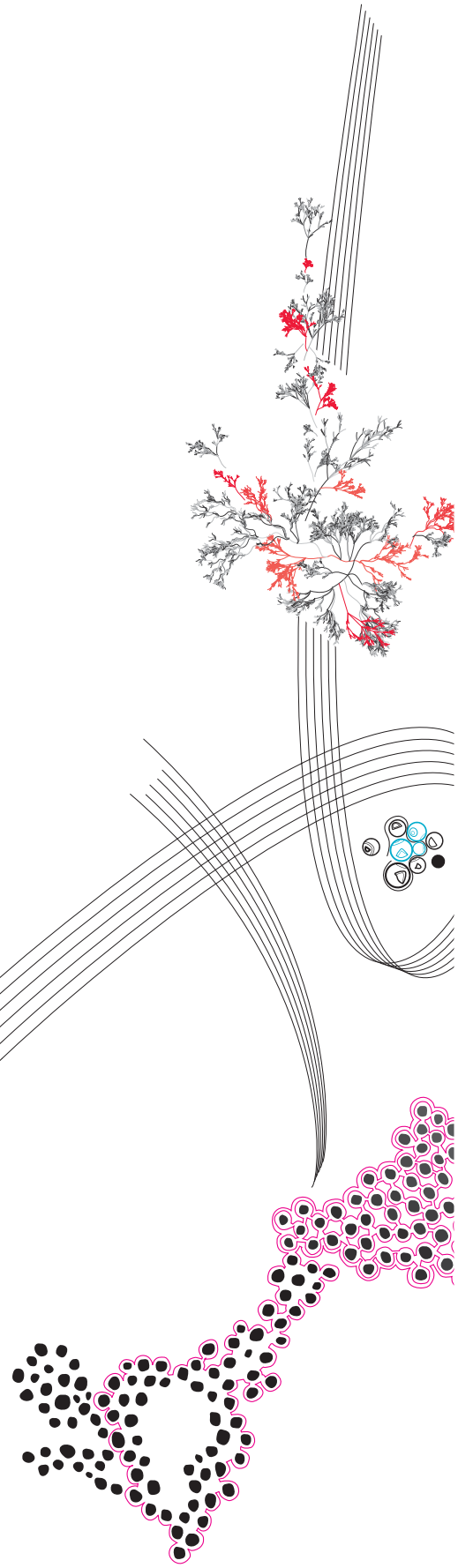

Graduation Project Creative Technology

Predicting the Formation of Disinfection Byproducts Using Environmental Parameters in Chlorinated Drinking Water

Julia Kersten

Supervisor: A. Kamilaris

Critical Observer: B. Guddanti

July 20, 2024

Department of Creative Technology
Faculty of Electrical Engineering,
Mathematics and Computer Science

UNIVERSITY OF TWENTE.

Abstract

Safe drinking water is crucial, yet 2.2 billion people lacked access to it in 2022. Chlorination, a common disinfection method, can produce harmful disinfection by-products (DBPs) like trihalomethanes (THMs) and halo-acetic acids (HAAs). These DBPs pose health risks, but only 30% of over 700 identified DBPs have been quantified, emphasizing the need for better predictive models. This project aims to further understand the relationship between water parameters such as, pH, alkalinity, DOC/TOC and SUVA/UVA, and the concentration of DBPs after chlorination; by doing a data analysis on publicly accessible data from the EPA and trying to fit various regression models for it. The best fit model achieved an R^2 of 0.5 which was a Ridge Regression model for the bromoform DBP. The findings reveal the challenges to making accurate predictive models without enough good quality data.

Acknowledgement

I would like to thank my supervisor Andreas Kamilaris and my critical observer Balaram Guddanti for their support on this graduation project. Without their feedback and generosity it would have not been possible.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Clean and safe drinking water should be an accessible resource to everyone. However, according to the UN, 2.2 billion people still lacked safely managed drinking services in 2022 [1]. Moreover, due to phenomena such as climate change, population growth and global industrialization, the quality and availability of our waters are put under considerable stress [2]. Currently, as ground water is becoming more limited and scarce, most of our drinking water is sourced from surface water, such as lakes and rivers, and or wastewater [3]. The water is then treated in several steps, to remove all kinds of undesired particles. Usually, disinfection is the last step where chemicals such as chlorine, chloramine or chlorine dioxide, are added to the water to help keep it safe and drinkable as it travels through potentially contaminated pipes to reach our homes and businesses [4].

Disinfection using chlorine – chlorination — is one of the cheapest and effective methods that is used across the world to achieve safe drinking water [5, 2]. It removes most of the pathogens present in the source water to prevent water borne illnesses, such as cholera, diarrhoea, dysentery and so on [6]. Disinfection is highly necessary, as according to the World Health Organization (WHO), nearly 80% of human diseases in developing countries were due to unsafe drinking water [7].

Despite this, an unintended side effect resulting from disinfection and chlorination is the formation of disinfection by products (DBPs). Chlorine reacts with natural organic matter (NOM) and other pollutants present in the water, producing all kinds of DBPs [8]. Alongside NOMs, other factors such as potential of hydrogen (pH), water temperature,

reaction time, amount of chlorine, bromide concentration etc. also affect the formation of DBPs.

In the 70s it has been discovered that certain DBPs, such as trihalomethanes (THMs) and haloacetic acids (HAAs), are likely to be carcinogenic, genotoxic and mutagenic [5]. Long term exposure of these DBPs have also been linked to an increased risk for several cancers including bladder, liver and colon cancers [9]. Since then over 700 different DBPs have been identified in drinking water treatment plants (DWTPs), but only about 30% has actually been quantified [10].

1.1 Problem Statement

The challenge of this project is that it is unclear which exact water conditions, prior to or while disinfecting, form particular DBPs. Moreover, with the number of newly discovered DBPs growing, the potential health effects of each of them are yet to be investigated. For example, toxicity data revealed that emerging nitrogenous DBPs (N-DBPs) are found to be more toxic than carbonaceous DBPs (C-DBPs) [2]. This shows that more research is needed on the formation factors and mechanisms of DBPs.

Due to the large variety of NOM and uncertain formations, slight changes in the water quality parameters can reduce the formation of one DBP, while increasing the formation of other DBPs, which may also pose additional cancer risks [5]. This makes it difficult to target specific DBPs and keep to regulation standards. According to Kali et al, the concentration of regulated DBPs surpassed the permissible limit in most of the regions investigated [2]. This further shows the need for research into predicting DBPs in a timely manner in order to let DWTPs know whether they have surpassed the limit before its deployed to the communities.

Furthermore, standard drinking water monitoring and DBP detection methods require complicated instruments such as gas chromatography (GC) and mass spectrometry (MS) that is time consuming and expensive [11]. Additionally, accurate models also require large datasets, but the lack of high quality data management and models being non-reproducible also contribute to the challenge [12]. Therefore, it is important to do further research into the conditions that make up the most harmful DBPs, try to centralize available data in order to make more accurate models that can predict the formation

of DBPs without the need for expensive equipment to eventually be able to mitigate the risks on human health.

1.2 Research Questions

The main research question of this project is:

How does the formation of DBPs in chlorinated drinking water correlate with water quality parameters?

The sub-questions will be:

1. What are the most significant water quality parameters in predicting DBPs?
2. What methods/models have been made to predict DBPs?
3. How effective are current methods/models that try to detect and predict the formation of DBPs?

1.3 Overview of the Report Structure

Chapter 2 contains the background information such as a deeper look into DBPs and the factors surrounding the formation of them from what is known in literature. Chapter 3 explains the methods and techniques that has been used to realise this project from literature search methodologies, cyclical methods to the software and coding languages. Chapter 4 illustrates the conceptual idea of the project and the steps to achieve it. Chapter 5 shows the process, results and reflection of each iteration of the data science cycle. Chapter 6 will evaluate the predictive models and discuss, as well as mention what can be done in the future. Lastly chapter 7 is the conclusion where previous chapters are briefly summarised and the overall picture of the project is illustrated.

Chapter 2

Background Research

This chapter aims to answer the research questions previously stated. Section 2.1 will briefly go over the water treatment process and the implications of it, Section 2.2 will dive into what DBPs are and the types of them, Section 2.3 explains the different factors affecting DBP formation, Section 2.4 highlights the different health effects caused by DBPs, Section 2.5 assesses the different methods used in predicting DBPs and lastly Section 2.6 gives a conclusion to this chapter.

2.1 The Water Treatment Process

The water treatment process is slightly different per country, regulation and water source, but to simplify, water undergoes roughly 5 procedures before water is sent out to communities. As Figure 2.1 points out, the first step in this process is coagulation. Here chemicals with a positive charge are added to the water to neutralize the negative charges of dirt and other dissolved particles [13]. This allows for particles to bind with the chemicals to form slightly larger particles. Next, it goes through flocculation which is a process where heavier particles called flocs are formed through gentle mixing of the water and adding more chemicals. After that, sedimentation occurs which is a step where solids are separated from the water by flocs sinking to the bottom of the tank due to them being heavier than water. Once that is done, the clear water on top of the flocs is filtered to separate additional solids from the water by going through differently sized pores of filters made out of different materials (e.g sand, gravel and charcoal). Lastly, the water gets disinfected

by one or more chemical disinfectants, ozone or ultraviolet (UV) light. The latter two work well for disinfection but they do not continue killing pathogens as the water travels through the pipes [4]. Therefore, it is recommended by regulatory authorities to maintain a residual chlorine concentration of 0.2 mg/L in the distribution system to mitigate future microbiological (re)contamination [14].

In the present world, due to stricter regulations, DWTPs have started to experiment with different ways of disinfection. For example the idea of sequential disinfection, where more than one disinfectant is used at different stages of the treatment process, has been experimented with such as the combination of chlorine dioxide followed chlorination. Results indicated that these methods had the potential to decrease DBP formation of THMs, HAAs, and HANs, but produced chlorite and chlorate in another multi disinfection method [14]. Furthermore, due to excessive costs of alternative methods [2], many developing countries do not have access to these advanced methods. This further shows the importance of researching DBP formation specific to chlorination as it will impact the most amount of people overall.

2.2 Introduction to Disinfection Byproducts

DBPs come in all kinds of chemical formations. Subsection 2.2.1 showcases the different types and major families of DBP, Subsection 2.2.2 will introduce the regulated DBPs and their current regulations and Subsection 2.2.3 will highlight the emerging DBPs that are recently discovered.

2.2.1 Types of DBPs

There are multiple families of DBPs which can be categorised in 3 types, namely aliphatic, alicyclic and aromatic DBPs. Aliphatic DBPS can be split into 2 additional types, nitrogenous DBPs (N-DBPs) and carbonaceous DBPs (C-DBPs). More C-DBPs are found in the water, while the toxicity of N-DBPs are higher and form in waters with more dissolved organic nitrogen (DON). Alicyclic DBPs are less explored while aromatic DBPs can act as precursors to aliphatic DBPs [9].

Out of DBPs that have been quantified in drinking water, the levels are typically present at small amounts, such as below a μg/L or low-to-mid μg/L [15]. The most

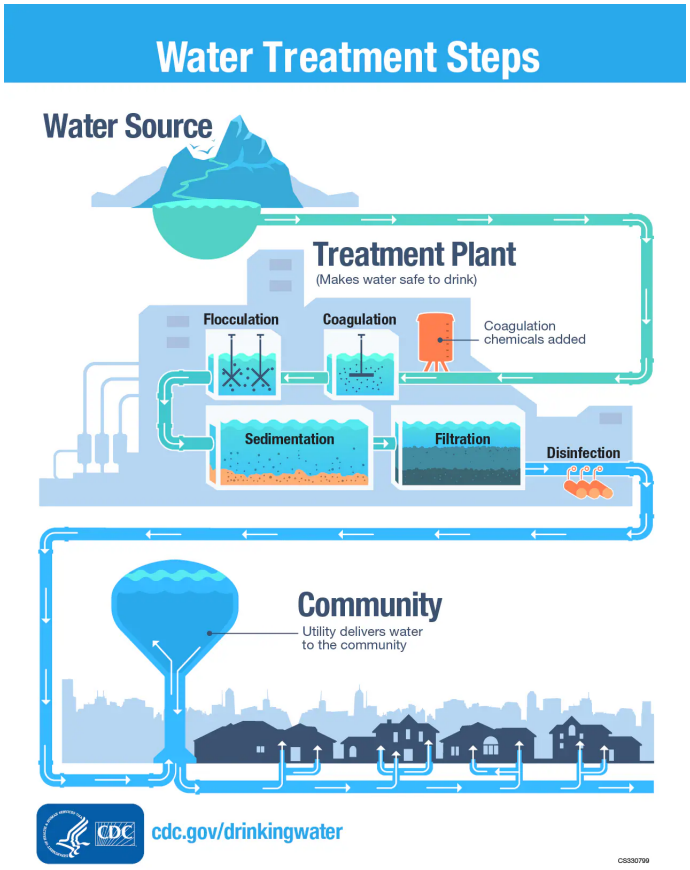

FIGURE 2.1: The Water Treatment Process [4]

common family of DBPs found in chlorinated drinking water are THMs [3] which include, chloroform/trichloromethane (CF/TCM, $CHCl_3$), dibromochloromethane (DBCM, $CHBr_2Cl$), bromodichloromethane (BDCM, $CHBrCl_2$) and bromoform (BF, $CHBr_3$). Another big family of DBPs are HAAs where such as Dichloroacetic acid ($Cl_2CH\text{-}CO_2H$), trichloroacetic acid (TCAA, $Cl_3C\text{-}CO_2H$), monochloroacetic acid (MCAA, $ClCH_2CO_2H$).

2.2.2 Regulated DBPs

About 25% of all known DBPs are from the THM and HAA family [2] and which is likely why it is widely regulated. The first regulation for DBPs came to the US in 1979 when the association between THMs and elevated chronic cancer risks were discovered. HAAs followed the next year due to their frequency in the waters [16]. These so called regulated DBPs (R-DBPs) have acronyms for them such as THM_4, where the number 4 refers to the 4 THM DBPs that are regulated (chloroform, DBCM, BDCM and bromoform). Similarly, HAA_5 and HAA_9 refer to the set of 5 and 9 HAA DBPS that are regulated.

For HAA_5 it is MCAA, DCAA, TCAA, MBAA and DBAA [13]. Likewise HAA_9 contains HAA_5 plus bromochloroacetic acid (BCAA), bromodichloroacetic acid (BDCAA), dibromochloroacetic acid (DBCAA) and tribromoacetic acid (TBAA) additionally [17].

DBP Group	DBP	U.S. EPA (μg/L)	WHO (μg/L)	EU (μg/L)
THMs	Chloroform	70 **	300	
	BDCM	45	60	
	DBCM	60 **	100	
	Bromoform	6	100	
	THM_4	80 *		100
HAAs	MCAA	70	20	
	DCAA	60	50	
	TCAA	20	200	
	MBAA	60		
	DBAA	60		
	HAA_5	60 *		60 *
HAN	DCAN	6	20	
	DBAN	20	70	
Inorganic DBPs	Bromate	10 *	10	10
	Chlorite	1000 *	700	
	Chlorate	1000	700	
NNA	NNDMA	0.01	0.1	

TABLE 2.1: Regulatory limits (*) or guideline values (non-regulatory limits) (**) for DBPs established by different organizations [13, 9]

Table 2.1 showcases the current regulations from the United States Environmental Protection Agency (U.S. EPA), WHO and the European Union (EU). It is noticeable that the EPA regulates the most amount of DBPs while EU only has 3 regulation guidelines of which 2 are aggregated values and not specific to a certain DBP.

While regulations for other countries do exist, it is also not as extensive as the U.S. or WHO and might also not be comparable due to the different kinds of source water low-income countries have to try and regulate. As such, a study conducted by Furst et al. [18], highlighted that although the THM4 levels in Rajasthan, India did not

exceed the international guidelines, other more toxicological DBPs were observed in high concentrations. Unlike high-income countries, the distribution systems can let sewage water be infiltrated and hence affecting the water quality. This showed that these THM4 regulations might not be an adequate indicator of overall DBP exposure when it comes to polluted water supplies in some low-income countries. Therefore, a predicative model that could take into account all the different potential water quality condition and accurately predict the amount of DBPs in necessary.

2.2.3 Emerging DBPs

Emerging DBPs are DBPs that are not regulated. Although it may seem like these DBPs are less harmful because they are not regulated yet, it is actually the opposite. For example halobenzoquinones (HBQs) are a new kind of DBP, that can cause damage to DNA and whose toxicity was higher than that of THMs and HAAs [14]. Several N-DBPs still need quantification and toxicity studies that conclude on the toxicity levels of different DBPs. Ionated DBPs are also an emerging DBPs that are hugely toxic to humans, animals, as well as aquatic life [19] [20].

2.3 Factors influencing DBPs

In this section several factors that influence DBP fomation are discussed.

2.3.1 Natural Organic Matter (NOM)

NOMs are an extremely complex mixture of organic compounds that vary in chemical and physical characteristics [22]. The composition and concentration of NOM in water can vary depending on factors such as the source of the water and seasonal changes [23] [9]. Generally, higher concentrations of NOM result in increased formation of DBPs due to the greater availability of organic precursors for chlorination [24].

As 2.2 illustrates, NOM is made up of dissolved organic matter (DOM) which contains dissolved organic carbon (DOC) and dissolved organic nitrogen (DON). DOC can be further broken down to its humic (humic acid, fulvic acid, and humin) and non-humic material [25]. Amino acids (AAs) are known as the important precursors of nitrogenous disinfection by-products (N-DBPs) and account for about 15-35% of dissolved organic nitrogen (DON) [15].

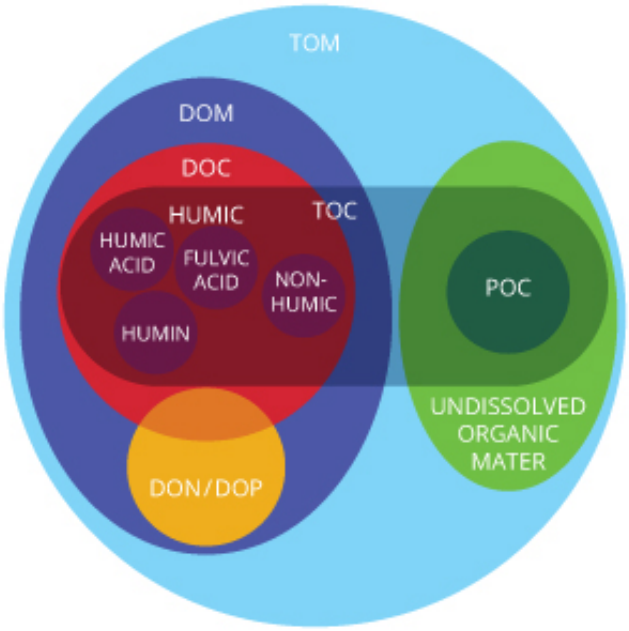

FIGURE 2.2: Venn diagram representation of the various forms of organic matter found in natural water [21]

Parameters that try to measure/quantify NOMs are dissolved organic nitrogen (DON), dissolved organic carbon (DOC), and UVA_2254, where DOC is the actual concentration of DOC while UVA_2254 measures the absorbance of UV light by organic matter at specifically 254nm of wavelength. It is often used as a surrogate for DOC. UV254 is higher for humic acids because of the higher aromatic content and greater molecular size of this type of compound [26].

2.3.2 Potential of Hydrogen (pH)

The pH, or potential of hydrogen, plays a crucial role in the formation of disinfection by-products (DBPs). The pH of the water affects the speciation of chlorine-based disinfectants, with different forms having varying reactivity towards organic precursors. In general, higher pH levels can lead to increased formation of certain DBPs, such as trihalomethanes (THMs), due to enhanced chlorination reactions [2]. Conversely, lower pH levels can favor the formation of other DBPs, such as haloacetic acids (HAAs) and N-DBPs [9]. Therefore, careful monitoring and control of pH during water treatment processes are essential to mitigate the formation of DBPs.

2.3.3 Temperature

Temperature also plays a significant role in the formation of disinfection by-products (DBPs) during chlorination processes in water treatment. Higher temperatures can accelerate the reactions between chlorine and organic matter, leading to increased formation of DBPs [27]. This is because elevated temperatures generally increase the kinetic energy of molecules, promoting more rapid chemical reactions [9]. Additionally, warmer water can stimulate microbial activity, resulting in higher concentrations of organic precursors available for chlorination [8].

2.3.4 Chlorine Dosage

Chlorine dosage is a critical factor in water treatment processes, especially in the context of disinfection by-product (DBP) formation. Balancing the need for effective disinfection with the minimization of DBP formation requires careful consideration of chlorine dosage. Insufficient chlorine dosage may result in inadequate disinfection, leaving harmful pathogens untreated, while excessive chlorine dosage can lead to the overproduction of DBPs [9].

2.3.5 Bromide

According to Oblensky and Singer, bromide had a significant influence on most DBPs that were tested [28]. Bromide naturally occurs in many water sources, and when chlorine is used as a disinfectant, it can react with bromide to form brominated DBPs, which often have higher toxicity than their chlorinated counterparts [29]. For example, brominated THMs (Br-THMs) and brominated HAAs (Br-HAAs) are among the most common brominated DBPs formed during chlorination. Moreover, concentrations of bromide are also generally considered a factor, because they can influence the distribution of the four THM compounds [30].

2.4 Health and Environmental Impacts

Health effects associated with DBPs vary depending on the specific compounds formed, their concentration in water, and individual susceptibility. Some DBPs, such as trihalomethanes (THMs) and haloacetic acids (HAAs), have been linked to adverse health effects including cancer, reproductive problems, and developmental disorders, especially

with long-term exposure [9].

Brominated DBPs, in particular, are of concern due to their higher toxicity compared to chlorinated DBPs. Brominated compounds, such as bromoform and bromodichloromethane, have been classified as probable human carcinogens by regulatory agencies [29]. Additionally, the combined activity of residual chlorine and DBPs may pose greater risks to aquatic ecosystems, affecting aquatic organisms and disrupting ecological balance [20].

2.5 State of the Art in DBP Prediction

To predict complex problems such as DBP formation, many factors need to be taken into account.

2.5.1 Multiple Linear and Non-Linear Regression

Multiple linear regression is a statistical technique used to model the relationship between a single dependent variable and two or more independent variables. It builds upon the foundation of simple linear regression by allowing for the consideration of multiple predictors simultaneously. The aim is to develop a linear equation that best fits the data, representing the relationship between the dependent variable and each independent variable, an example of what it graphically looks like can be seen in Figure 2.3. This equation includes coefficients for each independent variable, indicating the magnitude and direction of their impact on the dependent variable while holding other variables constant. The performance of linear regression is reliant on the correlation of the data and so it is used in conjunction with correlation metrics such Pearson's correlation [31].

Throughout the research, many equations of linear and non-linear regression models were identified. They were then compared by metrics such as R^2 as it is an indicator of how well the model fits the data. One equation was chosen out of all the models gathered by comparing the R^2 values, unless there was only one equation found for a specific DBP.

In table 2.2 we can see that various parameters were used in making linear as well as non-linear equations for each DBP where pH, UV_254, DOC/TOC, chlorine dosage and chlorine reaction time were the most common and impactful parameters considered.

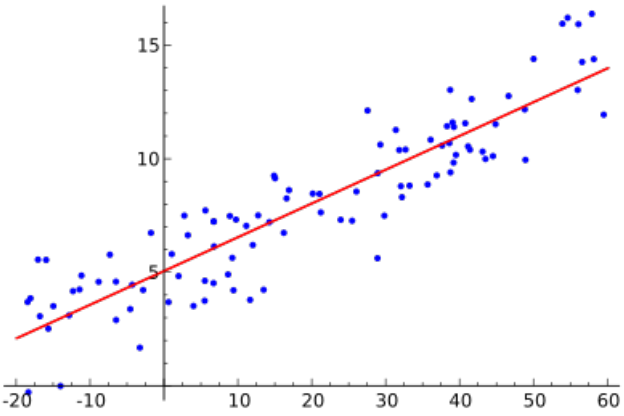

FIGURE 2.3: Linear Regression [31]

2.5.2 Parameter Values for THMs

As one of the most prevalent and researched DBPs, THMs had multiple models and hence there were more equations that were comparable to each other.

Table 2.3 is a table where focus was laid on THM and the values for the parameters of logarithmic regression equations. Regular linear regression equations were taken the log of in order to standardize the equations and make the values of the parameters comparable. Thus, non-linear equation where parameters were multiplied by each other or squared could not be evaluated together with the linear equations as they do not mean the same thing from a mathematical standpoint. From the table we can see that pH has a the most positive correlation with the formation of THMs, followed by UV254 and chlorine consumed/dose.

2.5.3 Machine Learning Methods

With developments in artificial intelligence and data science, supervised machine learning has been a increasingly common way for researchers to model the formation of DBPs. Many different techniques exists and the most commonly used ones were artificial neural networks (ANN) and fuzzy interference systems (FIS). Kulkarni and Chellam demonstrates that their ANN model identified DOC and bromide concentrations to be the most important parameter for both THMs and HAAs with slight differences in the numbers depending on whether the water was pre-treated/filtered or not [35]. Okoji et al. explored the use of FIS to predict DBPs and found that DOC and UV254 are the most important factors contributing to the formation of THMs along with pH and temperature [27].

DBP	Example Equation	R^2	Reference
THMs	$THM = 10^{-0.038} \times (Cl2)^{0.654} \times (pH)^{1.322} \times (time)^{0.174} \times (SUVA)^{0.712}$	0.88	Uyak et al. [32]
CF/TCM	$\log(CF) = -1.935 - 0.2393 \times \log(Br) + 0.0170 \times (temp) - 0.0012 \times (alk) + 0.1993 \times \log(toc) + 0.4450 \times \log(uv) + 0.3824 \times \log(cl2) + 0.0921 \times \log(t) + 0.1133 \times \log(pH)$	0.6854	Oblensky and Singer [28]
BDCM	$BDCM = 10^{-1.188} (Br)^{0.411} (UV254)^{1.042} (t)^{0.259} \times (Temp)^{0.560} (pH)^{1.732} (Cl2/DOC)^{0.238} \times (R2 = 0.972, p<0.0005, n = 36)$	0.972	Hong et al. [33]
HAA9	$HAA9\ (\mu g\ /\ L) = -345 + 1.695(Temperature) + 93.1(pH) - 226(UVA_{254}) + 4.95(Cl_2) + 5.66(NO_2^- - N) + 16.6(DOC) + 0.325(NH_4^+ - N) - 0.0693(Temperature)^2 - 6.41(pH)^2 + 190821(UVA_{254})^2 - 1.73(NO_2^- - N)^2 - 3.77(DOC)^2 - 0.01663(NH_4^+ - N)^2$	0.811	Okoji et al. [27]
TCAA	$TCAA\ (\mu g\ /\ L) = 11.47 - 1.42(Temperature) - 2.26(pH) + 5.71(UVA_{254}) + 1.39(Cl_2) - 3.11(NO_2^- - N) + 3.42(DOC) - 1.458\ (NH_4^+ - N) - 2.86(Temperature)^2 - 1.32(pH)^2 + 3.63(UVA_{254})^2 - 3.56(NO_2^- - N)^2 - 1.45\ (DOC)^2 - 3.196(NH_4^+ - N)^2 - 2.11(Br^-)^2$	0.818	Okoji et al. [27]
DCAA	$Ln(DCAA) = 6.256 + 0.643 \times Ln(UV254)$	0.768	Peng et al. [34]
HANs	$T\text{-}HANs = 10^{-1.065} (Br)^{0.346}(DOC)^{0.369} \times (Cl2/DOC)^{0.520} (t)^{0.238} (Temp)^{0.373} \times (R^2 = 0.943, p<0.0005, n = 36)$	0.943	Hong et al. [33]
DCAN	$DCAN = 10^{-0.583}(Br)^{-0.581} (t)^{0.297}(Cl2/DOC)^{0.577} \times (DOC)^{1.452} (Temp)^{0.472} \times (R2 = 0.933, p<0.0005, n = 36)$	0.933	Hong et al. [33]
CH	$Ln(CH) = 8.945 + 0.558 \times Ln(UV254)\ 2.37 \times Ln(pH) + 0.152 \times Ln(TOC)$	0.752	Peng et al. [34]

TABLE 2.2: The equations and R^2 value per DBP from literature

In general the parameters that are shown to be the most correlated are ones related to NOMs, pH and temperature. Overall, machine learning has been shown to score better in statistical values against testing data.

Parameter	Value
pH	1.576
UV254	0.3235
Temperature	0.015
Reaction time	0.1465
TOC	0.188
chlorine consumed	0.291
chlorine residual	0.167

TABLE 2.3: The parameters and their average values from logarithmic regression equations for THMs found in literature

Table 2.4 is a summary and comparison of methods of each ML method used for DBP formation and the benefits vs limitations of it.

2.6 Conclusion and Discussion

To conclude, this literature review aimed to answer the research questions stated in Subsection 1.2, and highlighted the critical importance of understanding the formation of disinfection byproducts (DBPs) in water treatment processes. By exploring the various environmental parameters implicated in DBP formation (NOMs, pH, bromide concentrations, chlorine dose, temperature) and evaluating the existing modelling techniques for prediction of DBPs (regression and supervised machine learning), valuable insights have been made in order to mitigate the risks associated with DBP formation.

The limitations of these studies however, is that a majority only focus on THMs and HAAs which are already regulated while there are only few models and parameters identified for emerging DBPs that also have the potential to be harmful to humans. Furthermore, the lack of openly available data make it harder to make and fully trust all kinds of mathematical and AI models as research conducted on small sample sizes tend to be unreliable as a general scientific practice. On top of that, climate change and pollution continuously affect our water quality from temperature to other kinds of NOMs and hence the models that were made a decade ago might not be relevant anymore.

For future research, scientists should incorporate models that can change over

Method	Benefits	Limitations
Multiple Linear Regression (MLR)	Simple, indicates strength and direction of each coefficient	Have to choose parameters carefully, assumes linearity, outliers shake it up
Artificial Neural Network (ANN)	For comple×nonlinear relationships, tolerant to missing values, automatic learning	Black box, need a lot of data, computationally intensive
Random Forest (RF)	Robust to overfitting, handles missing data well	Accuracy and robustness determined by the "density" of decision trees, More memory and resources needed
Adaptive Network-based Fuzzy Inference System (ANFIS)	Handles nonlinear relationships, good with vague or uncertain data as outputs and decisions are easy to interpret with a well-defined system	Applicability dependent on operator-defined parameters and experience-prone to human error, limited scalability for large datasets
Radial Basis Function Kernel (RBF)	Versatile, widely applicable	Performance depends on problem and choice of algorithm, can be impractical

TABLE 2.4: Machine learning methods and its benefits and limitations [12]

time based on the continuous data being collected to achieve real time monitoring of water treatment plants and also have a central database for water quality data in order to make more reliable prediction models. This should further improve the knowledge and understanding of DBP formation and help inform governments on what regulations are necessary.

Chapter 3

Methods and Techniques

In this chapter, the methods and techniques that are used for realising the background information, and overall project are described in each subsection. Section 3.1 will briefly go over the literature search strategy, Section 3.2 will dive into one of the processes used throughout the project, Section 3.3 explains the data science life cycle, Section 3.4 highlights the different metrics used to evaluate the models, Section 3.5 explains the model testing method and 3.6 mentions the software and tools used.

3.1 Literature Search Strategy

To find appropriate literature several methods were used. First and foremost google scholar was used by searching with keyword combinations where the main keyword was "DBP" or "disinfection byproduct" followed by others such as "model", "regression", "water parameters", "water quality", "machine learning", "prediction". Then filters were used to search through most recent papers first (especially for literature reviews) and each title was read, where papers with potential added up into the browser as a new tab.

From there on, a reference managing software named Zotero was used to "collect" papers/literature that seemed to be relevant after reading the abstract and scanning through them. It started out with a lot of papers which got narrowed down to much less after careful inspection of the actual subject matter of the papers, such as relevant disinfection method (chlorine), relevant methods of analysis (regression, ML) and whether the publication was to be trusted or not, from there on it reduced in number because some

papers had similar content and did not add much more to the research.

Furthermore, literature reviews were looked at first to get a broader idea of the context and past research that has already been concluded. While reading them the snowball method was used where papers were found through looking at the sources/papers that literature reviews were referring to and adding them to the list if deemed relevant.

Lastly, literature was searched for regression equations and metrics that validate the models. The model was checked whether it used water quality parameters or not and each equation was noted down.

3.1.1 Selection of DBPs

Since there are many DBPs, it was useful to narrow down a list of DBPs to consider while doing research. The list of priority of research was based on a previous study that assessed the health impacts of each DBP [9]. 'Table 5' in that paper shows a list of DBPs in order of THMs, HAAs, HANs and oxyhalide, NNAs, etc. This order was kept in mind when looking for papers, equations and data. The data available also only had data on regulated DBPs that are considered the most dangerous and hence it seemed like a good option to use despite the drawbacks.

3.2 CreaTe Design Process

Normally, the graduation projects from the Creative Technology study follow the steps from the *Design Process for Creative Technology* written by A. Mader and W. Eggink [36].

Projects following this design process begin with a divergent phase, where various potential solutions are explored. This is followed by a convergent phase to narrow down these ideas to a single solution. During the transition from divergence to convergence, reflection occurs at key stages, allowing earlier decisions to be revisited and adjusted as necessary. The process is not strictly linear; instead of taking one large divergent step and one large convergent step, several rounds of divergence and convergence are conducted. This approach allows for the integration of new information at each stage.

The Creative Technology design process specifically consists of four phases: Ideation, Specification, and Realisation, followed by an Evaluation phase. Each of these phases in-

volves repeated cycles of divergence and convergence, and earlier phases can be revisited if new insights are gained later on. This iterative process is visually represented in Figure 3.1.

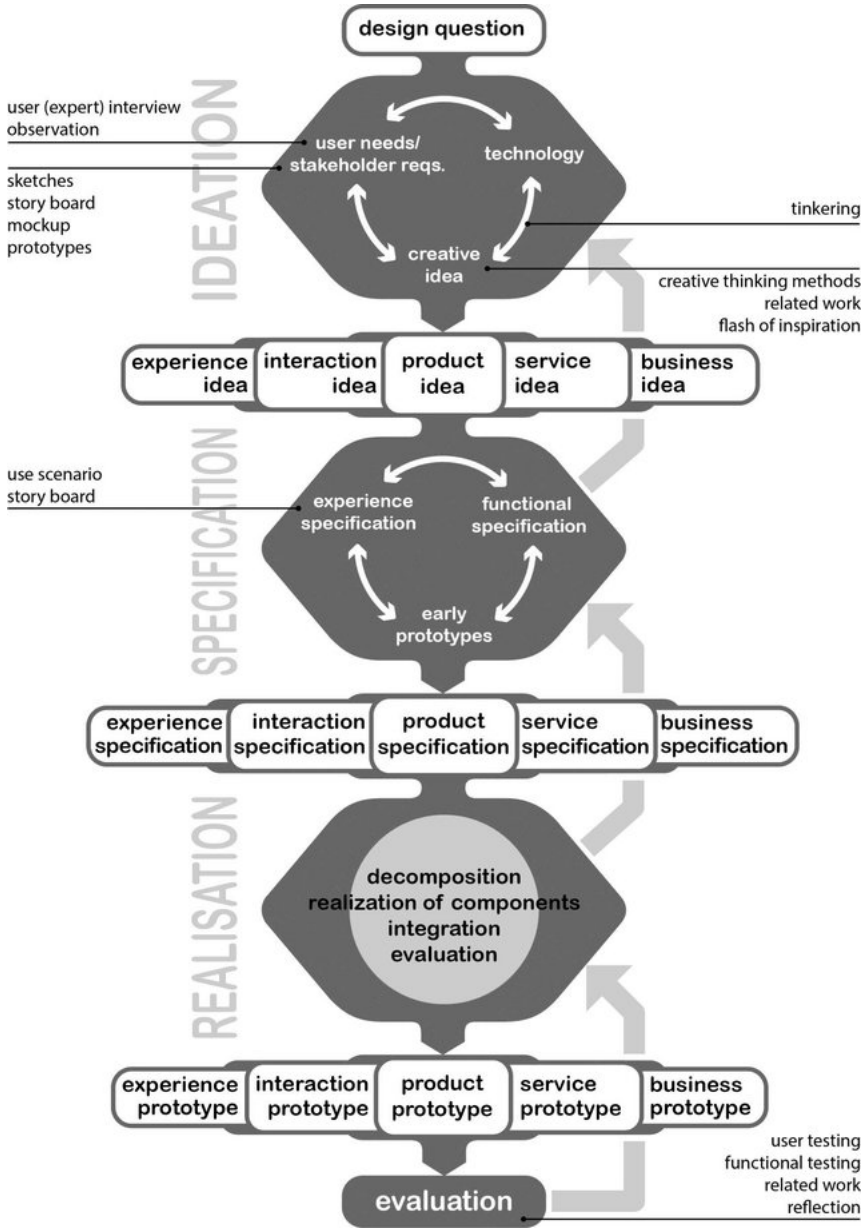

FIGURE 3.1: The CreaTe Design Process [36]

For this project, the stages are largely followed but with some consideration to the making of machine learning models. For the ideation stage, the concept of creating a machine learning model for predicting DBP formations was mostly led by the literature and state of the art research. Low fidelity prototypes such as sketches or paper prototypes did not seem very appropriate in conveying the design unlike other create projects, as it

was going to be a piece of software. The literature research helped in making decisions for which specific algorithms or methods could be successfully used for the problem and which metrics should be used to evaluate it. Lastly, this time was also spent on learning more about statistics and machine learning theory as well as the possible ways to implement it and thinking about the next steps.

The specification and realisation stages for this project blend into each other as selected machine learning models will be trained in iterations, tested and evaluated to make choices on the data and models in order to reach a better performing model which follows the steps of the data science life cycle. Lastly, the project will conclude with an evaluation of the best-performing model and against previous existing models.

3.3 Data Science Life Cycle

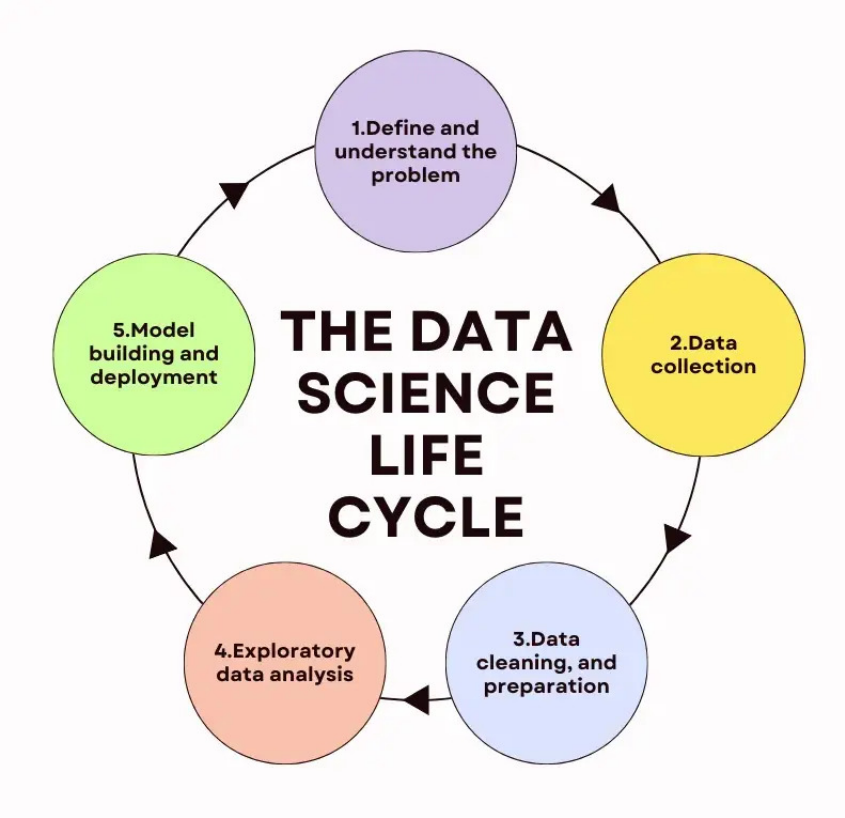

FIGURE 3.2: The Data Science Life Cycle by Madison Hunter [37]

As mentioned previously, the data science life cycle was kept in mind when doing this project. Just like the Creative Technology design process, machine learning development is a cycle where earlier stages will have to be re-examined to improve the performance

of the machine learning model.

3.3.1 Understanding the Problem

Looking at Figure 3.2, the first step is to define and understand the problem. It is important to know what the goal is and to think of the steps in order to achieve it. The goal in terms of this project is to understand the relationship/correlation between water parameters and DBPs better and make a machine learning model that can predict the concentration of a given DBP. The goal can also change for example in the earlier stages when data exploration was key to work out the best approach to making the model.

3.3.2 Data Collection

It is important to collect good data and databases were found through using *Dataset Search* by Google and also reading the methodology part of papers to see what data they used to make their models. Easily available databases were only on water quality of a specific region/treatment plant, which include some interesting fields such as pH and DOC. However, those did not include any information about DBP concentrations and hence it was left out as it would not be applicable for this project. An ideal database would have all the important water parameters mentioned in the literature, have a large amount of data that can be trained, be consistent in definition and labels and should be timely to the situation. The latter two features have been met in the dataset used explained further in chapter 4.

3.3.3 Data Cleaning and Preparation

The data that is collected might need some cleaning and organisation or some scaling such that it can be correctly interpreted by a machine learning model. For example, there might be incorrectly formatted data, corrupt data, duplicates or null values and extreme outliers. The data cleaning process should be done but its crucial not to spend too much time on perfecting the database and instead testing bit by bit to see how it affects the results.

3.3.4 Exploratory Data Analysis

The fourth step is the exploratory data analysis phase where it is done to summarize the main characteristics of a data set and it consists of making data visualisations in order to

quickly see patterns or anomalies in the data. It will be useful for developing the models later on. Examples that are used in the following chapter are summary statistics, scatter plots, pie charts and histograms.

3.3.5 Model Building

The last step is for building the actual model and deciding on what sort of model to make. Machine learning models can either be supervised or unsupervised, where supervised means it will use training data to 'learn' patterns in order to classify unseen data or forecast future trends while unsupervised models find similarities within the data, understand relationships between different data points and perform additional data analyses. This project will use supervised modelling as it might not have enough data for a good unsupervised model.

3.4 Metrics Used

The metrics used for evaluating the models are those most commonly found in literature on machine learning. Below is a list with brief explanations of the metrics used where the first one is just for general data analysis:

- Pearson Correlation (r): linear correlation between two sets of data (ranges from -1 to 1)
- Coefficient of Determination(R^2): proportion of variance in the dependent variable ; how well the model fits the data (from 0 to 1, but can be negative)
- Mean Squared Error (MSE): the average squared difference between the estimated values and the actual value (lower value is usually better)
- Root Mean Squared Error (RMSE): how well the model is able to predict the target value/accuracy (the lower the value, the better)
- Mean Absolute Percentage Error (MAPE): measures the prediction accuracy of the model; how far off predictions are on average (the lower the percentage, the better)

3.5 Model Testing Methodology

With machine learning models, in order to test the performance, the dataset is commonly split into a training dataset and a testing dataset. The training part is used to train the chosen model and the testing part is used for evaluating the performance of the trained model on data that is previously not seen by the model during training. This section describes how that was implemented in the project.

3.5.1 Splitting the Data

As previously stated, the models were initially evaluated using a simple train-test-split approach. This method involves dividing the entire dataset into two parts: a training set and a testing set. The larger training set is used to train the model, while the smaller testing set is used to assess the model's performance on unseen data.

For this project, the dataset was randomly divided into a 70% training set and a 30% testing set, a common ratio suitable for the dataset's relatively large number of rows. This was implemented in the Python code using the train_test_split() function from the sklearn.model_selection library. The test_size parameter was set to 0.3, allocating 30% of the data to the testing set. The random_state parameter was set to 42 to ensure reproducible results across multiple runs and iterations.

3.6 Software and Tools

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

from sklearn.linear_model import LinearRegression, Ridge
from sklearn.model_selection import train_test_split
from sklearn.metrics import r2_score
from sklearn.metrics import mean_absolute_percentage_error
from sklearn.metrics import mean_squared_error
from sklearn.metrics import root_mean_squared_error
```

FIGURE 3.3: The libraries and specific tools imported for realising the project

For developing prediction models and doing the data analysis, Python 3.11 will

be the main language used for programming as it has a lot of useful libraries for machine learning such as *numPy* and *scikit-learn*. Another library called *seaborn* [38] and *matplotlib* was also used for making plots and visualizations. From *scikit-learn*, out of the linear models available mainly the LinearRegression() and Ridge() functions were used for training the model, the reasons explained in chapter 5. Figure 3.3 shows the imports and libraries used for implementing the project.

Chapter 4

Ideation and Data Exploration

The aim of the ideation process for this graduation project was to identify the project needs and desired outcomes. Further it was to establish the steps needed to develop the data analysis and regression model.

4.1 Data Collection and Compilation

4.1.1 The EPA Database

Eventually what was found, was a lot of separate files on measurements taken ranging from water parameters to regulated DBPs at water treatment plants from the United States. The fourth Six-Year-Review database is a database collected by the Environmental Protection Agency (EPA) from January 3rd 2012 through December 31st 2019 [17]. It contains the data of water treatment plants, ranging across 50 states, 10 Regions, and several territories of the United States as seen in Figure 4.4. All DBP concentrations have a unit of μg/L.

4.1.2 Missing Data

As table 4.1 shows, a lot of data is missing where 3 out of 6 water parameters (indicated by 'p') have more than 95% missing/null data and 5 out of 13 DBPs have more than 77% of the data missing and hence pre-processing needed to be done in order to have workable/trainable dataset. This percentage was derived from the non-null count divided by 18499 which is the total amount of rows.

#	Column	Dtype	Non-Null Count	Percentage Missing
1	PWSID	object	18499 non-null	0%
2	DATE	object	18499 non-null	0%
3	p_PH	float64	6237 non-null	66.28%
4	p_DOC	float64	62 non-null	99.66%
5	p_TOC	float64	9877 non-null	46.61%
6	p_ALKALINITY	float64	13051 non-null	29.45%
7	p_SUVA	float64	421 non-null	97.72%
8	p_UVA	float64	62 non-null	99.66%
9	d_BROMATE	float64	397 non-null	97.85%
10	d_CHLORITE	float64	2004 non-null	89.17%
11	t_BROMOFORM	float64	5512 non-null	70.20%
12	t_CHLOROFORM	float64	12017 non-null	35.04%
13	t_BROMODICHLOROMETHANE	float64	12027 non-null	34.99%
14	t_DIBROMOCHLOROMETHANE	float64	10559 non-null	42.92%
15	t_TTHM	float64	12886 non-null	30.34%
16	h_DIBROMOACETIC_ACID	float64	4203 non-null	77.28%
17	h_DICHLOROACETIC_ACID	float64	8023 non-null	56.63%
18	h_HAA5	float64	11550 non-null	37.56%
19	h_MONOCHLOROACETIC_ACID	float64	2556 non-null	86.18%
20	h_MONOBROMOACETIC_ACID	float64	1466 non-null	92.08%
21	h_TRICHLOROACETIC_ACID	float64	7383 non-null	60.09%

TABLE 4.1: Overview of dataset with the percentage of missing data

4.2 Data Pre-Processing

The aforementioned EPA database had to be combined from several files into one dataset which was kindly done by Balaram Guddanti in a form of an excel spreadsheet. The columns consisted of the water treatment plant id ('PWSID'), the date, parameters ('p') and DBPs ('t' for THM family, 'h' for HAA and 'd' for inorganic DBPs). He further processed it to contain rows where each row has at least one parameter and one DBP concentration value, which is the same dataset that is shown in Table 4.1.

4.2.1 Outliers

Further data pre-processing was done such as taking out obvious outliers by filtering by the Inter-Quartile Range (IQR) from the summary statistics. Data was also filtered per column to see if any values were strange. Noticing that a maximum value for some DBPs were in the 1000s while the lowest and usual measurements were below 100s made it obvious that there were some outliers that need to be removed. The columns that had outliers are listed below with the value that it has been filtered on. The code below shows the final set of outliers:

```
data = data.drop(data[data['t_BROMODICHLOROMETHANE'] >= 200].index)
data = data.drop(data[data['h_MONOCHLOROACETIC_ACID'] >= 200].index)
data = data.drop(data[data['t_CHLOROFORM'] >= 400].index)
data = data.drop(data[data['p_ALKALINITY'] >= 600].index)
```

4.2.2 Further Methods

At this stage a plan was made for further tackling missing data. One solution was to take the mean of every column and fill that into the unknown data. From testing this and reflecting on it some other methods were proposed, where the details are written in Chapter 5.

4.3 Data Exploration and Analysis

4.3.1 Summary Statistics of the Parameters

Parameter	Mean	Std	Min	Max	Units
pH	7.55	0.58	0.43	13	n/a
DOC	2.71	1.25	0.73	5.55	mg/L
TOC	2.91	1.83	0.001755	31.6	mg/L
Alkalinity	125.32	103.12	0.0134	1430	mg/L
SUVA	2.38	0.70	0.94	8.975	L/mg-m
UVA	4.01	3.03	0.008	9.4	cm-1

TABLE 4.2: Summary Statistics of the Parameters

One of the first things to find out when doing data exploration is the summary

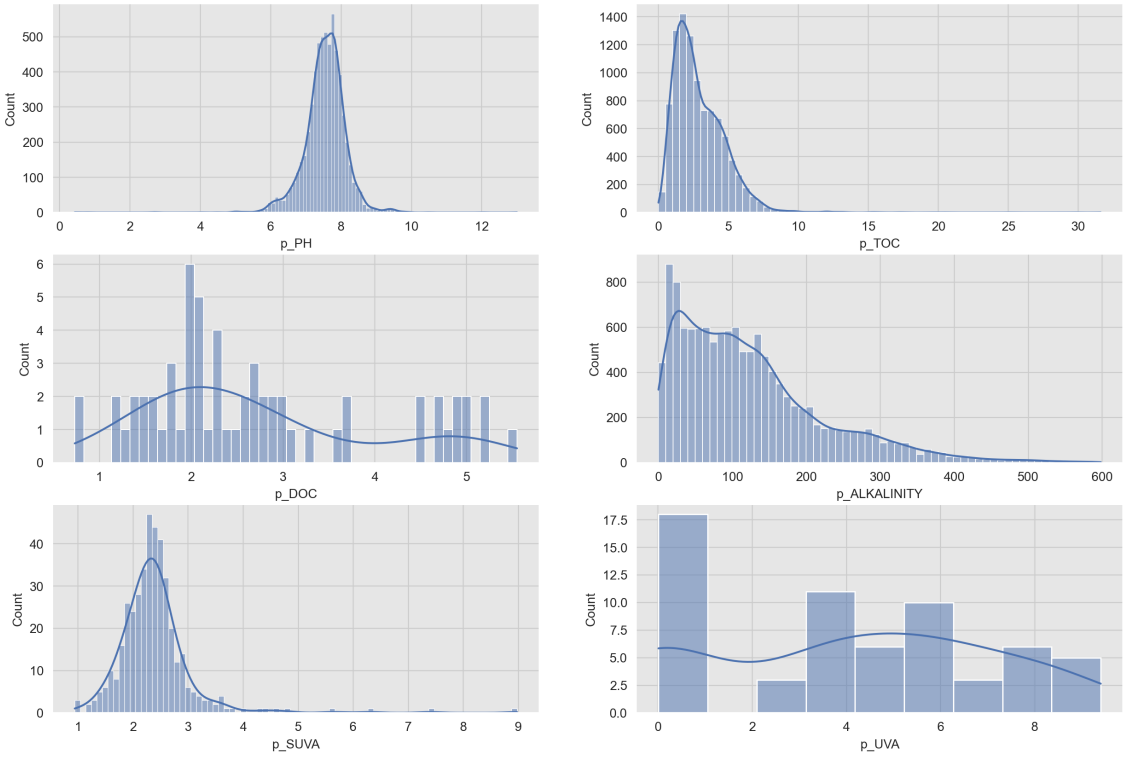

FIGURE 4.1: Distribution of Parameters

statistics. Table 4.2 is the summary statistics of the dataset after outliers have been weeded out. As mentioned it was also generated before the outliers in order to find any anomalities in the data. The range of pH is quite large where the minimum value is near 0 while the max is 13 but Figure 4.1 shows that the distribution is quite normal. SUVA also has a normal distribution but in the lower range while TOC and Alkalinity have slightly skewed distributions. DOC and UVA have very strange distributions and hence it might be a bottleneck later on.

4.3.2 Parameters vs DBPs

The second visualisation that was explored was scatterplots of parameters vs DBPs with a red line indicating the regulation limit for that DBP. Figure 4.2 is a curious example where most of the data points were above the regulation limit. SUVA seems to go in an upwards correlation motion while UVA has no correlation at all. (Perhaps at this stage considerations for dropping some parameters should have occured).

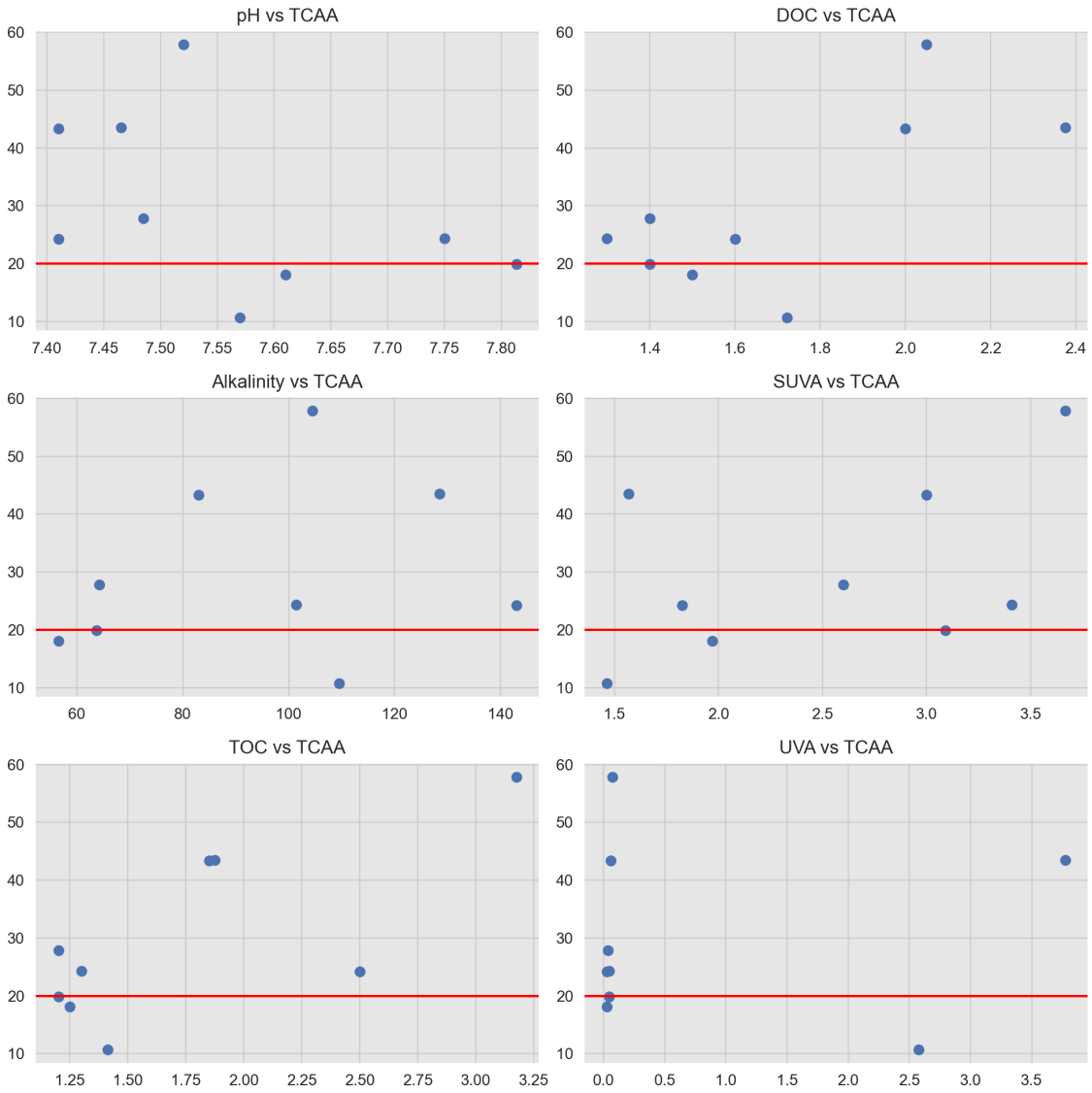

FIGURE 4.2: Parameters vs TCAA (μg/L)

4.3.3 Correlation Heatmap

To better understand the correlations between data a correlation heatmap was carefully crafted where attention was paid to the color so that it was more obvious which correlations were important and in which direction. As one can see in Figure 4.3, there are big correlations between water parameters which were only considered much later in the project as an issue/something to fix.

Furthermore, Chloroform has medium correlation with DCAA and TCAA which both contain chlorine in their chemical composition. The same observation can be made about DBPs with Bromide and so it is logical to assume that when one form of chlorine containing DBP is formed, others will follow too, same for bromide. HAA5 has a really high correlation with DCAA which could mean that there is some bias where there is more data of DCAA available than other HAAs.

4.3.4 Investigation on Location

By extracting the first 2 letters of the 'PWSID' the data could further be categorised into states and then into regions according to the EPA as seen in Figure 4.4.

When further investigating the categorisation of states, it was explored which state had the highest average concentration of a certain DBP. pH was indifferent to location as expected, as water treatment plants are usually advised to keep the pH at a certain level. Additionally, Colorado and Pennsylvania are the only states that have data on the parameters: DOC, UVA and SUVA which is not a great distribution and could potentially mean some bias for these States Texas makes up 21.3% of the datapoints followed by California with 11%. Furthermore, California and Arizona (Region 9) seemed to have the highest average DBCM, while regions with highest averages for bromoform all have access to the ocean. This is no coincidence as the sea is rich in bromide [40].

Lastly, Figure 4.5 shows the distribution of regions and its noticaeble that the 3 biggest regions make up more than half of the dataset which could indicate some bias for the water quality / environment from those regions.

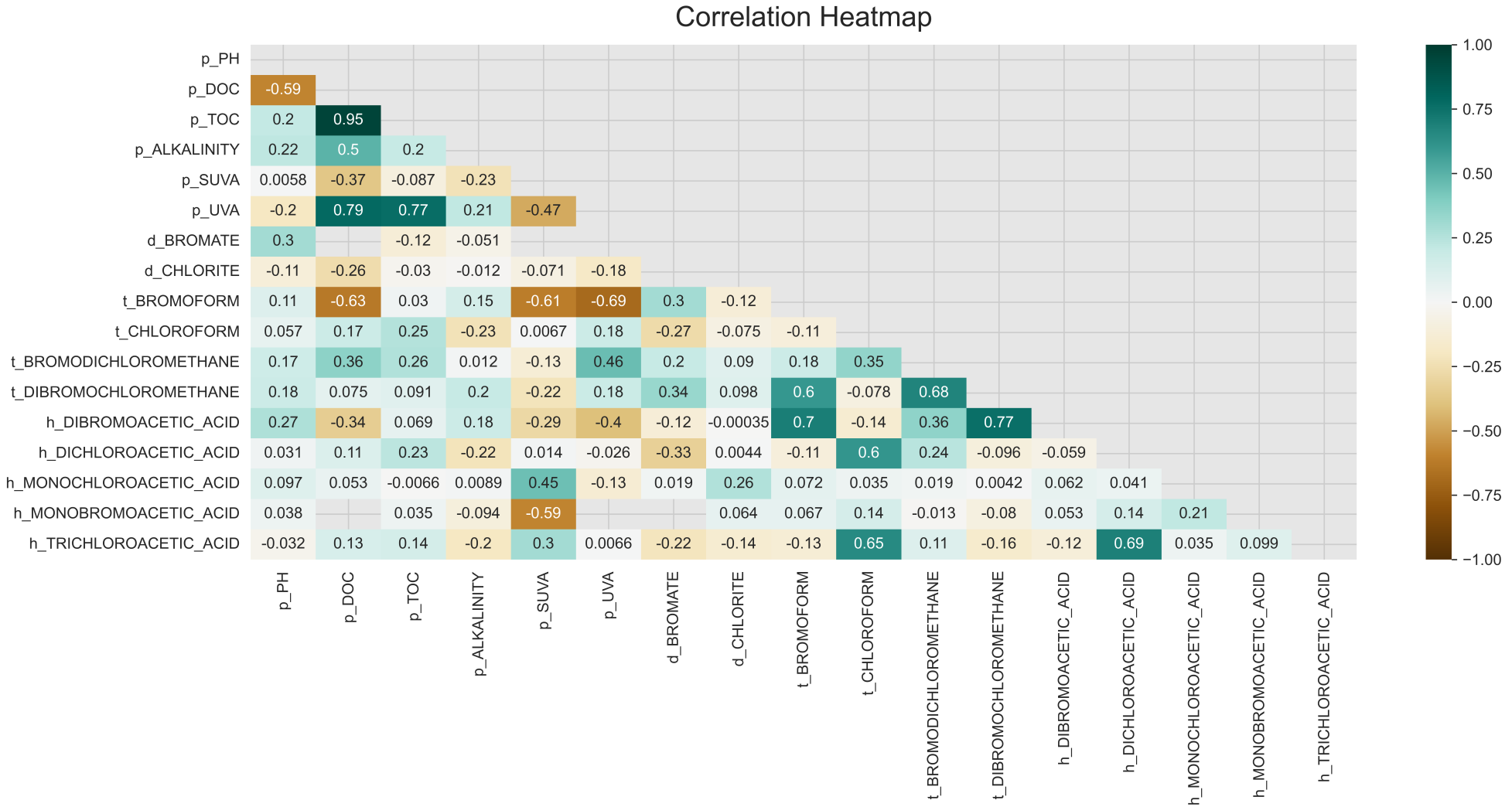

FIGURE 4.3: The Correlation Heatmap of all columns

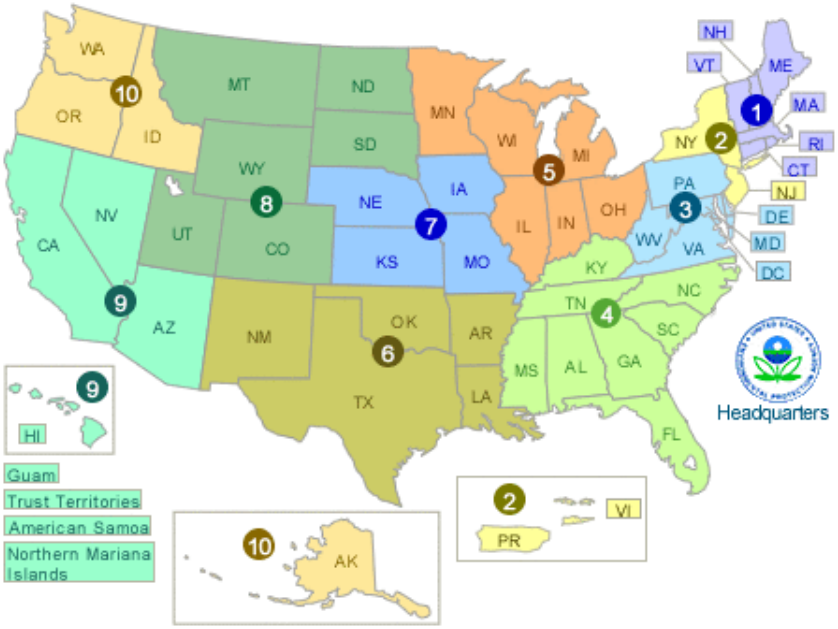

FIGURE 4.4: How the states are divided into regions [39]

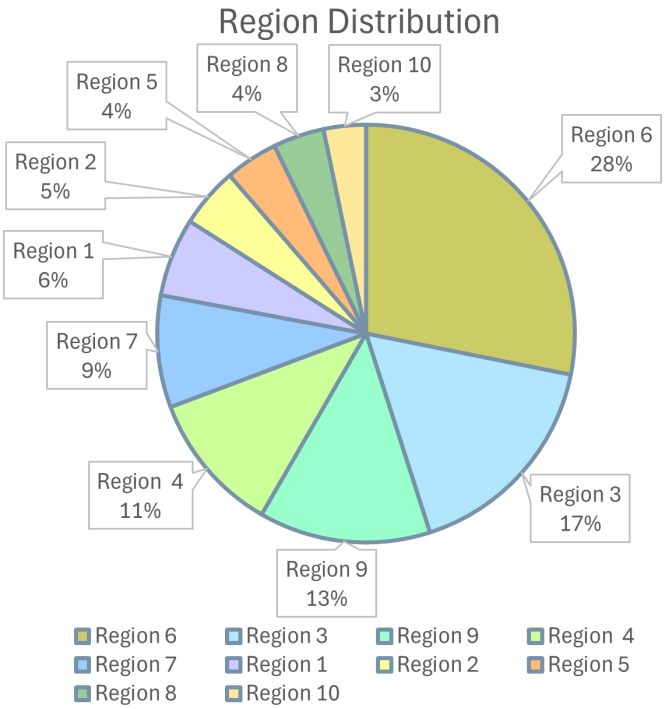

FIGURE 4.5: The Regions and the corresponding make up of the dataset [39]

4.4 Next Steps

The next steps that would be iterated to meet specific solutions to problems, concluded from this chapter for the realisation phase are:

1. Pre-processing the data of the EPA database (with different methods to mitigate missing data)

2. Train the data with various regression models

3. Calculate regression functions and measure metrics such as R^2, MAPE, MSE, RMSE [41]

4. Validate these models by comparing them to other models and test data to see effectiveness of predicting DBPs

Chapter 5

Realisation and Results

This chapter aims to describe the processes of each iteration and the results along with a reflection. Section 5.1 , Section 5.2 and Section 5.3

5.1 Iteration 1: Simple Regression

The first iteration was the most simple linear regression model I could make by replacing null values with the mean for each column of parameters done by the code:

```
# blindly fill parameter NAs with mean
data[PARAMETERS] = data[PARAMETERS].fillna(data[PARAMETERS].median())
```

5.1.1 Process

Python notebooks were used instead of regular python files as it is more efficient to run chucks during testing so that when you want to change a snippet of code not everything needs to be run again. The outliers were taken out from the data and a copy of the data was made. As you can see in Figure 5.1, After just isolating the parameters and targeted DBP (in this case the total THMs), NA (cells with no data) were dropped in order for the LinearRegression() function to work [42]. The train_test_split is done and the 'score' calculates the R^2 value.

```python
data_thm = data.copy()

# remove HAA column
data_thm = data_thm.drop(['h_HAA5','d_BROMATE', 'd_CHLORITE'], axis=1)

# drop target NAs
data_thm = data_thm.dropna()

# split the data
X = data_thm[PARAMETERS]
y = data_thm['t_TTHM']

X_train, X_test, y_train, y_test = train_test_split( X, y, test_size=0.3, random_state=42)

# fit a regression to the training data
reg_t = LinearRegression().fit(X_train, y_train)

# evaluate on the testing set
print(reg_t.score(X_test, y_test))
```

FIGURE 5.1: Snippet of code from the first iteration

5.1.2 Results

So far since it was just a test to see how to code linear regression models, the metrics collected are only that of the aggregated THM and HAA columns with R^2 as the only parameter. These correspond to 0.078 for THMs and 0.099 or 0.1 for HAAs. Since both values are more near the 0 than the 1, it is safe to assume that the model did not fit very well.

5.1.3 Reflection

While having a go at trying to see if models found in literature could be used to for my own dataset, I realised that I could only compare it against few models/equations as I did not have the same parameters as them. Upon consultation with my supervisor we decided to use additional data from a water treatment plan in Coimbra, Portugal to see if filling those gaps would improve the predictions. That process is described iteration 2.

5.2 Iteration 2: More Parameters

The second iteration thus included two additional parameters from Coimbra's data where the values were 0.174 for the chlorine dose (mg/L) and 20 degrees for temperature:

```
data['p_CHLORINE_MG_L'] = 0.174
data['p_TEMPERATURE'] = 20
```

5.2.1 Process

The same code was used as the previous iteration where the only difference being those two additional columns.

5.2.2 Results

Because the added values are constant they barely added anything to the results, which is logical because having a constant means that the regression line will always pass that point / would be on the same height as the y-intercept (if said constant was the only parameter). Nevertheless the results are 0.068 for THMs and 0.088 or 0.9 for HAAs which is technically worse than before.

5.2.3 Reflection

What I realised at this stage is the fact that there is a lot of correlation going on between the parameters and wondered whether that would be an issue. I stumbled upon the term multicollinearity and looked up ways to mitigate it / what to do about it. It led me to find out about another regression model, the Ridge Regression. When the issue of multicollinearity occurs, least-squares are unbiased, and variances are large, which results in predicted values being far away from the actual values [43].

Also since the results didn't improve I decided to entirely ditch those columns and went back to my original dataset.

5.3 Iteration 3: Testing Different Models

The last iteration consisted of comparing two different regression methods for all DBPs.

5.3.1 Process

To optimalise the process of testing many DBPs a general function was made with parameters such as the type of regression (regressor), DBP and an array of parameters to potentially remove in order to get results. In the case that there was no output due to not having enough data, the parameter columns that were removed were the ones with the least amount of data (and most NAs) which made less rows drop/dissapear when doing 'dropna()', hence resulting in a score.

```python
def regress_DBP(regressor, dbp, extra_removals=[]):
    data_dbp = data.copy()
    not_dbp = INDIVIDUAL_DBPS.copy()
    not_dbp.remove(dbp)
    # target only MBAA
    data_dbp = data_dbp.drop(TOTAL_DBPS, axis=1)
    data_dbp = data_dbp.drop(not_dbp, axis=1)
    data_dbp = data_dbp.drop(extra_removals, axis=1)

    # drop target NAs
    data_dbp = data_dbp.dropna()

    # split the data
    X = data_dbp[[param for param in PARAMETERS if param not in extra_removals]]
    y = data_dbp[dbp]

    split = train_test_split(X, y, test_size=0.3, random_state=42)

    X_train, X_test, y_train, y_test = split

    # fit a regression to the training data
    reg_dbp = regressor.fit(X_train, y_train)

    # evaluate on the testing set
    score = reg_dbp.score(X_test, y_test)

    return reg_dbp, score
```

FIGURE 5.2: The general regressing function

5.3.2 Results

Many calls to the general regression function with different DBPs and methods gave the result shown in Table 5.1.

DBP	R^2 value using LinearRegression()	R^2 value using Ridge()
THM	-1.32	-0.98
CF	-2.53	-1.55
BF	0.0673	0.501
BDCM	-0.302	-0.49
DBCM	-1.014	0.11
HAA	-6.918	-2.06
TCAA	0.085	0.13
DCAA	-10.728	-10.73
MCAA	-2780.64	-508.60
DBAA	-29.58	-47.59
MBAA	-54.98	-55.03
Bromate	0.0000045	-0.48
Chlorite	-0.062	-0.058

TABLE 5.1: The R^2 value per DBP per Method

5.3.3 Reflection

The results show that bromoform (BF) makes a huge jump in terms of how well the model fits, which has the highest score of 0.5; others don't differ much in the sense that the scores are pretty much equal except MCAA where it goes from -2780 to -508 which is still really bad but is a noticeable difference. We can not be certain on how reliable this 0.5 value is and maybe it is overfitting because it has a very small dataset size.

Chapter 6

Evaluation, Discussion and Future Work

This chapter aims to answer the research question of how effective current methods/models are in predicting DBPs. Section 6.1 shows how the results are evaluated, Section 6.2 discusses the evaluation results and lastly Section 6.3 will talk about future direction this project can go into.

6.1 Evaluation

For the evaluation phase, it is important to evaluate the models made against a state-of-the-art example. Therefore the equations derived from literature were tested out. The code in Figure 6.1 was used to see whether the equation applied to the dataset.

6.1.1 Applying Equations from Literature on Dataset

In order to evaluate whether the equations found in Chapter 2 apply universally to any dataset on water parameters and DBPs, a test was carried out. All equations are from a paper by Peng et al. [34], as that were the only paper that had matching parameters.

6.2 Discussion

Looking at Table 6.1 it is clear that the models are not universal because the R^2 values are very negative which means that the model's predictions are worse than a constant

```
# Ln(THMs)= 1.579 + 0.477 x Ln(UV254)+ 1.829 x Ln(pH) (R2 = 0.594)

data_thm = data.copy()

# target only totals
data_thm = data_thm.drop(INDIVIDUAL_DBPS, axis=1)

# remove HAA column
data_thm = data_thm.drop('h_HAA5', axis=1)

# drop target NAs
data_thm = data_thm.dropna()

data_thm_formula = data_thm.copy()
data_thm_formula = data_thm_formula[['p_UVA', 'p_PH', 't_TTHM']]
data_thm_formula = np.log(data_thm_formula)
data_thm_formula['pred_THM'] = 1.579 + 0.477 * data_thm_formula['p_UVA'] + 1.829 * data_thm_formula['p_PH']

r_squared = r2_score(data_thm_formula['t_TTHM'], data_thm_formula['pred_THM'])
rmse = root_mean_squared_error(data_thm_formula['t_TTHM'], data_thm_formula['pred_THM'])
mape = mean_absolute_percentage_error(data_thm_formula['t_TTHM'], data_thm_formula['pred_THM'])


corr = data_thm_formula['t_TTHM'].corr(data_thm_formula['pred_THM'])
mse = mean_squared_error(data_thm_formula['t_TTHM'], data_thm_formula['pred_THM'])
```

FIGURE 6.1: An example code used to derive the metrics while applying the equation

function that always predicts the mean of the data. RMSE and MAPE are higher than the original equation which also indicates a bad performance. The reasons for this could be that the original equations are fitted to a dataset which took samples from China that is not close to the sea, while the dataset that was used for this project came from the US which has many different climates across the country and has a lot of states that have access to the sea. Bromide is the seventh most abundant minerals in the sea [40] which is also a precursor for B-DBPs such as Bromoform and hence could be a reason as to why bromoform could be predicted semi accurately and also why it didn't work on this project's database.

6.2.1 Limitations

Reflecting on the project, there are several ways in which it could have been improved for the next time. Firstly, the search for a better database should have been conducted more intensively. More e-mails could have been sent to authors, try to scrape the web for similar databases etc. This is because the results heavily depend on the quality of the database. There is not much to train when there is not a lot of data and smaller datasets increases the likelihood that the model over-fits because it does not contain enough data samples to accurately represent all possible input data values [44]. Additionally, more advanced

Equation	R^2 (paper)	R^2	RMSE (paper)	RMSE	MAPE(%) (paper)	MAPE (%)
Ln(THMs)= 1.579 + 0.477 × Ln(UV254)+ 1.829 × Ln(pH)	0.594	-4.71	0.184	1.334	3.745	32.0
Ln(HAAs)= 6.681 + 0.645 × Ln(UV254)	0.706	-22.41	0.183	2.373	3.266	65.4
Ln(DCAA)= 6.256 + 0.643 × Ln(UV254)	0.768	-14.64	0.156	1.615	3.081	61.7
Ln(TCAA)= 5.224 + 0.608 × Ln(UV254)+ 0.134 × Ln(TOC)	0.706	-2.38	0.185	0.88	4.325	34.7

TABLE 6.1: Equations and metrics derived from literature [34] compared against applied metrics

methods could have been used in outlier detections such as K-Nearest Neighbours or clustering methods in order to increase the data quality although it might reduce the risk of overfitting as well.

Another limitation, also due to not having enough data, is that there wasn't a clear baseline or target to meet and so the evaluation process has not been able to be as thorough as it should be. Perhaps hyperparameter tuning could have been done to delete parameters that don't add much to the dataset such as DOC, UVA and SUVA, but then again not many parameters would be left to analyse the data. In terms of outliers, only the upper bounds were considered which may affect the overall skewed-ness and bias of the dataset.

The limitations of the project in general is that we could not collect data ourselves as it is too expensive and not suitable for the Netherlands since it uses Ozone and UV

radiation. An interesting approach would have been to compare the formations of ozone and uv radiation versus that of chlorine to see whether alternative methods are better at mitigating DBPs.

As water and the environment constantly changes the likelihood that predictions will not be accurate anymore after a certain period of time is likely. Therefore in a real world context, such as a water treatment plant in Southern California who implemented site-specific THM4 models, proved that real time monitoring could be achieved as long as the prediction models were updated frequently [45]. This further shows the limitation of this project in a real world context.

6.3 Future Work

In the future, more careful planning should be taken into consideration throughout the whole data science process as better results could have potentially be had if there was more quality data or exploring other methods to make the model more robust against missing data.

Some ways in which the project could be improved on are perhaps doing a cross-validation of suitable models with k-folds in order to see which chuck of training data performs best / is the most accurate. Furthermore, the project could have developed into a combination of a classification algorithm and an aggregate regression model where it will first classify which DBP might be formed and then of how much. Classification could also be used to predict, given the input of water quality parameters, whether the DBP amount would go over the regulation limit or not by only focusing on the data where it goes over the limit. Lastly, since the date of measurement was available an analysis on the effect of seasons on DBP formation could have also been explored in another project.

Chapter 7

Conclusion

In a world where drinking water availability is put under pressure by many factors such as climate change, population growth and global industrialization, disinfection using chlorine seemed to be the most cost-effective solution until the discovery of DBPs. This made it clear that more research needed to be done on how to mitigate/regulate them, what the exact health effects are, the chemical composition of every possible and existing DBP, and how they might form.

Knowledge gaps from literature show that the formation of DBPs are largely unknown, where it is unclear which exact water conditions form particular DBPs, the impact of each parameter on particular DBPs and making reproducible models for predicting DBPs. Hence, this project tried to solve the problem of the formation of DBPs by analysing the correlations between water parameters and DBP concentrations, and making a predictive model for it.

From the research done in this project it was found that location has an influence on some of the DBPs formed, Ridge Regression is a better alternative for datasets where parameters are multi-correlated, and lastly that some models from other studies are not universal in their application of predicting DBPs and hence that predicting DBPs is likely a case by case situation per certain geographical location or water treatment plant, as more factors such as chlorine dosages are at play other than just water parameters.

Although results from this project are not very good, it shows the importance of data quality in a dataset when trying to train a machine learning model as well as

understanding the bigger problem.

Overall, this project shows that there is still a long way to go for DBP research in terms of using machine learning in order to predict DBP formations. Without international cooperation and research into making a centralised and high quality database on water parameters (plus other factors influencing DBP formation) and DBP concentrations, it will be difficult to make reliable and accurate machine learning models that can automate and regulate water treatment plants to ensure safe drinking water.

Bibliography

[1] "Goal 6 | Department of Economic and Social Affairs," Apr. 2024. [Online; accessed 2. May 2024].

[2] S. Kali, M. Khan, M. S. Ghaffar, S. Rasheed, A. Waseem, M. M. Iqbal, M. Bilal khan Niazi, and M. I. Zafar, "Occurrence, influencing factors, toxicity, regulations, and abatement approaches for disinfection by-products in chlorinated drinking water: A comprehensive review," *Environ. Pollut.*, vol. 281, p. 116950, July 2021.

[3] "Water Disinfection with Chlorine and Chloramine | Public Water Systems | Drinking Water | Healthy Water | CDC," Nov. 2020. [Online; accessed 2. May 2024].

[4] "Water Treatment | Public Water Systems | Drinking Water | Healthy Water | CDC," May 2022. [Online; accessed 9. May 2024].

[5] S. Chowdhury, P. Champagne, and P. J. McLellan, "Models for predicting disinfection byproduct (DBP) formation in drinking waters: A chronological review," *Sci. Total Environ.*, vol. 407, pp. 4189–4206, July 2009.

[6] W. H. O. Who, "Drinking-water," *World Health Organization: WHO*, Sept. 2023.

[7] A. Sulehria, Y. Mustafa, B. Kanwal, and A. Nazish, "ASSESSMENT OF DRINKING WATER QUALITY IN ISLAMPURA, Distt. LAHORE (Local Report)," *Science International (Lahore)*, vol. 25, pp. 359–361, Jan. 2013.

[8] K. Zhang, Q. Chungen, A. Cai, J. Deng, and X. Li, "Factors affecting the formation ofDBPs bychlorine disinfection inwater distribution system," *Desalination and Water Treatment*, vol. 205, pp. 91–102, Nov. 2020.

[9] I. Kalita, A. Kamilaris, P. Havinga, and I. Reva, "Assessing the Health Impact of Disinfection Byproducts in Drinking Water," *ACS ES & T Water*, vol. 4, pp. 1564–1578, Apr. 2024.

[10] X.-F. Li and W. A. Mitch, "Drinking Water Disinfection Byproducts (DBPs) and Human Health Effects: Multidisciplinary Challenges and Opportunities," *Environ. Sci. Technol.*, vol. 52, pp. 1681–1689, Feb. 2018.

[11] C. N. Okoji, A. I. Okoji, M. S. Ibrahim, and O. Obinna, "Comparative analysis of adaptive neuro-fuzzy inference system (ANFIS) and RSRM models to predict DBP (trihalomethanes) levels in the water treatment plant," *Arabian J. Chem.*, vol. 15, p. 103794, June 2022.

[12] M. Lowe, R. Qin, and X. Mao, "A Review on Machine Learning, Artificial Intelligence, and Smart Technology in Water Treatment and Monitoring," *Water*, vol. 14, p. 1384, Apr. 2022.

[13] M. I. Roque, J. Gomes, I. Reva, A. J. M. Valente, N. E. Simões, P. V. Morais, L. Durães, and R. C. Martins, "An Opinion on the Removal of Disinfection Byproducts from Drinking Water," *Water*, vol. 15, p. 1724, Apr. 2023.

[14] S. Koley, S. Dash, M. Khwairakpam, and A. S. Kalamdhad, "Perspectives and understanding on the occurrence, toxicity and abatement technologies of disinfection by-products in drinking water," *J. Environ. Manage.*, vol. 351, p. 119770, Feb. 2024.

[15] J. Li, J. Chen, and J. Li, "The ideal model for determination the formation potential of priority DBPs during chlorination of free amino acids," *Chemosphere*, p. 142306, May 2024.

[16] S. Parvez, K. Frost, and M. Sundararajan, "Evaluation of Drinking Water Disinfectant Byproducts Compliance Data as an Indirect Measure for Short-Term Exposure in Humans," *Int. J. Environ. Res. Public Health*, vol. 14, May 2017.

[17] Ow, "Fourth Unregulated Contaminant Monitoring Rule," *US EPA*, Apr. 2024.

[18] K. Furst, R. M. Coyte, M. Wood, A. Vengosh, and W. A. Mitch, "Disinfection Byproducts in Rajasthan, India: Are Trihalomethanes a Sufficient Indicator of Disinfection

Byproduct Exposure in Low-Income Countries?," *Environ. Sci. Technol.*, vol. 53, pp. 12007–12017, Oct. 2019.

[19] A. Gonsioroski, M. Laws, V. E. Mourikes, A. Neff, J. Drnevich, M. J. Plewa, and J. A. Flaws, "Iodoacetic acid exposure alters the transcriptome in mouse ovarian antral follicles," *J. Environ. Sci.*, vol. 117, pp. 46–57, July 2022.

[20] J. B. da Costa, S. Rodgher, L. A. Daniel, and E. L. G. Espíndola, "Toxicity on aquatic organisms exposed to secondary effluent disinfected with chlorine, peracetic acid, ozone and UV radiation," *Ecotoxicology*, vol. 23, pp. 1803–1813, Nov. 2014.

[21] "Chromophoric Dissolved Organic Matter - Environmental Measurement Systems," Jan. 2019. [Online; accessed 13. May 2024].

[22] H. Canada, "Guidance on Natural Organic Matter in Drinking Water - Canada.ca," May 2024. [Online; accessed 2. May 2024].

[23] A. Alver, E. Başt ürk, and A. Kılıç, "Disinfection By-Products Formation Potential Along the Melendiz River, Turkey; Associated Water Quality Parameters and Non-Linear Prediction Model," *Int. J. Environ. Res.*, vol. 12, pp. 909–919, Dec. 2018.

[24] T. Bond, J. Huang, M. R. Templeton, and N. Graham, "Occurrence and control of nitrogenous disinfection by-products in drinking water – A review," *Water Res.*, vol. 45, pp. 4341–4354, Oct. 2011.

[25] T. Pagano, M. Bida, and J. Kenny, "Trends in Levels of Allochthonous Dissolved Organic Carbon in Natural Water: A Review of Potential Mechanisms under a Changing Climate," *Water*, vol. 6, pp. 2862–2897, Sept. 2014.

[26] T. Priya, B. K. Mishra, and M. N. V. Prasad, "Physico-chemical techniques for the removal of disinfection by-products precursors from water," in *Disinfection By-products in Drinking Water*, pp. 23–58, Oxford, England, UK: Butterworth-Heinemann, Jan. 2020.

[27] A. I. Okoji, C. N. Okoji, and O. S. Awarun, "Performance evaluation of artificial intelligence with particle swarm optimization (PSO) to predict treatment water plant DBPs (haloacetic acids)," *Chemosphere*, vol. 344, p. 140238, Dec. 2023.

[28] A. Obolensky and P. C. Singer, "Development and Interpretation of Disinfection Byproduct Formation Models Using the Information Collection Rule Database," *Environ. Sci. Technol.*, vol. 42, pp. 5654–5660, Aug. 2008.

[29] W. Chen, Z. Liu, H. Tao, H. Xu, Y. Gu, Z. Chen, and J. Yu, "Factors affecting the formation of nitrogenous disinfection by-products during chlorination of aspartic acid in drinking water," *Sci. Total Environ.*, vol. 575, pp. 519–524, Jan. 2017.

[30] M. J. Rodriguez, Y. Vinette, J.-B. Sérodes, and C. Bouchard, "Trihalomethanes in Drinking Water of Greater Québec Region (Canada): Occurrence, Variations and Modelling," *Environ. Monit. Assess.*, vol. 89, pp. 69–93, Nov. 2003.

[31] "linear regression in simple terms - Alps Academy," Mar. 2023. [Online; accessed 13. May 2024].

[32] V. Uyak, K. Ozdemir, and I. Toroz, "Multiple linear regression modeling of disinfection by-products formation in Istanbul drinking water reservoirs," *Sci. Total Environ.*, vol. 378, pp. 269–280, June 2007.

[33] H. Hong, Q. Song, A. Mazumder, Q. Luo, J. Chen, H. Lin, H. Yu, L. Shen, and Y. Liang, "Using regression models to evaluate the formation of trihalomethanes and haloacetonitriles via chlorination of source water with low SUVA values in the Yangtze River Delta region, China," *Environ. Geochem. Health*, vol. 38, pp. 1303–1312, Dec. 2016.

[34] F. Peng, Y. Lu, Y. Wang, L. Yang, Z. Yang, and H. Li, "Predicting the formation of disinfection by-products using multiple linear and machine learning regression," *J. Environ. Chem. Eng.*, vol. 11, p. 110612, Oct. 2023.

[35] P. Kulkarni and S. Chellam, "Disinfection by-product formation following chlorination of drinking water: Artificial neural network models and changes in speciation with treatment," *Sci. Total Environ.*, vol. 408, pp. 4202–4210, Sept. 2010.

[36] A. Mader and W. Eggink, "A DESIGN PROCESS FOR CREATIVE TECHNOLOGY," *ResearchGate*, Sept. 2014.

[37] M. Hunter, "The 5-Step Data Science Project Life Cycle You Need to Be an Effective Data Scientist," *Medium*, Oct. 2022.

[38] "seaborn: statistical data visualization — seaborn 0.13.2 documentation," Jan. 2024. [Online; accessed 18. Jul. 2024].

[39] Oa, "Regional and Geographic Offices," *US EPA*, Jan. 2024.

[40] "Maine Coast Sea Vegetables," July 2024. [Online; accessed 20. Jul. 2024].

[41] GeeksforGeeks, "Regression Metrics," *GeeksforGeeks*, Oct. 2023.

[42] "LinearRegression," July 2024. [Online; accessed 20. Jul. 2024].

[43] G. L. Team, "What is Ridge Regression?," *Great Learning Blog: Free Resources what Matters to shape your Career!*, June 2024.

[44] "What is Overfitting? - Overfitting in Machine Learning Explained - AWS," July 2024. [Online; accessed 19. Jul. 2024].

[45] O. Lu, S. W. Krasner, and S. Liang, "Modeling approach to treatability analyses of an existing treatment plant," *Journal -. American. Water. Works. Association.*, vol. 103, pp. 103–117, Apr. 2011.