Internship Report - Module MHSE 26

Institution:

Bundesanstalt für Gewässerkunde (BfG) / Federal Institute of Hydrology
Location: Am Mainzer Tor 1, 56068 Koblenz, Germany
P.O.Box 200253, 56002 Koblenz
Fon: +4926113060
Fax: +4926113065302
E-mail: posteingang@bafg.de
www.bafg.de

Topic/Task of the Internship:

Data analytics, Data visualization, River water quality

Advisors:

Dr. Jens Wyrwa, Dr. Marieke Frassl, Dr. Fabian Große

Intern and author:

Shayan, Kamali. Matriculation Number: 5034911

Supervisors at TU Dresden:

Prof. Dr. Peter Krebs, Geovanni Teran Velasquez

Study Course: MSc-HSE

Date: 31st March 2023

Table of Contents

List of Figures

List of Tables

List of Abbreviations

BfG	Bundesanstalt für Gewässerkunde
CO_2	Carbon dioxide
CPU	Central processing unit
E	East
ETR	External Timer Reference
N	North
O_2	Oxygen
UTM	Universal Transverse Mercator

1 Introduction

Microbial processes have a significant impact on the rivers in inland and coastal regions. The biological, chemical and physical characteristics have an impact on the microorganisms in the sediment, water, and shoreline. The Federal Institute of Hydrology (Bundesanstalt für Gewässerkunde, BfG) as a scientific institution ranking as a higher federal authority, is responsible for the German waterways in federal ownership. Department U2 ''Microbial Ecology'' is engaged in comprehensive investigation and advisory work (The department U2 website, 2020; The Federal Institute of Hydrology, 2020).

Microorganisms including bacteria, phyto- and zooplankton live in the water column of the waterways, and through their metabolism, they have a substantial impact on the oxygen, carbon, pH, and nutrient balance of water bodies and thus on the water quality. Department U2 investigates their populations and impact on inland waters, and estuaries. The investigations regularly take place on the Berlin waterways, the Elbe, the Rhine and Moselle, and the German North Sea estuaries (The department U2 website, 2020).

In this report, the data specifically from the Bunthaus station in the upper Elbe Estuary (Elbe-km 609) is analyzed to visualize and study the temporal evolution of O2, chlorophyll, and pH individually and in relation to each other.

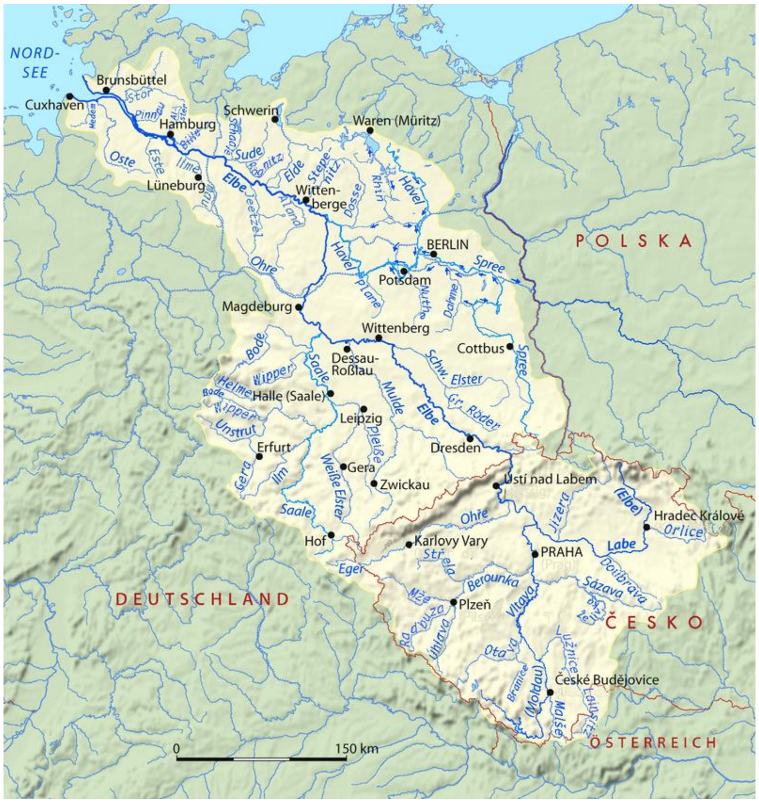


Figure 1: The Elbe River basin (Wikimedia Commons, the free media repository, 2021).

The Elbe has the fourth-largest river basin in Central and Western Europe with a population of around 25 million people. It originates in the Giant Mountains of the northern Czech Republic and flows into the North Sea near Cuxhaven (Elbe-km 727). 1,094 km make up its entire length (International Commission for the Protection of the Elbe River, 2018).

1.1 Objective

One of the initial phases in data analysis is called data exploration, and it involves looking at and visualizing data to find insights right away or point out regions or patterns that need further investigation (TIBCO Software, 2023).

The purpose of this internship was to visualize and understand the water quality data from the Bunthaus station. The structure of the work can be classified into three steps:

- Describe the data (introduction)
    - Describing data properties
- Explore the data (results)
    - Visualization
- Verifying data quality (discussion)
    - Checking data completeness and further investigations

1.2 Data source

- There are several stations along the Elbe where water quality and quantity is monitored. The data we analyzed was gathered at the Bunthaus station placed on the left side of the northern Elbe River.
- Data origin:
    - Up to the year 2021: http://www.portal-tideelbe.de/
    - For the year 2022: Hamburg Institute for Hygiene and the Environment (HU), Water Quality Monitoring Network (WGMN)
- Station name: Bunthaus
- Geo reference: ETRS89/UTM32N
- Location: 53°27'42.1"N 10°03'51.6"E
- River kilometer: 609.8
- Reference gauge: Neu Darchau, Elbe
- Observation start date: 1988
- Water quality: chlorophyll-*a* (from 2014), oxygen, electric conductivity, pH, water temperature
- Water level (starting 1950): high tide, low tide water
- Operator: Hamburg Institute for Hygiene and the Environment
- Station type: multiparameter station
- Water body: Elbe, main section Elbe km 607.5 to 639
- Remarks: Current online data and station information at:
    - http://www.hamburg.de/wasserguetemessnetz/

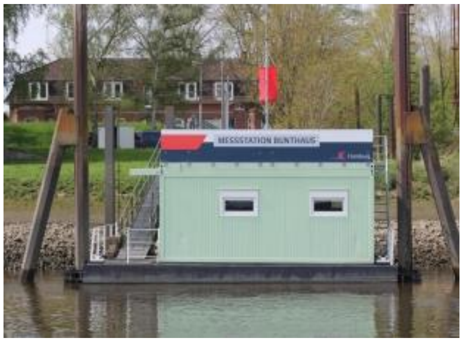


Figure 2: Bunthaus station (Messstation Bunthaus, Elbe, n.d.).

1.3 Data description

No prior knowledge of data existed and data were in raw text format, with often more than 1,600,000 records. With the given measurement interval, there were some missing values in the data set. For chlorophyll-*a* and pH, the data for the year 2016 were in a separate file and inserted into the main data. Our data includes date and time and numeric values. For pH, chlorophyll content, oxygen content, and water temperature, data for 2022 is appended separately. Table 1 shows individual parameter query dates and periods.

Table 1: The water quality parameters query date and period of measurement.

	Query date	Query period
Chlorophyll [µg/l]:	2022-03-08 11:43:10	2014-01-01 01:00:00 to 2021-12-31 23:50:00
Electrical Conductivity	2022-03-07 17:30:12	1988-06-19 01:00:00 to 2021-12-31 23:50:00
pH	2022-03-07 17:30:44	1988-06-19 01:00:00 to 2021-12-31 23:50:00
Oxygen	2022-03-07 17:31:06	1988-06-19 01:00:00 to 2021-12-31 23:50:00
Water temperature	2022-03-07 17:32:27	1988-06-19 01:00:00 to 2021-12-31 23:50:00

2 Material and methods

2.1 Software

Python programming language version 3.11.2 64bit was used as the main software for analyzing the data. As an environment for programming in Python, Jupyter Notebook was chosen. Jupyter Notebook is an interactive document and was created to facilitate reproducible research by lowering several barriers to replication and giving scientists the tools to create easily shared computational narratives that combine code, results, and language (Rule et al., 2018).

The main packages used:

- **Pandas** is a powerful, flexible, and open-source library written for Python, and mainly used for data manipulation and analysis (*Pandas - Python Data Analysis Library*, 2023).
- **NumPy** is the cornerstone Python module for scientific computing. It provides multidimensional array objects, as well as other derivative objects like matrices (NumPy Developers, 2022).
- **Matplotlib** provides a complete tool for building static and animated visualizations (Matplotlib development team, 2023).
- **Seaborn** is a library based on Matplotlib, which offers a sophisticated drawing tool for drawing informative statistical graphics(Waskom, 2021).

2.2 Methods used for data visualization

The initial phase in data analysis is called data exploration, which involves visualizing the data to find insights right away or point out regions or patterns that need further investigation. In this project, we made two main types of graphs:

- Line graph (time series chart) for each water quality parameter, which shows their changes over years.
- Scatter plot, which helps to recognize relationships and dependencies.

3 Results

In this part, line graphs and scatter plots help us to explore the data.

3.1 Line graphs (time series charts)

3.1.1 Oxygen and pH

As Figure 3 and Figure 4 show, the O_2 and pH were low before 1990 compared to the rest of the years. The main reason for this O_2 and pH behavior is the input of high organic loads and contaminants before the German reunification, which caused O2 consumption and a decrease in pH. After the reunification, the yearly average of the pH increased and seasonally fluctuated around the approximate average value of 8.

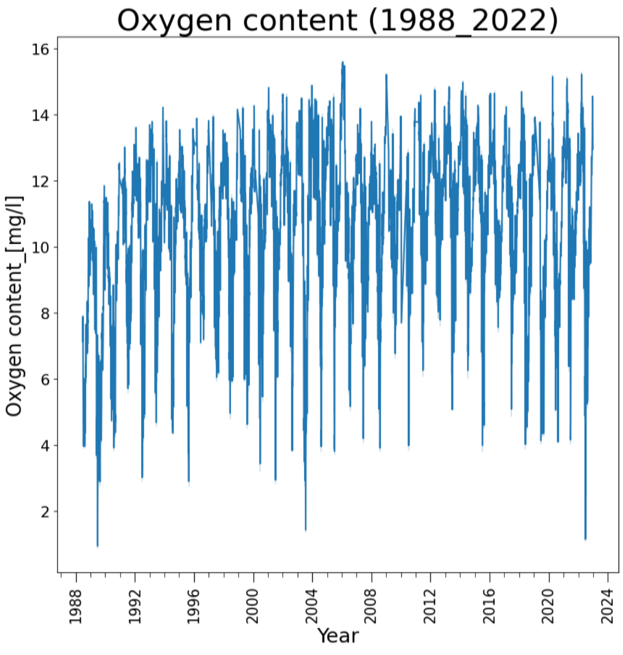


Figure 3: Oxygen content (mg/l) measured at Bunthaus station from 1988 to 2022.

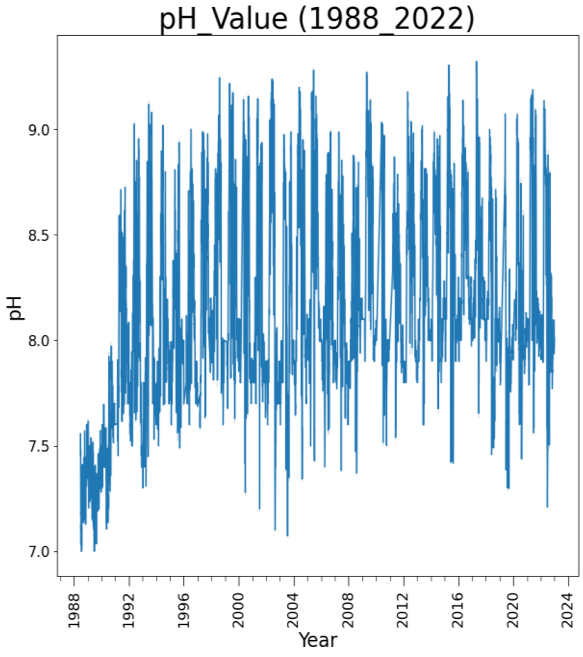


Figure 4: pH values measured at Bunthaus station from 1988 to 2022

In Figure 3, three years with the lowest oxygen content stand out. The first trough is in 1989 which reason explained in the previous paragraph. The second is in 2003 during a drought period. Water flows tend to decrease and there is less mixing when there is a drought. As a result, the water system's dissolved oxygen level may decrease (BUNCH, 2018). And the third trough is in the year 2022, in which we assume the extreme growth of zooplankton caused a lack of oxygen. data recorded at Geesthacht weir (Elbe-km 585) in concert with mechanistic modeling (both not shown) suggest that favorable growth conditions in the river basin resulted in stronger than usual zooplankton growth driving the production of readily biodegradable organic matter, which in turn caused high O2 consumption.

3.1.2 Chlorophyll and Oxygen

In lakes and rivers, microscopic plants called algae create chlorophyll, which gives plants their green color. Since it is difficult for algae to grow in the seasons without sufficient light, the amount of chlorophyll in water is often the highest in the summer and lowest in the winter, and it has a cyclic behavior (Figure 5).

The information above can explain Figure 6, which is the time series show the year 2022 as an example for a (more or less) typical seasonal cycle. The fast increase begins in March when there is enough light and warm temperature to allow the chlorophyll to grow, until the beginning of

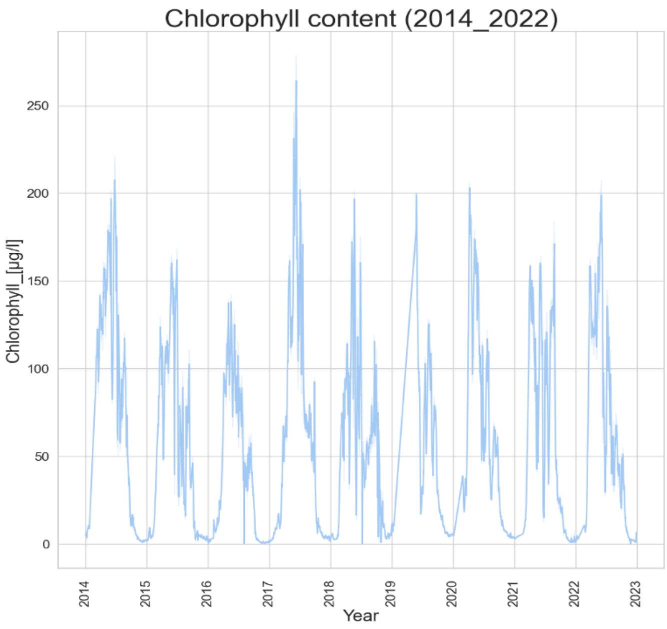


Figure 7: Chlorophyll content (µg/l) measured at Bunthaus station from 2014 to 2022

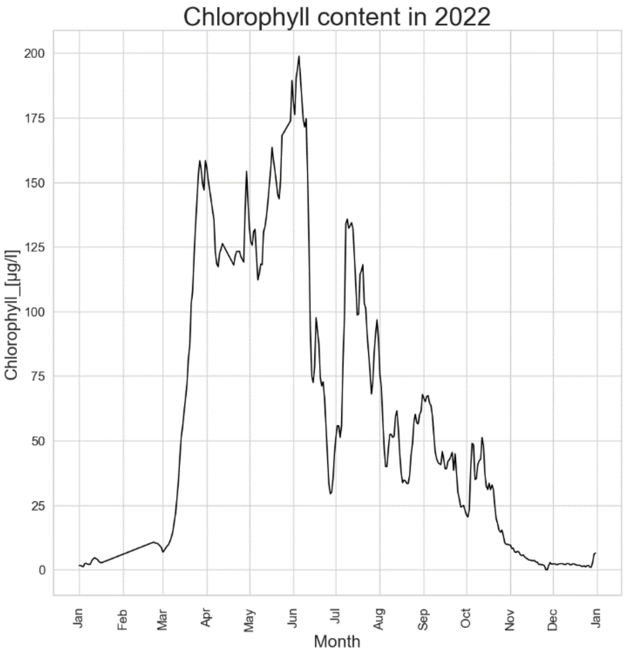


Figure 6: Chlorophyll content (µg/l) measured at Bunthaus station in the year 2022

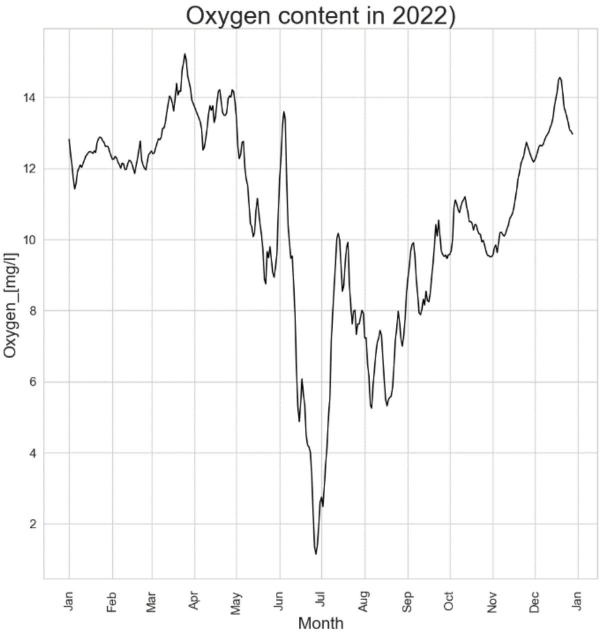


Figure 5: Oxygen content (mg/l) measured at Bunthaus station in the year 2022

April, when the water nutrients are consumed to a degree that makes the chlorophyll content reach a plateau. This bump in April could also be due to a "bad weather" period (less light or higher discharge). And from the beginning of May, it increases again till it peaks in June. From June to July, there is a significant decrease caused by a lack of oxygen (Figure 7) due to zooplankton growth, which feed on phytoplankton, and oxygen consumption via heterotrophic bacterial respiration (as a result of the production and further decay of biodegradable organic matter from zooplankton fecal pellets). when zooplankton numbers drop, the phytoplankton can start growing again, and accordingly the chlorophyll contents in mid-July increase. Finally, it generally decreases from August as the weather gets darker and colder.

3.1.3 Electrical conductivity and water temperature

Water temperate seasonally fluctuates around 13°C (Figure 8). The minimum temperature is -2.2 recorded on 13th February 1994, and it might be due to measurement or human error. However, according to (Christopher, 2013), water can continue to be liquid below zero degrees Celsius, which can have several reasons such as salinity or other additives. Thus, the measurement could be correct.

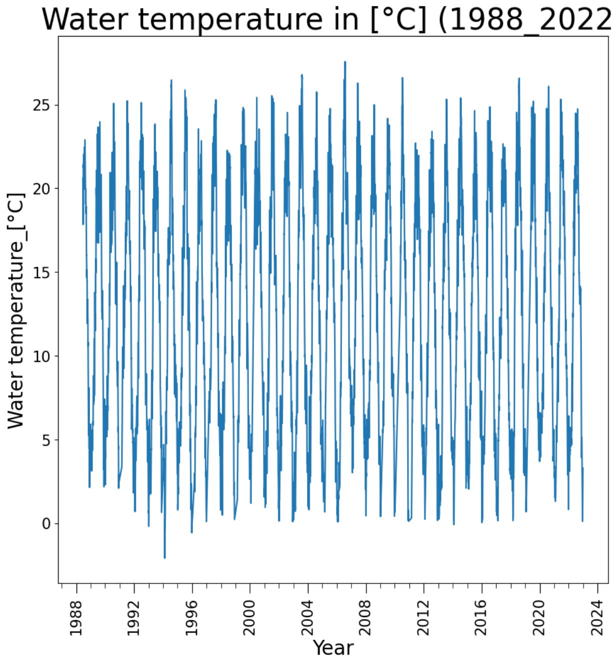


Figure 9: Water temperature (°C) measured at Bunthaus station from 1988 to 2022

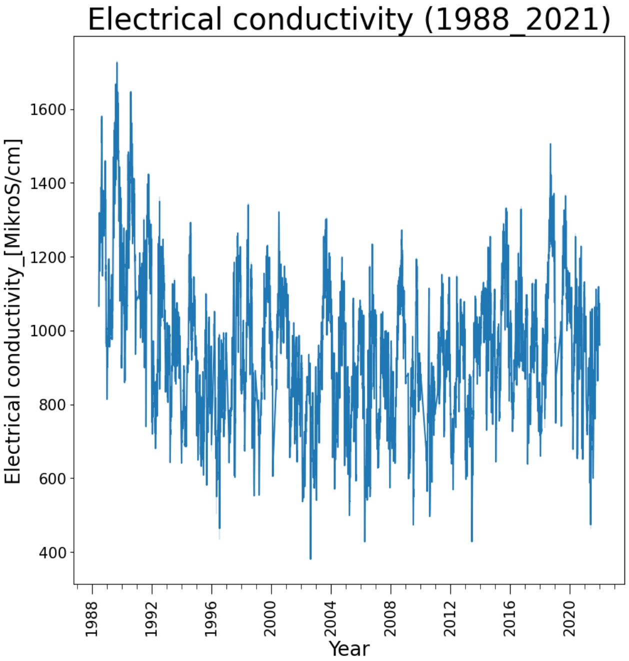


Figure 8: Electrical conductivity measured at Bunthaus station from 1988 to 2021

The ability of water to conduct an electrical current is measured by its conductivity. Conductivity rises with salinity because dissolved salts and other inorganic compounds carry electrical currents. Temperature also has an impact on conductivity; the warmer the water, the higher the conductivity (US EPA, 2013). Based on the mentioned knowledge, we can interpret that the high electrical conductivity before 1994 is accompanied by a high concentration of dissolved ions in water (Figure 9). However, the focus of the project is not on the electrical conductivity behavior.

3.2 Scatter plot

For each scatter plot, the data of the two parameters were measured on the same date merged into a single table and then plotted.

3.2.1 Chlorophyll vs. pH:

As we can also see in Figure 10, the pH is directly proportional to chlorophyll content. Also, we can see data are less scattered in recent years.

Figure 11 is a monthly-separated graph that shows the correlation between pH and chlorophyll content depends on the month of the year. A direct relation can be seen more clearly in the summer months (productive period). During the winter months, the chlorophyll content becomes lower and the pH is almost fixed around 8 (non-productive period).

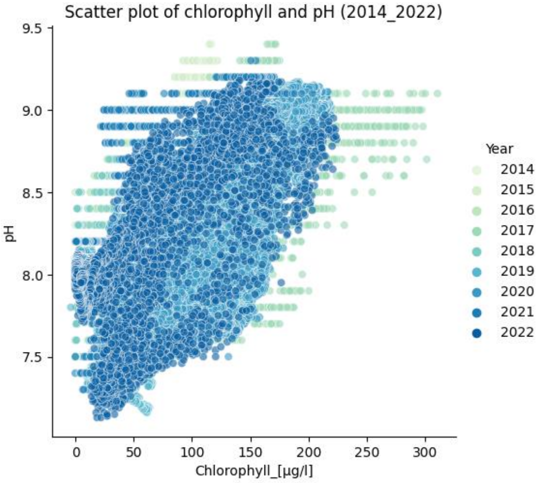


Figure 10: pH values and chlorophyll content (µg/l) scatter plot measured at Bunthaus station from 2014 to 2022

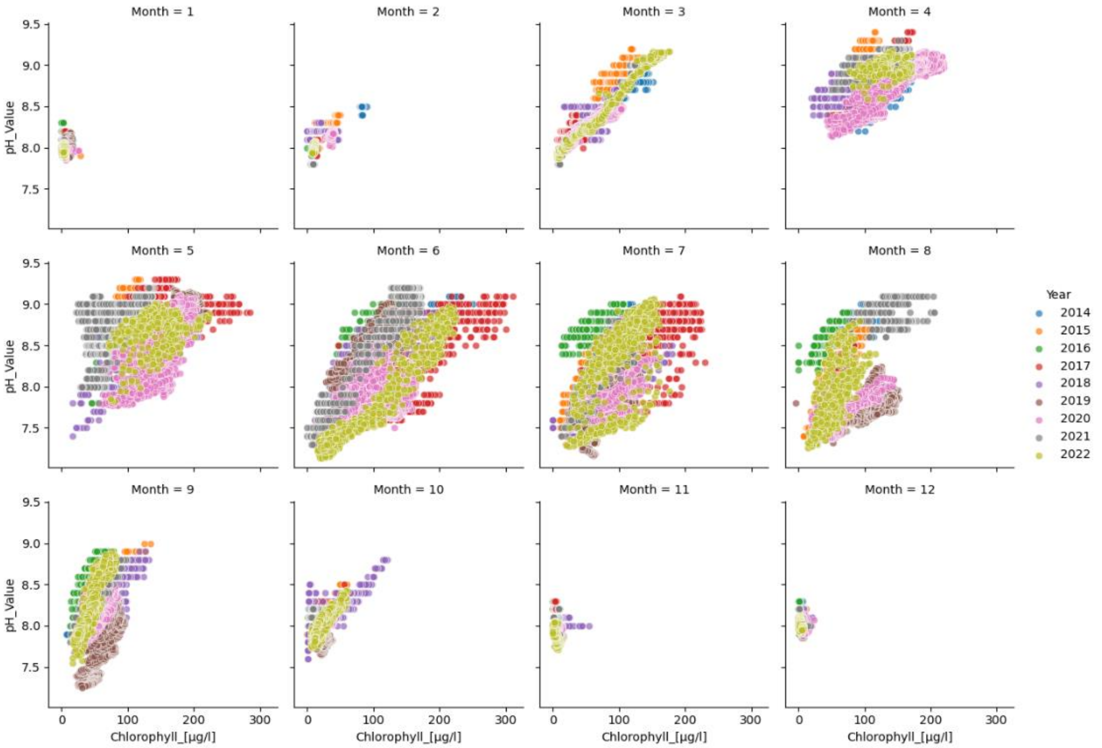
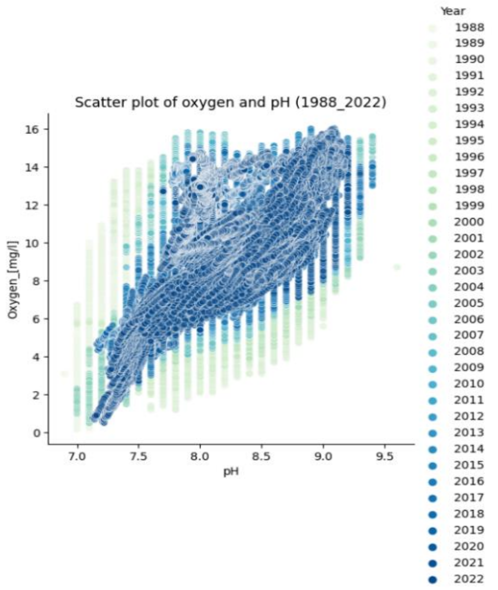


Figure 11: pH values and chlorophyll content (µg/l) relationship in each month of the year measured at Bunthaus station from 2014 to 2022

3.2.2 Oxygen vs. pH:

The data for both pH and O2 existed from 1988 till 2022. However, Figure 12 shows a high scatter in the earlier years. Also, for comparison with the chlorophyll plots, only the data after 2014 were plotted (Figure 13).



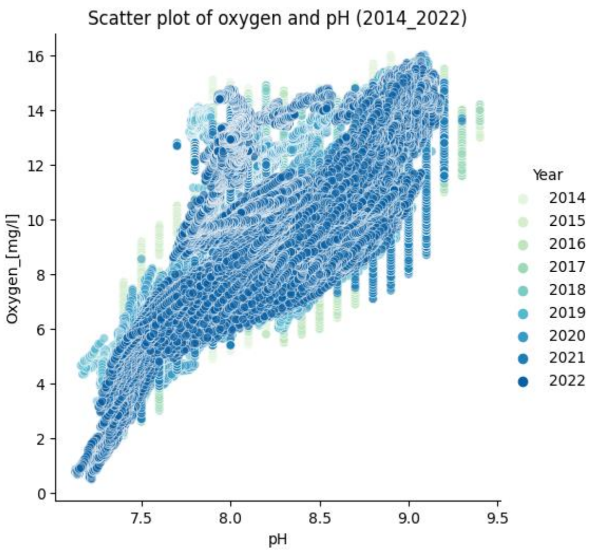


Figure 13: Oxygen content (mg/l) and pH relationship measured at Bunthaus station from 1988 to 2022

Figure 14: Oxygen content (mg/l) and pH relationship measured at Bunthaus station from 2014 to 2022

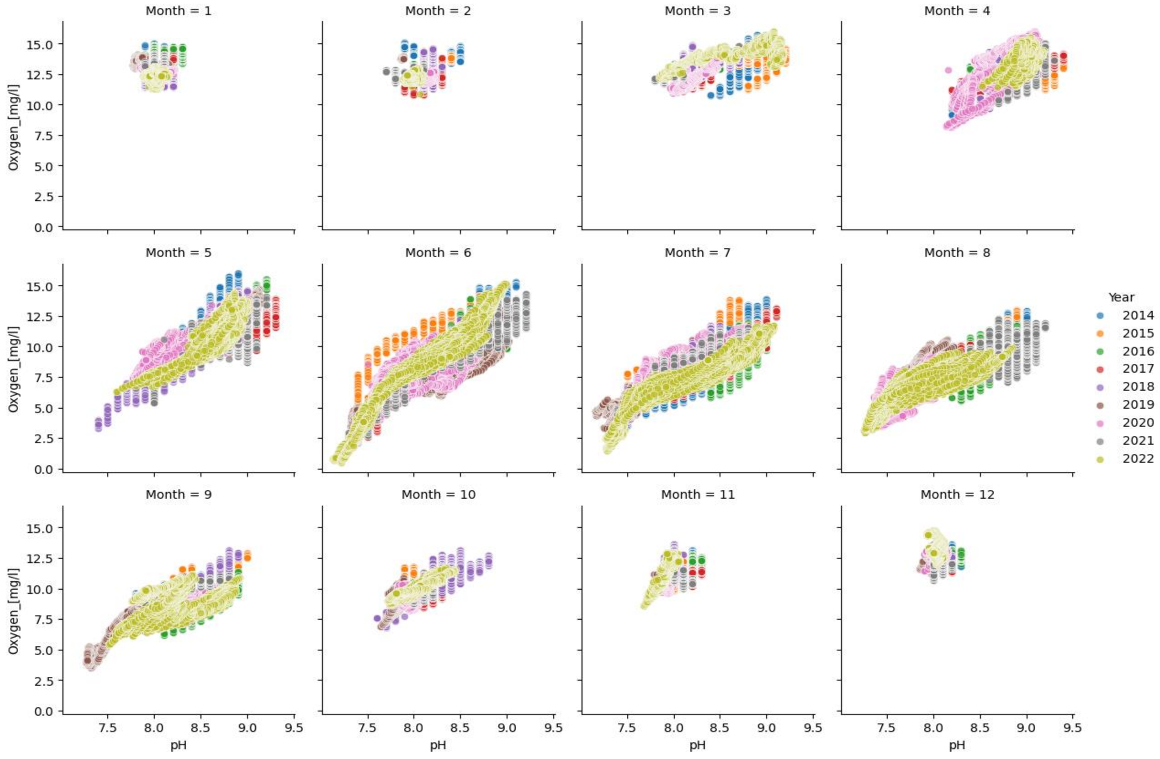


Figure 12: Oxygen content (mg/l) and pH monthly relationship measured at Bunthaus station from 2014 to 2022

As pH and oxygen have huge data spread differences based on the month of the year, the graph in Figure 14 is plotted to show the monthly values, and it is color-coded by year. The above graph clearly states that in spring to fall the range of pH and oxygen grows in an approximately linear shape due to the dependency on primary production for which chlorophyll-*a* is a proxy. And in the colder months of the year, pH and oxygen values are focused around 8 and 12 mg/l respectively.

3.2.3 Oxygen vs. Chlorophyll:

The scatter plot for O2 and chlorophyll was drawn with an extra feature. We assumed that there could be a dependency on chlorophyll and O2 contents on high and low tides. The reasons for our assumption are first stronger riverine influence during the ebb phase (low tide), with water enriched with higher chlorophyll/oxygen/pH due to high production in the river, and second stronger estuarine influence during the flood phase (high tide), with water with lower chlorophyll/oxygen/pH due to lower production (as a result of high turbidity and strong zooplankton) grazing in the deeper downstream parts of the estuary. Therefore, the water level data were merged with the O2 and chlorophyll data into one table. The high and low tides periods were identified and color-coded for the oxygen-chlorophyll scatter plot (Figure 15)

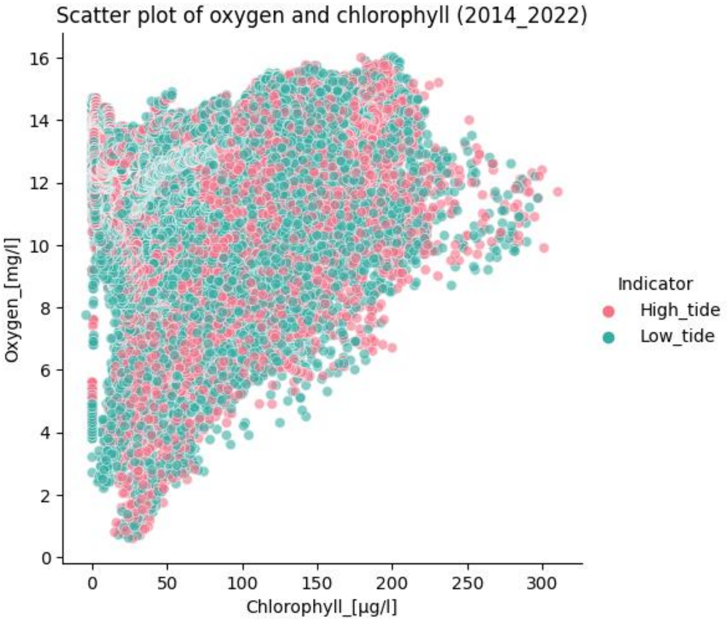


Figure 15: Oxygen (mg/l) and chlorophyll (μg/l) relationship measured at Bunthaus station from 2014 to 2022

In principle, tides are long-period waves that travel through the oceans in reaction to the moon's and sun's gravitational pull. Tides start in the oceans and move toward the coasts, where they manifest as the regular rise and fall of the water's surface. High tide is when the wave's highest point, or crest, reaches a specific area; low tide is when the wave's lowest point, or trough, happens. The tidal range is the height between high tide and low tide (US Department of Commerce National Oceanic and Atmospheric Administration, 2023).

The Elbe estuary is tidally influenced and the major tidal constituent is the so-called principal lunar semi-diurnal (M2) tide, with a period of 12 hours and 25.2 minutes. We consider the time period from when the trough is recorded until the tide peak is recorded as a flood phase (high tide period), and the time from when the peak is recorded till the trough is recorded, as the ebb phase (low tide period).

After visualizing Figure 15, we recognized that the pattern in both low and high tide periods are very similar, and no significant difference can be illustrated. Thus, to reach the maximum difference, we decided to only extract the first O2 and chlorophyll measured values right after the peak time in high tides and trough time in the low tides (there is no O2, and chlorophyll measures at the exact time of tide peaks and troughs).

Figure 16 illustrates the O2 vs. chlorophyll contents in two groups. One group is the measured contents right after the peaks and another group is the measured values right after the troughs of the tidal waves. These two groups are named “High_tide” and “Low_tide” respectively. We can observe a difference between these two groups, especially from March to October.

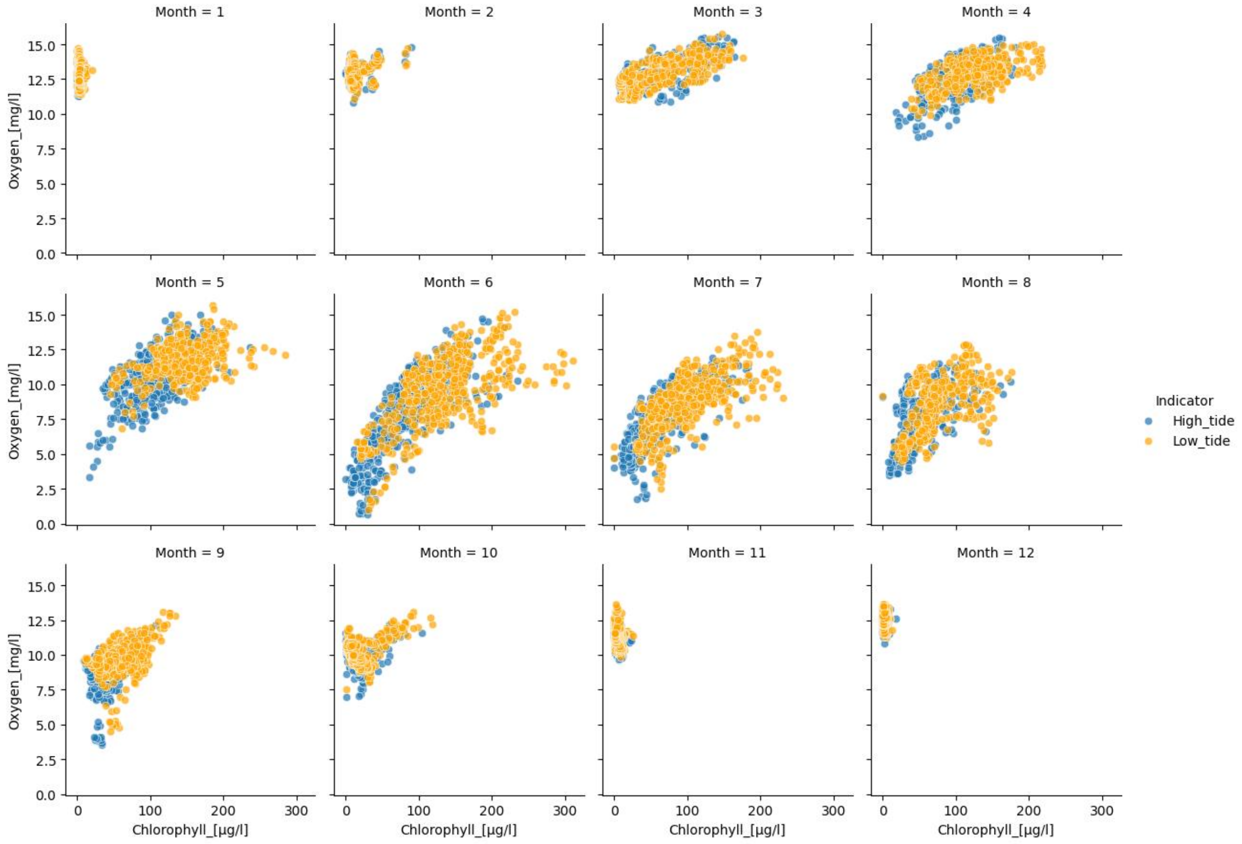
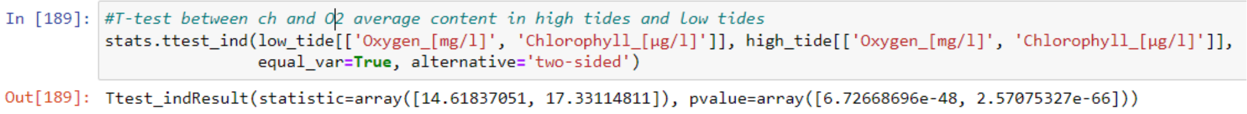


Figure 16: Oxygen (mg/l) and chlorophyll (µg/l) monthly contents measured at Bunthaus station during the peak of high and low tides from 2014 to 2022

To become more certain, a statistical t-test was run between the two groups' values. The null hypothesis is that the difference between the values in high and low tides is zero. Figure 17 shows the t-test result, which rejects the null hypothesis. Therefore, there is a significant statistical difference between the values of the two groups.



```
In [189]: #T-test between ch and O2 average content in high tides and low tides
          stats.ttest_ind(low_tide[['Oxygen_[mg/l]', 'Chlorophyll_[µg/l]']], high_tide[['Oxygen_[mg/l]', 'Chlorophyll_[µg/l]']],
                          equal_var=True, alternative='two-sided')

Out[189]: Ttest_indResult(statistic=array([14.61837051, 17.33114811]), pvalue=array([6.72668696e-48, 2.57075327e-66]))
```

Figure 17: Result of t-test for checking the difference between oxygen (mg/l) and chlorophyll (µg/l) average values during high tides and low tides from 2014 to 2022

In May, June, July, and August, the difference in high tide and low tide oxygen and chlorophyll contents is even more visible. As was assumed, in high tides we have lower chlorophyll and oxygen contents due high turbidity and strong zooplankton. Thus, lower growth of chlorophyll and accordingly oxygen is expected.

4 Reflections

During the 3-week internship, I managed to make informative plots from raw data, but 3 weeks for going deep into the data understanding was not enough, and more time is needed to interpret each and every graph precisely, formulate the hypothesis, and conduct the necessary statistical tests for recognizing any special relationships.

For the scatter plots the data were cleaned, formatted, and necessary data were constructed. However, firstly, each data individually has some missing values which need to be dealt with efficiently. To handle this task, it needs to be clear that for what data will be used, which questions or problems will be answered or solved by the data, and which type of analytics will be needed. Second, two types of data merged with inner join for creating scatter plots, which keep only the ones that are measured at the same time. Since the data are from one station and measured at the same time, this will not cause any issue, however, if there were lots of data measured at different times, there would be lots of missing values in the merged data, which leads to missing information for creating any efficient scatter plots.

For further investigation, the existence of the correlation between Oxygen content, chlorophyll, and pH could be discussed. For such tasks running statistical tests are essential as the initial step. Also, establishing and testing relationships between parameters at a single location to develop gap-filling models for missing data can be done. Furthermore, a model could be developed to predict the conditions at one station based on other locations by testing relationships between parameters at different locations, which needs access to data from other stations alongside the Elbe River.

5 Conclusions

In conclusion, water quality data strongly depends on the time of the year due to the seasonal cycle of phytoplankton growth and respiratory processes. Time series data show a cyclic behavior of each parameter every year, and no trend is visible. We observed a linear shape in the scatter plots in May, June, July, August, and September. However, correlation is not measured as it is beyond the scope and time of this study.

In line with our assumption, tidal waves influence chlorophyll and oxygen content. High tides are accompanied by high turbidity and strong zooplankton grazing in the deeper downstream parts of the estuary. Thus, a lack of chlorophyll content happens which leads to lower oxygen production. On the other hand, aerobic organisms in water consume oxygen for their respiration and cause a lower oxygen content in the water.

6 Acknowledgment

I would like to express my sincere gratitude to Dr. Jens Wyrwa, Dr. Fabian Große, and Dr. Marieke Frassl for introducing me to the BfG and providing me with an opportunity to apply my knowledge in a real-world project. Your guidance and expertise have been invaluable in enhancing my understanding and practical skills in the field of data analytics and water quality. Your dedication and enthusiasm for teaching have been evident throughout the internship, and I greatly appreciate the effort you have put into making the material accessible and engaging. The project you assigned, allowed me to work with the Python programming language and apply the concepts I learned to practical problems, which was an extremely valuable experience.

I would like to sincerely thank Prof. Dr. Peter Krebs and Geovanni Teran Velasquez for their encouragement and their time in supervising my work at TU Dresden.

A big appreciation to the management team of Python for giving us the platform to carry out my project without any limitations.

Once again, thank you all for your exceptional teaching and mentorship. Your contributions have had a significant impact on my education and professional development.

7 References

BUNCH, K. (2018, November 15). *Droughts Can Exacerbate Water Quality Problems*. International Joint Commission. https://www.ijc.org/en/droughts-can-exacerbate-water-quality-problems

Christopher, S. (2013). *Can water stay liquid below zero degrees Celsius?* Science Questions with Surprising Answers. https://wtamu.edu/~cbaird/sq/2013/12/09/can-water-stay-liquid-below-zero-degrees-celsius/

International Commission for the Protection of the Elbe River. (2018). *Elbe River basin*.

Matplotlib development team. (2023). *Matplotlib—Visualization with Python*. https://matplotlib.org/

NumPy Developers. (2022). *What is NumPy? —NumPy v1.24 Manual*. https://numpy.org/doc/stable/user/whatisnumpy.html

pandas—Python Data Analysis Library. (2023). https://pandas.pydata.org/

Rule, A., Birmingham, A., Zuniga, C., Altintas, I., Huang, S.-C., Moshiri, N., Nguyen, M. H., Rosenthal, S. B., & Pérez, F. (2018). *Ten Simple Rules for Reproducible Research in Jupyter Notebooks*.

The department U2 website. (2020). *BfG - Referat U2—Mikrobielle Ökologie, Stoffhaushalt*. https://www.bafg.de/DE/08_Ref/U2/01_mikrobiologie/mikrobiologie_node.html

The Federal Institute of Hydrology. (2020). *Bfg_brochure*. https://www.bafg.de/EN/03_The_%20BfG/bfg_brochure.pdf?__blob=publicationFile

TIBCO Software. (2023). *What is Data Exploration?* TIBCO Software. https://www.tibco.com/reference-center/what-is-data-exploration

US Department of Commerce National Oceanic and Atmospheric Administration. (2023). *Tides and Water Levels: NOAA's National Ocean Service Education*. https://oceanservice.noaa.gov/education/tutorial_tides/tides01_intro.html

US EPA, O. (2013, November 21). *Indicators: Conductivity* [Overviews and Factsheets]. https://www.epa.gov/national-aquatic-resource-surveys/indicators-conductivity

Waskom, M. (2021). seaborn: Statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/joss.03021