. 2020. ‘An Optimal NARX Neural Network Identification Model for a Magnetorheological Damper With Force-Distortion Behavior’. *Frontiers in Materials* 7 (February): 1–12. https://doi.org/10.3389/fmats.2020.00010.

Liu, Yangguang, Yangming Zhou, Shiting Wen, and Chaogang Tang. 2014. ‘A Strategy on Selecting Performance Metrics for Classifier Evaluation’. *International Journal of Mobile Computing and Multimedia Communications* 6 (4): 20–35. https://doi.org/10.4018/IJMCMC.2014100102.

Lu, Jianwei. 2020. ‘Industrial Pollution Governance Efficiency and Big Data Environmental Controlling Measures: A Case Study on Jiangsu Province, China’. *Nature Environment and Pollution Technology* 19 (4): 1743–48. https://doi.org/10.46488/NEPT.2020.v19i04.046.

Maaten, Laurens van der, and Geoffrey Hinton. 2008. ‘Visualizing Data Using T-SNE Laurens’. *Journal of Machine Learning Research* 9: 2579–2605. https://doi.org/10.1007/s10479-011-0841-3.

Maimon, Oded, and Lior Rokach. 2006. *Data Mining and Knowledge Discovery Handbook : A Complete Guide for Researchers and Practitioners*. 1st ed. Tel-Aviv: Springer US.

Mamandipoor, Behrooz, Mahshid Majd, Seyedmostafa Sheikhalishahi, Claudio Modena, and Venet Osmani. 2020. ‘Monitoring and Detecting Faults in Wastewater Treatment Plants Using Deep Learning’. *Environmental Monitoring and Assessment* 192 (2): 1–12. https://doi.org/10.1007/s10661-020-8064-1.

MathWorks. 2016. ‘Getting Started with Machine Learning’.

———. 2020. ‘Design Time Series NARX Feedback Neural Networks’. 2020. https://www.mathworks.com/help/deeplearning/ug/design-time-series-narx-feedback-neural-networks.html.

McCoy, W.F., and B.H. Olson. 1987. ‘Analysis of the Microbiological Particulates in Municipal Drinking-Water by Scanning Electron Microscopy/X-Ray Energy Spectroscopy’. *Zentralblatt Fur Bakteriologie, Mikrobiologie Und Hygiene. Serie B, Umwelthygiene, Krankenhaushygiene, Arbeitshygiene, Praventive Medizin* 183 (5–6): 511–29.

Medema, G.J., P.W.M.H Smeets, E.J.M. Blokker, and J.H.M. van Lieverloo. 2014. ‘Safe Distribution without a Disinfectant Residual’. In *Microbial Growth in Drinking-Water Supplies: Problems, Causes, Control and Research Needs*, edited by Paul W J J van der Wielen and Dick van der Kooij, 95–125. London: IWA Publications.

Met Office. 2020. ‘MIDAS Open: UK Daily Rainfall Data, V202007. Centre for Environmental Data Analysis’. 2020. https://doi.org/10.5285/ec9e894089434b03bd9532d7b343ec4b.

Meyers, Gregory, Zoran Kapelan, and Edward Keedwell. 2017. ‘Short-Term Forecasting of Turbidity in Trunk Main Networks’. *Water Research* 124: 67–76. https://doi.org/10.1016/j.watres.2017.07.035.

Miettinen, Ilkka T, and Terttu Vartiainen. 1997. ‘Phosphorus and Bacterial Growth in Drinking Water’. *Applied and Environmental Microbiology* 63 (8): 3242–45. https://doi.org/0099-2240/97/.

Mitchell, Tom. 1997. *Machine Learning*. Vol. 17. WCB/McGraw-Hill Science. https://doi.org/10.1007/978-3-642-21004-4_10.

Mohammed, Hadi, Ibrahim A. Hameed, and Razak Seidu. 2017. ‘Random Forest Tree for Predicting Fecal Indicator Organisms in Drinking Water Supply’. In *Proceedings of 4th International Conference on Behavioral, Economic, and Socio-Cultural Computing, BESC 2017*. https://doi.org/10.1109/BESC.2017.8256398.

Moore, G. F., B. C. Dunsmore, S. M. Jones, C. W. Smejkal, J. Jass, P. Stoodley, and H. M. Lappin-Scott. 2000. ‘Microbial Detachment from Biofilms’. In *Community Structure and Co-Operation in Biofilms*, edited by D. G. Allison, P. GIlbert, H. M. Lappin-Scott, and M. Wilson, 107–27. Cambridge University Press.

Morrow, J. B., J. L. Almeida, L. A. Fitzgerald, and K. D. Cole. 2008. ‘Association and Decontamination of Bacillus Spores in a Simulated Drinking Water System’. *Water Research* 42 (20): 5011–21. https://doi.org/10.1016/j.watres.2008.09.012.

Mounce, S., I. Douterelo, R. Sharpe, and J. Boxall. 2012. ‘A Bio- Hydroinformatics Application of Self- Organizing Map Neural Networks for Assessing Microbial and Physico-Chemical Water Quality in Distribution Systems’. *Proceedings of 10th International Conference on Hydroinformatics, Hamburg, Germany*.

Mounce, S. R., E. J. M. Blokker, S. P. Husband, W. R. Furnass, P. G. Schaap, and J. B. Boxall. 2016. ‘Multivariate Data Mining for Estimating the Rate of Discolouration Material Accumulation in Drinking Water Distribution Systems’. *Journal of Hydroinformatics*, jh2015140. https://doi.org/10.2166/hydro.2015.140.

Mounce, S. R., J. B. Boxall, and J. Machell. 2010. ‘Development and Verification of an Online Artificial Intelligence System for Detection of Bursts and Other Abnormal Flows’. *Journal of Water Resources Planning and Management* 136 (3): 309–18. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000030.

Mounce, S. R., K. Ellis, J. M. Edwards, V. L. Speight, N. Jakomis, and J. B. Boxall. 2017. ‘Ensemble Decision Tree Models Using RUSBoost for Estimating Risk of Iron Failure in Drinking Water Distribution Systems’. *Water Resources Management* 31 (5): 1575–89. https://doi.org/10.1007/s11269-017-1595-8.

Mounce, S. R., W. Shepherd, G. Sailor, J. Shucksmith, and A. J. Saul. 2014. ‘Predicting Combined Sewer Overflows Chamber Depth Using Artificial Neural Networks with Rainfall Radar Data’. *Water Science and Technology* 69 (6): 1326–33. https://doi.org/10.2166/wst.2014.024.

Mounce, Stephen. 2018. ‘Visualising Smart Water Meter Dataset Clustering with Parametric (Deep Learning) t-Distributed Stochastic Neighbour Embedding’. In *ICNC-FSKD 2017 - 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery*, 1940–45. Guilin. https://doi.org/10.1109/FSKD.2017.8393065.

Mounce, Stephen, John Machell, and Joby Boxall. 2012. ‘Water Quality Event Detection and Customer Complaint Clustering Analysis in Distribution Systems’. *Water Science & Technology: Water Supply* 12 (5): 580. https://doi.org/10.2166/ws.2012.030.

Mounce, Stephen, Richard Mounce, and Joby Boxall. 2011. ‘Novelty Detection for Time Series Data Analysis in Water Distribution Systems Using Support Vector Machines’. *Journal of Hydroinformatics* 13 (4): 672–86. https://doi.org/10.2166/hydro.2010.144.

Mueller, J.P., and L. Massaron. 2016. *Machine Learning for Dummies*. New Jersey: John Wiley & Sons, Inc.

Myrans, Joshua, Zoran Kapelan, and Richard Everson. 2018. ‘Automatic Identification of Sewer Fault Types Using CCTV Footage’. Proceedings *of 13th International Conference on Hydroinformatics, Palermo, Italy*.

Najah Ahmed, Ali, Faridah Binti Othman, Haitham Abdulmohsin Afan, Rusul Khaleel Ibrahim, Chow Ming Fai, Md Shabbir Hossain, Mohammad Ehteram, and Ahmed Elshafie. 2019. ‘Machine Learning Methods for Better Water Quality Prediction’. *Journal of Hydrology* 578 (August). https://doi.org/10.1016/j.jhydrol.2019.124084.

Nevel, S. Van, S. Koetzsch, C. R. Proctor, M. D. Besmer, E. I. Prest, J. S. Vrouwenvelder, A. Knezev, N. Boon, and F. Hammes. 2017. ‘Flow Cytometric Bacterial Cell Counts Challenge Conventional Heterotrophic Plate Counts for Routine Microbiological Drinking Water Monitoring’. *Water Research* 113: 191–206. https://doi.org/10.1016/j.watres.2017.01.065.

Nevel, Sam Van, Benjamin Buysschaert, Bart De Gusseme, and Nico Boon. 2016. 'Flow Cytometric Examination of Bacterial Growth in a Local Drinking Water Network'. *Water and Environment Journal* 30 (1–2): 167–76. https://doi.org/10.1111/wej.12160.

Niquette, Patrick, Pierre Servais, and Raoul Savoir. 2000. 'Impacts of Pipe Materials on Densities of Fixed Bacterial Biomass in a Drinking Water Distribution System'. *Water Research* 34 (6): 1952–56. https://doi.org/10.1016/S0043-1354(99)00307-3.

Nourani, Vahid, Gozen Elkiran, and S. I. Abba. 2018. 'Wastewater Treatment Plant Performance Analysis Using Artificial Intelligence - An Ensemble Approach'. *Water Science and Technology* 78 (10): 2064–76. https://doi.org/10.2166/wst.2018.477.

Okada, Mihoko. 2021. 'Big Data and Real-World Data-Based Medicine in the Management of Hypertension'. *Hypertension Research* 44 (2): 147–53. https://doi.org/10.1038/s41440-020-00580-3.

Osman, Ahmed M.Shahat. 2019. 'A Novel Big Data Analytics Framework for Smart Cities'. *Future Generation Computer Systems* 91: 620–33. https://doi.org/10.1016/j.future.2018.06.046.

Ozturk, Tulin, Muhammed Talo, Eylul Azra Yildirim, Ulas Baran Baloglu, Ozal Yildirim, and U. Rajendra Acharya. 2020. 'Automated Detection of COVID-19 Cases Using Deep Neural Networks with X-Ray Images'. *Computers in Biology and Medicine* 121 (April): 103792. https://doi.org/10.1016/j.compbiomed.2020.103792.

Parkhurst, David, Kristen Brenner, Alfred Dufour, and Larry Wymer. 2005. 'Indicator Bacteria at Five Swimming Beaches—Analysis Using Random Forests'. *Water Research* 39: 1354–60. https://doi.org/10.1016/j.watres.2005.01.001.

Parsons, Simon, and Bruce Jefferson. 2009. *Introduction to Potable Water Treatment Processes*. 2nd ed. Oxford: Blackwell Publishing.

Patterson, Josh, and Adam Gibson. 2017. *Deep Learning: A Practitioner's Approach*. 1st ed. Sebastopol: O'Reilly Media Inc.

Payment P. 2003. 'Health Effects of Water Consumption and Water Quality'. In *Handbook of Water and Wastewater Microbiology*, edited by D. Mara and N. Horan, 203–19. Elsevier Inc.

Pedrycz, Witold, and Zenon A Sosnowski. 2001. 'The Design of Decision Trees in the Framework of Granular Data and Their Application to Software Quality Models'. *Fuzzy Sets and Systems* 123: 271–90.

Peng, Ching Yu, Gregory V. Korshin, Richard L. Valentine, Andrew S. Hill, Melinda J. Friedman, and Steve H. Reiber. 2010. 'Characterization of Elemental and Structural Composition of Corrosion Scales and Deposits Formed in Drinking Water Distribution Systems'. *Water Research* 44 (15): 4570–80. https://doi.org/10.1016/j.watres.2010.05.043.

Pennine Water Group; ARC Consultancy. 2017. '5011840000 – Network Investigations - Water Quality Investigation Into Deteriorating Cast Iron Mains - Final Report'.

Petrova, Olga E., and Karin Sauer. 2016. 'Escaping the Biofilm in More than One Way: Desorption, Detachment or Dispersion'. *Current Opinion in Microbiology* 30: 67–78. https://doi.org/10.1016/j.mib.2016.01.004.

Pisoni, Enrico, Marcello Farina, Claudio Carnevale, and Luigi Piroddi. 2009. 'Forecasting Peak Air Pollution Levels Using NARX Models'. *Engineering Applications of Artificial Intelligence* 22 (4–5): 593–602. https://doi.org/10.1016/j.engappai.2009.04.002.

Praveena, M., and V. Jaiganesh. 2017. 'A Literature Review on Supervised Machine Learning Algorithms and Boosting Process' 169 (8): 32–35.

Prest, E. I., F. Hammes, S. Kötzsch, M. C.M. Van Loosdrecht, and J. S. Vrouwenvelder. 2016. 'A Systematic Approach for the Assessment of Bacterial Growth-Controlling Factors Linked to Biological Stability of Drinking Water in Distribution Systems'. *Water Science and Technology: Water Supply* 16 (4): 865–80. https://doi.org/10.2166/ws.2016.001.

Prest, Emmanuelle I., Frederik Hammes, Mark C.M. van Loosdrecht, and Johannes S. Vrouwenvelder. 2016. 'Biological Stability of Drinking Water: Controlling Factors, Methods, and Challenges'. *Frontiers in Microbiology* 7 (FEB): 1–24. https://doi.org/10.3389/fmicb.2016.00045.

Quinlan, J R. 1987. 'Simplifying Decision Trees'. *International Journal of Man-Machine Studies* 27 (August 1986): 221–34. https://doi.org/10.1016/S0020-7373(87)80053-6.

Rojek, Izabela. 2014. 'Models for Better Environmental Intelligent Management within Water Supply Systems'. *Water Resources Management* 28 (12): 3875–90. https://doi.org/10.1007/s11269-014-0654-7.

Rokach, Lior. 2010. 'Ensemble-Based Classifiers'. *Artificial Intelligence Review* 33 (1–2): 1–39. https://doi.org/10.1007/s10462-009-9124-7.

Romano, Michele, Zoran Kapelan, and Dragan A. Savić. 2014. 'Automated Detection of Pipe Bursts and Other Events in Water Distribution Systems'. *Journal of Water Resources Planning and Management* 140 (4): 457–67. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000339.

Romero, Juan Manuel Ponce, Stephen H. Hallett, and Simon Jude. 2017. 'Leveraging Big Data Tools and Technologies: Addressing the Challenges of Thewater Quality Sector'. *Sustainability (Switzerland)* 9 (12). https://doi.org/10.3390/su9122160.

Ronao, Charissa Ann, and Sung Bae Cho. 2016. 'Human Activity Recognition with Smartphone Sensors Using Deep Learning Neural Networks'. *Expert Systems with Applications* 59: 235–44. https://doi.org/10.1016/j.eswa.2016.04.032.

Rosin, T, M Romano, K Woodward, E Keedwell, and Z Kapelan. 2018. 'Prediction of CSO Chamber Level Using Evolutionary Artificial Neural Networks'. Proceedings *of 13th International Conference on Hydroinformatics, Palermo, Italy*. Palermo.

Sarin, Pankaj, Jonathan Clement, Vernon Snoeyink, and Waltraud Kriven. 2003. 'Iron Release from Corroded, Unlined Cast-Iron Pipe'. *Journal - American Water Works Association* 95 (11): 85–96.

Sarin, Pankaj, Darren A Lytle, and Waltraud M Kriven. 2004. 'Iron Corrosion Scales : Model for Scale Growth , Iron Release , and Colored Water Simon W . Freese Environmental Engineering Lecture Iron Corrosion Scales : Model for Scale Growth , Iron Release , and Colored Water Formation'. *Journal of Environmental Engineering ASCE* 9372 (April): 1488–97. https://doi.org/10.1061/(ASCE)0733-9372(2004)130.

Schmidhuber, Jürgen, Daan Wierstra, and Faustino Gomez. 2005. 'Evolino: Hybrid Neuroevolution / Optimal Linear Search for Sequence Learning'. *IJCAI International Joint Conference on Artificial Intelligence*, 853–58.

Scornet, Erwan. 2018. 'TUNING PARAMETERS IN RANDOM FORESTS'. *ESAIM: PROCEEDINGS AND SURVEYS* 60: 144–62. https://doi.org/https://doi.org/10.1051/proc/201760144.

Seiffert, Chris, Taghi M. Khoshgoftaar, Jason Van Hulse, and Amri Napolitano. 2010. 'RUSBoost: A Hybrid Approach to Alleviating Class Imbalance'. *IEEE Transactions on Systems, Man, and Cybernetics Part A:Systems and Humans* 40 (1): 185–97. https://doi.org/10.1109/TSMCA.2009.2029559.

Seiffert, Chris, Taghi M Khoshgoftaar, Jason Van Hukse, and Amri Napolitano. 2008. 'RUSBoost: Improving Classification Performance When Training Data Is Skewed'. In the

19th *International Conference on Pattern Recognition (ICPR 2008), December 8-11, 2008, Tampa, Florida, USA*. https://doi.org/10.1109/ICPR.2008.4761297.

Seth, A., R. T. Bachmann, J. B. Boxall, A. J. Saul, and R. Edyvean. 2004. 'Characterisation of Materials Causing Discolouration in Potable Water Systems'. *Water Science and Technology* 49 (2): 27–32. https://doi.org/10.2166/wst.2004.0080.

Sharpe, Rebecca L., Catherine A. Biggs, and Joby B. Boxall. 2019. 'Hydraulic Conditioning to Manage Potable Water Discolouration'. *Proceedings of the Institution of Civil Engineers: Water Management* 172 (1): 3–13. https://doi.org/10.1680/jwama.16.00038.

Shu, Chang, and Donald H. Burn. 2004. 'Artificial Neural Network Ensembles and Their Application in Pooled Flood Frequency Analysis'. *Water Resources Research* 40 (9): 1–10. https://doi.org/10.1029/2003WR002816.

Sly, L I, M C Hodgkinson, and Vullapa Arunpairojana. 1990. 'Deposition of Manganese in a Drinking Water Distribution System'. *Applied and Environmental Microbiology* 56 (3): 628–39.

Smith, Lindsay I. 2002. 'A Tutorial on Principal Components Analysis Introduction'. University of Otago. http://hdl.handle.net/10523/7534.

Speight, Vanessa, Stephen Mounce, and Joseph B. Boxall. 2019. 'Identification of the Causes of Drinking Water Discolouration from Machine Learning Analysis of Historical Datasets'. *Environmental Science: Water Research and Technology* 5.4: 747–55. https://doi.org/10.1039/c8ew00733k.

Standing Committee of Analysts. 2002. 'The Microbiology of Drinking Water (2002) -Part 1 - Water Quality and Public Health Methods for the Examination of Waters and Associated Materials'.

———. 2009. 'The Microbiology of Drinking Water (2009) - Part 4 - Methods for the Isolation and Enumeration of Coliform Bacteria and Escherichia Coli (Including E. Coli O157:H7) Methods'.

———. 2015a. 'Standing Committee of Analysts'. *The Microbiology of Recreational and Environmental Waters (2015) – Part 5 – Methods for the Isolation and Enumeration of Sulphite-Reducing Clostridia and Clostridium Perfringens*. http://standingcommitteeofanalysts.co.uk/library/MoDW Part 4 (2016) - Coliforms & E. coli (FINAL June 2016).pdf.

———. 2015b. 'The Microbiology of Recreational and Environmental Waters (2015) – Part 4 – Methods for the Isolation and Enumeration of Enterococci'. http://standingcommitteeofanalysts.co.uk/library/MoDW Part 4 (2016) - Coliforms & E. coli (FINAL June 2016).pdf.

———. 2016. 'The Microbiology of Recreational and Environmental Waters (2016) – Part 3 – Methods for the Isolation and Enumeration of Escherichia Coli (Including E. Coli O157:H7)'. http://standingcommitteeofanalysts.co.uk/library/MoDW Part 4 (2016) - Coliforms & E. coli (FINAL June 2016).pdf.

Sunny, Iftekhar, Stewart Husband, Graeme Moore, Nick Drake, Kevan Mckenzie, and J. Boxall. 2017. 'Discolouration Risk Management and Chlorine Decay'. In *Proceedings of the Computing and Control for the Water Industry 2017 Conference, Sheffield, UK*. Sheffield.

Telfer A. 2014. 'NITRIFICATION IN CHLORAMINATED DRINKING WATER SUPPLIES'. In *77 Th Annual WIOA Victorian Water Industry Operations Conference and Exhibition Bendigo Exhibition Centre 2 to 4 September, 2014*, 42–48. Victoria.

The U.S. Department of the Interior, Bureau of Reclamation. 2013. 'Effect of Chlorine vs.

Chloramine Treatment Techniques on Materials Degradation in Reclamation Infrastructure'. Denver, Colorado.

Tornevi, Andreas, Olof Bergstedt, and Bertil Forsberg. 2014. 'Precipitation Effects on Microbial Pollution in a River: Lag Structures and Seasonal Effect Modification'. *PLoS ONE* 9 (5). https://doi.org/10.1371/journal.pone.0098546.

United Nations. 2010. '64/292. The Human Right to Water and Sanitation'. *General Assembly* 64 (292): 3. http://www.un.org/es/comun/docs/?symbol=A/RES/64/292&lang=E.

Usama, Muhammad, Junaid Qadir, Aunn Raza, Hunain Arif, and Kok-lim Alvin Yau. 2017. 'Unsupervised Machine Learning for Networking : Techniques , Applications and Research Challenges', no. September.

USEPA. 2001. 'Controlling Disinfection By-Products and Microbial Contaminants in Drinking Water'. Washington.

———. 2002a. 'Effects of Water Age on Distribution System Water Quality'. *Usepa*. Washington DC.

———. 2002b. 'Nitrification'. Washington D.C. https://doi.org/10.1007/978-3-662-13187-9_80.

Vatanen, T., M. Osmala, T. Raiko, K. Lagus, M. Sysi-Aho, M. Orešič, T. Honkela, and H. Lähdesmäki. 2015. 'Self-Organization and Missing Values in SOM and GTM'. *Neurocomputing* 147 (1): 60–70. https://doi.org/10.1016/j.neucom.2014.02.061.

Villanueva Zacarias, Alejandro Gabriel, Peter Reimann, and Bernhard Mitschang. 2018. 'A Framework to Guide the Selection and Configuration of Machine-Learning-Based Data Analytics Solutions in Manufacturing'. *Procedia CIRP* 72: 153–58. https://doi.org/10.1016/j.procir.2018.03.215.

Vreeburg, Ir J.H.G., and Dr J.B. Boxall. 2007. 'Discolouration in Potable Water Distribution Systems: A Review'. *Water Research* 41 (3): 519–29. https://doi.org/10.1016/j.watres.2006.09.028.

Vreeburg, J. H.G. 2014. 'Optimization of Design and Operation of Distribution Systems to Preserve Water Quality and Maintain Customer Satisfaction'. In *Microbial Growth in Drinking-Water Supplies: Problems, Causes, Control and Research Needs*, edited by Paul W J J van der Wielen and Dick; van der Kooij, 403–20. London: IWA Publications.

Vreeburg, J. H.G., P. G. Schaap, and J. C. Van Dijk. 2004. 'Particles in the Drinking Water System: From Source to Discolouration'. *Water Science and Technology: Water Supply* 4 (5–6): 431–38. https://doi.org/10.2166/ws.2004.0135.

Vreeburg, J. H.G., D. Schippers, J. Q.J.C. Verberk, and J. C. van Dijk. 2008. 'Impact of Particles on Sediment Accumulation in a Drinking Water Distribution System'. *Water Research* 42 (16): 4233–42. https://doi.org/10.1016/j.watres.2008.05.024.

Vreeburg, J. H G. 2007. 'Discolouration in Drinking Water Systems: A Particular Approach'. *Civil Engineering and Geosciences*. Delft. http://resolver.tudelft.nl/uuid:20db1587-3383-4dd3-81a4-2a0ef276e127.

Vries, D., B. Van Den Akker, E. Vonk, W. De Jong, and J. Van Summeren. 2016. 'Application of Machine Learning Techniques to Predict Anomalies in Water Supply Networks'. *Water Science and Technology: Water Supply* 16 (6): 1528–35. https://doi.org/10.2166/ws.2016.062.

Wang, Hong, Sheldon Masters, Yanjuan Hong, Jonathan Stallings, Joseph O. Falkinham, Marc A. Edwards, and Amy Pruden. 2012. 'Effect of Disinfectant, Water Age, and Pipe Material on Occurrence and Persistence of Legionella, Mycobacteria, Pseudomonas

Aeruginosa, and Two Amoebas'. *Environmental Science and Technology* 46 (21): 11566–74. https://doi.org/10.1021/es303212a.

Wang, Yingying, Frederik Hammes, Karen De Roy, Willy Verstraete, and Nico Boon. 2010. 'Past, Present and Future Applications of Flow Cytometry in Aquatic Microbiology'. *Trends in Biotechnology* 28 (8): 416–24. https://doi.org/10.1016/j.tibtech.2010.04.006.

Water Research Centre. 1976. 'Deterioration of Bacteriological Quality of Water during Distribution'.

Wei, Bao, Jun Yue, and Yulei Rao. 2017. 'A Deep Learning Framework for Financial Time Series Using Stacked Autoencoders and Long- Short Term Memory'. *PLoS ONE* 7. https://doi.org/10.1371/journal.pone.0180944.

West, Jarrod, and Maumita Bhattacharya. 2016. 'Intelligent Financial Fraud Detection: A Comprehensive Review'. *Computers and Security* 57: 47–66. https://doi.org/10.1016/j.cose.2015.09.005.

WHO. 2000. 'WHO Guidelines for Drinking Water Quality: Training Pack'. https://apps.who.int/iris/handle/10665/66218.

———. 2011. 'Strategies for the Safe Management of Drinking-Water for Human Consumption'. Geneva. http://apps.who.int/gb/ebwha/pdf_files/WHA64/A64_24-en.pdf.

———. 2017. 'WATER QUALITY AND HEALTH - REVIEW OF TURBIDITY: Information for Regulators and Water Suppliers'. Geneva. http://www.who.int/water_sanitation_health/publications/turbidity-information-200217.pdf.

Wu, Wenyan, Graeme C. Dandy, and Holger R. Maier. 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'. *Environmental Modelling and Software* 54: 108–27. https://doi.org/10.1016/j.envsoft.2013.12.016.

Wu, Zheng Yi, and Atiqur Rahman. 2017. 'Optimized Deep Learning Framework for Water Distribution Data-Driven Modeling'. *Procedia Engineering* 186: 261–68. https://doi.org/10.1016/j.proeng.2017.03.240.

Xenochristou, Maria. 2019. 'Water Demand Forecasting Using Machine Learning on Weather and Smart Metering Data'. University of Exeter.

Xenochristou, Maria, Chris Hutton, Jan Hofman, and Zoran Kapelan. 2021. 'Short-Term Forecasting of Household Water Demand in the UK Using an Interpretable Machine Learning Approach'. *Journal of Water Resources Planning and Management* 147 (4). https://doi.org/10.1061/(asce)wr.1943-5452.0001325.

Xenochristou, Maria, Zoran Kapelan, and Chris Hutton. 2020. 'Using Smart Demand-Metering Data and Customer Characteristics to Investigate Influence of Weather on Water Consumption in the Uk'. *Journal of Water Resources Planning and Management* 146 (2): 1–12. https://doi.org/10.1061/(ASCE)WR.1943-5452.0001148.

Xenochristou, Maria, Zoran Kapelan, Chris Hutton, and Jan Hofman. 2018. 'Smart Water Demand Forecasting : Learning from the Data', no. July: 1–7.

Xu, Runze, Jiashun Cao, Fang Fang, Qian Feng, E. Yang, and Jingyang Luo. 2021. 'Integrated Data-Driven Strategy to Optimize the Processes Configuration for Full-Scale Wastewater Treatment Plant Predesign'. *Science of the Total Environment* 785: 147356. https://doi.org/10.1016/j.scitotenv.2021.147356.

Yang, Jian, Mark W. LeChevallier, Peter F M Teunis, and Minhua Xu. 2011. 'Managing Risks from Virus Intrusion into Water Distribution Systems Due to Pressure Transients'.

Journal of Water and Health 9 (2): 291–305. https://doi.org/10.2166/wh.2011.102.

Yura, Eisaku, Kohji Tanaka, Yeonjoong Kim, Hiroki Tsujikura, and Tatsuya Yoshida. 2018. ‘Development of the Similar Typhoon Search System Based on the Deep Neural Network Using Deep Learning’, no. July.

Zekić-Sušac, Marijana, Saša Mitrović, and Adela Has. 2020. ‘Machine Learning Based System for Managing Energy Efficiency of Public Sector as an Approach towards Smart Cities’. *International Journal of Information Management*. https://doi.org/10.1016/j.ijinfomgt.2020.102074.

Zhang, Kejiang, Gopal Achari, Hua Li, Amin Zargar, and Rehan Sadiq. 2013. ‘Machine Learning Approaches to Predict Coagulant Dosage in Water Treatment Plants’. *International Journal of Systems Assurance Engineering and Management* 4 (2): 205–14. https://doi.org/10.1007/s13198-013-0166-5.

Zhang, Yan, and Marc Edwards. 2009. ‘Accelerated Chloramine Decay and Microbial Growth by Nitrification in Premise Plumbing’. *Journal / American Water Works Association* 101 (11): 51–62. https://doi.org/10.1002/j.1551-8833.2009.tb09990.x.

Zhou, Xiao, Zhenheng Tang, Weirong Xu, Fanlin Meng, Xiaowen Chu, Kunlun Xin, and Guangtao Fu. 2019. ‘Deep Learning Identifies Accurate Burst Locations in Water Distribution Networks’. *Water Research* 166: 115058. https://doi.org/10.1016/j.watres.2019.115058.

Zlatanović, Lj, J. P. van der Hoek, and J. H.G. Vreeburg. 2017. ‘An Experimental Study on the Influence of Water Stagnation and Temperature Change on Water Quality in a Full-Scale Domestic Drinking Water System’. *Water Research* 123: 761–72. https://doi.org/10.1016/j.watres.2017.07.019.