Data Mining study on data collected in Arctic Oceanographic Campaigns

Tânia Isabel Alexandre Mestre Ferreira
Engenharia de Redes e Sistemas Informáticos
Departamento de Ciência de Computadores
2020

Orientador
Rita Paula Almeida Ribeiro, Professora Auxiliar, Faculdade de Ciências da Universidade do Porto

Coorientador
Catarina Maria Pinto Mora Pinto de Magalhães, Professora Auxiliar Convidada, Faculdade de Ciências da Universidade do Porto

Todas as correções determinadas pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto, _____/_____/_____

Abstract

From 2016 to 2019, the Norwegian Polar Institute organized four Arctic expedition campaigns in the context of the monitoring program Svalbard and Jan Mayen (MOSJ program). From these expeditions, valuable data was collected by different research areas with three main purpose levels: biological, biogeochemical, and environmental. One important possibility that this data set collected at extreme environments opens is the study of the effects of global warming on polar ice and in the evolution of microbiological communities existing in these waters. Typically, this data is subject to purely statistical analysis, performed independently by the researchers involved in these campaigns.

The goal of this study is to allow the acquisition of new and insightful analysis of the collected data, not only through statistical analysis in space and time but also by some machine learning descriptive methods. Through our developed web application, we provide a dynamic way to analyze the data and interpret its results at a graphical level. The application offers three main types of analysis: temporal and spatial visualization and further statistical analysis and descriptive modelling. Our objective is to enable a spatial and temporal perception of the evolution of a set of selected parameters. It is possible to observe how the values for the various parameters change during the years under investigation, for a given station. A similar exercise can be performed for a set of stations, for a specific year. Correlation analysis between the parameters, subject to year and station is also available. Additionally, we generated association rules to assess the environmental parameters that are influencing the higher occurrence of primary producers in the system.

In the end, we believe that this dynamic data analysis platform can be considered a possible relevant contribution to further research studies on the collected data.

Resumo

De 2016 a 2019, o Instituto Polar Norueguês organizou quatro campanhas de expedição ao Ártico no contexto do programa de monitorização Svalbard e Jan Mayen (programa MOSJ). A partir dessas expedições, dados relevantes foram recolhidos por diferentes áreas de investigação com três propósitos de estudo principais: biológico, biogeoquímico e ambiental. Este conjunto de dados recolhidos em contextos extremos oferece a importante possibilidade de estudar osefeitos do aquecimento global no gelo polar e na evolução das comunidades microbiológicas existentes nessas águas. Normalmente, esses dados estão sujeitos a análises puramente estatísticas, realizadas, de forma independente, pelos investigadores envolvidos nessas campanhas.

O objetivo deste estudo é permitir a elaboração de análises novas e criteriosas dos dados recolhidos, não apenas por meio de análises estatísticas no espaço e no tempo, mas também por alguns métodos descritivos de *machine learning*. Através do desenvolvimento de uma aplicação web, fornecemos uma maneira dinâmica de analisar os dados e interpretar os seus resultados a um nível gráfico. A aplicação oferece três tipos principais de análise: visualização temporal e espacial, análise estatística e modelação descritiva. Desta forma, possibilitamos uma percepção espacial e temporal da evolução de um conjunto de parâmetros selecionados. É possível observar como os valores dos diversos parâmetros mudam ao longo dos anos em investigação, para uma determinada estação. Um exercício semelhante pode ser realizado para um conjunto de estações, para um ano específico. A análise de correlação entre os parâmetros, sujeita ao ano e estação também está disponível. Além disso, permite gerar regras de associação para avaliar os parâmetros ambientais que estão a influenciar a maior ocorrência de produtores primários no sistema.

Em conclusão, acreditamos que esta plataforma de análise dinâmica de dados pode ser considerada uma possível contribuição relevante para futuros estudos de investigação sobre novos dados.

Agradecimentos

Agradeço todo o apoio dado pelas minhas orientadoras e pelos conhecimentos que foram transmitindo ao longo de todo este processo. À minha mãe agradeço o incentivo que me deu e aos meus amigos por me apoiarem nos dias mais complicados. O meu muito obrigada.

This project was financed by the Portuguese Polar Program, the Norwegian Polar Institute (NPI), the MOSJ (Environmental Monitoring of Svalbard and Jan Mayen) and the Portuguese Foundation for Science and Technology through the NITROLIMIT project (PTDC/CTA-AMB/30997/2017).

Contents

List of Tables

List of Figures

Acronyms

CTD Conductivity, Temperature and Depth

PAR Photosynthetically Active Radiation

SPAR Surface Photosynthetically Active Radiation

EDA Exploratory Data Analysis

DBSCAN Density Based Spatial Clustering of Applications with Noise

Chapter 1

Introduction

Over the past years, due to warming effects on the Arctic climate, the thickness and extent of Arctic summer sea-ice has been severely decreasing. As a result, the ice pack has become younger, and the older multi-layered ice (MYI) that survived the summer melting has been disappearing and being replaced by first-year ice (FYI). These modifications in the ice cause changes in the dynamics of Arctic phytoplankton and biogeochemistry [36].

The biogeochemical implications of the alterations in the Arctic sea-ice regime must be monitored in detail at different trophic levels to assess the consequences that result from that, in the sustainability of ecosystems as well as the primary production.

The lowest of sea-ice extent was recorded in October of 2016, and it is assumed that due to the rapid decrease in sea-ice cover and its thickness over the last few years, the Arctic will be completely ice-free in the summer by the end of the 21st century [32].

1.1 Motivation

Over the course of the Norwegian Polar Institute monitoring program of Svalbard and Jan Mayen (MOSJ program), a tremendous amount of data is gathered with three different purpose levels: biological, biogeochemical, and environmental. These areas acquire data differently, either at different depths or at different times. Some of the parameters can be measured on board related to near-surface meteorological conditions, such as air temperature and wind speed, sea surface conditions, such as sea surface temperature and salinity, as well as subsurface water characteristics, like oxygen, salinity, fluorescence depth profiles, ocean currents or dissolved nutrients. Regarding the biogeochemical data (nutrients) they include spatio-temporal context and samples were processed in laboratory after the oceanographic campaigns.

As there is so much data that needs to be analysed and visualised, there is considerable interest from the domain experts in reducing the gap that exists between the data collected and data analysis and publication. The aim of this study is to be able to accelerate the process of extracting relevant temporal and spatial information from an Arctic Ocean monitoring data set.

For that purpose, appropriate data mining techniques need to be applied.

1.2 Goals

The ultimate goal of this thesis is to create a Shinny application [13] that allows for the integration and visualization of all the collected data throughout the campaigns by relating them in a spatial and temporal dimensions. In particular, the aim is to provide an interactive and responsive tool that allows a set of dynamic analysis to be performed. Still, for this goal to be accomplished is important to define what are the best statistical analysis and descriptive modeling techniques to obtain meaningful insights over the data. Our study will be guided by the exploration of known important parameters on the collected data and how they relate to each other.

1.3 Organization

After this introductory chapter, the remain of the thesis is organized as follows.

In Chapter 2, we present a review of the literature related to this study. The goal is both to understand the collected data and propose some techniques for data exploration.

In Chapter 3, we describe the set of parameters that compose our data set and the performed pre-processing steps. We also discuss some exploratory analysis made on the data set.

In Chapter 4, we present the developed Shiny App [13] to integrate the visualization and some of the statistical analysis over the data. We describe the layout of the application, its components and respective organization. The aim is to provide a user manual on how to interact with the application.

Finally, in Chapter 5, we conclude with the results we have reached, as well as suggestions for future work.

Chapter 2

Literature Review

In this chapter, we start by contextualizing the problem from the biological and chemical perspective. We then present the data mining techniques and tools used throughout this study.

2.1 Arctic Oceanographic Campaigns

One of the biggest manifestations of climate change is the radical changes observed in the Arctic sea-ice seasonal regimes. The reduction in the extent of summer sea-ice and the modifications in the layers of the sea-ice, turning the multiyear ice which is thicker into a first-year ice that is severely thinner. This has a documented impact on the microplankton communities that are a key component of the Arctic marine food web and maintain the stability of biogeochemical cycles [36]. The samples collected throughout the Arctic campaigns allow the study of Arctic's water relevant parameters and the dynamics of microbial community. This is possible through parallel sequencing of small subunit ribosomal RNA amplicon and metagenomic data (environmental DNA) [18].

During the Arctic Ocean 2016, 2017, 2018 and 2019 campaigns, as part of the long term environmental monitoring program of Svalbard region (MOSJ) led by the Norwegian Polar Institute (NPI), an enormous amount of in situ and laboratory physical, biochemical and biological data has been collected. To accomplish this, a total of 20 stations along two transects (West and North Svalbard and Kongsfjorden and Rijpfjorden) were characterized using relevant deep profiles sensors together with the collection of water samples at 3 different depths (surface, maximum chlorophyll and above seafloor) for complementary biogeochemical variables [32].

Collected water samples were stored on board and analysed at NPI laboratories for compounds such as nitrite, nitrate and ammonium among others [32]. The results from previous analysis of the environmental variables showed higher temperatures and higher concentration of nitrogenous compounds, especially ammonia, on the surface waters of the Kongsfjorden transect. In both Kongsfjorden and Rijpfjorden there is an increase in the concentration of nitrogenous compounds mainly nitrate and silica hydroxide (Si(OH)4) with depth [32]. The correlations between environmental and biogeochemical variables with respective depths were identical in both transects [32]. At biological level, previous studies also registered differences in their spatial

(between different stations and different depths) and temporal distribution in both Rijpfjorden and Kongsfjorden transect.

Our data set consists of the data collected during the expeditions MOSJ-ICE2016, MOSJ-ICE2017, MOSJ-ICE2018 and MOSJ-ICE2019, where samples were taken from oceanographic stations from both transects, Kongsfjorden and Rijpfjorden.

The majority of the previous analysis carried out over the data collected on MOSJ monitoring program were at a statistical level. Our study goes beyond this type of analysis by including a statistical analysis component which shows how the correlations change over the time period range of the expeditions and also include other analysis that integrate spatio-temporal data in order to draw conclusions about how a parameter or a station can influence others.

2.2 Data Mining

Data mining is the process of discovering information from data. It is built on several fields including statistics, mathematical modeling, database activities, machine learning and artificial intelligence. Data mining can extract useful insights to both summarize the available data and provide actionable information to make crucial decisions [25].

One of the definitions of data mining was given by David J. Hand [26] as follows:

> *"Data mining is the analysis of (often large) observational data sets to find unexpected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner."*

Data Mining comprises the process of collecting, cleaning, processing, analyzing and gaining useful insights from data. According to Aggarwal [3] this process has three main phases: Data Collection, Data Preprocessing and Analytical Processing, as illustrated in Figure 2.1. The process is not always done one-way, in most cases, the feedback provided by a phase to a previous one can be quite useful in order to improve the method. We describe each of these phases in more detail next.

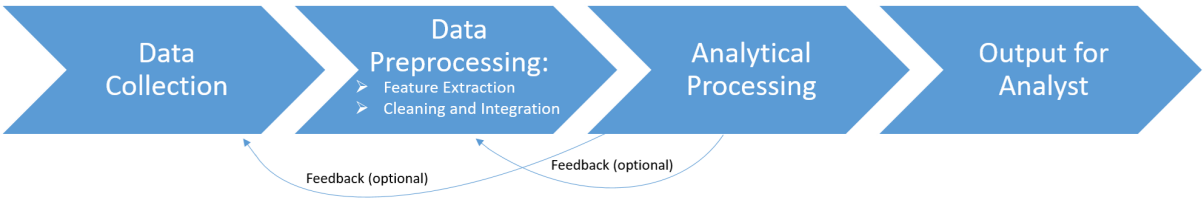


Figure 2.1: Data Mining phases according to Aggarwal [3].

Data Collection - It is the process of gathering information, which is where the data is obtained.

Data Preprocessing - Once the data is collected, it is necessary to process it. This can be considered the most important phase, since this is where it may be necessary to perform a set of operations before doing any type of analysis. At this phase the data can come from multiple sources and, in order for this data to be handled, it must be transformed into an acceptable format. This stage is dependent on the analyst to extract the features that may be most relevant to the project, so it is important that those who do this analysis understand exactly the data they have to work with, so that they would figure out what could be more interesting to explore. During this process it may be needed to have to address some data quality issues by cleaning the data (e.g. impute missing values). It also includes feature selection and transformation, in order to obtain the relevant data, which although smaller than the collected data, still maintaining the integrity. In the transformation part, the data is consolidated to make the data mining process more efficient and easier to further recognize patterns.

Analytical Processing - This last phase corresponds to the mining process, which is the creation of analytical methods to solve the problem in question from the processed data. The major challenge is the difficulty to create techniques that are general and reusable for different applications. However, many data mining formulations are used in different applications. It is up to the analyst's experience to determine whether or not these different formulations can be used in the specific context of an application [3].

Nowadays it is easy to obtain a huge amount of data through remote sensors, satellites, where, for example, environmental events are observed without the need for contact with the phenomena itself. Even so, we have other forms of data collection, such as field campaigns that involve manual observations and measurement of a range of environmental phenomena [9]. When dealing with environmental data and since it is relatively simple to have access to a large amount of data, where we have a wide variety of sources, the pre-processing and analytical processes become even more complex. We do not know exactly how each type of data was collected and in what situations it was collected. These leads us to question the veracity of the data: it is different to have an observation from a satellite than an observation made by a professional on the spot [9]. Therefore, it is possible to identify the arising of some data challenges [9], namely:

- managing the variety and heterogeneity in underlying sources of data, including achieving interoperability across data sets;
- making all data open and accessible through environmental data centers;
- ensuring all data are enhanced with appropriate semantic meta-data capturing rich semantic information about the data and inter-relationships;
- ensuring mechanisms are in place to both record and reason about the veracity of data;
- finding appropriate mechanisms and techniques to support integration of different data sets to enhance scientific discovery and constrain uncertainty.

In this dissertation we aimed to address some of these data challenges. In particular, we focused on building a platform to integrate different data sources to enhance exploratory analysis and support knowledge discovery. In the next subsections we will explain some techniques that we used during this study. Namely, exploratory data analysis, summarization and descriptive modeling techniques, and techniques specially suited for handling spatio-temporal data.

2.2.1 Exploratory Data Analysis

Exploratory Data Analysis (EDA) is based on statistics. It is considered to be one of the most important steps when analyzing data and is fundamental to the pre-processing phase presented by Aggarwal [3]. After collecting the data, the analysis should not be carried out immediately [16], i.e. without any initial exploratory analysis. EDA provides a way to understand what kind of data is available for the analysis, mainly using graphical techniques, in order to understand the limitations of the data sets, essential to define the methodologies of data mining analysis. It is based on organizing and summarizing the collected data and identifying irregularities or patterns in the observations [17].

To begin with, it is necessary to perform a pre-analysis of the data, identifying different data types and viewing the data with basic plots. The goal is to identify the missing data, incorrect entries, outliers and inconsistencies. The existence of these problems can significantly affect the possible conclusions that may be drawn from the data, jeopardizing the quality of the decisions taken on the basis of this analysis [3]. After this pre-analysis, it may be possible to identify if the data follows any known model that allows the study of the question under analysis or if other techniques need to be found in order to solve it [17].

According to [14], the purposes of the EDA are:

- give a general structure of the data;
- compute summary statistics;
- obtain simple descriptive summaries;
- check the data quality;
- draw the appropriate graphs;
- evoke ideas for more complex analysis.

Various graphic techniques are used in EDA. These techniques are used to study patterns and relationships between data in a visual and simplified way, allowing users the possibility to retrieve information more easily. Some of the most commonly used graphics are histograms and heatmaps.

A histogram is a visual representation of the distribution of values assumed by a numerical variable. It consists of two axes (x and y) and various bars with different weights. The x-axis

groups the values of the variable in a set of consecutive range of values. The y-axis shows how often each range of values in the x-axis occur in the sample. Figure 2.2, shows a representation of a histogram.

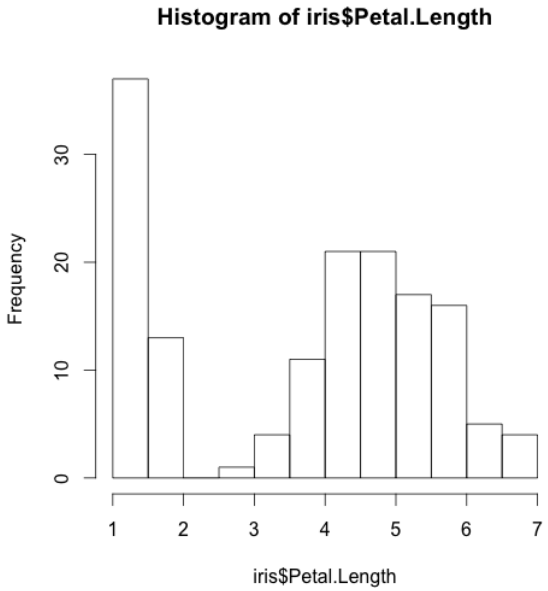


Figure 2.2: Histogram of Petal Length variable in the iris data set [22].

Heatmap is a graphical representation where individual values are represented in a matrix and are characterized by colours. This type of graphs are efficient to visualize relations between different variables. Figure 2.3 exhibits an example of a heatmap, where the lines represent IDs and the columns are parameters. The color index corresponds to the values for each of the matrix entries it represents

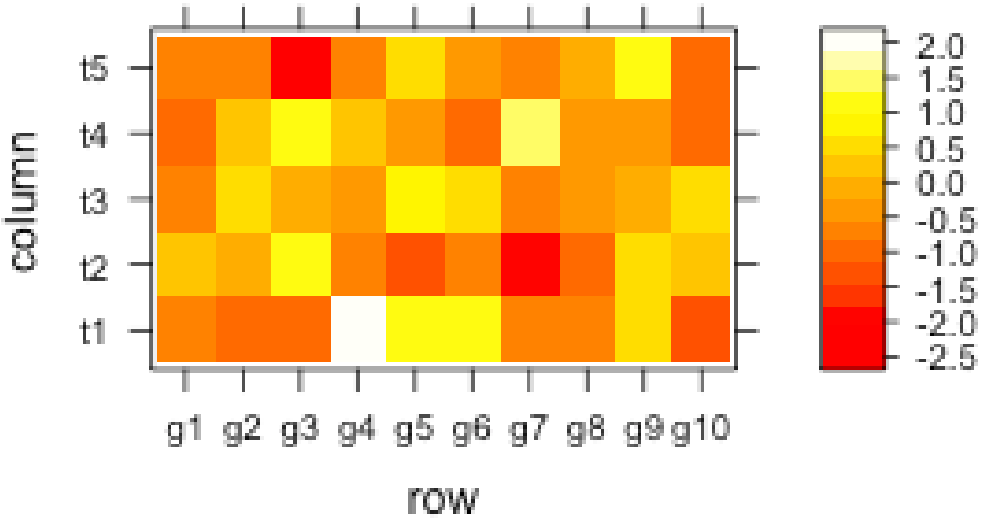


Figure 2.3: Heatmap example for a data set [24].

Heatmaps can also be used to represent correlations between variables at a visual level. The

correlation analysis is useful to examine the nature of relationships between the pairs of variables. There are different correlation coefficients, but there are two most common used when treating environmental data [30]: *Pearson* and *Spearman* coefficients.

The *Pearson* coefficient is a common measure of association between two continuous variables, measures the linear dependence between two variables [39]. Is defined as a ratio between the covariance of the two variables and the product of their respective standard deviations, commonly denoted by ρ [15].

When we apply the *Pearson* correlation coefficient to a sample it is represented by r_{xy} and can be called a sample correlation coefficient. Given a set of pairs $\{(x_1, y_1), ..., (x_n, y_n)\}$ consisting of n pairs r_{xy} is defined by [41]:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{2.1}$$

where n is the sample size, x_i and y_i are the individual sample points indexed with i and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the sample mean, the same for $\bar{y}$ [41].

The *Pearson* correlation coefficient ranges from -1 to +1. A $\rho > 0$ means that two variables tend to increase or decrease simultaneously and a $\rho < 0$ implies that one variable tends to increase when the other variable decreases [15]. An example is shown in Figure2.4.

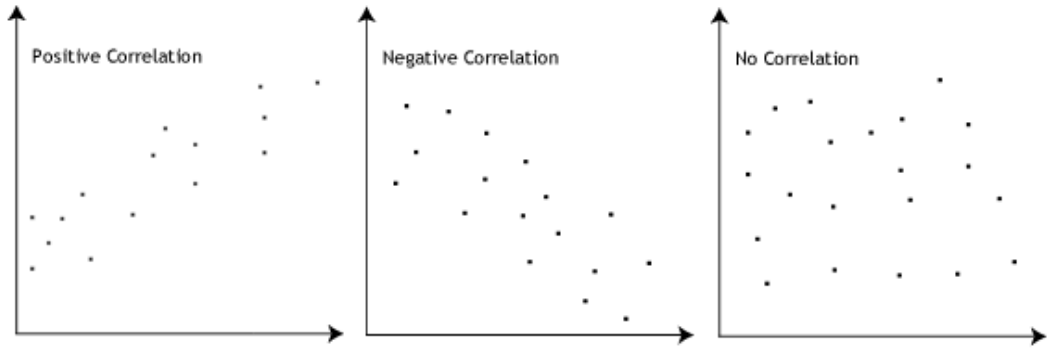


Figure 2.4: Three main Pearson correlation scenarios [37].

The *Spearman* correlation coefficient is a rank-based version of *Pearson* correlation coefficient. It is estimated by the following equation:

$$rs_{xy} = \frac{\sum_{i=1}^{n}((rank(x_i) - \overline{rank(x)})(rank(y_i) - \overline{rank(y)}))}{\sqrt{\sum_{i=1}^{n}(rank(x_i) - \overline{rank(x)})^2 \sum_{i=1}^{n}(rank(y_i) - \overline{rank(y)})^2}} \tag{2.2}$$

where $rank(x_i)$ and $rank(y_i)$ are the ranks of the observations in the sample.

This correlation coefficient also varies from -1 to +1 and the absolute value of ρ describes the strength of the monotonic relationship, which occurs when one variable increases and the other also increase or when a variable decreases and the other decreases too [15].

Spearman correlation coefficient differs from *Pearson* correlation coefficient in that it can be

1 not only when variables are linearly related but also when variables are related according to a type of non-linear but still monotonic relationship [15]. An example of the results that can be obtained is in Figure2.5. When analyzing results from correlations we need to be aware of the statistical significance, that is the correlation depends on the sample we are analyzing. In a small sample one correlation between two variables might not be as strong as in a larger sample.

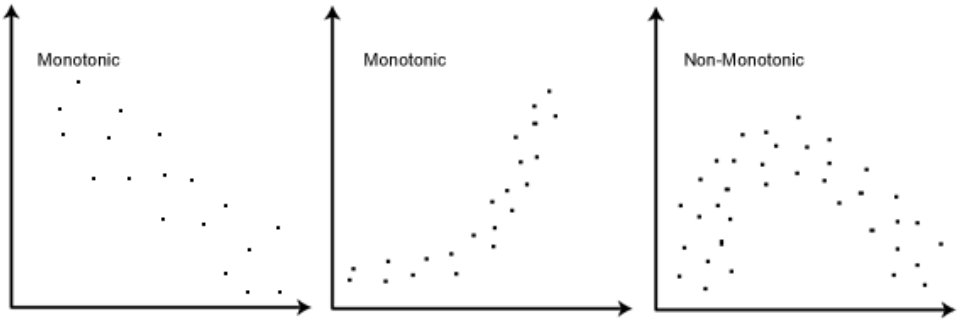


Figure 2.5: Three main Spearman correlation scenarios [38].

2.2.2 Machine Learning

Machine learning is one of the components of data mining and is included in the analytical process as presented by [3]. It is considered a subfield of artificial intelligence and statistics. It refers to algorithms that automatically learn to recognize complex patterns in new data sets, improving their performance from experience. Machine Learning is one of the most important steps in the last stage, where we already have clean data and we want to achieve the results we set out to reach. According to [23], Machine Learning is about using the right features to build the right model for the task we are trying to execute.

There are two major machine learning tasks: supervised and unsupervised learning tasks. The Table 2.1 summarises the different machine learning settings, according to [23].

Table 2.1: Overview of different machine learning settings [23].

	Predictive model	Descriptive model
Supervised learning	Classification Regression	Subgroup Discovery
Unsupervised learning	Predictive Clustering	Descriptive Clustering Association Rules

In supervised learning a target value that is associated with each example, i.e. for each example of the training data set, is given a target with the answer of how the algorithm should respond. The objective of this model is to learn a function (model) that maps each example to its target variable (e.g. predictive modeling).

In unsupervised learning there is no value to be associated with each example, a training data set is a collection of examples for which there is no correct answer, i.e. the goal is to get a description of the data set (e.g. descriptive/predictive clustering, association rules).

It is also possible to distinguish whether the model output involves the target data or not. Predictive model is the name given when the target variable is included, and descriptive model if the target variable is not used [23].

Classification and Regression are both predictive modeling techniques in a supervised learning context. Classification is the approximation of a mapping function that receives variables as input and as output returns discrete variables (categories). This mapping function predicts the class or category of a given observation [10]. Regression represents a mapping function that receives variables as input and returns continuous variables as output [10].

Subgroup Discovery is a data mining technique that extracts interesting rules with respect to a target variable [28]. It aims to find subsets of the data with different distribution of the target variable [23].

In Predictive Clustering the mapping domain is the entire space of instances and is therefore generalized to instances that have not yet been seen, that is it constructs classes in an unsupervised manner after which the learned model can be applied to unseen data in the usual form [23].

Descriptive Clustering and Association Rules are two examples of descriptive modeling within an unsupervised learning scenario. Descriptive Clustering is about knowing a certain domain from which we want to separate into similar areas. It is only applied to data at hand [23]. Association rules are used to discover item sets that frequently appear in transactions together [23].

In this study, we will explore some descriptive modeling techniques in an unsupervised learning scenario, i.e. clustering and association rules.

2.2.2.1 Clustering

Clustering aims to discover "natural" groups in the data by grouping a singular set of objects based on characteristics joining the ones with given resemblance [2]. The process of clustering consists on, given a set of data points, create partitions that contain very similar data points [3].

There are several clustering algorithms that splits the data points according to specific distance metrics like Euclidean and Manhattan. However, in high dimensional spaces, when there is a lot of features, these metrics can become meaningless and uninformative in cluster analysis. Thus, it is important to assess the features that can be removed at the beginning of the clustering process [3].

The main clustering methods can be divided in to two types:

- partition - divides observations into n partitions according to a pre-specified maximization or minimization criterion (e.g. k-means algorithm [27]);

- hierarchical - generates a hierarchy group, from 1 to n groups, where n is the number of observations in the data set and each level represents a possible solution for grouping. This type can still be divided into two approaches, depending on how the groups of observations are merged or splitted. It can be:
  - agglomerative: where the hierarchy is generated from bottom to top (from n to 1 groups);
  - divisive: where the hierarchy is created from top to bottom (from 1 to n groups).

DBSCAN is a density based partition algorithm, whose objective is to discover clusters and noise in a spatial database [20]. The density of an object is calculated by the number of objects in its ε-neighbours of that object [34]. It will be discussed latter in subsection 3.3.4 as a spatial-temporal technique used in this thesis.

2.2.2.2 Association Rules

Association rules is a method that discovers a co-relation between variables of a data set [31] based on frequent pattern mining, which is an analytical process that finds frequent patterns or associations.

The Apriori [31] is the base algorithm used to generate association rules. Given a transaction database, this algorithm relies on two important metrics to find the rules: support and confidence.

Let X and Y be itemsets, $X \rightarrow Y$ an association rule and τ a set of transactions of a given database. The support (cf. Equation 2.4) measures the importance of a set in the context of all the transactions.

$$sup(X) = \frac{|t \in \tau : X \subseteq t|}{|\tau|} \tag{2.3}$$

the support of the itemset X related to τ is defined as a proportion of transactions t in the dataset which contains X.

The confidence (cf. Equation 2.4) measures the strength of the rule.

$$conf(X \rightarrow Y) = \frac{sup(X \cup Y)}{sup(X)} \tag{2.4}$$

the confidence of a rule $X \rightarrow Y$ regarding τ is the proportion of transactions that contains X and also contains Y.

The support threshold determines which are the frequent itemsets. Afterwards, these are used to generate association rules, based on the confidence threshold.

The Apriori algorithm works in two steps: frequent itemset generation, i.e. itemsets with support $\geq$ minsup; and rule generation, i.e. generate all confident association rules from the frequent itemsets, i.e. rules with confidence $\geq$ minconf.

The frequent itemsets generation, has itself two steps: a join step and a prune step. This algorithm uses the downward closure property in order to prune the candidate search space.

A set of items is considered to be frequent when it meets the minimum threshold required for support values. To find the most frequent itemsets in the database the algorithm makes quite a few searches where k-itemsets are used to generate k+1-itemsets (joining step). Each k-itemset must be equal to or greater than the minimum support threshold to be frequent. Otherwise they are called candidate itemsets[4]. Based on the found frequent itemsets, rules are generated, selecting those rules whose confidence is higher than minconf.

It is expected that, even with the support and confidence threshold, an exponential number of association rules are still generated. In this context, other metrics were proposed to allow the selection of rules based on a measure of interest of the rules. One such metric is the lift (cf. Equation 2.5), which compares the confidence and the expected confidence [12].

$$lift(X \rightarrow Y) = \frac{sup(X \cup Y)}{sup(X)sup(Y)} \tag{2.5}$$

If a lift value is greater than 1, it indicates that A and B appear more often together than expected. This means that the occurrence of A has a positive effect on the occurrence of B, i.e. the occurrence of A increases the likelihood of occurrence of B. If a lift value is less than 1 implies that indicates that A and B appear less frequently than expected together, i.e. the occurrence of A has a negative effect on the occurrence of B. If the lift value is approximately 1 indicates that A and B often appear together, i.e. the occurrence of A has no effect on the occurrence of B.

2.2.3 Spatio-temporal Data Mining Techniques

Most common data mining techniques are based on the assumption that data instances are independent and identically distributed (i.i.d). In environmental data this does not happen because instances are structurally related to each other in the context of space and time and show varying properties in different spatial regions and time periods. Ignoring these dependencies during data analysis can lead to precision and interpretability problems [5].

This type of data differs from relational data because besides the actual measurements or attributes, we still have to deal with the spatial and temporal attributes, which means it will also depend on temporal and spatial dimensions [5].

Spatio-temporal data contains both spatial and temporal attributes, and depends on which of the attributes is contextual or behavioral. If the contextual attribute is temporal while the spatial attribute is behavioral we may consider that we have a time series [3] which is a collection of sequential observations over a period of time, where the data order is fundamental. If both spatial and temporal attributes are contextual, we can see it as a direct generalization of spatial and temporal data. It is particularly useful when the behaviour of a given attribute is measured at the same time. For example, when considering variations in sea-surface temperature measured over time, temperature is a behavioral attribute while space and time are contextual attributes.

Dealing with spatio-temporal data is challenging. There is an increasing need to create or

adapt methods that can handle large spatio-temporal data sets and improve the spatial and temporal resolution of predictions. In [5] we find the reference to some techniques that allow us to analyze spatio-temporal (ST) data. They are outlined as follows.

- **Clustering:** for the process of grouping data set instances of ST data that share similar characteristics, there are several methods: Clustering Points, Clustering Trajectories, Clustering Time Series, Clustering Spatial Maps, Finding Dynamics ST Clusters and Clustering ST Rasters.
- **Predictive Learning:** for learning to map from the input features to the output variables using a representative training set, the objective is to predict future observations of the saptio-temporal data based on previous data. Some of the predictive learning techniques that are commonly encountered in ST applications are Time Series, Spatial Maps, ST Rasters and ST Reference Points.
- **Frequent Pattern Mining:** for the process of discovering patterns in a ST data set that occur frequently over multiple instances in a data set, we can use Co-occurrence Patterns in ST Points, Sequential patterns in ST Points, Sequential patterns in Trajectories, Motif Patterns in Time-Series and Network Patterns in ST Rasters. One algorithm we can use to do Frequent Pattern Mining is the Spatio-Temporal Co-Occurrence Pattern (STCOP). It finds a subset of events whose instances often co-occur together in space and time [6].
- **Anomaly Detection:** for detecting instances that are remarkably different from the majority of instances in the data set. In ST detecting anomalies or outliers can help identifying interesting but rare phenomena. Methods used in this setting are ST Point Anomalies, Trajectory Anomalies and Group Anomalies in ST Rasters.
- **Change Detection:** involves identifying the time point when the behavior of a system undergoes a significant deviation from its past behaviour. Change detection in time series has been studied extensively with the objective of determining time intervals (segments) that exhibit homogeneous properties.
- **Relationship Mining:** relationships among pairs of time series can be discovered using any of the similarity measures defined over time series instances.

To handle spatio-temporal data using clusters, we can use the ST-DBSCAN algorithm, which is based on another algorithm called Density-Based Spatial Clustering of Applications with Noise (DBSCAN). This algorithm is density based and is an improvement over DBSCAN due to 3 factors [8]: (i) can cluster temporal spatial data according to non-spatial, spatial, and temporal attributes; (ii) DBSCAN cannot detect noise points when there are clusters of different densities, this problem is solved in this new algorithm by assigning a density value to each cluster; (iii) the values at the opposite-end barriers can be quite different, if the non-spatial values of the neighbour objects have small differences and the clusters are adjacent; this is solved by comparing the average value of a cluster with the new coming value.

2.3 Tools

In this thesis we use R [33] programming language. R is a free and open-source programming language, it provides a computer environment that facilitates the manipulation, analysis and graphical display of the data. Has a wide range of statistical analysis, such as correlation analysis, and machine learning methods, such as clustering and classification.

To implement the framework for the dynamic exploratory analysis of data, we use the *Shiny* package [13]. It is an R package that makes it easier to create interactive web applications directly through R without having to have web development skills. These applications can be hosted locally or on a remote server. Our goal is to use this package to build a tool for dynamic visualization and statistical analysis of the data. We can achieve this by creating graphs for a better understanding of the data and make an analysis based on algorithms considered relevant for the study.

For plotting the graphs, we use *Plotly* [1] because it allows to make interactive plots, where is possible to zoom on it, download as an image, and by hovering the mouse over the graph it is possible to see the data values. For us this was the obvious choice because it is user friendly, allowing us to do everything we could do in *ggplot2* [40] but more interactive. We also use *pheatmap* which is a function that draws clustered heatmaps, using hierarchical clustering.

Chapter 3

Case Study on Arctic Oceanographic Campaign Data

In this chapter we describe our case study. In particular, we characterize our data set, explaining how and which data was collected. We also illustrate the data mining techniques applied to this data set, to gain useful insights in extracting relevant information from the data set.

3.1 Data set Characterization

Our data set is composed by 2 different components: (i) a biogeochemical data set generated from water samples collected at different but defined depths, making it possible to comprehend the concentrations of nutrients, such as nitrogen and ammonia; and (ii) an environmental in situ data set which generate depth profiles of different physical (e.g. salinity), biogeochemical (e.g. Oxygen) and biological (e.g. fluorescence) water masses characteristics. To generate this profiles it was used a CTD device, which stands for Conductivity, Temperature and Depth. The samples were either analyzed on the research vessel itself or in laboratories. These water samples were collected by using a Rosette with several Niskin bottles (8L) coupled with a Conductivity, Temperature and Depth (CTD) device with multiple sensors that preserve information, for example, on the water temperature at each depth. Figure 3.1 shows a photograph of a CTD taken during one of the expeditions.

The CTD is an oceanographic instrument and its main function is to detect how the conductivity and temperature of the water column changes regarding the depth. For the purpose of getting this parameters, CTD uses modular and interchangeable sensors and small probes that are attached to a large metal rosette wheel, which is lowered by a cable down to the seawater. It is possible to add other sensors to measure extra parameters such as ph, dissolved oxygen, turbidity and fluorescence. This device grants the opportunity of having data recorded every second, giving a lot to explore [21].

The integration of biological, chemical and environmental components gives support to better figure out how these variables are interconnected and the influence of a certain variable in all

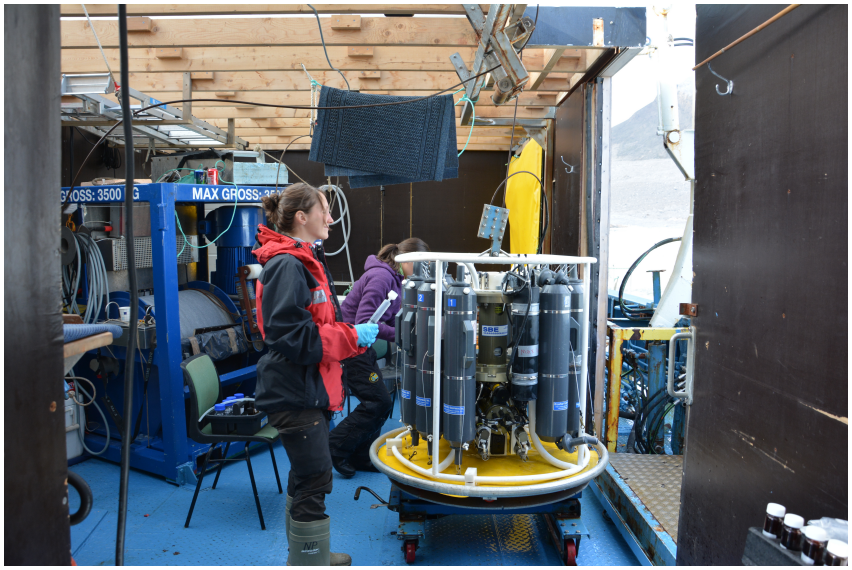


Figure 3.1: Photograph of a CTD taken during one of the expeditions.

relevant compartments. This includes the different trophic levels, which are groups of organisms within an ecosystem that occupy the same level in the food chain [19].

The samples that form our data set were collected during the arctic expeditions that took place between the years 2016 and 2019, and are composed of mainly environmental and biogeochemical data, but also includes one biological variable (fluorescence). During the expeditions (MOSJ-ICE2016, MOSJ-ICE2017, MOSJ-ICE2018 and MOSJ-ICE2019), the samples were taken from different oceanographic stations that constitute the two transects (Kongsfjorden and Rijpfjorden), as shown in Figure 3.2. The composition of each transect is defined in Table 3.1.

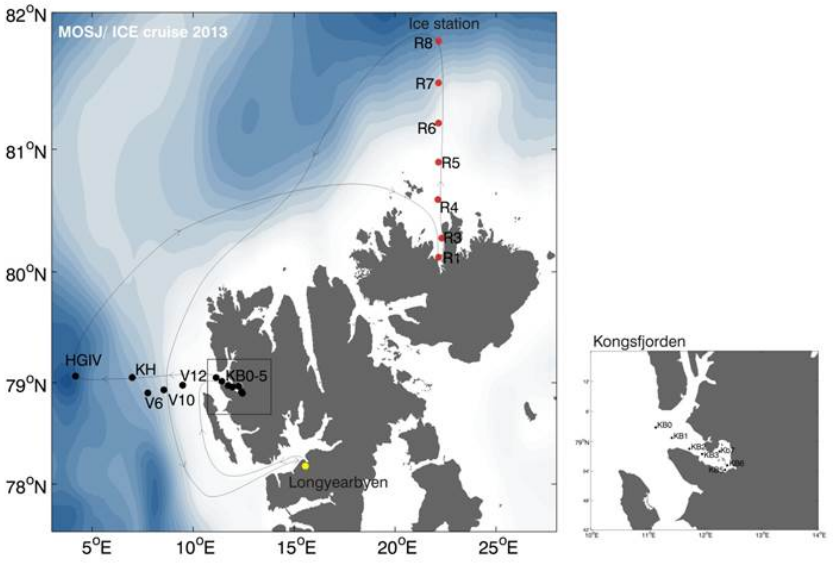


Figure 3.2: Location of the stations from which the data was collected.

During the MOSJ monitoring program it is not always possible to collect samples from all the stations in every campaign, related to logistic causes. It may not be possible to visit some

Table 3.1: Transects and respective oceanographic stations.

Transect Kongfjorden	**Transect Ripfjorden**
KB0	R1
KB1	R2
KB2	R3
KB3	R4
KB4	R5
KB5	R6
KB6	R7
KB7	R8
V6	-
V10	-
V12	-

stations every year, either because they are covered with ice or because it is necessary to skip some stations due to limited time issues. For these reasons, not all stations have the same amount of information. This situation is shown in Table 3.2 where we clarify which stations were sampled in each campaign.

Table 3.2: Oceanographic stations where data was collected for each year, from 2016 until 2019.

2016	**2017**	**2018**	**2019**
HGIV	HGIV	KB0	HG1
KB0	KB0	KB1	HGIV
KB3	KB1	KB2	KB0
KB6	KB2	KB3	KB1
R1	KB3	KB5	KB2
R4	KB4	KB6	KB3
R6	R1	KB7	KB5
R7	R2	V6	KB6
V6	R3	V10	KB7
V12	R4	V12	KH
-	R5	-	V6
-	R6	-	V10
-	R7	-	V12
-	V6	-	-
-	V10	-	-
-	V12	-	-

For all samples collected in the 20 oceanographic stations, regardless of the year in which they were taken, we obtain more than one file with the information received by the CTD. For the same stations we have multiple files because several CTD's are made per station to be able to collect enough water samples for the various parameters that will be analyzed in the laboratory.

The environmental data refers to data that characterizes the seawater, such as temperature

and salinity. The biogeochemical/chemical data includes the compounds that can be used or produced by the biological compartment (e.g. nutrients) and the biological data includes the parameters that give information about the biological communities (e.g. fluorescence). Our data set is composed by the data collected through the CTD, which can collect information regarding the water column profile, providing knowledge at all different depths from the surface until long depth distances, and will help contextualize the rest of the data that is not possible to be collected at the same rate. The parameters acquired are described in Table 3.3.

Table 3.3: Environmental data collected by the CTD

Parameter	**Units of measure**
Pressure	$Digiquartz, db$
Potential Temperature	$ITS-90, ^{\circ}C$
Salinity	PSU
Density	Kg/m^3
Fluorescence	mg/m^3
Oxygen	ml/l
Temperature	$ITS-90, ^{\circ}C$
Photosynthetically Active Radiation (PAR)	-
Surface Irradiation (SPAR)	$Einsteins/m^2/sec$

We have a file with more summarized information consisting of a table indicating the values found per parameter whether it is nutrient information or data collected by the CTD found at four main depths: surface (5m), 25m, 50m and bottom, which varies from station to station. These 4 depths are not the only ones for which data are collected but they are the ones that are always considered. This is because at the surface (5m) is where the water mass normally has different characteristics in relation to other depths, since it is in contact with the atmosphere, suffering great influence from the winds and temperatures. At 25 and 50m is usually where the largest amount of phytoplankton (primary ecosystem producers) is concentrated. Lastly it is also used information collected at the bottom because it is the area of the water column that is closest to the sediment and in this sense has higher concentrations of nutrients and less oxygen, since it is influenced by mineralization processes (microbial processes in which organic matter is converted into nutrients and other inorganic constituents). Table 3.4 provides a glimpse of environmental and chemical data for KB0 station, at the four main levels of depth (Surface, 25 meters, 50 meters and Bottom) for the year 2016.

3.2 Data Pre-Processing

We started by pre-processing the data collected during the arctic expeditions throughout the years of 2016 until to 2019. This is an important step in the data mining process, because as mentioned before, if there are too many missing values, irrelevant or redundant values it may

Table 3.4: Extract of environmental and chemical data for KB0 station, at the four main levels of depth (Surface, 25 meters, 50 meters and Bottom) for the year 2016.

Sample Code	**KB0_R1_S**	**KB0_R1_25m**	**KB0_R1_50m**	**KB0_R1_B**
Sample Depth (meters)	5	25	50	320
NGS_code	k1	k2	k3	k4
NH4+[μM]	0.43	0.82	1.06	0.28
NO2- [μM]	0.06	0.06	0.11	0.45
NO3- [uM]	0.0	1.3	1.5	10.7
PO4- [uM]	0.09	0.15	0.19	0.87
Si(OH)4 [μM]	0.9	1.4	1.5	6.3
Chl [mg/ m3]	0.15	0.77	0.52	1.34
Phaeopigment [mg/ m3]	0.26	0.63	0.43	1.27
Temperature [ITS-90.° C]	8.1147	5.7402	4.8304	1.8655
Salinity. Practical [PSU]	33.2313	34.4985	34.768	34.9376
Fluorescence.Wetlab ECO-AFL/FL [mg/m^3]	0.9532	1.0469	0.4556	-0.0805
Oxygen. SBE 43 [ml/l]	2.49831	2.47481	2.46697	2.5131
PAR/Irradiance.Biospherical/Licor	8.65	0.47334	0.02422	1E- 12
SPAR/Surface Irradiance	49.521	53.253	54.076	59.425

affect the final result. The main objective of this step is to prepare the data for exploratory analysis, basically we transform raw data into a format that is perceptible and useful for the analysis that we are going to do next.

First, we have to understand the variety of data and what type of data we have for each year and station and how they can or cannot be interpreted by R [33].

The files we had ended up in a .cnv extension and turned them into a file with a .csv extension. Each file ended up with 13 columns whose names are Depth, Potential_Temperature, Salinity, Density, Fluorescence, Oxygen, Temperature, PAR_Irradiance, SPAR_Surface_Irradiance, flag, Year, CTD and finally Station. The number of lines depends on the number of observations existing for each station. In Table 3.5 we have an example of a file created from one of the CTD files for KB0 station where we exemplify the first entries in the file. In Figures 3.3 and 3.4 we have two graphs that show the number of observations we have per year and per station, respectively. These graphs serve as an illustration to understand how the weight of each station and each year is distributed when we perform analyzes in greater detail.

In this new data set, we removed information such as which sensors collected the data, as it does not matter for the analysis we are going to do. We also eliminated the maximum and minimum values for each specific parameter because it is something that we can obtain using formulas in R. In some files we have data collected during the descent and ascent of the CTD, but this only happens in one of the years for which we have information. Taking into account that the values are quite similar, it ended up being redundant analysis and, therefore, we removed

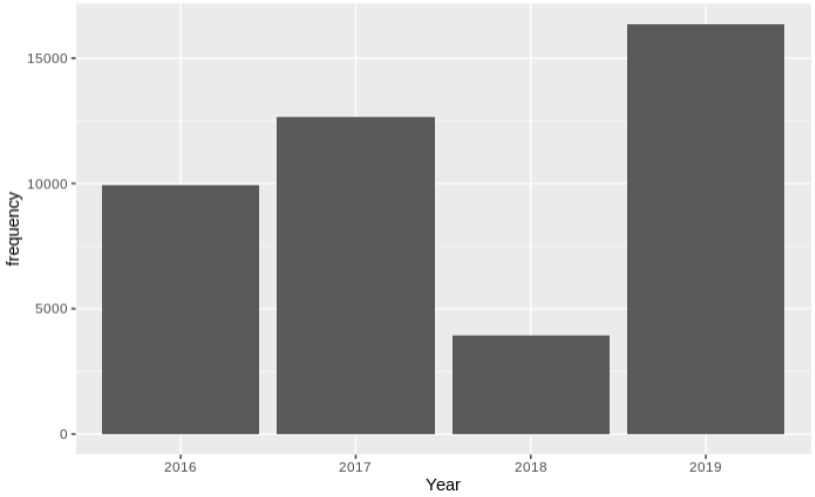


Figure 3.3: Number of observations per year for all stations.

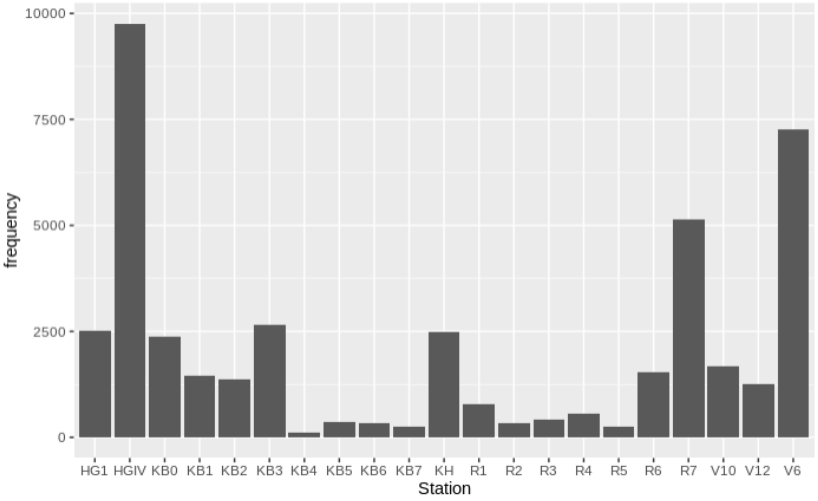


Figure 3.4: Number of observations per station for the 4 years.

these data and only consider data collected when the CTD goes down.

Table 3.5: Extract of the data set created from the CTD for KB0 station in 2016.

Depth	Potential_Temperature	Salinity	Density	Fluorescence	Oxygen	Temperature	PAR_Irradiance	SPAR_Surface_Irradiance	flag	Year	CTD	Station
2.000	7.3473	30.6148	23.9191	1.0287	2.43251	7.3474	1.7963e+01	5.2822e+01	0.0000e+00	2016	ctd_Kb0_1_2016	KB0
3.000	8.1570	31.9286	24.8395	0.9653	2.46291	8.1572	1.2559e+01	5.0410e+01	0.0000e+00	2016	ctd_Kb0_1_2016	KB0
4.000	8.1667	32.9408	25.6324	0.7850	2.44572	8.1671	9.2346e+00	4.9401e+01	0.0000e+00	2016	ctd_Kb0_1_2016	KB0
5.000	7.5006	33.1175	25.8668	0.7180	2.45293	7.5010	7.0239e+00	5.0590e+01	0.0000e+00	2016	ctd_Kb0_1_2016	KB0
6.000	6.5401	33.1386	26.0139	0.9442	2.44222	6.5406	5.7894e+00	5.5020e+01	0.0000e+00	2016	ctd_Kb0_1_2016	KB0
7.000	6.0042	33.2556	26.1737	1.3316	2.42859	6.0048	4.6883e+00	4.9803e+01	0.0000e+00	2016	ctd_Kb0_1_2016	KB0
8.000	6.8602	33.4312	26.2008	1.4434	2.36376	6.8609	3.9341e+00	4.9474e+01	0.0000e+00	2016	ctd_Kb0_1_2016	KB0
9.000	7.7408	33.7374	26.3202	1.1924	2.39367	7.7417	3.1377e+00	5.2657e+01	0.0000e+00	2016	ctd_Kb0_1_2016	KB0
10.000	7.8704	33.8076	26.3564	1.1873	2.41929	7.8714	2.6536e+00	5.6481e+01	0.0000e+00	2016	ctd_Kb0_1_2016	KB0

After creating these new files we loaded them into R and started to do some analysis, such as, seeing the maximum and minimum overall value for each parameter. This allowed us to realize that not all data collected starts or ends at the same depth level regardless of whether it was the same station or the same year.

We also realized that the years 2018 and 2019 do not have all the parameters. For 2018 we do not have information regarding the PAR_Irradiance parameter, and for 2019 we do not have the data for SPAR_Surface Irradiance. In the year of 2019 the sensors used were different and that is why in the Depth parameter there are more fine-grained values like 5.937 which differs from the other observations that are discrete values like 5.000.

During the initial analysis we conclude that the potential temperature parameter would not be very interesting for the rest of our analysis. This is because it represents the predictable temperature for a certain seawater element in which the salinity does not change in relation to a certain depth. And, since the values obtained are similar to those we have for the temperature parameter, we decided to exclude it from further analysis. Figure 3.5 represents the similarity described above.

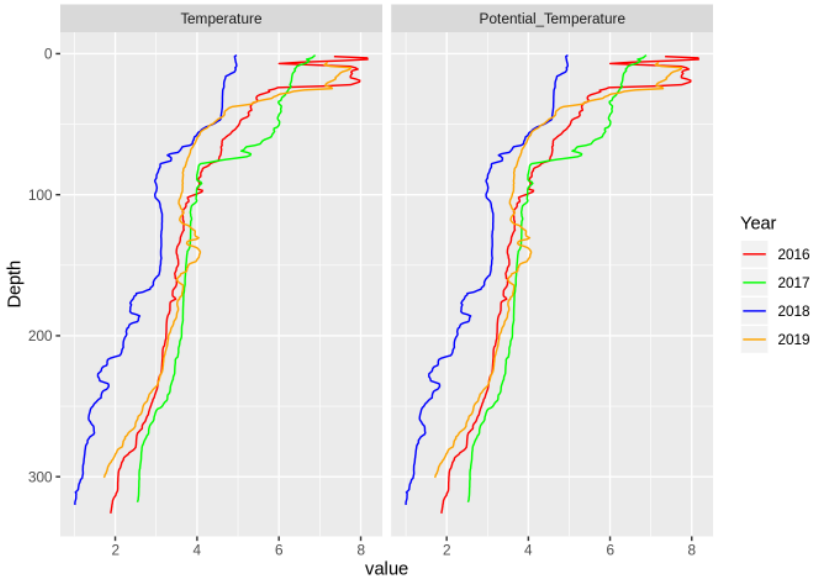


Figure 3.5: Temperature and Potential Temperature parameters for KB0 station for all years.

The parameter flag has multiple uses like for verifying data quality [7] but doesn't give us relevant information for our analysis, so we decided not to use it.

As for the Photosynthetically Active Radiation (PAR) and Surface Photosynthetically Active Radiation (SPAR) parameters, they give us the same information in the same unit of measure ([m$Einsteins/m^2/sec$]). Basically, there is a sensor that is coupled to the CTD and measures the radiation of the solar spectrum between 400-700nanometers, which is used by the primary producers during the process of photosynthesis [11]. These parameters give us an indication of whether the conditions in terms of luminosity are optimal for primary production or not.

After some investigation, we concluded that the PAR parameter is the one which is more applied when comes to analysing the CTD data, so this is the one we will be using for future analysis.

Fluorescence is an optimal phenomenon that occurs when light is absorved by a material and creates a molecular excitation that causes the material to re-emit light as a different

wavelength, the parameter from the CTD measures indirectly the pigment concentration. This is a very important parameter because it gives relevant biological information in terms of primary producers biomass. In Figure 3.6 we have the comparison between the parameters Temperature, Salinity and Fluorescence for KB3 station, for all years under analysis. Thus, we have to explore more on how both parameters change the fluorescence throughout the years or the stations.

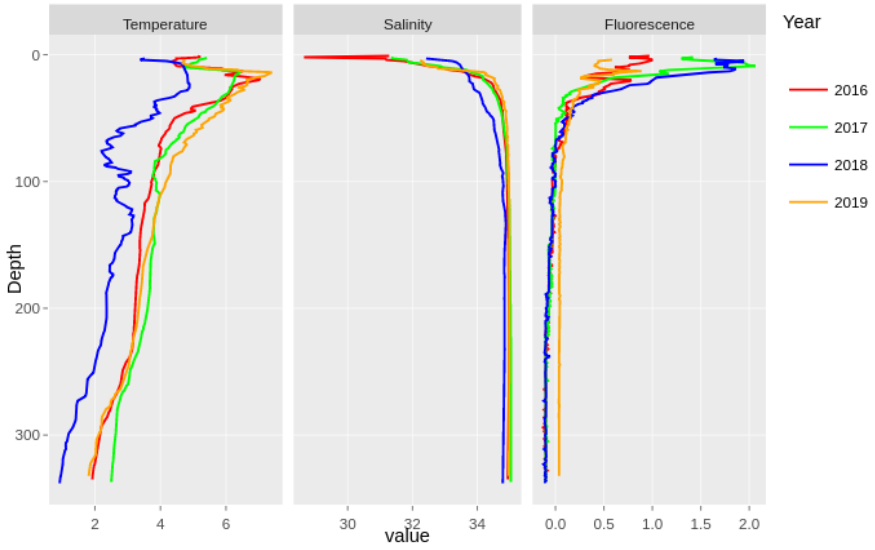


Figure 3.6: Temperature, Salinity and Fluorescence parameters for KB3 station for all years.

The density parameter informs us about how thick is a certain seawater mass. This parameter is not relevant to explain the biological and biogeochemical parameters distribution, and it is also possible to obtain it through other parameters calculated from the CTD data, based on temperature, salinity and depth measures.

Oxygen is also a parameter which is affected by the temperature, salinity and biological activity, and has also an influence on the biological life on those waters.

Then we are left with the parameters Depth, Temperature, Salinity, Oxygen, Fluorescence and PAR Irradiance. These are parameters that we are interested in evaluating because some are related to each other or influence the environment in some way. An example of a plot generated for all parameters for the KB0 station during the years 2016 to 2019 is represented by the Figure 3.7

In some other files we had the coordinates of the stations from which the data were taken, and so we decided that we should create a map. To be able to make the map we first had to adapt some of the coordinates because they were not all in the same format. Part of the coordinates were in the form of degrees, minutes and seconds (DMS) and we needed them to be in a decimal form in order to use them for the leaflet map3.8.

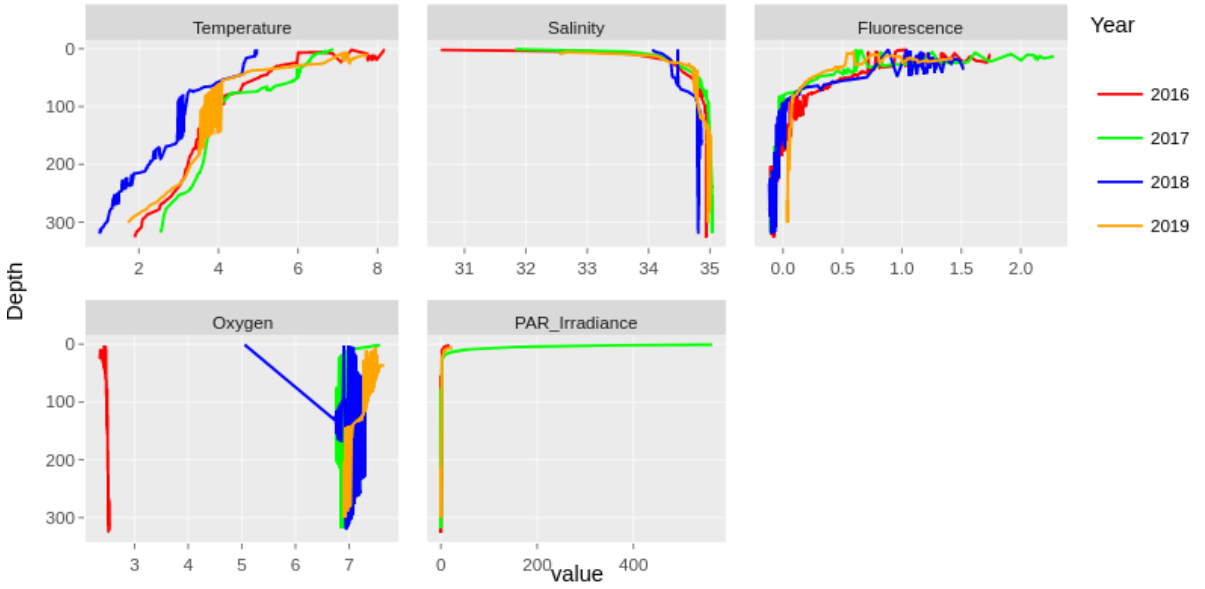


Figure 3.7: Temperature, Salinity, Fluorescence, Oxygen, PAR Irradiance parameters for all years for KB0 Station.

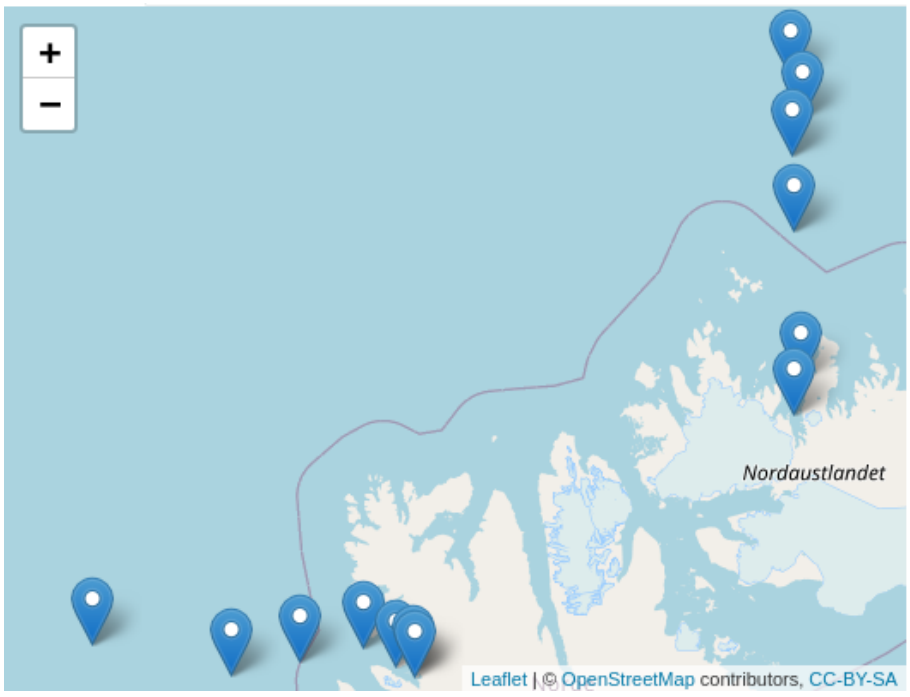


Figure 3.8: Map created for the year of 2016.

3.3 Data Analysis

After the pre-processing phase, we began analysing the data. This analysis was mainly focused on a visual form, using graphs because we are interested in understanding the pattern of temporal and spatial evolution of each parameter, how each parameter is influencing others and what kind of relationships there are among the different parameters throughout the years. In resume,

we want to inspect the spatial and temporal patterns and how they relate to one another, for example, we want to see if there are any changes when analyzing the information related to a station throughout the years. If there is, we want to understand maybe why that happened and how may influence others.

3.3.1 Temporal Analysis

The purpose of the temporal analysis was to have the possibility of choosing different years and a station for which we could analyze how it changes or not over the chosen years.

In this analysis we perform two forms of visualization. In the first one, we see how the parameter values change for a given station for the years we choose to analyze. As a second option, we have a graph that represents the chosen CTDs, thus being able to perceive for each of the years if there were changes in the values and making comparisons with other chosen CTDs. This will enable to verify if changes in a certain parameter occurred in one station or during all the years.

We decided to identify stations for which we have information regarding the four years under investigation, in this case the stations with this information available were only KB0, KB3, V6 and V12. Then, we selected one of these stations. We choose V6 because it was the one with highest depth values.

Figure 3.9, represents the first option of visualization described above, where we can see the aggregation of the CTDs chosen for this station by year. We can see that the values are similar to each other except for the oxygen parameter for the year 2016.

Figure 3.10 illustrates the second form of analysis. It is possible to see how those stations have evolved over time by analyzing its CTDs. In this case we have chosen for each station and for each year a CTD resulting in 16 samples. The previous analysis was verified in this plot because we see the evolution of the values taken from the CTD which maintains the same trend shown.

3.3.2 Spatial Analysis

In order to understand the evolution of the parameters described above for the different stations we decided to do an analysis were we were able to choose which stations we wanted to compare. From the resulting graph, we can understand if there are similar values between stations or if they are different. This can give us information about the spatial differences of a certain parameter in the target region.

In Figure 3.11 we present the graph obtained for stations in Kongfjorden transect which are close to the coast. We can see that the values are similar. As we can see on the map (cf. Figure 3.2), stations KB3 and KB6 are near to each other and KB0 is a little further away. As for stations V12, located in an ocean area outside of the fjord, we can see that the values

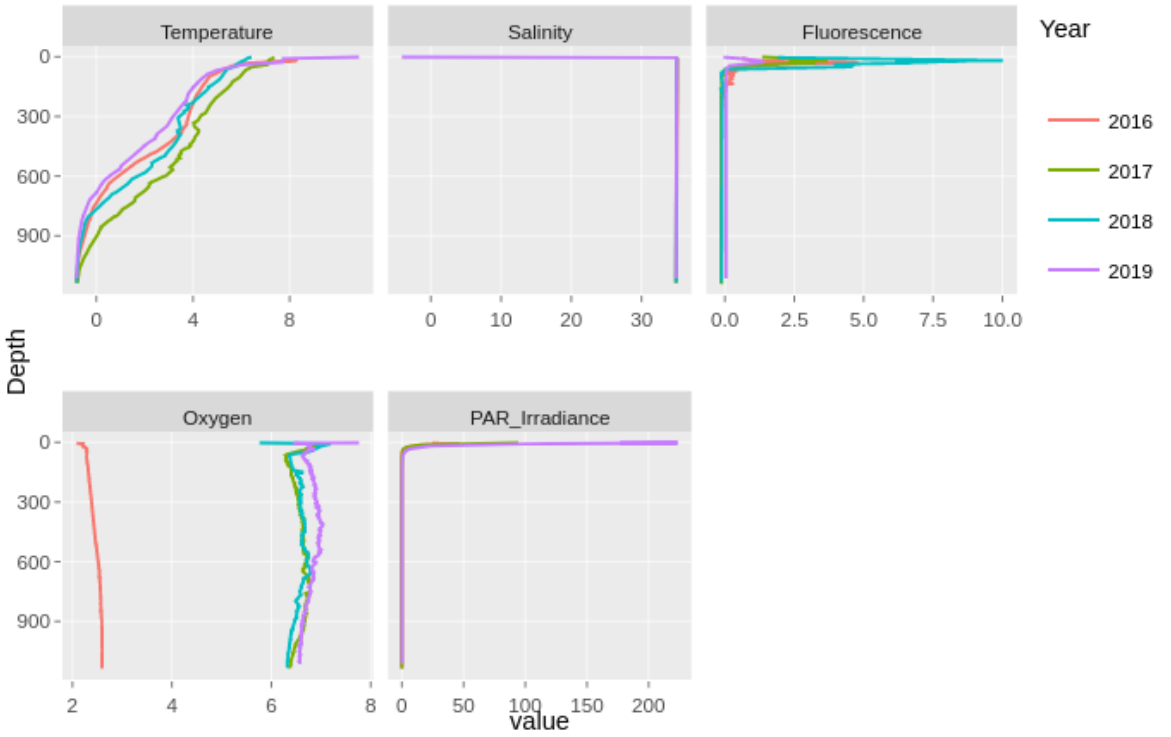


Figure 3.9: Analysis for the four years under study, for V6 station.

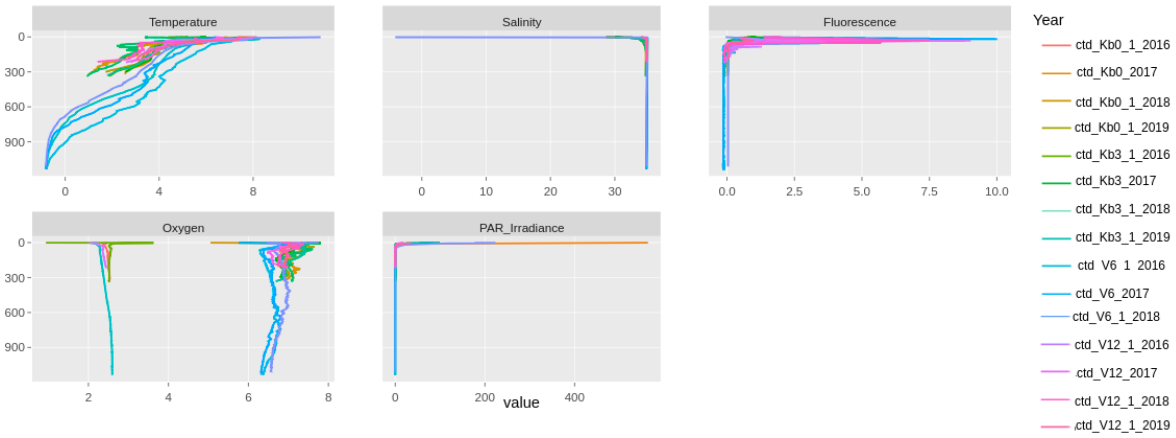


Figure 3.10: Evolution of the CTDs over the four years for stations KB0, KB3, V6 e V12.

presented a greater variation in the Fluorescence parameter.

On the other hand, and still regarding the year of 2016, we have the other transect, in Figure 3.12. Here we are able to notice that we have an opposite relationship. The stations that are further off the coast have similar values like R1, R4 and R6. In contrast R7, which is a more oceanic station, has the most disparate values in relation to the others.

In Figure 3.13 we have the comparison between stations from both transects and we can see that stations belonging to transect Kongfjorden it have very distinct values, mainly in the Temperature and Oxygen parameters, when compared with the other stations under analysis of transect Ripfjorden. We have higher values for the temperature and lower oxygen values at

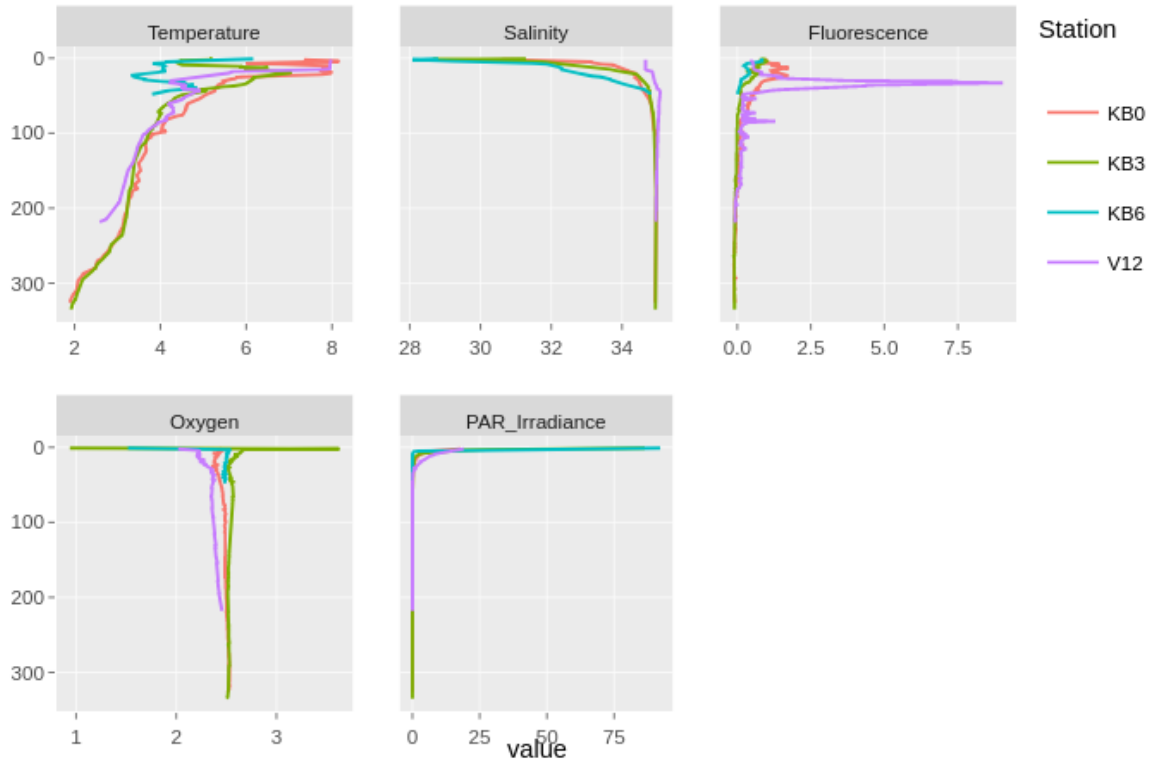


Figure 3.11: Comparison between stations KB0, KB3, KB6 and V12, for the year of 2016, for Temperature, Salinity, Fluorescence, Oxygen, PAR Irradiance parameters.

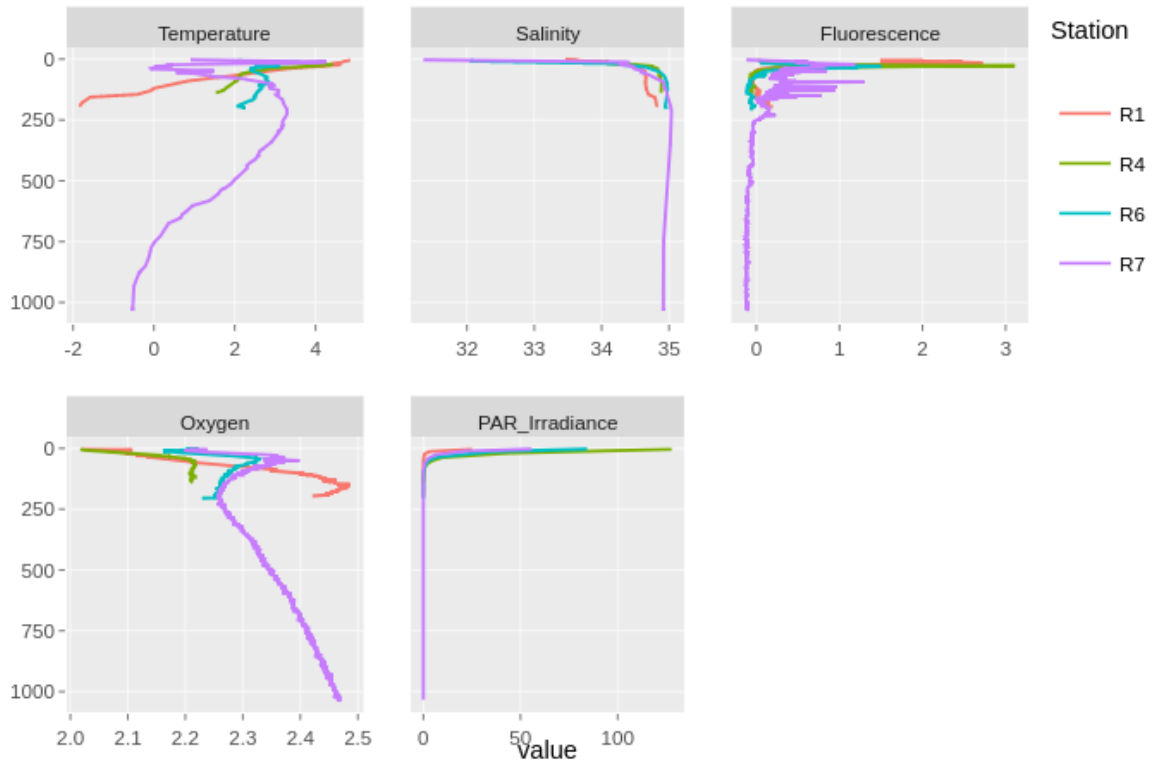


Figure 3.12: Comparison between stations R1,R4,R6, and R7, for the year of 2016, for Temperature, Salinity, Fluorescence, Oxygen, PAR Irradiance parameters.

stations KB0 and KB3 when compared to stations R2 and R3 in particular.

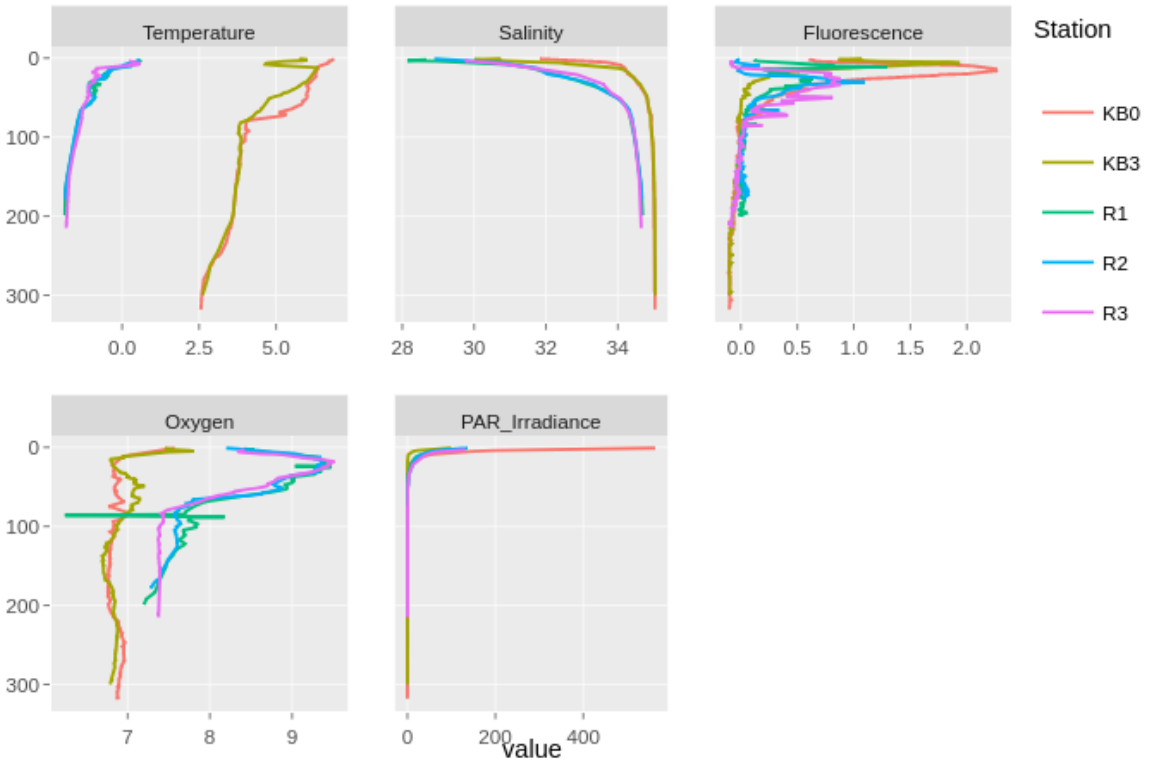


Figure 3.13: Comparison between some stations from both transects, for the year of 2017, for Temperature, Salinity, Fluorescence, Oxygen, PAR Irradiance parameters.

3.3.3 Statistical Analysis

Through some statistical analysis our goal was to check if there are relations between stations of the same transect or even between different transects. To achieve this we did a correlation analysis. Correlation is used to evaluate the relationship between two or more variables. We also decided to define a set of depths for which it would make sense to perform the analysis. Initially, we started analyzing according to intervals but this quickly became unfeasible. All CTDs have different maximum values for the depth parameter and for the year 2019 as the sensor was different from what used to be used, the depth turned out to be a little more fine-grained, such as 5.168m instead of 5.00m and it ended up not being able to do the best management of these depths. This fact can be verified by the histogram of the Depth variable for all CTDs shown in Figure 3.14.

In this context, we decided to choose the depths that would be more relevant for the study, which are:

- Surface - where all values will be considered where the Depth parameter is less than 6.
- Maximum Fluorescence - where we select the 10 values where the Fluorescence value is maximum. The ten depth values will not be continuous because the fluorescence depends

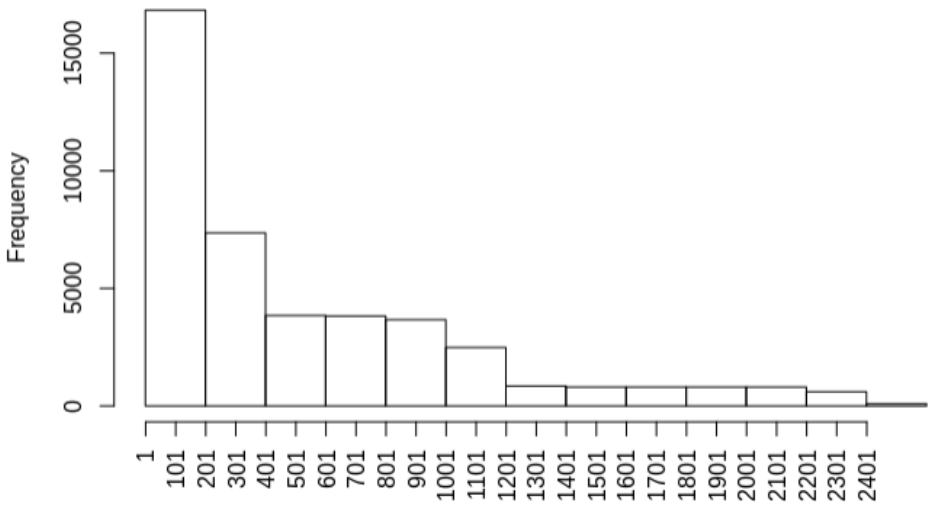


Figure 3.14: Distribution of the frequency values for Depth parameter.

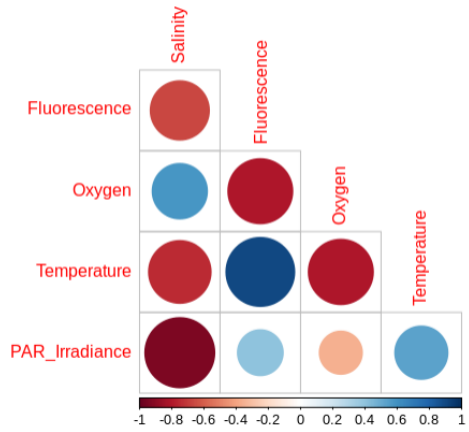
on how the sun propagates in the water.

- Bottom - here we picked the 10 highest values for the depth parameter to be considered.

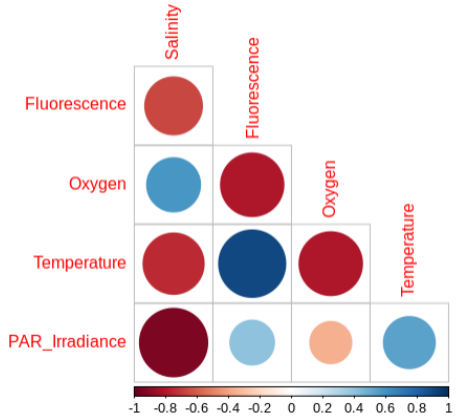
After choosing one of these possibilities, the values are collected and then the average for each parameter is made and used to calculate the correlation between variables because when making correlations we have to have the same number of observations for all parameters. In this analysis, we use one of two correlation measures: *Pearson* or *Spearman* coefficients. As stated in 2.2.1, these are the most commonly used to deal with this type of data [30].

In Figure 3.15 we have the analysis of the correlations using *Pearson* coefficient between the parameters for the years 2016, 2017, 2018 and 2019 for KB0 station. For 2016, the negative correlation that stands out the most is between the PAR_Irradiance parameter and Salinity, which also occurs in the years of 2017 and 2019, except 2018, since we do not have information about the PAR_Irradiance parameter for this year. This might have occurred due to the influence of higher salinity water masses at higher depths where fluorescence and light penetration is limited. As for the positive relationships, what stands out most for 2016 is between the Temperature and Fluorescence parameter. This positive correlation happens every year for this station, which is supported by the context in which we know that an increase in temperature may stimulate the growth of phytoplankton communities and thus the fluorescence in the water.

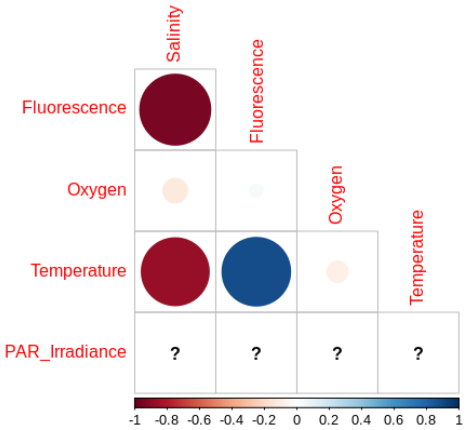
We have also performed a correlation analysis of the parameters separately for each transect. Results are shown in Figure 4.9. Regarding the transect Kongfjorden (cf. Figure 3.16a) and, in particular, the Temperature and Fluorescence parameters, we can see that they are positively correlated. This means that when the temperature increases, the fluorescence values also increase. In Ripfjorden transect (cf. Figure 3.16b), the Temperature and Oxygen parameters are negatively influenced by each other, that is, when the temperature increases, the oxygen decreases and vice versa. The PAR_Irradiance and Salinity parameters also have a negative influence, despite not being such a negative correlation.



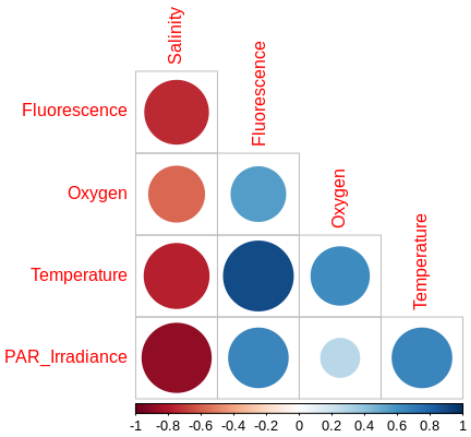
(a) Year 2016.



(b) Year 2017.

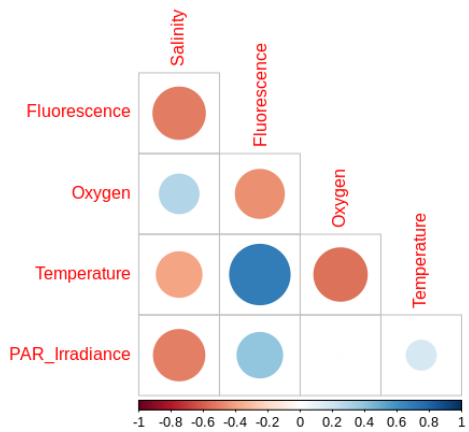


(c) Year 2018.

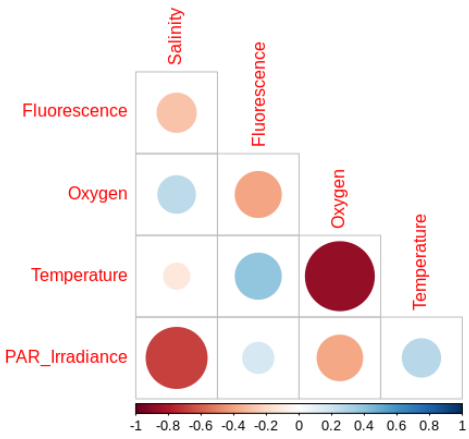


(d) Year 2019.

Figure 3.15: *Pearson* correlation between Temperature, Salinity, Oxygen, Fluorescence and PAR_Irradiance parameters for KB0 station, by year.



(a) Transect Kongfjorden.



(b) Transect Ripfjorden.

Figure 3.16: *Pearson* correlation between Temperature, Salinity, Oxygen, Fluorescence and PAR_Irradiance, by transect.

3.3.4 Descriptive Modeling

We also created an hierarchical clustering by calculating the distance which is computed using all complete pairs of observations on the variables chosen, and using the correlation measure selected. All values were normalized by columns because as the values vary a lot, we guarantee a normalization with a mean of 0 and standard deviation of 1.

An example of this analysis can be seen in Figure 3.17 using the *Pearson* correlation coefficient and in Figure 3.18 using *Spearman* correlation coefficient. These graphic was generated using `pheatmap`, which also created the hierarchical cluster. The hierarchical cluster is constructed by using `hclust`. This function builds the hierarchical cluster using a set of dissimilarities for the n objects that will be organized by clusters [35]. The matrix of the data is represented using the mean values in both plots. As we can see there are differences between both measures. *Spearman* correlation works with rank-ordered variables instead *Pearson* correlation works with raw data values.

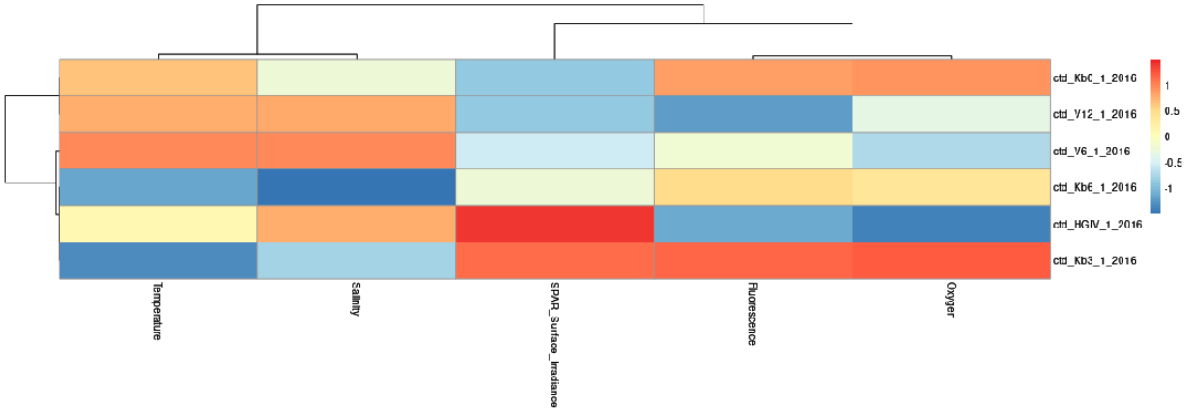


Figure 3.17: Hierarchical clustering for Transect Kongfjorden using *Pearson* correlation coefficient and a depth level of Maximum Fluorescence.

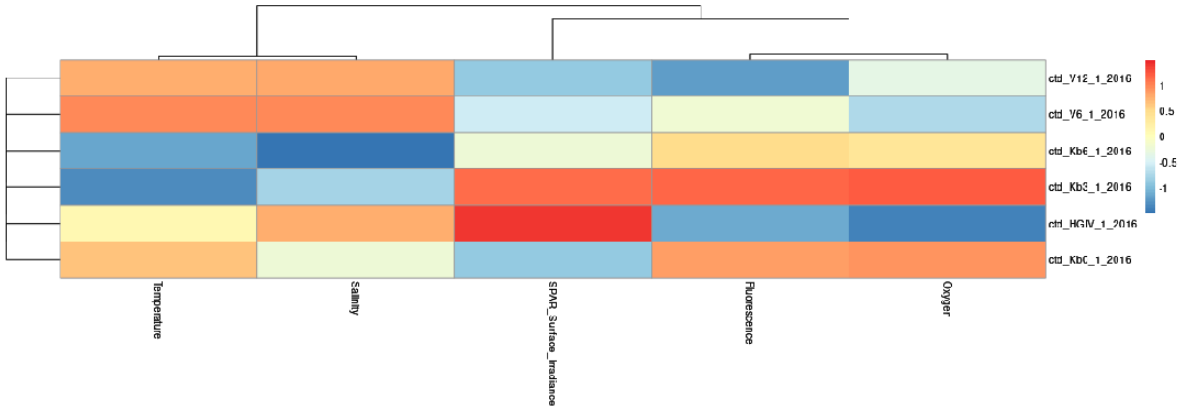


Figure 3.18: Hierarchical clustering for Transect Kongfjorden using *Spearman* correlation coefficient and a depth level of Maximum Fluorescence.

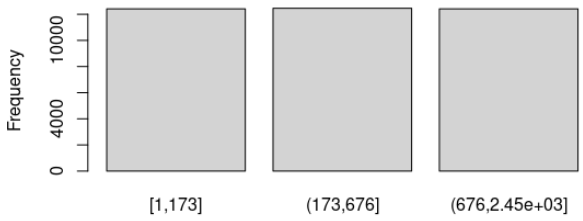
For cluster analysis, we tried to use the ST-DBSCAN [8] algorithm, but we realized that due to the low spatio-temporal granularity of our data set, the results obtained were not meaningful.

Firstly, we started by using a data set that only contained the coordinates of the stations for one of the years. With this, we realized that the places where we thought the clusters would be placed were not where they were. This led us to create a new data set in which we include latitude and longitude for all stations for all years. This time, and as shown in Figure 3.19, we have already managed to get closer to the locations of the stations, it resembles a little more with the expected locations of the stations, however we still do not have enough data to achieve the expected representation.

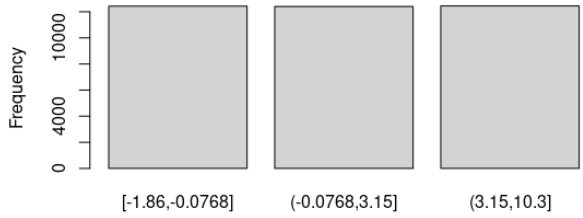
Figure 3.19: Map of the clusters created with all the locations for all stations by ST-DBSCAN.

Regarding association rules, we had to adapt our data. All of our parameters (Depth, Temperature, Salinity, Fluorescence, Oxygen, Par_Irradiance) needed to be described in intervals and these intervals also have to be similar, because we need to merge the values so we could make comparisons between the frequency of different items. We divided the values into intervals of equal frequency. Figure 3.20 represents the histograms generated for each parameter and the respective bins.

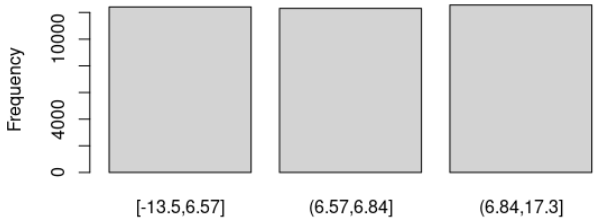
As we have anticipated, a large set of rules was generated using the apriori algorithm [4] with the default support and confidence values, 0.1 and 0.8, respectively, resulting on a total of 105 rules. After generating the rules, we use a function called `is.redundant` [29], that looks for redundant rules, returning which are. A rule is redundant if there is a more general rule with the same or greater confidence value. So we are left with 75 rules. Figure 3.21 shows how



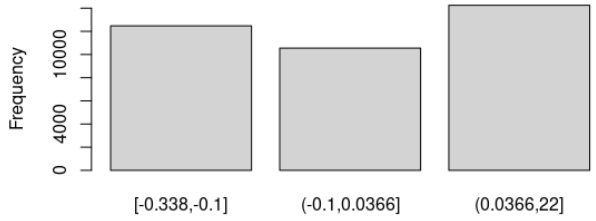
(a) Depth



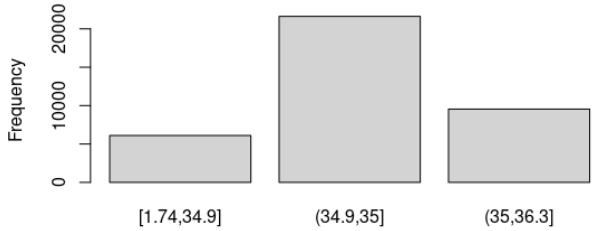
(b) Temperature .



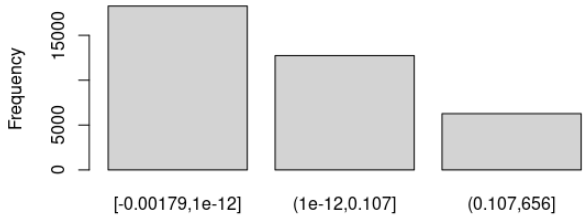
(c) Oxygen



(d) Fluorescence



(e) Salinity



(f) PAR_Irradiance

Figure 3.20: Equal-frequency histograms of each parameter.

the 75 resulting rules are distributed by support, confidence and lift measures. In appendix A, the 75 rules are presented in Tables A.1, A.2 and A.3 ordered by the lift measure.

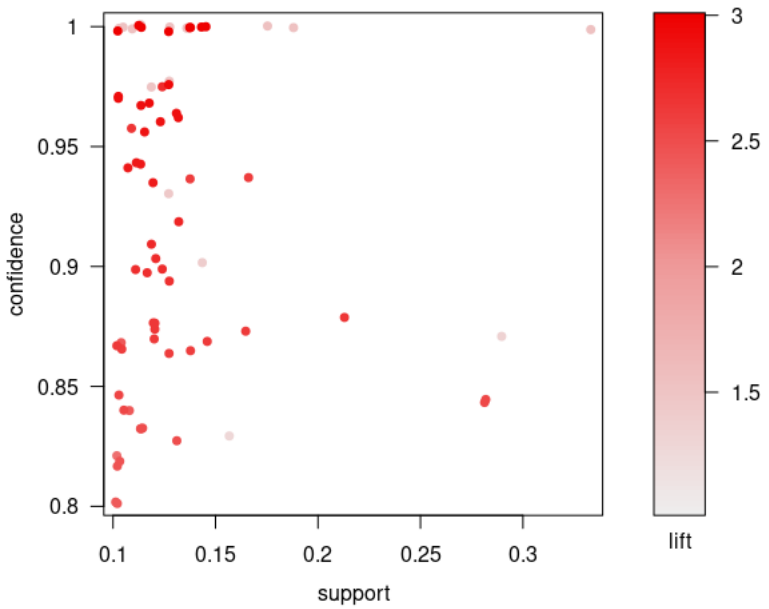


Figure 3.21: Association rules by support, confidence and lift measures.

With these rules, our main objective was to understand if environmental variables (e.g. oxygen, salinity, etc.) could influence the distribution of phytoplankton biomass, which is an important component of the system since it functions as the base of the trophic chain.

To accomplish this we decided that it would be more interesting to filter a specific interval on the Fluorescence parameter. This interval is where this parameter reaches the maximum values. We made this choice because it is in this interval that the biomass from primary producers is higher, which is an issue that we know to be relevant in biological terms.

To discover this range we had to look for the 10 maximum values of fluorescence for each of the stations. We concluded that the interval is fixed at [0.149, 22]. In the histogram with equal frequency distribution, the corresponding interval is [0.0366, 22]. These intervals do not have a uniform density, most of the values are on the lower side of the range and there are only a few higher values.

The resulting rules add up to a total of 10 (cf. Table A.4) after this filtering, and are shown in the graph of Figure 3.22.

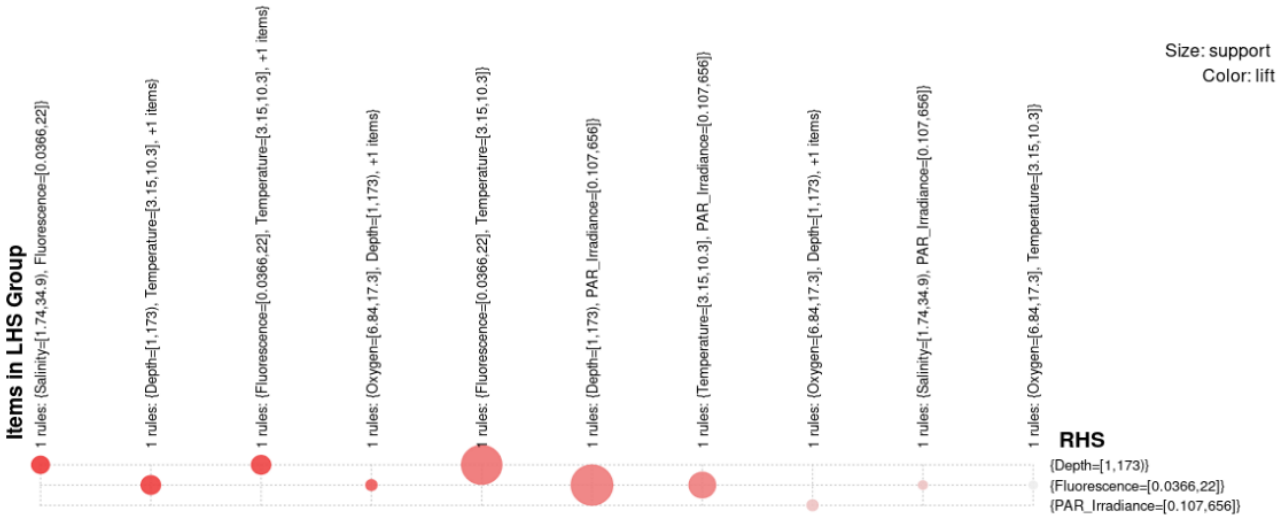


Figure 3.22: Association rules filtered by the 10 maximum values of fluorescence for each of the stations.

Chapter 4

Shiny App

The main objective of this thesis is to build a framework that allows to dynamically analyze and visualize the data collected during the Arctic expeditions that took place during the years of 2016 to 2019 through the monitoring program of Svalbard region (MOSJ). The shiny app was created with the additional purpose of having a user friendly interface. In this chapter we are going to show how the application is organized, and how it is possible to perform the analysis and visualization shown in the previous chapter, through the shiny app.

4.1 Presentation

When the shiny application is launched we can see a set of tab panels at the top of the page. These panels are divided into Interactive Map, Spatial Analysis, Temporal Analysis, Statistical Analysis and Descriptive Modeling.

The first thing we see is the first tab related to the map (cf. Figure 4.1). This turns out to be the first panel because it is an essential component. During the data exploration process it is crucial to have the map always available so that if there is any doubt as to the location of a station in a given year, it is always available.

This map depends on the year chosen, because for different years the latitude and longitude coordinates for each station available also change, they are similar but still there is a slight discrepancy.

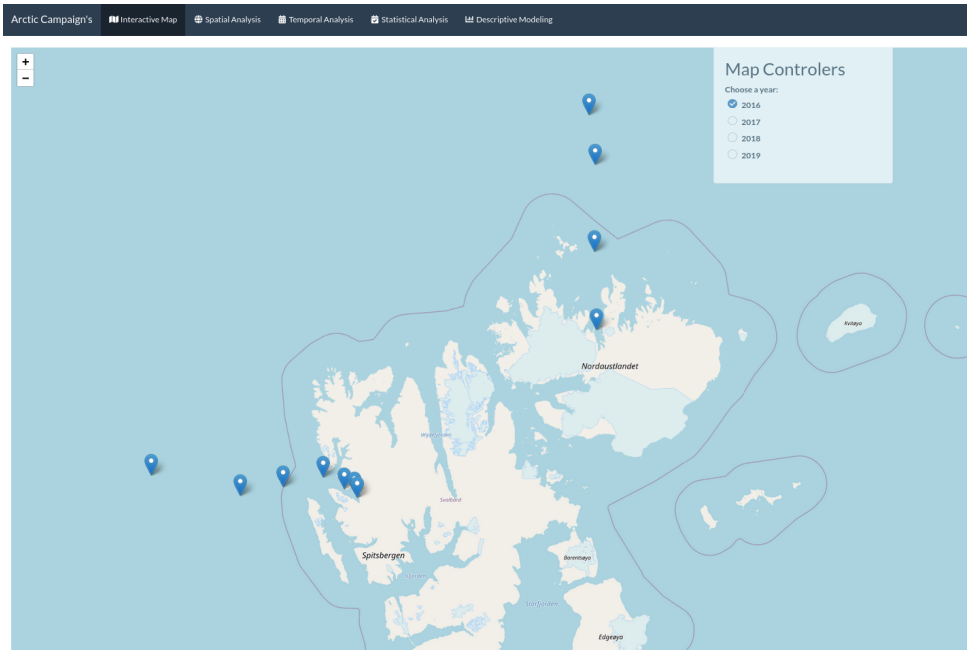


Figure 4.1: Interactive Map Panel: initial screen of the application.

The entire design of the application is chosen around being easy to use and understandable by the users, in the way to interact and draw conclusions. For this, we used one of the themes provided by shiny and made only a few changes. The layout of the application is shown in Figure 4.2, and is composed by:

- Header (A) - the header section of the page is a navigation bar that has the name of the application and it is used to switch between tabPanels,which are Interactive Map, Spatial Analysis, Temporal Analysis, Statistical Analysis and Descriptive Modeling;
- SidebarPanel (B) - the section on the left side of the page, which contains shiny widgets where users must introduce what they want to analyze;
- Main Panel (C) - the section on the right side of the sidebarPanel; it is used to display the elements that were analyzed in the form of plots or tables.

As previously described, the application has a header section that serves as a navigation menu where we have a variety of panels that we can choose from. There are 5 tabs: **Interactive Map**, **Spatial Analysis**, **Temporal Analysis**, **Statistical Analysis** and **Descriptive Modeling**.

The first tab is the only one that differs a bit in terms of layout from the rest. A map corresponding to the area for which data was collected in the arctic is displayed across the width of the page, and also have a selection menu on the right side where it is possible to drag and move around, and also choose the year for which to present the stations location (cf. Figure 4.1).

In the remaining four tabs, the layout is no longer similar to the one in the Interactive Map. In these tabs we have on the left side of the page the side bar panel, this is the place where we can select from among the possibilities, the options we want to examine in close detail. We also

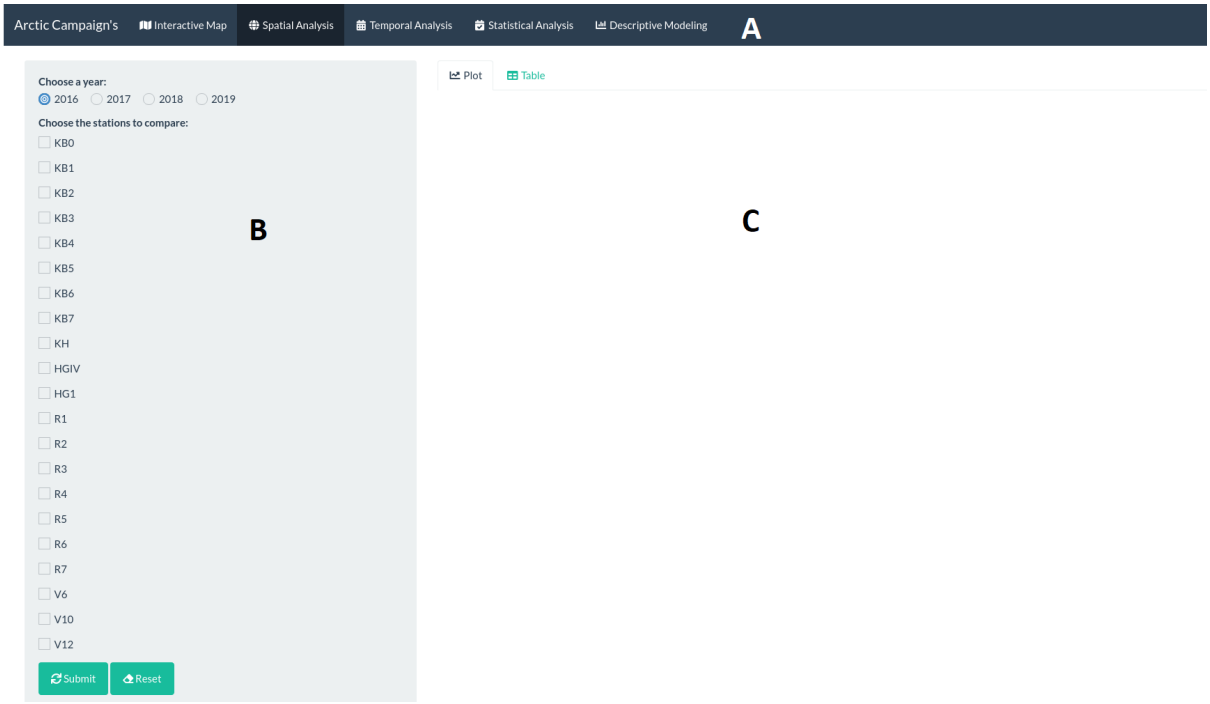


Figure 4.2: The Shiny app layout illustration.

have the main panel which is one the right side of the side bar panel, this is where we can see the graphics resulting from the analyzes. We can see an example of this layout in Figure 4.2.

The main panel still has other tabs (sub-tabs) that are at the top of the page just below the header section. The sub-tabs allows us to access different pages within the same main tab, and will vary depending on the main tab selected, we will explain in more detail below.

4.1.1 Spatial Analysis

This tab analyzes the spatial level and shares some widgets with **Temporal Analysis**, such as the choice of year or years and choice of stations. Regarding the choice of year or years, we have four options available from 2016 until 2019, which are the years we have data available.

The choice of stations depends on the applicant for the utility. All stations are available for selection, as a rule after selecting a year and a station a new widget appears with options available for choosing the desired CTD, as shown in Figure 4.3. For those that are selected and there is no CTD data available, there will be displayed a message in order to inform that for that year and for that station provided there are no data available, Figure 4.4 exemplifies it.

This analysis can be done by choosing a year and one or more stations as shown in Figure 4.5a. As stated earlier, the main panel is divided 4.5b, allowing you to choose between two sub-tabs:

- Plot
  In the sub-tab plot we can see the graph created by the parameters previously chosen in the side bar panel. This graph is interactive and we can zoom-in by selecting the area we

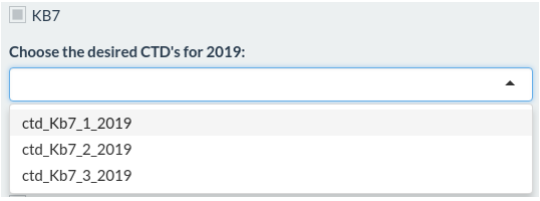


Figure 4.3: Example of a selection menu for station KB7.

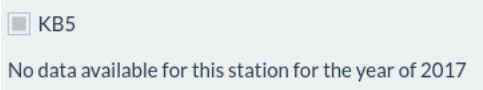


Figure 4.4: Example of the message displayed when there is no data available.

want to see in greater detail and zoom-out using a double click, and hovering over the graph we can see the value, the depth at which we found that value and which station we are getting this information for, it also helps that the box where this information appears is the same color as the station in the chart legend. It is also possible to download the image created by the graphic directly from the application.

- Table
  In this sub-tab it is possible to see the data that was selected in the side bar panel by the following categories:
  - Depth
  - Station
  - Variable
  - Value

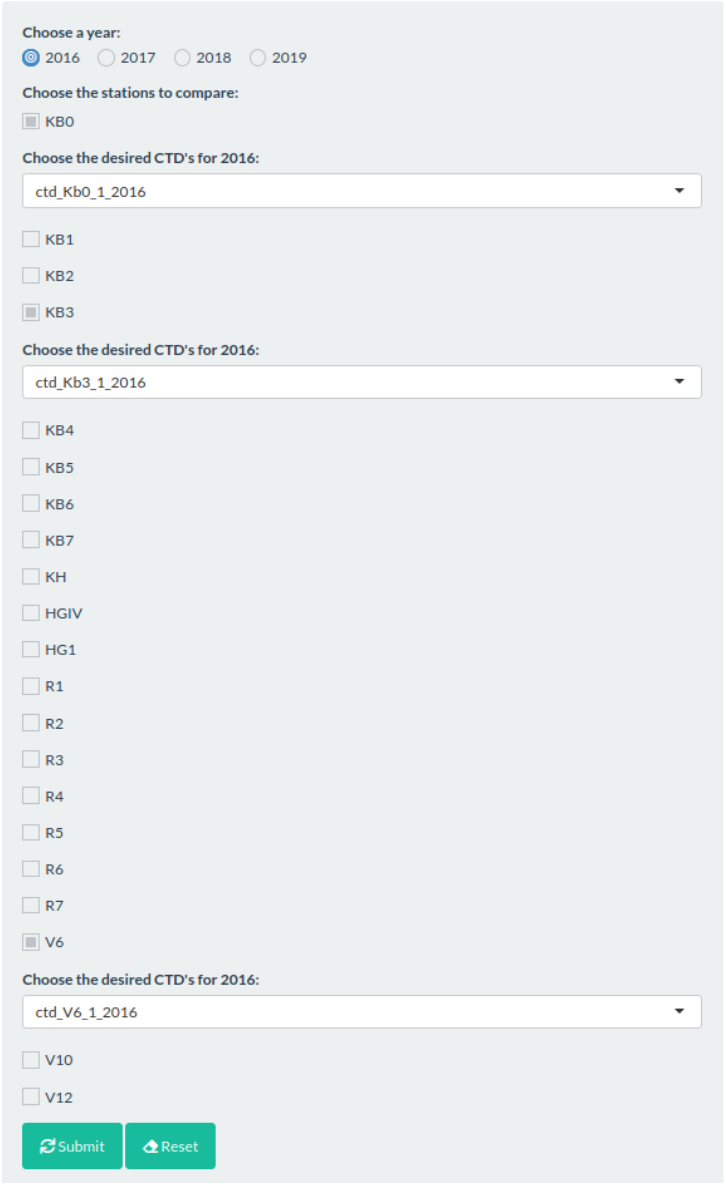
  It is still possible to do a search, for example for a certain parameter such as Fluorescence or station. In Figure 4.6 we have an example of how this table looks like.

4.1.2 Temporal Analysis

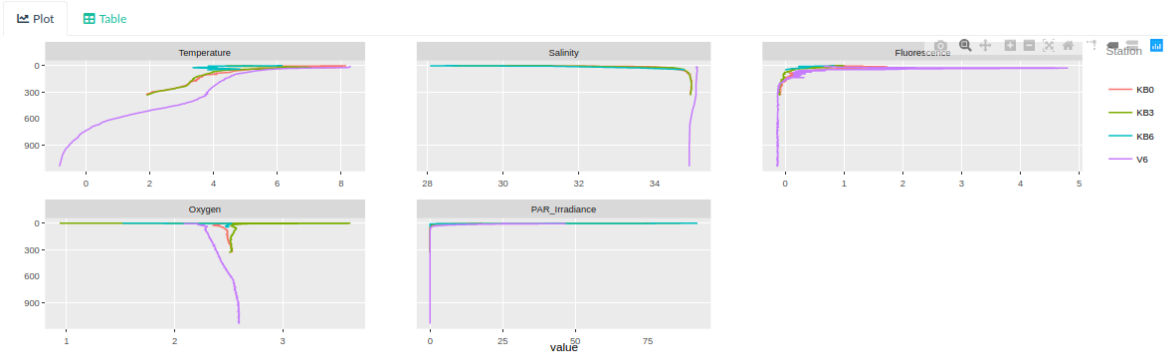
In this tab we do an analyze focused on how some years compare with others and how the stations diverge over the years. In the main panel we have 3 sub-tabs:

- Plot by Year
- Plot by CTD
- Table

In both sub-tabs, Plot by Year and Plot by CTD, the characteristics of the graph are the same as in the plot sub-tab, here the difference is that we analyze by year, where we can choose more than one year, for one station and a CTD per year to analyze, like in Figure 4.7a were we



(a) SideBar Panel.



(b) Main Panel.

Figure 4.5: Spatial Analysis Panel.

Plot Table

Show 10 entries Search:

	Depth	Station	variable	value
1	1	KB0	Temperature	7.2267
2	2	KB0	Temperature	7.5939
3	3	KB0	Temperature	8.5865
4	4	KB0	Temperature	8.2663
5	5	KB0	Temperature	7.6814
6	6	KB0	Temperature	6.6885
7	7	KB0	Temperature	6.8381
8	8	KB0	Temperature	7.3468
9	9	KB0	Temperature	7.6116
10	10	KB0	Temperature	7.7251

Showing 1 to 10 of 1,925 entries Previous 1 2 3 4 5 ... 193 Next

Figure 4.6: Example of the Table sub-tab for the Spatial Analysis Panel.

select the input, and Figure 4.7b shows the graph. We can also analyze all CTDs available for the station selected allowing the possibility of see how values vary over the years by CTD.

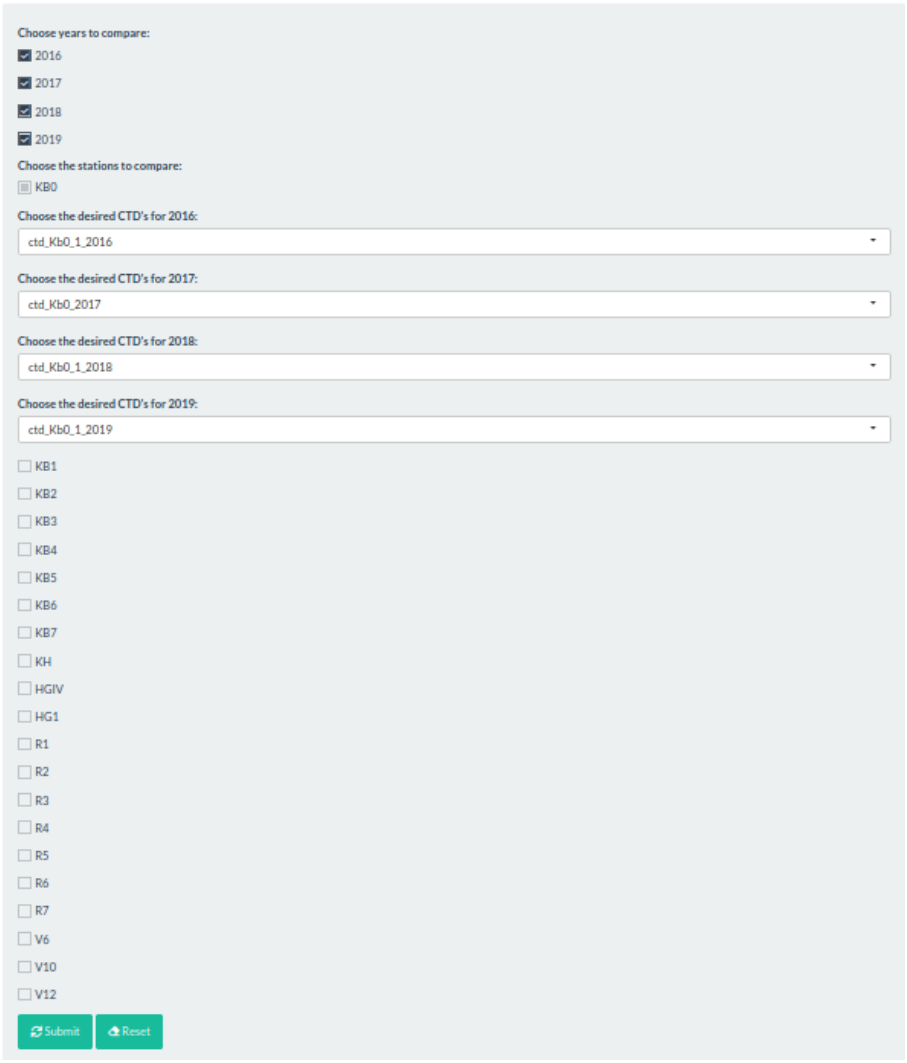
4.1.3 Statistical Analysis

This tab as a similar layout as the others. The side bar panel is used to choose a year and the station that we want to analyze in greater detail. Here we no longer choose the CTD we want to analyze, but only the station, as all CTD's for that station and year are used to create the graph, an example of how the side bar panel is is shown in Figure 4.8a. In the main panel shown in Figure 4.8b, we have the correlation plot for all parameters we are analysing for the station KB0 for the year of 2016.

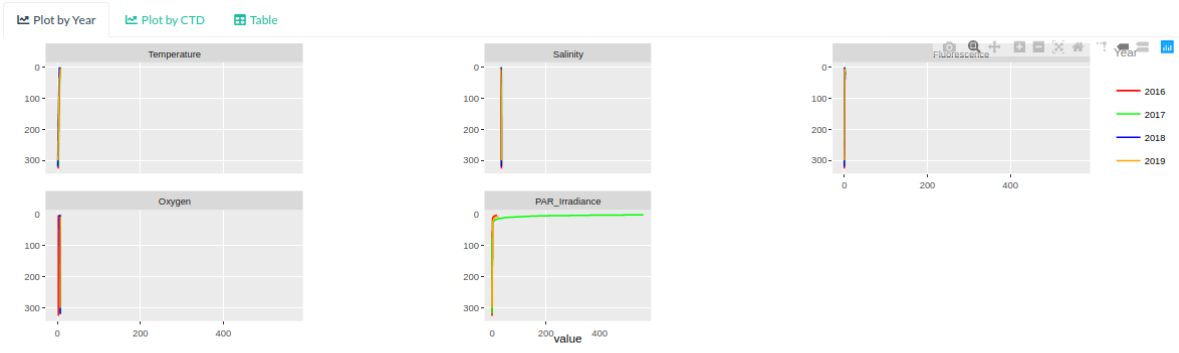
4.1.4 Descriptive Modeling

In the descriptive modeling tab, the options of the side bar panel are a little different from the others. We start by having 3 options regarding the stations available for analysis. Here we have to select whether we want the transect Kongsfjorden, Rijpfjorden or choose the options manually. When we make this selection, if we choose the transect Kongsfjorden, only the stations that belong to this transect previously selected appeared, the same happens if we choose the transect Rijpfjorden. After this selection, we return to having the same as we had previously regarding the selection of the year or years to be analyzed. Subsequently after choosing the year, we can then choose the CTD's for each station.

In addition to this options, in this tab, we can also choose the level of depth we intend to analyze, there are only 3 depths that have been defined because they are the most interesting. The depths are divided as follows:

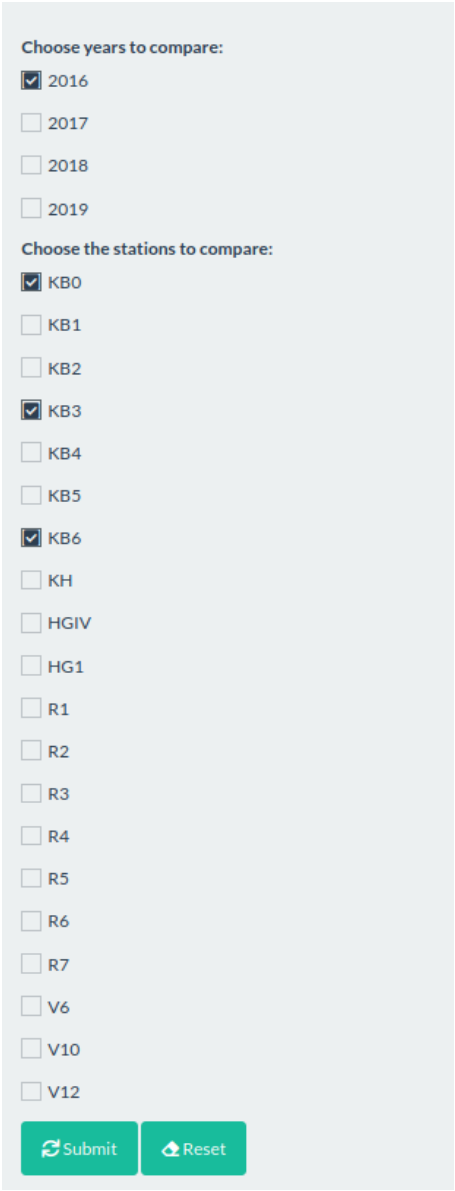


(a) SideBar Panel.

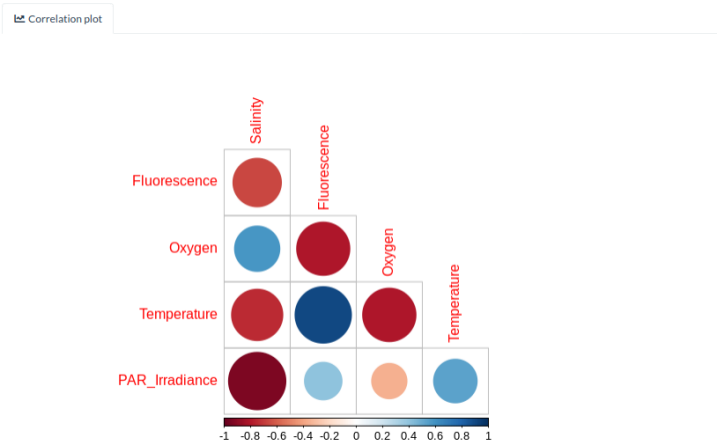


(b) Main Panel.

Figure 4.7: Temporal Analysis Panel.



(a) SideBar Panel.

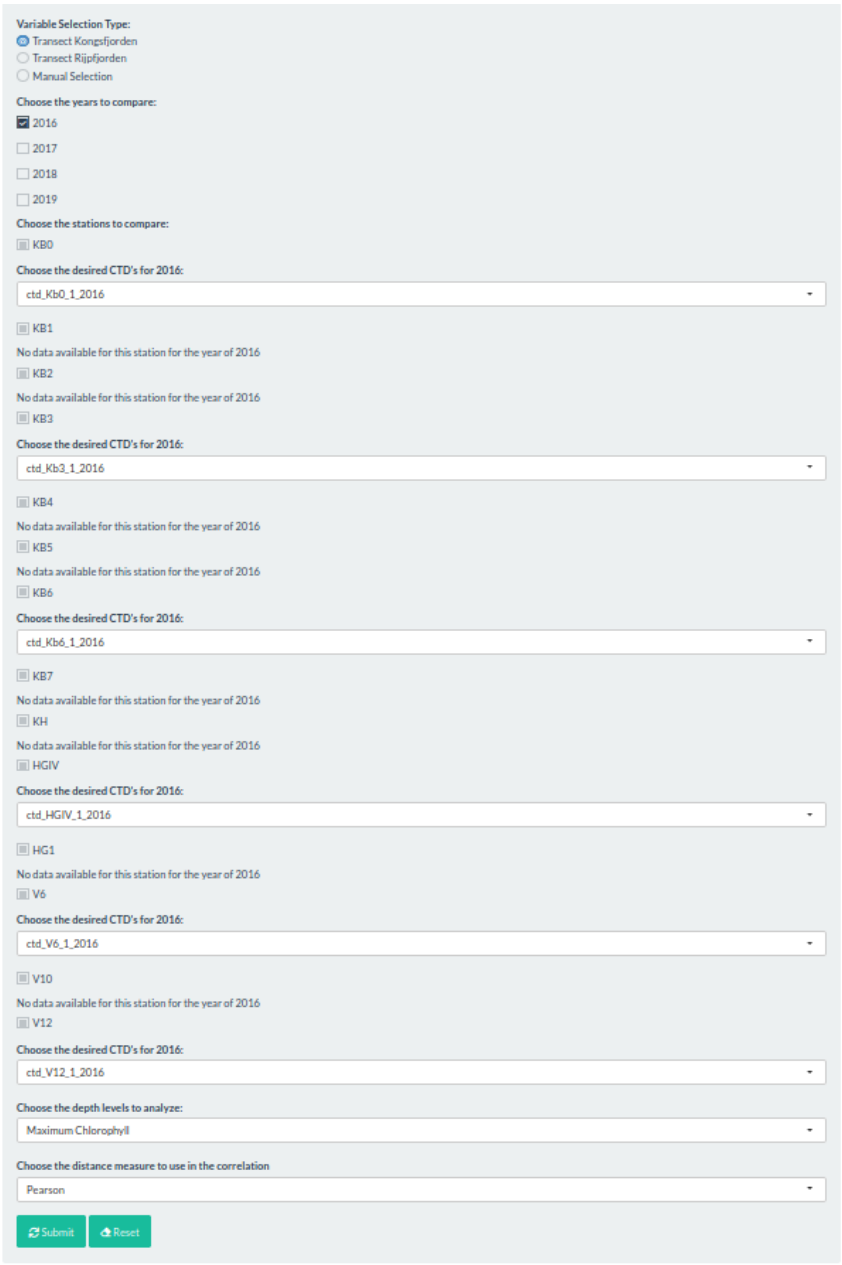


(b) Main Panel.

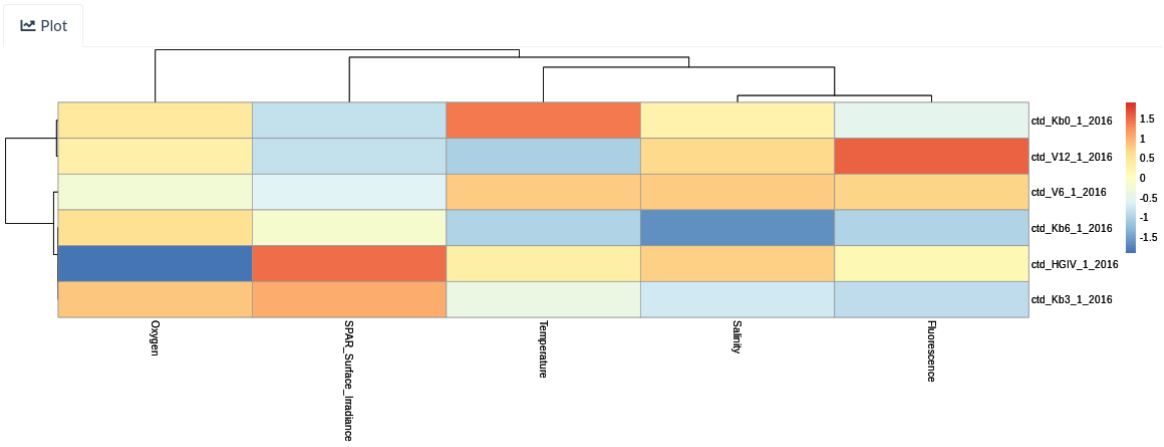
Figure 4.8: Statistical Analysis Panel.

- Surface - comprises depth levels of less than 6m.
- Maximum Chlorophyll - includes the 10 values where the fluorescence parameter is maximum, when compared to the remaining data from that CTD.
- Bottom - 10 highest values for the depth parameter.

Lastly, we can still choose which correlation measurement unit we want the chart to use when doing hierarchical clustering. For this feature we only have two options available Pearson and Spearman. Figure 4.9a is the example of how the side bar panel for this tab looks like. This section also has a sub-tab named Plot, which is different from the one in Spatial Analysis tab. In this case this plot has the values normalized for each pre-defined parameter by scaling and centering the values by the columns direction, and also have the correlation between stations through the hierarchical clustering, but doesn't allow you to download, or zoom in/out the graph created. This is exemplified in the following image 4.9b.



(a) SideBar Panel.



(b) Main Panel.

Figure 4.9: Descriptive Modeling.

Chapter 5

Conclusions

In this chapter we will discuss the conclusions reached from this study. We also present the contributions and some limitations we encountered, as well as proposals to develop as future work.

5.1 Contributions

The main objective of this thesis was the creation of a web platform where it was possible to dynamically analyze and visualize the data collected during the monitoring program of Svalbard and Jan Mayen (MOSJ program). This objective was achieved and, therefore, the following contributions were made:

- analysis of the evolution of the parameters and stations under study in time;
- analysis of the evolution of the parameters and stations under study in space;
- study of parameters correlation at specific depth measures;
- generation of association rules for parameters at a specific interval to make possible the analysis of the influence of environmental standards (i.e. Oxygen, Salinity,etc) in the distribution of phytoplankton biomass;
- development of a web app which allows a dynamic and interactive form of graphic visualization and interpretation of exploratory data analysis.

5.2 Limitations

As for the limitations we find, there are essentially two. The first was a matter of time. We would like to have explored the biological component to which we had access but did not have the time to explore and also to incorporate into the application the association rules that ended up not being possible. The second is that some methods that we enumerated ended up not

being used. This was due to the fact that, despite being a data set with a good number of observations, for this type of methods is too limited. It does not work properly with the lack of fine space and time granularity present in the data set.

5.3 Future Work

Regarding future work, it would be adequate to start by including the association rules in the application as it is a component that we have already proven to be valuable. Next, we suggest the analysis of other components, such as the biological and the chemical components.

Another idea is to explore in detail more methods that would be able to cope with this data, taking into account its spatio-temporal nature so that other type of conclusions can be drawn.

In addition to this, the application can also be improved so that it is possible to follow online the progress of the campaigns. For that purpose, a new functionality needs to be developed to allow the upload of the collected data into the application during the campaigns. This would allow a more expeditious analysis of the data. New panels can also be created to perform chemical and biological analysis of the data, thus making the application more multi-disciplinary.

Appendix A

Association Rules

In Tables A.1, A.2 and A.3 we have the set of association rules generated by the Apriori [4] algorithm, and which relate environmental variables. As for Table A.4 we have the set of association rules where we relate environmental variables with the maximum values for the fluorescence parameter. These rules are ordered by their lift value and were the result of filtering the original set of rules eliminating those that are considered redundant.

Table A.1: Apriori association rules involving all environmental variables width minimum support of 0.1 and minimum confidence of 0.8, ordered by decreasing order of lift.

Id	rules	support	confidence	lift	count
01	{Temperature=[-1.86,-0.0767),Oxygen=[-13.5,6.57),Fluorescence=[-0.338,-0.1)} =>{Depth=[676,2.45e+03]}	0.114043557558202	1	3	4252
02	{Temperature=[-1.86,-0.0767),Salinity=[34.9,35),Oxygen=[6.57,6.84)} =>{Depth=[676,2.45e+03]}	0.112031970818582	1	3	4177
03	{Temperature=[-1.86,-0.0767),Oxygen=[6.57,6.84)} =>{Depth=[676,2.45e+03]}	0.145719343418088	0.999815973500184	2.99944792050055	5433
04	{Temperature=[-1.86,-0.0767),Salinity=[34.9,35)} =>{Depth=[676,2.45e+03]}	0.143922325930694	0.999813676169182	2.99944102850755	5366
05	{Temperature=[-1.86,-0.0767),Fluorescence=[-0.338,-0.1),PAR_Irradiance=[1e-12,0.107)} =>{Depth=[676,2.45e+03]}	0.136868361763759	0.999608227228208	2.99882468168462	5103
06	{Depth=[676,2.45e+03],Salinity=[1.74,34.9),Oxygen=[-13.5,6.57)} =>{Fluorescence=[-0.338,-0.1)}	0.102108142903122	0.998688352570829	2.99847773693435	3807
07	{Temperature=[-1.86,-0.0767),Fluorescence=[-0.338,-0.1)} =>{Depth=[676,2.45e+03]}	0.136868361763759	0.999216761308009	2.99765028392403	5103
08	{Depth=[676,2.45e+03],Oxygen=[-13.5,6.57),PAR_Irradiance=[1e-12,0.107)} =>{Fluorescence=[-0.338,-0.1)}	0.127373672352752	0.998108448928121	2.99673662504719	4749
09	{Depth=[676,2.45e+03],Oxygen=[-13.5,6.57)} =>{Fluorescence=[-0.338,-0.1)}	0.127373672352752	0.975955610357583	2.93022459144565	4749
10	{Salinity=[1.74,34.9),Oxygen=[-13.5,6.57),Fluorescence=[-0.338,-0.1),PAR_Irradiance=[1e-12,0.107)} =>{Depth=[676,2.45e+03]}	0.102108142903122	0.970183486238532	2.9105504587156	3807
11	{Salinity=[1.74,34.9),Oxygen=[-13.5,6.57),Fluorescence=[-0.338,-0.1)} =>{Depth=[676,2.45e+03]}	0.102108142903122	0.969936305732484	2.90980891719745	3807
12	{Temperature=[-1.86,-0.0767),Oxygen=[-13.5,6.57)} =>{Depth=[676,2.45e+03]}	0.117181632872009	0.968521392152516	2.90556417645755	4369
13	{Temperature=[-1.86,-0.0767),Oxygen=[-13.5,6.57),PAR_Irradiance=[1e-12,0.107)} =>{Fluorescence=[-0.338,-0.1)}	0.114043557558202	0.966363636363636	2.90142549671298	4252
14	{Salinity=[1.74,34.9),Fluorescence=[-0.338,-0.1),PAR_Irradiance=[1e-12,0.107)} =>{Depth=[676,2.45e+03]}	0.131504130458105	0.963829368979752	2.89148810693926	4903
15	{Salinity=[1.74,34.9),Fluorescence=[-0.338,-0.1)} =>{Depth=[676,2.45e+03]}	0.131504130458105	0.962693893579423	2.88808168073827	4903
16	{Depth=[1,173),Salinity=[35,36.3]} =>{Temperature=[3.15,10.3]}	0.123189571934342	0.960275977420029	2.88082793226009	4593
17	{Depth=[173,676),Salinity=[34.9,35)} =>{Temperature=[-0.0767,3.15)}	0.1154650788542	0.955817051509769	2.86745115452931	4305
18	{Depth=[676,2.45e+03],Salinity=[34.9,35),Oxygen=[6.57,6.84)} =>{Temperature=[-1.86,-0.0767)}	0.112031970818582	0.943741527338455	2.83122458201536	4177
19	{Temperature=[-1.86,-0.0767),Oxygen=[-13.5,6.57)} =>{Fluorescence=[-0.338,-0.1)}	0.114043557558202	0.942584792728885	2.83003151973778	4252
20	{Temperature=[-0.0767,3.15),Salinity=[35,36.3]} =>{Depth=[173,676)}	0.106721381825984	0.941775147928994	2.82305391665739	3979
21	{Depth=[676,2.45e+03],Temperature=[-1.86,-0.0767),Salinity=[1.74,34.9),PAR_Irradiance=[1e-12,0.107)} =>{Fluorescence=[-0.338,-0.1)}	0.119622358116082	0.935402684563758	2.80846784436102	4460
22	{Depth=[676,2.45e+03],Salinity=[1.74,34.9),PAR_Irradiance=[1e-12,0.107)} =>{Fluorescence=[-0.338,-0.1)}	0.131504130458105	0.918508804795804	2.75774539201214	4903
23	{Depth=[676,2.45e+03],Salinity=[1.74,34.9),Fluorescence=[-0.338,-0.1)} =>{Temperature=[-1.86,-0.0767)}	0.119622358116082	0.909647154803182	2.72894146440954	4460
24	{Salinity=[1.74,34.9),Fluorescence=[0.0366,22]} =>{Depth=[1,173)}	0.12104387941208	0.903141885131079	2.7116075088764	4513

Table A.2: Apriori association rules involving all environmental variables width minimum support of 0.1 and minimum confidence of 0.8, ordered by decreasing order of lift (cont.I).

Id	rules	support	confidence	lift	count
25	{Depth=[1,173),Temperature=[3.15,10.3],PAR_Irradiance=[0.107,656]} =>{Fluorescence=[0.0366,22]}	0.124101491256303	0.974515585509688	2.70522217929739	4627
26	{Temperature=[3.15,10.3],Fluorescence=[0.0366,22],PAR_Irradiance=[0.107,656]} =>{Depth=[1,173)}	0.124101491256303	0.898795648795649	2.69855830002391	4627
27	{Salinity=[1.74,34.9),Oxygen=[6.84,17.3]} =>{Depth=[1,173)}	0.111736938096771	0.898619499568594	2.698029426793	4166
28	{Depth=[676,2.45e+03],Oxygen=[-13.5,6.57)} =>{Temperature=[-1.86,-0.0767)}	0.117181632872009	0.897862720920674	2.69358816276202	4369
29	{Depth=[676,2.45e+03],Salinity=[1.74,34.9),PAR_Irradiance=[1e-12,0.107)} =>{Temperature=[-1.86,-0.0767)}	0.127883274326789	0.893218433870363	2.67965530161109	4768
30	{Depth=[1,173),Oxygen=[6.84,17.3],PAR_Irradiance=[0.107,656]} =>{Fluorescence=[0.0366,22]}	0.108089260808926	0.95792726408367	2.65917356221395	4030
31	{Temperature=[-1.86,-0.0767),PAR_Irradiance=[1e-12,0.107)} =>{Depth=[676,2.45e+03]}	0.212691771269177	0.879059971178362	2.63717991353509	7930
32	{Salinity=[1.74,34.9),Fluorescence=[-0.338,-0.1),PAR_Irradiance=[1e-12,0.107)} =>{Temperature=[-1.86,-0.0767)}	0.119676000429139	0.877137802241006	2.63141340672302	4462
33	{Salinity=[1.74,34.9),Fluorescence=[-0.338,-0.1)} =>{Temperature=[-1.86,-0.0767)}	0.119729642742195	0.876497152955036	2.62949145886511	4464
34	{Temperature=[-1.86,-0.0767),Fluorescence=[-0.338,-0.1)} =>{Salinity=[1.74,34.9)}	0.119729642742195	0.874094380262385	2.6237609591581	4464
35	{Temperature=[3.15,10.3],Fluorescence=[0.0366,22]} =>{Depth=[1,173)}	0.164279583735651	0.872507122507122	2.61962921207566	6125
36	{Depth=[676,2.45e+03],Temperature=[-1.86,-0.0767),Salinity=[1.74,34.9)} =>{Fluorescence=[-0.338,-0.1)}	0.119622358116082	0.870413739266198	2.61334400505725	4460
37	{Depth=[676,2.45e+03],Oxygen=[6.57,6.84)} =>{Temperature=[-1.86,-0.0767)}	0.145719343418088	0.869419107057129	2.60825732117139	5433
38	{Temperature=[-1.86,-0.0767),Oxygen=[-13.5,6.57),PAR_Irradiance=[1e-12,0.107)} =>{Salinity=[1.74,34.9)}	0.102403175624933	0.867727272727273	2.60464887177873	3818
39	{Depth=[1,173),PAR_Irradiance=[0.107,656]} =>{Fluorescence=[0.0366,22]}	0.166961699388478	0.936653626241348	2.60011866583147	6225
40	{Temperature=[3.15,10.3],PAR_Irradiance=[0.107,656]} =>{Fluorescence=[0.0366,22]}	0.138075313807531	0.935829849118342	2.59783188850631	5148
41	{Salinity=[1.74,34.9),PAR_Irradiance=[0.107,656]} =>{Depth=[1,173)}	0.103476021886064	0.865216416236824	2.59773948002688	3858
42	{Depth=[676,2.45e+03],Salinity=[1.74,34.9)} =>{Temperature=[-1.86,-0.0767)}	0.137431606050853	0.865248226950355	2.59574468085106	5124
43	{Temperature=[3.15,10.3],PAR_Irradiance=[0.107,656]} =>{Depth=[1,173)}	0.127346851196224	0.863115797127795	2.59143254792339	4748
44	{Temperature=[-1.86,-0.0767),Oxygen=[-13.5,6.57)} =>{Salinity=[1.74,34.9)}	0.102483639094518	0.847040567501663	2.54255378139699	3821
45	{Depth=[676,2.45e+03]} =>{Temperature=[-1.86,-0.0767)}	0.281353931981547	0.844061795944641	2.53218538783392	10490
46	{Temperature=[-1.86,-0.0767)} =>{Depth=[676,2.45e+03]}	0.281353931981547	0.844061795944641	2.53218538783392	10490
47	{Salinity=[1.74,34.9),Oxygen=[-13.5,6.57),PAR_Irradiance=[1e-12,0.107)} =>{Fluorescence=[-0.338,-0.1)}	0.10524621821693	0.840797085922434	2.52442249569432	3924
48	{Depth=[676,2.45e+03],Temperature=[-1.86,-0.0767),Fluorescence=[-0.338,-0.1)} =>{Oxygen=[-13.5,6.57)}	0.114043557558202	0.83323535175387	2.49970605526161	4252
49	{Temperature=[-1.86,-0.0767),Fluorescence=[-0.338,-0.1),PAR_Irradiance=[1e-12,0.107)} =>{Oxygen=[-13.5,6.57)}	0.114043557558202	0.832908912830558	2.49872673849167	4252

Table A.3: Apriori association rules involving all environmental variables width minimum support of 0.1 and minimum confidence of 0.8, ordered by decreasing order of lift (cont.II).

50	{Temperature=[-1.86,-0.0767),Fluorescence=[-0.338,-0.1)} =>{Oxygen=[-13.5,6.57)}	0.114043557558202	0.832582729586842	2.49774818876052	4252
51	{Depth=[676,2.45e+03],Salinity=[1.74,34.9)} =>{Fluorescence=[-0.338,-0.1)}	0.131504130458105	0.827929753461668	2.48578941279311	4903
52	{Salinity=[1.74,34.9),Oxygen=[-13.5,6.57),PAR_Irradiance=[1e-12,0.107)} =>{Temperature=[-1.86,-0.0767)}	0.102403175624933	0.818084422541247	2.45425326762374	3818
53	{Salinity=[1.74,34.9),Oxygen=[-13.5,6.57),PAR_Irradiance=[1e-12,0.107)} =>{Depth=[676,2.45e+03]}	0.102242248685763	0.816798800085708	2.45039640025712	3812
54	{Depth=[1,173),Oxygen=[6.84,17.3],Fluorescence=[0.0366,22]} =>{PAR_Irradiance=[0.107,656]}	0.108089260808926	0.839408456571548	2.41858615879549	4030
55	{Salinity=[1.74,34.9),PAR_Irradiance=[0.107,656]} =>{Fluorescence=[0.0366,22]}	0.103931981547044	0.86902893025342	2.41239480571577	3875
56	{Depth=[676,2.45e+03],Oxygen=[-13.5,6.57),Fluorescence=[-0.338,-0.1)} =>{Salinity=[1.74,34.9)}	0.102108142903122	0.801642451042325	2.40628267809855	3807
57	{Depth=[676,2.45e+03],Oxygen=[-13.5,6.57),PAR_Irradiance=[1e-12,0.107)} =>{Salinity=[1.74,34.9)}	0.102242248685763	0.801176965111391	2.4048854333156	3812
58	{Temperature=[3.15,10.3],Oxygen=[6.84,17.3]} =>{Fluorescence=[0.0366,22]}	0.102617744877159	0.821735395189003	2.28110955805426	3826
59	{Oxygen=[6.57,6.84),Fluorescence=[-0.338,-0.1)} =>{PAR_Irradiance=[1e-12,0.107)}	0.127641883918035	1	1.53432098765432	4759
60	{Depth=[676,2.45e+03],Fluorescence=[-0.338,-0.1)} =>{PAR_Irradiance=[1e-12,0.107)}	0.174498444372921	1	1.53432098765432	6506
61	{Depth=[173,676),Fluorescence=[-0.338,-0.1)} =>{PAR_Irradiance=[1e-12,0.107)}	0.136600150198477	1	1.53432098765432	5093
62	{Temperature=[-1.86,-0.0767),Oxygen=[-13.5,6.57),Fluorescence=[-0.338,-0.1)} =>{PAR_Irradiance=[1e-12,0.107)}	0.114043557558202	1	1.53432098765432	4252
63	{Depth=[676,2.45e+03],Salinity=[1.74,34.9),Oxygen=[-13.5,6.57)} =>{PAR_Irradiance=[1e-12,0.107)}	0.102242248685763	1	1.53432098765432	3812
64	{Salinity=[35,36.3],Fluorescence=[-0.338,-0.1)} =>{PAR_Irradiance=[1e-12,0.107)}	0.110074026392018	0.999756394640682	1.53394721883881	4104
65	{Salinity=[1.74,34.9),Oxygen=[-13.5,6.57),Fluorescence=[-0.338,-0.1)} =>{PAR_Irradiance=[1e-12,0.107)}	0.10524621821693	0.999745222929936	1.53393007784855	3924
66	{Temperature=[-1.86,-0.0767),Fluorescence=[-0.338,-0.1)} =>{PAR_Irradiance=[1e-12,0.107)}	0.136922004076816	0.999608380654004	1.53372011787259	5105
67	{Oxygen=[-13.5,6.57),Fluorescence=[-0.338,-0.1)} =>{PAR_Irradiance=[1e-12,0.107)}	0.187292136036906	0.999284487693188	1.53322316210505	6983
68	{Temperature=[-1.86,-0.0767),Salinity=[1.74,34.9),Oxygen=[-13.5,6.57)} =>{PAR_Irradiance=[1e-12,0.107)}	0.102403175624933	0.999214865218529	1.53311633888097	3818
69	{Fluorescence=[-0.338,-0.1)} =>{PAR_Irradiance=[1e-12,0.107)}	0.332796910202768	0.999194717345788	1.53308542557697	12408
70	{Depth=[676,2.45e+03],Oxygen=[-13.5,6.57)} =>{PAR_Irradiance=[1e-12,0.107)}	0.127615062761506	0.977805178791615	1.50026700765706	4758
71	{Temperature=[-1.86,-0.0767),Oxygen=[-13.5,6.57)} =>{PAR_Irradiance=[1e-12,0.107)}	0.118013088724386	0.975393482598094	1.49656669157149	4400
72	{Depth=[676,2.45e+03],Temperature=[-1.86,-0.0767),Salinity=[1.74,34.9)} =>{PAR_Irradiance=[1e-12,0.107)}	0.127883274326789	0.930523028883685	1.42772101271191	4768
73	{Depth=[676,2.45e+03],Salinity=[1.74,34.9)} =>{PAR_Irradiance=[1e-12,0.107)}	0.143171333547903	0.901384667342114	1.38301341305281	5338
74	{Oxygen=[-13.5,6.57)} =>{PAR_Irradiance=[1e-12,0.107)}	0.290258555948933	0.870775667846798	1.33604938271605	10822
75	{Temperature=[-1.86,-0.0767),Salinity=[1.74,34.9)} =>{PAR_Irradiance=[1e-12,0.107)}	0.157145156099131	0.829769154510693	1.27313222867394	5859

Table A.4: Apriori association rules involving environmental variables and maximum fluorescence values width minimum support of 0.1 and minimum confidence of 0.8, ordered by decreasing order of lift.

Id	rules	support	confidence	lift	count
01	{Salinity=[1.74,34.9),Fluorescence=[0.0366,22]} =>{Depth=[1,173)}	0.12104387941208	0.903141885131079	2.7116075088764	4513
02	{Depth=[1,173),Temperature=[3.15,10.3],PAR_Irradiance=[0.107,656]} =>{Fluorescence=[0.0366,22]}	0.124101491256303	0.974515585509688	2.70522217929739	4627
03	{Temperature=[3.15,10.3],Fluorescence=[0.0366,22],PAR_Irradiance=[0.107,656]} =>{Depth=[1,173)}	0.124101491256303	0.898795648795649	2.69855830002391	4627
04	{Depth=[1,173),Oxygen=[6.84,17.3],PAR_Irradiance=[0.107,656]} =>{Fluorescence=[0.0366,22]}	0.108089260808926	0.95792726408367	2.65917356221395	4030
05	{Temperature=[3.15,10.3],Fluorescence=[0.0366,22]} =>{Depth=[1,173)}	0.164279583735651	0.872507122507122	2.61962921207566	6125
06	{Depth=[1,173),PAR_Irradiance=[0.107,656]} =>{Fluorescence=[0.0366,22]}	0.166961699388478	0.936653626241348	2.60011866583147	6225
07	{Temperature=[3.15,10.3],PAR_Irradiance=[0.107,656]} =>{Fluorescence=[0.0366,22]}	0.138075313807531	0.935829849118342	2.59783188850631	5148
08	{Depth=[1,173),Oxygen=[6.84,17.3],Fluorescence=[0.0366,22]} =>{PAR_Irradiance=[0.107,656]}	0.108089260808926	0.839408456571548	2.41858615879549	4030
09	{Salinity=[1.74,34.9),PAR_Irradiance=[0.107,656]} =>{Fluorescence=[0.0366,22]}	0.103931981547044	0.86902893025342	2.41239480571577	3875
10	{Temperature=[3.15,10.3],Oxygen=[6.84,17.3]} =>{Fluorescence=[0.0366,22]}	0.102617744877159	0.821735395189003	2.28110955805426	3826

Bibliography

[1] Plotly R open source graphing library. Accessed on 10/07/20.

[2] What is clustering in data mining?, 2015. Accessed on 27/11/19.

[3] Charu C. Aggarwal. *Data Mining: The Textbook*. Springer, 2015.

[4] Mohammed Al-Maolegi and Bassam Arkok. An improved apriori algorithm for association rules. *International Journal on Natural Language Computing*, 3, 03 2014. doi:10.5121/ijnlc.2014.3103.

[5] Gowtham Atluri, Anuj Karpatne, and Vipin Kumar. Spatio-temporal data mining: A survey of problems and methods. *ACM Comput. Surv.*, 51(4), 2018. doi:10.1145/3161602.

[6] Berkay Aydin and Rafal Angryk. Spatiotemporal frequent pattern mining on solar data: Current algorithms and future directions. pages 575–581, 11 2015. doi:10.1109/ICDMW.2015.10.

[7] Biological and Chemical Oceanography Data Management Office. Parameter:flag. Accessed on 20/09/20.

[8] Derya Birant and Alp Kut. ST-DBSCAN: An algorithm for clustering spatial–temporal data. *Data Knowledge Engineering*, 60(1):208 – 221, 2007. ISSN: 0169-023X. Intelligent Data Mining. doi:10.1016/j.datak.2006.01.013.

[9] Gordon S. Blair, Peter Henrys, Amber Leeson, John Watkins, Emma Eastoe, Susan Jarvis, and Paul J. Young. Data science of the natural environment: A research roadmap. *Frontiers in Environmental Science*, 7:121, 2019. ISSN: 2296-665X. doi:10.3389/fenvs.2019.00121.

[10] Jason Brownlee. Difference between classification and regression in machine learning. Accessed on 23/09/20.

[11] Timothy J.B. Carruthers, Ben J. Longstaff, William C. Dennison, Eva G. Abal, and Keiko Aioi. Chapter 19 - measurement of light penetration in relation to seagrass. In Frederick T. Short and Robert G. Coles, editors, *Global Seagrass Research Methods*, pages 369 – 392. Elsevier Science, Amsterdam, 2001. ISBN: 978-0-444-50891-1. doi:https://doi.org/10.1016/B978-044450891-1/50020-7.

[12] IBM Knowledge Center. Lift in an association rule. Accessed on 05/12/19.

[13] Winston Chang, Joe Cheng, JJ Allaire, Yihui Xie, and Jonathan McPherson. *shiny: Web Application Framework for R*, 2019. R package version 1.4.0.

[14] Chris Chatfield. Exploratory data analysis. *European Journal of Operational Research*, 23 (1):5 – 13, 1986. ISSN: 0377-2217. doi:10.1016/0377-2217(86)90209-2.

[15] Nian Shong Chok. Pearson's versus spearman's and kendall's correlation coefficients for continuous data. September 2010.

[16] Victoria Cox. *Exploratory Data Analysis. In: Translating Statistics to Make Decisions.* Apress, Berkeley, CA, 2017. ISBN: 978-1-4842-2255-3. doi:10.1007/978-1-4842-2256-0_3.

[17] Prof. Marcelo Menezes Reis Manoel de Oliveira Lino. Introdução e análise exploratória de dados. Slides accessed on 10/12/19.

[18] António Gaspar Gonçalves de Sousa. *Arctic microbiome and N-functions during the winter-spring transition.* PhD thesis, Faculdade de Ciências da Universidade do Porto, 2017.

[19] BD Editors. Trophic level - definition. Accessed on 11/09/20.

[20] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, page 226–231. AAAI Press, 1996.

[21] Ocean Exploration and Research. What does "CTD" stand for? Accessed on 03/12/19.

[22] R.A. Fisher. UCI machine learning repository, iris data set, 1936.

[23] Peter Flach. *Machine Learning: The Art and Science of Algorithms that Make Sense of Data*. Cambridge University Press, 2012. doi:10.1017/CBO9780511973000.

[24] Thomas Girke. R bioconductor manual. Accessed on 05/12/20.

[25] A. F. Guerra D. C. E. Bakker C. Canchaya E. Curry F. Foglini J. Irisson K. Malde C. T. Marshall M. Obst Guidi, L. European marine board: Big data in marine science. *Future Science Brief*, 6, 04 2020. doi:10.5281/zenodo.3755793.

[26] David Hand, Heikki Mannila, and Padhraic Smyth. *Principles of data mining*. A Bradford Book, 2001. ISBN: 978-0-262-08290-7.

[27] J. A. Hartigan and M. A. Wong. Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979. ISSN: 00359254, 14679876.

[28] Francisco Herrera, Cristóbal J. Carmona, Pedro González, and María José Del Jesus. An overview on subgroup discovery: Foundations and applications. *Knowledge and Information Systems*, 29:495–525, 12 2011. doi:10.1007/s10115-010-0356-2.

[29] Christian Buchta Michael Hahsler. is.redundant:find redundant rules. Accessed on 18/11/20.

[30] University of Massachusetts. Analysis of environmental data conceptual foundations: Data exploration, screening adjustments. Accessed on 31/08/20.

[31] Mauro Pichiliani. Data mining na prática: Regras de associação, 2008. Accessed on 05/12/19.

[32] Benedita Portugal, Catarina Magalhães, and Maria Tomasino. Exploring the nitrogen cycle in the arctic ocean, 2017-2018. Internship Report, Instituto de Ciências Biomédicas Abel Salazar.

[33] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018.

[34] Anant Ram, Jalal Sunita, Anand Jalal, and Kumar Manoj. A density based algorithm for discovering density varied clusters in large spatial databases. *International Journal of Computer Applications*, 3, 06 2010. doi:10.5120/739-1038.

[35] RDocumentation. hclust function. Accessed on 07/12/20.

[36] António Sousa, Maria Tomasino, Pedro Duarte, Mar Fernández-Méndez, Philipp Assmy, Hugo Ribeiro, Jaroslwa Surkont, Ricardo Leite, José Pereira-Leal, Luís Torgo, and Catarina Magalhães. Diversity and composition of pelagic prokariotic and protist communities in a thin arctic sea-ice regime. *Microbial Ecology*, 78(2), 2019. doi:10.1007/s00248-018-01314-2.

[37] Laerd Statistics. Pearson's product moment correlation, . Accessed on 23/09/20.

[38] Laerd Statistics. Spearman's rank-order correlation, . Accessed on 23/09/20.

[39] Statistical tools for high-throughput data analysis (STHDA). Correlation test between two variables in r. Accessed on 31/08/20.

[40] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN: 978-3-319-24277-4.

[41] Wikipedia. Pearson correlation coefficient. Accessed on 23/09/20.