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. Fabian Große, Dr. Marieke Frassl

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

µg/l	Micrograms per liter
BfG	Bundesanstalt für Gewässerkunde
CO2	Carbon dioxide
CPU	Central processing unit
E	East
ETR	External Timer Reference
Km	Kilometer
m3/d	Cubic meter per day
mg/l	Milligram per liter
mm	Millimeter
N	North
O2	Oxygen
UTM	Universal Transverse Mercator
vs.	Versus

1 Introduction

Microbiological processes have a significant impact on the ecological systems of federal rivers in inland and coastal regions. The biological, chemical, and physical characteristics interact with the microorganisms in the sediment, water, and shoreline. Bundesanstalt für Gewässerkunde (BfG) as a scientific institution ranking as a supreme federal agency, is responsible for the German waterways in federal ownership. (BfG), Unit U2 is engaged in the comprehensive investigation and advisory work (M. Frassl (BfG) & M. Mannfeld (BfG), 2020; The Federal Institute of Hydrology & Die Bundesanstalt für Gewässerkunde – BfG, 2020).

Microorganisms including bacteria and phyto- and zooplankton live in the water, in the sediment, and in the shore areas of federal 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. Unit U2 investigates their populations and influence in inland waters, estuaries, and coastal waters. The investigations regularly take place on the Berlin waterways, the Elbe, the Rhine and Moselle, and the North German estuaries (M. Frassl (BfG) & M. Mannfeld (BfG), 2020).

In this report, the data specifically from the Bunthaus station at Elbe River is analyzed to check and understand the total behavior of some of the water quality elements (O2, chlorophyll, pH, water temperature, and electrical conductivity) separately, and in comparison with each other.

The Elbe has the fourth-largest river basin in Central and Western Europe with a population of around 25 million. It originates in the Giant Mountains of the northern Czech Republic and flows into the North Sea near Cuxhaven, 110 km (68 miles) northwest of Hamburg, after passing through much of Bohemia (the western half of the Czech Republic). 1,094 km make up its entire length (International Commission for the Protection of the Elbe River, 2018).

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 (*What Is Data Exploration?*, 2023).

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

1.1 Objective

The purpose of this internship is to visualize and understand the data given by 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 alongside the Elbe where water content measurements normally occur. The data we analyzed was gathered at the Bunthaus station placed on the left side of the northern Elbe River.
- Data origin: http://www.portal-tideelbe.de/
- Station name: Bunthaus
- Geo reference: ETRS89/UTM32N
- Location: 53°27'42.1"N 10°03'51.6"E
- River kilometer: 609.8 (below Germany / Czech border)
- Catchment area: 138380 km²
- Reference level: Neu Darchau, Elbe
- Observation start date: 1988
- Water quality: Chlorophyll (from 2014), oxygen, electric conductivity, pH, water temperature
- Water level (from 1950): Tidal flood, Tidal low water
- Operator: Institut für Hygiene und Umwelt, Hamburg
- station type: multiparameter station
- water body: Elbe, main section Elbe km 607,50 to 638,98
- stream kilometer: 609,8
- 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 totally raw in a text file, with often more than 1,600,000 records. There are some missing values due to the interval between measuring times. For Chlorophyll 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 separately appended. 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 is 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 are:

- 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, animated, and 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. As mentioned in the objective, data exploration happened via data visualization. In this project we made two main types of graphs:

- Line graph (time series chart) for each water quality parameter (O2, chlorophyll, pH, water temperature, and electrical conductivity), which shows their changes over years.
- Scatter plot for Oxygen vs. Chlorophyll content, Chlorophyll content vs. pH, and Oxygen vs. pH, which helps to recognize any relationships and dependencies.

3 Results

In this part, line graphs and scatter plots help us to explore the data more. Data exploration enables users to more effectively choose which areas of the data to investigate further and to gain a general picture of the situation before posing more specific queries (*What Is Data Exploration?*, 2023).

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 heterotrophs that consumed O_2, and the increase of CO_2 due to contaminants before reunification, which caused the relatively lower pH. After reunification, the yearly average of the pH increased and seasonally fluctuated over the approximate average value of 8.

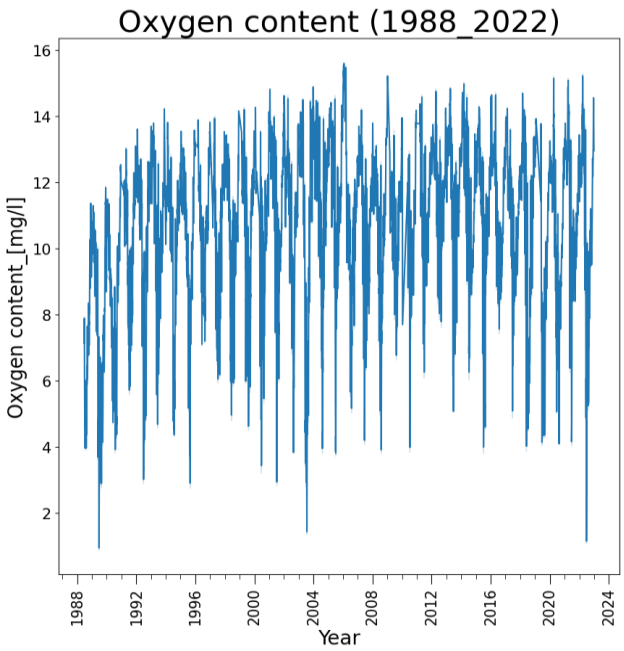

Figure 3: Oxygen content (mg/l) measured at Bunthaus station from 1988 to 2022.

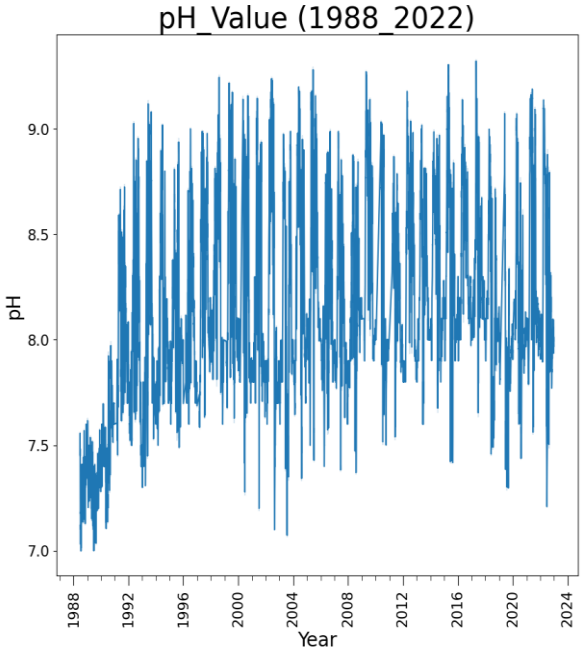

Figure 4: pH values measured at Bunthaus station from 1988 to 2022

In Figure 3, 3 years with the lowest oxygen content are the most important ones. The first trough is in 1989 which might be due to lots of organic matter dissimilation and contamination in water before reunification. The second is in 2003 during the 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 (*Droughts Can Exacerbate Water Quality Problems*, 2018). And the third trough is in the year 2022, in which we assume the extreme growth of zooplankton caused a lack of oxygen.

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 obviously difficult for plants to grow in the winter, the amount of chlorophyll in water is often highest in the summer and lowest in the winter, and it has a cyclic behavior (Figure 5). Chlorophyll in water is impacted by a variety of human activities, including wastewater discharge and the erosion of lake and river shorelines (Craig, 2013).

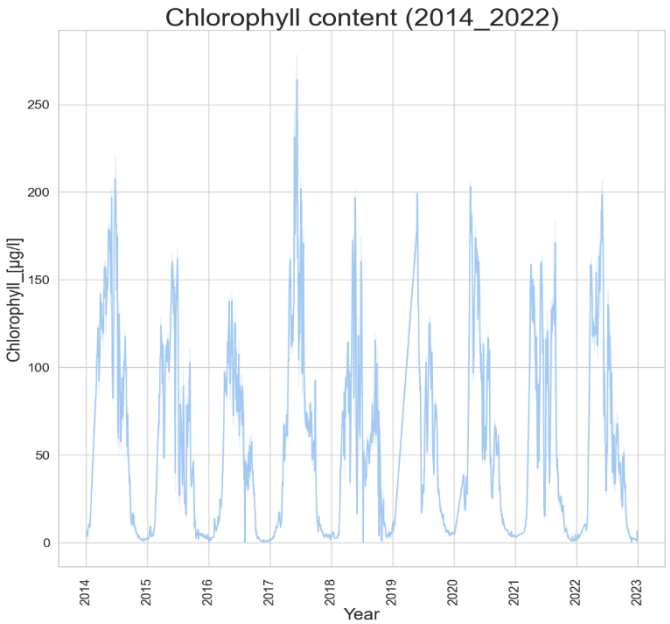

Figure 5: Chlorophyll content (µg/l) measured at Bunthaus station from 2014 to 2022

The information above can explain Figure 6, which shows that chlorophyll content is low during winter and high in summer months with a peak in June. The fast increase begins in March when there is enough light and warm temperature to allow the chlorophyll to grow, until the beginning of April, when the water nutrients are consumed to a degree that makes the chlorophyll content

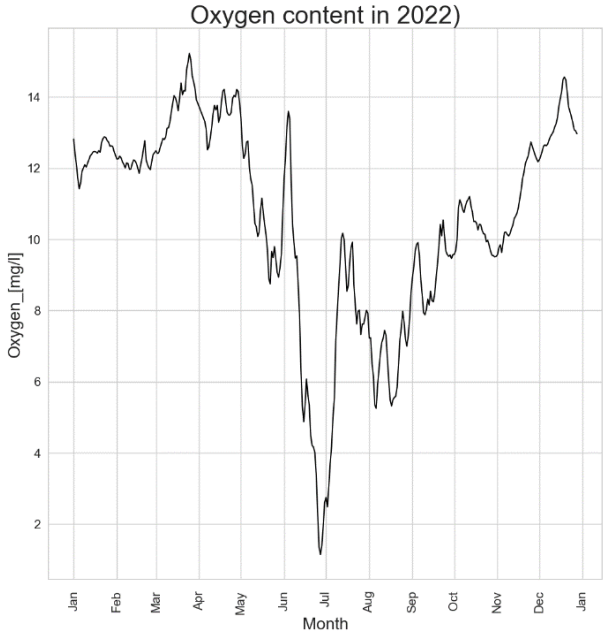

Figure 7: Oxygen content (mg/l) measured at Bunthaus station in the year 2022

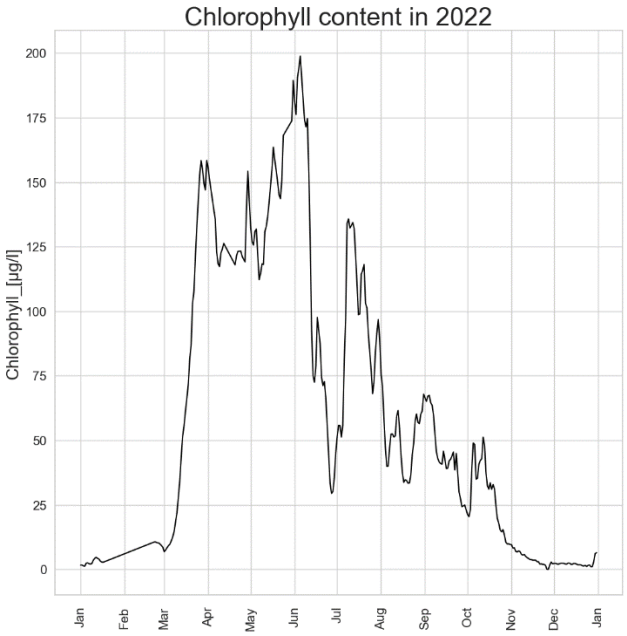

Figure 6: Chlorophyll content (µg/l) measured at Bunthaus station in the year 2022

reaches a plateau. And from the first 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 consumption overtook the phytoplankton production. Release of nutrients due to

dissimulation of phytoplankton by zooplankton causes the growth of phytoplankton again and chlorophyll contents in mid-July. Finally, it generally decreases from August as the weather gets darker and colder.

3.1.3 Electrical conductivity and water temperature

Water temperate doesn't have any trend and it's seasonally fluctuating 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 S. Baird, 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.

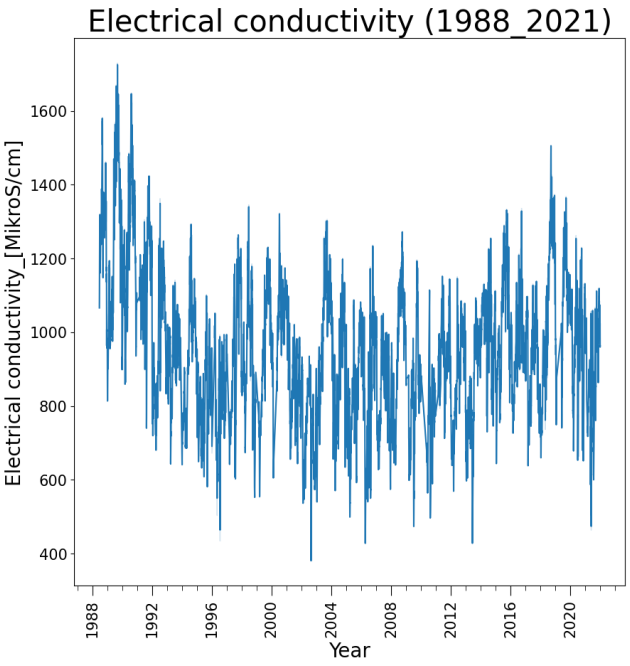

Figure 9: Electrical conductivity measured at Bunthaus station from 1988 to 2021

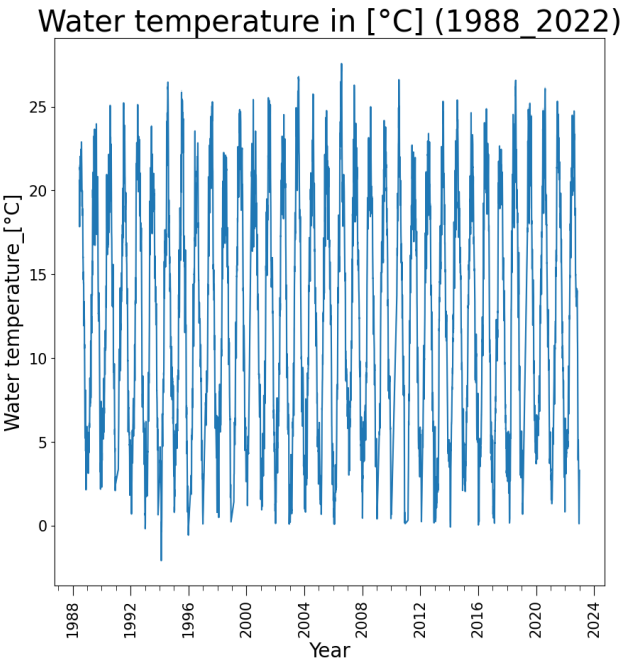

Figure 8: Water temperature (°C) measured at Bunthaus station from 1988 to 2022

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 for now.

3.2 Scatter plot

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

3.2.1 Chlorophyll vs. pH:

As we expected and we can also see in Figure 10, the pH is directly proportional to chlorophyll content. Also, we can see this relationship becomes stronger in recent years.

Figure 11 is a monthly-separated graph that clearly shows that pH and chlorophyll content depends on the month of the year. A direct relation can be seen more clearly in the summer months. As the cold months come, the chlorophyll content becomes lower and the pH is almost fixed around 8.

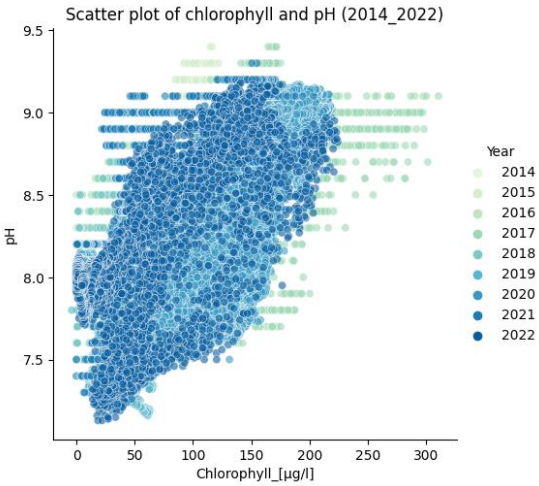

Figure 10: pH values and chlorophyll content (µg/l) scatter plot measured at

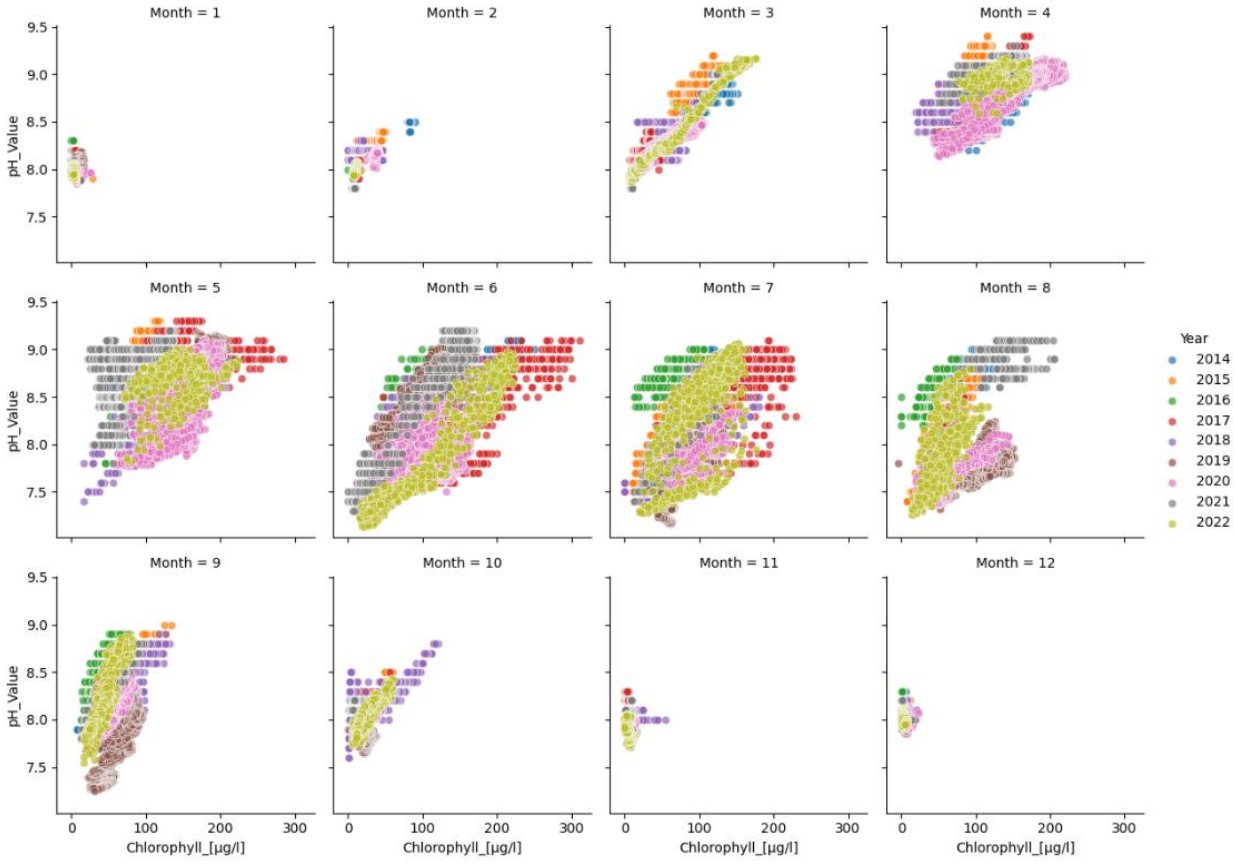

Figure 11: pH values and chlorophyll content (µg/l) relationship in each month of the year measured at Bunthaus station from 2014 to 2022

3.2.2 Oxygen vs. pH:

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

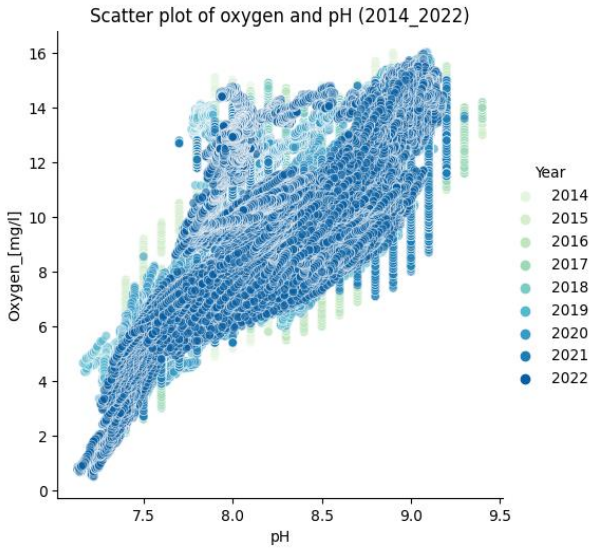

Figure 12: Oxygen content (mg/l) and pH values relationship measured at Bunthaus station from 2014 to 2022

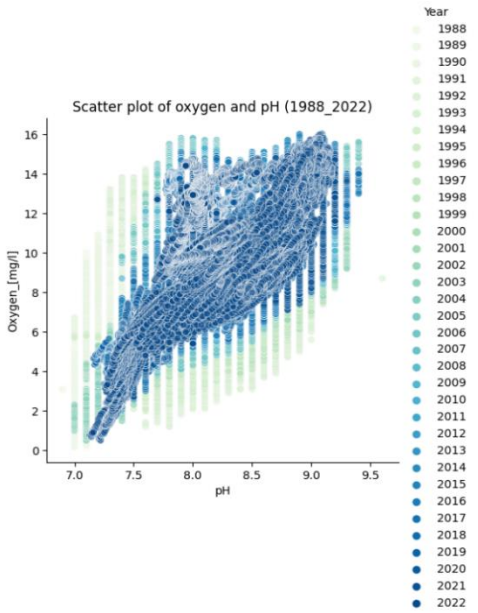

Figure 13: Oxygen content (mg/l) and pH values relationship measured at Bunthaus station from 1988 to 2022

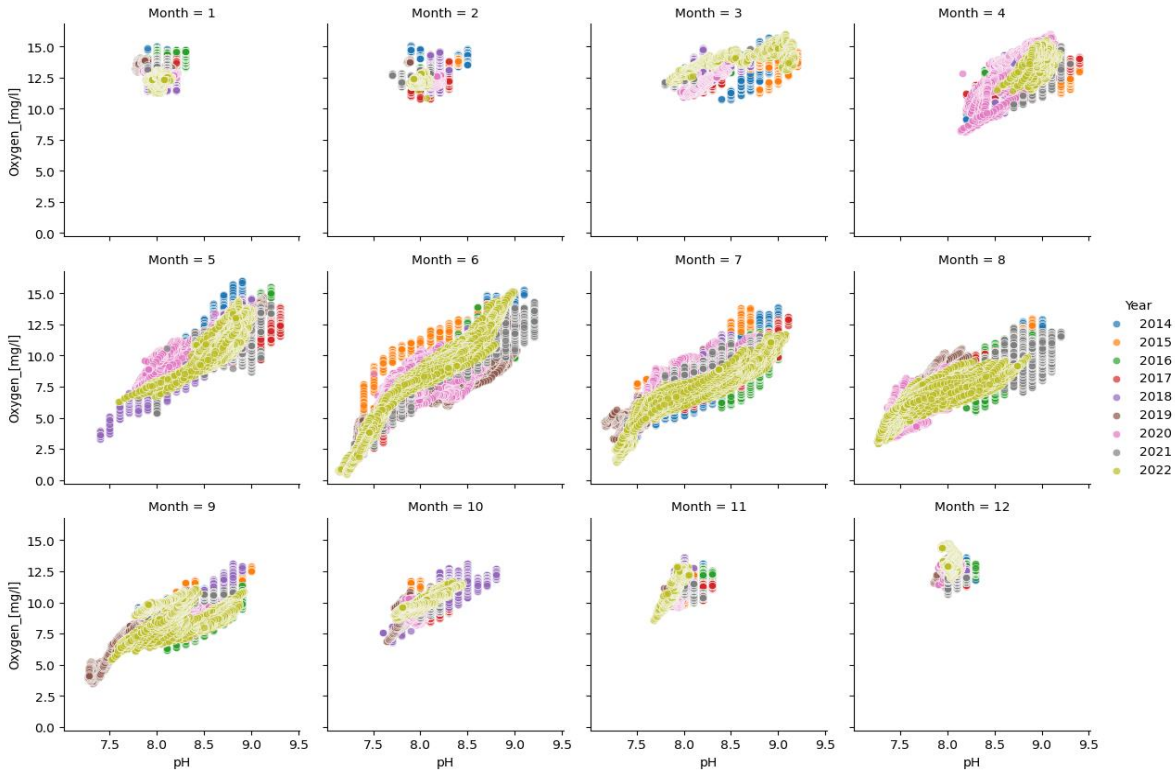

Figure 14: Oxygen content (mg/l) and pH values monthly relationship measured at Bunthaus station from 2014 to 2022

As pH and oxygen have huge dispersion 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 as the weather becomes warmer the range of pH and oxygen grows in an approximately linear shape. 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. 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)

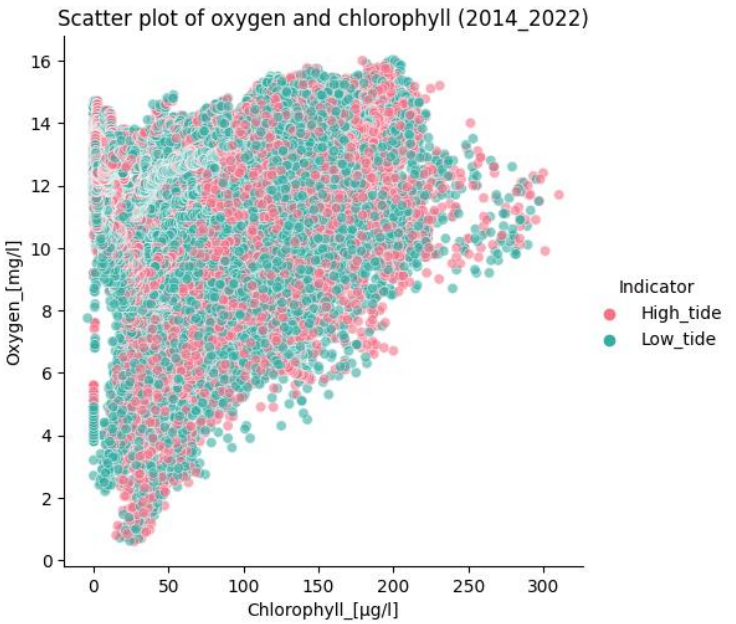

Figure 15: Oxygen (mg/l) and chlorophyll (µg/l) relationship measured at Bunthaus station from 2014 to 2022

In principle, tides are extremely 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, 2023). We consider the time period from when the trough is recorded until the tide peak is recorded as a high tide period, and the time from when the peak is recorded till the trough is recorded, as a 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 is 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.

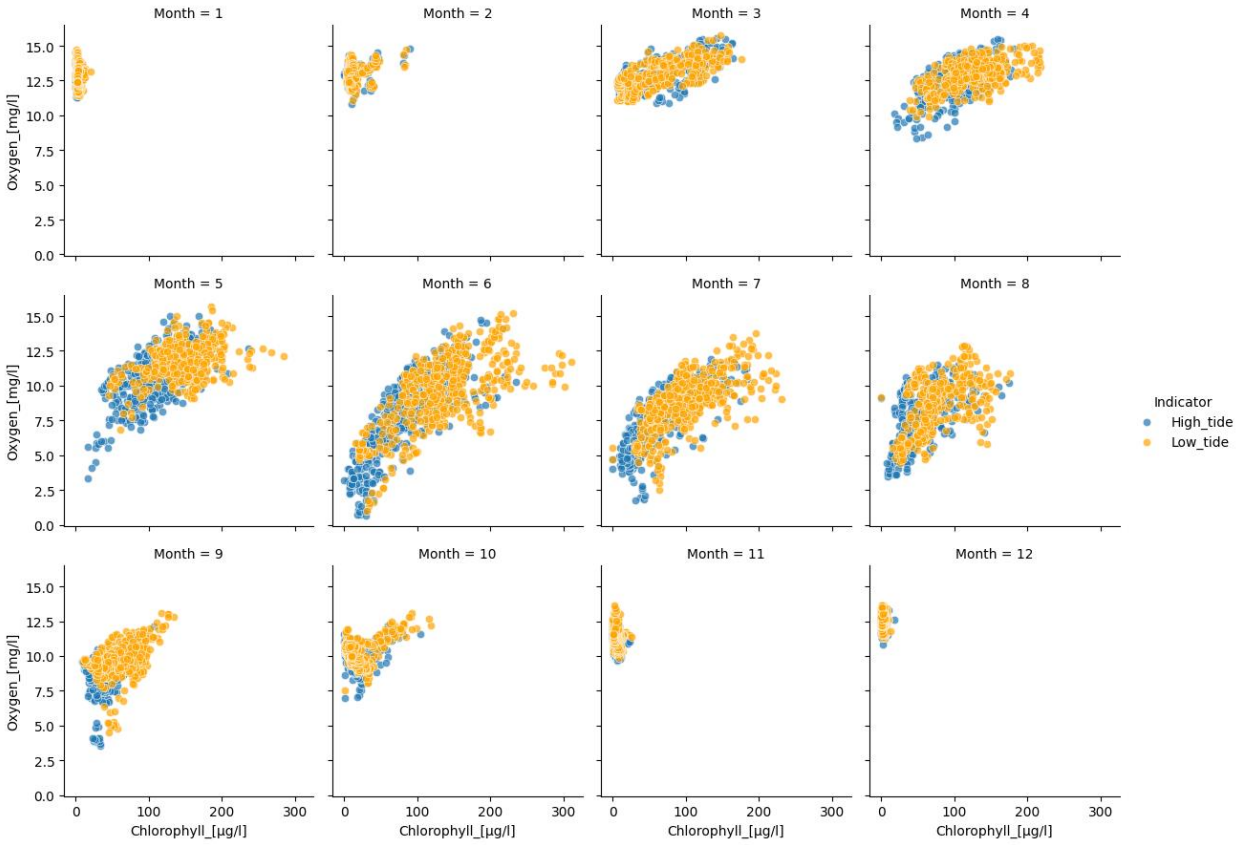

Figure 16: Oxygen (mg/l) and chlorophyll (µg/l) monthly contents measured at Bunthaus station during the peak of high and low tides from 2014 to 2022

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 expected, in high tides we have lower contents. Because in high tides, the sunlight hardly goes through the water due to high turbidity caused by ships and pollutants. Thus, lower growth of chlorophyll and accordingly oxygen is expected.

4 Discussion

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, perform 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. 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.

Analyzing data, especially when it comes to visualization, needs modern hardware with a relatively high CPU. For the data I worked with, the current system was totally sufficient but still not much fast. It is worth mentioning that when it comes to big data analytics then the current system must be changed with a suitable data science system. Accordingly, the more data, the higher the system performance should be.

For further investigation, the existence of the correlation between Oxygen content, chlorophyll, and pH and whether this correlation means causation could be discussed. For such tasks running statistical tests are essential as the initial step. Furthermore, a model could be built to predict a specific and intended parameter behavior. Moreover, if access to data from other stations alongside the Elbe River is possible, there will be the possibility of checking the locational relationship among data and also calibrating and validating the predictive model.

5 Conclusions

In conclusion, water quality data are highly dependent on the time of the year. Time series data show a cyclic behavior of each parameter every year, and no trend is visible. Based on the scatter plots, we observed a long-range linear shape between the two parameters in May, June, July, August, and September. However, correlation is not measured as it is out of the purpose and time of this study.

In line with our assumption, tidal waves influence chlorophyll and oxygen content. High tides are accompanied by high turbidity, which prevents sunlight penetration into the water. 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.

These findings can be useful for researchers working in the field of river water quality management, as they provide valuable insights into the impact of tidal waves on water quality. Further research is needed to understand the underlying mechanisms behind these observations and to identify potential solutions to mitigate the impact of tidal waves on water quality.

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 institution 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

Christopher S. Baird. (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/

Craig. (2013). *Chlorophyll*.

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

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

M. Frassl (BfG), & M. Mannfeld (BfG). (2020). *BfG - Referat U2—Mikrobiologie, Stoffhaushalt*. https://www.bafg.de/DE/08_Ref/U2/01_mikrobiologie/mikrobiologie_node.html

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 Federal Institute of Hydrology & Die Bundesanstalt für Gewässerkunde – BfG. (2020). *Bfg_brochure*.

US Department of Commerce, N. O. and A. A. (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

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