UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat d'Informàtica de Barcelona

UNSUPERVISED PHYTOPLANKTON COMMUNITY DETECTION AND ANALYSIS OF ENVIRONMENTAL AND SATELLITE PARAMETERS ON EACH COMMUNITY

MOHANA FATHOLLAHI

Thesis supervisor: GABRIEL VALIENTE FERUGLIO (Department of Computer Science)

Thesis co-supervisor: RAMIRO LOGARES

Degree: Master Degree in Innovation and Research in Informatics ()

Thesis report

Facultat d'Informàtica de Barcelona (FIB)

Universitat Politècnica de Catalunya (UPC) - BarcelonaTech

26/06/2023

Unsupervised phytoplankton community detection and analysis of environmental and satellite parameters on each community

By

Mohana Fathollahi

Departament Informatic
Universitat Politècnica de Catalunya

Supervisor:
Gabriel Valiente
Co-supervisor:
Ramiro Logares

A thesis submitted for the degree of

Master in Innovation and Research in Informatics, Data Science

Barcelona, June 19, 2023

This thesis is submitted to the Informatic Department, Universitat Politècnica de Catalunya in fulfilment of the requirements for the degree in Innovation and Research in Informatics (Data Science).

Mohana Fathollahi, June 19, 2023

Acknowledgements

I would like to express my sincere gratitude to my advisors for their invaluable guidance, support, and expertise throughout this thesis project. Their insightful feedback and encouragement have been instrumental in shaping the direction of my research.

I would also like to extend my heartfelt thanks to my family for their unwavering support and encouragement. Their love, understanding, and belief in my abilities have been a constant source of motivation throughout this journey.

Abstract

Marine dynamics largely affects the phytoplankton community composition. The distribution characteristics of phytoplankton can reflect spatio-temporal variability in the marine ecosystem, on the other way around. In this work, we study the relation between remote sensing satellite observations, environmental factors and phytoplankton communities.

First, we employ network-based unsupervised clustering approaches to identify representative communities using metabarcoding data that has been collected both across depth and surface. Next, we investigate the relation between the detected phytoplankton communities and the environmental parameters (e.g., temperature, salinity, nutrients and so on). Our results show that phytoplankton communities are segregated based on the depth and basin. Additionally, for communities where the majority of samples are gathered from the Atlantic ocean, the nutrient levels are much higher than other communities.

To extend this analysis to other years, a scientific ship should collect water samples in different years. This would be very costly and even infeasible for many applications such as analyzing the seasonal changes in plankton communities. Therefore, in the second part of our work, we utilize the fact that the reflected light from the ocean's surface that is captured by a remote sensing satellite has a specific relationship with the plankton composition.

To this end, we first cluster the samples that are collected at the surface of the ocean. Next, we apply several machine learning algorithms to classify these representative communities from satellite data. Our top performing classifier reached 0.94 accuracy in leave-one-out cross validation setting. The results show three top important features in predicting communities are surface temperature, chlorophyll and particulate organic carbon.

Contents

1 Introduction 1

1.1 overview . . . 1

1.2 Aims and objectives . . . 2

1.3 Thesis outline . . . 2

2 Literature Review 3

3 Dataset and preprocessing 5

3.1 Datasets . . . 5

3.1.1 In-situ sample collection . . . 5

3.1.2 Satellite data . . . 8

3.2 Preprocessing . . . 8

3.2.1 Environment dataset . . . 9

3.2.2 Meta-barcoding dataset . . . 9

3.2.2.1 Analyse on all-depth samples . . . 9

3.2.2.2 Analyse on surface samples . . . 10

3.2.3 Satellite dataset . . . 11

4 Phytoplankton communities in all-depth samples 12

4.1 Methodology . . . 12

4.1.1 Constructing ASV network . . . 12

4.1.2 Community detection algorithms . . . 14

4.1.3 Edge satisfaction . . . 16

4.1.4 Dimension reduction technique . . . 17

4.2 Implementation on Prokaroyote . . . 17

4.2.1 Network of Prokaroyote . . . 17

4.2.2 Analyse environmental parameters . . . 19

4.2.3 Analysis of communities . . . 22

4.3 Implementation on Eukaryote . . . 24

4.3.1 Network of Eukaryote 24
4.3.2 Analysis on environmental parameters 26
4.3.3 Analysis of communities 28

5 Phytoplankton communities in surface samples 30
5.1 Correlation between in-situ and satellite 30
5.2 Methodology for community detection(surface water) 33
5.3 Implementation of community detection(surface water) 35
5.4 Community composition 40
5.5 Classify communities based on satellite features 42

6 Conclusion and Future work 45
6.1 Conclusion 45
6.2 Future work 46

Bibliography 48

List of Figures

3.1 Stations in the map Sebastián et al. (2021) 6
3.2 Sample distribution 7
3.3 Accumulate abundance for different rank 10

4.1 Prokaryote communities 18
4.2 Number of samples in each cluster of prokaryote 19
4.3 Depth for prokaryote 20
4.4 Temperature for prokaryote 20
4.5 Fluorescence for prokaryote 21
4.6 Salinity for prokaryote 21
4.7 NO3 for prokaryote 21
4.8 PO4 for prokaryote 21
4.9 SiO3 for prokaryote 22
4.10 Shepard diagram for prokaryote 22
4.11 NMDS for prokaryote labeled by basin & depth 23
4.12 NMDS for prokaryote labeled by basin & cluster 24
4.13 Eukaryote communities 25
4.14 Number of samples in each community of eukaryote 25
4.15 Depth of each eukaryote 26
4.16 Temperature for eukaryote 26
4.17 Fluorescence for eukaryote 27
4.18 Salinity for eukaryote 27
4.19 NO3 for eukaryote 27
4.20 PO4 for eukaryote 27
4.21 SiO3 for eukaryote 28
4.22 Shepard diagram for eukaryote 28
4.23 NMDS for eukaryote labeled by the basin & depth 29
4.24 NMDS for eukaryote based on the basin & cluster 29

5.1 Correlation between satellite and in-situ for Diato and green 32
5.2 Correlation between satellite and in-situ for Crypto and hapto 32
5.3 Correlation between satellite and in-situ for dino and prokar 33
5.4 Dendrogram for surface samples(hclust) 36
5.5 silhouette Width for hclust 37
5.6 DR for hclust 37
5.7 Dendrogram for surface samples(recluster) 38
5.8 silhouette Width for recluster 39
5.9 DR for recluster 39
5.10 Stations with their labels in the map 39
5.11 Composition in east 40
5.12 Composition in west_1 40
5.13 Composition in west_2 41
5.14 Composition in west_3 41
5.15 Composition in west_4 41
5.16 Composition in atlantic 41
5.17 Comparison between accuracy and f1 score of classification algorithms 44
5.18 Importance of features based on RF algorithm 44

Chapter 1

Introduction

1.1 overview

Phytoplankton are microscopic organisms that live in the sea. They are bases of aquatic food webs and play a critical role in maintaining the health and balance of the ocean and its complex food web. Biological production by phytoplankton facilitates the slow movement of carbon dioxide from the atmosphere into the deep ocean Lindsey (2010). Therefore, phytoplanktons have an important role in the earth's carbon cycle. Distinct phytoplankton taxa impact these essential ecosystem processes in a different way. Thus, to model the role of phytoplankton in the earth system, it is necessary to model the distribution and abundance of phytoplankton communities in the ocean.

Historically, scientific cruises travel in the ocean and collect and analyze water samples to infer the composition of phytoplankton. This is called in-situ measurement which provides precise information about abundance of different Phytoplankton taxonomies in different depths and locations. Additionally, the environment features, such as salinity, nutrient level, temperature and so on have been gathered as well. We should notice to the importance of different environmental features or physical barrier in the sea on distribution of phytoplanktons.

However, because of high expenses of research ships and expanse of the ocean it is not possible to collect the considerable amount of samples from different parts of the ocean in various time intervals. Additionally, we know that phytoplankton's characteristics such as size, pigment composition and distinct morphology along with their physiological state impact on how much sunlight they absorb and scatter. Therefore, satellite ocean color at hyperspectral resolution is expected to provide estimates of phytoplankton community composition from the space.

1.2 Aims and objectives

The initial objective of this study is finding communities of phytoplanktons in vertical and horizontal scales of the sea. Subsequently, we will determine the specific environmental features or barriers that are associated with each phytoplankton community.

The second goal is finding correlation between common features in in-situ and remote sensing satellite data.

The third goal is analyzing the samples that collected from surface water and detect communities of phytoplanktons in surface water. Then, we will find the phytoplancton composition for each community.

The fourth goal is mapping satellite observation to representative phytoplankton communities. In another word, we want to determine which satellite features play a crucial role in accurately identifying different phytoplankton communities.

1.3 Thesis outline

To reach the goals of this project, first we need to take a look at previous works in this field. In the Chapter 2 literature survey about related articles have been provided. After gaining insight about what other researchers did in this field, we can identify appropriate methodology to achieve project goals.

Before describing methods, we need to dig deeper about our dataset. In the Chapter 3, detail about each part of dataset and preprocessing strategies have been provided. In the section 4 and 5, method and implementation on all-depth samples and surface samples have been presented respectively. Finally, in the chapter 6, summary of this project and future work have been provided.

Chapter 2

Literature Review

Related works in this field can be categorized into two parts: analysing all-depth and surface samples. One of the most relevant works that has been done in all-depth samples is Sebastián et al. (2021). Authors in this study, tried to find which environmental features or physical barrier lead to changes in community structure of prokaryote. They found that physical barrier in Mediterranean Sea, Sicily and Gibraltar strait, have considerable effect on creating different prokaryote communities.

For surface samples, most of recent works focus not only on the information that gathered from water but also, consider satellite data and they demonstrated that many satellite-derived optical variables has strong correlation with phytoplankton composition. For example, satellite signals at regions with high phytoplankton abundance reduce light absorption coefficient, at 670 nm close to the pure water absorption coefficient. The study presented at Magnuson et al. (2004), has shown that the diatom, dominant phytoplankton type, has significantly lower absorption coefficient per unit [Chl a] compared with non-diatom species because of differences in their pigment packaging effect and intra-cellular pigment composition. In Zheng and DiGiacomo (2018), the authors have shown that the red-to-blue band can have a linear relation with the diatom fraction in Chesapeake Bay. The drive linear regression is used to assess the monthly changes in satellite-derived diatom fraction for the upper, middle, and lower Chesapeake Bay.

Kramer et al. (2022) modeled the global open ocean phytoplankton pigment composition from concurrent in-situ HPLC pigments and hyperspectral remote sensing reflectance, $R_{rs}(\lambda)$, datasets. HPLC or high performance liquid chromatography is a well established approach to measure Chlorophyll a and photosynthetic accessory pigments in water samples to determine phytoplankton composition and biomass of different algal groups.

The authors in Kramer et al. (2022) employed optimized principal components regression modeling to reconstruct phytoplankton pigments from $R_{rs}(\lambda)$. This study demonstrated that thirteen phytoplankton pigments that represents five phytoplankton pigments group (e.g., diatoms, dinoflagellates, haptophytes, green algae, and cyanobacteria) can be modeled from hyperspectral $R_{rs}(\lambda)$.

Although, it is very important to predict the size or taxonomic groups of phytoplanktons, understanding the global plankton network that includes both zooplankton and phytoplankton is getting more attention in recent years. In Kaneko et al. (2022), six plankton communities were identified from a global co-occurance network. Then a supervised classifier model was trained to predict these representative communities from satellite data. The trained model then used to predict the global distribution of these six representative communities over 19-years to understand seasonal changes at different geographical locations.

Chapter 3

Dataset and preprocessing

3.1 Datasets

To study the composition and diversity of phytoplankton communities, we study and analyze two different datasets. In this section, each of these datasets have been explained.

3.1.1 In-situ sample collection

To collect this kind of dataset, scientific cruises travel in the ocean and gather water samples at different depths. These samples will be analyzed at laboratories to specify the abundance of phytoplankton taxonomies in each sample. DNA metabarcoding and high-performance liquid chromatography (HPLC) pigment are two widely used methods for assessing phytoplankton composition. In this study we use the dataset that has been analyzed via the meta-barcoding approach.
Metabarcoding refers to the process of sequencing specific regions of DNA to identify the abundance of different taxonomies within the sample Cristescu (2014). The term ″meta″ refers to the comprehensive analysis of multiple genetic sequences to understand the broader species community rather than focusing on individual species or organisms.

Our in-situ dataset includes data collected from 29 stations during the period of April 29th to May 28th, 2014. These stations have been displayed in the figure 3.1. These stations started from east of Mediterranean sea to adjacent subtropical northeast (NE) Atlantic ocean near to the Canary Islands. At each station, sample of water collected at different depths, for example minimum depth in most of stations is 3 meters and maximum depth is 10 m above the seafloor.

The depth measurements, taken at various stations, show that the smallest maximum depth recorded is 300 meters, while the largest maximum depth recorded is 4539 meters. In total we have 188 samples collected at various depths and stations. In-situ dataset has been divided to three tables that are described in below.

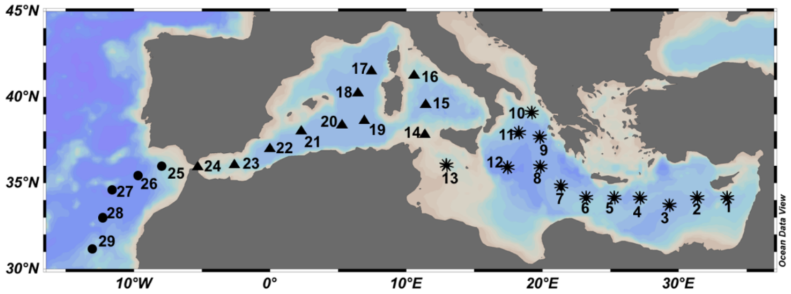

Figure 3.1: Stations in the map Sebastián et al. (2021)

- Environmental information

  The first table refers to the environmental information for each of the collected samples. Environmental features consist of Temperature, Salinity, Fluorescence, chlorophyll, NO3, PO4 and SiO3. As mentioned before, samples collected in different depths, each depth categorised based on this article Junger et al. (2023). In the table 3.1 range of depths related to each category have been provided. The first part which refers to Epipelagic divided to two parts: surface of the water (SRF) and deep chlorophyll maxima (DCM).

Depth category	Range
Epipelagic (SRF)	less than 50 m
Epipelagic (DCM)	between 50 m and 200 m
Mesopelagic (MES)	between 200m and 1000 m
Bathypelagic (BAT)	higher than 1000 m

Table 3.1: Range of depth for each category

On the other side, stations that samples have been collected categorised based on various basins. In the table 3.2, these basins and their geographic locations have been presented.

In the figure 3.2 distribution of samples based on different basins and depth categories provided. As we can see, the number of samples collected from east of Mediterranean are more than other basins. While, we have less number of samples from Atlantic ocean.

Basin name	Stations	Geographic area
East of med sea	stations from 1 to 13	from east to Sicily island
West of med sea	stations from 14 to 24	from sicily island to Gibraltar
Atlantic	stations from 25 to 29	Atlantic ocean

Table 3.2: Range of basin

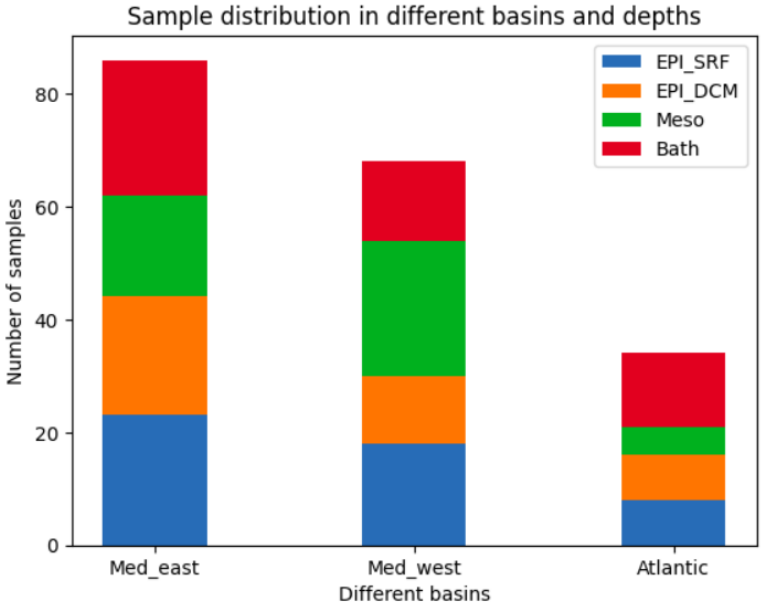

Figure 3.2: Sample distribution

- Prokaryote
  This dataset refers to metabarcoding of prokaryote and provides information on the diversity and abundance of bacteria in samples. The statistics related to prokaryotes have been derived through the sequencing of the 16S rRNA gene. The output of this analysis typically is a table that amplicon sequence variants (ASVs)[1] exist in rows and samples are in columns. Additionally, the dataset provides taxonomic classifications of each ASVs based on their sequence similarity. The output consists 50,762 ASVs and their abundance in 188 samples.

- Eukaryote
  Eukaryotes refers to the microbes in the sea. This is another in-situ dataset that has been obtained via metabarcoding on 18s rRNA gene. Eukaryote

[1] ASVs are derived from the sequencing of individual DNA molecules

abundance table is similar to prokaryote table and it gives us abundance of ASVs in different samples. In the eukaryote table, 103,421 ASVs for 179 samples have been considered.

3.1.2 Satellite data

With respect to latitude and longitude of each station and the date that samples have been collected, we can download the remote sensing satellite data.
Since, satellites can give us information of surface [2] water, we should consider samples that collected in shallower part of the sea. In this study, the maximum depth that satellite can capture is 20 meters. Satellite data has been downloaded from Copernicus website. Copernicus is one of the satellites that is launched by the European Space Agency (ESA).
To retrieve satellite features, a python script has been prepared which product id, service id, the desired features in specific latitudes, longitudes and dates defined in this script. After running this script the desired features will be downloaded. The features that will study in this project are listed in below.

- Remote sensing reflectance (Rrs (λ)) from 5 visible light wavelength 412, 555, 443, 670 and 490 website (reflectance Atlantic) website (reflectance Med sea).
- chlorophyll-a (CHL) website (poc chl product).
- concentration of particulate organic matter expressed as carbon in sea water (poc) website (poc chl product).
- Sea surface temprature (SST) website (SST product).
- Volume attenuation coefficient of downwelling radiative flux in sea water (KD490) website (KD490 Atlantic ocean) website (KD490 Med sea).
- Amount of chlorophyll in the taxonomies such as Diatom, Dino, Crypto, Hapto, Prokar and Green website (Plankton Med sea)

3.2 Preprocessing

In his section, the strategies that utilized to deal with missing values and some filtration that should be applied on the aforementioned dataset have been described.

[2] In the environment information the surface water refers to the depth lower than 50 m, while when we are considering satellite dataset the surface water related to depth lower than 20 m.

3.2.1 Environment dataset

There are NaN values in some columns of environment dataset. We have eight samples which have missing values in some features such as SiO3, PO4, NO3, temperature and salinity. Based on the knowledge of experts in this domain, missing values imputed by mean of nearest neighbors. These neighbors should have two features:

- Their stations should be near to the station that has sample with missing values. For example, we should not use information of Atlantic ocean to impute missing values from east of Mediterranean sea.
- Their depth should be in the same range, with regard to the depth table 3.1.

Based on the Righetti et al. (2019), stations that their sea floor is less than 200 meters considered as coastal area and the samples that collected from these stations should be removed. In the environmental table, smallest maximum depth is 300 meters. Therefore, none of the samples classified as coastal area and all of them have been kept.

3.2.2 Meta-barcoding dataset

Based on the goals that defined in the section 1.2, we need to find communities in whole samples and in surface samples. Steps to prepare metabarcoding dataset for each of these analysis are different and will be described in below.

3.2.2.1 Analyse on all-depth samples

Metabarcoding tables are very sparse and there are some ASVs that their abundances are zero in most of samples. Such ASVs will bring noise to the analysis and make it hard to explain variation and pattern in the data Pedro (2021).
Based on the Kaneko et al. (2022), to remove ASVs that have zero abundance in most of samples, we can apply filtration on abundance of ASVs. The filtration that has been applied in the prokaryote and eukaryote metabarcoding has some minor changes with each other. In the Prokaryote dataset, ASVs with minimum occurrence larger than 5 (5% reads) in at least 10% of samples (19 samples) were considered. Therefore, number of ASVs in prokaroytes reduced from 50762 to 1241. In the figure 3.3 accumulate abundance of top 100 ASVs has been presented. Based

on this plot, number of ASVs that have very high accumulate abundance are significantly lower than the number of ASVs with small accumulate abundance. It is a reason for sharp decline in y axis. Additionally, most of accumulate abundances for ASVs are less than 1000. It is a reason for long tail in this plot.

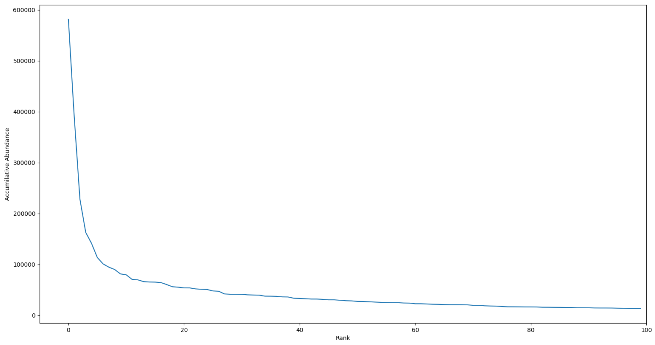

Figure 3.3: Accumulate abundance for different rank

In the eukaryotes metabarcoding dataset, there are eight missing samples. In total, we have 179 samples instead of 188 samples in this table.

In this dataset, ASVs with a minimum occurrence larger than 5 (5% reads) in at least 5% of samples (8 samples) are considered. Therefore, number of ASVs reduced from 103421 to 1760 ASVs.

3.2.2.2 Analyse on surface samples

To generate a metabarcoding dataset for the purpose of identifying communities in surface water, several steps need to be followed. First, the samples that collected in surface water (depth lower than 20 m) should be filtered. Second, metabarcoding for prokaroyte and eukaroyte should be combined in one table. Third, we need to filter ASVs that do photosynthesis for living. If a taxonomy do photosynthesis, it has chlorophyll (chl) to capture sunlight and then in the photosynthesis procedure the sunlight turn into chemical energy Lindsey (2010). Therefore, for metabarcoding dataset we consider the taxonomies that have chl and can be observed from space due to their chl content.
After applying these three steps, we will have 23 columns that are pointing to the number of samples and 3036 rows or ASVs that fulfilling the above condition.

3.2.3 Satellite dataset

There are some nan values for satellite dataset as well. For example, for specific latitude and longitude or specific dates there were not data.

Based on the El Hourany et al. (2022), we can apply two approaches to deal with missing values. First, we can consider a box around the location that sample has been collected. Second, considering temporal window. In this approach instead of considering exact day that sample has been collected, a period of n days before and after that specific date will be taken into account. Indeed, we want to estimate missing data by nearest spatial and temporal data.

In this study, the matchups between satellite observations and in-situ samples selected by considering 5x5 pixel boxes[3] around the in-situ coordinates and ∓ 2 days around the day of the in-situ measurement.

[3] 1 pixel is equal to 1 km.

Chapter 4

Phytoplankton communities in all-depth samples

4.1 Methodology

In this section, we will focus on the all-depth samples and try to find communities for prokaryotes and eukaryotes separately. In the first step, we will build a network of ASVs that will show ecological connections between ASVs. Second, different community detection algorithms will be run on the network to find which one of ASVs have stronger connection and will be in the same community. After finding communities, we need to know the dominant community for each sample. Therefore, the Edge satisfaction (ES) formula has been used to assign a sample to a community based on the highest ES value . In the final step, we will analyse and compare environmental features in each community.

4.1.1 Constructing ASV network

In the first step, ecological network[1] of ASVs has been created based on co-occurrence or co-abundance patterns. This network created with FlashWeave package that is implemented in Julia Programming Language Git. FlashWeave is based on local-to-global learning (LGL) framework and tries to estimate pairwise association between nodes. In the first step, it tries to derive all directly associated neighbors of a target node and then connect these nodes by a combinator strategy to create global association graph . This strategy will connect two nodes when one of them is in the derived directly associated neighborhood of another one Tackmann et al.

[1] When we have enough number of samples it does make sense to create a network of ASVs

(2019). The setting that used for creating a network is based on the below parameters:
"heterogeneous=False", "sensitive=True", "alpha=0.05".
When heterogeneous selected as false, means that the selected mode is homogeneous. This mode is selected when we have moderate[2] number of samples. Therefore, because we have thousand number of ASVs in each group this mode has been selected. For the sensitive parameter the True value has been selected and it means the network created based on the fine grained association that focus on localize relationship between nodes. On the other side, when the False value assigned to this parameter, the algorithm will run faster but in the coarse grained associations. Another parameter that is defined is alpha, it is a threshold which is used to determine statistical significance.
At the end, FlashWeave will give us a table that consists source and destination node and the edge weight that connect these two nodes. Weights are in the range between -1 and +1 and considered as an attribute for an edge that connect two ASVs.
If FlashWeave assigns positive weight to an edge, it means that if abundance of one of ASVs increases, abundance of another ASV will increase as well. This situation will happen when one of below relations exists between ASVs:

- Commensalism: It is a form of symbiosis that involves a long-term biological interaction where one species obtains benefits from another species, while the other species neither benefits nor suffers any harm through another speciesWilson (1975).

- Mutualism: It refers to an ecological interaction involving two or more species, where each species derives a net benefit from relation that has with another species Bronstein (2015).

On the other side, if the weight between two ASVs is negative, it means that their abundance have negative correlation to each other. In another word, with increasing abundance of one of them, the abundance of other one will decrease. This situation will happen when one of below relations exists between ASVs:

- Competition: It is a biological interaction between species or organisms that share a limited resource, such as water, food, or other necessities and they are living in the same place Haslett (1997).

[2] In this case moderate means hundreds or thousands of samples

- Predation: Where one organism, the predator, kills and eats another organism.

4.1.2 Community detection algorithms

Based on the output of flashweave, we have some negative weights. In the study Kaneko et al. (2022), authors removed negative weights. The reason behind removing negative edges is related to the Edge Satisfaction (ES) formula that will be used in the 4.1.3. If we consider negative edges, the summation over weight of edges in a community can be equal or less than zero. In the case that summation is equal to zero, the denominator will be zero and the fraction will be undefined. Moreover, removing negative edged does not affect on the main purpose, because we want to consider nodes that have strong connection together in one community. Strong connection means that the edge that connects two nodes has positive weight. If the edge between two nodes has negative weight, these two nodes should not be in a same community.

In the next step, the community detection algorithms should be applied. A key characteristic of a community is the dense interconnectedness among its internal nodes. In another word, nodes in one community should have strong connection with each other.

One of the measurements for the quality of a network partition is modularity index Newman and Girvan (2004). Modularity index will help us to measure how nodes in each community, connected together or how partition could separate nodes and assign them to a suitable community Clauset et al. (2004). The question that arises is how to achieve a high modularity index. When nodes with strong connection assigned to one community and nodes with poor connection assigned into different communities.

Tow different type of algorithms have been applied on the network and the best one selected based on modularity index value. In the below, formula for modularity index has been provided:

$$Q = \frac{1}{2m} \sum_{u,v} \left[\sigma(u,v) - \frac{k_u k_v}{2m} \right] \delta(c_u, c_v) \tag{4.1}$$

Q is the modularity index.

$\sigma(u,v)$ is an edge weight between node u and v

k is the degree of the nodes in the network

m is the total number of edges in the network

$\delta(c_u, c_v)$ is the Kronecker delta function, which equals 1 if nodes u and v belong to the same community (c) and 0 otherwise.

In below a short description of algorithms that applied on our dataset provided.

- Louvain
  The Louvain algorithm is an unsupervised, agglomerative (or bottom-up) clustering algorithm that does not require the number of communities as a priori information. In the first phase, each node represent as a community. Therefore, number of communities is equal to number of nodes. Then, the algorithm iteratively merges communities to improve the modularity score. This phase will run until the merging communities do not increase the modularity score.
  In the second phase, the communities found in the first phase are treated as nodes in a new network, where the edge weight between two communities is the sum of the weights of the edges between their constituent nodes in the original network. The first phase is repeated on this new network and the resulting communities are considered as the final partition. Blondel et al. (2008). One of the advantages of the Louvain algorithm is its efficiency and relatively low time complexity, which is $O(n.logn)$, n refers to number of nodes.

- Girvan-newman algorithm
  The Girvan-Newman algorithm, like the Louvain algorithm, is an unsupervised approach that does not rely on prior knowledge of the number of communities. However, it follows a divisive (top down) strategy instead of an agglomerative one. It considers whole network as a single community and iteratively removes edges that have highest betweenness centrality. Betweenness centrality quantifies how often an edge lies on the shortest path between pairs of nodes.
  In each iteration, modularity index has been calculated and when it does not improve compare to previous modularities, the algorithm will stop. The time complexity of this algorithm is much higher than Louvain algorithm, it is $o(m^2.n)$, m and n are referring to the number of edges and nodes respectively. Watson (2022).

4.1.3 Edge satisfaction

Edge satisfaction index introduced in Kaneko et al. (2022) and determined which community dominated in each sample. In below, formula for edge satisfaction has been provided.

$$ES_{c,i} = \frac{\sum_{u,v \in C} \sigma(u,v) min(p_i(u), p_i(v))}{\sum_{u,v \in c} \sigma(u,v)} \tag{4.2}$$

c is representative of a community and i is representative of a sample.
$\sigma(u, v)$ is an edge weight between node u and v (u and v refer to ASVs)
$p_i(u)$ is sigmoid transformation of centered log-ratio (clr) transformation of node u in sample i.
The question can be arose is that why we need to apply log-ratio transformation? First, we want to capture relation between features. Then, we will take the logarithm of these ratios. At the end, log-ratio makes the data symmetric and linearly related. Additionally, we should mention that the clr formula introduced by Aitchison (1986) and it presented in below. After applying clr transformation, we need to map retrieved values in a range between 0 and 1, therefore sigmoid transformation has been used.

$$x_{clr} = (\ln(x_1/G(x)), \ln(x_2/G(x)), \ldots, \ln(x_D/G(x))) \tag{4.3}$$

$$G(x) = \sqrt[D]{x_1 x_2 \cdots x_D} \tag{4.4}$$

G(x) is geometric mean of vector $x = [x_1, x_2, ..., x_D]$. Vector x refers to the sample and each element of this vector, $x_1, x_2, .., x_D$, refers to the abundance of different taxonomies in this sample. In this case, imagine we have D taxonomies or ASVs. As previously mentioned, in some samples, the abundance of certain ASVs is zero. This leads to the geometric mean of this sample being zero. To deal with this problem, zero abundances replaced with very small value.
Briefly, the Edge satisfaction index measures the ratio of the number of edges between existing nodes or ASVs in a given sample and the number of all the edges within a given community. This index will give us value between 0 and 1. If the $ES_{c,i}$ is equal to 1 means that the cluster c is dominant in sample i. On the other side, if this index is equal to 0, means that the sample i does not belong to cluster c. Es index should be calculated for all samples(s) and communities(c). Therefore, we will have s*c values for edge satisfaction. For each sample the ES with highest value should be selected and it shows that which community is dominant for that sample.

4.1.4 Dimension reduction technique

To find relation between samples and how they are close to each other, dimension reduction technique on the metabarcoding table or abundance of ASVs in different samples has been implemented. One of the techniques that is widely used to reduce dimensions in ecological dataset is Non-metric Multidimensional Scaling (NMDS). This technique is particularly useful when dealing with sparse abundance data Sebastián et al. (2021).

There are two popular dissimilarity functions in NMDS, Jaccard and Bray-Curtis. The Jaccard distance is based on presence/absence of taxonomies and it does not consider abundance of taxonomies. While, Bray-Curtis considers not only presence or absence of specific taxonomies, but also consider relative abundance of them in a sample. Therefore, this feature makes Bray-Curtis more accurate compare to Jaccard dissimilarity ZACH (2022). In the 4.5, formula for Bray-Curtis dissimilarity has been provided.

$$BC_{ij} = 1 - \frac{2c_{ij}}{s_i + s_j} \tag{4.5}$$

BC_{ij}: Bary-Curtis dissimilarity between sample i and j.
C_{ij}: Sum of the smaller abundances in sample i and j.
S_i, S_j : Accumulative abundance for sample i and j respectively.

To measure the goodness of the NMDS technique, we can calculate the stress value. It will give us intuition that how much the NMDS preserves the similarity between samples in reduced dimensions. The lower stress we get, the original distance and projected distance are closer together. In the Clarke (1993) a guideline for acceptable stress value has been suggested. According to this guideline, $strees \leqslant 0.05$ provides excellent preservation in NMDS dimension, $strees \leqslant 0.1$ indicates good representation, $strees \leqslant 0.2$ falls into the category of usable representation and $strees \geq 0.2$ provides poor representation.

4.2 Implementation on Prokaroyote

4.2.1 Network of Prokaroyote

After running FlashWeave on prokaryote metabarcoding dataset with 1241 ASVs, we will have a table of source, destination nodes and edge weight that connect source to destination nodes. Before creating a graph, we should remove edges with negative weight, therefore we will have 4595 edges instead of 4685 edges. To build a graph from output table, the networkx package in python has been used.

We applied two community detection algorithms, Louvain and Girvan-newman. Among these algorithms, Louvain selected because its result was so close to Girvan-newman algorithm and it runs much faster than it. We could reach to modularity index equal to 0.64.

Based on the Figure 4.1, Louvain algorithm, clustered prokaroyote dataset into 7 communities. Nodes or ASVs that belong to one community have same color and shape. Therefore, various communities have different shapes and colors. In the right hand side of this figure, we have community 1 and 5 that are close to each other and separated from left hand side communities.
After community 5, we have community 2 that spreads in a wide range and has some overlapping with community 6. Then, we have community 0 and 4 that are very close to each other and separated from other communities. At the end, we have community 3 that has few members compare to other communities. Some of its members are close to community 4 and some of them are near to community 5.

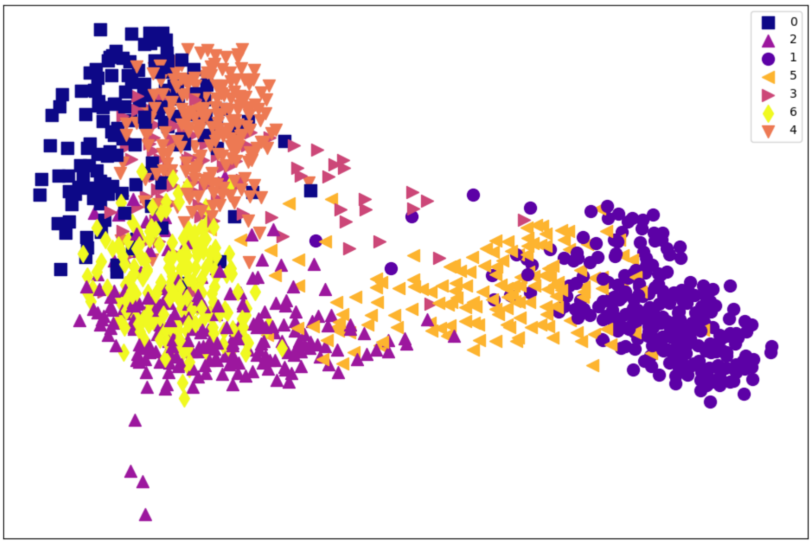

Figure 4.1: Prokaryote communities

After creating the network of ASVs, edge satisfaction formula that introduced in the section 4.2 has been calculated[3] for all samples and communities. For each sample we have 7 possible communities or 7 Edge satisfaction. Between these values the community with the maximum ES will be selected as dominant community for that sample. In the figure 4.2, number of samples that belong to each community or cluster have been presented. Based on this figure, we have more than 50 samples that community 1 is dominant on them. On the other side, we have lowest member of samples in community 4.

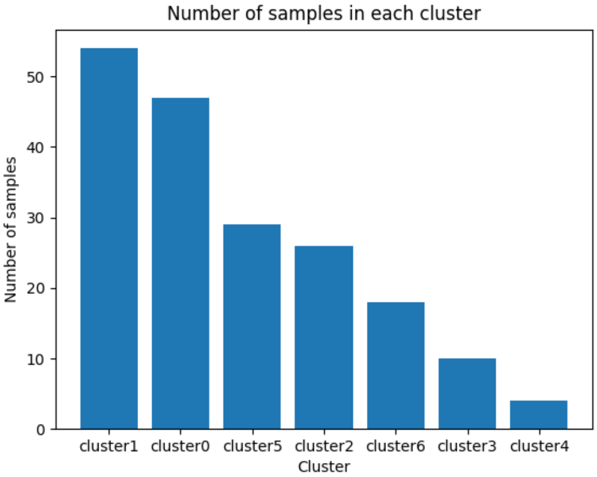

Figure 4.2: Number of samples in each cluster of prokaryote

4.2.2 Analyse environmental parameters

In this section, we will try to answer this question, what environmental features are driving the formation of specific communities within a network? Based on the environmental features for the samples that belong to a specific community, we draw a box plot. Then, we will compare these features for different communities.

Depth: In the figure 4.3, distribution of depth of all samples that assigned to each cluster has been provided. As we can see, cluster 1 and 5 are representative for samples in shallower part of the sea. When we are moving to the right hand side of the plot, the mean of depth will increase. Highest depth that sample have been collected belong to cluster 3. Additionally, this cluster has wider range of depths compare to other clusters. As a result, if two communities were close in the figure 4.1, the depth that their samples collected is in the same range.

[3] Edge satisfaction formula has been implemented in the python

Temperature: In the figure 4.4, we can observe that temperature is decreasing when we are moving from cluster 1 to the cluster 3. This agree with our previous plot, because the temperature of surface water is higher than deeper part of the sea. It does worth to mention that in cluster 3 we have 2 outliers that are representing the samples that collected from shallower part and their temperature are higher than other points in this cluster.

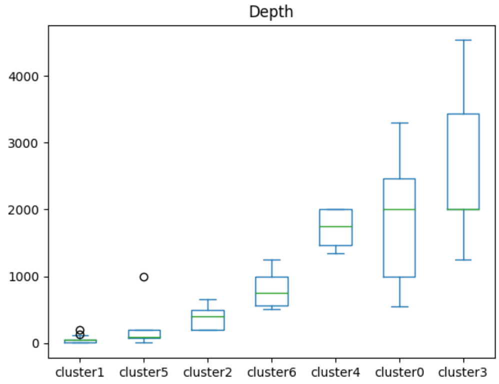

Figure 4.3: Depth for prokaryote

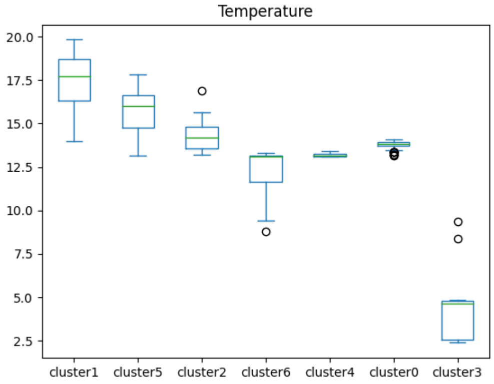

Figure 4.4: Temperature for prokaryote

Fluorescence: In the figure 4.5, fluorescence range in each cluster has been shown. Based on our observation from the previous figures, cluster 1 and 5 are representing for shallower part of the sea. In this figure, these two clusters show higher range of fluorescence, which makes sense due to the presence of sunlight in the surface water. Consequently, we observe more fluorescence. On the other side, other clusters that are representative for deeper part of the sea have near zero fluorescence.

Salinity: In the figure 4.6, salinity of different clusters show that for cluster 1, 5, 2, 6, 4 and 0 we have high salinity and for cluster 3 we have considerable lower salinity compare to other clusters.

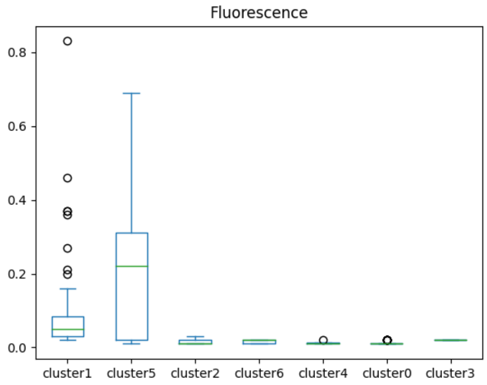

Figure 4.5: Fluorescence for prokaryote

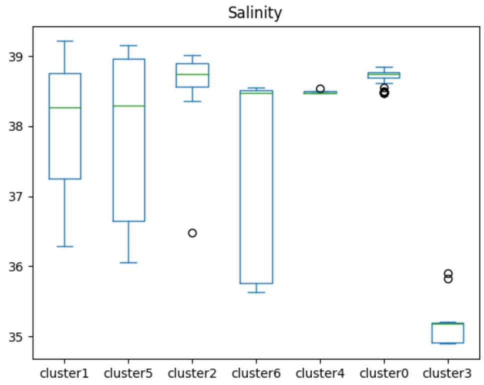

Figure 4.6: Salinity for prokaryote

Nutrients: In the figures 4.7 , 4.8 and 4.9, nutrients such as NO3, PO4 and SiO3 are analyzed respectively. These nutrients show similar behaviour in clusters. When we are moving from shallower part of the sea to the depth near to 1000 m, the amount of nutrient will increase. On the other side, when the depth increased more than 1000 m, cluster 4 and 0, the level of nutrient is decreasing. While, in cluster 3 that has samples collected from deeper part of the sea, the nutrient level is higher than other clusters.

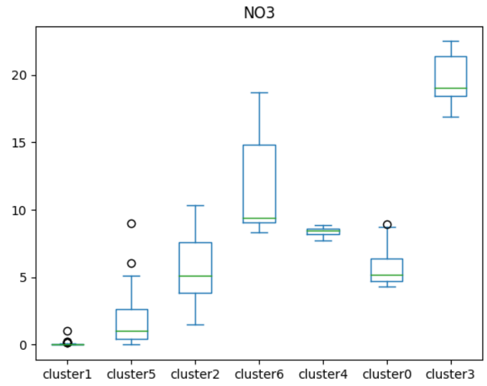

Figure 4.7: NO3 for prokaryote

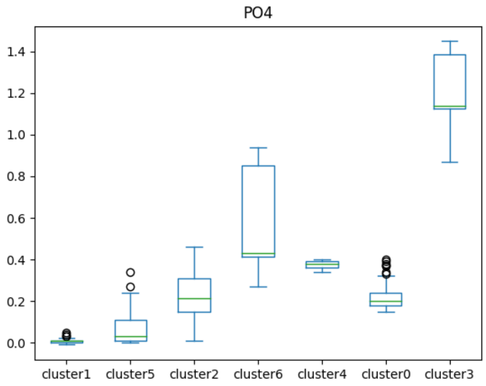

Figure 4.8: PO4 for prokaryote

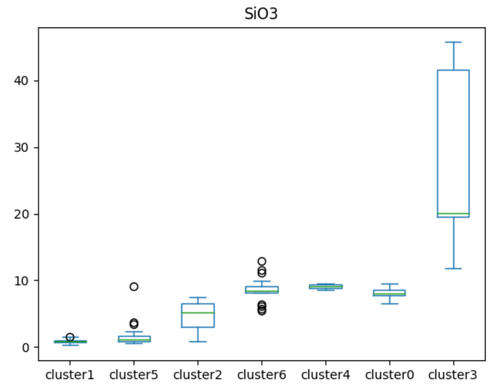

Figure 4.9: SiO3 for prokaryote

4.2.3 Analysis of communities

Each sample in the prokaryote matabarcoding dataset has 1241 dimensions or ASVs. To visualize this multi-dimensional dataset, we first need to reduce dimensions. The metaMDS function from vegan package in R has been used to reduce dimensions from 1241 to two dimensions. After 100 iterations, stress reaches to 0.096 that shows NMDS is good representation for our dataset. To visualize the relation between dissimilarity in real observation and dissimilarity in reduced dimensions, we can also utilize Shepard diagram. In the figure 4.10, the Shepard diagram has been provided. In this plot, with increasing dissimilarity between real observation, dissimilarity between reduced dimensions, ordination distance, will increase as well. Therefore, it shows good performance of NMDS in prokaryote.

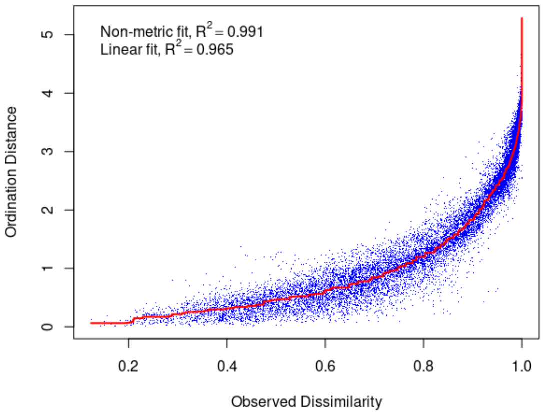

Figure 4.10: Shepard diagram for prokaryote

In addition to showing samples in two dimensions, we can assign label for each of samples to clarify how samples separated from each other. This label can be dominant community for each sample or depth and basin of samples.
In the figure 4.11, shape and colour of sample points are specified based on basin and depth that samples gathered respectively. According to this figure, the samples on the right hand side of the plot correspond to the surface and DCM ($depth \leq 200m$) area. While, we have a sample from DCM part of Atlantic ocean that exist very far from other points.
When we are moving to the left hand side of the plot, there are samples from deeper part of the sea, Mesopelagic ($200m \leq depth \leq 1000m$) and Bathypelagic ($depth \geq 1000m$).

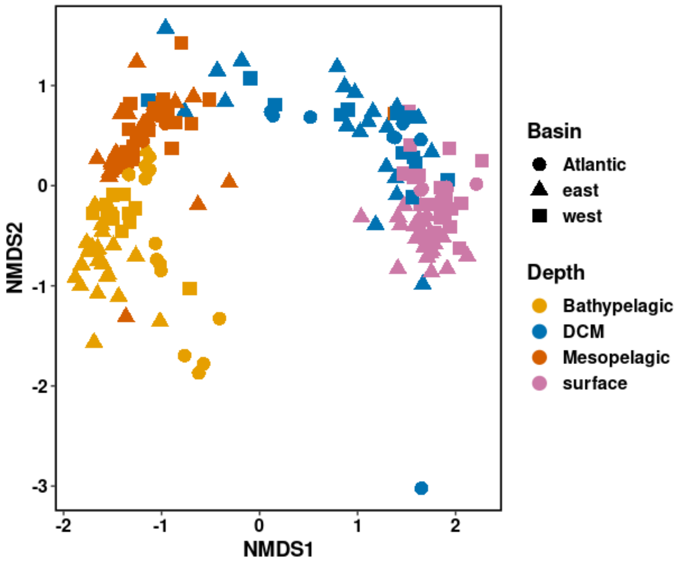

Figure 4.11: NMDS for prokaryote labeled by basin & depth

In the figure 4.12, label of samples are based on their dominant cluster and basin. For the shallower part of the sea there are three clusters, cluster 1, 5 and 2, and most of samples in these clusters belong to east and west part of Mediterranean sea. These clusters were also close in the ASV network in the figure 4.1.
Samples in the left side of the plot belong to 4 communities, cluster 6, 4, 0 and 3. Most of samples in cluster 6 are collected from west of Med sea and Atlantic ocean. While, in cluster 0 we have samples from east and west of Med sea that their depth is much higher than other clusters. At the end, we have cluster 3 that belongs to samples in Atlantic ocean. Samples in this cluster are widely spreading in the plot. The samples that collected from shallower part of Atlantic are near to cluster 6

and 0 and the samples collected from deeper part are located further than other samples. Based on the domain knowledge of biologist, this behaviour is related to the water flow between Atlantic ocean and Med sea in the shallow part of the sea. Additionally, we can conclude that the different behaviour of cluster 3 in previous section is related to the basin that the samples of this cluster collocated.

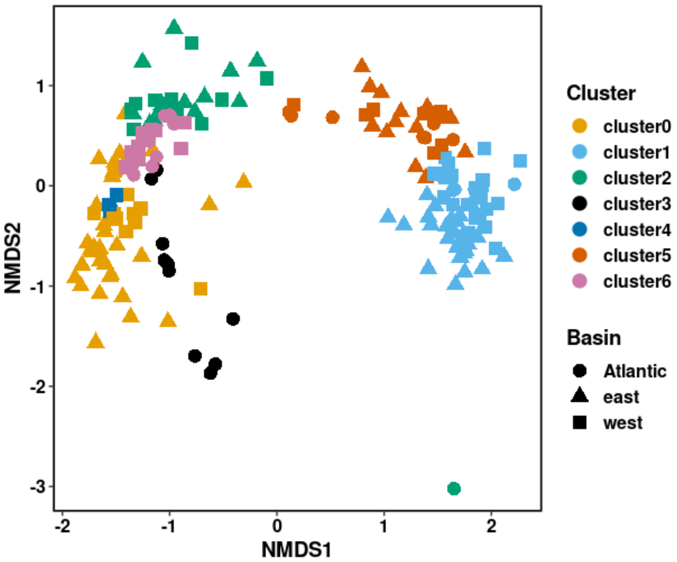

Figure 4.12: NMDS for prokaryote labeled by basin & cluster

4.3 Implementation on Eukaryote

4.3.1 Network of Eukaryote

Like the steps that have been done for prokaryote dataset, we will do these steps on eukaryote dataset as well. After running FlashWeave and filtering negative weights, the network of ASVs created by networkx. This network has 1760 nodes and 7464 edges. Then, we need to know which one of community detection algorithm give us better modularity score. Louvain gives higher score and in the figure 4.13 the communities that specified by this algorithm has been presented. In the upper right side of this plot, we have three communities, 6, 5 and 1 that are located very close to each other. On the other side, in the bottom left side of the plot, we have another three communities, 2, 3 and 0 that are connected to communities in the right hand side by a bridge that is community 4.

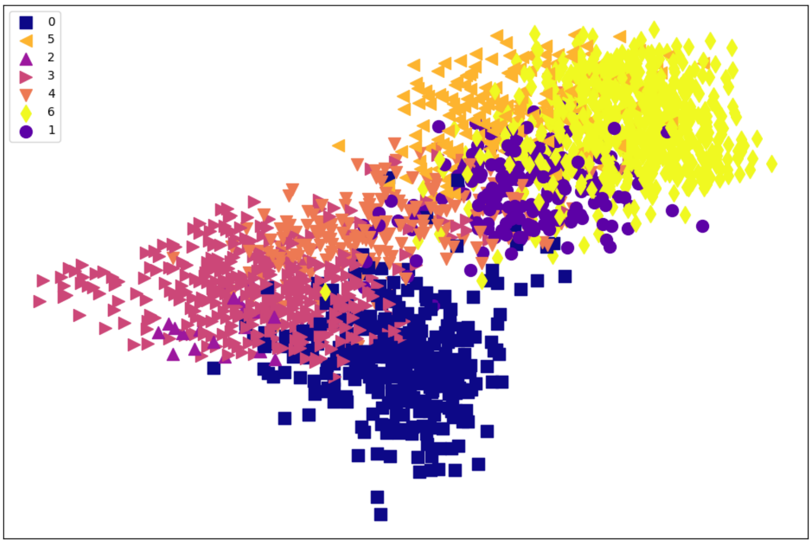

Figure 4.13: Eukaryote communities

In the next step, we should know that which one of samples belong to which one of communities. Therefore, edge satisfaction has been calculated for each of samples. In the figure 4.14, number of samples that assigned to each cluster have been presented. Based on this diagram, community 0 is dominant for most of samples and community 1 is less dominant.

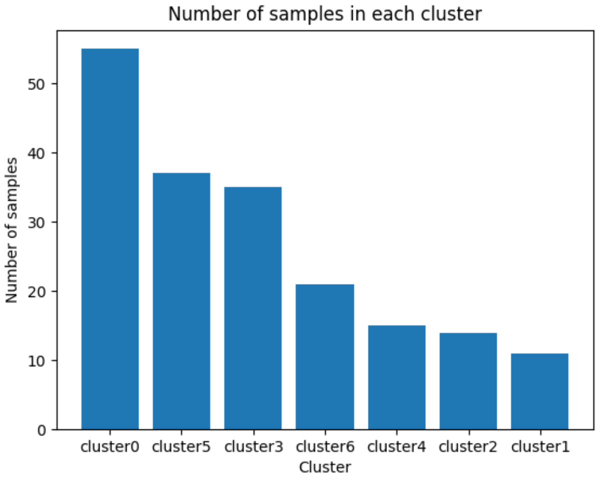

Figure 4.14: Number of samples in each community of eukaryote

4.3.2 Analysis on environmental parameters

Similar to the analysis that had been presented for prokaryote communities, in this part we will analyse environmental features for eukaryote communities.

Depth: In the figure 4.15, depth of samples in each community presented. Based on its result, cluster 6, 5, 1 and 4 are representing for lower depth of the sea. while, cluster 3, 2 and 0 are dominant clusters for deeper part of the sea.
Temperature: In the figure 4.16, temperature for each cluster provided. As we are moving to the deeper part, temperature is decreasing. For example, cluster 2 is represented for the depth higher than 1000 m and the temperature associated with this cluster is significantly lower than other clusters.

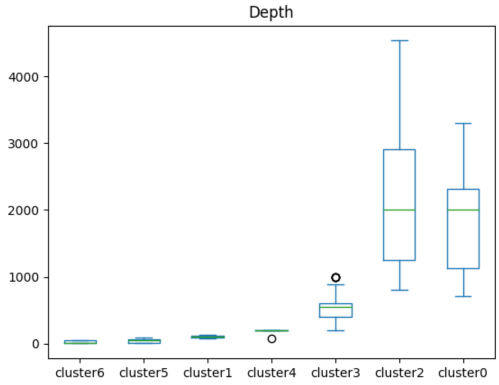

Figure 4.15: Depth of each eukaryote

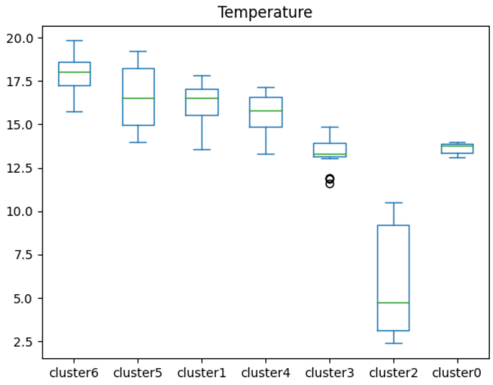

Figure 4.16: Temperature for eukaryote

Fluorescence: In figures 4.17, fluorescence and salinity for different clusters have been presented. The average of fluorescence for cluster 5 and 1 are higher than other clusters, because these clusters consist of samples that are collected from DCM and surface part of the sea.
Salinity: Based on the figure 4.18, salinity for cluster 2 is much lower than other clusters and cluster 1 has highest salinity.

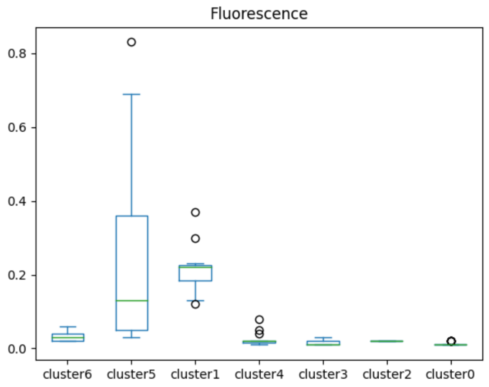

Figure 4.17: Fluorescence for eukaryote

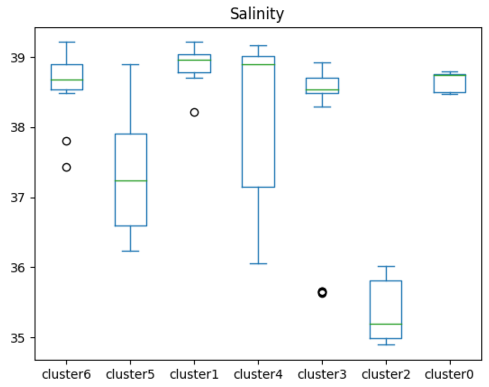

Figure 4.18: Salinity for eukaryote

Nutrients: In the figures 4.19, 4.20 and 4.21, amount of NO3, PO4 and Sio3 for different communities have been provide respectively. For all these three nutrients, we have the same behaviour in all communities. When the depth increases to 1000 m the nutrient level increases as well. In the cluster 2 there is sharp increase in the level of nutrients and in cluster 0 the nutrient level decreases and reaches to even lower value compare to cluster 3.

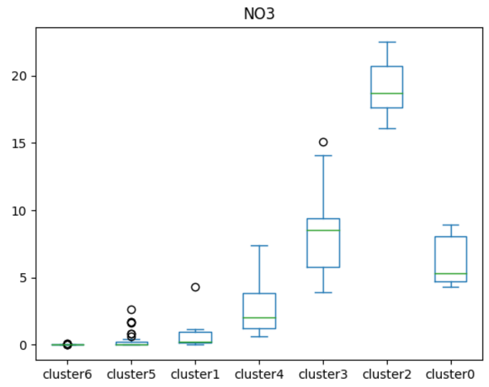

Figure 4.19: NO3 for eukaryote

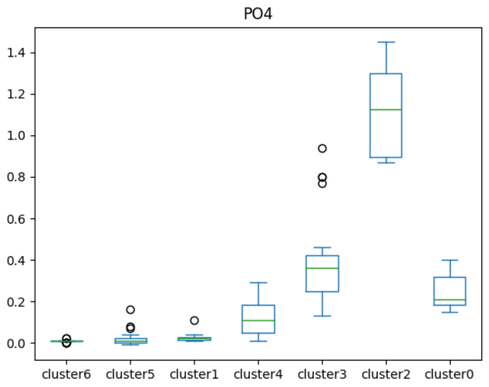

Figure 4.20: PO4 for eukaryote

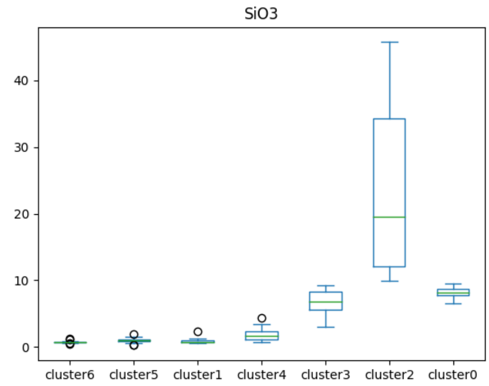

Figure 4.21: SiO3 for eukaryote

4.3.3 Analysis of communities

To reduce dimensions for eukaryote, the metaMDS function has been used to reduce dimensions from 1760 to two dimensions. After 100 iterations, stress reaches to 0.107 that shows NMDS is approximately good representation for our dataset.

Additionally, Shepard diagram that is provided in the figure 4.22 shows that with increasing dissimilarity between observed values, the dissimilarity between reduced dimensions values increases too.

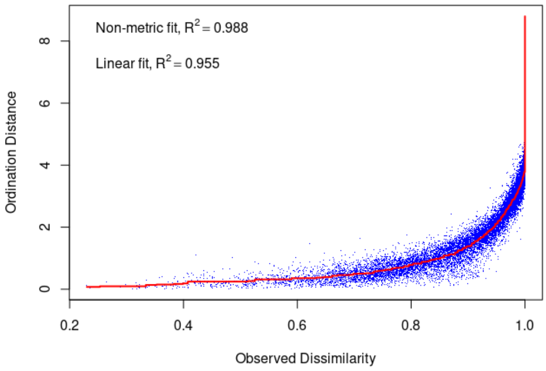

Figure 4.22: Shepard diagram for eukaryote

In the figure 4.23, the samples are presented in 2 dimensions based on their depth and basin. The distribution of samples in this plot is approximately as same as prokaryote structure that presented in the figure 4.11. Distribution of samples starts from surface water in the right hand side of the plot and when we are moving to the left side the depth is increasing. In the prokaryote, samples that were collected from deeper part of Atlantic were exist far from other samples. While, in the eukaryote most of these samples are close to deeper part of west of Med sea.

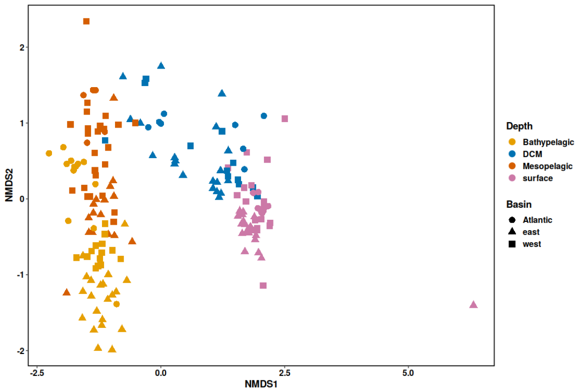

Figure 4.23: NMDS for eukaryote labeled by the basin & depth

Based on the figure 4.24, samples in NMDS are specified by the cluster and basin. In the right side of the plot, there are four communities, 6, 5, 1 and 4 that are representing for shallower part of the sea, surface and DCM. In the left side of plot, three clusters, 3,2 and 0, are presenting the deeper part of the sea.

As we saw in the section 4.3.2, cluster 2 that had different behaviour compare to other clusters, has samples that collected from deeper part of the Atlantic ocean.

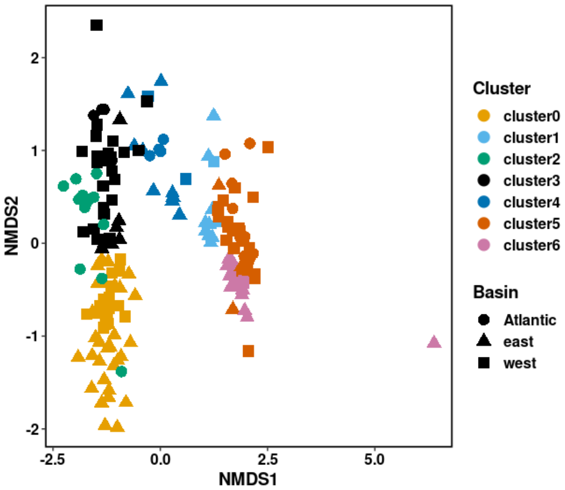

Figure 4.24: NMDS for eukaryote based on the basin & cluster

Chapter 5

Phytoplankton communities in surface samples

In this chapter, we will try to find behaviour of samples in surface water. Additionally, we will use satellite features for this part, because as mentioned before satellite can capture information of surface water. In the first part, correlation between satellite and in-situ data will be studied. Then, we will find communities of surface water by in-situ dataset. At the end, we will analyse possibility of prediction presented communities with satellite features.

5.1 Correlation between in-situ and satellite

In one of the product of Copernicus satellite dataset website (Plankton Med sea), amount of chlorophyll that represented in some of taxonomies such as Diatom, Dino, Crypto, Hapto, green algae and prokar for Mediterranean sea presented. Therefore, we can retrieve these values for each of stations in Med sea. In the case that the value for specific location did not exist, because of cloudy weather or other reasons, the preprocessing steps, described in the section 3.2.3, should be applied to impute missing values.
On the other side, we can retrieve relative abundance of these taxonomies from in-situ dataset. To do that we need to do below steps:

1. Preprocess metabarcoding dataset based on the steps that described in the section 3.2.2.2.

2. Find accumulative abundance of taxonomies for each station. It means, we should apply summation over abundance of all ASVs for specific station.

3. Filter ASVs for each of desire taxonomies, Diatom, Dino, Crypto, Hapto, green algae and prokar. Then, calculate accumulate abundance for each group in each station.

4. Divide accumulate abundance for each specific group to accumulative abundance, for each station. In another word, divide result of step 3 to result of step 2 for each station and taxonomy.

Then, we can compare result of satellite with in-situ in each station. In the case that there is high correlation between these two features, we can say that changes in this feature in in-situ dataset is compatible with changes in satellite dataset. Additionally, to find a correlation between two variables, the Pearson correlation has been used, because we want to evaluate linear relationship between two features. The formula for Pearson correlation has provided in below. The range of changes for Pearson correlation is between -1 and +1.

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}} \tag{5.1}$$

In this formula, x and y are referring to the two variables that we want to calculate correlation between them. In the numerator of this equation, covariance between two variables calculated and it shows how these two variables vary together. Then, in the denominator, the covariance will be normalized by dividing it to the multiplication of standard deviations.

In the below figures, correlation between two datasets for different taxonomies have been presented. The x-axis presents the stations and the right hand side y-axis shows the values in the satellite dataset which specified by red colour and the left hand side y-axis shows the in-situ values that coloured with blue colour.
Each figure has 2 plots, the blue plot shows the changes in the taxonomy in different stations for the in-situ dataset and the red one shows that changes in satellite dataset. As we can see, in the 5.1 comparison for Diato and green has been provided. Based on these plots and the Pearson correlation value that specified with pcc in the plot, there is high correlation between in-situ and satellite dataset for these two taxonomies.

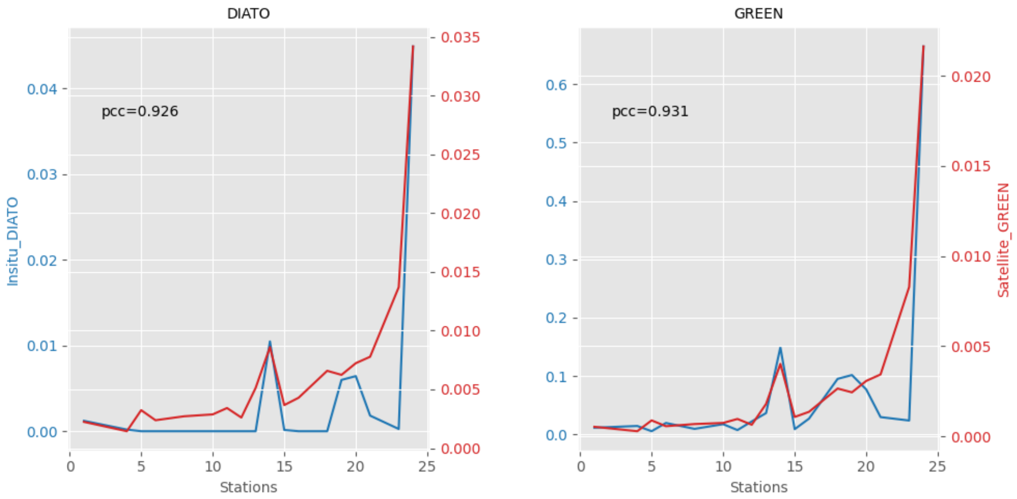

Figure 5.1: Correlation between satellite and in-situ for Diato and green

Based on the figures 5.2 and 5.3, other taxonomies such as crypto, hapto, dino and prokar do not show high correlation between in-situ and satellite dataset. The possible reason for that is related to the filtration that scientist used at laboratory to retrieve different taxonomies. For some taxonomies, this filteration is so big and scientist could not get whole of this taxonomies from the samples.
In the table 5.1, correlation between satellite and in-situ dataset for different taxonomies have been provided.

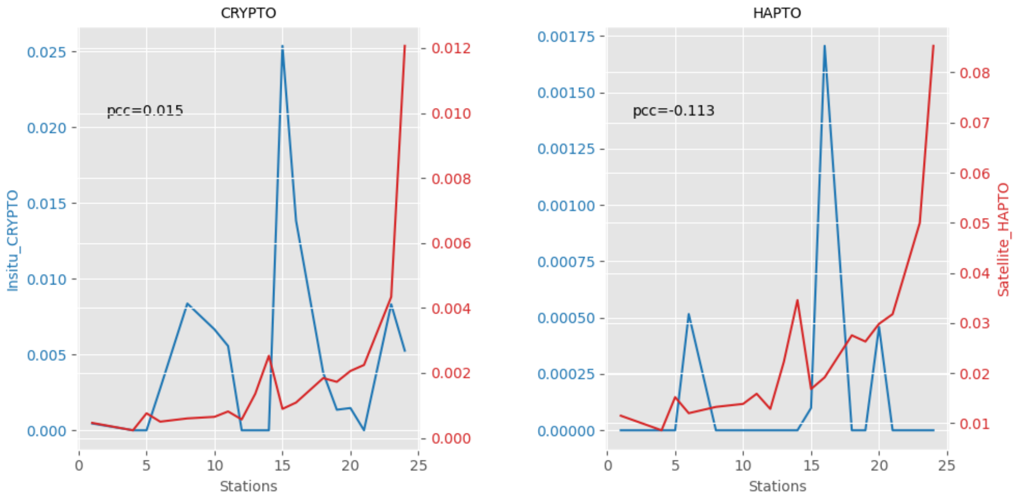

Figure 5.2: Correlation between satellite and in-situ for Crypto and hapto

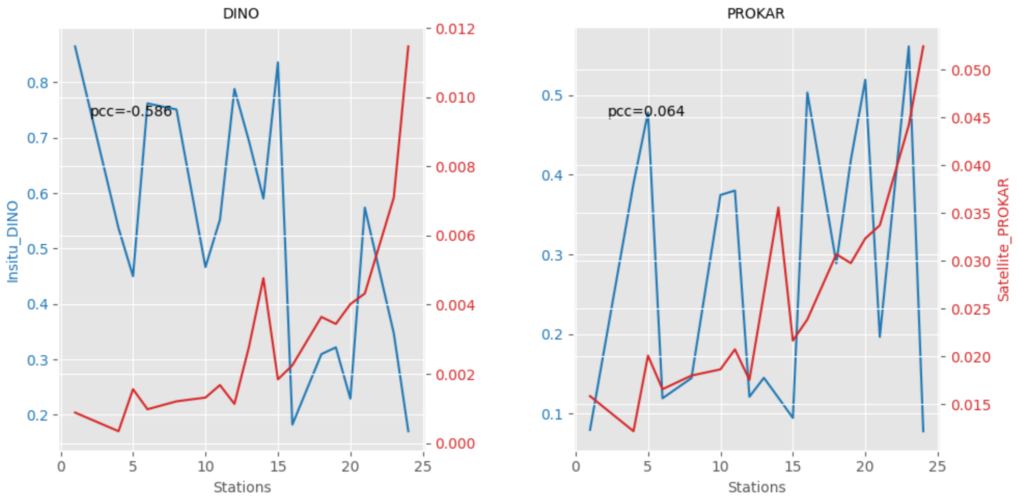

Figure 5.3: Correlation between satellite and in-situ for dino and prokar

	DITO	GREEN	CRYPTO	HAPTO	DINO	PROKAR
pcc	0.926	0.931	0.015	0.113	-0.586	0.064

Table 5.1: Pearson correlation between satellite and in-situ dataset

5.2 Methodology for community detection(surface water)

The number of samples that collected from surface water was 23 samples. To find the communities of surface samples, creating network of ASVs is not an efficient way because the number of samples is small. In this case, the probability of finding same pattern in abundance of ASVs in different samples will be lower than having higher number of samples. Therefore, the output network will not be one connected component and we will have different connected components. Additionally, as mentioned in github page of FlashWeave Git, this package is optimized for large-scale data sets.

The approach that is applied to detect communities in surface water is based on Hierarchical clustering. This technique divide into two types: Agglomerative and Divisive.

Agglomerative clustering is a bottom-up approach. In this approach each node considers as an individual cluster and then in each iteration nodes with higher

similarity merged into one cluster. This procedure continues until we reach to one cluster that has all nodes. In the divisive type, the procedure is opposite to Agglomerative approach. It means that all the nodes consider as one cluster and in each iteration the node that is not similar to other nodes will be separated from cluster.
In this study, agglomerative hierarchical clustering has been used to put similar samples in one cluster. The similarity between samples is assessed by considering the metabarcoding table, which represents the abundance of ASVs in each sample. The approach to calculate similarity between samples is based on Bray-Curtis that described in the section 4.1.4.

Another thing that we should take into account is how to calculate distance between two clusters, in the case that at least one of clusters has more than one member. To calculate this distance there are some linkage methods, such as Single, Complete, Average and ward linkage. In below, formula for each linkage method has been provided. In these formulas, a and b are the members in the cluster A and B respectively. The goal in single linkage is finding minimum distance between members in cluster A and B, while in complete linkage we prefer to find maximum distance between members in cluster A and B.
In the average linkage, the pairwise distance between members in cluster A and B calculated and then considered the average of these distances. Since, average linkage considers all the points in the cluster, it is less sensitive to outliers and noise compared to complete and single linkage methods Clu. On the other side, ward method will minimize the total within cluster variance. In each iteration, two clusters will merge if we could reach to the minimum total within variance after merging these two clusters Raieli (2021).

$$Single\ linkage(A,B) = min(d(a,b) : a \in A, b \in B) \tag{5.2}$$

$$Complete\ linkage(A,B) = max(d(a,b) : a \in A, b \in B) \tag{5.3}$$

$$Average\ linkage(A,B) = \frac{1}{N_A.N_B}.\sum_{i=1}^{N_A}\sum_{j=1}^{N_B} D(x_{Aa}.x_{Bb}) \tag{5.4}$$

$$Ward\ linkage(A,B) = \frac{N_A.N_B}{N_A+N_B}.||m_A - m_B||^2 \tag{5.5}$$

N_A, N_B: number of points in cluster A and B respectively.
m_A, m_B: center of cluster A and B respectively.

To decide which method has better performance on the dataset, agglomerative coefficient (AC) value for each method has been calculated. When AC value gets closer to one, it suggests a stronger clustering. The AC is calculated based on formula 5.6. In this formula, m(i) is the dissimilarity of observation i to the first cluster it merges with, divided by the dissimilarity of the last merger in the algorithm. The agglomerative coefficient is obtained by averaging the values of 1 - m(i) for all observations Documentation (2022).

$$AC = \frac{1}{n}(1 - m_i) \tag{5.6}$$

After finding the best linkage method, we will run the agglomerative hierarchical clustering algorithm. It will give us dendrogram[1] that leafs refer to the samples. Then, to measure the goodness of clustering technique, we can use silhouette score.

This metric measures the cohesion of a cluster, which refers to the within-cluster similarity, as well as the separation of a cluster from other clusters. It ranges between -1 and +1, when it is near to 1 it shows that observations are very well assigned to clusters. When this index is decreasing, it shows the points are poorly assigned to the cluster. In below formula for Silhouette width has been provided.

$$S_i = \frac{b_i - a_i}{max(a_i, b_i)} \tag{5.7}$$

a_i: Average dissimilarity between i and other points that are in the same cluster that i exists.
$d(i, c)$: Average dissimilarity of point i to the points in the cluster c.
$b_i = min_c d(i, c)$: Minimum distance between observation i and points in the cluster c.

5.3 Implementation of community detection(surface water)

Based on the steps that mentioned in methodology, first we need to find similarity between points or stations. The vegan package in R has been used to find

[1] The main role of dendrogram is showing the hierarchical relationship between objects or samples.

Bray-curtis dissimilarity between samples. Then, hclust in stats package or agnes in cluster package in R can put similar stations in one cluster. To find the best approach for assigning nodes to clusters, different linkage methods have been tried and the result provided in the table 5.2. In accordance with the result, Ward method has highest AC and it shows stronger clustering structure compare to other methods.

	Single	**Average**	**Complete**	**Ward**
AC	0.440	0.443	0.486	0.686

Table 5.2: AC for different linkage methods

After specifying the distance between stations and best linkage method we can get the graphical representation of hierarchical clustering algorithm as a tree, known as a dendrogram. It is presented in the figure 5.4. The height in this dendrogram is a measure of similarity between leafs or stations[2]. When the height is increasing, the stations are less similar. For example, station 28 and 26 have low height that shows these stations are very similar. while, station 15 is not so similar to the station 16 and 18, because their height is high.

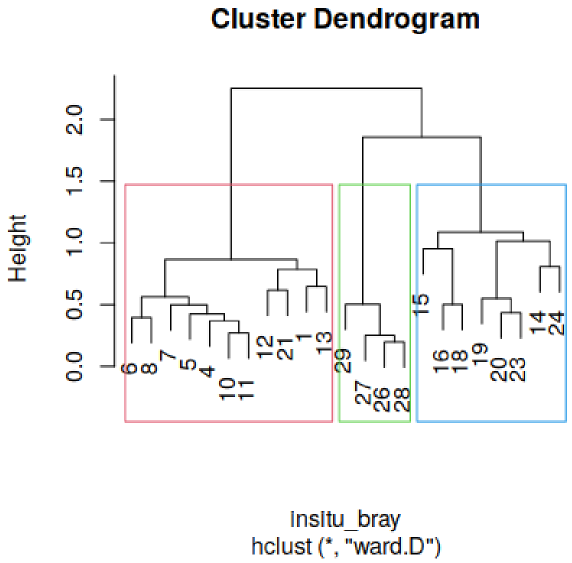

Figure 5.4: Dendrogram for surface samples(hclust)

[2] Each station has only one surface sample, therefore instead of using the word sample we are using the station.

To evaluate generated clusters, silhouette width has been used and the result has been provided in the figure 5.5. Moreover, to dig deeper about relation between stations and cluster boundaries, dimension reduction technique on the abundance of ASVs for different stations has been implemented by fviz_cluster function from the factoextra package in R. The result of dimension reduction presented in the figure 5.6.

Based on the figure 5.5, cluster 3 has highest average silhouette width, 0.59, compare to other clusters. It shows that samples in this cluster are very similar. Additionally, in the figure 5.6, the cluster 3 that has been specified with blue colour is very compact and its members are very close to the center. The members of cluster 3 are pointing to the last stations that are located in the Atlantic ocean, based on table 3.2 that has information of stations.

The second position belongs to the cluster 1 that its silhouette width equals to 0.29. The dominant stations in this cluster are located in the eastern part of the Mediterranean Sea, with only one station located in the western region. Cluster 1 is not as compact as cluster 3 and there are some stations such as 14 and 21 that their distance to the center of the cluster is higher than other stations in this cluster.

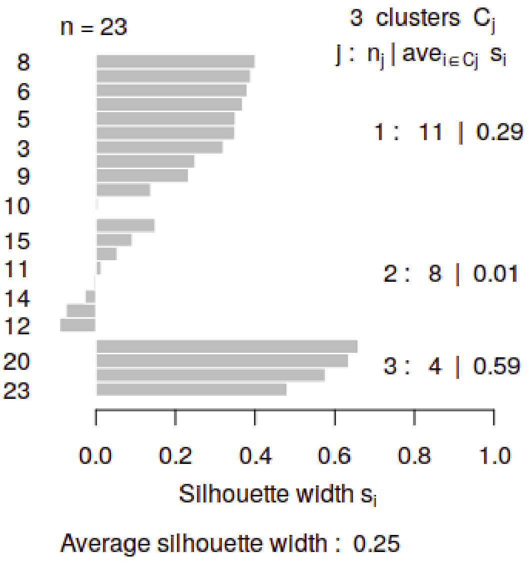

Figure 5.5: silhouette Width for hclust

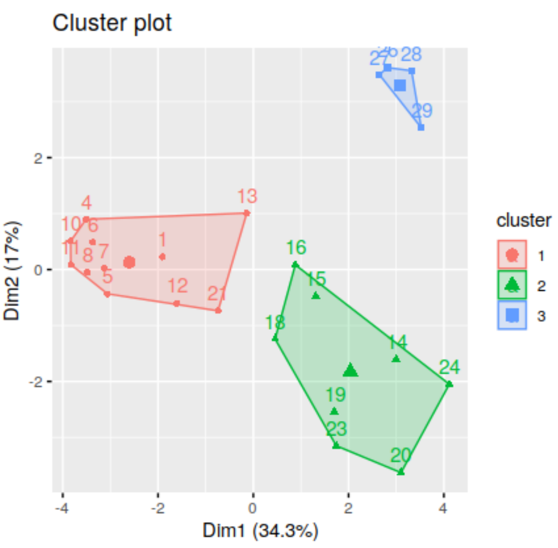

Figure 5.6: DR for hclust

The last position belongs to the cluster 2 that has lowest silhouette width. This cluster is dominant with samples from west of Med sea. There are some samples in this cluster that their sillouhette score are negative. It shows that behaviour of these samples are different with other members in this cluster. Additionally, based on the result of 5.6, the area of polygon related to the this cluster is bigger than other clusters and it shows the diversity of samples in western part of Med sea.

Another approach to build hierarchical clustering is based on reculter.cons from recluster package in R. In this approach a series of trees have been created based on resampling the order of stations in the similarity matrix. Then, to find robust clusters, consensus among trees has been computed Dapporto et al. (2013). The result of this approach provided in the figure 5.7.

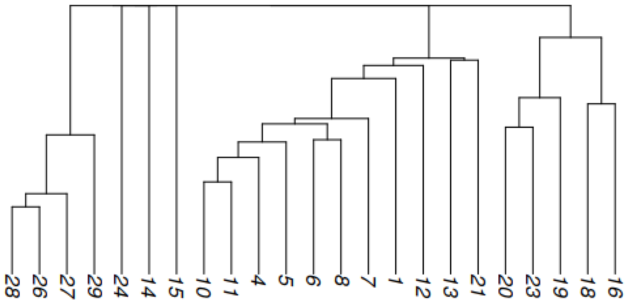

Figure 5.7: Dendrogram for surface samples(recluster)

As we can see in the figure 5.7, we have some stations that do not belong to any cluster and they have their own cluster. Similarly, to evaluate the effectiveness of the recluster approach, silhouette score and dimension reduction were employed. Based on the result in the figure 5.8 the average silhouette width increased to 0.26. The good point of the new clustering is that we do not have negative silhouette width. The three stations that had negative score in previous case, station 14, 15 and 24, now have zero silhouette score [3].

On the other side in the figure 5.9 the distribution of samples in clusters presented. Clusters that were related to the east and Atlantic did not change. While the cluster that was related to the west divided to the four clusters.

[3] The silhouette score for a cluster with one member is zero

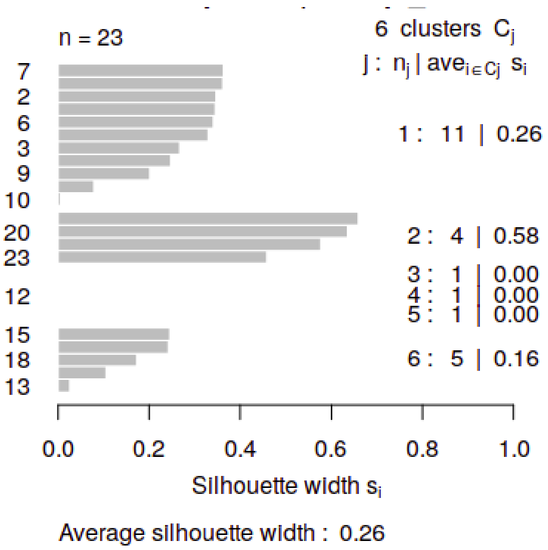

Figure 5.8: silhouette Width for recluster

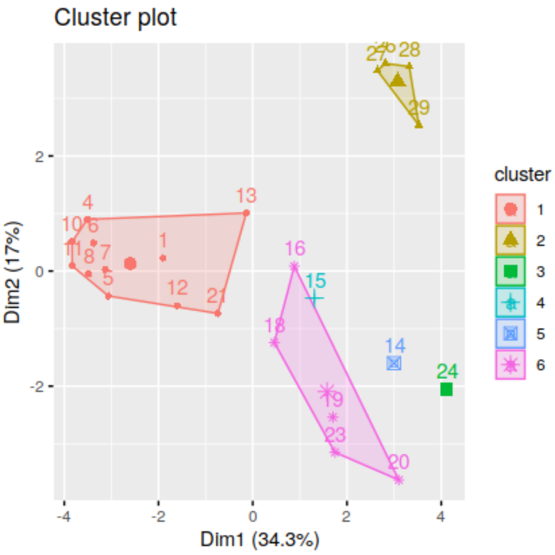

Figure 5.9: DR for recluster

In the figure 5.10, clusters and their location on the map have been presented. This plot has been drawn by ggmap in the R Kahle and Wickham (2013) .As we can see, the samples that collected from east of Mediterranean sea are grouped in one cluster. While in the west of med we have 4 different clusters. Moreover, samples in the Atlantic ocean grouped in one cluster.

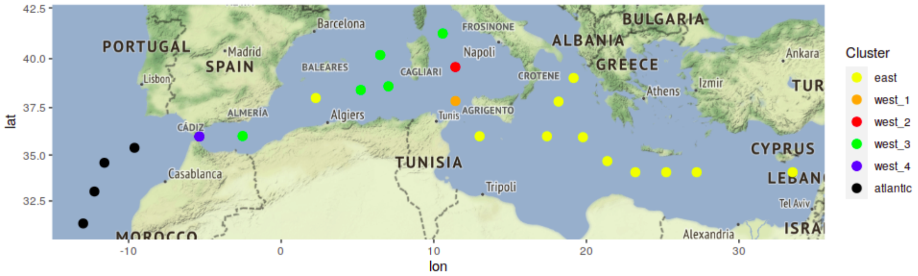

Figure 5.10: Stations with their labels in the map

5.4 Community composition

Community composition refers to finding relative abundance of different taxonomies in each community. Even though, we expect to see specific pattern in composition of taxonomies in each community, but there might be some similarity between composition in some communities. In the figure 5.11, relative abundance of different taxonomies in the east cluster presented. Based on this plot, Dinoflag has higher abundance in this cluster and next position belong to the prokar that its mean relative abundance is getting higher than 0.2. Relative abundance of other taxonomies in this cluster are near zero.

When we are moving from east to west of Med sea, the abundance of phytoplanctons starts to change. As mentioned before for west of Med sea, we have 4 clusters and for each of these clusters composition of taxonomies shows different behaviour. For example, in cluster west_1 in figure 5.12 the relative abundance of Dinoflag and Prokar starts decreasing compare to the east cluster and the amount of Green algae starts increasing.

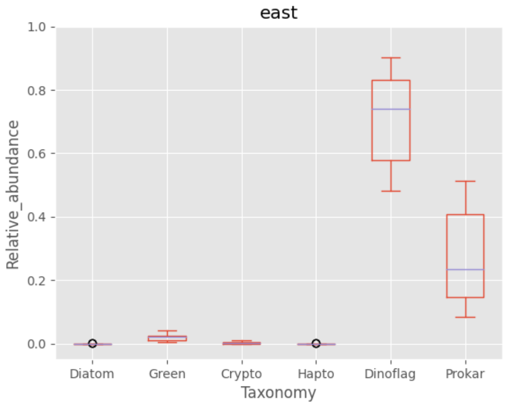

Figure 5.11: Composition in east

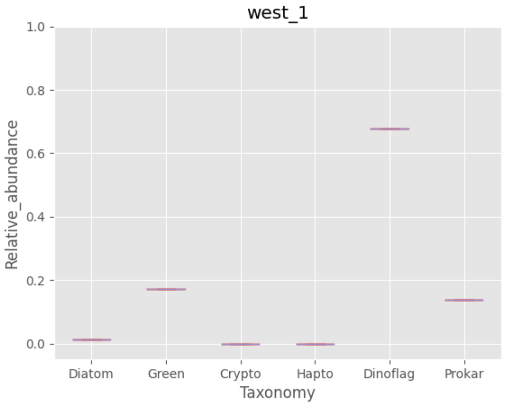

Figure 5.12: Composition in west_1

For cluster west_2 that presented in the figure 5.13, the relative abundance of Dinoflag reaches to 0.9, that is highest among other clusters. Additionally, abundance of other taxonomies in this cluster is very low and we can say dinoflag is a dominant taxonomy for cluster_2.

On the other side, in the west_3 cluster the abundance of dinoflag considerably decreased compared to west_2 cluster and abundance of prokar and green increased.

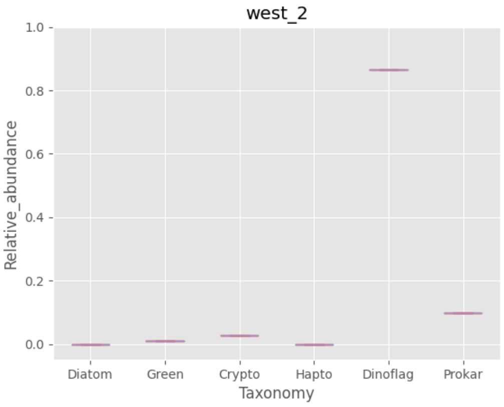

Figure 5.13: Composition in west_2

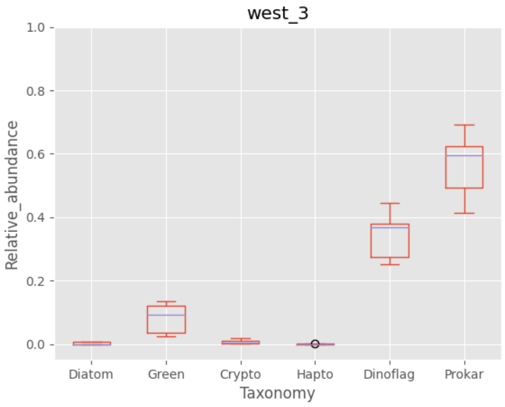

Figure 5.14: Composition in west_3

For the last cluster in the west of Med sea, cluster west_4, we can observe a different behaviour in abundance of taxonomies. With respect to figure 5.15 the amount of Dinoflag and prokar decreased compared to the west_3 cluster and the abundance of green algae increased and reached to 0.7 and became a dominant taxonomy in this cluster.

Finally, behavior of taxonomies in atlantic cluster has been provided in the figure 5.16. It shows that relative abundance of prokar reaches to highest value compare to other clusters. The second and third highest relative abundances belong to Dinaflag and Green respectively. It does worth to mention that, the behaviour of phytoplancton composition in the atlantic cluster and west cluster follow the same pattern, prokar is dominant taxonomy and dinoflag and green algae are in the second and third position.

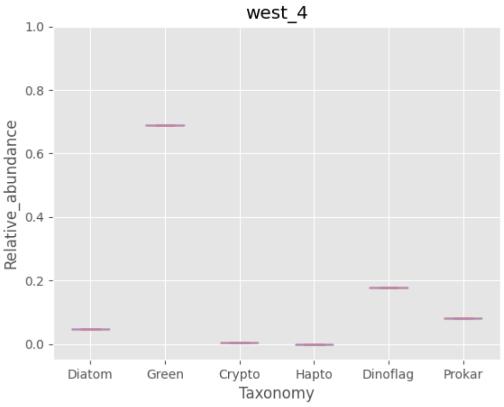

Figure 5.15: Composition in west_4

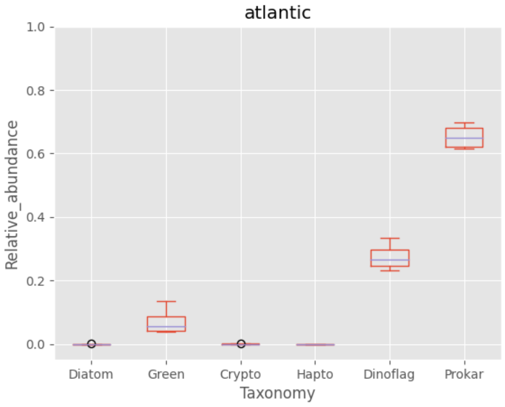

Figure 5.16: Composition in atlantic

5.5 Classify communities based on satellite features

After finding communities and composition of each community for the surface water, our next objective is to explore the feasibility of satellite-based features in predicting these communities. In order to do that, we will assign labels that generated from hierarchical agglomerative clustering to the corresponding satellite features obtained from that station.

Then, we can train different classification algorithm on the satellite features and find which one of algorithms can achieve higher accuracy. We should consider that there are 23 stations with surface samples which is not sufficient to train classification algorithm. As mentioned in the preprocessing part, a box of 5*5 km considered around the station to impute missing values. Therefore, we expect to not have considerable changes in distribution of phytoplanktons inside this box. As a result, we can increase number of samples by considering imaginary samples inside the box of each station. Then, the label that is defined for that station will be assigned to the imaginary samples as well. At the end, number of samples increased to 100.

The approach that is used to train and test algorithms is based on Leave one out cross validation (Loocv), because our dataset is small and we want to utilize all available data points for training and test evaluation. As a result, LOOCV leading the model to have more reliable performance.

While, train-test split may not perform well in small dataset because the limited amount of data allocated for training can be insufficient and potentially biased towards specific groups. Therefore, the model may not learn patterns from other groups adequately.

In the next step, we should test performance of different classification algorithms on the dataset. These algorithms have been presented in the table 5.3 that scikit-learn package has been used to train each of them.

The criteria that are used to select the best model are accuracy score and F1 score. In the accuracy score that its formula provided in the 5.8, we want to know the proportion of the data that correctly predicted. Therefore, it focuses on the true positive and true negative classes. In another word, all the classes are equally important in accuracy score. While, in F1 score that its formula provided in the 5.9, a combination of precision and recall has been considered and it pays more attention to the cases that incorrectly classifies, false negative and false positive cases.

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{True Positives} + \text{True Negatives} + \text{False Positives} + \text{False Negatives}} \tag{5.8}$$

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5.9}$$

Algorithm	**Accuracy score**	**F1 score**
Random Forest (n trees = 100)	0.945	0.928
SVM(linear kernel)	0.890	0.873
K-nearest Neighbors (k =3)	0.824	0.817
K-nearest Neighbors (k =2)	0.818	0.817
K-nearest Neighbors (k =1)	0.818	0.821
K-nearest Neighbors (k =4)	0.809	0.797
K-nearest Neighbors (k =5)	0.796	0.780
Naive Bayes(Gaussian)	0.781	0.773
SVM(RBF kernel)	0.745	0.706
SVM(sigmoid kernel)	0.690	0.629

Table 5.3: Result of classification algorithms on test set

Based on the table 5.3, the accuracy and f1 score in test set for random forest are higher than classifiers. On the other side, between different variant of SVM algorithm such as rbf, sigmoid and linear kernel, svm with linear kernel has better performance. It means that the dataset can be linearly separated and we do not need to map data to higher dimension and find a hyperplane to separate dataset. For the k nearest neighbor algorithm , when we choose k = 3 we could reach to higher accuracy compare to other k values. Additionally, the visualization format of this plot has been presented in the figure 5.17.

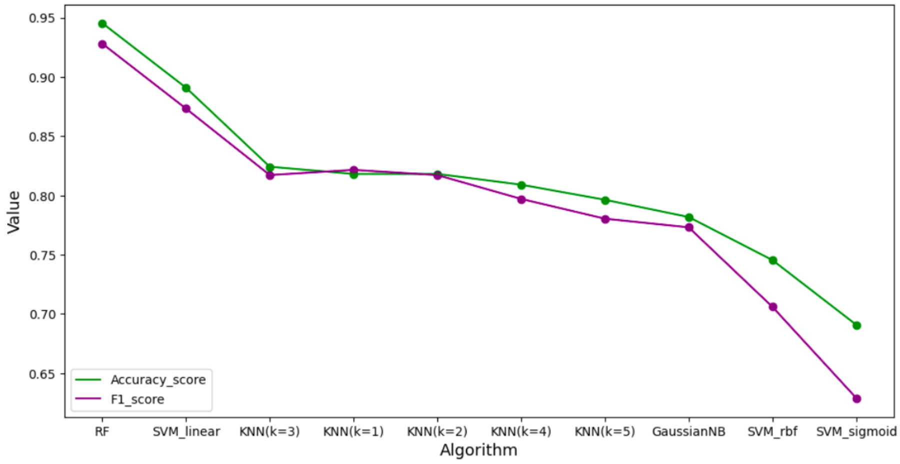

Figure 5.17: Comparison between accuracy and f1 score of classification algorithms

Afterward, we want to know which one of Satellite features, Reflectance(in 5 wave length), KD490, CHL, SST and POC, has considerable effect on classification of communities. In the figure 5.18 importance of features based on the Random forest algorithm has been provided. The first three important features are SST, poc and chl and if we train the random forest algorithm with just these three features, accuracy score reaches to 0.927 and f1 score reaches to 0.91.

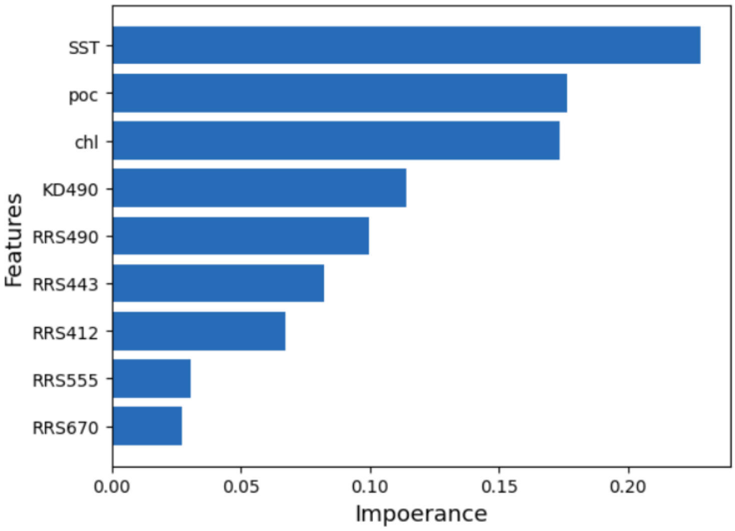

Figure 5.18: Importance of features based on RF algorithm

Chapter 6

Conclusion and Future work

6.1 Conclusion

In this project, we detected phytoplankton communities in Mediterranean sea and adjacent eastern Atlantic ocean. The community detection divided into two parts, finding communities for all-depth water and surface water. Conclusion for each part provided separately in below:

- All-depth water
  In below we have three main conclusions for all-depth water:
  First: For each of phytoplankton types, prokaryote and eukaryote, we could detect 7 communities by applying Louvain algorithm. The most important features that affect on creating communities were the depth and basin that sample collected.
  Second: There is a difference between eukaroyte and prokaryte communities, in prokaryote we have 3 clusters in the $depth \leq 200m$, while in the eukaryote communities, we have 4 clusters in this depth. On the other side, in the $depth \geq 200m$, 4 clusters identified for prokaryote, whereas eukaryotes exhibit 3 clusters in this depth. Therefore, it seems that eukaroyte has higher distribution in shallower part of the sea compare to prokaroyte.
  One possible explanation for this behaviour is that eukaryotes consist a more diverse range of taxonomies, including dino, diato, hapto, crypto and green algae, which rely on sunlight for photosynthesis. In contrast, prokaryote primarily consist of cyanobacteria as the main photosynthetic organisms.
  Third: One of the environmental features that affect on creating new community is the amount of nutrient levels such as NO3, PO4 and Sio3. For

example, in the Atlantic ocean the nutrient levels are higher than Med sea and it causes to have separate community in the Atlantic ocean.

- Surface water
  The main conclusion related to the studying surface water divided to three parts:
  First: In the surface samples, both in-situ data and satellite data are utilized to explore the correlation of common features between these two datasets. Notably, a significant correlation observed in diatoms and green algae. Therefore, in the long term, satellite features can be employed to track the fluctuations in these two taxonomies.
  Second: We could detect 6 communities in the surface water, which are differentiated based on basin characteristics and taxonomic composition.
  Third: We proved that with satellite features, we can predict representative communities with high accuracy. The most important features that have highest affect on classification are SST, chl and poc.

6.2 Future work

The possible future work also can be divide into two parts, all-depth analysis and surface analysis.

- all-depth water
  We could find number of communities and important environmental or physical features that affect on creating communities. Another possible work is finding community composition and analyse the amount of taxonomies that exist in each of communities. To do that we need to have domain knowledge of biology.

- Surface water
  Future works related to surface water divided to two parts:
  First: Correlation between satellite and in-situ data has been calculate based on the metabarcoding table and as we saw there were not correlation for some taxonomies. Therefore, one possible work can be considering another dataset, like metagenomic dataset, to explore the potential correlation for desire taxonomies between satellite and in-situ dataset.
  Second: Based on trained algorithm with satellite features, we can study

changes in distribution of communities for different years. Therefore, this study can be step toward further analysis of long term changes in communities.

Bibliography

Important Notes on Cluster Analysis. `https://medium.com/@kirtitambe17/important-notes-on-cluster-analysis-ac0abc65a337`.

Github FLashWeave. `https://github.com/meringlab/FlashWeave.jl`.

J. Aitchison. The analysis of compositional data (london. *Chapman and*, 1986.

V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

J. L. Bronstein. *Mutualism*. Oxford University Press, USA, 2015.

K. R. Clarke. Non-parametric multivariate analyses of changes in community structure. *Australian journal of ecology*, 18(1):117–143, 1993.

A. Clauset, M. E. Newman, and C. Moore. Finding community structure in very large networks. *Physical review E*, 70(6):066111, 2004.

M. E. Cristescu. From barcoding single individuals to metabarcoding biological communities: towards an integrative approach to the study of global biodiversity. *Trends in ecology & evolution*, 29(10):566–571, 2014.

L. Dapporto, M. Ramazzotti, S. Fattorini, G. Talavera, R. Vila, and R. L. Dennis. recluster: an unbiased clustering procedure for beta-diversity turnover. *Ecography*, 36(10):1070–1075, 2013.

R. Documentation. Agglomerative / Divisive Coefficient for 'hclust' Objects, 2022. URL `https://stat.ethz.ch/R-manual/R-devel/library/cluster/html/coef.hclust.html`.

R. El Hourany, J. P. Karlusich, L. Zinger, H. Loisel, M. Levy, and C. Bowler. Linking satellites to genes with machine learning to estimate major phytoplankton groups from space. 2022.

J. R. Haslett. Insect communities and the spatial complexity of mountain habitats. *Global Ecology and Biogeography Letters*, pages 49–56, 1997.

P. C. Junger, H. Sarmento, C. R. Giner, M. Mestre, M. Sebastian, X. A. G. Moran, J. Aristegui, S. Agusti, C. M. Duarte, S. G. Acinas, et al. Global biogeography of the smallest plankton across ocean depths. *bioRxiv*, pages 2023–01, 2023.

D. Kahle and H. Wickham. ggmap: Spatial visualization with ggplot2. *The R Journal*, 5(1):144–161, 2013. URL `https://journal.r-project.org/archive/2013-1/kahle-wickham.pdf`.

H. Kaneko, H. Endo, N. Henry, C. Berney, F. Mahé, J. Poulain, K. Labadie, O. Beluche, R. El Hourany, T. O. Coordinators, et al. Global observation of plankton communities from space. *bioRxiv*, pages 2022–09, 2022.

S. J. Kramer, D. A. Siegel, S. Maritorena, and D. Catlett. Modeling surface ocean phytoplankton pigments from hyperspectral remote sensing reflectance on global scales. *Remote Sensing of Environment*, 270:112879, 2022.

R. Lindsey. What are Phytoplankton?, 2010. URL `https://earthobservatory.nasa.gov/features/Phytoplankton#:~:text=Like%20land%20plants%2C%20phytoplankton%20have,energy%20by%20consuming%20other%20organisms`.

A. Magnuson, L. W. Harding Jr, M. E. Mallonee, and J. E. Adolf. Bio-optical model for chesapeake bay and the middle atlantic bight. *Estuarine, Coastal and Shelf Science*, 61(3):403–424, 2004.

M. E. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.

Pedro. INTRO TO DATA CLUSTERING, 2021. URL `https://ourcodingclub.github.io/tutorials/data-clustering`.

S. Raieli. Clustering techniques with Gene Expression Data for Acute Myeloid Leukemia, 2021. URL `https://medium.com/leukemiaairesearch/clustering-techniques-with-gene-expression-data-4b35a04f87d5`.

D. Righetti, M. Vogt, N. Gruber, A. Psomas, and N. E. Zimmermann. Global pattern of phytoplankton diversity driven by temperature and environmental variability. *Science advances*, 5(5):eaau6253, 2019.

M. Sebastián, E. Ortega-Retuerta, L. Gómez-Consarnau, M. Zamanillo, M. Álvarez, J. Arístegui, and J. M. Gasol. Environmental gradients and physical barriers drive the basin-wide spatial structuring of mediterranean sea and adjacent eastern atlantic ocean prokaryotic communities. *Limnology and Oceanography*, 66(12): 4077–4095, 2021.

J. Tackmann, J. F. M. Rodrigues, and C. von Mering. Rapid inference of direct interactions in large-scale ecological networks from heterogeneous microbial sequencing data. *Cell systems*, 9(3):286–296, 2019.

C. Watson. Girvan-Newman and Louvain Algorithms for Community Detection, 2022. URL `https://medium.com/smucs/girvan-newman-and-louvain-algorithms-for-community-detection-f3feb7c`
`~:text=The%20Louvain%20algorithm%20is%20an,fastest%20complexity%20in%20community%20detection`.

C. website(KD490 Atlantic ocean). KD490 for Atlantic ocean, 1999. URL `https://data.marine.copernicus.eu/product/OCEANCOLOUR_ATL_BGC_L3_MY_009_113/download?dataset=cmems_obs-oc_atl_bgc-transp_my_l3-multi-1km_P1D`.

C. website(KD490 Med sea). KD490 for Med sea, 1999. URL `https://data.marine.copernicus.eu/product/OCEANCOLOUR_MED_BGC_L3_MY_009_143/download?dataset=cmems_obs-oc_med_bgc-transp_my_l3-multi-1km_P1D`.

C. website(Plankton Med sea). Plankton for Med sea, 1999. URL `https://data.marine.copernicus.eu/product/OCEANCOLOUR_MED_BGC_L3_MY_009_143/download?dataset=cmems_obs-oc_med_bgc-plankton_my_l3-multi-1km_P1D`.

C. website(poc chl product). poc and chl in ocean, 1999. URL `https://data.marine.copernicus.eu/product/MULTIOBS_GLO_BIO_BGC_3D_REP_015_010/download?dataset=cmems_obs_glo_bgc3d_rep_weekly`.

C. website(reflectance Atlantic). reflectance for Atlantic in ocean, 1999. URL `https://data.marine.copernicus.eu/product/OCEANCOLOUR_ATL_BGC_L3_MY_009_113/download?dataset=cmems_obs-oc_atl_bgc-reflectance_my_l3-multi-1km_P1D`.

C. website(reflectance Med sea). reflectance for Med sea, 1999. URL https://data.marine.copernicus.eu/product/OCEANCOLOUR_MED_BGC_L3_MY_009_143/download?dataset=cmems_obs-oc_med_bgc-reflectance_my_l3-multi-1km_P1D.

C. website(SST product). SST in ocean, 1999. URL https://data.marine.copernicus.eu/product/SST_GLO_SST_L4_NRT_OBSERVATIONS_010_001/download?dataset=METOFFICE-GLO-SST-L4-NRT-OBS-SST-V2.

E. Wilson. Sociobiology: The new synthesis. harvard universitypress.[rhg, tg, gew](1998) consilience: The unity of knowledge, 1975.

ZACH. How to Calculate Bray-Curtis Dissimilarity in R, 2022. URL https://www.statology.org/bray-curtis-dissimilarity-in-r/.

G. Zheng and P. M. DiGiacomo. Detecting phytoplankton diatom fraction based on the spectral shape of satellite-derived algal light absorption coefficient. *Limnology and Oceanography*, 63(S1):S85–S98, 2018.