Report on Selected Water Data Parameters for the Region of Cambridgeshire and Bedfordshire

Course: Introduction to Data Science with Python for Engineers and Researchers

by

Fabiane Fantinelli Franco (2120215F)

Infrastructure and Environment Research Division

James Watt School of Engineering

University of Glasgow

July 2022

1. Introduction

Water quality monitoring (pH, inorganic nitrogen, chloride, dissolved oxygen, etc) is a multiparameter approach that requires intensive sample preparation and analysis. By understanding trends in analyte concentration, resources can be better allocated. The parameters are usually correlated by non-linear and complex relationships, making it difficult to understand using only statistical analysis. Machine learning can help to develop a forecasting model to improve the accuracy and precision of water quality monitoring in river water and to detect clusters that are not readily identifiable. Artificial neural network (ANN) is usually employed to train the data for analyte concentration prediction over periods of time (e.g., daily, weekly, or monthly depending on the frequency of measurement) [1]–[3].

2. Description of Data

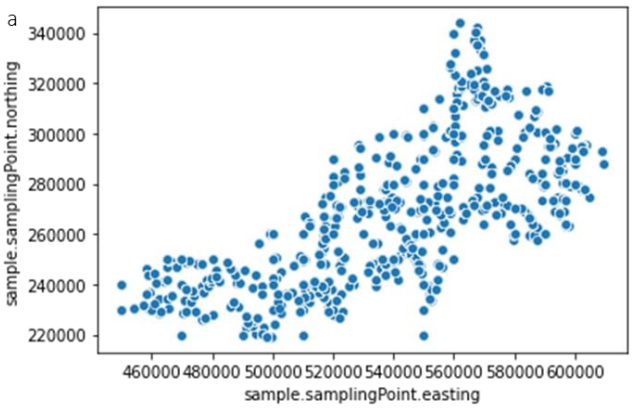
The data was obtained from the open water quality archive datasets (https://environment.data.gov.uk/water-quality/view/download#). It consists of datapoints from various water quality parameters measured in different regions in England. The region chosen was the Cambridgeshire and Bedfordshire for monitoring purposes only in 2016. It consists of 61905 rows x 17 columns in .csv format, in a long format – the parameters are stacked in one single column. The original data includes unique IDs to each row, IDs and coordinates for the sampling region, the type of water body the measure was taken, the name of the parameters (e.g., ammonia, pH, temperature, etc), the date-time of measurement, the measured value of the parameter and the unit it was measured in (Table 1).

Table 1. Original data frame.

@id	sample.samplingPoint	sample.samplingPoint.no	sample.samplingPoint.labe	sample.sampleDateTime	determinand.labe	determinand.de	determinar	resultC	result	codedR	determin	sample.sa	sample.is	sample.pi	sample.sa	sample.sam
http://en	http://environment.di	AN-01M02	SYRESHAM STR.TRIB.OUSE./	2016-01-13T11:13:00	Orthophospht	Orthophosphate	180		0.072		mg/l	RIVER / RU	FALSE	ENVIRONI	463800	241400
http://en	http://environment.di	AN-01M02	SYRESHAM STR.TRIB.OUSE./	2016-01-13T11:13:00	Oxygen Diss	Oxygen, Dissolve	9924		10.7		mg/l	RIVER / RU	FALSE	ENVIRONI	463800	241400
http://en	http://environment.di	AN-01M02	SYRESHAM STR.TRIB.OUSE./	2016-01-13T11:13:00	O Diss %sat	Oxygen, Dissolve	9901		84.5		%	RIVER / RU	FALSE	ENVIRONI	463800	241400
http://en	http://environment.di	AN-01M02	SYRESHAM STR.TRIB.OUSE./	2016-01-13T11:13:00	N Oxidised	Nitrogen, Total C	116		4.26		mg/l	RIVER / RU	FALSE	ENVIRONI	463800	241400
http://en	http://environment.di	AN-01M02	SYRESHAM STR.TRIB.OUSE./	2016-01-13T11:13:00	Ammonia(N)	Ammoniacal Nitr	111	<	0.03		mg/l	RIVER / RU	FALSE	ENVIRONI	463800	241400
http://en	http://environment.di	AN-01M02	SYRESHAM STR.TRIB.OUSE./	2016-01-13T11:13:00	Cond @ 25C	Conductivity at 2	77		546		us/cm	RIVER / RU	FALSE	ENVIRONI	463800	241400
http://en	http://environment.di	AN-01M02	SYRESHAM STR.TRIB.OUSE./	2016-01-13T11:13:00	Alky pH 4.5	Alkalinity to pH 4	162		160		mg/l	RIVER / RU	FALSE	ENVIRONI	463800	241400
http://en	http://environment.di	AN-01M02	SYRESHAM STR.TRIB.OUSE./	2016-01-13T11:13:00	Temp Water	Temperature of '	76		5.1		cel	RIVER / RU	FALSE	ENVIRONI	463800	241400
http://en	http://environment.di	AN-01M02	SYRESHAM STR.TRIB.OUSE./	2016-01-13T11:13:00	Phosphorus-P	Phosphorus, Tot	348		0.153		mg/l	RIVER / RU	FALSE	ENVIRONI	463800	241400
http://en	http://environment.di	AN-01M02	SYRESHAM STR.TRIB.OUSE./	2016-01-13T11:13:00	Nitrite-N	Nitrite as N	118		0.0142		mg/l	RIVER / RU	FALSE	ENVIRONI	463800	241400
http://en	http://environment.di	AN-01M02	SYRESHAM STR.TRIB.OUSE./	2016-01-13T11:13:00	NH3 un-ion	Ammonia un-ion	119	<	0.00038		mg/l	RIVER / RU	FALSE	ENVIRONI	463800	241400
http://en	http://environment.di	AN-01M02	SYRESHAM STR.TRIB.OUSE./	2016-01-13T11:13:00	pH	pH	61		8.26		phunits	RIVER / RU	FALSE	ENVIRONI	463800	241400
http://en	http://environment.di	AN-01M02	SYRESHAM STR.TRIB.OUSE./	2016-01-13T11:13:00	Nitrate-N	Nitrate as N	117		4.25		mg/l	RIVER / RU	FALSE	ENVIRONI	463800	241400
http://en	http://environment.di	AN-01M02	SYRESHAM STR.TRIB.OUSE./	2016-01-13T11:13:00	Sld Sus@105C	Solids, Suspende	135		46.4		mg/l	RIVER / RU	FALSE	ENVIRONI	463800	241400
http://en	http://environment.di	AN-01M02	SYRESHAM STR.TRIB.OUSE./	2016-02-18T13:51:00	Temp Water	Temperature of '	76		5.4		cel	RIVER / RU	FALSE	ENVIRONI	463800	241400
http://en	http://environment.di	AN-01M02	SYRESHAM STR.TRIB.OUSE./	2016-02-18T13:51:00	Cond @ 25C	Conductivity at 2	77		502		us/cm	RIVER / RU	FALSE	ENVIRONI	463800	241400
http://en	http://environment.di	AN-01M02	SYRESHAM STR.TRIB.OUSE./	2016-02-18T13:51:00	O Diss %sat	Oxygen, Dissolve	9901		87.1		%	RIVER / RU	FALSE	ENVIRONI	463800	241400
http://en	http://environment.di	AN-01M02	SYRESHAM STR.TRIB.OUSE./	2016-02-18T13:51:00	Ammonia(N)	Ammoniacal Nitr	111		0.035		mg/l	RIVER / RU	FALSE	ENVIRONI	463800	241400
http://en	http://environment.di	AN-01M02	SYRESHAM STR.TRIB.OUSE./	2016-02-18T13:51:00	pH	pH	61		7.68		phunits	RIVER / RU	FALSE	ENVIRONI	463800	241400
http://en	http://environment.di	AN-01M02	SYRESHAM STR.TRIB.OUSE./	2016-02-18T13:51:00	Phosphorus-P	Phosphorus, Tot	348		0.17		mg/l	RIVER / RU	FALSE	ENVIRONI	463800	241400
http://en	http://environment.di	AN-01M02	SYRESHAM STR.TRIB.OUSE./	2016-02-18T13:51:00	N Oxidised	Nitrogen, Total C	116		2.73		mg/l	RIVER / RU	FALSE	ENVIRONI	463800	241400
http://en	http://environment.di	AN-01M02	SYRESHAM STR.TRIB.OUSE./	2016-02-18T13:51:00	Nitrate-N	Nitrate as N	117		2.72		mg/l	RIVER / RU	FALSE	ENVIRONI	463800	241400
http://en	http://environment.di	AN-01M02	SYRESHAM STR.TRIB.OUSE./	2016-02-18T13:51:00	Nitrite-N	Nitrite as N	118		0.0088		mg/l	RIVER / RU	FALSE	ENVIRONI	463800	241400

To better explore the data, the dataset was transformed to wide, as each parameter is unique as it has a different measurement unit. The data was reduced to measurements in river as they consisted of most of the measurements, with 2219 datapoints while the other water bodies had less than 100 datapoints. The columns kept were coordinates, date-time, parameters of interest (e.g., NH_4, NH_3, conductivity, NO_3, NO_2, diss. O_2, temperature, and pH), and values measured. A new column for seasons and region was created. The region column was created using the coordinate points and three regions were chosen based on the east coordinate. Figure 1a shows the coordinate map used to select the regions and Figure 1b shows the clean dataset used for data analysis.

a



b

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1406 entries, 0 to 2218
Data columns (total 13 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   sample.samplingPoint.notation  1406 non-null   object
 1   Location_east                 1406 non-null   object
 2   sample.sampleDateTime         1406 non-null   datetime64[ns]
 3   NH4                           1406 non-null   float64
 4   NH3                           1406 non-null   float64
 5   Conducivity                   1406 non-null   float64
 6   NO3                           1406 non-null   float64
 7   NO2                           1406 non-null   float64
 8   Diss.O2                       1406 non-null   float64
 9   Temperature                   1406 non-null   float64
 10  pH                            1406 non-null   float64
 11  Seasons                       1406 non-null   object
 12  Region                        1406 non-null   object
dtypes: datetime64[ns](1), float64(8), object(4)
memory usage: 153.8+ KB
```

Figure 1. a) Coordinate plot of the sampling points location. b) Clean dataset description using the info() function.

3. Methodology

i. Supporting Literature

The reported literature usually employs a form of ANN to predict the parameters combined with de-noising techniques and basic statistical analysis [1]–[3].

Some of the approaches used are:

- Statistical analysis including mean, variance, minimum, maximum, etc.
- Correlation matrix between the parameters.
- Denoising of the data with wavelet de-noising techniques.
- ANN training, validation, and testing using neuro-fuzzy inference systems, multi-layer perceptron neural network, radial basis function neural network, etc.

These approaches rely on ANN to give a prediction of the selected parameter over time.

ii. My Approach

For this course, I planned to use only the course material to analyse the dataset and to not utilise ANN. The objective was to find parameters that could either be clustered or correlated to other parameters and to understand more of each of the techniques taught in the coursework. For this, I did the following steps:

- Clean the dataset and transform it to wide so each parameter is a column.
- Add a season and location data to see if there are patterns in the data.
- Check if the parameters follow a gaussian curve and how many outliers are present by plotting the histogram and the boxplot, and by analysing the data statistical description.
- Clean some of the outliers while keeping the original data for further analysis.
- Perform a MANOVA test on each parameter based on season and region. Select parameters of interest.
- Plot a pairplot on the data with the outliers and the clean data. Compare the correlation matrix to find pair of parameters that are correlated. Colour the pairplot by region and location to visually find clusters.
- Perform a principal component analysis (PCA) with elbow-test using different scaling methods to decrease the dimensionality.
- Cluster the data by season.

4. Results and Discussion

4.1. Statistical Analysis

Firstly, the histogram and the boxplot of the original data was analysed (Figure 2). As there are quite a few outliers, the data was cleaned accordingly while still maintaining some of the outliers (Figure 3). It can be observed that many parameters are gaussian, although some of them are left skewed, such as NH_4 and NH_3. Then, a Tukey test was performed to see if the parameters correlated either with season or region. For example, diss. O_2 correlated to seasons (Table 2), as diss. O_2 has an inverse correlation to temperature.

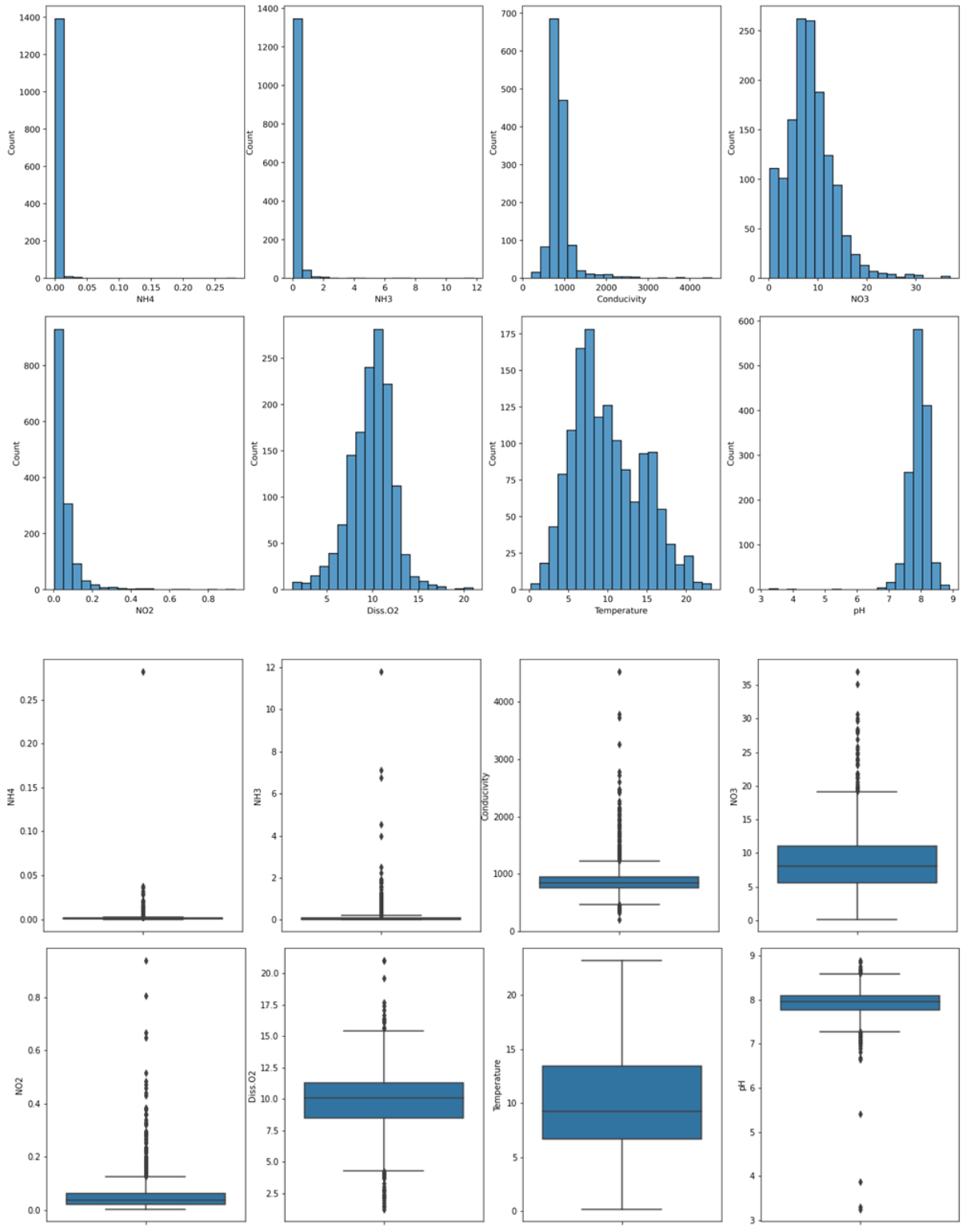


Figure 2. Histogram and boxplot for parameters of original data with outliers.

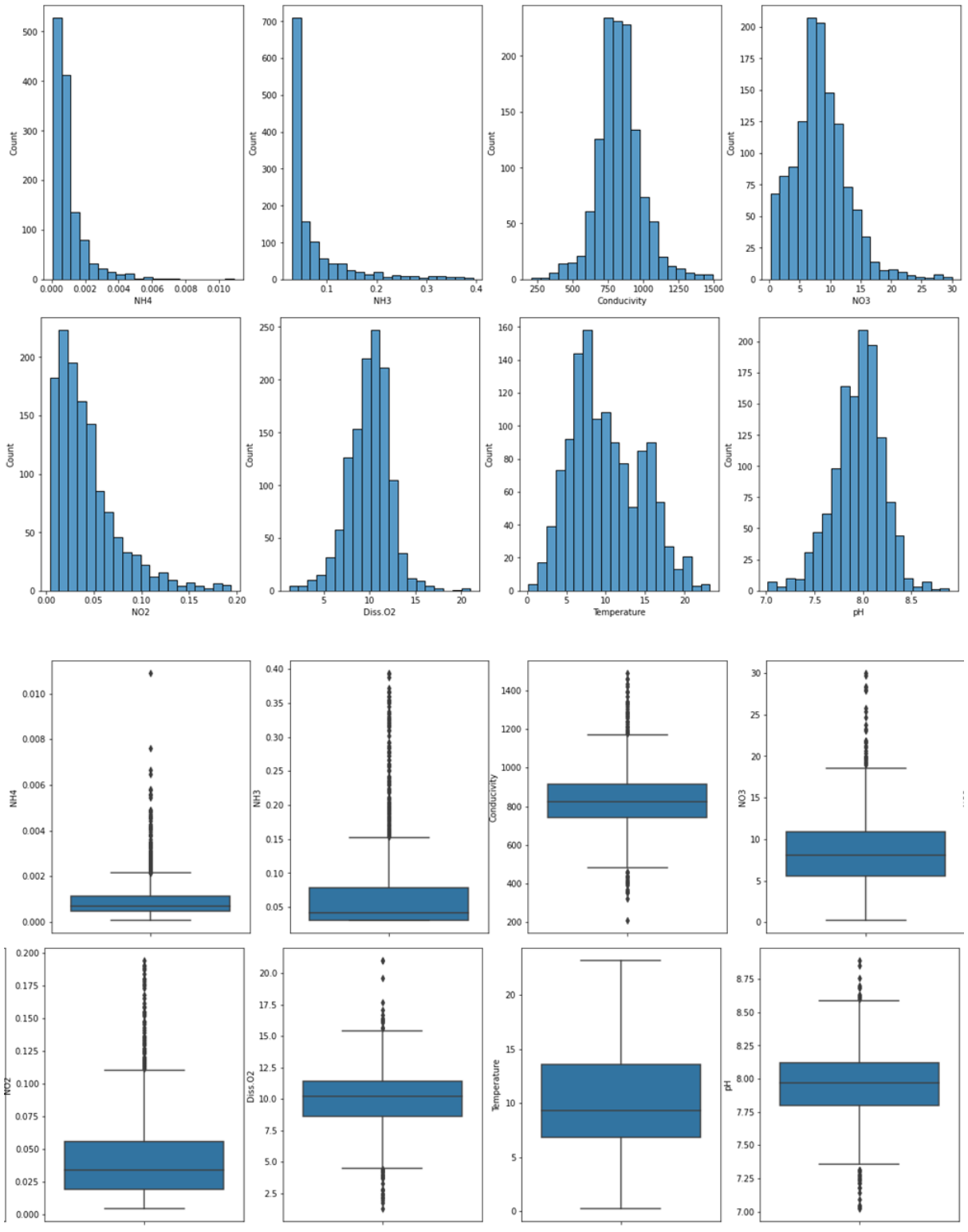


Figure 3. Histogram and boxplot for parameters of cleaned data. Some outliers

Table 2. Example of Tukey test perform on diss. O_2 based on seasonality.

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Autumn	Spring	2.0968	0.001	1.7046	2.489	True
Autumn	Summer	-0.666	0.001	-1.0984	-0.2335	True
Autumn	Winter	1.7425	0.001	1.3623	2.1227	True
Spring	Summer	-2.7627	0.001	-3.2188	-2.3067	True
Spring	Winter	-0.3542	0.1134	-0.7611	0.0527	False
Summer	Winter	2.4085	0.001	1.9627	2.8543	True

A pairplot was then performed on the original data (Figure 4) and the clean data (Figure 5). It was easier to visually identify correlated parameters on the scaled clean data (Figure 5), with NH_3 and NH4 linearly correlating and diss. O_2 and temperature inversely correlating. However, most of the parameters do not seem to correlate with each other. These parameters were chosen as NH_3/NH_4 concentrations are known in literature to correlate with temperature and pH. However, this is not obvious in this dataset and further analysis is needed to understand it. From the correlation matrix (Table 3) diss. O_2 seems to correlate with temperature. Seasons (Figure 5) seem to form clusters while regions do not (Figure 6). Naturally, season is related to temperature.

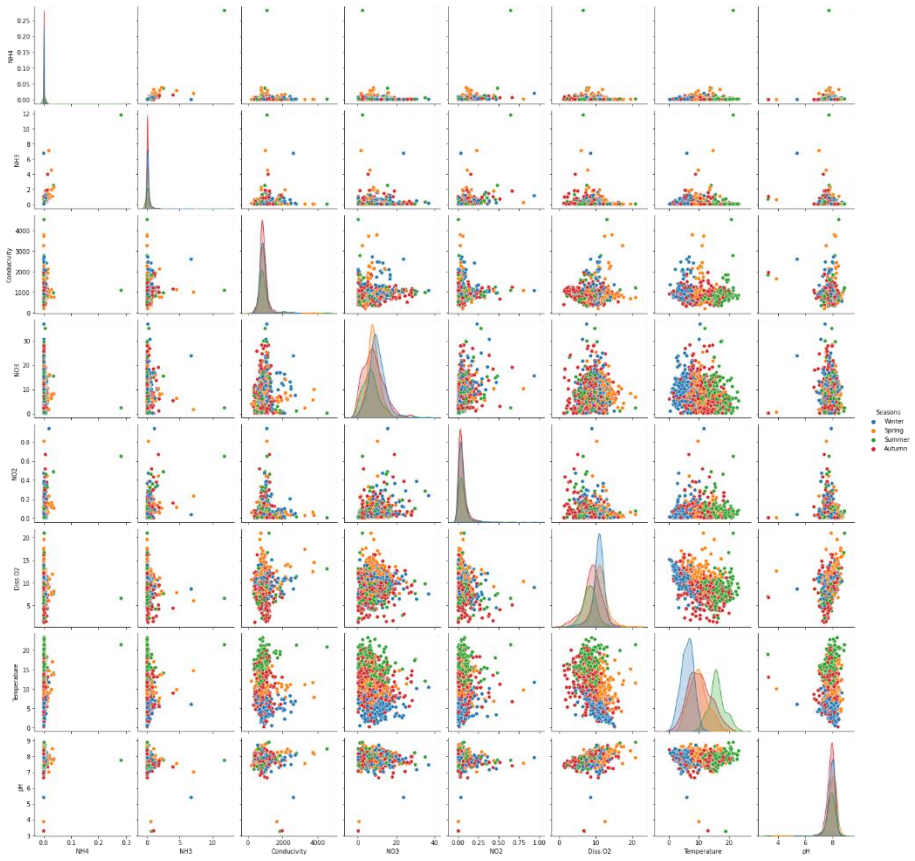


Figure 4. Pairplot using original, not scaled data colouring according to the season.

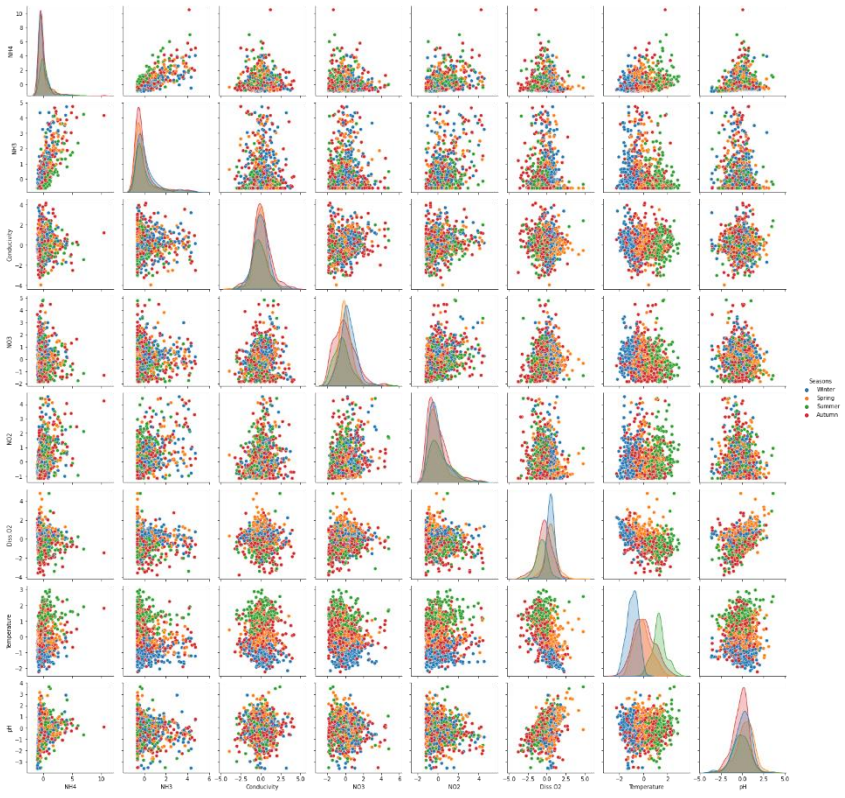
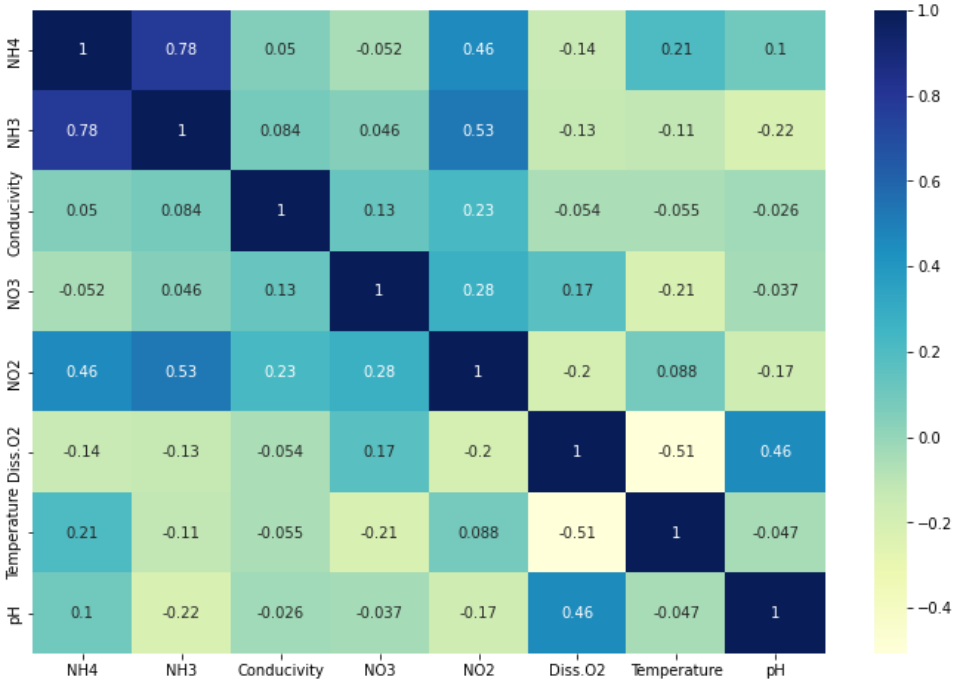


Figure 5. Pairplot using data with less outliers and scaled colouring according to the season.

Table 3. Correlation matrix for the clean data.



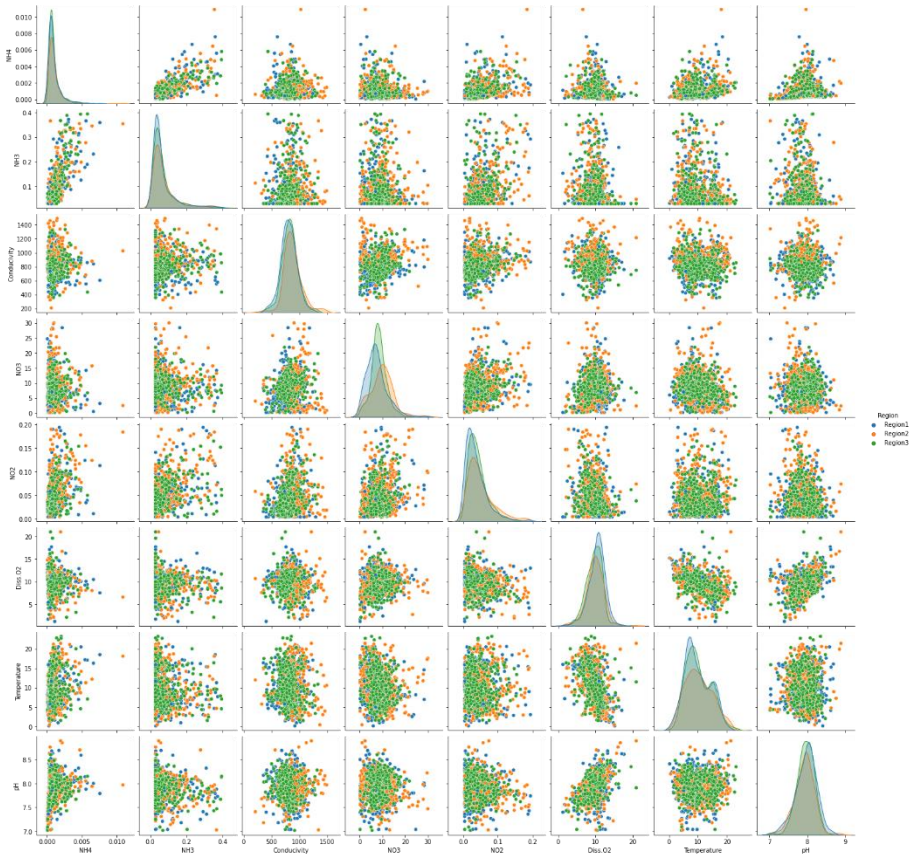


Figure 6. Pairplot for the clean data colouring according to the region.

4.2. Unsupervised Machine Learning

Firstly, PCA on two variables was performed (diss. O_2 and temperature) to understand if the data could be better separated. The PCA seemed to work better for separating summer from winter than spring from autumn (Figure 7).

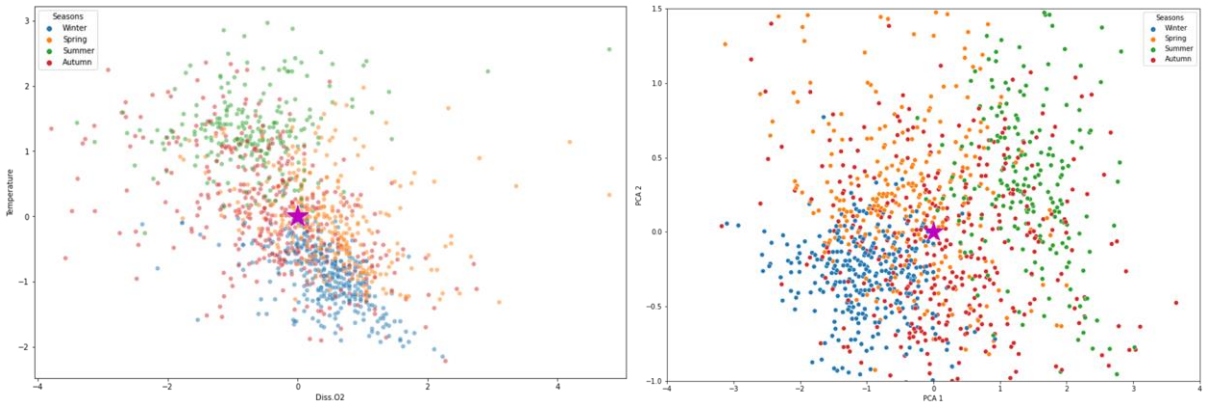


Figure 7. PCA on two variables (diss. O_2 and temperature).

To see if the dimensions could be reduced, an elbow test on both the original data and the clean data was performed. The test was repeated for different scaler options (two examples on Figure 8). The results were different for the different scalers, but most seemed to include up to 4 components. A pairplot of the PCA can be seen in Figure 9, but this data might not be suitable for this test.

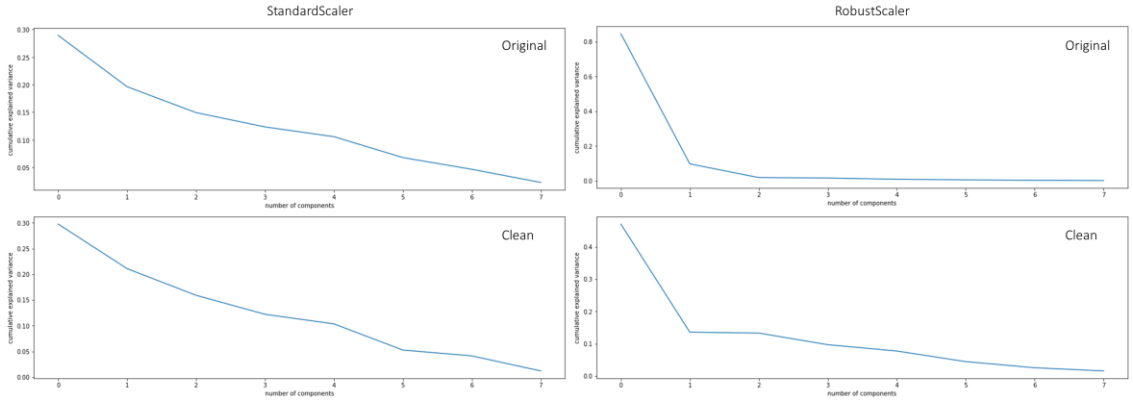


Figure 8. Example of elbow plots for two different scalers using the original and the clean data.

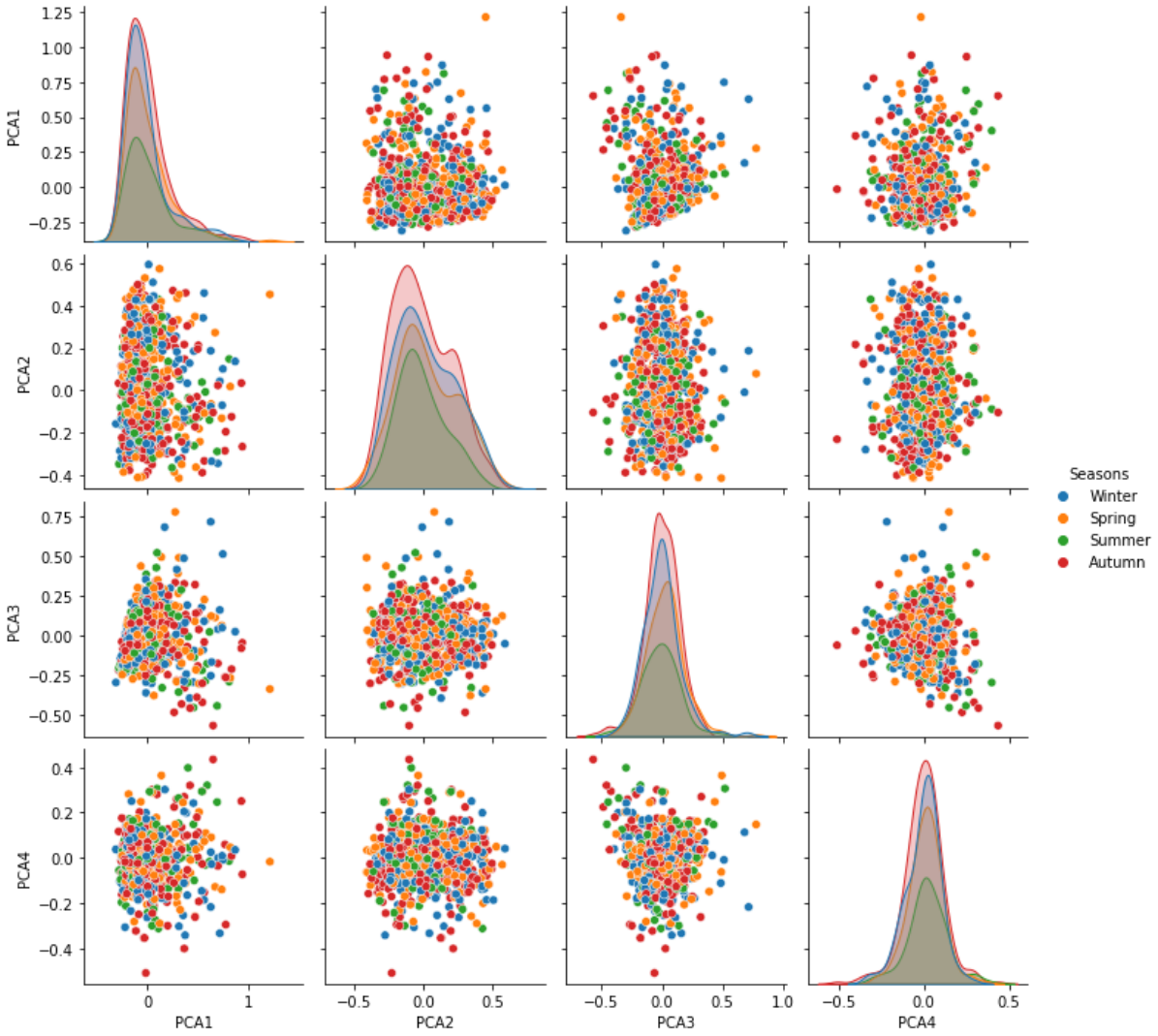


Figure 9. Pairplot for the PCA using 4 components and the MaxAbsScaler.

Clustering of the data was performed using K-means. The clusters were formed based on the seasons (Figure 11). To validate the model, a closer comparison was done using the diss. O_2

vs temperature (Figure 12). The confusion matrix shows that the model can be refined but works best for summer. The confusion matrix is not labelled correctly.

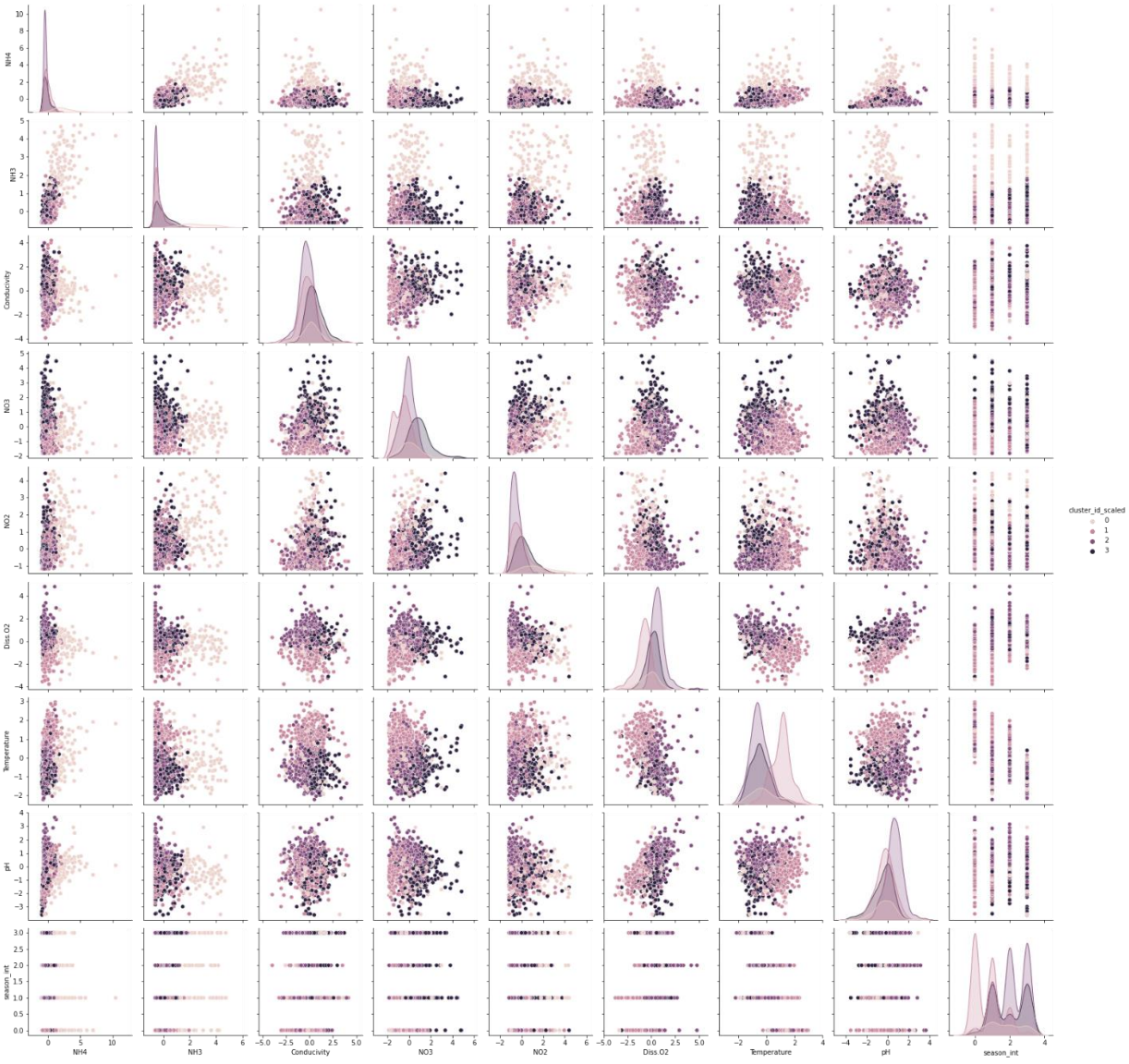


Figure 10. Pairplot of all the parameters clustered according to season.

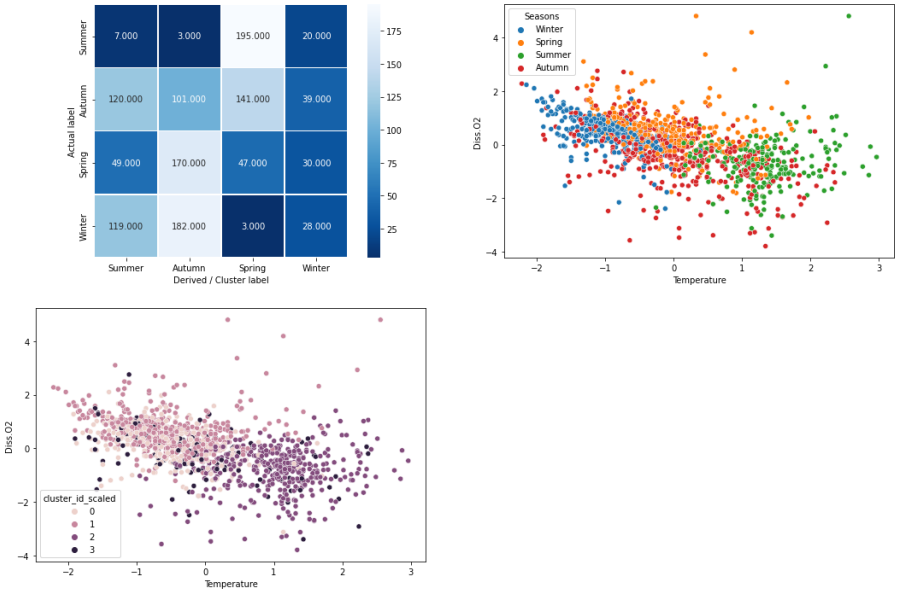


Figure 11. Confusion matrix of the cluster model according to season for diss. O_2 and temperature.

5. Future Perspectives

The modelling was not as successful as expected, but it did produce some interesting results. Some of the future analysis that could be employed is to separate the region into rivers and analyse if there are differences in the data collected. There is also a large range of yearly datasets, which could be combined for a more robust model and to detect if there are yearly changes. There also many other parameters that have not been included in the analysis but could improve the prediction. Furthermore, there are clear regulations for each parameter, so that could be included to investigate if the measured value is in compliance with current regulations. Finally, ANN can be employed to predict complex parameters.

References

[1] A. Najah Ahmed *et al.*, "Machine learning methods for better water quality prediction," *Journal of Hydrology*, vol. 578, 2019, doi: 10.1016/j.jhydrol.2019.124084.

[2] A. H. Haghiabi, A. H. Nasrolahi, and A. Parsaie, “Water quality prediction using machine learning methods,” *Water Quality Research Journal*, vol. 53, no. 1, 2018, doi: 10.2166/wqrj.2018.025.

[3] K. B. Newhart, C. A. Marks, T. Rauch-Williams, T. Y. Cath, and A. S. Hering, “Hybrid statistical-machine learning ammonia forecasting in continuous activated sludge treatment for improved process control,” *Journal of Water Process Engineering*, vol. 37, 2020, doi: 10.1016/j.jwpe.2020.101389.