



UNIVERSITY OF  
CAMBRIDGE

Department of Engineering

MASTER'S PROJECT REPORT

# Quantification of Free Residual Chlorine for Water Purification

C-AJK61-3

**Author: Neelay R. Sant**

**Supervisor: Prof. Alexandre Kabla**

1st June 2022

I hereby declare that, except where specifically indicated, the work submitted herein is my own original work.

Signed:  Date: 01/06/2022

# Acknowledgements

I want to thank my supervisor, Prof. Alexandre Kabla, for his guidance and encouragement during this project. His patience in helping me understand what I should be aiming for and managing difficulties during the project has been beyond what I could hope for from a supervisor, and I am very grateful for that.

I also want to thank both Rebecca Donaldson and Tom Stakes from Blue Tap for their support in the project. Their guidance in helping me frame my objectives, and aid with their own experience in water quality, was tremendously helpful in my progress and deeply appreciated.

# Quantification of Free Residual Chlorine for Water Purification

Neelay Sant, *Emmanuel College*

## Technical Abstract

The subject of water quality is extremely important, as access to clean water is essential for human survival. Chlorination is a common method of disinfection, but dosing water samples with appropriate amounts of chlorine is an issue; too little can leave the water still infected, but too much can make the water unpotable. Measuring the Free Residual Chlorine (FRC) of a sample identifies whether a sample has been dosed appropriately, but current methods of measuring FRC are expensive and have limitations.

The aim for this project was to establish whether it is feasible to use the measurement of common water parameters to quantify the FRC of drinking water samples that have been treated with chlorine. The purpose of this was to explore possible ways of developing a low-cost FRC sensor to use in low-income areas.

The first stage of this project concerned the collection of data, in order to build a dataset that could be used for analysis. This involved the development of a method to produce predictable chlorine doses and consistent measurements across sensors. This meant we could be confident in using the method for building the dataset. Samples over a range of water sources were gathered and measured, aiming to get a general representation of waters that might be analysed in the field.

The parameters that were analysed for use in this project were water conductivity, pH, dissolved oxygen concentration, turbidity and Oxidation-Reduction Potential (ORP). The measurement of the ORP presented unexpected results. One was that the ORP probe's reading had a strong dependency on its recent history (in the previous hour). The measurement also had long stabilisation times (longer than 5 minutes). This led to developing a technique for taking measurements with the ORP probe involving long periods between measurements, and measuring samples of similar compositions together, to mitigate these effects. This was also an indicator that this measurement might be less reliable to take in the field.

Each parameter measured was evaluated on whether it would be useful for predicting the FRC. This was partly done by looking at the correlation coefficient ( $R^2$ ) between the parameter and the measured FRC for each water source.  $R^2$  gives a number between 0 and 1, and higher magnitudes show stronger correlation, so it was used as a 'score' for the correlations. On looking at the direct relationships between each measured parameter, ORP provided the most valuable information, as it held a relationship with FRC that was reasonably independent of the source. The conductivity measurements had some correlation with the FRC (on average,  $R^2 = 0.506$  for each source), but were very source-dependent. This suggested they would be possibly useful measurements for determining FRC, but less important than the ORP. Both dissolved oxygen and turbidity had little to no direct correlation with FRC, and dissolved oxygen also had a very small range of variation. This suggested that these measurements would be less useful than the others mentioned, or useless, when trying to predict FRC.

The second stage of this project involved using different data analysis techniques to find mappings from the measured parameters to a continuous quantification of FRC, or to

form classifiers that could classify samples into having too little, enough, or too much chlorine. Scores of 0.75 for correlation, and 0.8 for classification, were decided as thresholds for success on trials conducted. These values were chosen to show that a method had significant viability in being assessed further for practical application.

For classification, the score was the proportion of correctly classified points. Classification was done in two approaches, both by splitting the total FRC range into ranges for each class. The first approach was a 3-class split, with the classes being ' $0 \rightarrow 0.5$  FRC' (not enough chlorine), ' $0.5 \rightarrow 2$  FRC' (enough chlorine), and ' $> 2$  FRC' (too much chlorine). The second approach was a 2-class split with classes of greater than or less than 0.5 FRC. This had the purpose of being able to label whether a sample had enough chlorine or not.

Linear Discriminant Analysis (LDA) performed best at finding a relationship between other common water parameters that correlated to the FRC. It obtained a correlation of 0.703, which was lower than the defined threshold, suggesting that it was not possible to find a continuous quantitative measure of FRC. Quadratic Discriminant Analysis (QDA) performed best at classifying points, and both scores, especially the classification score on the 2-class split (deciding whether a sample's FRC was above or below 0.5), were very promising. The 3-class score was 0.777, which was close to, but lower than, the desired threshold of 0.8. The 2-class score of 0.891 was very strong, and well above the desired threshold. This means the analysis correctly classified 89.1% of the test datapoints, which is strong evidence for the use of this method in the future. All the analysis techniques showed turbidity and dissolved oxygen to have a negligible impact, as they both reduced the correlation and classification scores when they were included. This suggested they are not useful for determining FRC.

In conclusion, this project shows that there is scope for using the measurements of other water parameters to be able to classify water samples into having enough chlorine or not, but that the complexity of the problem is too high for providing a continuous quantitative measure of FRC. With a much larger dataset, and exploring non-linear data analysis techniques, there is possibility for finer-grain classification. However, this outcome is highly dependent on reliable measurements of ORP, as the data analysis showed high dependency on this measurement. If it is too difficult to get reliable measurement of ORP in the field, the outcome of this project suggests it would not be possible to use measurements of other water parameters to determine FRC.

# Contents

<b>1</b>	<b>Introduction and Background</b>	<b>3</b>
1.1	The Importance of Water Purification . . . . .	3
1.2	Disinfection Methods . . . . .	3
1.3	What is Free Residual Chlorine? . . . . .	4
1.4	Current FRC Measurement Methods . . . . .	6
1.5	The Relation of FRC to Other Parameters . . . . .	7
1.6	Project Aim and Objectives . . . . .	9
<b>2</b>	<b>Data Collection</b>	<b>10</b>
2.1	Experimental Design for Data Collection . . . . .	10
2.1.1	Producing Accurate Chlorine Doses . . . . .	10
2.1.2	Verifying ORP Measurements . . . . .	11
2.1.3	Experimental Method . . . . .	13
2.2	Data Collection Results and Analysis . . . . .	14
2.2.1	Conductivity . . . . .	14
2.2.2	pH . . . . .	16
2.2.3	Oxidation-Reduction Potential (ORP) . . . . .	17
2.2.4	Turbidity (Tb) . . . . .	18
2.2.5	Dissolved Oxygen (DO) . . . . .	20
2.3	Data Collection Conclusions . . . . .	20
<b>3</b>	<b>Data Analysis Approaches</b>	<b>22</b>
3.1	ORP as a baseline . . . . .	23
3.2	Principal Component Analysis (PCA) . . . . .	24
3.2.1	Background . . . . .	24
3.2.2	Methods of Implementation . . . . .	24
3.2.3	Results and Discussion . . . . .	25
3.2.4	Conclusions . . . . .	27
3.3	Linear Discriminant Analysis (LDA) . . . . .	28
3.3.1	Background . . . . .	28
3.3.2	Methods of Implementation . . . . .	28
3.3.3	Results and Discussion . . . . .	29
3.3.4	Conclusions . . . . .	33
3.4	Quadratic Discriminant Analysis (QDA) . . . . .	33
3.4.1	Background . . . . .	33
3.4.2	Methods of Implementation . . . . .	33
3.4.3	Results and Discussion . . . . .	33

---

3.4.4	Conclusions . . . . .	35
3.5	Comparisons . . . . .	35
<b>4</b>	<b>Summary and Conclusions</b>	<b>37</b>
4.1	Summary of Aims and Indicators of Success . . . . .	37
4.2	Conclusions and Future Work . . . . .	37
4.3	Risk Assessment . . . . .	39

# Chapter 1

## Introduction and Background

Clean water is essential for human survival, and the topic of water purification is extremely important for ensuring a sustainable supply of clean water for everyone. The most pressing health concern associated with drinking water is contamination with pathogens that can cause deadly diseases. Due to its low cost, ease of use, and safety, chlorination is a popular method of disinfection for drinking water. The main issue that arises from using this method is dosing samples appropriately [13]. Identifying an appropriate dose requires measuring the Free Residual Chlorine (FRC) of samples. Using a traditional in-line sensor (a sensor in a pipe taking measurements) has limitations, with the main issue being its expense. This report evaluates the possibility of measuring other common water parameters in order to determine the FRC to an appropriate precision. This could then lead to the development of a low-cost FRC sensor to use in low-income areas.

### 1.1 The Importance of Water Purification

3 in 10 people lack access to clean drinking water worldwide [1]. Nearly half of the global population already live in potentially water-scarce areas for at least one month each year, which could rise to 5.7 billion people by 2050 [24]. These statistics show how access to clean drinking water is an extremely important and growing issue.

There are several processes involved in cleaning water to produce drinking water. Contaminants can be in the form of particles, dissolved organics and inorganics, and microorganisms. Removing particles from the water can be done by filtration or sedimentation. Removing dissolved inorganic compounds can be done using ion-exchange or precipitation, and dissolved organic compounds can be removed by adsorption. The most prevalent contaminants in drinking water concerning health are microorganisms. There are 5-10 million water-related deaths worldwide each year, mainly from pathogens causing water-borne diseases, such as cholera and typhoid [1]. Pathogens in water are killed by disinfection.

### 1.2 Disinfection Methods

The main methods of disinfecting water supplies are ozonation, using UV light, using chloramines, and chlorination [11]. Ozone ( $O_3$ ) is an unstable form of oxygen that de-

composes very quickly to produce highly reactive free radicals. It has a higher oxidising power than even chlorine, and so is a very effective disinfectant [11]. It also does not leave behind any unpleasant odour or taste. However, the production of ozone is very expensive, requiring high amounts of energy and specialised equipment [9]. Ozonation also has no way of protecting treated water samples from re-contamination.

UV light can be used to disinfect water as it can break down the chemical bonds in DNA and proteins in microorganisms [9]. Its advantages are that it doesn't produce any byproducts, and it can limit potential regrowth of microorganisms. However, it also has a very high energy requirement, and its efficacy can be limited in more turbid water samples [12]. This method also has no way of protecting treated water samples from re-contamination.

Chlorination is the process of disinfecting water samples by reacting them with chlorine. It kills microorganisms by oxidising bonds in their structure, and it is very effective at eliminating a wide range of pathogens due to its high oxidising power [16]. It is much cheaper to use than UV or ozone, and the dosing rate can be changed easily and flexibly. As opposed to ozonation and using UV light, chlorination can protect water from recontamination, as once the chlorine is reacted, a small residual amount (the free residual chlorine) remains in the water. However, it can leave behind byproducts that increase the toxicity of the water, such as trihalomethanes [11]. Furthermore, the dose amount is of high importance, as too much can leave the water unpotable, and too little means that the water could still contain pathogens.

Chloramines can be used as an alternative to chlorine, and work in a similar way. They are produced by reacting free chlorine with ammonia and are also oxidising agents. They are more stable than chlorine, so can provide long-lasting residual disinfection. They also produce fewer byproducts [11]. However, due to their lower oxidising power, they are less effective than free chlorine at disinfecting samples.

This project will only be concerned with chlorination as a disinfection method. This is because it is a very popular form of disinfection, especially in low-income areas, due to being cheap and highly effective.

## 1.3 What is Free Residual Chlorine?

Chlorine can be added to water in two ways - as a gas ( $\text{Cl}_2$ ), or in a hypochlorite ( $\text{OCl}^-$ ) salt, such as sodium hypochlorite ( $\text{NaOCl}$ ) [16]. The gas disproportionates in water into hydrochloric acid ( $\text{HCl}$ ) and hypochlorous acid ( $\text{HOCl}$ ).  $\text{HOCl}$  and  $\text{OCl}^-$  are both strong oxidising agents and the sum of their concentrations in a water sample is the Free Residual Chlorine (FRC).

The process by which contaminants can affect the concentration of FRC is shown in Figure 1.1. In the reduction stage, chlorine will immediately react with reducing agents in the water, such as iron, manganese, and nitrites. No residual will form until these agents are completely oxidised. Once these are all reduced, chlorine reacts with nitrogen-based compounds (organic matter and ammonia), forming *Combined Chlorine Residuals*, such as chloramines. These residuals are weaker oxidising agents than FRC. As more chlorine is added, it will oxidise the chloramines further into compounds that are not oxidising agents, therefore reducing the residual. This occurs until the 'breakpoint' is reached,



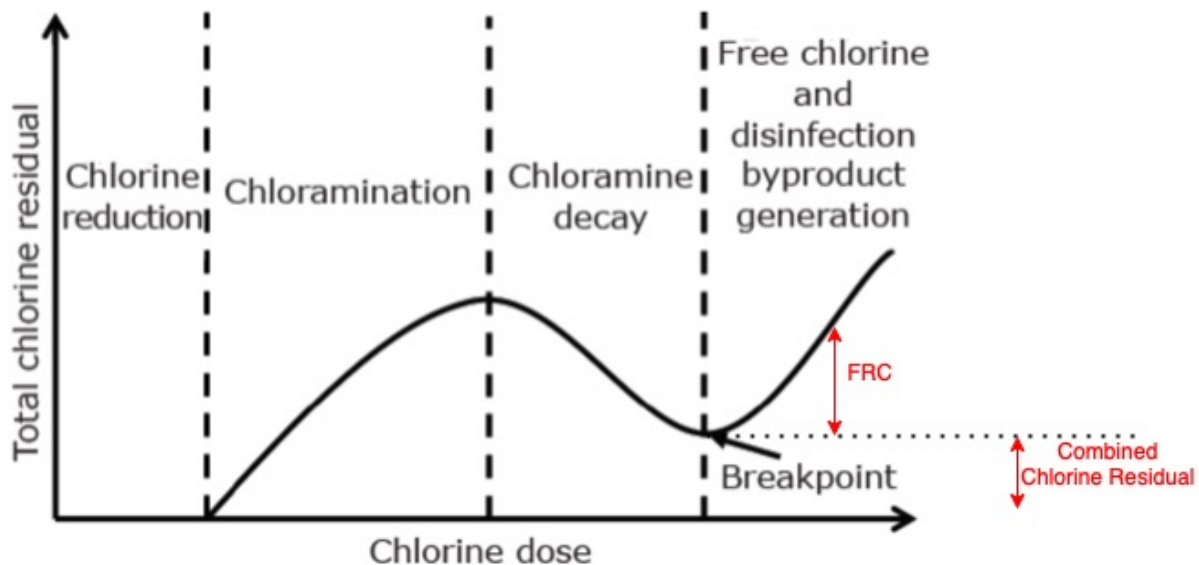


Figure 1.1: The generation of residual chlorine in untreated water samples. Taken from [26]

where there is a minimum amount of reducible chloramines in the sample. After this point, any chlorine added forms a proportional amount of FRC [26, 16].

The presence of FRC in a chlorinated water sample shows that the sample is past the breakpoint, which means it is fully disinfected. The sample also then has a strong oxidising power, which is important as it can still disinfect the water more, in case of further contamination. The oxidising power of  $\text{OCl}^-$  is around 1% that of  $\text{HOCl}$ , but both are much stronger oxidising agents than any chloramines [21]. Due to this, FRC can be used to indicate whether a water sample is safe to drink. The guidelines on how to treat a water sample based on the measured FRC, as set by the WHO [27], are detailed in Table 1.1.

FRC Range (mg/L)	Short Label	Details
$FRC < 0.2$	FRC too low	This is not high enough to be sure it is past the breaking point.
$0.2 \leq FRC < 0.5$	safe	This water is safe to drink immediately, but might not be safe to drink after storing, since it does not have much disinfection capability left.
$0.5 \leq FRC < 2$	safe to store	This water can be stored in a bucket or jerry can and should be safe to drink for 24 hours.
$2 \leq FRC$	FRC too high	This water can become unpotable due to unpleasant taste and odor from chlorine, and is above the WHO recommended level of FRC.

Table 1.1: A Table summarising the meaning, in terms of safety of consumption, behind the FRC of a sample according to the range in which the FRC fits. These are in accordance with WHO guidelines [27].

Fig 1.1 gives an indication of the difficulty of predicting the FRC in a sample indirectly. It shows how the FRC is not directly correlated to the chlorine added, and how different levels of contaminants can affect the amount of FRC produced by a specified dose.

## 1.4 Current FRC Measurement Methods

Reacting chlorine with water gives three metrics to help keep track of how much is needed. The dose is the amount of chlorine added. The demand is the amount that reacts with substances in the water through oxidation. The FRC is the chlorine left uncombined. This gives the simple relationship,  $dose = demand + FRC$ .

In order to achieve complete disinfection whilst ensuring that the water remains potable and non-toxic, it is vital to measure FRC and maintain appropriate levels. [21]. Figure 1.2 shows the distinction between different compounds present in water when chlorine is added. These groups can make the measurement of FRC in particular difficult, as they can all cause interference and affect all measurement methods.

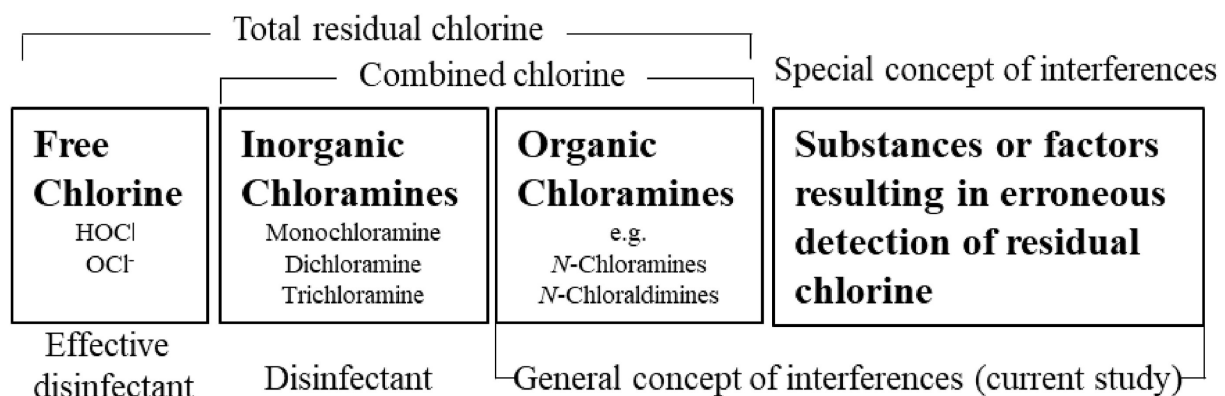
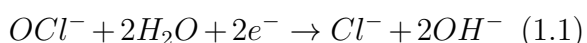


Figure 1.2: Forms of chlorine and interfering compounds in chlorinated water. Taken from [13]

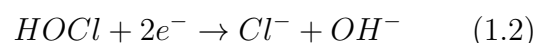
The oldest method of FRC measurement is *titration*, which involves reacting samples with starch iodide [13]. It is no longer commonly used, as it is disadvantaged relative to modern instruments in accuracy, difficulty and measurement time. However, modern methods still derive from the principle of titration [13].

The *electrochemical* (EL) method is one commonly used for FRC measurement. Sensors using this method are often referred to as amperometric sensors. The method involves passing a small current between an anode and cathode. Due to chemical reactions involving FRC, there is a change in current which can be measured, that can then be used to determine the FRC concentration [23]. It is based on the reduction of FRC at varying potentials and summarised by the following reactions:

in basic conditions:



in acidic conditions:



The EL method can also distinguish free chlorine from combined residual chlorine by controlling the pH and potassium iodide concentration in the solution. It is therefore suitable for measuring FRC in-line continuously [10]. Li, 2021 [13] states that the determining range for the EL method is 0.01 to 1 mg/L. However, the company Sensorex sell an amperometric sensor that has a determining range of 0.01 to 10mg/L [22]. This leaves ambiguity in the suitable FRC range for this method. Nevertheless, the sensor is expensive (over \$1000 [22]) and requires routine calibration against EPA-approved N,

N-diethyl-p-phenylenediamine (DPD) colorimetric methods [15]. This is controversial within the water quality industry, as amperometric sensors have been shown to have large variance in the measured FRC when other water parameters are varied, such as pH [15].

The *colorimetric-photometric* method can be used for in-line as well as on-site analysis of FRC. It involves reacting chlorinated samples with one of two reagents - DPD or o-toluidine (OT). In this method, the FRC reacts with the colorimetric reagents, and the resulting change in colour is measured to identify the FRC concentration. Both reagents have determining ranges of greater than 4mg/L [13]. However, OT has been found to be potentially carcinogenic, so it is no longer commonly used as a reagent [20]. DPD is recognised to be used in the standard colorimetric method for FRC, but also has drawbacks in being sensitive to unexpected interference by other substances, such as a high concentration of manganese or chromium [14]. It also has the issue of requiring an added reagent for measurement that can't be mixed into the main supply.

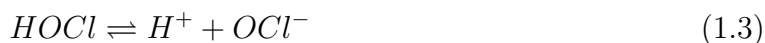
For our purpose, a low-cost sensor is required that can measure in the range of 0-2.5 FRC at least. The expense and potential drawbacks of the methods described have motivated this project to look at whether measuring other common water parameters are enough to serve as a proxy for measuring the FRC directly. The company 'PCE Instruments' [8] sell probes to measure each of the parameters at approximately £50 each. This gives us confidence that a low-cost sensor could be developed if these measurements can be used for determining FRC. There would obviously be more costs associated with manufacturing the sensor, but this shows that it is at least worth looking into.

## 1.5 The Relation of FRC to Other Parameters

The FRC has potential to have strong relationships with other commonly measured water parameters. This project aims to evaluate whether the measurements of these other parameters can be used to quantify the FRC with enough precision. The Oxidation-Reduction Potential (ORP) is expected to hold the strongest direct relationship with FRC, due to the components of FRC being strong oxidising agents. ORP is a measure (in millivolts) of the tendency for a substance to oxidise or reduce others [4]. It measures the electron activity in the sample, so depends on concentrations of all substances in the water, as well as temperature and pH.

However, it can also be considered as a measurement of the oxidising power of a substance. Since FRC has a high oxidising power, a strong correlation between the FRC and ORP of samples is expected, especially since the presence of FRC means the chlorine has already oxidised strong reducing agents in the water. This relationship is signified in Figure 1.3. The corner in the graph shows that the oxidising power of the samples is much less affected by more FRC after a certain threshold is reached. The location of this corner, and how sharp the corner is, are dependent on substances other than FRC in the water.

Adding chlorine to water should not have much effect on the pH of the sample. However, the composition of FRC in the water will be affected by different pH. HOCl exists in water in an equilibrium with hypochlorite ions following Equation 1.3.



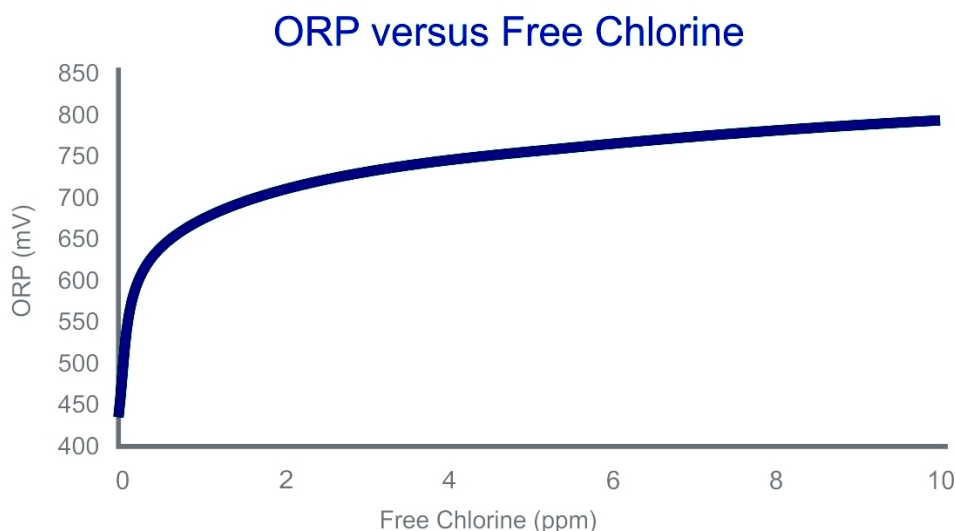


Figure 1.3: The relationship between ORP and FRC. Taken from [17]

The equilibrium point is highly dependent on the pH of the sample. Acidic solutions will favour the  $\text{HOCl}$  form, and basic solutions will favour the  $\text{OCl}^-$  form. Both are strong oxidising agents, but the oxidising power of  $\text{OCl}^-$  is around 1% that of  $\text{HOCl}$  [26]. This means that the ORP of a specific amount of FRC would be significantly different at varying pH. Therefore, measuring the pH could prove useful in conjunction with measuring the ORP, as well as providing an indication of the level of contamination.

Conductivity is a measure of the capability of water to pass an electrical current. It is directly related to the concentration of ions present in the water, as well as their form, oxidation state and mobility [2]. Since chlorine introduces ions when added to water, the amount added, and therefore the FRC, should have an effect on the measured conductivity of a sample. However, the conductivity will predominantly be a good indicator of the level of contamination of a sample, which could be related to the chlorine demand, as high contamination would likely indicate a high amount of reducing agents or pathogens in the water.

Another commonly measured water parameter is the Dissolved Oxygen (DO) level, although it is usually measured in the context of waters for aquatic life, as it is an essential component for those use cases. DO is not involved directly in reactions with FRC itself, but since it is an oxidising agent, it could feasibly also have an effect on the reducing agents in the water and the measured ORP in a sample. Therefore, it is possible that it will add value when measured in conjunction with the ORP.

Turbidity is a measure of the relative clarity of water. It measures the amount of light scattered by water, and is an indicator of water quality as it is related to the number of particulates in the water [25]. Since FRC is soluble, it will not have a direct effect on the turbidity. However, turbidity can have a negative impact on the effectiveness of chlorine disinfection [27]. This could be because it reduces the ORP, so a fixed FRC could have a lower ORP in more turbid waters. Therefore, there may be benefit to measuring the two in conjunction with each other.

The meaning and potential relevance of each parameter is summarised in Table 1.2. This

will be referenced later in the report.

Parameter	Meaning	Relation to Chlorine
Oxidation-Reduction Potential (ORP)	The oxidising strength of the water	High correlation due to the components of FRC being strong oxidising agents
Conductivity (Cd)	Measures the capability of water to pass a current	Chlorine forms ions in water so there should be a positive correlation. Also indicates the level of contamination in the water.
pH	$-\log[H^+]$	Controls the state of FRC in water, so should effect the ORP measurement for a specified FRC.
Dissolved Oxygen (DO)	Concentration of oxygen dissolved in the water	Oxygen is also an oxidising agent, so having a high DO could change the ORP at a certain FRC as the oxidation power would be higher.
Turbidity (Tb)	How transparent the water is	Negative impact on the effectiveness of chlorine disinfection, so may change the ORP at a specified FRC.

Table 1.2: Summary of parameters and their possible uses in evaluating the FRC of a water sample.

## 1.6 Project Aim and Objectives

The overarching aim of this project is to evaluate whether it is feasible to use the measurement of common water parameters to be able to quantify the FRC of drinking water samples that have been treated with chlorine. The project is divided into a set of objectives that are combined to evaluate this. These objectives are:

1. Develop a reliable method for creating and measuring water samples from a variety of water sources that can give reproducible results.
2. Observe the direct relationships between each individual parameter and chlorine, as well as their accuracy and reproducibility when being measured. This will identify which parameters are most likely to add information about the FRC when used in data analysis together. It will also identify the ease of measuring these parameters in practice.
3. Explore different data analysis techniques to try to determine the FRC from the data. In the best case, a continuous mapping from the measured parameters to the FRC will be found, i.e. an estimated FRC could be outputted from an input of the parameters measured. In this project, this would be finding a vector in the space defined by the measured input parameters. Methods trying to do this will be evaluated by calculating the  $R^2$  value between the estimated FRC of the datapoints and their actual measured FRC.
4. Investigate classification techniques on the datapoints, where each class will be a specified FRC range. The ranges chosen will be based on those detailed in Table 1.1, as this will tell us how to treat each water sample based on their class, i.e. the sample needs more chlorine, it is safe to drink, or the sample has too much chlorine. Classifiers will be evaluated by their success rate in classifying datapoints.

The outcomes of these objectives will aid in determining whether it might be possible to develop an FRC sensor from the output of measuring common water parameters.

# Chapter 2

## Data Collection

This chapter is focused on establishing a reliable method for creating and measuring water samples from a variety of sources over a range of FRC, to then later be used in data analysis. It will also look at the direct relationships between each measured parameter and the FRC, to identify how they can be used for data analysis and how useful they might be.

### 2.1 Experimental Design for Data Collection

Samples from a variety of water sources were dosed with varying amounts of chlorine, and measured for each of the included parameters, to build up a reliable dataset to be used in analysis. Each sample collected varied in pH, conductivity, etc. due to different contamination levels before they were dosed with chlorine. This allowed the identification of trends associated with changing FRC for a variety of different conditions.

#### 2.1.1 Producing Accurate Chlorine Doses

The first step was to establish a robust methodology for creating and measuring samples, in order to give strong reproducibility of measurements. This involved establishing a method that produced repeatable chlorine doses, and finding the precision to which these doses could be made.

For this step, samples of distilled water with varying FRC concentrations were made using serial dilutions. Using distilled water for this stage made the process simpler and made the output more predictable, since the chlorine demand would be 0, so  $dose = FRC$ . Using a simple solution at first also meant readings from the other sensors were only being influenced by the one parameter being changed (the FRC), so their trends could be verified more easily.

After iterating the process by which samples were made, we arrived at the procedure detailed in Figure 2.4, which is discussed in more detail Section 2.1.3. This involved dissolving sodium dichloroisocyanurate (NADCC) tablets in distilled water to make a ‘mother’ solution. Samples were made with doses in the expected range of 0 to 2.5 mg/L, as this was the most relevant range for data analysis. The expected concentration was calculated from the proportion of mother solution added to plain distilled water. The resulting chlorine concentrations were measured using a colorimetric FRC sensor. For this, the sample was reacted with a reagent (DPD) to form a pink colour that could be measured. The resulting measured chlorine doses are shown in Figure 2.1.

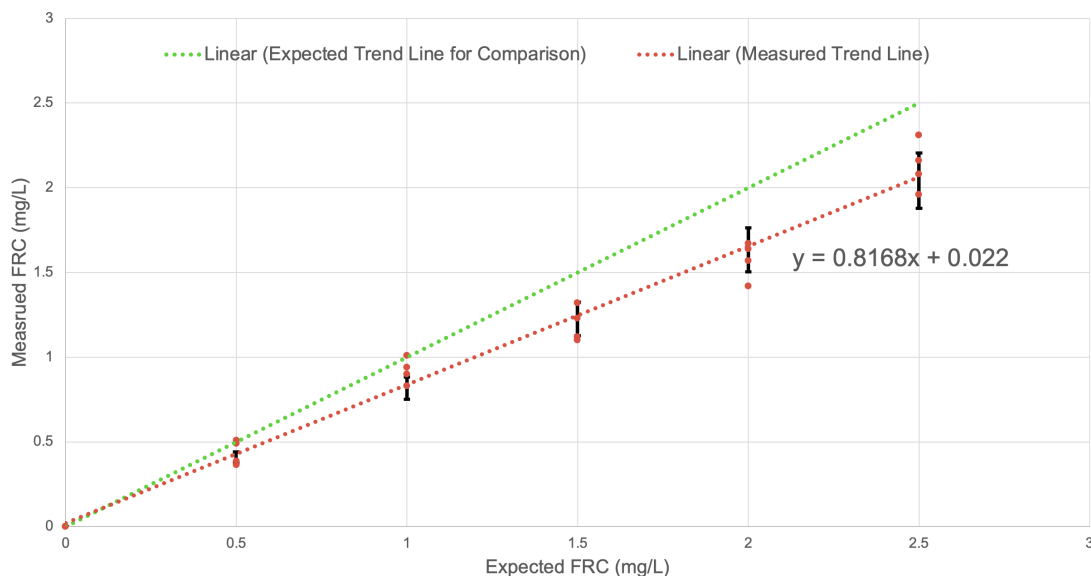


Figure 2.1: Expected vs measured FRC for distilled water samples. Error bars are shown in black, and a line for the expected FRC is also plotted to compare actual values against.

The error bars indicate an error margin of 6%, which was the maximum percentage error associated with the diluting process and the FRC sensor. Figure 2.1 gives promising results, as a linear relationship is seen that passes through the origin, which means the FRC measured is proportional to the dose. The procedure still gave a systematic error, with the mean 18.3% lower than expected, but the spread of values was more consistent. The average percentage error from this new mean was 7.95%. The systematic error and remaining percentage error from the mean could be due to the FRC delivered by the tablets being variable from the amount stated in the manual.

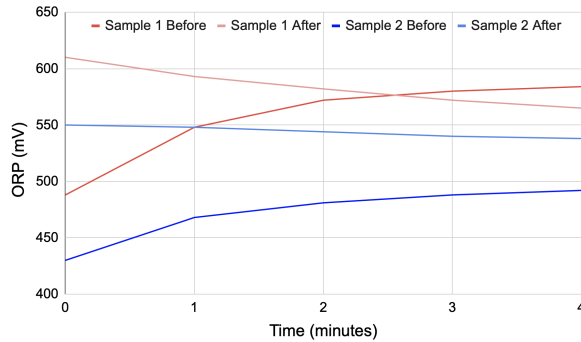
Through these iterations, we discovered potential issues in the sample-making process. These were that measured FRCs were consistently lower than expected, and had very high variation. These issues were either from the tablets used, the FRC sensor, or the mixing after dilutions not being as thorough as required. The sensor output was cross-checked with other FRC sensors to ensure it wasn't faulty. They had very close agreement of measurements, which gave confidence that this wasn't the primary issue.

The mixing was evaluated by varying the volume, the FRC concentration of the mother solution, and time spent mixing, to see which gave the best results. From this, a method was established of making samples to 100ml, using a 50mg/L mother solution, and shaking in glass bottles for a minute before and after the reaction time, as shown in Figure 2.4. As a result, research using chlorine dosing should follow these guidelines set to produce accurate chlorine doses.

### 2.1.2 Verifying ORP Measurements

The outputs of the sensors used were verified by checking their consistency over time and seeing if samples of similar composition (e.g. similar FRC) and conditions (e.g. room temperature) would give similar readings. This verified the output of all sensors except ORP, which gave an interesting variation of results. The manual of the ORP probe stated it would take around 30 seconds to stabilise. In reality, it took much longer and

(a) Readings from the ORP probe over time for two samples before and after measuring other samples. Note the different starting values and how the readings don't tend to the same value for either sample.



(b) Readings from the ORP probe over time for a distilled water sample before and after dipping the probe in a concentrated FRC solution.

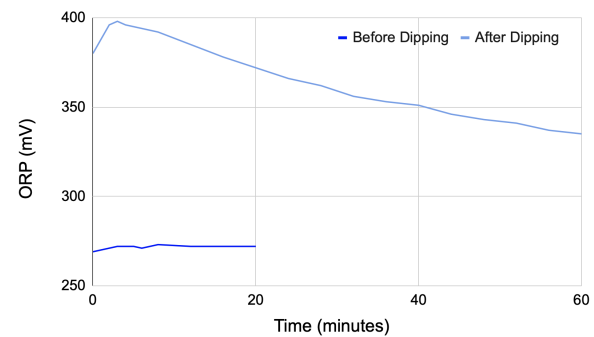


Figure 2.2: Graphs of initial ORP measurement characteristics

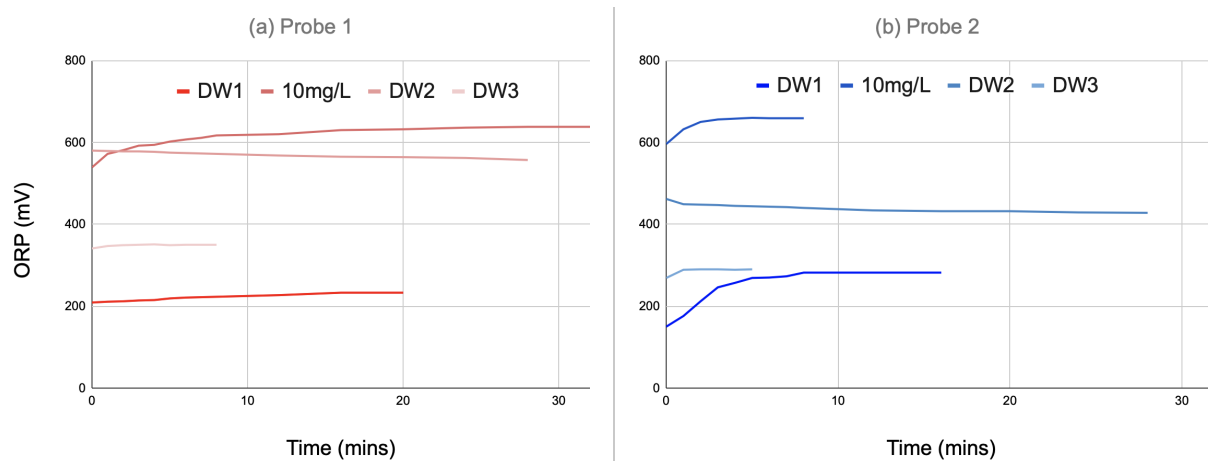


Figure 2.3: A comparison of the measurement characteristics for 2 different probes in different samples. Both probes were initially placed in a distilled water sample (DW1). They were then placed in a 10mg/L FRC solution (10mg/L), followed by returning to the distilled water sample immediately (DW2). Finally, the probes were unused for 30 minutes and then measured the distilled water again (DW3).

the measurement varied depending on what the probe had been measuring beforehand. Figure 2.2a expresses these issues, as the time taken for the measurement to stabilise was much longer than expected, and had depended on the history of the sample.

To analyse this behaviour, the ORP probe was placed in some samples in quick succession, to see the time taken to stabilise, and the magnitude of influence on the reading. The results are shown in figure 2.2b. To start, the probe was placed in distilled water, when it had not been in any sample for more than 12 hours prior. This gave a reading of around 300mV, as expected for distilled water. The probe was then dipped in a 10mg/L FRC solution for a couple of minutes. It was then washed thoroughly and returned to measuring the distilled water sample again. Figure 2.2b shows the probe took around 10 minutes to stabilise at first, but took longer than an hour when it was heavily influenced by previous measurement. However, this was resolved with enough time between measurements, as it



would return to taking 10 minutes after not being in use for over an hour.

To ensure that the characteristics were not just due to a faulty probe, the output was compared to another ORP probe. Again, both were placed in different samples in quick succession, with the results shown in Figure 2.3. The second probe still exhibits the same unexpected patterns of being highly affected by recent measurements, and taking a significant amount of time to converge to a value. However, the convergence time is less, and the probe returns to having an unaffected measurement after 30 minutes, which isn't the case for the first probe. Due to this, all measurements were taken using the second probe. To mitigate the issues found, sets of samples were made in increasing concentrations, so that the ORP variation should be less. At least 30 minutes were allowed between ORP measurements so that the influence of history should be removed. These changes were effective in giving results that matched samples of similar concentrations.

### 2.1.3 Experimental Method

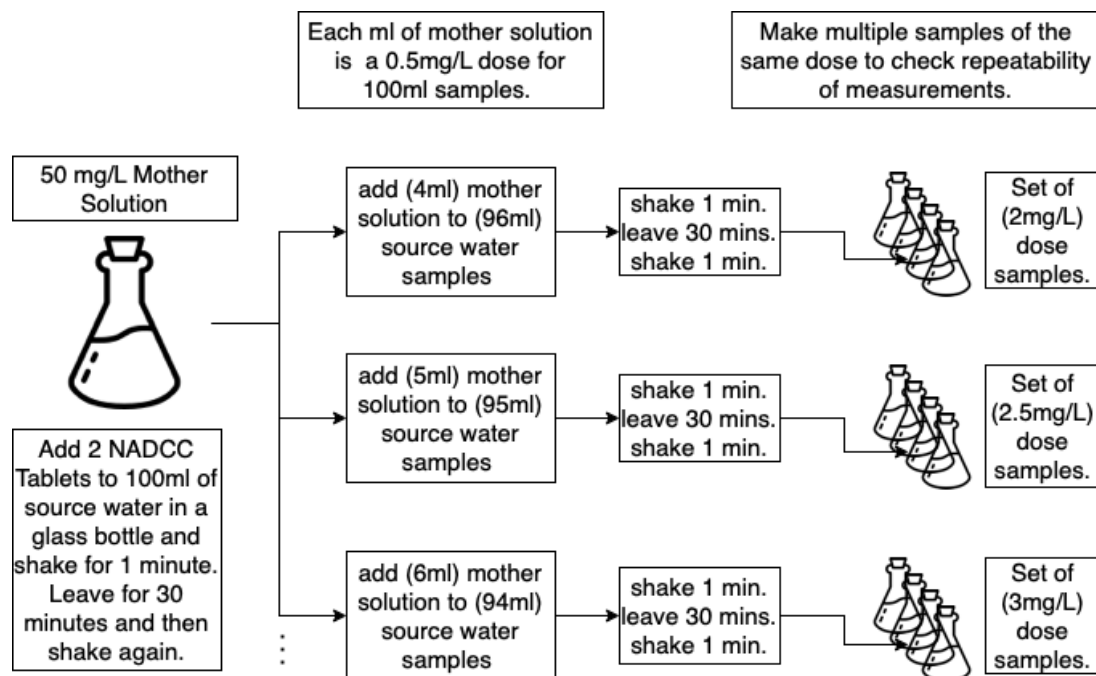


Figure 2.4: A diagram showing the process of creating samples for one water source. The number of dose sets made would be enough to get a spread of FRC values in the 0 to 2.5 mg/L range (3 to 4 sets of 4 samples each). Numbers in brackets indicate these values would be changed in order to get FRCs in the range of 0 to 2.5.

Once these issues were overcome, a process was established for making and measuring samples, detailed in figure 2.4. The high amount of shaking and long shaking periods were necessary, as less than this still gave high variation in delivered chlorine doses. The thirty-minute reacting periods were due to the fact that this is the time required for chlorine to react fully in water [27]. Making samples of similar doses together was also crucial for being able to check consistencies and identify anomalies. This method was used for creating samples from all the sources detailed in table 2.1, except for some measurements taken in Kenya, which were from chlorine-dosed water dispensers.

Data Source	Naming Convention	Details
Distilled Water 11/21	Distilled	Distilled water from the laboratory.
CUED Tap Water 11/21	Tap	Tap water from the laboratory.
Midsummer Common 12/21	MC 1	Water from the river Cam at a place called Midsummer Common. It is downstream from tourist areas and close to animal habitats.
Midsummer Common 01/22	MC 2	This was again river Cam water from the same place but at a different time. This water required extremely high chlorine doses suggesting that it contained a high amount of reducing agent.
Jesus Lock 01/22	JL	This was another set of river Cam water collected upstream from Midsummer Common and closer to the tourist area.
College Pond 01/22	Pond	This was water collected from a college pond. The pond hosted a variety of aquatic life.
Swimming Pool 12/21	Pool	This was water collected from a swimming pool.
Ivonangya, Kenya: 09/21	Kenya Ivonangya 1	Blue Tap were able to collect water measurements in Kenya on two separate occasions. This set was from the first.
Ivonangya, Kenya: 03/22	Kenya Ivonangya 2	This was water collected in Kenya on the second occasion.
Mumo, Kenya: 09/21	Kenya Mumo 1	This was water collected in a different location in the first trip to Kenya.
Mumo, Kenya: 03/22	Kenya Mumo 2	This was water collected in the second trip to Kenya.
Ivonangya, Kenya Jerry Cans 03/22	Kenya jerry cans	This was water collected from jerry cans filled from the water supply in Ivonangya.
Kyuso, Kenya: Rock Catchment 03/22	Kenya Kyuso	This was water from a rock catchment in Kyuso, Kenya
Hostel in Kenya 03/22	Kenya Hostel	This was water from a rainwater collector in a hostel in Kenya.

Table 2.1: A summary of water sources included in the dataset with naming conventions used in the report graphs and details about the source

## 2.2 Data Collection Results and Analysis

Each parameter measured was evaluated on whether it would be useful for predicting the FRC. This was partly done by looking at the resulting correlation coefficient ( $R^2$ ) between the parameter and the measured FRC for each water source.  $R^2$  is a score for how good a fit the line-of-best-fit (LOBF) is for two variables  $x$  and  $y$ , and is calculated as follows:

$$R^2 = \frac{(\sum(x_i - \bar{x})(y_i - \bar{y}))^2}{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2} \quad (2.1)$$

### 2.2.1 Conductivity

Figure 2.5 shows the evolution of conductivity over different water sources for varying FRC. Across the different sources, there are commonly slightly positive gradients, with varying values of  $R^2$ . This is in line with expectations from table 1.2. The distilled water dataset has a high  $R^2$ . This can be expected due to the simplicity of the makeup of the water. The only change being made is additional ions in the water, so it makes sense that the correlation is strong. The variation in this relationship across all samples is most

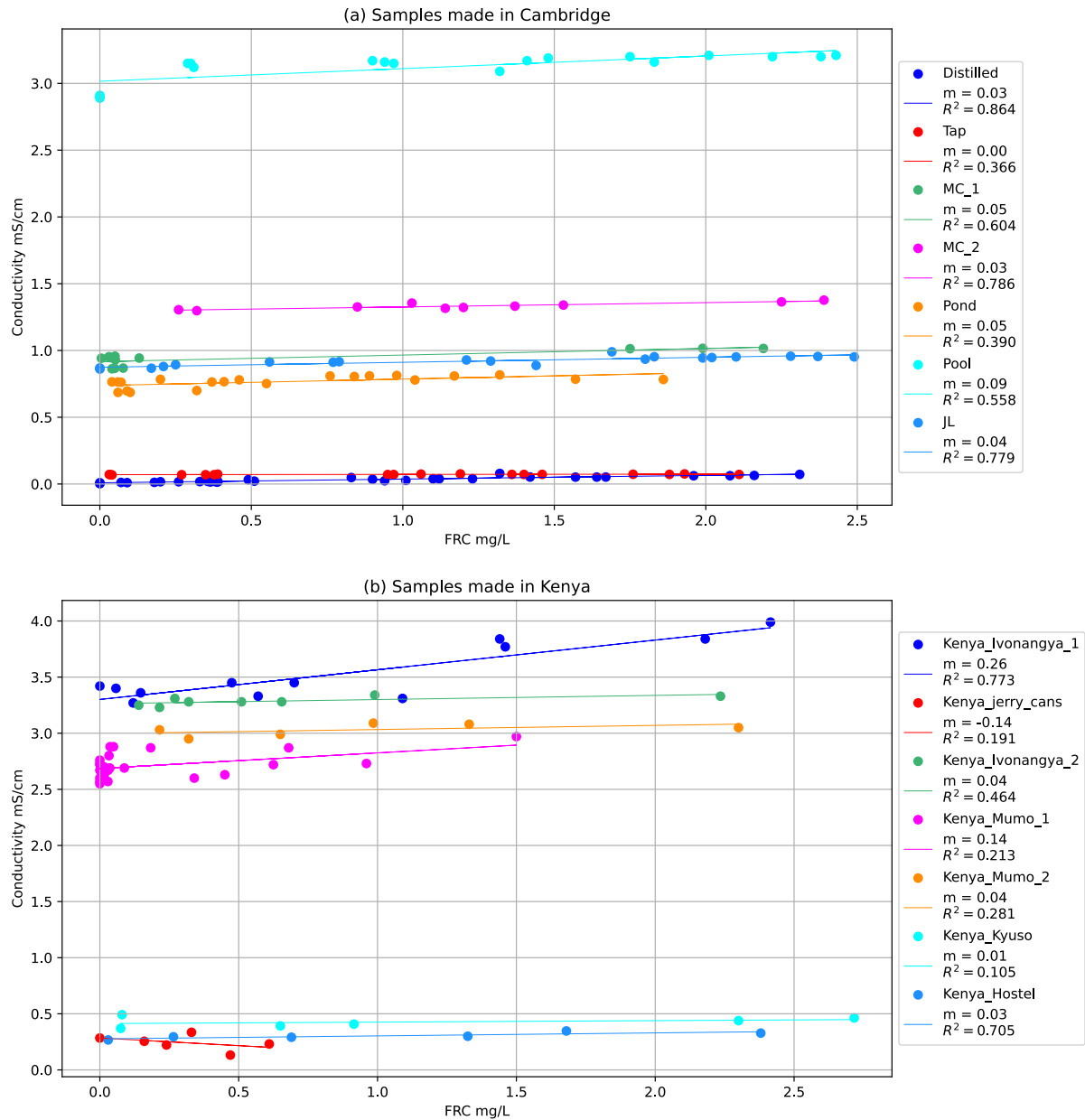


Figure 2.5: Conductivity Trends for different sets of samples, with gradients ( $m$ ) and determination coefficients ( $R^2$ ) shown in the legend. (a) and (b) have been separated so individual trends are easier to see. Naming conventions of sets are from table 2.1.

likely due to varying interference being caused by the different waters when the ions are added. Whilst conductivity is positively correlated with inorganic dissolved solids, organic compounds have poor conductivity, and could cause enough interference to mitigate the influence of increased FRC. Tap water can contain contaminants such as nitrates that might have this effect [6]. The highlight from this graph is that conductivity does have an identifiable relationship with the FRC, but it is very source-dependent, i.e. in a model, more information on the source would be needed to be able to use the conductivity to deduce the FRC.

## 2.2.2 pH

The components of FRC exist in an equilibrium in water (shown in eq. 1.3) that is highly dependent on pH. Higher pH levels (over pH 8) cause the equilibrium to lie further to the right. Figure 2.6 shows the evolution of pH over different water sources for different FRC values, where there are varying correlations in the data. Distilled water has a convincing positive correlation with a high  $R^2$  value, but many sample datasets have much lower or negative gradients with varying  $R^2$  values. It is therefore difficult to conclude whether any information on the FRC can be extracted from the pH measurement directly. However, as stated in table 1.2, the ORP will likely have dependency on the pH when being used to predict an FRC, due to the equilibrium of FRC.

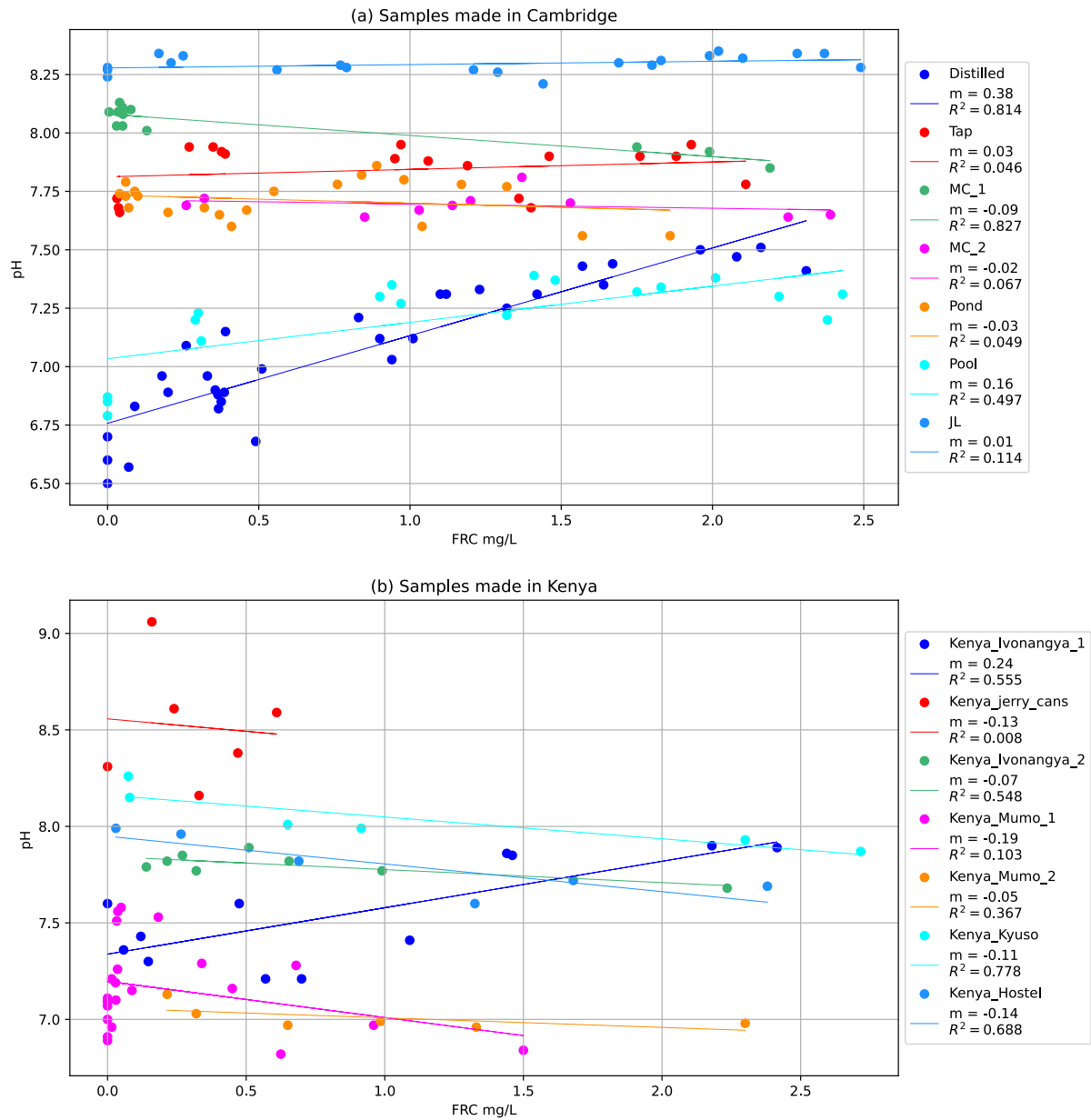


Figure 2.6: pH Trends for different sets of samples, with gradients (m) and determination coefficients ( $R^2$ ) shown in the legend. (a) and (b) have been separated so that individual trends are easier to see. Naming conventions of sets are detailed in table 2.1.

### 2.2.3 Oxidation-Reduction Potential (ORP)

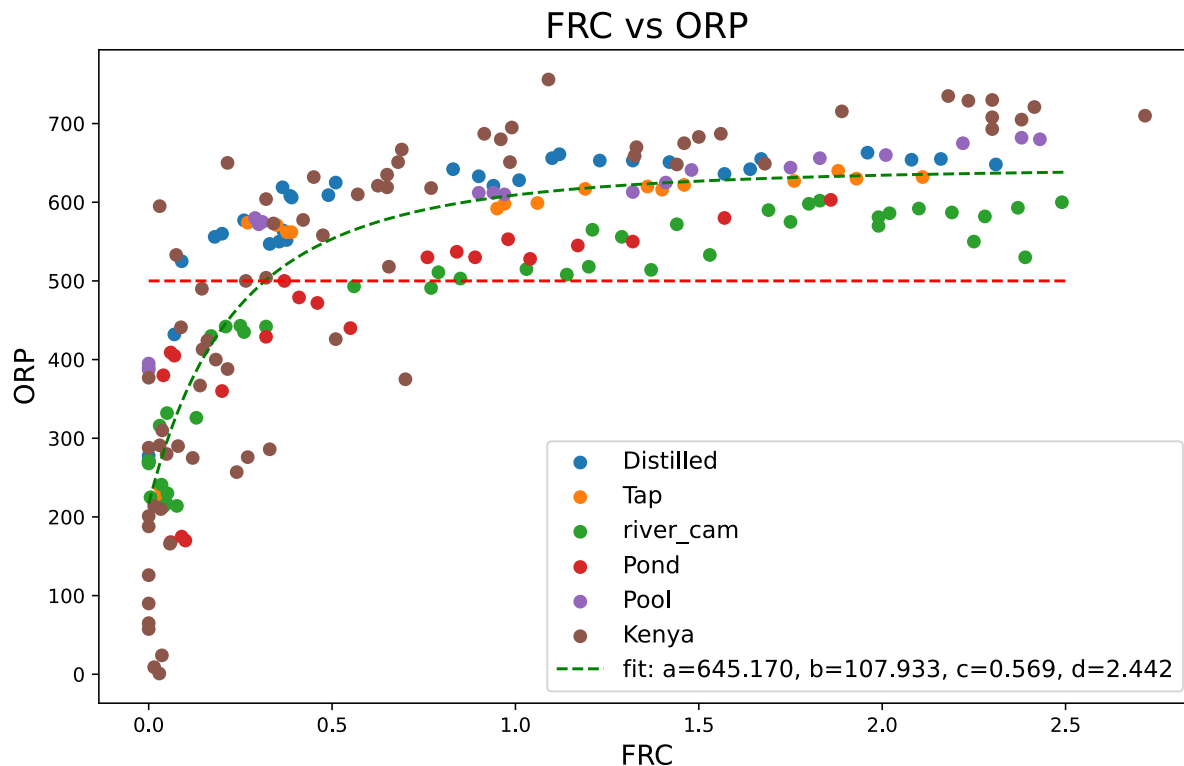


Figure 2.7: ORP Trends over sets of samples with a plotted trend line. The coefficients given in the legend follow those detailed in eqn. 2.2. The dotted red line is a marker for the cutoff between two possible ‘linear’ regions, which are used later in the report. Note that water sources have been combined to avoid cluttering the figure.

Since  $\text{HOCl}$  and  $\text{OCl}^-$  are both strong oxidants, ORP should have a correlation with the FRC. This is confirmed in figure 2.7, which shows the evolution of ORP over different water sources for different FRC values. Although there is slight separation between datasets, they all follow a general trend that seems to be fairly independent of the source. The graph shows a plateau for  $\text{ORP} > 500$  (indicated by the red dashed cutoff line), with a steep rise before that. This is promising for the potential of this measurement to predict the FRC. Since the aim is to predict the FRC from the ORP, a trend line has been fitted to the data of the form in eqn 2.2 as an estimate of the true relationship. The inverted relationship is given in eqn 2.3, which shows the mapping from ORP to FRC. In both equations, ‘x’ represents FRC, and ‘y’ represents ORP, relating to figure 2.7.

$$y = a - \frac{b}{(x + c)^d} \quad (2.2)$$

$$x = \left( \frac{b}{a - y} \right)^{\frac{1}{d}} - c \quad (2.3)$$

This form of equation has been chosen for the following reasons. ‘a’ controls where the plateau occurs. ‘b’ acts as a scaling factor. ‘c’ allows a translation in the x direction and ‘d’ controls the sharpness of the curve. In order to predict an FRC, this relationship must be inverted, to take an input of an ORP value and output a predicted FRC. One issue is that ORP values above the plateau would not map to an FRC, as the graph won’t exist in that region. To get around this, these points are outputted as an FRC of 2.5mg/L,

as this gives the same meaning as a value higher than 2.5mg/L i.e. too much chlorine. The inverted relationship then has a root-mean-square error of 0.791 FRC. This is a high error margin, as it is 32% of the total range of 2.5 FRC; it indicates that ORP may not be accurate enough to predict FRC alone. Using ORP alone to predict FRC is analysed in section 3.1.

In addition, the challenges faced using the ORP sensor suggest it is a less reliable measurement than the other variables, as it is more difficult to reproduce, and getting a dependable reading in the field may prove more difficult. Also, for a specific FRC, ORP can vary for two reasons [2]. The first is due to  $\text{OCl}^-$  having a lower oxidation power than  $\text{HOCl}$ . If FRC is made up more of  $\text{OCl}^-$ , then the ORP will give a lower reading. This can occur at higher pH values as described before, so them being used in conjunction might yield more information. The second reason is that other oxidants, such as DO, and reductants, such as sodium, will also influence the ORP. Again this could mean that using DO and conductivity measurements in conjunction with ORP could provide a stronger model.

### 2.2.4 Turbidity (Tb)

Figure 2.8 shows the evolution of turbidity values for different water sources over changing FRC. All sets have very low  $R^2$  values for their linear relationships between FRC and turbidity, indicating little to no correlation. Since chlorine fully dissolves in water in the range this project is concerned with (so it doesn't affect how transparent the water is), it makes sense that there is no obvious trend. Turbidity is really a measure of particulates in the water, which are usually removed by filtration. Overall, this graph shows that no direct information is given on the FRC from the turbidity measurement alone. However, the reasons detailed in table 1.2 about turbidity affecting the efficacy of chlorine mean there may be some use in its measurement when used in conjunction with ORP.

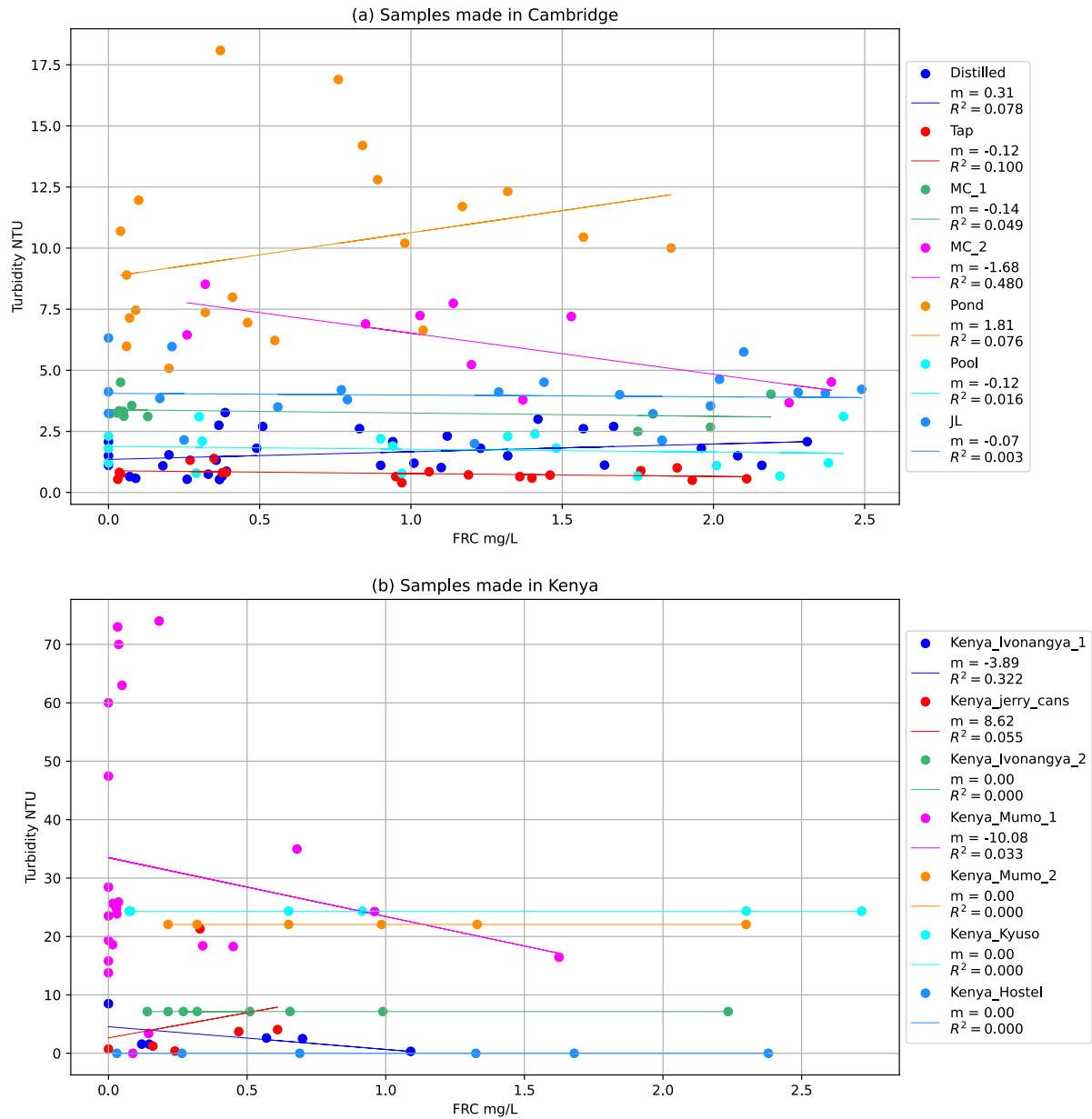


Figure 2.8: Turbidity trends for different sets of samples, with gradients ( $m$ ) and determination coefficients ( $R^2$ ) shown in the legend. (a) and (b) have been separated so that individual trends are easier to see. Naming conventions of sets are detailed in Table 2.1. (NTU = Nephelometric Turbidity Units).

## 2.2.5 Dissolved Oxygen (DO)

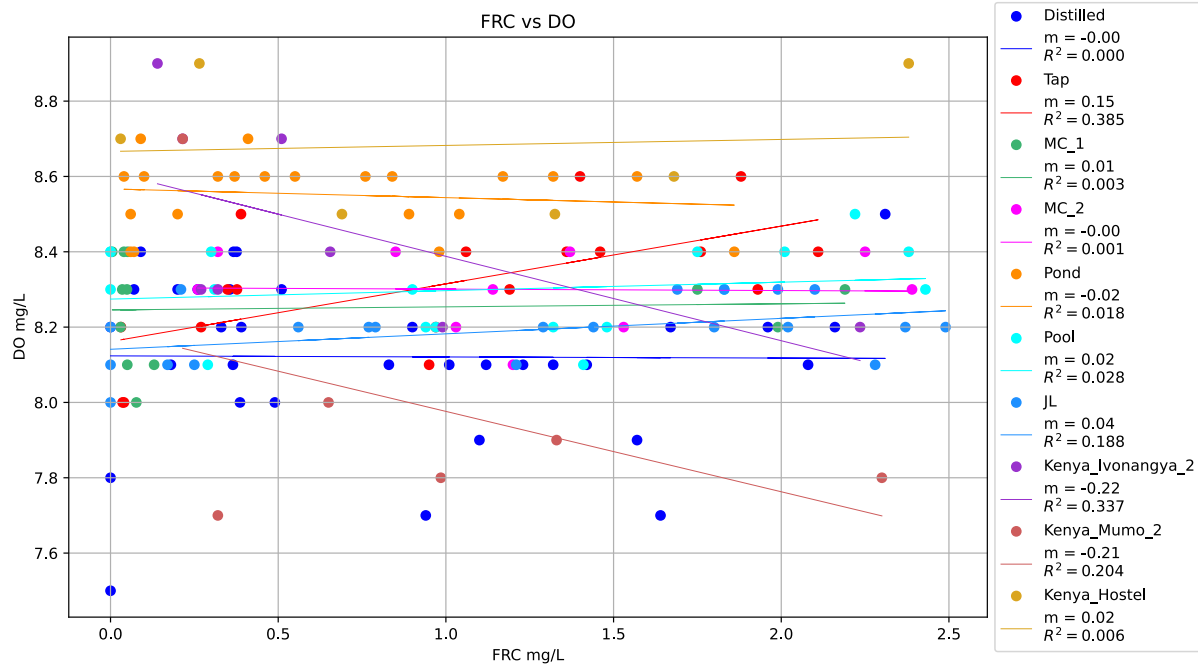


Figure 2.9: Dissolved Oxygen (DO) trends for different sets of samples, with gradients ( $m$ ) and determination coefficients ( $R^2$ ) shown in the legend. Naming conventions of sets are detailed in table 2.1.

Figure 2.9 shows the evolution of dissolved oxygen levels for different water sources for different FRC values. No sets hold a strong correlation between FRC and dissolved oxygen levels, which is expected as they are not involved directly in reactions together. The range over which the dissolved oxygen varies is also quite small. Most datapoints are in the  $8 \rightarrow 8.6$  mg/L range, which is quite tight. Waters analysed for aquatic life can appear in the range of 3 to 12 mg/L [5]. This suggests that water sources that this project is concerned with will not vary much in dissolved oxygen levels, and that this is representative of waters used for drinking water in general. Therefore, it is likely it will not be a useful measurement to determine FRC. High dissolved oxygen levels will push the ORP curve higher as oxygen is also an oxidant, so the measurement could have been relevant; the small variance in these waters are probably not enough for this effect to have an impact.

## 2.3 Data Collection Conclusions

From the analysis of individual parameters against the changing FRC, the ORP provides the most valuable information, as it holds a relationship with FRC that is reasonably independent of the source. This suggests it will be the most useful parameter involved in data analysis for identifying the FRC. However, measuring the ORP also had the most difficulty compared to other sensors, so this will impact its reliability and could mean it will be difficult to use in the field. The conductivity held some relationship with the FRC that was somewhat consistent over the different sources, but the relationship was very source-dependent. The pH did not seem to have any direct correlation, but it is still expected to be important when used in conjunction with the ORP. This suggests that



these parameters may still add value in determining FRC, but to a lesser extent than ORP.

The turbidity measurement had a very weak correlation with the FRC. Due to possibly affecting the effectiveness of chlorine, there may be some use of measuring it in conjunction with the ORP, but otherwise it will likely not be a useful measurement to include.

The dissolved oxygen had a vary small range of variation over all the samples, and also did not seem to have any correlation with the FRC. Because of this, it is likely that it will not be useful in determining the FRC.

# Chapter 3

## Data Analysis Approaches

The aim of the data analysis is to be able to determine the FRC from a range of different measured variables. In the best case, it will produce a mapping of the variables measured, to one continuous dimension that is aligned with the FRC (i.e. input conductivity value, pH value, ORP value etc. and output FRC). This is a dimensionality reduction problem starting with the number of dimensions equal to the number of measured parameters, and aiming to reduce that to one dimension. As this is unlikely to be achieved with sufficient accuracy, datapoints should also be classified into groups defined by ranges of FRC. This is relaxing the aim of having a continuous measure of FRC, but still allows us to mark datapoints as being in a certain FRC range.

For the classification, the FRC range is divided into groups detailed in table 3.1. These are based on the WHO guidelines [27] mentioned in table 1.1, but the  $(0 \rightarrow 0.2)$  and  $(0.2 \rightarrow 0.5)$  groups have been combined. This approach will be referred to as ‘3-class’ classification. This is because it is still reasonable to treat these samples as having too little, enough, and too much chlorine, and also gives better separation between the classes. Distinction between the two groups that have been combined proves very difficult in analysis, so the improved simplicity of classification justifies combining the groups.

Class	FRC Range (mg/L)	Short Label	Details
<b>1</b>	$FRC < 0.5$	too low	should be dosed with more chlorine
<b>2</b>	$0.5 \leq FRC < 2$	safe	safe to drink and store for 24 hours
<b>3</b>	$2 \leq FRC$	too high	water unpotable

Table 3.1: Classification ranges for the 3-class classification approach in data analysis.

Another classification will also be conducted purely on identifying whether the sample is below or above 0.5 FRC. This will only be able to serve as a marker on whether the sample is fully disinfected or not. This method will be referred to as the ‘2-Class Classification’. This distinction is still useful in the field as a consumer could use the classifier to be happy the sample is disinfected, and judge the potability based on smell and taking a sip.

Since this analysis is on a small number of datapoints, it is infeasible to look at deep learning models for analysing the data. This report looks at using Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) to solve the problem. All these methods work on using Singular Value

Decomposition (SVD) on matrices produced from the data, which is feasible for a small dataset. They all involve first normalising the data. The effectiveness of each decomposition is evaluated by:

1. Finding the correlation between the datapoints mapped onto vectors found in the space of the measured parameters, and the FRC of the datapoints. This will be scored by the resulting  $R^2$  value.
2. Making 3-class and 2-class classifiers based on the decomposition from each method, and calculating accuracy scores on the classification (proportion of points correctly classified).

Both of these scores will be calculated based on training/test data splits, with a random 80% of the datapoints being used in the training set, and the remaining 20% being used in the test set to calculate correlation or classification scores. These splits and calculations will be repeated 50 times, and the average scores over all the runs will be reported. This is in attempt to not have the results be based on over-fitting, and give a reflection on the scores obtained seeing new data. We chose to conduct 50 runs, as it showed to be past the point where the average score over runs had converged.

The decision on what would count as a ‘successful score’ for correlation or classification is not concrete and quite arbitrary. This is because the final dataset is quite small (less than 200 datapoints), so is unlikely to get scores high enough to be acceptable to use in the field (less than 1% error perhaps), and would be subject to field needs. Therefore, the scores obtained should show whether there is strong correlation between the method output and FRC, to justify taking the analysis forward and optimising it in the future (e.g. using a much larger dataset). A correlation threshold of greater than 0.75, or a classification accuracy higher than 0.8, is used for this purpose.

### 3.1 ORP as a baseline

The measurement of the ORP holds significant correlation to the FRC by itself; the correlation over all the datapoints of measured ORP and FRC was 0.704. This suggests using the ORP alone might give a reasonably accurate FRC, or be used to classify points. This can be evaluated using the trend fit of the ORP calculated (eqn 2.3) to give predicted FRC values that can correlate to actual FRC values. It can also give ORP bounds for each class, based on the ORP values that map to the FRC bounds of each class.

For this analysis, the training/test splits will give training data to be used to find values for the parameters in eqn 2.3, as well as ORP bounds for classification in the three class and two class case. The test data will be used to calculate correlation and classification scores from the trend fit and ORP bounds calculated. This will be repeated 50 times to get average scores, which can be used to compare against other data analysis methods. The results are shown in table 3.2. One point to note is that because the ORP-FRC curve has a plateau, ORP values above the plateau will not map to an FRC. Consequently, these points are assumed to have  $> 2.5$  FRC, and so will be classified as such.

Table 3.2 shows the ORP scores are below the desired thresholds. The reduced correlation when compared to the full dataset’s correlation (0.704) is expected. Less data is used in each run to calculate the curve fit, and the correlation is calculated on previously unseen

	$R^2$	3-Class	2-Class
<b>ORP Method</b>	0.527	0.576	0.727

Table 3.2: ORP  $R^2$  and classification scores. Note ‘3-class’ refers to the classification score with three classes, and ‘2-class’ is the score with two classes.

datapoints, compared to seen datapoints for the full dataset value reported earlier. Overall, these scores show that the ORP does not provide sufficient information to quantify the FRC accurately, but do present a baseline that we can compare our analysis techniques against.

## 3.2 Principal Component Analysis (PCA)

### 3.2.1 Background

PCA is an unsupervised learning technique that aims to reduce the dimensionality of a dataset whilst minimising information loss [19]. It does so by finding unit vectors, in the space of the variables measured, that maximise the variance between datapoints. This is in order to maximise the separation between datapoints, so that it is easier to identify the class they belong to. These unit vectors are called ‘principal components’ (PCs). The first step is normalising the data so that the dataset has a mean of 0 and unity variance for each variable. The first principal component (PC1) is then found to be the direction in the space that, when all datapoints are mapped on to the direction (the dot product between the direction and the datapoints is taken), maximises the sum of the magnitudes. This is equivalent to finding a line of best fit over all the data in the normalised space. The second principal component (PC2) then does the same but over all directions perpendicular to the first. This repeats for as many principal components as required, or until the number of dimensions is reached [3].

### 3.2.2 Methods of Implementation

Our hypothesis when using PCA is that the changing FRC would be a strong cause of variance between samples, because it is a purposefully varied parameter, so a principal component might be encouraged to align with it. From the data collection results, it appeared that the ORP, pH and conductivity (Cd) were the best parameters to begin with in analysis, as they appeared to deliver the most information on the FRC. After this, the effect of including dissolved oxygen (DO) and turbidity (Tb) in analysis was observed, to see if they improved the results gained.

PCA does not have any classification method associated with it directly. It is used for dimensionality reduction. However, for the purpose of comparing the different analysis techniques explored, it is feasible to implement an ad-hoc method of classifying the datapoints after finding a PC. Training datapoints will be labelled with the class they belong to, and a mean for each class will be calculated. This mean is projected on to the PC found. Test datapoints are classified by projecting on to the PC, and assigning them to the class whose mean is closest. One issue is this method assumes identical variances in each class, which is known not to be the case, as the FRC ranges are not equal in size. However, PCA does not have any attributed classification method, so this can serve as a

simple approach to compare to in-built classification methods used in the other analysis techniques.

PCA first normalises the data, and can only involve linear mappings from variables to potential principal components. Figure 2.7 shows that the ORP does not have a linear relationship with the FRC. However, it is feasible to consider dividing the ORP trend into two ‘linear’ regions. One for the steep trend at the start, and one for the shallow trend for greater FRC. PCA could then be used to find a mapping for the separate regions, using the cutoff line marked in Figure 2.7.

The idea behind this would be that a sensor in the field could use the initial ORP measurement to separate datapoints into a ‘potentially still infected’ group and a ‘potentially too high FRC’ group, based on if the ORP was below or above the cutoff. The mapping provided by the PCA could then identify if the estimated FRC was an issue (i.e. under 0.2mg/L, or over 2mg/L). For this analysis, correlations of each region’s principal component with the FRC were used, and then a classifier score on marking samples ‘problematic’ (too low or too high FRC) or ‘unproblematic’ (safe FRC).

An important note to make is that even if the FRC does not align with PC1, it could still be useful if it aligns with a subsequent one. This would mean that the FRC isn’t the main cause of variance among samples, but the prior principal components could be accounting for variance between sources, that want to be removed anyway. For example, if two sources have vastly different ‘baseline’ conductivities, before any chlorine is added, but hold a similar trend for change in conductivity with changing FRC, then PC1 may align with the conductivity between sources, and PC2 could align with the conductivity change due to FRC.

### 3.2.3 Results and Discussion

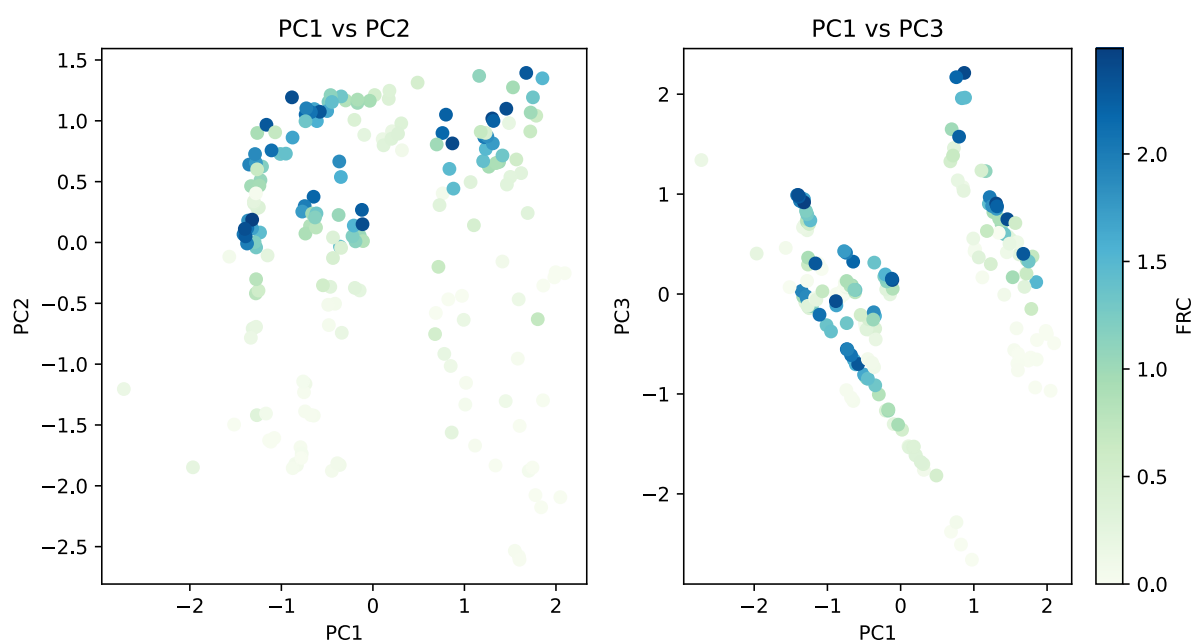


Figure 3.1: PCA results on analysis using the parameters Cd, pH and ORP.

Figure 3.1 shows the datapoints plotted in the space of the three PCs found using the

	PC1	PC2	PC3
<b>R<sup>2</sup></b>	0.193	0.610	0.355
<b>unit vector</b>	[ 0.72 -0.69 -0.12]	[-0.12 -0.29 0.95]	[0.69 0.66 0.29]

Table 3.3: Initial PCA correlations between PCs found and the FRC of all datapoints, within a 0 to 2.5 FRC range and including Cd, pH and ORP parameters.

ORP, conductivity and pH parameters. The shading indicates the FRC, as shown on the right. Out of the three PCs, PC2 appears to be the best aligned with changing FRC. This is supported by the PC2 having the largest  $R^2$  value. However, although a strong correlation, it is far from a direct relationship, and still has a lot of error within it. For the plot of PC1 vs PC3, there is a distinct separation between two sets of datapoints. Both PC1 and PC3 have a high weight on the conductivity, as seen in table 3.3, by the high magnitude of the conductivity component of both of their vectors. Therefore, the conductivity is expected to be the cause of the separation. There are no datapoints in the 1.5 to 2.5 mS/cm range, as seen in Figure 2.5, which is highlighted in this plot. However, this separation is not helpful in identifying the FRC.

Trial	Parameter Weightings					R <sup>2</sup>	3-Class	2-Class
	Cd	pH	ORP	DO	Tb			
<b>1</b>	0.18	0.26	0.92	-	-	0.603	0.601	0.786
<b>2</b>	0.33	0.44	0.62	0.37	-	0.536	0.560	0.730
<b>3</b>	0.49	0.57	0.57	-	0.27	0.475	0.549	0.715

Table 3.4: PCA  $R^2$  and classification scores on different combinations of parameters, labelled as ‘trials’. Parameter weightings indicate the average weight applied to each parameter in the PC with the highest correlation to FRC for each run. A dash indicates the parameter was not included. Note ‘3-class’ refers to the classification score with three classes, and ‘2-class’ is the score with two classes. Both classifications have been done using the ad-hoc method detailed in section 3.2.2.

Table 3.4 summarises the average scores obtained for different settings of PCA conducted. In trial 1, the ORP ended up being very heavily weighted in the PC that was best aligned with the FRC. This is expected, since the two parameters are strongly correlated. Both the pH and conductivity are given low weights, which means the variance in them was largely not associated with the FRC. This was expected from the results in the data collection. The  $R^2$  and 3-class classification scores are slightly improved from the ORP alone, although not significantly.

However, even though a simple ad-hoc method was used, the 2-class classification score has dramatically improved, coming close to the threshold value of 0.8. This could be due to using a linear relationship rather than the non-linear relationship used for the ORP alone. Imagine a linear line-of-best-fit plotted over Figure 2.7; the new ORP cutoff for 0.5 FRC would be less. This would make more misclassifications on (high ORP, low FRC) datapoints, but fewer misclassifications on (low ORP, high FRC) datapoints.

In trial 2, incorporating the dissolved oxygen measurement had a negative effect on all the scores. This means that the measurement included much more variance that wasn’t associated with the FRC than variance that was associated. PCA is just trying to find directions of maximum variance. Here it is used in the hope that the variance is being

caused by the FRC, as it is a parameter purposefully changed throughout the dataset. The dissolved oxygen seems to vary much more in ways not dependent on the FRC, so it makes the analysis output worse. From the reasons detailed in table 1.2, and the results from data collection, this outcome is not surprising.

Trial 3 shows that the turbidity behaved similarly to the dissolved oxygen, and the variation introduced by it had little correlation to the FRC. From table 1.2, it was hypothesized that using the turbidity in conjunction with the ORP may have added value, but within PCA this did not happen. In both the dissolved oxygen and turbidity trials, the ORP was given smaller weights than when they weren't included, so it is not surprising that the correlations were worse.

ORP Range	Parameter Weightings			R <sup>2</sup>	Classification
	Cd	pH	ORP		
$< 500mV$	0.42	-0.23	0.88	0.452	0.714
$\geq 500mV$	0.44	0.85	0.28	0.517	0.627

Table 3.5: PCA  $R^2$  and classification scores for the two ‘linear’ regions of the ORP vs FRC relationship.

PCA can only involve linear mappings. Figure 2.7 shows the relationship between ORP and FRC is definitely not linear, but could be seen as two ‘linear’ regions, so this analysis observed the effectiveness of using PCA in these two regions. In this case, the classification was based on whether the FRC was predicted to be too low for the low ORP group ( $< 0.2mg/L$ ), or if it were too high for the high ORP group ( $> 2mg/L$ ). The correlation scores ended up being significantly lower than the threshold. For the low ORP group, there is still a high weight on the ORP, which is expected due to the strong, steep linear trend for this region in the ORP vs FRC relationship. The lack of a strong correlation is because the data had too much variance unrelated to the FRC.

The ORP was given a low weight in the high ORP group. This could be due to the very shallow trend in this linear relationship. The pH was instead heavily weighted, which is unexpected given there was no strong relationship between the pH and FRC directly. Correlation was seen between the conductivity and FRC in the data collection, but the conductivity was still not given a large weight. The outcome had a low correlation anyway, so this is not providing strong support for the weightings used in this method. High conductivity and ORP weightings may have performed better, but PCA did not find enough variance along those weights to identify it.

### 3.2.4 Conclusions

The scores obtained from PCA were higher than those gained by using ORP alone, but none of the scores achieved by PCA came close to the thresholds defined. This shows that it did not provide a correlation that could be used as a quantitative measure of FRC, or classification scores that would suggest it could be used as a classifier. This is likely due to the high variation in the measured parameters that isn't linked to changing FRC, and PCA is aimed at finding directions of maximum variance. The best performance of PCA was when conductivity, pH and ORP were included, and ORP was the most important out of these. This supports the prediction from the data collection that these parameters would be the most useful. Dissolved oxygen and turbidity were both not useful parameters

to be included in the analysis, as they both had a negative impact on the scores when included. This supports the results from data collection, which had a similar conclusion.

As it is unsupervised, PCA cannot be influenced to align with a direction correlated to the FRC. This is where Linear Discriminant Analysis may become more useful, as the groups chosen for it to separate can influence choosing a direction correlated to FRC.

### 3.3 Linear Discriminant Analysis (LDA)

#### 3.3.1 Background

LDA is similar to PCA, in that it involves creating matrices from the data, which can be decomposed into unit vectors that can provide dimensionality reduction. However, the matrices made and the objectives of the methods are different. LDA is a supervised learning method, so each datapoint in the training set is labelled with its class. Instead of trying to maximise the variation of datapoints along an axis, it aims to maximise the separability of different classes in the dataset [3]. The unit vectors align to maximise separation between classes from a trade-off of two metrics:

1. Maximise the difference in means of classes
2. Minimise the variation (scatter) within a class

When there are more than 2 categories, a central point in the data is found, and the distance of means to that point is maximised. LDA is a Gaussian Maximum-Likelihood Classification Technique. This means each class is assumed to have a multivariate Gaussian probability distribution with an associated mean and variance; classification is built into the method because datapoints can be classified based on which Gaussian they are most likely to come from (whichever Gaussian gives the maximum probability for that datapoint). In addition, LDA produces unit vectors, that can themselves be used to produce a continuous mapping from measured parameters to a predicted FRC.

LDA is supervised so it can be trained to specifically find a direction aligned with changing FRC, since classes are defined by different FRC ranges. This means the resulting unit vectors (Linear Discriminants (LDs)) might align better with the variation due to FRC itself than in PCA.

One notable issue with LDA is that it assumes normally distributed data, statistically independent features, and identical covariance matrices for every class. This is not the case for this problem, as ORP variation is different for different classes, and these variables can't be assumed to be independent. However, this only applies for LDA as a classifier and LDA for dimensionality reduction can also work reasonably well if those assumptions are violated [18].

#### 3.3.2 Methods of Implementation

Similar to PCA, analysis first involved just the conductivity, pH and ORP parameters, and then looked at incorporating dissolved oxygen and turbidity and comparing the output. As opposed to PCA, including more measurements in the training should not adversely affect the output, as the method only aligns with maximum variance between classes. Therefore, if the turbidity does not provide variance attributed to the FRC, LDA should



ignore the contribution from it (place a small weighting on the turbidity in the unit vector). Another important distinction between PCA and LDA is that in LDA, the best aligned unit vector that is obtained from the decomposition will always be the first linear discriminant (LD1); it is a supervised learning technique so it is purposefully aligning with the FRC. Because of this, only LD1 (and maybe LD2) are important when seeing how effective the analysis was.

### 3.3.3 Results and Discussion

	Vector	R <sup>2</sup>	Classification
<b>LD1</b>	[0.18 0.29 0.94]	0.691	0.690
<b>LD2</b>	[0.77 0.61 -0.21]	-0.024	

Table 3.6: Initial LDA correlations between LDs found and the FRC of all datapoints for a 3-Class split, within a 0 to 2.5 FRC range and including DO, Tb and Cd parameters. The classification is done using both LDs, which is signified by the merged cell. These values relate to the LDs in Figure 3.2.

Table 3.6 shows the resulting unit vectors for LDA over the whole dataset using pH, conductivity and ORP in a 3-class split. The correlation and classification scores do not meet the thresholds defined. Figure 3.2 shows the difficulties in separating the classes using the LDs found, by plotting the actual classifications against their predictions. Actual classifications maintain a significant overlap between classes, particularly in the  $(0 \rightarrow 1)$  region of LD1, which has prevalence of all three classes, although mainly from the  $(0.5 \rightarrow 2)$  FRC) class. This overlap is what the method is trying to avoid, as it cannot distinguish between classes. However, this decomposition was the best achievable by the method, meaning the overlap was found to be very difficult to remove.

Figure 3.2 also shows that there seems to be a non-insignificant difference between the direction to separate class 1 from class 2, compared to the direction to separate class 2 from class 3. This is likely due to the stark change in gradient in the ORP curve. From Figure 2.7, it is evident that ORP is more helpful at distinguishing between ‘ $FRC < 0.5$ ’ vs ‘ $FRC > 0.5$ ’, compared to distinguishing ‘ $FRC < 2$ ’ vs ‘ $FRC > 2$ ’.

Some of the datasets hold a significant number of the overlapping points seen. One of these is the pool water samples, which contains a number of class 1 points with a value greater than 0 in LD1. It also contains a large number of the points in the cluster seen in  $0.5 \rightarrow 1$  range for LD1 and around 0.75 in LD2. Looking at the data collection results, the pool water was among the groups with high mean conductivity, and the plateau in ORP can also be seen to be quite high. These characteristics could explain the typically higher values (in the LD1 direction) for the class 1 points, and why the class 2 and class 3 points get clumped.

The dataset ‘Kenya Invonangya 1’ was also a big contributor to the overlapping datapoints, and can be explained in a similar way to the pool water, due to having a high mean conductivity and high ORP in the plateau. Since this dataset is representative of a source that might use the FRC sensor, this indicates that the method would be less reliable.

The tap water data also contributed to some of the datapoints from class 1 with higher values in LD1. This is surprising considering the low mean conductivity. However, Figure

2.7 shows there are some tap water datapoints with a high ORP but less than 0.5 FRC, which most likely contributed to these overlapping datapoints. Although there are distilled water datapoints in the same region of the ORP graph, they had an extremely low conductivity, and also lower pH than the tap water, which would be why they don't overlap. Many of the (low FRC, high ORP) datapoints come from the tap water and distilled water samples. The distilled water is a very simple solution, and the tap water is likely to have a simpler composition than water in Kenya. This could mean that these datapoints

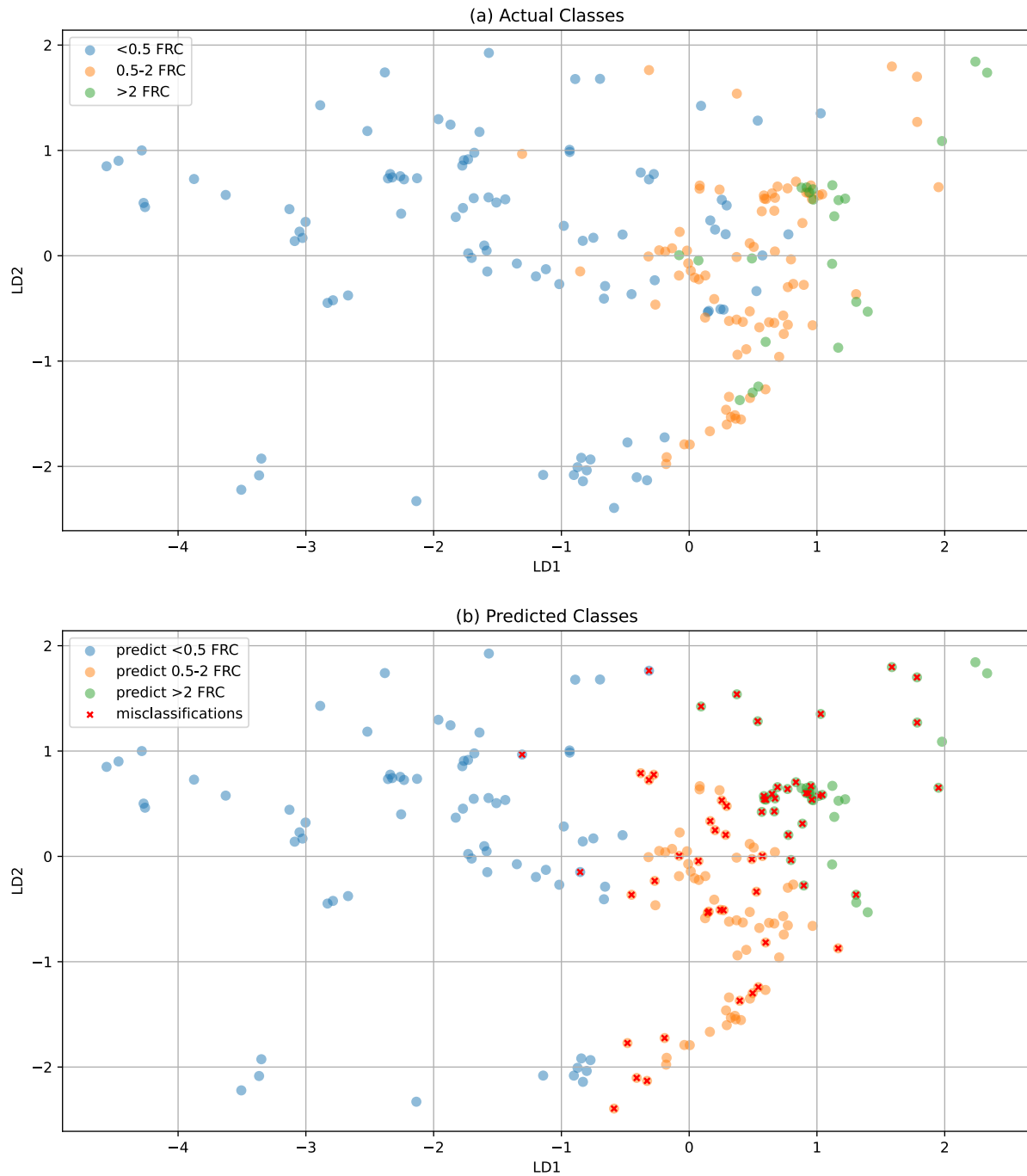


Figure 3.2: **LDA results** with data split into **3 classes** using **Cd, pH and ORP** parameters. (a) shows the training labels on the datapoints plotted, and (b) shows the labels predicted by the LDA. This shows which points have been misclassified.

are not so representative of values expected in water samples treated. However, this range of water samples would ideally still be represented in the output of a classifier, so they are still relevant in evaluating the performance of the classification.

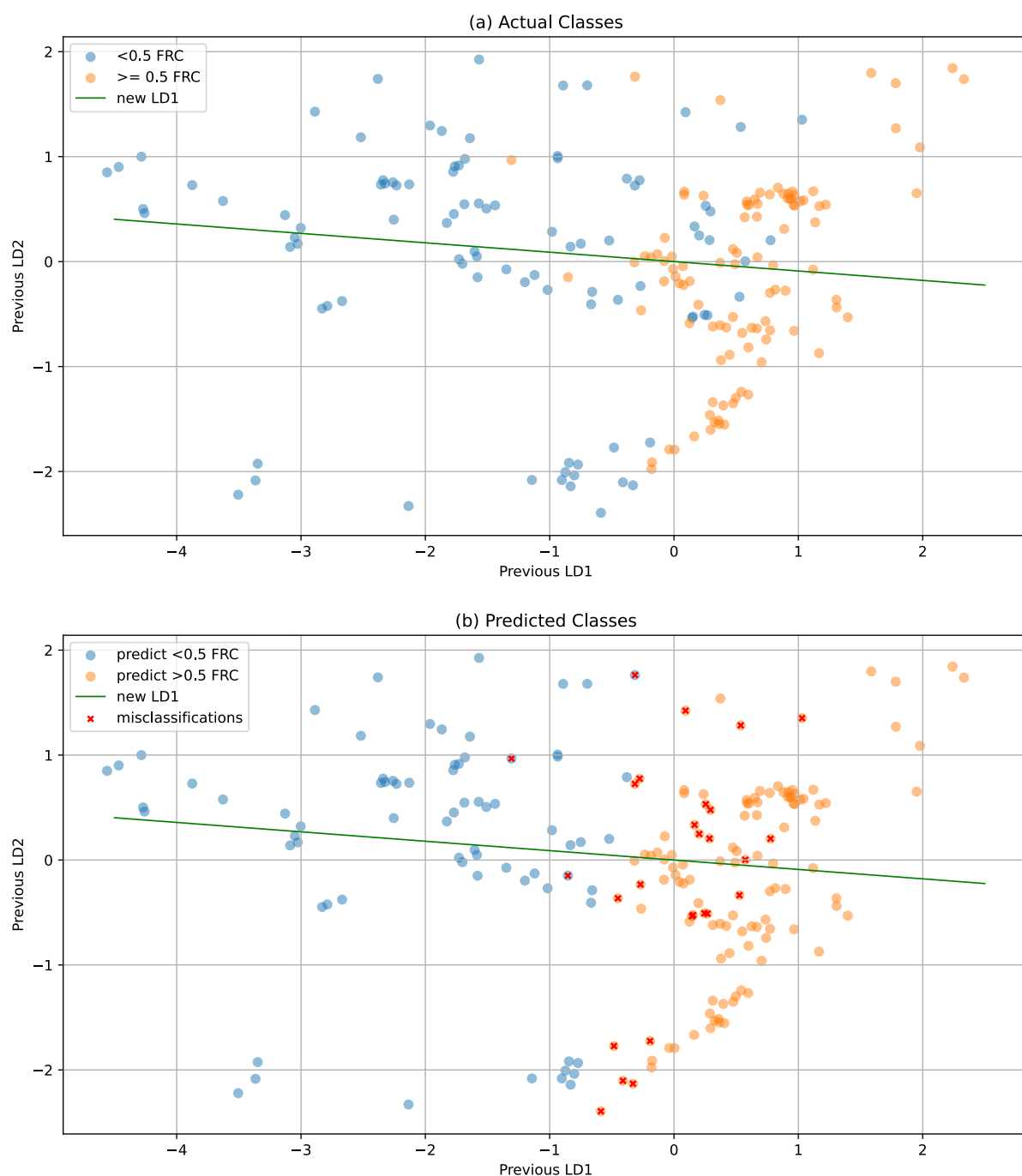


Figure 3.3: **LDA results** with data split into **2 classes**, using **Cd, pH and ORP** parameters. This has been plotted on the LDA directions found in the 3 class-split for easy comparison, with the LD from the 2-class split plotted. (a) shows the training labels on the datapoints plotted, and (b) shows the labels predicted by the LDA. The plots show which points have been misclassified.

Table 3.7 shows the resulting unit vectors for LDA over the whole dataset using pH, conductivity and ORP in a 2-class split. Interestingly, there is a slightly improved correl-

	Vector	R <sup>2</sup>	2-Class
<b>LD1</b>	[0.13 0.25 0.96]	0.713	0.863

Table 3.7: Initial LDA correlations between LDs found and the FRC of all datapoints for a 2-Class split, within a 0 to 2.5 FRC range and including Cd, pH and ORP parameters.

ation of this unit vector compared to the 3-class split, although it still does not meet the threshold. The difference between the two unit vectors is a slight increased weight on the ORP for the 2-class unit vector. Figure 3.3 shows the new LD1 plotted in the space of LD1 and LD2 found from the 3-class split. This change in direction supports the difference in directions found for separating different sets of classes in Figure 3.2. The distinction between class 1 and class 2 is more important than distinguishing class 2 from class 3, as the former distinguishes potentially infected water from disinfected water; using LD1 from the 2-class split makes sense as well.

Classification on 2-classes is most-likely going to be less challenging than on 3-classes, so it is not surprising that this classification score is higher. That being said, the classification score surpasses the threshold assigned, which indicates this method could be useful for classification.

Trial	# Classes	LD1 Weightings					R <sup>2</sup>	Classification
		Cd	pH	ORP	DO	Tb		
<b>1</b>	3	0.18	0.29	0.94	-	-	0.70258	0.659
<b>2</b>	2	0.13	0.25	0.96	-	-	0.70261	0.852
<b>3</b>	3	0.14	0.42	0.89	0.11	-	0.68305	0.647
<b>4</b>	2	0.09	0.42	0.89	0.13	-	0.68306	0.843
<b>5</b>	3	0.17	0.28	0.94	-	0.04	0.69108	0.666
<b>6</b>	2	0.13	0.26	0.96	-	0.03	0.69112	0.862

Table 3.8: LDA  $R^2$  and classification scores for different settings. Each trial is done on a 3-class or 2-class split. The ‘weightings’ are the average magnitude (over the 50 runs) placed on each measured parameter in LD1.

Table 3.8 summarises the average scores obtained for different settings of LDA conducted. Across all settings, the average correlation obtained is extremely similar between the 2-class and 3-class splits, although it is always slightly higher for the 2-class split. This is most-likely due to the difference in direction highlighted from Figures 3.2 and 3.3. In trials 1 and 2, a very high weight placed on the ORP, which further suggests this is the most important measurement for determining FRC. The 2-class classification scores surpass the threshold, which suggest this method is likely to be useful as a classifier.

From trials 3 and 4, including the dissolved oxygen and turbidity is seen to be mostly negative. The one positive influence has been on the average classification score of the 2-class split including the turbidity. This is therefore further evidence that these measurements are not providing useful information for determining the FRC, especially considering that the weights on dissolved oxygen and turbidity are all small.

### 3.3.4 Conclusions

The correlation scores achieved by LDA (0.703 at best) were higher than those achieved by PCA, but did not meet the desired thresholds. This suggests LDA could not provide a quantitative measure of FRC. From the large overlap of regions seen in Figures 3.2 and 3.3, the problem appears too complex for a linear analysis to provide a strong correlation. However, the classification score from the 2-class split (0.863) exceeded the threshold, and suggests that this method was successful in distinguishing between low and high FRC samples. It is important to note that this score would still likely be too low to justify direct implementation in the field, as that would require an accuracy of perhaps 99% or higher. Nevertheless, it is a strong signal that this method could be analysed further, with larger datasets, to find a result that could meet field needs.

Turbidity and dissolved oxygen were again not useful when included in the analysis, as they both had a mostly negative impact on scores obtained. ORP was also again most highly weighted. These observations are all in line with what was expected from the outcomes in data collection.

The results also show that this problem is too complex for linear separation methods such as LDA, as the decision boundaries appear to be non-linear. We would require a much larger dataset to train methods that can use non-linear mappings for finding a quantitative measure of FRC, however this also motivates looking at non-linear decision boundaries for classification, to see if the scores can be improved. For this reason, Quadratic Discriminant Analysis may prove more useful, as it can have non-linear decision boundaries.

## 3.4 Quadratic Discriminant Analysis (QDA)

### 3.4.1 Background

QDA is a variation of LDA, in which each class has a unique covariance matrix, rather than assuming that all classes have equal covariance matrices [3]. Because of this, QDA can find curved decision boundaries rather than only linear decision boundaries, which was the case for LDA. However, the disadvantage is that QDA cannot be used for dimensionality reduction and can't be used to map to a continuous measurement like LDA or PCA, as it does not output unit vectors aligned with the separation. Nevertheless, the curved decision boundaries mean that it has the potential to provide better classification [7].

### 3.4.2 Methods of Implementation

Similar to LDA, analysis involved varying the parameters included and the number of classes used, in order to explore the performance. However, due to QDA not outputting any vectors, there was no way of finding some direction that could be correlated to the FRC. As a result, the only thing to compare is the classification scores.

### 3.4.3 Results and Discussion

Table 3.9 shows the resulting classification scores using QDA in a 3-class and 2-class split on the whole dataset with the pH, conductivity and ORP parameters included. Both scores are quite high, with the 2-class score surpassing the threshold of 0.8. This is evidence for the method having practical application as a classifier.

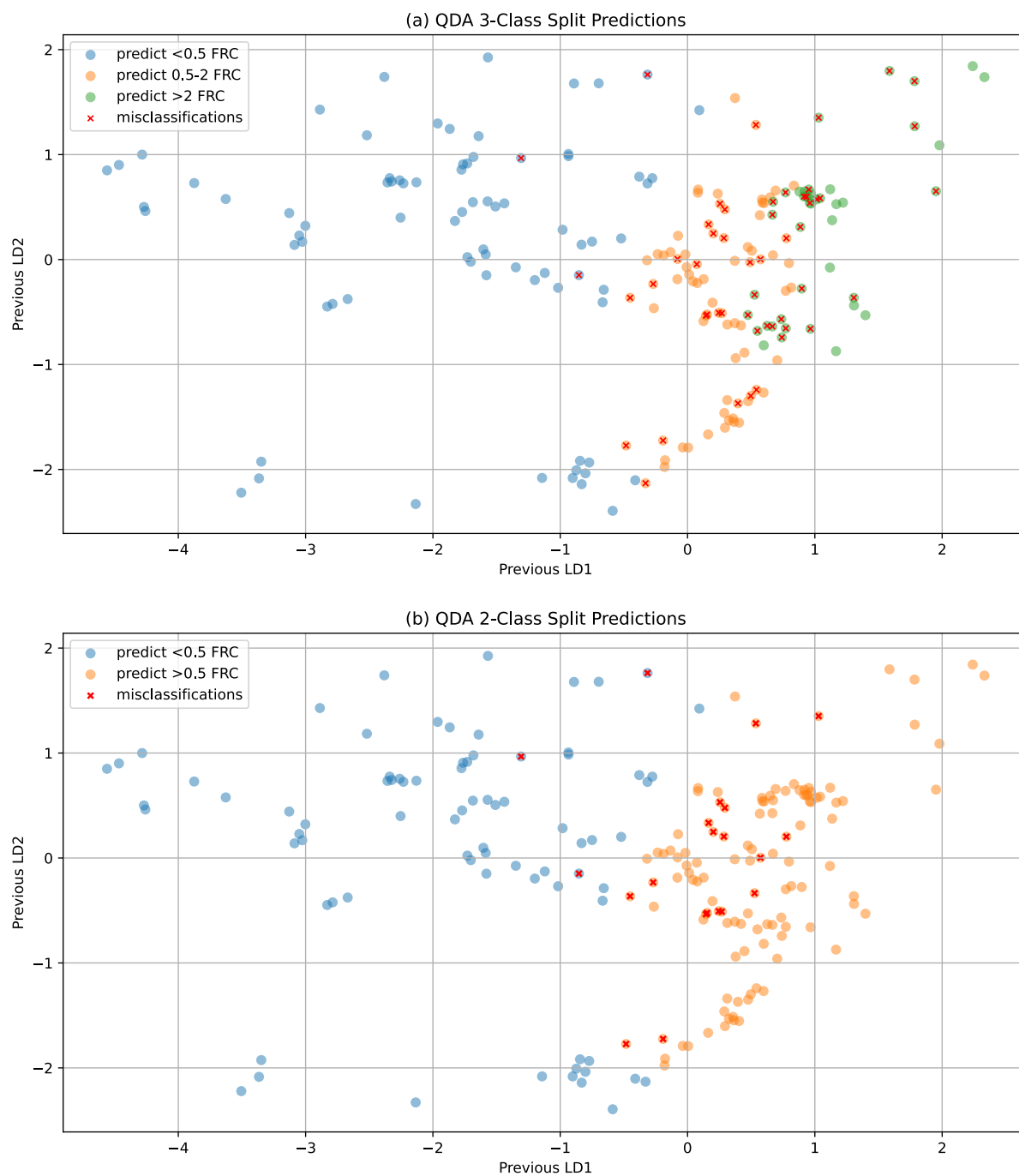


Figure 3.4: **QDA results** for **3-class** and **2-class** splits of the data. These have been plotted on the vectors found in the 3-class LDA from Figure 3.2, to allow for easy comparison. These figures show the misclassifications from the actual assignments seen in Figures 3.2 and 3.3.

	3-Class Score	2-Class Score
<b>All Data</b>	0.777	0.893

Table 3.9: Initial QDA classification scores for a 3-Class and 2-Class split over the whole dataset.

Because QDA does not output any vectors related to the decomposition, the classifications made by QDA have been plotted on the LDA directions found from the 3-class split, to compare the decision regions. QDA is able to have non-linear decision boundaries, which allows it to overcome the overlapping regions slightly better. This is seen in the improved classification scores.

<b>Trial</b>	<b>Parameters</b>	<b>3-Class Score</b>	<b>2-Class Score</b>
<b>1</b>	Cd, pH, ORP	0.762	0.891
<b>2</b>	Cd, pH, DO, ORP	0.728	0.865
<b>3</b>	Cd, pH, Tb, ORP	0.74	0.873

Table 3.10: QDA classification scores for different sets of included parameters under 2-Class and 3-Class splits.

Table 3.10 summarises the average classification scores for 2-class and 3-class splits on data including different sets of parameters. The high classification scores have been mostly retained despite splitting into training and test data, which suggests the method can generalise over varying water sources. There are again reduced scores when including the dissolved oxygen and turbidity parameters. This is a strong indication that these parameters are not useful for determining FRC.

### 3.4.4 Conclusions

The classification score achieved by QDA for a 2-class split beat the threshold for acceptance, and also beat the score achieved by LDA. This supports QDA for being viable in classifying water samples into ‘not enough’ or ‘enough’ chlorine, although, similar to LDA, more analysis on a much larger dataset would be required to justify direct application. The 3-class split score (0.762) also improved from LDA, and came close to the desired threshold. This suggests that QDA may also have practical application in classifying water samples into having ‘too much chlorine’ as well, although this is less conclusive. Dissolved Oxygen and Turbidity again had a negative impact on scores, which suggests they are not useful parameters to include.

## 3.5 Comparisons

<b>Analysis Method</b>	<b>Scoring Method</b>	<b>Randomised Test Data</b>	<b>Site-Specific Test Data</b>
<b>LDA 2-class</b>	correlation	0.703	0.733
<b>QDA 2-class</b>	2-class classification	0.891	0.861
<b>QDA 3-class</b>	3-class classification	0.762	0.694

Table 3.11: Evaluations of the analysis techniques for each scoring method with a new test set of the ‘Kenya Ivonangya 2’, ‘Kenya Mumo 2’ and ‘Kenya Mumo 1’ datasets.

Across all the analysis techniques explored, dissolved oxygen and turbidity measurements did not provide significant useful information in determining the FRC. Within the scores obtained from using the pH, Conductivity, and ORP parameters, LDA gave the best correlation score (0.703) from trial 2 in table 3.8. QDA gave the best 3-class (0.762) and 2-class (0.891) classification scores, from trial 1 in table 3.10. To see how they might

perform in a real-life scenario (i.e. with no data on the water source being evaluated), their performance has been further evaluated by excluding some of the full datasets from the training set, and using them as the test sets. These sets are ‘Kenya Ivonangya 2’, ‘Kenya Mumo 2’ and ‘Kenya Mumo 1’. The results are shown in table 3.11. The QDA scores have both reduced. The decomposition is no longer training on any data from the source that it is testing on, so it may not be representing that set as well. However, the results have not dropped very significantly, which is a positive sign that the method can generalise well.

Surprisingly, the correlation score from LDA improved significantly. The new unit vector found was [0.13 0.3 0.94] for the parameters [Cd pH ORP], which is very similar to the vector before, except for a slightly larger weight on pH and slightly lower weight on ORP. This suggests that the sets used for testing are well represented by the training sets. Another reason for this performance could be that using a full dataset to train on means it is guaranteed that the test set will have a full range of FRCs included. When done randomly, it is possible some of the test sets are all low FRC values, which might lead to lower scores due to having less variation over which to compare.

QDA is the only method evaluated here that is able to find non-linear decision boundaries. Since it has performed better than the other methods, this motivates using more complex, non-linear mappings of the measured parameters, where it might have better results.



# Chapter 4

## Summary and Conclusions

### 4.1 Summary of Aims and Indicators of Success

The aim of this project was to identify whether it is feasible to use the measurement of common water parameters to quantify the FRC of drinking water samples that have been chlorinated. The quantification was done by correlating an outputted mapping of parameter measurements to FRC measurements, to assess whether the mapping could act as a quantitative measure of FRC. Quantification was also explored by creating classifiers on classes defined by FRC ranges. The first set of ranges was for a 3-class split detailed in table 3.1, which had the meaning of classifying samples into having ‘not enough chlorine’, ‘enough chlorine’ or ‘too much chlorine’. The second set of ranges was for a 2-class split, predicting whether samples had FRC below or above 0.5. This was to serve as just identifying whether a sample was disinfected or not.

Due to the constraint of a limited dataset, the success was determined by whether a method could do well enough to justify that further analysis may make it feasible to use in practical application. This is because it was extremely unlikely that scores high enough for direct application ( $< 1\%$  error perhaps) would be achieved. Scores of 0.75 for correlation and 0.8 for classification were decided as thresholds for success.

### 4.2 Conclusions and Future Work

Referring back to the project objectives, we were able to produce a method for creating and measuring samples reliably that gave repeatable results. Measuring most parameters was simple and outputs were reproducible. The one exception was measuring ORP, which presented difficulties when using the ORP probe, such as dependency on the probe’s recent (in the previous hour) history, and long stabilisation times. This could mean that this measurement is less reliable and more difficult to take in the field.

The direct relationships between each measured parameter and the FRC were measured and evaluated. This showed ORP to likely be the most useful parameter for determining the FRC, due to having less dependency on the source. The turbidity and dissolved oxygen parameters showed the least correlation with FRC, which suggested they would be the least useful parameters for determining the FRC.

A method that could provide a continuous quantitative measure of FRC could not be

found. None of the methods explored achieved a correlation score ( $R^2$ ) above the threshold. Using LDA on a 2-class split came the closest, with a score of 0.703, which improved to 0.733 when using a site-specific test dataset rather than a randomised dataset. The proximity to the threshold, as well as the volatility in the score when the test set was changed, may suggest that this outcome is inconclusive. It would be useful to be able to conduct the analysis on a larger dataset, to be more sure of the outcome, especially including more datasets that can be used as test sets.

We were able to find a method that could be viable to use in classifying water samples into FRC ranges. QDA achieved a classification score for the 2-class split of 0.891, which means that 89.1% of points were classified correctly. This is well above the desired threshold, and strongly supports using QDA in further analysis for classifying water samples, to eventually have a practical application. QDA achieved a classification score of 0.777 for the 3-class split, which is below the desired threshold but does come close. This suggests that the method may also have application for this classification if analysed further, but is less likely.

All the analysis methods indicated that ORP was the most important parameter for correlating to FRC, which was expected from the data collection results. Turbidity and dissolved oxygen were shown not to be useful when included in the analysis, as they had negative impacts on the scores.

Each method analysed gave improvements in scores and held different advantages. This indicates that it could be useful to look at more methods, to see if the output correlations and classifications could be further improved. QDA was the only method that could have non-linear decision boundaries. This suggests that further methods explored should incorporate non-linearities, such as auto-encoders (similar to PCA, but mappings can be non-linear), or deep neural networks (large networks that can find very complex decision boundaries). These methods could also incorporate non-linear uses of turbidity or dissolved oxygen, that couldn't be explored in our scope, as training these methods requires many more datapoints.

Further data analysis techniques will certainly require a much larger dataset, which would be on the order of tens of thousands of datapoints or more. For this, it would be interesting to look into a method of collecting data on a much larger scale. One option could be to design an in-line sensor to take all the measurements, i.e. an in-line FRC sensor with a conductivity probe, pH probe and ORP probe that collects data over time on a particular chlorinated water source. Only a few units would be needed in order to build a much larger dataset, that could then be used for a much wider range of analysis techniques.

In conclusion, we have shown that there is scope for using the measurements of other water parameters to be able to classify water samples into having enough chlorine or not, but that the complexity of the problem is too high for providing a continuous quantitative measure of FRC. With a much larger dataset, and exploring non-linear data analysis techniques, there is possibility for having finer-grain classification. However, this outcome is highly dependent on reliable measurements of ORP, as the data analysis was highly dependent on this measurement. If it becomes too difficult to get reliable measurement of ORP in the field, the outcome of this project suggests it would not be possible to use other measurements of other water parameters to determine FRC.

## 4.3 Risk Assessment

The risks in this project were primarily in the lab, handling concentrated chlorine solutions and quantities of unknown water samples taken from natural sources. The process of always wearing gloves, safety goggles, and a labcoat when handling the chemicals proved very effective at eliminating risk of contact, as well as only handling the concentrated solutions in a fume cupboard.

# References

- [1] Satinder Ahuja. *Chapter 1 - Overview: modern water purity and quality*. Ed. by Satinder Ahuja. Second Edition. Amsterdam: Academic Press, 2021, pp. 1–11. ISBN: 978-0-12-821057-4. DOI: <https://doi.org/10.1016/B978-0-12-821057-4.00014-8>. URL: <https://www.sciencedirect.com/science/article/pii/B9780128210574000148>.
- [2] Becca Bartleson. *What Is a Conductivity Meter?* Mar. 2018. URL: <https://sciencing.com/conductivity-meter-5134852.html> (visited on 15/05/2022).
- [3] Lars Buitinck et al. “API design for machine learning software: experiences from the scikit-learn project”. In: *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*. 2013, pp. 108–122.
- [4] Emerson Process Management. *Fundamentals of ORP Measurement*. May 2008. URL: <https://www.emerson.com/documents/automation/application-data-sheet-fundamentals-of-orp-measurement-rosemount-en-68438.pdf> (visited on 15/05/2022).
- [5] Environment and Natural Resources. *Dissolved Oxygen (DO)*. URL: <http://www.h2ou.com/h2wtrqual.htm#Oxygen> (visited on 17/05/2022).
- [6] *How to remove chemicals from water — Waterlogic*. URL: <https://www.waterlogic.com/en-gb/resources/whats-in-my-tap-water/how-to-remove-chemicals-from-water/> (visited on 07/01/2022).
- [7] Stefan Hrouda-Rasmussen. *Intro to QDA with theory and Python implementation — Towards Data Science*. URL: <https://towardsdatascience.com/quadratic-discriminant-analysis-ae55d8a8148a> (visited on 27/05/2022).
- [8] PCE Instruments. *Dissolved Oxygen Meter PCE-PHD 1 — PCE Instruments*. URL: [https://www.pce-instruments.com/english/measuring-instruments/test-meters/oxygen-meter-pce-instruments-dissolved-oxygen-meter-pce-phd-1-det\\_5886654.htm?\\_list=kat&\\_listpos=3](https://www.pce-instruments.com/english/measuring-instruments/test-meters/oxygen-meter-pce-instruments-dissolved-oxygen-meter-pce-phd-1-det_5886654.htm?_list=kat&_listpos=3) (visited on 13/10/2021).
- [9] Sassi J et al. “Experiments with ultraviolet light, ultrasound and ozone technologies for onboard ballast water treatment”. In: *VTT Research Notes* 2313 (Jan. 2006), pp. 1–86.
- [10] Jiye Jin et al. “A Miniaturized FIA System for the Determination of Residual Chlorine in Environmental Water Samples”. In: *Analytical Sciences* 20.1 (2004), pp. 205–207. ISSN: 0910-6340. DOI: [10.2116/ANALSCI.20.205](https://doi.org/10.2116/ANALSCI.20.205).
- [11] Sher Bahadar Khan and Kalsoom Akhtar. *Photocatalysts - Applications and Attributes*. IntechOpen, Mar. 2019, p. 156. ISBN: 978-1-78985-476-3. DOI: [10.5772/INTECHOPEN.75848](https://doi.org/10.5772/INTECHOPEN.75848). URL: <https://www.intechopen.com/books/7478>.

- [12] A Khayyat. “Study of point of use treatment methods for the disinfection of drinking water in Nepal”. In: 2000, pp. 36–40.
- [13] Peng Li. “Concise review on residual chlorine measurement: Interferences and possible solutions”. In: *Journal of Cleaner Production* 323 (Nov. 2021), p. 129119. ISSN: 0959-6526. DOI: [10.1016/J.JCLEPRO.2021.129119](https://doi.org/10.1016/J.JCLEPRO.2021.129119).
- [14] Peng Li, Takeshi Furuta and Takuya Kobayashi. “Micro-particles as interfering substances in colorimetric residual chlorine measurement”. In: *Ecotoxicology and Environmental Safety* 207 (Jan. 2021), p. 111279. ISSN: 10902414. DOI: [10.1016/J.ECOENV.2020.111279](https://doi.org/10.1016/j.ecoenv.2020.111279). URL: <https://doi.org/10.1016/j.ecoenv.2020.111279>.
- [15] Dr. Vadim B. Malkov et al. “Comparison of On-line Chlorine Analysis Methods and Instrumentation Built on Amperometric and Colorimetric Technologies”. In: 2010, pp. 2–12.
- [16] Nicholas G. Pizzi. “Chapter 1 - Water Treatment Processes”. In: *Water treatment*. American Water Works Association, 2010, pp. 1–6.
- [17] ProAutomation. *PoolSense — PoolSchool - Difference between ORP and Chlorine/ppm*. URL: <https://www.proautomation.co/difference-between-orp-and-chlorine-ppm/> (visited on 15/05/2022).
- [18] Sebastian Raschka. *Linear Discriminant Analysis*. Aug. 2014. URL: [sebastianraschka.com/Articles/2014\\_python\\_lda.html](http://sebastianraschka.com/Articles/2014_python_lda.html) (visited on 19/05/2022).
- [19] Sebastian Raschka. *Principal Component Analysis*. URL: [sebastianraschka.com/Articles/2015\\_pca\\_in\\_3\\_steps.html#pca-vs-lda](http://sebastianraschka.com/Articles/2015_pca_in_3_steps.html#pca-vs-lda) (visited on 27/05/2022).
- [20] G. J. Schieferstein et al. “Carcinogenicity study of 3,3dimethylbenzidine dihydrochloride in BALBc mice”. In: *Food and Chemical Toxicology* 27.12 (Jan. 1989), pp. 801–806. ISSN: 0278-6915. DOI: [10.1016/0278-6915\(89\)90111-7](https://doi.org/10.1016/0278-6915(89)90111-7).
- [21] National Academy of Sciences. “Chapter 2 - The Disinfection of Drinking Water”. In: *Drinking water and health*. Vol. 2. National Academy Press, 1980, pp. 5–41.
- [22] Sensorex. *Free Chlorine (FCL)*. URL: <https://sensorex.com/free-chlorine-fci/> (visited on 20/05/2022).
- [23] Ian Seymour et al. “Electrochemical Detection of Free-Chlorine in Water Samples Facilitated by In-Situ pH Control Using Interdigitated Microelectrodes”. In: (Feb. 2020), pp. 1–11. DOI: [10.26434/CHEMRXIV.11898126.V1](https://doi.org/10.26434/CHEMRXIV.11898126.V1). URL: <https://chemrxiv.org/engage/chemrxiv/article-details/60c74857567dfef14dec49cf>.
- [24] United Nations. *Scarcity — UN-Water*. URL: <https://www.unwater.org/water-facts/scarcity/> (visited on 12/05/2022).
- [25] US Geological Survey. *Turbidity and Water — U.S. Geological Survey*. 2018. URL: <https://www.usgs.gov/special-topics/water-science-school/science/turbidity-and-water> (visited on 15/05/2022).
- [26] Robert Euan Wilson, Ivan Stoianov and Danny O’Hare. “Continuous chlorine detection in drinking water and a review of new detection methods”. In: *Johnson Matthey Technology Review* 63.2 (Apr. 2019), pp. 103–118. ISSN: 20565135. DOI: [10.1595/205651318X15367593796080](https://doi.org/10.1595/205651318X15367593796080).
- [27] World Health Organization. *Rapid Assessment of Drinking Water Quality. Country Report Nigeria*. Ed. by World Health Organization. 2012, pp. 68–75. ISBN: 789241504683. URL: <https://www.who.int/publications/i/item/789241504683>.