Application of Machine Learning Methods to Predict Phytoplankton Blooms and Determine Microbial Biomarkers using Marine Microbiomes

Nuria Fernández González

MU Bioinformática y Bioestadística

Bioinformática Estadística y Aprendizaje Automático

Nombre de la Tutora de TF

Romina Astrid Rebrij

Profesor responsable de la asignatura

Carles Ventura Royo

20/06/2023

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Application of Machine Learning methods to Predict phytoplankton blooms and Determine microbial Biomarkers using Marine Microbiomes*
Nombre del autor:	*Nuria Fernández González*
Nombre del consultor/a:	*Romina Astrid Rebrij*
Nombre del PRA:	*Carles Ventura Royo*
Fecha de entrega (mm/aaaa):	*20706/2023*
Titulación o programa:	*MU en Bioinformática y Bioestadística*
Área del Trabajo Final:	*Statistical Bioinformatics and Machine Learning*
Idioma del trabajo:	*Inglés*
Palabras clave	*Coastal blooms, biomarkers, random forest*

Resumen del Trabajo

El conocimiento de las relaciones entre el bacterioplancton y las proliferaciones de fitoplancton es clave para entender el funcionamiento de los ecosistemas, como también predecir y mitigar los efectos del cambio global sobre estos ecosistemas. Estas comunidades microbianas son gobernadas por relaciones complejas. Además, los datos para estudiar la diversidad del bacterioplancton (Variantes de secuencias de amplicones del gen del ARNr 16S) son altamente dimensionales, dispersos y ruidosos. En este proyecto, los clasificadores Random Forest basados en datos de diversidad se utilizaron para predecir proliferaciones costeras de fitoplancton y buscar biomarcadores de estos. Tras unir los datos de dos campañas oceanográficas, las muestras se clasificaron entre las categorías Bloom y normal según la concentración de clorofila. Los datos resultantes eran altamente dimensionales (166 muestras, 7593 variables) y desbalanceados (31 muestras bloom, 135 normales). Para reducir la dimensionalidad, las variables biológicas con abundancias relativas menores al 0,01% se eliminaron. Alternativamente, se agruparon a nivel de género. Los modelos Random Forest se entrenaron valorando diferente número de variables en los árboles individuales. El proceso se repitió con cien divisiones diferentes de los datos en los grupos de entrenamiento y test para asegurar la representatividad de los resultados. Los modelos sólo alcanzaron buenos niveles de desempeño (kappa, sensibilidad y especificidad medias > 0.8) tras utilizar la técnica de sobre muestreo sintético de la clase minoritaria, bloom, para balancear los datos. Finalmente, se determinaron los biomarcadores como las variables más importantes según su error predictivo.

Abstract

Understanding the relationship between bacterioplankton and coastal phytoplankton blooms is key to understand coastal ecosystems functioning, which are the most productive areas for fisheries. With that knowledge, we could predict and may be mitigate, the effects of global change or contamination events in these productive ecosystems. However, these microbial communities are governed by very complex relationships. In addition, the data used to study bacterioplankton diversity (Amplicon Sequence Variants of 16S rRNA gene) is highly dimensional, sparse, and noisy. In this project, Random Forest classifiers based on diversity data were used to predict coastal phytoplankton blooms and search for their biomarkers. After joining two oceanographic campaigns data, samples were classified as bloom or normal depending on the total chlorophyll concentrations. The resulting dataset was highly dimensional (166 instances, 7593 features) and imbalanced (31 instances bloom, 135 – normal). To reduce dimensionality, biological features with relative abundances below 0.01 were removed, or they were grouped into clusters at genus level. Random forest models were trained and tuned with a grid-search of the number of features included in the individual trees. The process was repeated using one hundred different data splits into train and test groups to ensure results' representativity. Good performance values (kappa, sensitivity, and specificity > 0.8) were achieved only after using the synthetic minority oversampling technique to level the number of instances between the two categories. Using those models, the topmost important features, according to the predictive error rate of features, were selected as biomarkers.

CONTENTS

List of figures vi

List of tables...vii

1. Introduction 1

1.1. Background and Project Rationale 1

1.2. Goals 3

1.3. Impacts on Sustainability, Ethical Behaviour, Social Responsibility and Diversity 4

1.4 Approach and Methodology 5

1.4. Working Plan 8

1.5. Summary of Obtained Products 11

1.6. Short description of Other Chapters 11

2 Materials and methods. 12

2.1 Data search and selection. 12

2.1.1 Search of public databases for additional data. 12

2.1.2 Envision and Dimension campaigns. 12

2.2 Environmental metadata and sequence files preparation and preprocess. ... 14

2.3 Classification of samples based on chlorophyll concentrations. 15

2.4 Amplicon Sequence Variants generation. 16

2.4.1 Sequences pre-processing 16

2.4.2 Generation and taxonomic assignment of ASVs. 16

2.5 Data preparation 16

2.6 ASVs filtering, clusters generation and exploratory analysis. 17

2.7 Random Forest Models. 17

2.7.1 Feature selection and preprocess 17

2.7.2 Model training, validation, and performance. 18

2.8 Improving model performance. 19

2.8.1 A Support Vector Machine model. 19

2.8.2 Synthetic data generation. 20

2.9 Determination of feature importance. 20

2.10 Hardware, Software and Code Availability 21

3 Results and discussion 22

3.1 Data included in the Project. 22

3.2 Classification of samples according to chlorophyll concentrations 23

3.2.1 PCAs of selected environmental variables for both campaigns. 23

3.2.2 Distribution of samples between bloom and normal events. 26

3.3 Sequences and ASVs preparation 27

3.4 ASVs exploratory analysis 27

3.5 Clusters exploratory analysis 30

3.6 Random Forest models training and validation 33

3.6.1 Feature preprocess and selection 33

3.6.2 Random forest models training 35

3.7 Attempts to improve model performance 39

3.7.1 SVM model 39

3.7.2 Synthetic data 40

3.8 Determining biomarkers 41

4 Future Work and conclusions 44

4.1 Conclusions 44

4.2 Future Work 45

4.3 Impacts on Sustainability, Ethical Behaviour, Social Responsibility and Diversity. 46

5 Glossary 47

6 Bibliography 48

7 Appendices 53

7.1 Other PCAs of Dimension campaign. 53

7.2 Total chlorophyl data exploration 54

7.3 Sequence counts per sample 58

7.4 Exploratory analysis of unfiltered ASVs 62

7.5 Features with mean importance values over 10 65

7.6 Repository contents 66

LIST OF FIGURES

Figure 1. a) Diagram of a coastal upwelling event, when deep cold water reaches the surface near a coast driven by wind changes (By Lichtspiel). b) Effects of upwelling on surface chlorophyll concentrations in the Western Pacific Ocean (NASA) (Source: https://en.wikipedia.org/wiki/Upwelling, accessed on Jun 6th 2023) 1

Figure 2. Phytoplankton bloom in the Barents Sea, notice the size of the blooming area compared with the northern Norwegian coast (Envisat satellite, ESA, CC BY-SA 3.0 IGO, https://creativecommons.org/licenses/by-sa/3.0/igo/).. 2

Figure 3. Gantt diagram. 10

Figure 4. Location of coastal and oceanic sampling points in the North-West coast of the Iberian Peninsula. (Courtesy of Prof. Sandra Martínez-García). 13

Figure 5. Initial pre-process outline for ENVISION and DIMENSION data and files. 15

Figure 6. Diagram of the protocol used to train and tune the models. 19

Figure 7. PCAs for Envision (left column) and Dimension (right column) calculated with environmental variables. The samples have been coloured according to different qualitative variables indicated on each panel. The percentage of variance explained by each comp component is indicated between parentheses... 24

Figure 8. Primary production for the different size-fraction Eukaryotes in Dimension campaign. 26

Figure 9: Boxplots of ASVs counts per sample before (a) and after (b) removing low abundant ASVs.. 28

Figure 10: NMDS calculated with filtered ASVs with data transformed by the square-root of proportions. First panel contains the stress plot. The other panels contain the NMDS plot in which samples are labelled with their names or coloured by the categorical variables. 29

Figure 11. NMDS using filtered ASVs transformed with rclr. First panel contains the stress plot. The other panels contain the NMDS plot in which samples are labelled with their names or coloured by the categorical variables. 30

Figure 12. Boxplots of clusters counts per sample. 31

Figure 13. NMDS calculated with clusters count data transformed by the square-root of proportions. First panel contains the stress plot. The other panels contain the NMDS plot in which samples are labelled with their names or coloured by the categorical variables. 32

Figure 14. NMDS using clusters count data transformed with rclr. First panel contains the stress plot. The other panels contain the NMDS plot in which samples are labelled with their names or coloured by the categorical variables. 33

Figure 15. Boxplots of feature values from the clusters (a) or filtered ASVs (b) datasets after preprocessing and selection... 35

Figure 16. Mean AUC for the mtry values in the initial RF initial train using the default (a, c) and the extended (b, d) grid searches for the clusters (a,b) and filtered ASVs (c, d) datasets. Values from the 5-fold RCV of the initial single run. 36

Figure 17. Mean AUC for the different mtry values explored. a) clusters with 5 k-fold RCV, b) clusters with 10 k-fold RCV, c) filtered ASVs with 5 k-fold RCV, d) filterd ASVs with 10 k-fold RCV,

e) unfiltered ASVs with 5 k-fold RCV, f) unfiltered ASVs with 10 k-fold RCV. Error bars indicate one standard deviation. 37

Figure 18. Performance for RF models. All metrics were calculated with the test group except "cv_metric_AUC", that indicates the results of the 5 or 10 k-fold repeated cross-validation. Datasets: asvF - filtered ASVs, asvUnF – unfiltered ASVs, cl: Clusters. 38

Figure 19. Mean AUC for the different values for hyperparameters C (a) and sigma (b) on the SVM models. 39

Figure 20. Performance of SVM and RF models. All metrics were calculated with the test group except "cv_metric_AUC", that indicates the results of the 5 or 10 k-fold rcv. The SVM model is: asvs-SVM_k5. The rest are the RF models. Datasets: asvF - filtered ASVs, asvUnF – unfiltered ASVs, cl: Clusters. 40

Figure 21. Performance of the RF models that used the synthetic dataset. 41

Figure 22. Biomarkers. Features with median scaled importance value across all 100 RF models over 20. Bars indicate the 25 and 75 percentiles. 42

LIST OF TABLES

Table 1. Temporal plan of tasks. 9

Table 2. Correlations of selected environmental variables with PC1 and PC2 axis of PCA analysis for Envision and Dimension campaigns. Results are shown in decreasing order. Variable names in Envision: large protists biomass (BBlp), small protists biomass (BBsp), large protists abundance (AFlp), small protists abundance (AFsp), Synechococcus abundance (AFSyne), Synechococcus biomass (BFSyne), bacterial biomass (BB), bacterial biomass (AB), total chlorophyll a (chla Total), Prochlorococcus abundance (AFProc), Prochlorococcus biomass (BFProc). Dimension – Chlorophyll a (Chla.t), Primary production (PP) of total microeukaryotes(.m), nanoeukaryotes (.n) or picoeukaryotes(.p), community respiration (PP.h.t) and Biomass Prochlorococcus (BP). 25

Table 3. Number of samples belonging to bloom and normal events for different chlorophyll percentiles and sampling stations. 26

Table 4. Averaged values of different performance metrics for RF models 38

Table 5. Taxonomy of ASVs selected as biomarkers. No species could be determined except for asv0008. 43

1. Introduction

1.1. Background and Project Rationale

Coastal areas represent only the 7% of the total area of the oceans but produce the 14-30% of the primary production of the ocean (Pontiller et al. 2022). This productivity supports the coastal marine ecosystem, having environmental and sociological implications. Coastal fisheries play an important socioeconomic role worldwide. For instance, in Europe the small-scale fisheries, which are the 84% of the fleet, provide direct employment for 100,000 people (Lloret et al. 2018). Among coastal areas, Easter Boundary Coastal Zones (EBCZ) are of particular interest. Upwelling events occurs along EBCZ when the winds induce the rise of nutrient-rich subsurface waters causing phytoplankton blooms that sustain high levels of productivity (Figure 1) (Joglar et al. 2021; Pontiller et al. 2022).

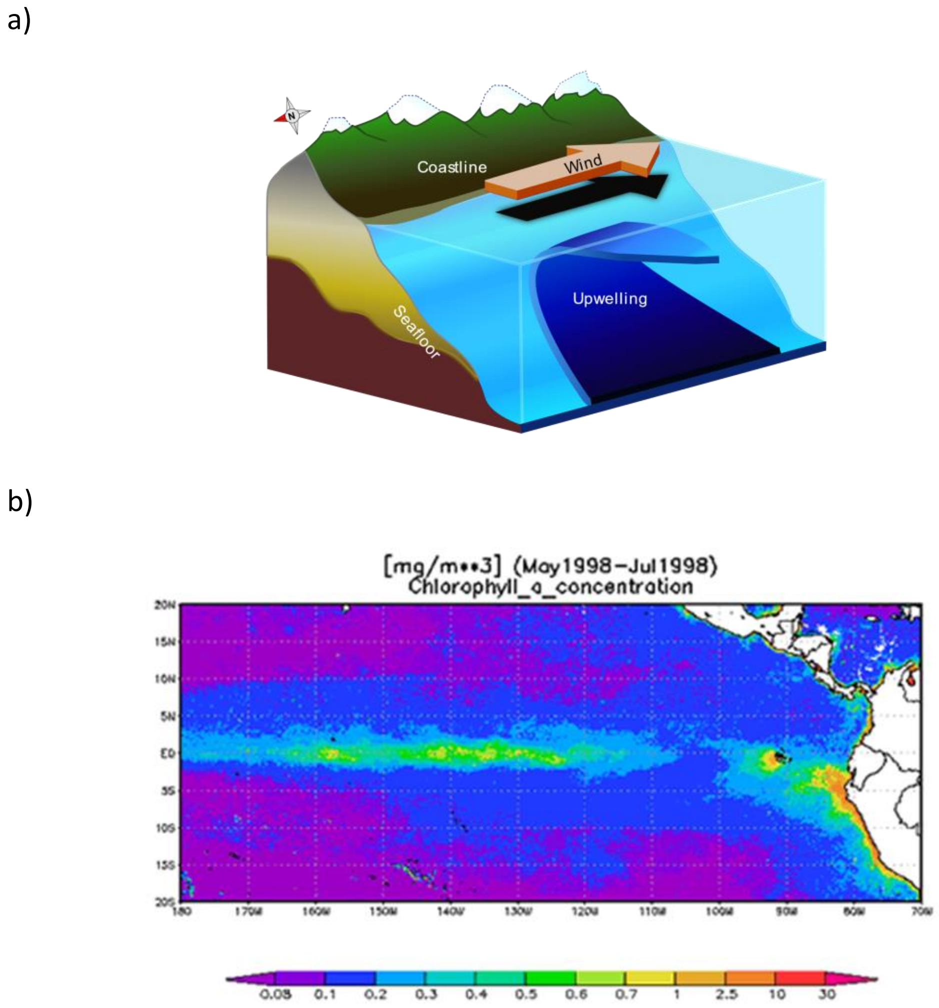


Figure 1. *a) Diagram of a coastal upwelling event, when deep cold water reaches the surface near a coast driven by wind changes (By Lichtspiel). b) Effects of upwelling on surface chlorophyll concentrations in the Western Pacific Ocean (NASA) (Source: https://en.wikipedia.org/wiki/Upwelling, accessed on Jun 6th 2023)*

Microorganisms play a central role in marine biogeochemical processes and therefore, in the marine ecosystem support. Bacterioplankton (bacteria and archaea living in the oceanic water column) are key members of marine food webs due to their high abundance and activity. Indeed, heterotrophic prokaryotes are the only organisms that efficiently transform the dissolved organic matter (DOM) in marine ecosystems (Deng, Vallet, Pohnert 2022). This is a very important process because marine photosynthetic microorganisms (cyanobacteria and single-cell eukaryote phytoplankton) are responsible for half of the primary production of Earth (figure 2) (Deng, Vallet, Pohnert 2022). Roughly, 50% of the organic carbon produced by phototrophs is transformed by heterotrophic prokaryotes, which also regulate the cycle of nitrogen and phosphorous among other elements.

Microbial communities affected by upwelling events are complex systems characterized by close interconnections of phytoplankton with bacteria, viruses, oomycetes, and herbivores (Deng et al., 2022). It is well known that temporal and spatial fluctuations of abiotic variables exert a control on ocean microbial communities. In the case of coastal marine ecosystems, the supply of inorganic nutrients impacts their productivity heavily (Paul et al. 2022). However, recent studies have recognized that biotic factors, specifically the interactions between bacterioplankton and phytoplankton, are relevant to understand the dynamics of these communities (Costas-Selas et al. 2022; Gronniger et al. 2022; Hernández-Ruiz et al. 2018).

Given the importance of marine microorganisms and their ecological and physiological differences among the different species, understanding microbial seasonal succession, interactions their consequences is highly relevant to determine and predict the adaptiveness of these communities but also, to better understand ocean functioning and the impacts of current global changes on EBCZ ecosystems (Bunse, Pinhassi 2017; Deng, Vallet, Pohnert 2022).

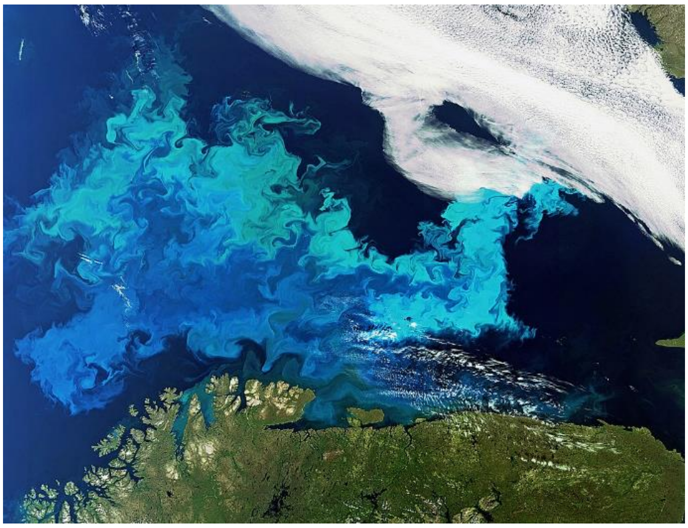


Figure 2. *Phytoplankton bloom in the Barents Sea, notice the size of the blooming area compared with the northern Norwegian coast (Envisat satellite, ESA, CC BY-SA 3.0 IGO, https://creativecommons.org/licenses/by-sa/3.0/igo/)*

In the last decades, molecular and computational advances have allowed the study of the phylogenetic and genomic diversity, and community composition patterns of marine microorganisms through the sequencing of marker genes (18S rRNA and 16S rRNA genes) or full metagenomes or metatranscriptomes (Costas-Selas et al. 2022). The expansion of high throughput sequencing is increasing the number of available microbial ecology datasets that have the potential to provide key insights into environmental phenomena. To analyse them, microbial ecology has relied on traditional statistical analysis. However, high throughput sequencing microbiome data are highly dimensional, noisy, sparse and, compositional and usually, they cannot meet the assumptions of classical statistics methods.

Machine Learning (ML) has been used to find patterns in data that can be predictive of different phenomena in biological disciplines such as neuroscience or drug discovery. However, investigation into microbial ecology applying ML models is lagging behind (Ghannam, Techtmann 2021; Marcos-Zambrano et al. 2021). ML methods present some advantages over classical statistics that are very attractive for the ecology field. Among them, the capacity of ML methods to perform robust interrogation of complex association patterns in microbial communities stands out (Ghannam, Techtmann 2021).

The main goal of this project is to take advantage of ML methods to identify the main bacterioplankton biomarkers that might be related to phytoplankton blooms and, that could serve to explore the role of biotic interactions in the highly productive EBCZ areas. The complexity of microbial communities of EBCZ areas and, the remaining questions about their functioning and capacity to adapt to the current global change and contamination issues, make them a very interesting target for ML studies. The ML methods that allow the search for biomarkers are of particular interest as they can help to disentangle the biotic interconnections of these systems.

1.2. GOALS

General objective.

Produce a ML model able to predict phytoplankton blooms events using microbial community diversity data and find relevant microbiological biomarkers of such ecological process.

Specific objectives.

1. To produce an ML model to predict phytoplankton blooms:
   1.1. Generate Amplicon Sequencing Variants (ASVs) counts per sample from raw sequencing data that will serve as initial data for ML approaches.
   1.2. Determine the range of Chlorophyll concentration that correspond to a phytoplankton bloom. This initial objective was changed through the project to

the classification of samples into two classes, bloom events or normal situations, depending on total Chlorophyll concentrations.

1.3. Generate a regression ML model capable to predict chlorophyll concentration from ASVs data with a precision at least over 70%. This initial goal was modified to create a classifier ML model capable to determine if a sample correspond to a bloom or normal situation with a performance over 70%.

2. To find relevant microbiological biomarkers:

2.1. Determine the threshold to consider a feature as a biomarker.

2.2. Generate a list of microbial species considered as biomarkers of phytoplankton blooms.

1.3. Impacts on Sustainability, Ethical Behaviour, Social Responsibility and Diversity

Sustainability

One of the main motivations to develop this project was its potential positive impacts on environmental sustainability. Achieving the main goal of this TFM, would help to better understand the coastal ecosystems functioning. That knowledge is key to better understand and predict the impacts of climate change and contamination events on coastal ecosystems; but also, it is part of the base to develop methods to mitigate those impacts, if possible. Coastal microalgae blooms are highly related with water temperature, which is rising due to global change. Blooms are also related to nutrient concentrations, which can dramatically increase due to run-off of land contaminants causing severe eutrophication events as the recent episodes occurred in the Mar Menor area of the southeast of Spain.

A direct negative impact of the development of this project is the use of computationally expensive methods. Given the current main sources of energy, the use of computationally intensive methodologies implies generation of green-house gas CO_2.

Ethical Behaviour and Social Responsibility

The deontological principles of research and academic work are followed in this project. Specifically, the data used are public or are being used with the permission of the authors; methods and results are all explained and showed, and all citations and previous works are acknowledged.

The Project is highly technical, focused in the areas of prediction and microbial ecology; therefore, it does not have direct economic or social dimensions. However, through its potential sustainability impacts, it might have a positive impact in sustainable development goals related to cero hunger (2), and decent work and economic growth (8). As the main goal of the TF can help to better understand the coastal ecosystems, it might have a future positive impact in the development of more sustainable fisheries, that would serve as a reliable source of food, and with a better distribution of resources between the different social agents. In any case, it would require the intervention of many external agents and the improvements of many other areas. A better understanding of the microbial mechanisms during the blooms is only a small portion of the knowledge required to achieve those positive impacts.

It is also related to the sustainable development objective 6, clean water and sanitation, because the water quality of the coastal areas is very related to the microbial communities that inhabits them, especially in the events of water eutrophication as already explained.

Diversity and Human Rights

Through the development of more sustainable fisheries with better resources distribution, this project might have a positive impact in the sustainable development objective 10, reduced inequalities. But as already indicated, the role of this TFM in this impact would be small and would require many changes far beyond the area of this project.

Human diversity is respected as much as possible during the TF. Inclusive language is used when needed. Search and inclusion of previous works are done regardless the sex, class, race, disability, sexual orientation, and gender of the authors. The ISO 690 citation style is used to ensure showing full names of authors. The obtained results are not directed or focused on any specific group of people.

1.4 Approach and Methodology

In the context of this project, a biomarker could be either a taxonomic marker or a gene related to any relevant biological process, where the most interesting ones could be those related to metabolic or cell signalling pathways. Because of that, the analysis would be initially focused on metagenomic data. Taking advantage of a set of metagenomes and corresponding environmental metadata of a seasonal study of Northwest coast of the Iberian Peninsula (Envision campaign). This data is available through a collaboration with oceanographers and marine ecologists: professors Sandra Martínez-García and Eva Teira from “Centro de Investigación Mariña da Universidade de Vigo, Departamento de Ecoloxía e Bioloxía Animal (CIM-UVigo)” in a wider research project in which bioinformatics and classical statistics are being applied to study the N-cycle on the EBCZ of Galicia (An approach to study the ecology of marine microbial communities based on traits and guilds – TRAITS). Some results using part of this dataset have been published already (Pontiller et al. 2022).

In microbial ecology studies, especially those using metagenomics, the number of samples that can be used as instances in ML approaches, is usually very low compared to other datasets. Envision dataset is not an exception as it comprises only 23 samples. An alternative could be search for biomarkers using only the prokaryotic taxonomic marker gene (16S rRNA gene) as the microbial ecology studies using that kind of data usually comprises a higher number of samples due to the lower costs of sequencing and data analysis. And again, Envision dataset is not an exception, as it contains 130 samples of 16S rRNA gene amplicon data. In addition, collaborators from U. of Vigo have other dataset of marker genes of a seasonal study developed in the same EBCZ in the previous years (Dimension, (Hernández-Ruiz et al. 2018) that can complement Envision dataset. In any case, the number of samples available is low, what is a concern for ML methods. Therefore, the project must start with a search on bibliography and several databases (i.e: European Nucleotide Archive (ENA), Joint Genome Institute Integrated Microbial Genomes and Microbiomes (JGI-IMG/M) or TARA Oceans databases (Sunagawa et al. 2015; Salazar et al. 2019) to find similar and available datasets of coastal phytoplankton blooms.

However, even though the sequences are usually available, most studies do not make available the corresponding environmental metadata, at least in the detailed level needed in this project. Because of that, it is highly probable that no similar datasets will be found. In the case of finding a set of available metagenome samples there are more difficulties. The initial steps of a metagenome analysis (assembly, gene identification and annotation) are characterized by long computing times even in high performance scientific clusters. For instance, the assembly and annotation of the Envision metagenomes, already done in the context of the TRAITS project, took a couple of weeks in the same server used in this project. Due to the time limitations in this project, performing again those analysis with a new group of metagenomes including Envision samples is not possible. Consequently, the search for compatible metagenomes must be focused on data already assembled and annotated with a pipeline compatible with the data analysis already performed with Envision metagenomes, which is highly improbable. In the case of finding appropriate metagenomic datasets, a table of sequence counts of genes vs. samples would be the data used in the ML approaches.

In the case of marker gene amplicon data, the computational times are significantly lower, which would allow to start the analysis from raw sequences. In the context of ecological studies, the first step when dealing with raw sequences of taxonomic marker genes is to produce tables of species vs. counts on each sample. This is the starting point for the standard statistical analysis, and the data that will be used in the ML approaches if no metagenomic data is found. To create those tables, sequences must undergo a series of processing steps to calculate a proxy for microbial species. Currently, there are two approaches for that. First, calculate Operational Taxonomic Units (OTUs) based on sequencing similarity (usually 97% or higher). A more recent approach is to calculate Amplicon Sequence Variants (ASVs) that discriminate biological sequences from sequencing error and have finer resolution (Callahan, McMurdie, Holmes 2017). Although there is some debate in the community regarding the use of both approaches (Glassman, Martiny 2018), ASVs will be used in this project because this approach has been claimed as more precise, reusable, reproducible and comprehensive (Callahan, McMurdie, Holmes 2017). Each ASV will be a feature of the dataset. This diversity data can be aggregated in different taxonomic levels (specie, genus, family, order…) that could be used as features instead the ASVs. However, a recent work has shown that ML algorithms perform better using more detailed diversity data (i.e., ASVs, species or genera) than lower taxonomic ranks that aggregate the information (i.e., order or class) (Wilhelm, van Es, Buckley 2022). All these steps can be performed with Mothur pipeline (Schloss 2020).

The microbiome sequencing data (either rRNA marker genes, metagenomes or metatranscriptomes) is characterized by noise, compositional nature, and sparsity, factors that can impact the performance of ML models. Despite of that, environmental microbiology studies usually fail to acknowledge these problems (Busato et al. 2023). Therefore, exploratory analysis will be employed to reduce the possible source of noise (i.e., batch effects); appropriate normalizations (i.e., log-ratio transformations) will be applied to avoid compositional problems; and addition of pseudo-counts will be used to reduce the sparsity problem. Also, the number of ASVs (features) of this kind of datasets are usually in the range of few thousands, making these datasets highly dimensional. A feature selection step will be used (i.e., ASVs with low abundances and present only a low fraction of the samples would be discarded).

In addition to ASVs or genes, other data describing the environment is crucial to achieve the goals of the project. The chlorophyll concentration of the water that is a proxy for phytoplankton growth and therefore, its blooms (Deng, Vallet, Pohnert 2022). A regression model to predict chlorophyll concentrations from DNA data is more interesting than using a classification approach because it can allow to infer more detailed links between the microbiota and the Chlorophyll concentrations. Therefore, this kind of models will be the target of the project. In any case, samples can be also divided between those corresponding to the bloom events or normal situation using chlorophyll concentrations, may be with the help of other environmental metadata. The input of the collaborators, professors Sandra Martínez-García and Eva Teira, would be crucial for this step.

Given the nature of the goal and data, the project will use a supervised learning methodology. Among the different possibilities capable of regression analysis (i.e., k-nearest neighbours, support vector machines or artificial neural networks), random forest (RF) is the ML method preferred when dealing with microbiome data (Busato et al. 2023; Ghannam, Techtmann 2021; Marcos-Zambrano et al. 2021; Janßen et al. 2019). The reasons are that RF are less prone to overfitting, can consistently identify true effects in complex and heterogeneous data, usually obtains better accuracies and in general, are considered interpretable and meaningful information can be extracted from their classification steps (Liu et al. 2022; Ghannam, Techtmann 2021; Janßen et al. 2019). Due to these advantages, RF will be used to predict the blooms, using chlorophyll as a proxy, as well as to determine the biomarkers associated to them. Although support vector machines, artificial neural networks and deep learning has some prevalence on studies of microbial communities, mainly from human microbiome using ML methods (Marcos-Zambrano et al. 2021), those approaches have been discarded for this project. First, the black-box nature of artificial neural networks and support vector machines make them a bad choice to search for biomarkers among the features. Second, the low number of microbiome samples available from marine environments limits the application of artificial neural networks or deep learning on the project. There are few examples that applies data augmentation or transfer learning methods to microbiome studies but, an in all cases authors used the largest available dataset of microbial community samples (human gut microbiomes) including thousands of samples in their methods (Tataru, David 2020). Again, due to the limited number of samples available for oceanic environments, these methods are initially discarded.

After training, the RF model will be adequately evaluated and refined. Then, the importance of different features in the RF model will be calculated and used to identify the best biomarkers. There are limited examples of biomarker search using ML on environmental microbiomes (Wilhelm, van Es, Buckley 2022; Janßen et al. 2019). For that the variable importance of the top features to predict the Chlorophyll concentrations will be determined using information during the forest construction (Liu et al. 2022; Yuan et al. 2022). All the work related to ML algorithms will be done in the R environment using `caret` package (R Core Team 2023).

1.4. WORKING PLAN

Resources

The required resources needed are:

- Data, both DNA sequences and environmental metadata, already available through a collaboration with professors Sandra Martínez-García and Eva Teira from University of Vigo, and on public databases.
- Computational resources, available through access to the scientific computation servers of the National Center of Biotechnology – CSIC.
- Personal computer.

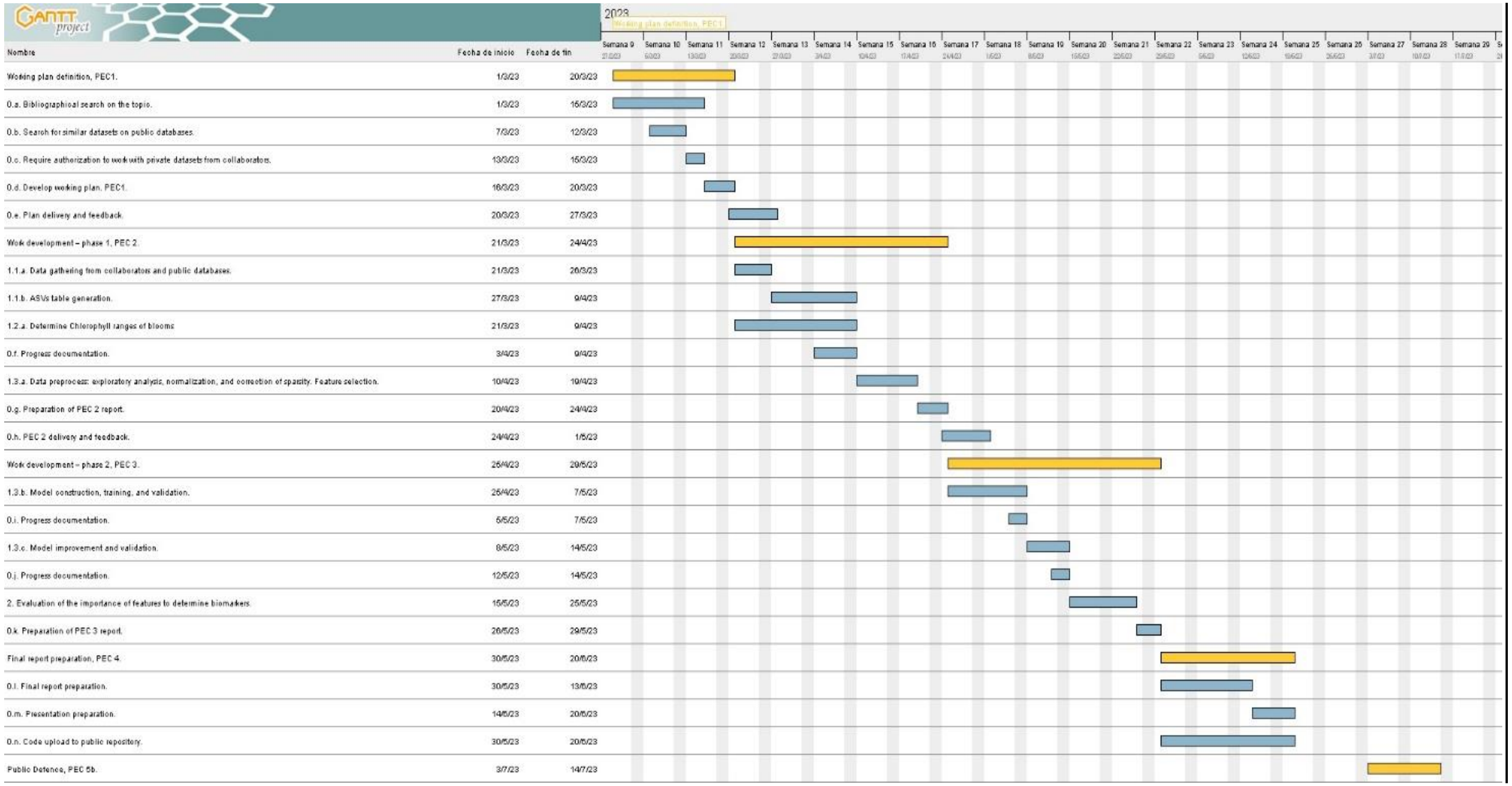
Tasks

Tasks are enumerated according to the objectives of the project: the numeric part corresponds to the specific objective, then a letter enumerate the task itself. For those tasks relate to documentation and preparation of reports, the numeric part of the label has been set to 0. Note that the date format used in this project is dd/mm/yyyy.

***Table 1**. Temporal plan of tasks.*

DESCRIPTION	**START**	**END**
Working plan definition, PEC1.	**01/03/2023**	**20/03/2023**
0.a. Bibliographical search on the topic.	01/03/2023	15/03/2023
0.b. Search for similar datasets on public databases.	07/03/2023	12/03/2023
0.c. Require authorization to work with private datasets from collaborators.	13/03/2023	15/03/2023
0.d. Develop working plan. PEC1.	16/03/2023	20/03/2023
0.e. Plan delivery and feedback.	20/03/2023	27/03/2023
Work development – phase 1, PEC 2.	**21/03/2023**	**24/04/2023**
1.1.a. Data gathering from collaborators and public databases.	21/03/2023	26/03/2023
1.1.b. ASVs table generation.	27/03/2023	09/04/2023
1.2.a. Determine Chlorophyll ranges of blooms.	21/03/2023	09/04/2023
0.f. Progress documentation.	03/04/2023	09/04/2023
1.3.a. Data preprocess: exploratory analysis, normalization, and correction of sparsity. Feature selection.	10/04/2023	19/04/2023
0.g. Preparation of PEC 2 report.	20/04/2023	24/04/2023
0.h. PEC 2 delivery and feedback.	24/04/2023	01/05/2023
Work development – phase 2, PEC 3.	**25/04/2023**	**29/05/2023**
1.3.b. Model construction, training, and validation.	25/04/2023	07/05/2023
0.i. Progress documentation.	05/05/2023	07/05/2023
1.3.c. Model improvement and validation.	08/05/2023	14/05/2023
0.j. Progress documentation.	12/05/2023	14/05/2023
2. Evaluation of the importance of features to determine biomarkers.	15/05/2023	25/05/2023
0.k. Preparation of PEC 3 report.	26/05/2023	29/05/2023
Final report preparation, PEC 4.	**30/05/2023**	**20/06/2023**
0.l. Final report preparation.	30/05/2023	13/06/2023
0.m. Presentation preparation.	14/06/20203	20/06/2023
0.n. Code upload to public repository.	30/05/2023	20/06/2023
Public Defence, PEC 5.	**03/07/2023**	**14/07/2023**

Gannt diagram



Nombre	Fecha de inicio	Fecha de fin
Working plan definition, PEC1.	1/3/23	20/3/23
0.a. Bibliographical search on the topic.	1/3/23	16/3/23
0.b. Search for similar datasets on public databases.	7/3/23	12/3/23
0.c. Require authorization to work with private datasets from collaborators.	13/3/23	15/3/23
0.d. Develop working plan. PEC1.	16/3/23	20/3/23
0.e. Plan delivery and feedback.	20/3/23	27/3/23
Work development – phase 1, PEC 2.	21/3/23	24/4/23
1.1.a. Data gathering from collaborators and public databases.	21/3/23	26/3/23
1.1.b. ASVs table generation.	27/3/23	9/4/23
1.2.a. Determine Chlorophyll ranges of blooms	21/3/23	9/4/23
0.f. Progress documentation.	3/4/23	9/4/23
1.3.a. Data preprocess: exploratory analysis, normalization, and correction of sparsity. Feature selection.	10/4/23	19/4/23
0.g. Preparation of PEC 2 report.	20/4/23	24/4/23
0.h. PEC 2 delivery and feedback.	24/4/23	1/5/23
Work development – phase 2, PEC 3.	26/4/23	29/5/23
1.3.b. Model construction, training, and validation.	26/4/23	7/5/23
0.i. Progress documentation.	6/5/23	7/5/23
1.3.c. Model improvement and validation.	8/5/23	14/5/23
0.j. Progress documentation.	12/5/23	14/5/23
2. Evaluation of the importance of features to determine biomarkers.	15/5/23	25/5/23
0.k. Preparation of PEC 3 report.	26/5/23	29/5/23
Final report preparation, PEC 4.	30/5/23	20/6/23
0.l. Final report preparation.	30/5/23	13/6/23
0.m. Presentation preparation.	14/6/23	20/6/23
0.n. Code upload to public repository.	30/5/23	20/6/23
Public Defence, PEC 5b.	3/7/23	14/7/23

Figure 3. *Gantt diagram.*

Milestones

Description	Date
Working Plan delivery	20/03/2023
Work development phase 1 delivery	24/04/2023
Work development phase 2 delivery	29/05/2023
Final report delivery	20/06/2023
Presentation delivery	20/06/2023
Public defense	03-14/07/2023

1.5. SUMMARY OF OBTAINED PRODUCTS

Working plan.
A PDF document that includes the problem to address, its background and importance, the objectives of the project and, de detailed working plan to achieve them.

Final report.
A full report on PDF format of the work developed during the project. It details the methods developed, the results obtained, and conclusions deduced from them.

Product.
A public repository where the code developed during the project is available.

On-line presentation.
Presentation document in ppt format to show the performed work and results.

1.6. SHORT DESCRIPTION OF OTHER CHAPTERS

The other chapters of this memory include:

2. Materials and methods. Detailed explanation of data sources and methods used, including the generation of Amplicon Sequence Variants, the training and improvement of Random Forest models and the selection of important features.
3. Results and discussion. Section that includes a description of the data finally included in the project, the results of sequences and data preprocessing, the performance of classifiers and the biomarkers found.
4. Future work and conclusions. Conclusions drawn from current results, discussion about future extensions of the work, and impacts on sustainable development goals.
5. Glossary. Explanation of the most used terms and acronyms.
6. Bibliography. Full description of cited articles, books, and websites.
7. Appendices. Section that includes extra plots and tables.

2 MATERIALS AND METHODS.

2.1 DATA SEARCH AND SELECTION.

2.1.1 Search of public databases for additional data.

Initially, the search of biomarkers was planned for metagenomic data from the oceanographic campaign Envision (Pontiller et al. 2022). However, Envision metagenomic data consisted of only 23 samples to serve as instances. As an alternative, 16S rRNA gene data was also considered. The two available campaigns (Envision and Dimension) comprised only 166 samples (see section 2.1.2), which still is a low number of instances. Consequently, similar datasets were searched on bibliography and public databases, all accessed several times between March 7th to 12th, 2023:

- Bibliography searches were preformed using Google Scholar (Google 2004) and Scopus (Elsevier 2004).
- Public repositories:
    - European Nucleotide Archive (EMBL-EBI 2023).
    - Joint Genome Institute Integrated Microbial Genomes and Microbiomes (DOE-JGI, 2006; Markowitz et al., 2006).
    - Qiita (Gonzalez et al. 2018).
    - Earth Microbiome Project (Thompson et al. 2017)
    - TARA Oceans databases (Sunagawa et al. 2015; Salazar et al. 2019)

2.1.2 Envision and Dimension campaigns.

Once metagenomic data was discarded as an option (see Results and Discussion section), two kinds of data were needed for the project. The partial 16S rRNA gene sequences (fastq files) and the environmental data including the chlorophyll concentrations and other relevant variables. The 16S rRNA gene sequences and environmental data were collected in two oceanographic campaigns developed in the North-West coast of the Iberian Peninsula in an upwelling system near Ría de Vigo (Dimension and Envision) (Joglar et al. 2020; Hernández-Ruiz et al. 2018). Sampling was done at two different locations of the east Atlantic Ocean: one coastal station (st 3) (42° N, 8.88° W), and one oceanic or offshore station (st 6) (42° N, 9.06° W) (Figure 4) (Joglar et al. 2020; Hernández-Ruiz et al. 2018). In both projects, samples to determine microbial community composition along with several other environmental variables were collected before, during and after phytoplankton blooms on an EBCZ.

In Dimension campaign, monthly seawater sampling was carried out only in the coastal station from January 2014 to November 2015. Between July and August 2014 sampling was not possible due to ship technical issues. Two depths were sampled, 1 m. and 30 m.

In Envision, both stations were sampled during different seasons along 2016. For that, three one weeklong cruises took place on February, April, and August. On them, samples were taken each other day at several depths (from 5m to 200 m) depending on the ocean conditions. More details about the sampling strategy and methodology can be found in (Hernández-Ruiz et al., 2018; Joglar et al., 2020).

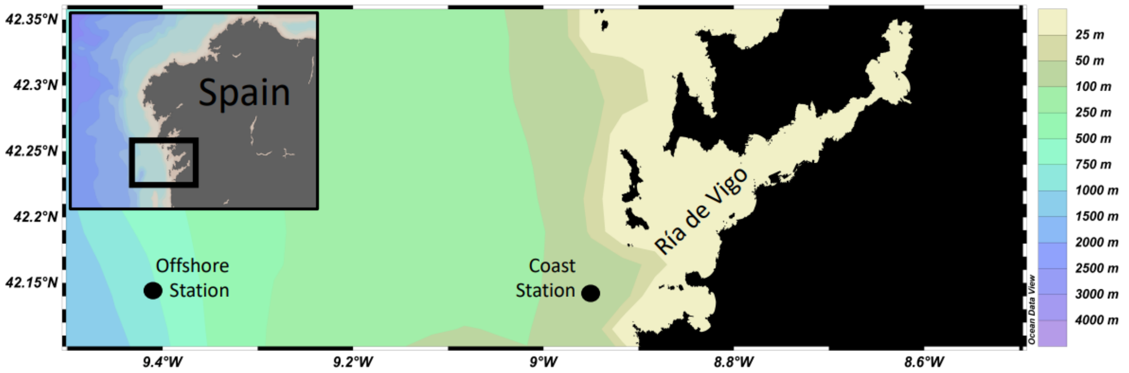


***Figure 4**. Location of coastal and oceanic sampling points in the North-West coast of the Iberian Peninsula. (Courtesy of Prof. Sandra Martínez-García).*

All data, except the raw partial 16S rRNA gene sequences of Envision, were directly obtained from the University of Vigo collaborators. The raw sequences of the Envision project were obtained from the Sequence Read Archive (SRA) public database (bioproject ID PRJEB36188). SRA-explorer (Phil Ewels 2014) and Aspera 4.1.9.93 (IBM) were used to download the fastq files of the raw sequences. Also, the table with the information of the bioproject was directly downloaded from the SRA website as a .csv file.

Initially, the available files included:

Envision:

- 321 fastq files of raw partial rRNA sequences. This set contained samples taken in the field but also from incubation experiments that were not of relevance in this project. Also, they included either 16S or 18S rRNA partial genes targeting Prokaryotic and Eukaryotic diversity.
- The sequencing form used when submitting the samples to a company for Illumina sequencing. This file was the only soured of information to correctly identify fastq files with the samples of interest (16S rRNA gene sequences taken in the field).
- 3 metadata tables with sampling and environmental information such as date and depth of sampling, chlorophyll concentrations or several inorganic and organic measurements (nitrate, nitrite, or prokaryotic biomass).

- 6 files with data from a CTD, an oceanographic instrument that measures conductivity (C), Temperature (T) and Depth (D). Each one contained data from one sampling station taken during one of the months studied in this campaign.

Dimension:

- 74 fastq files of partial 16S rRNA gene sequences.
- Sample – fastq files table correspondence for the first year of the project (2014).
- Metadata table with information about sampling date, depth, chlorophyll concentrations and, other environmental variables including those measured with the CTD.

2.2 ENVIRONMENTAL METADATA AND SEQUENCE FILES PREPARATION AND PREPROCESS.

Data gathered from collaborators and SRA lacked a system to unequivocally identify samples between fastq files and metadata tables because different naming systems were mixed. To avoid that problem, a common sample naming system, compatible with downstream analysis tools (Mothur), was created and applied to all files in the project:

XXXProk###

where XXX corresponded to either ENV or DIM for each project and ### indicated the sample number. Then, sample names were changed in all files (fastq files, SRA bioproject information table and, metadata tables) using information from different sources depending on the file. For instance, in the Dimension project, names of fastq files did not match the sample names in the metadata table. The available table with correspondences between fastq file names and sample names only covered the first year (2014). To match the fastq files and sample names on metadata table of the second year (2015) information of the sampling date had to be extracted from the fastq file names and contrasted with the information contained in the metadata table. Then, the name of both fastq files and samples could be changed to the new common system.

The initial pre-process included the recodification or renaming of several environmental variables because they were named or recorded with different coding systems even within the same project. For example, in the three initial metadata tables of the Envision project, one for each month, the depth level, a categorical variable, was encoded as *prof 1* - *prof 7* in February table, and as *p1* - *p7* in the other cases. Finally, environmental variables with missing data that could not be imputed were removed. For instance, that was the case of primary production measurements that were completely missing for ocean station and not measured several times in the coastal station.

The files preprocess was complex and dependent on the campaign and file that was being processed. An overall outline can be found in figure 5, and details of the process can be found in the companion code (see section 2.10 and appendix 7.6).

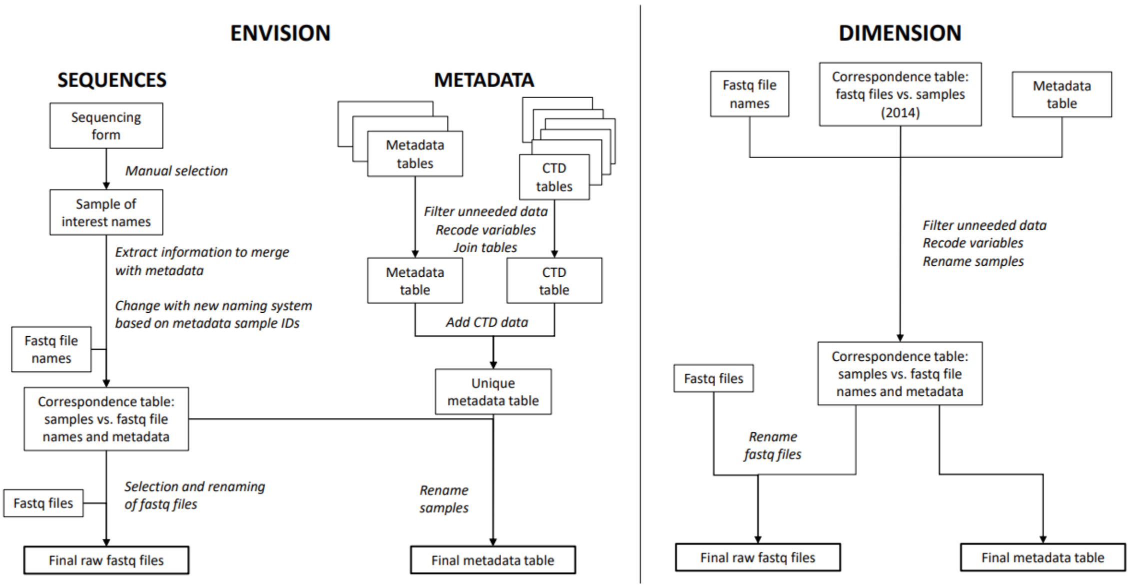


Figure 5. *Initial pre-process outline for ENVISION and DIMENSION data and files.*

2.3 CLASSIFICATION OF SAMPLES BASED ON CHLOROPHYLL CONCENTRATIONS.

Despite that the initial plan was training a regression model using chlorophyll concentrations as outcome variable, finally oceanographers decided that to classify the samples between normal and bloom situations would be a better approach. To help oceanographers to classify samples between each event, an exploratory analysis of some environmental variables was performed.

For that, PCA analyses for both Envision and Dimension campaigns were performed using numeric environmental data and not the biological data (chlorophyll and abundances and biomass of Prokaryotes and Eukaryotes in Envision, chlorophyl and primary productivity in Dimension). Prior to PCA computation, samples with missing values were filtered and variables were scaled. Then the correlations of environmental variables to the first couple of components were explored. Categorical variables (month, depth) were used to colour the samples in the PCA plot to explore sample distribution according to them.

Collaborators from U. of Vigo also requested an analysis of sample frequency distribution through months, seasons and depths using different percentiles of total chlorophyll concentrations to divide the data in groups. Sample distribution among groups defined by 50, 75 and 90 percentiles of total chlorophyll concentration were examined. This later approach was preferred by the oceanographers to decide the division of samples between bloom and no-bloom categories. The 75 percentile was used in the sampling station closer to the coast while the 90 percentile was used for the oceanic station. Samples with chlorophyll concentrations above the corresponding percentile were classified as “bloom” and samples below it as “normal”.

2.4 AMPLICON SEQUENCE VARIANTS GENERATION.

To generate Amplicon Sequence Variants (ASVs), raw partial 16S rRNA gene sequences from Envision and Dimension were jointly treated. Those sequences were obtained with 515F-Y and 926R primers (Parada, Needham, Fuhrman 2016) and sequenced in an Illumina platform (Hernández-Ruiz et al., 2018; Joglar et al., 2020).

2.4.1 Sequences pre-processing

The starting data were partially overlapping paired sequences. Initially, the presence of 16S rRNA primers used to generate the data, on raw sequences was checked with TagCleaner (Schmieder et al. 2010). As the primers were present and removing the highly conserved primer sequences is important to correctly generate ASVs, primers were trimmed from paired reads using the tools of Mothur pipeline (Schloss 2020).

Next, trimmed reads were filtered to remove low quality sequences. In this step, paired reads that are missing a partner or sequences with low quality values or sequencing errors were removed using Moira (Puente-Sánchez, Aguirre, Parro 2016). Moira applied a Poisson binomial distribution to estimate sequencing errors. Then, Moira merged paired reads into consensus sequences. High quality paired sequences that partially overlap, were merged to generate a single and longer sequences that were used as input data for ASVs generation.

2.4.2 Generation and taxonomic assignment of ASVs.

DADA2 was used to calculate ASVs (Callahan et al. 2016). Apart from calculating ASVs, DADA2 pipeline also detected and removed chimeric sequences. Due to the conserved regions within the 16S rRNA gene, chimeric sequences can be generated during DNA library preparations (i.e., during Polymerase Chain Reaction), sequencing or data analysis. The taxonomic inconsistency of different sequence regions is the key to remove them from the dataset. Finally, DADA2 assigned taxonomy of ASVs using a native implementation of the naïve Bayesian classifier method (Wang et al. 2007) and SILVA v.138 database (Quast et al. 2013).

2.5 DATA PREPARATION

DADA2 results needed some preprocessing prior to work on R. DADA2 does not create ASVs names, but instead uses the sequence as ID. Therefore, ASVs were numerically named while the sequence match with those IDs was recorded. Then, ASVs corresponding to eukaryotes, mitochondria and chloroplasts were removed from the ASVs vs. samples counts table using taxonomic assignments to locate them. A problematic Envision sample (ENVProk141) was also removed.

Next, metadata from both Envision and Dimension were merged. Most of the environmental variables were removed as many of them were not present in both datasets and others were not further needed. Also, a new variable with depth levels was added. It classified the numeric value of depth (in meters) into three different levels: surface (5-30 m), intermediate (40-55 m) and deep (60-200 m). Finally, the sample classification between *bloom* and *normal* situations based on chlorophyll concentrations was encoded into a variable called *event*.

2.6 ASVS FILTERING, CLUSTERS GENERATION AND EXPLORATORY ANALYSIS.

To reduce dimensionality, ASVs were filtered based on relative abundances. The ASVs with a mean relative abundance below 0.01% were removed. Also, not filtered ASVs were joined according to their taxonomic assignments at the genus level into clusters.

These preprocessing steps produced three different datasets of biological features that were used throughout the project:

- Filtered ASVs (relative abundance > 0.01%).
- Clusters (ASVs joined at genus level).
- Unfiltered ASVs.

Data characteristics of either filtered or unfiltered ASVs and clusters were explored visualizing data distributions. Also, NMDS ordinations were calculated. Two options were included. First, ASVs or clusters count data was transformed with robust centered log-ratio transformations (rclr), Euclidean distance was calculated and the best solution for a NMSD ordination was searched with at least 500 random starts. A second approach was to apply a total sum transformation to the ASVs or clusters count data, that is equivalent to calculate proportions, and then calculate their square root followed by the search of the best NMDS ordination using Bray-Curtis dissimilarities with 500 random starts.

2.7 RANDOM FOREST MODELS.

2.7.1 Feature selection and preprocess

Features used to train the models included both metadata variables (i.e., depth, season) and biological features. Therefore, the full datasets were a mix of categorical and numeric variables. Categorical variables were recoded using one-hot encoding.

Biological features (ASVs or Clusters) were transformed as previously described to deal with differences in sequencing depth. Then, all features were centred and scaled.

Several criteria for feature selection were applied in all cases considered:

- Perfectly correlated features were joined into groups and only one member of each group was included in the dataset.
- Features with variances near zero were filtered from the dataset.
- Biological features were filtered by prevalence, removing those that appeared only in few samples. Different percentage of samples were explored for this filter (1% ~ 2 samples, 2% ~ 3, 3% ~ 5, 5% ~8,10% ~17 and, 20% ~33).
- As already mentioned, ASVs with relative abundances lower than 0.01% were removed from the data in one of the cases considered.

2.7.2 Model training, validation, and performance.

Datasets were split into train and test groups, keeping the 80 % of instances in the train group and the other 20 % in the test group. This division was done considering the imbalanced nature of the dataset (see results section 3.2.2); therefore, the train and test group kept the proportion between the two classes of the outcome variable.

Initially, datasets were split one time and a RF model was trained (figure 6). The main goal of this initial model training was to explore the hyperparameters values to consider; but also, to estimate the computing times needed to train the models using several datasets splits (see below). The RF function used (`rf` function of R `caret` package, wrapped in the `mikropml` package) had one hyperparameter that allowed tunning: *mtry*, that is the number of randomly selected features to consider on each decision tree of the forest. The number of trees on the forest was fixed by the function to 500. During this first train, a grid search was performed to explore different values for the hyperparameter *mtry*. The range of values was:

[sqrt(F)/2, sqrt(F), sqrt(F)x2]

where sqrt(F) is the square root of the number of features. The different *mtry* values were validated using repeated k-fold cross-validation (RCV). Two values of k were wxpored in the k-fold RCV: 5 and 10. The number of repetitions were 10 for ASVs datasets and 100 for clusters datasets due to the computational cost. Area Under the Curve (AUC) was used as the performance metric to select the best *mtry* value. Then, the best model performance was evaluated with the test data.

The results of the k-fold RCV were used to redefine the grid-search for hyperparameter *mtry*. Then, the same pipeline to train, validate the new *mtry* values and calculate the performance of the best model, was applied to 100 different datasets splits to obtain

100 RF models (figure 6). Finally, the averaged performance of this set of RF models was calculated using several metrics: AUC, Accuracy, Kappa, Specificity and, Sensitivity considering the “normal” level of the outcome variable as positive. This approach was based on the one proposed by Begüm D. Topçuoglu et al. that implemented a pipeline according to good practices in the literature (Topçuoğlu et al. 2021).

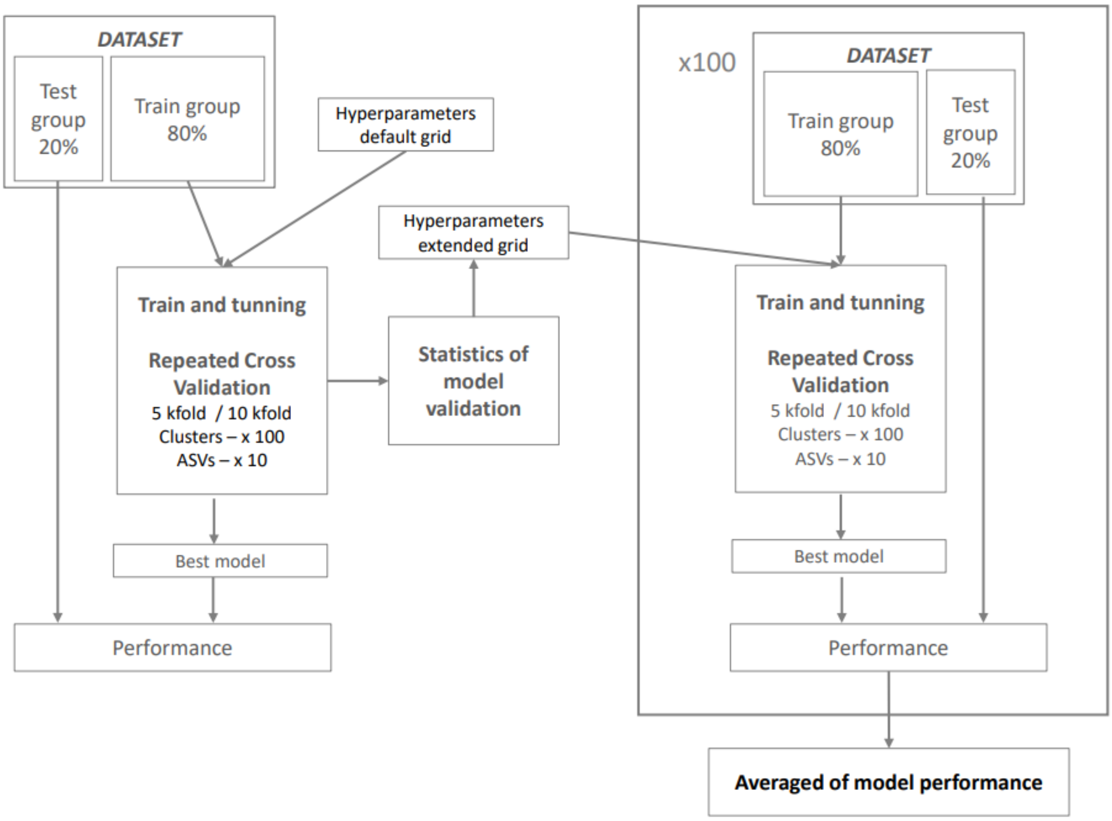


Figure 6. *Diagram of the protocol used to train and tune the models.*

2.8 Improving model performance.

2.8.1 A Support Vector Machine model.

A Support Vector Machine model (SVM) was trained using the non-linear Radial Basis Function kernel. In this case only the filtered ASVs dataset was used. The same pipeline used in the RF models for data preprocess, model training, validation and evaluation of performance was applied for the SVM case. The only exception was that the initial model train step, in which only one dataset split was considered to explore the validation of the hyperparameters was skipped. The SVM validation was performed using the hyperparameters grid-search defined by default by the function used (`run-ml` in `mikropml` package): cost hyperparameter (*C*): 1x10^{-3}, 1x10^{-2}, 1x10^{-1}, 1, 1x10^{1}, 1x10^{-3} and *sigma*: 1x10^{-6}, 1x10^{-5}, 1x10^{-4}, 1x10^{-3}, 1x10^{-2}, 1x10^{-1}. The validation strategy was a 5 k-fold RCV repeated 10 times.

2.8.2 Synthetic data generation.

Synthetic Minority Oversampling Technique (SMOTE) method was used to deal with the imbalanced nature of the datasets (Kovács 2019). The approach was applied only to the filtered ASVs dataset. The pipeline in this case was a modification of the one already described (section 2.7.2) with the following modifications:

- The initial RF forest model was not trained and the *mtry* hyperparameter grid search was based on the results obtained for the previous RF models trained on the filtered ASVs dataset. The values tested were 8,16,32,48,64.
- Only a 5 k-fold repeated cross-validation (x10) approach was used for validation, with Accuracy as metric to select the best value for *mtry*.
- Dataset was split into train and test groups 100 times as previously described. However, the data on the train group was used to generate new synthetic data with the SMOTE method to sample up the ”bloom” level to the same frequency to the “normal” level of the outcome variable.
- These models were trained in R as in previous cases, but the functions used to train the model differed. For convenience, prior models were trained with the `run_ml` function of the `mikropml` package that also performs validation and calculate models’ performance. Although this function is a wrapper of several `caret` tools, it did not allow a direct implementation of the SMOTE method. Therefore, the new models were trained and validated using `caret train` function directly. Therefore, SMOTE method was applied through the `trainControl` function using the default options, which called the `smote` function from `themis` package.
- Also, performance was directly calculated with `confusionMatrix` function after making predictions for the test group. In this case, the positive class was “bloom”.

2.9 DETERMINATION OF FEATURE IMPORTANCE.

Feature importance was obtained from the RF models trained with synthetic data. The protocol used was the one described by Leo Breiman, 2001 (Breiman 2001). To determine feature importance the prediction error rate or each feature variable was calculated from permuting out-of-bag data during the forest building procedure. Then, the importance was scaled from 0 to 100. A feature was considered as important if its median scaled importance value across all 100 RF models was over 20.

2.10 Hardware, Software and Code Availability

Hardware used.

Most of the work has been developed in a personal computer with the following characteristics:

- Processor: AMD Ryzen 7 4800H with 16 cores.
- RAM memory: 64 Gb.
- Operational system: Windows 11.

The most intense computing steps (sequences processing, ASVs calculation and most models training with 100 data splits x 100 repeated cross validation) were performed in a scientific computing server located at National Center of Biotechnology (CSIC):

- Processor: 104 CPUs
- Total RAM memory: 1.48 Tb
- Operational system: Ubuntu 20.04. Environments managed with Conda.

Softare used.

All scripts and code were developed on Bash (Free Software Foundation 2023), python 3.8.10 (Python Software Foundation 2023) or R 4.1.3 (R Core Team 2023). Other software:

UBUNTU LTS 20.04(UBUNTU community 2020)

Data management:

SRA-explorer (Phil Ewels 2014), Aspera 4.1.9.93 (IBM)

Sequences processing and ASVs generation:

TagCleaner 0.16 (Schmieder et al. 2010); Moira v1.3.2 (Puente-Sánchez, Aguirre, Parro 2016), Mothur 1.36 (Schloss 2020)

Main R libraries used. For individual references, please see CRAN-repository (CRAN Team 2023):

- Data management: tidyverse 2.0.0, purrr 1.0.1, tibble 3.2.1
- Statistical analysis: vegan 2.6-4
- Plots: ggplot2 3.4.2, RColorBrewer 1.1-3, ggpubr 0.6.0
- Parallel computing: future.apply 1.10.0, doFuture 1.0.0, future 1.32.0, tictoc 1.2
- Machine learning methods: caret 6.0-94, mikropml 1.6.0, themis 1.0.1

Code availability.

All code developed during this project is available on the GitHub repository:

https://github.com/micronuria/envisdim_ml

A description of repository content is included in the Appendix 7.6 of this memory.

3 RESULTS AND DISCUSSION

3.1 DATA INCLUDED IN THE PROJECT.

The initial idea was to apply ML algorithms to search for bloom biomarkers among different microbial processes (i.e., metabolic pathways and cell-signalling) and taxonomic groups using metagenomic data available within the Envision project (Pontiller et al. 2022). However, the low number of samples available, only 23, was a concern. A search on bibliography and several databases to find similar and available large metagenomic datasets of coastal phytoplankton blooms (including both sequence data and the corresponding metadata) was unsuccessful. Few promising microbiome studies which included coastal datasets during blooms were found. The metagenomes of TARA Oceans project (Sunagawa et al. 2015; Salazar et al. 2019) were discarded as the data analysis was different from the pipeline used in Envision metagenomes, which raised concerns about compatibilities between both datasets. To solve this problem, all data had to be realized together starting with metagenomes assembly, which was not possible within the temporal frame of this project. A second option was a work on a succession of bacterioplankton populations induced by a phytoplankton bloom (Teeling et al. 2012). Unfortunately, sequences data was not available in any database.

There are previous studies on environmental microbiomes that applied ML methods to low number of metagenomes (in the order of few dozen samples, (see for instance several references cited in (Ghannam, Techtmann 2021; Marcos-Zambrano et al. 2021; Li et al. 2022), but the reported results and model accuracy were quite poor. Therefore, using metagenomic data was considered too risky and discarded. Instead, the objective was focused on finding biomarkers among prokaryotic taxa using 16S rRNA gene sequences, for which more samples (130) were available in the Envision project (Joglar et al. 2020). To increase the number of instances to train the ML models, a second dataset of an oceanographic campaign of collaborators from U. of Vigo in the same EBCZ, Dimension, was included. Dimension contained 36 samples, increasing the number of samples to 166. As that number of instances was relatively low, a new search for similar 16S rRNA gene datasets on bibliography and public databases was performed to complement the data from Dimension and Envision. However, the search was unsuccessful again. The main reason was the lack of appropriate metadata tables that unequivocally indicated the chlorophyll concentrations and other variables for the sequenced samples. Therefore, the project proceeded only with the data facilitated by collaborators from U. of Vigo. This initial number of samples could be considered low, but it is in the order of samples of similar studies already published that reported excellent results (Li et al. 2022; Ghannam, Techtmann 2021; Marcos-Zambrano et al. 2021).

3.2 CLASSIFICATION OF SAMPLES ACCORDING TO CHLOROPHYLL CONCENTRATIONS

3.2.1 PCAs of selected environmental variables for both campaigns.

To aid oceanographers to classify the samples as belonging to bloom or normal events, several environmental variables related to blooms were explored using PCA. The analyses were done for each campaign because many environmental variables differed between Envision and Dimension datasets. While chlorophyll measurements were present in both datasets, primary production was consistently determined only in Dimension campaign, whereas biomass and cell abundances of different microorganisms were recorded in Envision.

Results (Figure 7, see also appendix 7.1) showed that samples differentiate according to depth. A trend according to the sampling month (or season because February corresponds to Winter, April to Spring and August to Summer) could be observed in Envision along principal component 1. That trend was not obvious in Dimension campaign. This could be related to the sampling frequency in both campaigns. While in Envision samples were taken intensively during one week of each month followed by a long period without samples; in Dimension, ocean water was sampled monthly. Samples grouped according to depth in both campaigns, in this case mostly along PC1 but also along PC2. Other variables such as sampling site (only for Envision, Dimension was sampled only on the coast station), or year did not group the samples.

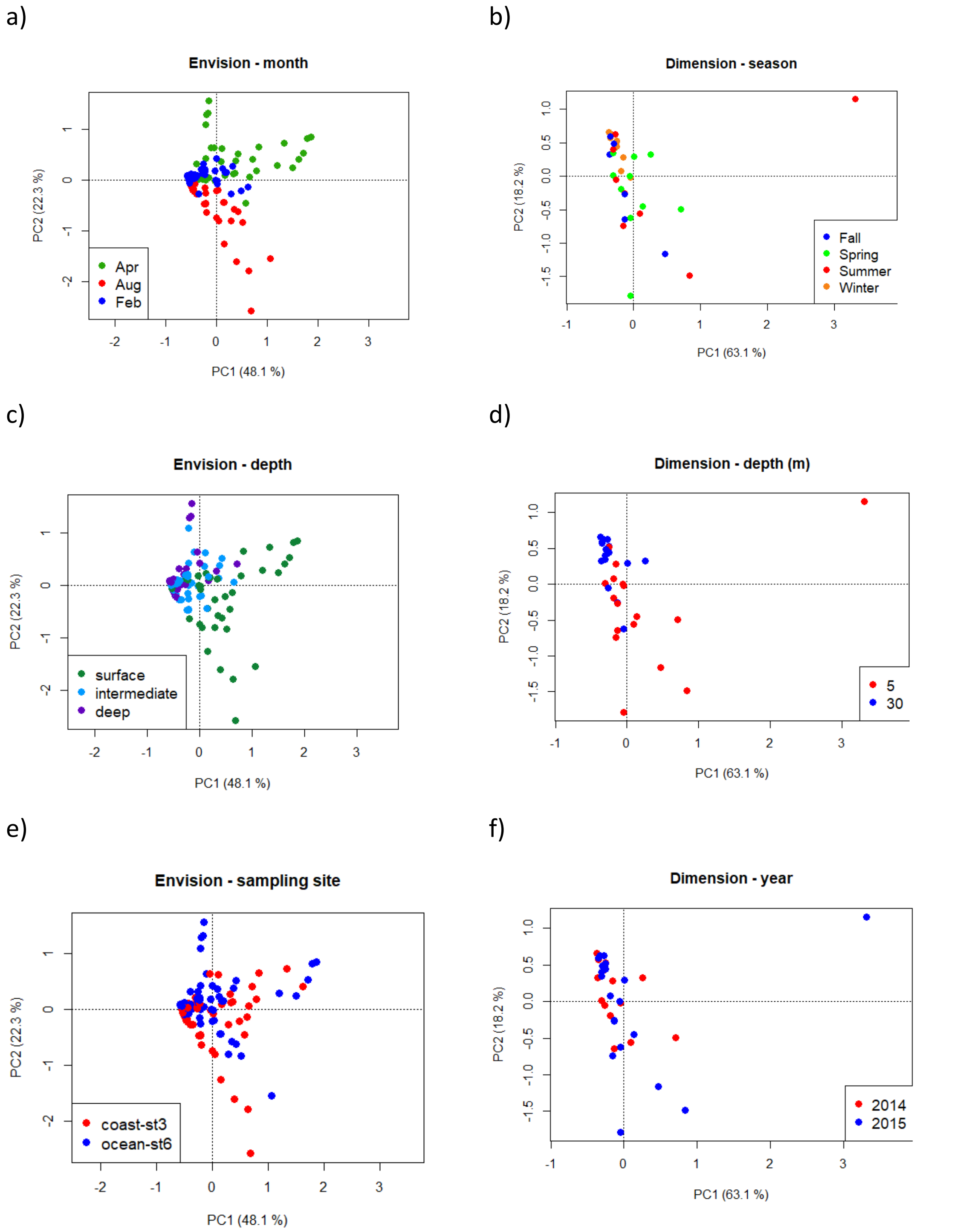


***Figure 7**. PCAs for Envision (left column) and Dimension (right column) calculated with environmental variables. The samples have been coloured according to different qualitative variables indicated on each panel. The percentage of variance explained by each comp component is indicated between parentheses.*

To further explore this data, correlations of variables with principal components were analysed (Table 2). Variables more correlated with both principal components were related with biomass or cell abundances of protists and bacteria in Envision, and primary production in Dimension. In both cases, Chlorophyll concentrations were not so strongly correlated with principal components, although it could be related with the samples order according to depth.

***Table 2**. Correlations of selected environmental variables with PC1 and PC2 axis of PCA analysis for Envision and Dimension campaigns. Results are shown in decreasing order. Variable names in Envision: large protists biomass (BBlp), small protists biomass (BBsp), large protists abundance (AFlp), small protists abundance (AFsp), Synechococcus abundance (AFSyne), Synechococcus biomass (BFSyne), bacterial biomass (BB), bacterial biomass (AB), total chlorophyll a (chla Total), Prochlorococcus abundance (AFProc), Prochlorococcus biomass (BFProc). Dimension –Chlorophyll a (Chla.t), Primary production (PP) of total microeukaryotes(.m), nanoeukaryotes (.n) or picoeukaryotes(.p), community respiration (PP.h.t) and Biomass Prochlorococcus (BP).*

Envision	Variable	Env PC1	Variable	Env PC2
	BBlp	0.393	*AFProc*	0.444
	AFlp	0.385	*AB*	0.429
	BBsp	0.380	*BB*	0.429
	AFsp	0.366	*chlaTotal*	0.425
	AFSyne	0.341	*BFProc*	0.389
	BFSyne	0.330	*BFSyne*	0.276
	BB	0.261	*BBsp*	0.136
	AB	0.253	*AFlp*	0.058
	chlaTotal	0.162	*AFsp*	0.049
	AFProc	0.159	*BBlp*	0.036
	BFProc	0.108	*AFSyne*	0.023
Dimension	**Variable**	**PC1**	**Variable**	**PC2**
	PP.t.h	0.467	*BP*	0.989
	PP.n	0.462	*PP.n*	0.094
	PP.p	0.459	*Chla.t*	0.086
	PP.m	0.439	*PP.t.h.*	0.073
	Chla.t	0.398	*PP.p*	0.064
	BP	0.078	*PP.m*	0.003

In Dimension PCA, a potential outlier could be observed. That sample was taken during September 2015 in the surface, and it corresponds to an event of intense phytoplankton primary production (Figure 8), probably related to a bloom. Because of that, that sample was kept in the analysis.

Despite the observed trends, oceanographers discarded these results to classify samples between bloom and normal conditions.

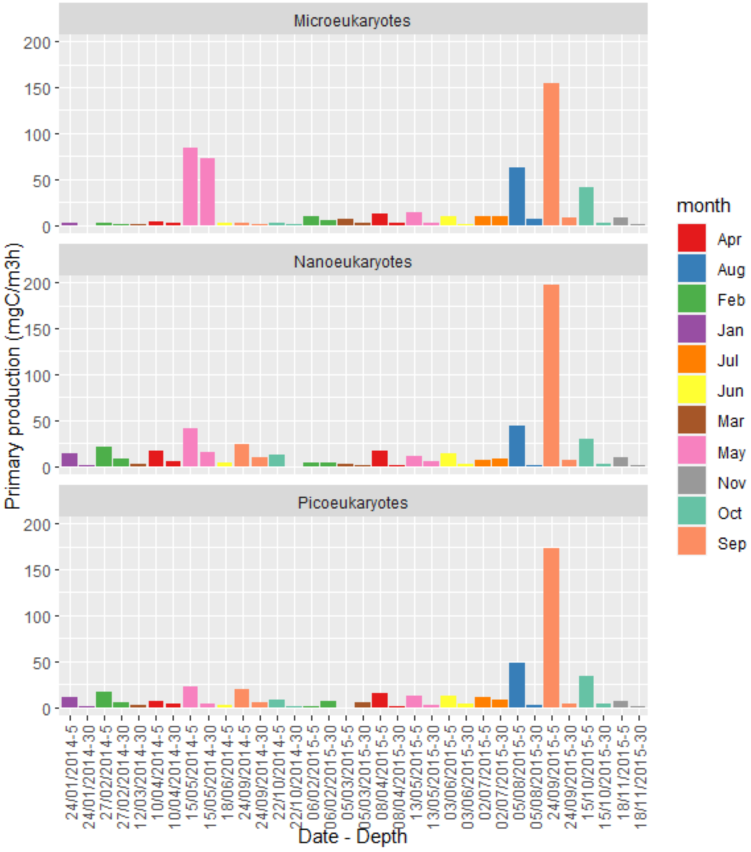


Figure 8. *Primary production for the different size-fraction Eukaryotes in Dimension campaign.*

3.2.2 Distribution of samples between bloom and normal events.

To determine which samples correspond to bloom and normal events, samples were divided into groups according to different percentiles of total chlorophyll a concentrations (table 3). On those divisions, samples below the cutoff were considered as normal situations and samples above the cutoff were labelled as bloom events. Three different percentiles were requested by the oceanographers: 50 and 75 percentiles for the full dataset, and 90 percentile only for the ocean station samples.

Table 3. *Number of samples belonging to bloom and normal events for different chlorophyll percentiles and sampling stations.*

	Percentile 50	**Percentile 75**	**Percentile 90**
Ocean_bloom	35	18	7
Ocean_normal	35	52	63
Coast_bloom	48	24	NA
Coast_normal	48	72	NA

Finally, oceanographers decided to apply the division based on the 75 percentile for the coastal station and the 90 percentile for the oceanographic station. The difference on sample number between the two categories was large, with 31 instances labelled as bloom and 135 as normal. As consequence, the dataset was highly imbalanced with the 78.9 % of the samples belonging to the "normal" class and the other 18.7 % of the samples in the "bloom" class.

3.3 SEQUENCES AND ASVS PREPARATION

The number of original raw sequences was 3,979,906. Of those, the 36.18% was discarded due to low quality issues and the 17.77% were filtered by chimera removal or by removing pairs of reads that were missing a partner. Likewise, the number of ASVs before chimera removal, 15,176, was reduced to 8,184 after chimera removal.

These sequences needed further refinement. 16S rRNA gene primers can amplify other sequences that are not *Bacteria* and *Archaea* such as some eukaryotic sequences as well as the 16S rRNA gene from chloroplasts and mitochondria. The data contained 5 Eukaryotes, 537 chloroplasts and 73 mitochondria. Also, the Envision sample 141, which was present in the original dataset, (Oceanic station, August, day 7, depth 1) was problematic because there were doubts about its DNA quality. To avoid problems, the sample was removed from the dataset. These filtering steps reduced the number of ASVs to 7,569 and the number of sequences to 1,673,164, the 42.04 % of the original sequences.

3.4 ASVS EXPLORATORY ANALYSIS

As part of the feature selection process low abundant ASVs, those with relative abundances below 0.01%, were removed to reduce dimensionality and sparsity (figure 9). This filter reduced the number of ASVs to 1203. Such practice is common when applying ML methods to this kind of data (Marcos-Zambrano et al. 2021; Topçuoğlu et al. 2020). It is also a common practice to reduce noise when classical analyses are used (Poretsky et al. 2014).

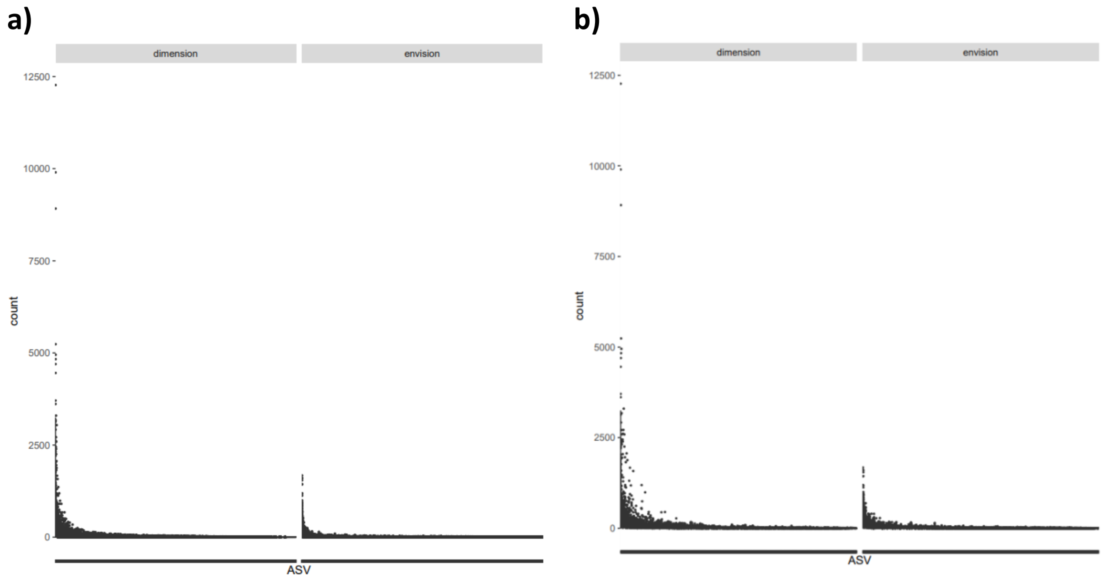


Figure 9: *Boxplots of ASVs counts per sample before (a) and after (b) removing low abundant ASVs.*

The ASVs distributions were highly skewed even after removing low abundant ASV. This observation is typical of microbial diversity studies using molecular tools, where few ASVs or microorganisms are highly abundant accompanied by a large cohort of low abundant organisms (Pedrós-Alió 2012). Therefore, the highly abundant ASVs were not removed as they are important members of the microbial community.

Samples were also explored with NMDS. In this case, two methods were used. First, counts were transformed by the square-root of their proportions, a widely used method on microbial ecology also used in ML approaches (Ghannam, Techtmann 2021; Marcos-Zambrano et al. 2021). Second, as ASVs counts are compositional data, they were transformed using rclr transformation that is more appropriate for this kind of data than other transformations. (figures 10 and 11).

The NMDS using square-root of proportions transformation (figure 10) indicated no batch effects due to the campaigns, year or sampling site. Other spatial-temporal variables (month, season and depth level) indicated slight trends in the data, although samples did not clearly separate according to the different levels of those variables. Because of that, these categorical variables were also considered as features in the ML model. In addition, some separation between bloom and normal samples (event variable) was observed although it was not very clear.

When using the rclr transformation, samples clustered according to campaign (figure 11). In addition to this observation, it is not well known how centered log ratio transformations interact with ML algorithms (Busato et al. 2023). For those reasons, this kind of transformation was discarded for downstream analyses.

This exploratory analysis was also performed using unfiltered ASVs with very similar results for square-root of proportions transformation, what reinforced the idea of removing low abundant ASVs from the analysis (see appendices).

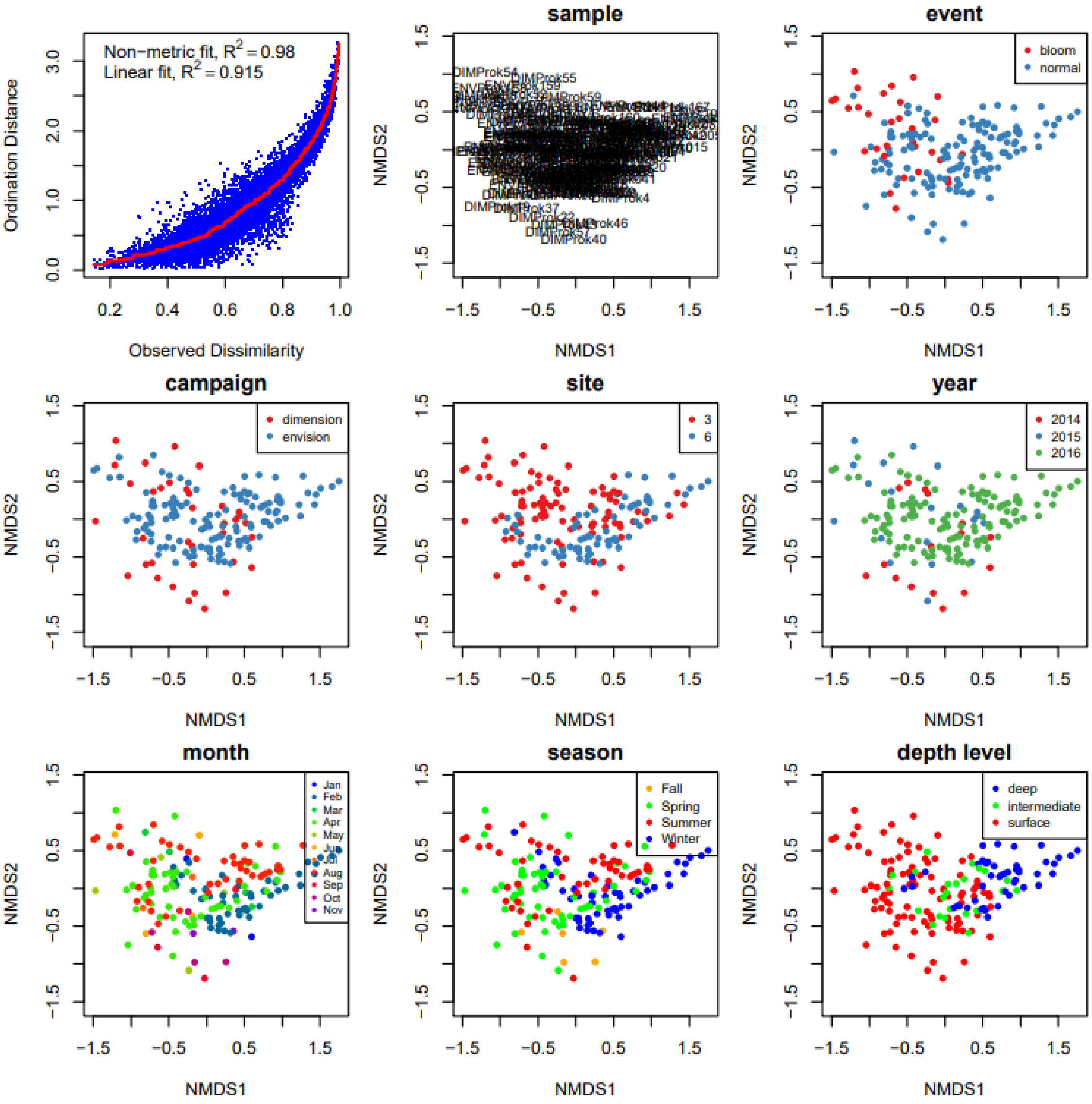


***Figure 10**: NMDS calculated with filtered ASVs with data transformed by the square-root of proportions. First panel contains the stress plot. The other panels contain the NMDS plot in which samples are labelled with their names or coloured by the categorical variables.*

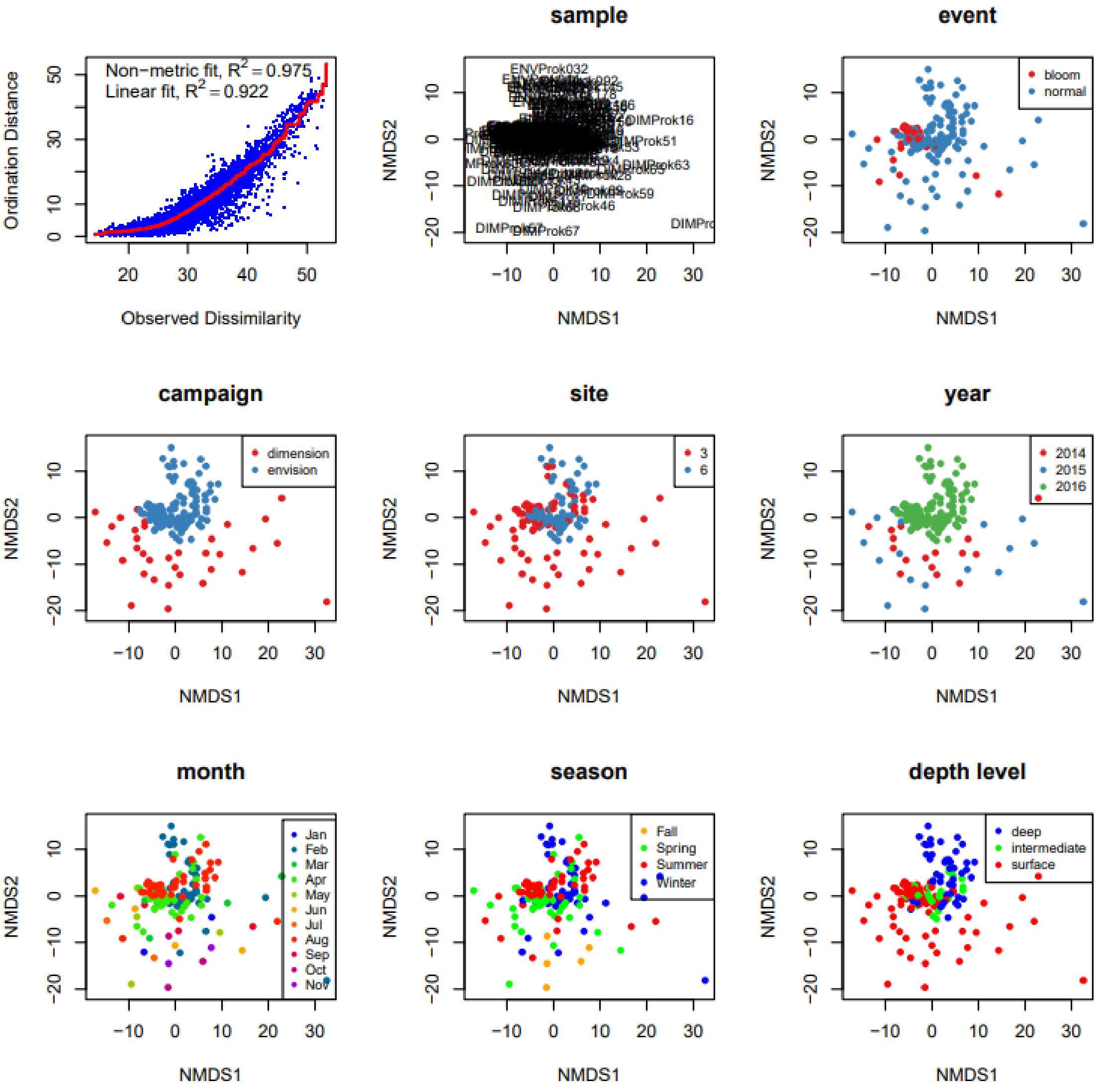


Figure 11. *NMDS using filtered ASVs transformed with rclr. First panel contains the stress plot. The other panels contain the NMDS plot in which samples are labelled with their names or coloured by the categorical variables.*

3.5 CLUSTERS EXPLORATORY ANALYSIS

To further reduce the dimensionality of the data, ASVs were joined into clusters according to their taxonomic annotation to genus level (Janßen et al. 2019). In total, 688 clusters were obtained with a highly skewed distribution as expected (Figure 12).

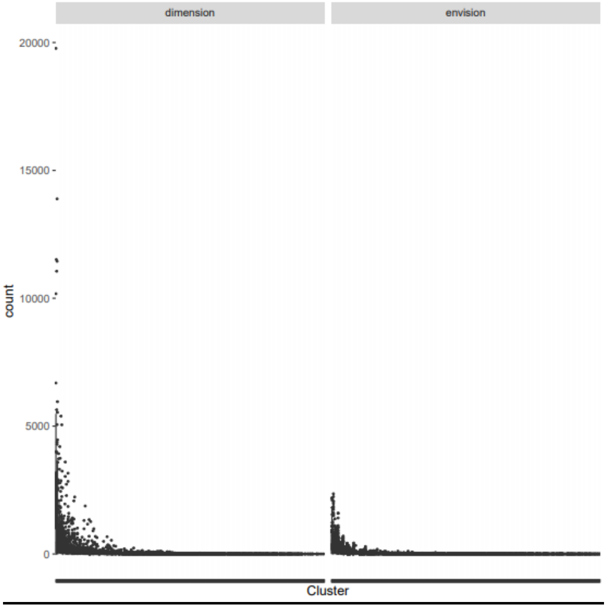


***Figure 12**. Boxplots of clusters counts per sample.*

As with ASVs, samples were analysed with NMDS but in this case using clusters applying both kind of transformations (Figures 13 and 14). Results found were like those obtained with ASVs. NMDS with root-square of proportions transformations showed slight trends in samples according to season, month, and depth level. Whereas NMDS with rclr transformed data showed undesirable grouping of samples for campaigns.

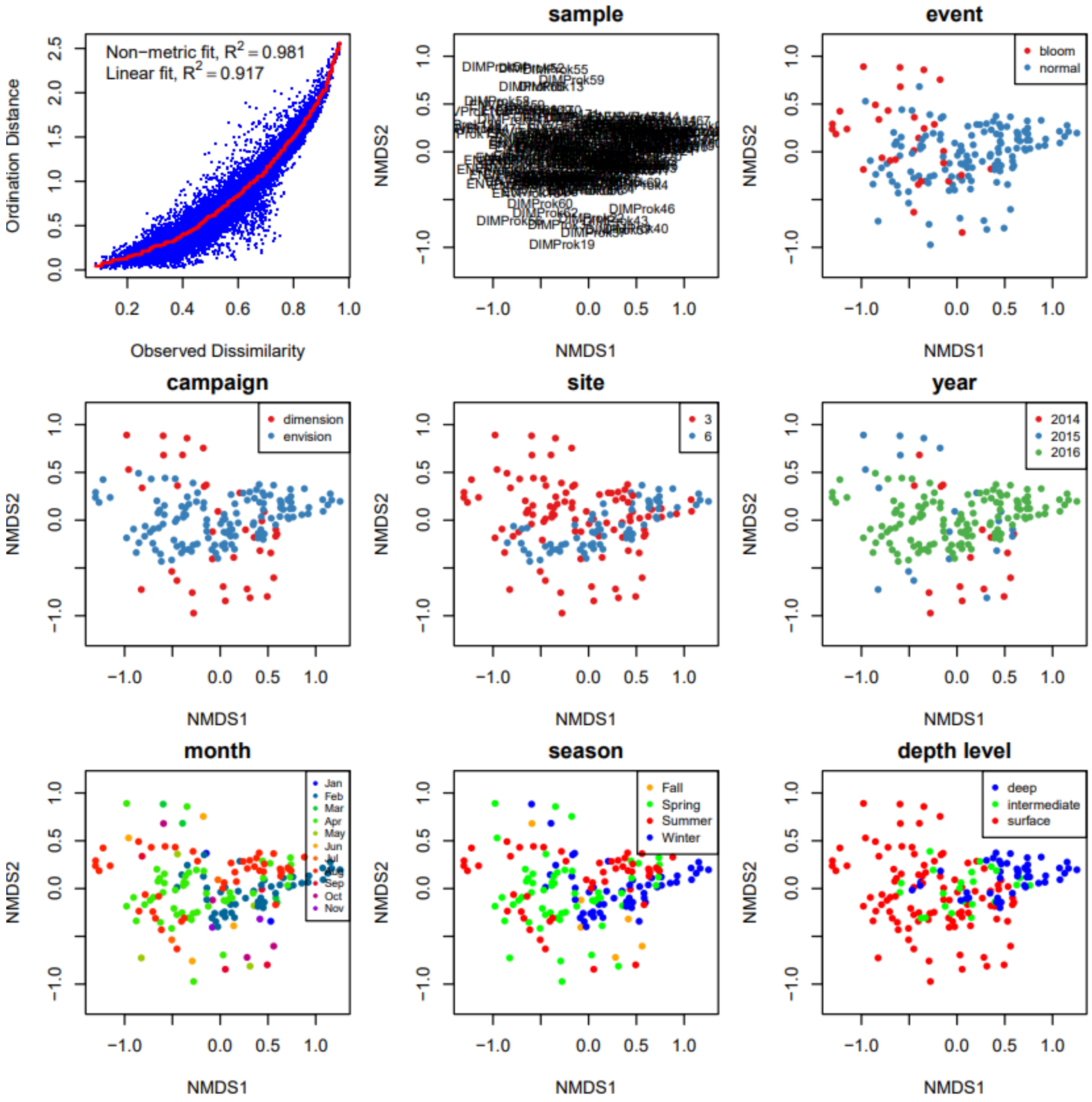


Figure 13. *NMDS calculated with clusters count data transformed by the square-root of proportions. First panel contains the stress plot. The other panels contain the NMDS plot in which samples are labelled with their names or coloured by the categorical variables.*

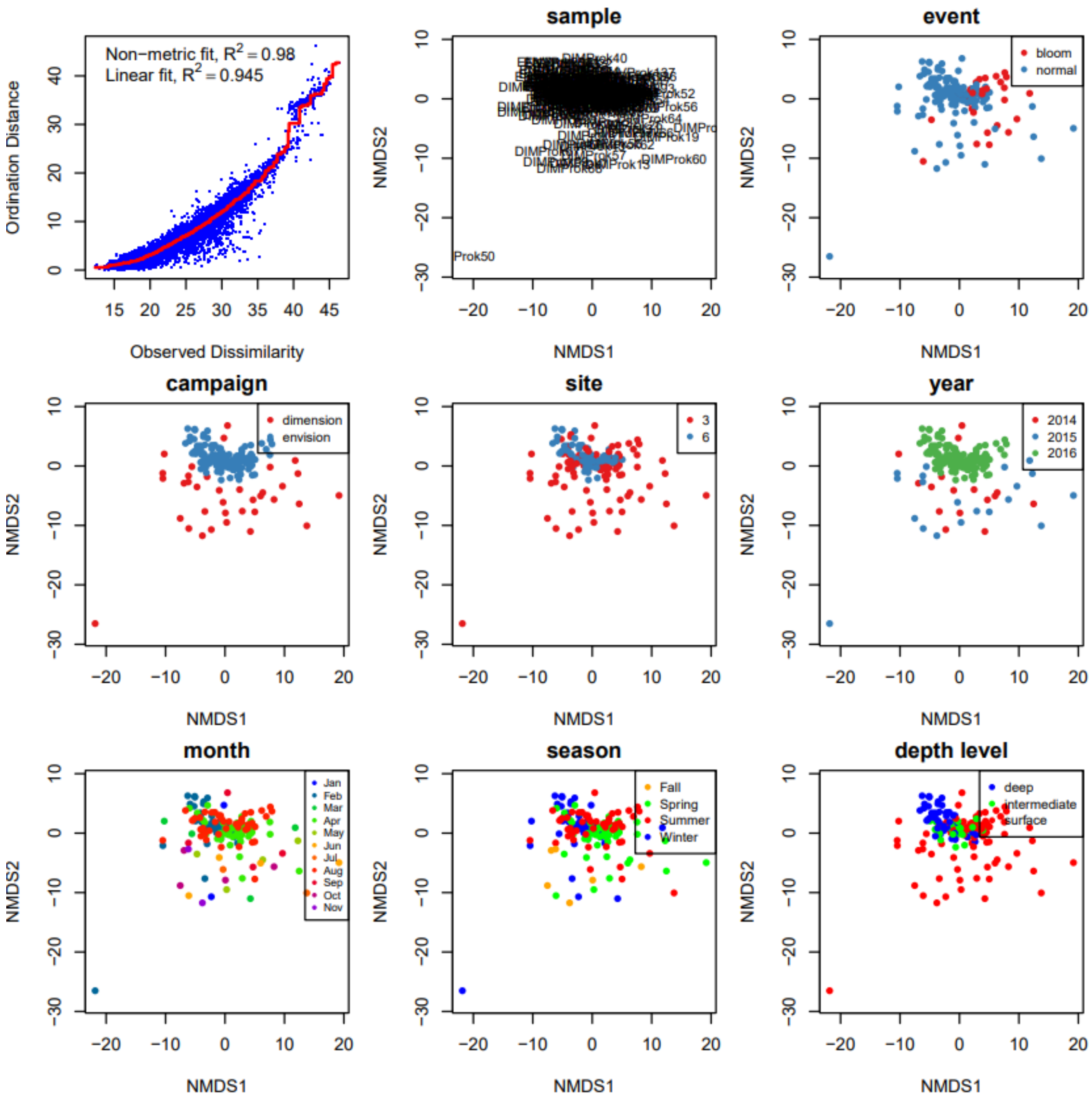


Figure 14. *NMDS using clusters count data transformed with rclr. First panel contains the stress plot. The other panels contain the NMDS plot in which samples are labelled with their names or coloured by the categorical variables.*

3.6 RANDOM FOREST MODELS TRAINING AND VALIDATION

3.6.1 Feature preprocess and selection

Features used to train the RF models consisted of the biological variables (either ASVs or Clusters) and other variables related to the environment of sampling characteristics, many of which were categorical: depth and depth level, year, month, sampling station and, campaign.

Features used in model training and evaluation were selected using several criteria. A criterion to select features has been already mentioned. In the case of ASVs, those with relative abundances lower than 0.01% were removed from the data in one of the cases considered. In addition, perfectly correlated features were grouped, keeping only one member on the dataset. On all analysed datasets, the campaign Envision and year_2016 were the only correlated features found. Next, features with variances near zero were filtered. 394 and 184 features were removed from clusters and filtered ASVs datasets respectively whereas, 3515 features were removed in the unfiltered ASVs dataset.

A prevalence filter was also explored. This approach consisted of removing features that appeared only in a low number of samples. It is a filter usually applied in microbial ecology studies to remove transient and not relevant species. To define the threshold of low prevalence, different percentage of samples were explored. Initially, low percentages were considered (1% ~ 2 samples, 2% ~ 3 samples and, 3% ~ 5 samples), but no features were removed in any case. Then, higher percentages were tested (5% ~ 8 samples, 10% ~17 samples and, 20% ~ 33 samples) to ensure the filter worked. Only the percentages above 10% removed some features. However, using those cutoffs could not be considered as a low prevalence filter. Therefore, no features were removed from the datasets due to this criterion.

The final size of the datasets analysed were 166 instances of:

- 318 features for the clusters dataset.
- 1033 features for the filtered ASVs dataset.
- 4078 features for the unfiltered ASVs dataset.

Features were preprocess prior to model training. Biological features were transformed applying the square root of their proportions as already described in section 3.4 and 3.5.

The data distributions with centered and scaled features are shown in figure 15. (See appendices for unfiltered ASVs results).

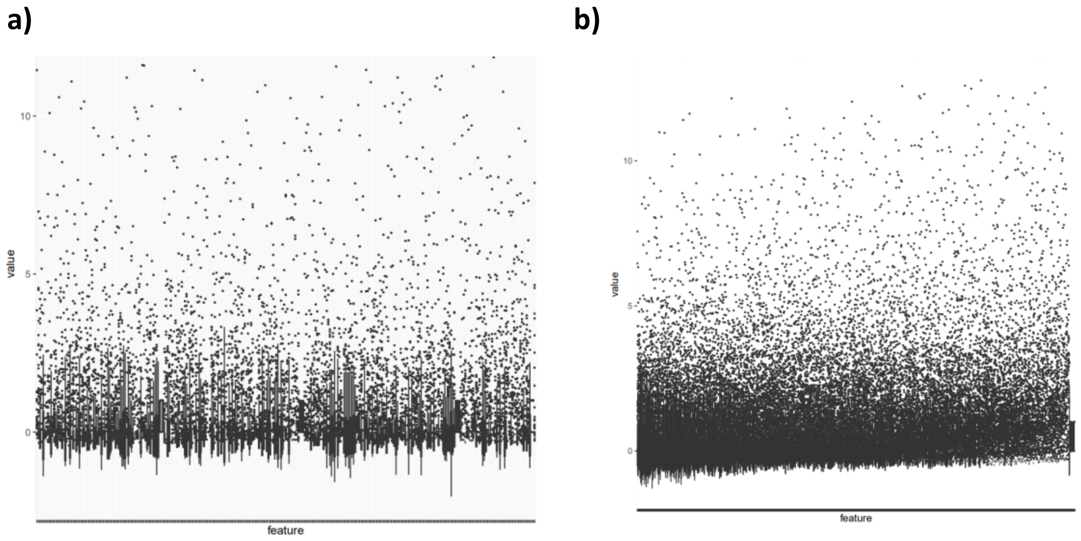


Figure 15. *Boxplots of feature values from the clusters (a) or filtered ASVs (b) datasets after preprocessing and selection.*

3.6.2 Random forest models training

The instances in the dataset were differentially distributed between the two categories of the outcome variable (event). Only 31 instances corresponded to the category of interest (bloom) and 135 to the normal category.

The dataset was split considering the unbalanced character of the dataset. The train group included the 80 % of the instances and the test group contained the other 20% of the samples. This division resulted in:

- Train group: 25 bloom and 108 normal instances.
- Test group: 6 bloom and 27 normal instances.

3.6.2.1 Hyperparameter tunning.

The default grid search for *mtry* hyperparameter was used in the initial test to train the RF models (Figure 6, in methods). The values validated with 5 and 10 k-fold RCV were:

- Clusters datasets: 9, 18, 36.
- Filtered ASVs dataset: 16, 32, 64.
- Unfiltered ASVs datasets: 32, 64, 128.

After evaluating the mean AUC values (Figure 16) obtained during the k-fold repeated validation, the range of *mtry* values was extended:

- Clusters datasets: 3,6,9,18,24,36.
- Filtered ASVs dataset: 8,16,32,48,64.
- Unfiltered ASVs datasets: 16,32,48,64,128.

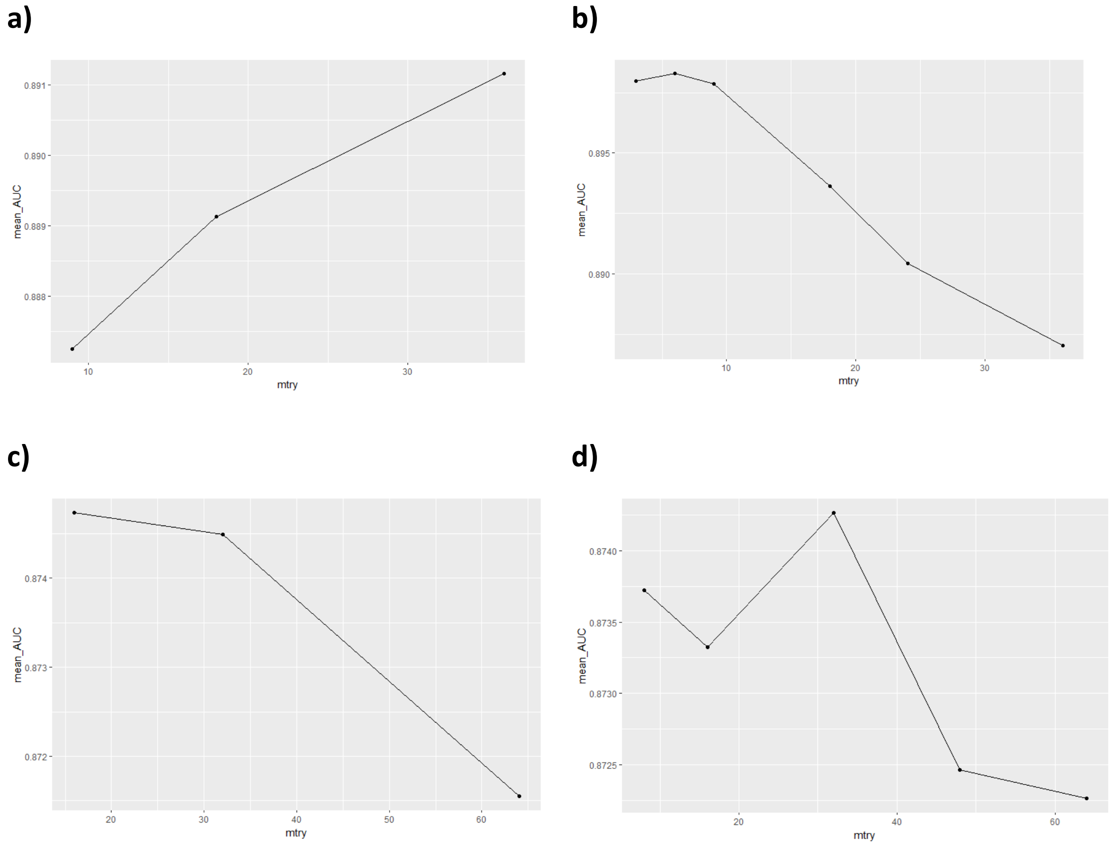


Figure 16. *Mean AUC for the mtry values in the initial RF initial train using the default (a, c) and the extended (b, d) grid searches for the clusters (a,b) and filtered ASVs (c, d) datasets. Values from the 5-fold RCV of the initial single run.*

3.6.2.2 Random Forest performance results.

After extending the grid for *mtry*, 100 models were trained and tuned for each dataset. For that, the dataset split was randomly repeated 100 times considering the distribution between both categories. This strategy was applied to obtain a robust interpretation of model performance (Topçuoğlu et al. 2020) given the low number of instances. This approach was applied to the clusters, filtered ASVs and unfiltered ASVs. This last case was run for comparison expecting lower performance than the models in which data dimensionality and sparsity was reduced.

The validation results are shown in figure 17. Overall, the best *mtry* value was 6 for the clusters, 16 for filtered ASVs, and 128 for unfiltered ASVs. In the case of the unfiltered

ASVs, the maximum value of AUC was located on one of the range sides, indicating that probably the best hyperparameter solution for those models was not found. To solve that issue, the range of *mtry* should have been incremented again. However, given the computational cost of running those models and that these cases were not of interest, the model tunning was not repeated for the unfiltered ASVs.

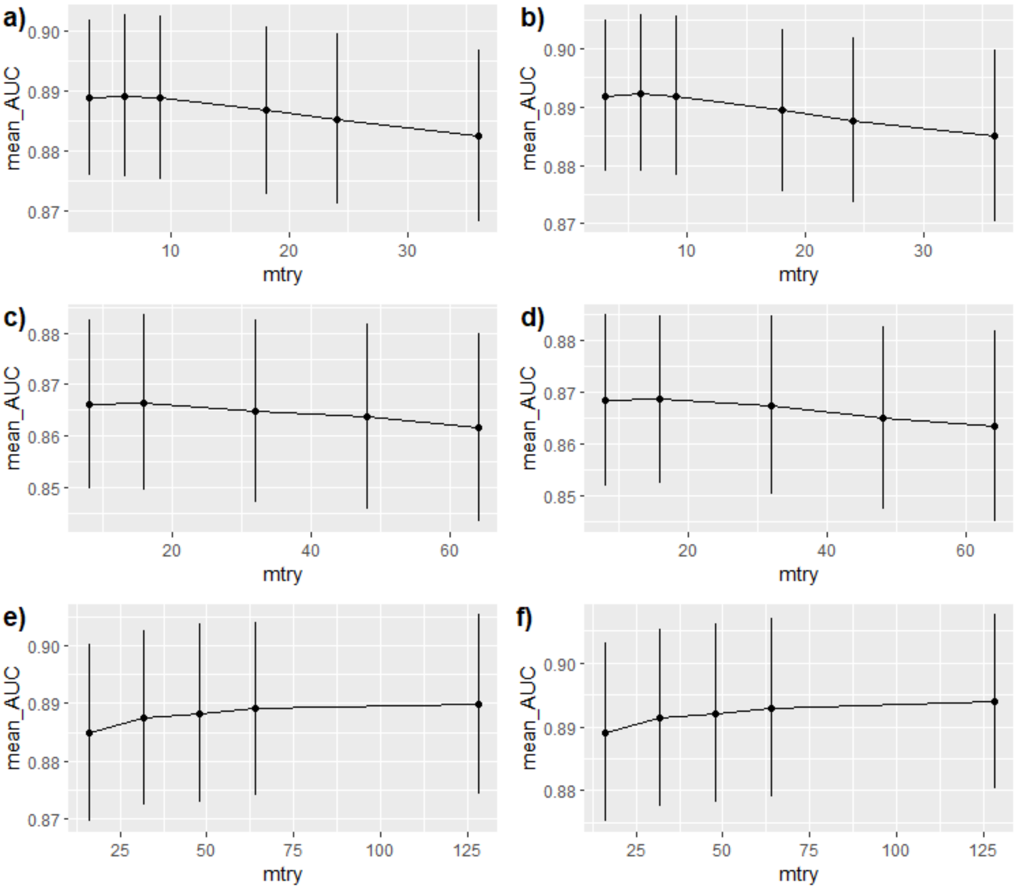


***Figure 17**. Mean AUC for the different mtry values explored. a) clusters with 5 k-fold RCV, b) clusters with 10 k-fold RCV, c) filtered ASVs with 5 k-fold RCV, d) filterd ASVs with 10 k-fold RCV, e) unfiltered ASVs with 5 k-fold RCV, f) unfiltered ASVs with 10 k-fold RCV. Error bars indicate one standard deviation.*

The performance results were very similar for all cases (figure 18). The similar values obtained for AUC during training (variable cv_metric_AUC, figure 18) and during model performance evaluation with the test data (variable AUC, figure 18) indicated that the models were not suffering from overfitting. Mean Accuracy and AUC were higher than 0.85 (Table 4). However, Kappa metric, that considers the possibility of a correct prediction by chance alone was quite poor with averaged values below 0.40 in all cases.

Table 4. Averaged values of different performance metrics for RF models

Model*	AUC	Accuracy	Kappa	Sensitivity	Specificity
cl-k5	0.899	0.856	0.371	0.917	0.387
cl-k10	0.898	0.855	0.364	0.924	0.374
ASVs F k5	0.885	0.854	0.366	0.894	0.400
ASVs F k10	0.883	0.850	0.346	0.890	0.390
ASVs Unf k5	0.898	0.8545	0.365	0.920	0.379
ASVs Unf k10	0.901	0.859	0.381	0.925	0.384

**) cl: Clusters, ASVs F: filtered ASVs, ASVs Unf: unfiltered ASVs, k5: 5 k-fold RCV, k10: 10 k-fold RCV.*

The results of Specificity and Sensitivity explained the problem. In this evaluation, the category “normal” was considered the positive class while “bloom” was the negative. Sensitivity, a metric that measures the true positive rate, reached high averaged values (> 0.88) showing that the models correctly classified most of the “normal” instances. In contrast, the low averaged values for Specificity (< 0.41), which measures the percentage of negative samples correctly classified, indicated that the models could not classify the “bloom” samples.

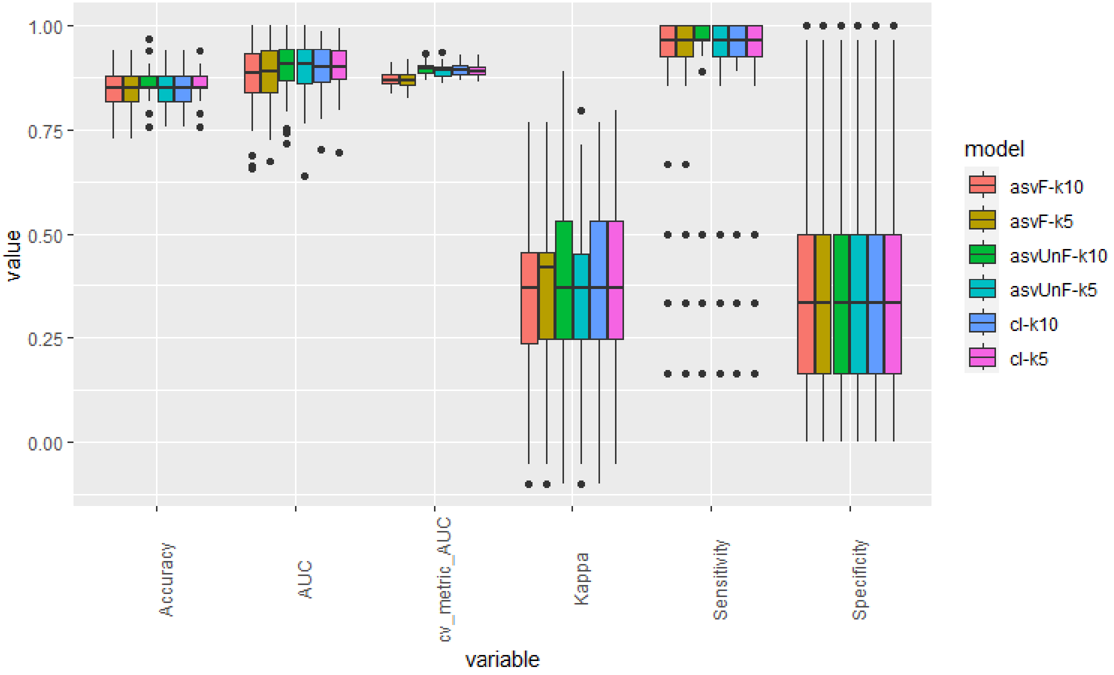


***Figure 18**. Performance for RF models. All metrics were calculated with the test group except "cv_metric_AUC”, that indicates the results of the 5 or 10 k-fold repeated cross-validation. Datasets: asvF - filtered ASVs, asvUnF – unfiltered ASVs, cl: Clusters.*

Given the highly unbalanced nature of the datasets (normal: 78.9 %, bloom: 18.7% the of instances), the high values achieved by Accuracy and AUC could be explained by correct classification of the instances by chance. As Kappa considers that possibility, it showed lower values.

3.7 Attempts to improve model performance

3.7.1 SVM model

To test the choice of ML algorithm, a SVM model with a Radial Basis Function kernel was also trained using the filtered ASVs dataset and the hyperparameters grid-search defined by default with a 5 k-fold RCV (see methods 2.8.1). The best hyperparameters were *C*=1 and *sigma* = 1x10^{-6}, that was the lower value explored in the grid-search for this hyperparameter (Figure 19). The search could have been extended to include lower values of sigma. However, it was not repeated since the AUC value was reaching a plateau in the lower values of *sigma*, therefore the expected benefits of repeating the process were smaller than its computational cost.

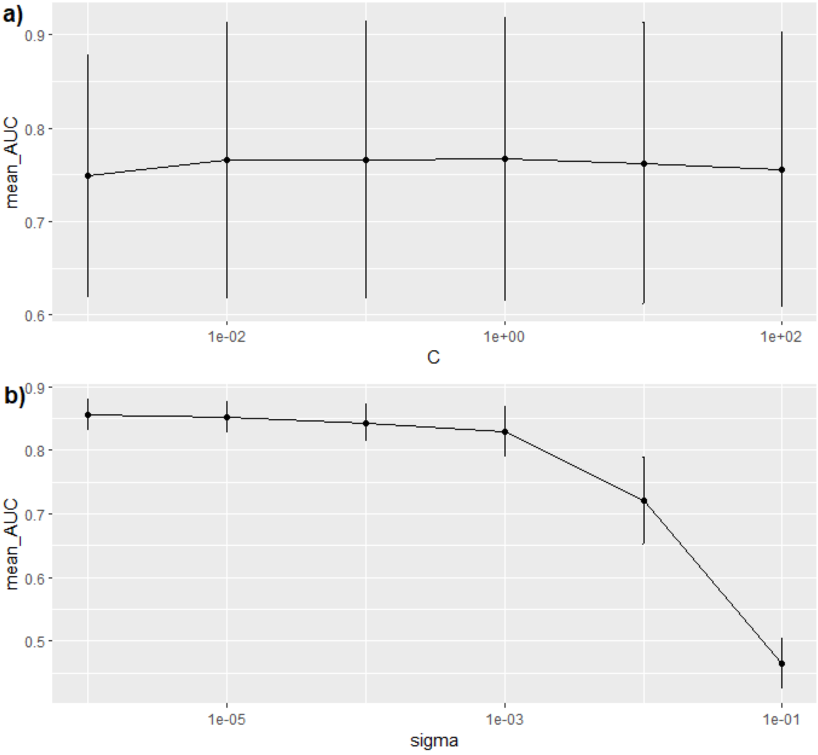


Figure 19. *Mean AUC for the different values for hyperparameters C (a) and sigma (b) on the SVM models.*

In any case, the SVM model did not improve the performance obtained with the RF models (Figure 20). On the contrary, the averaged performance metrics were worse than in the RF models (AUC: 0.86, Kappa 0.24, Sensitivity: 0.90, Specificity: 0.32). These results were not surprising as it has been reported that SVM models perform worse than RF models for this kind of data (Busato et al. 2023; Ghannam, Techtmann 2021; Marcos-Zambrano et al. 2021).

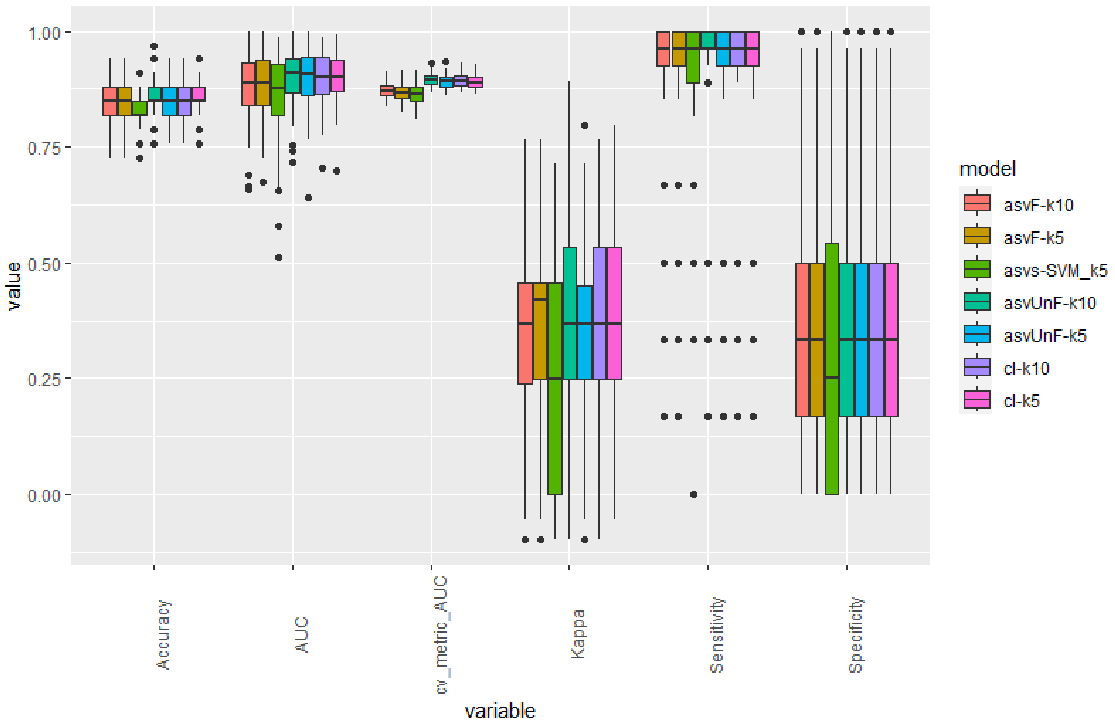


Figure 20. *Performance of SVM and RF models. All metrics were calculated with the test group except "cv_metric_AUC", that indicates the results of the 5 or 10 k-fold rcv. The SVM model is: asvs-SVM_k5. The rest are the RF models. Datasets: asvF - filtered ASVs, asvUnF – unfiltered ASVs, cl: Clusters.*

3.7.2 Synthetic data

When dealing with imbalanced datasets, ML methods tend to overfit majority classes (He, Garcia 2009). That seems to be the case of the previous results. A common way to increase the performance of models dealing with imbalanced datasets is the synthetic data generation. Among the different methods available, the SMOTH algorithm is commonly applied (Kovács 2019).

This methodology was tested using the filtered ASVs dataset. The number of instances in the “bloom” class was incremented through SMOTE method to level the instances included in the “normal” class. This approach was applied only to the train data, not to the test data. The process was repeated for the 100 different datasets splits between the train and test groups used to train a set of RF models as previously done.

The performance of the models increased considerably using the synthetic dataset (figure 21) with values for Kappa, Sensitivity and Specificity over 0.8 in most cases. Please, notice that in this case, “bloom” was considered the positive class during the calculation of the performance metrics. Even though that the capacity of these models to detect the "bloom" samples was still worse than for the "normal" samples, these results indicated that the classification performance was good.

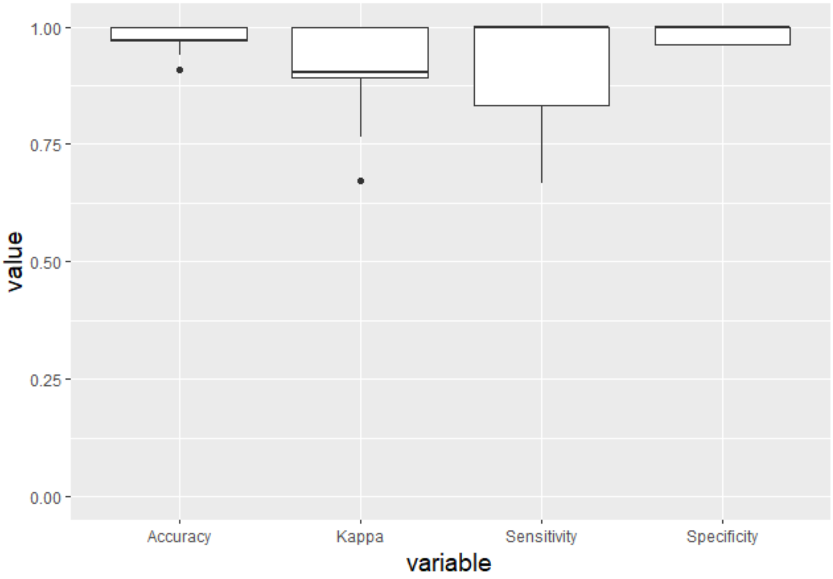


Figure 21. *Performance of the RF models that used the synthetic dataset.*

3.8 Determining biomarkers

As the performance reached with the RF models using synthetic datasets were good, these results were used to search for biomarkers. The feature importance was scaled from 0 to 100. Among them, the features with median values over 20 were selected as biomarkers (Figure 22). This cut-off was selected by convenience to focus the results on the features with the larger influence. Features with median values over 10 are shown in the Appendix 7.5.

Considering a stringent biomarker definition, environmental variables should not be considered so. Among the 37 important features, only one was an environmental variable, depth. This could be expected as blooms occurs closer to the surface. The other 36 features were ASVs that could be considered as biomarkers. Most biomarkers were bacteria and only one archaea was found (table 5). Bacteria belonged mostly to *Proteobacteria* and *Bacteroidota* phyla. The five biomarkers with the largest importance included quite diverse microorganisms belonged to four different families (*Pseudohongiella, Flavobacteriaceae, Rhodobacteriaceae and Alteromonas*) and included some genera known for being associated with microalgal blooms like the genus *Pseudohongiella* and *Polaribacter.* The analysis of the potential role of these biomarkers is beyond the goals of this project.

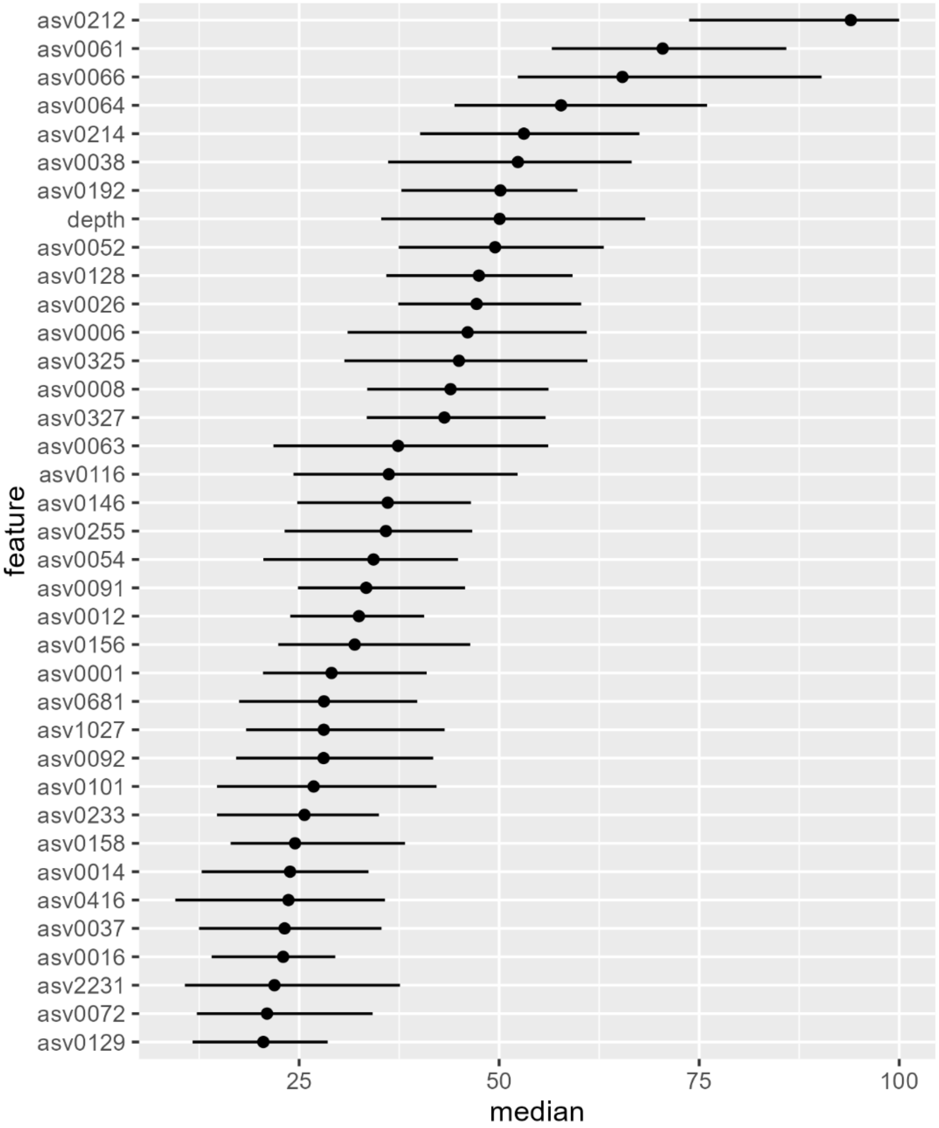


Figure 22. *Biomarkers. Features with median scaled importance value across all 100 RF models over 20. Bars indicate the 25 and 75 percentiles.*

Table 5*. Taxonomy of ASVs selected as biomarkers. No species could be determined except for asv0008.*

ASV	Kingdom	Phylum	Class	Order	Family	Genus
asv0001	*Bacteria*	*Proteobacteria*	*Alphaproteobacteria*	*Rhodobacterales*	*Rhodobacteraceae*	*Amylibacter*
asv0006	*Bacteria*	*Bacteroidota*	*Bacteroidia*	*Flavobacteriales*	*Flavobacteriaceae*	*Aurantivirga*
asv0008	*Bacteria*	*Proteobacteria*	*Alphaproteobacteria*	*Rhodobacterales*	*Rhodobacteraceae*	*Planktomarina**
asv0012	*Bacteria*	*Proteobacteria*	*Alphaproteobacteria*	*Rhodobacterales*	*Rhodobacteraceae*	*Ascidiaceihabitans*
asv0014	*Archaea*	*Thermoplasmatota*	*Thermoplasmata*	Marine Group II	NA	NA
asv0016	*Bacteria*	*Bacteroidota*	*Bacteroidia*	*Flavobacteriales*	NS9 marine group	NA
asv0026	*Bacteria*	*Bacteroidota*	*Bacteroidia*	*Flavobacteriales*	*Cryomorphaceae*	NA
asv0037	*Bacteria*	*Bacteroidota*	*Bacteroidia*	*Flavobacteriales*	*Flavobacteriaceae*	NA
asv0038	*Bacteria*	*Bacteroidota*	*Bacteroidia*	*Flavobacteriales*	*Flavobacteriaceae*	*Polaribacter*
asv0052	*Bacteria*	*Bacteroidota*	*Bacteroidia*	*Flavobacteriales*	*Cryomorphaceae*	NA
asv0054	*Bacteria*	*Proteobacteria*	*Gammaproteobacteria*	*Pseudomonadales*	*Porticoccaceae*	SAR92 clade
asv0061	*Bacteria*	*Bacteroidota*	*Bacteroidia*	*Flavobacteriales*	*Flavobacteriaceae*	*Polaribacter*
asv0063	*Bacteria*	*Marinimicrobia (SAR406 clade)*	NA	NA	NA	NA
asv0064	*Bacteria*	*Proteobacteria*	*Gammaproteobacteria*	*Enterobacterales*	*Alteromonadaceae*	*Glaciecola*
asv0066	*Bacteria*	*Proteobacteria*	*Alphaproteobacteria*	*Rhodobacterales*	*Rhodobacteraceae*	*Yoonia-Loktanella*
asv0072	*Bacteria*	*Proteobacteria*	*Gammaproteobacteria*	*Pseudomonadales*	*Halieaceae*	OM60(NOR5) clade
asv0091	*Bacteria*	*Bacteroidota*	*Bacteroidia*	*Flavobacteriales*	*Flavobacteriaceae*	*Winogradskyella*
asv0092	*Bacteria*	*Proteobacteria*	*Gammaproteobacteria*	*Pseudomonadales*	*Halieaceae*	*Luminiphilus*
asv0101	*Bacteria*	*Proteobacteria*	*Gammaproteobacteria*	*Thiotrichales*	*Thiotrichaceae*	NA
asv0116	*Bacteria*	*Bacteroidota*	*Bacteroidia*	*Flavobacteriales*	NS9 marine group	NA
asv0128	*Bacteria*	*Bacteroidota*	*Bacteroidia*	*Flavobacteriales*	*Flavobacteriaceae*	NS5 marine group
asv0129	*Bacteria*	*Bacteroidota*	*Bacteroidia*	*Flavobacteriales*	*Flavobacteriaceae*	NS5 marine group
asv0146	*Bacteria*	*Bacteroidota*	*Bacteroidia*	*Flavobacteriales*	*Flavobacteriaceae*	NS5 marine group
asv0156	*Bacteria*	*Bacteroidota*	*Bacteroidia*	*Flavobacteriales*	*Cryomorphaceae*	NA
asv0158	*Bacteria*	*Proteobacteria*	*Alphaproteobacteria*	*Thalassobaculales*	*Nisaeaceae*	OM75 clade
asv0192	*Bacteria*	*Proteobacteria*	*Gammaproteobacteria*	*Pseudomonadales*	*Pseudohongiellaceae*	*Pseudohongiella*
asv0212	*Bacteria*	*Proteobacteria*	*Gammaproteobacteria*	*Pseudomonadales*	*Pseudohongiellaceae*	*Pseudohongiella*
asv0214	*Bacteria*	*Bacteroidota*	*Bacteroidia*	*Flavobacteriales*	*Flavobacteriaceae*	NS3a marine group
asv0233	*Bacteria*	*Proteobacteria*	*Alphaproteobacteria*	SAR11 clade	Clade I	NA
asv0255	*Bacteria*	*Proteobacteria*	*Gammaproteobacteria*	*Pseudomonadales*	*Halieaceae*	OM60(NOR5) clade
asv0325	*Bacteria*	*Bacteroidota*	*Bacteroidia*	*Flavobacteriales*	*Flavobacteriaceae*	NS3a marine group
asv0327	*Bacteria*	*Proteobacteria*	*Alphaproteobacteria*	SAR11 clade	Clade I	Clade Ib
asv0416	*Bacteria*	*Bacteroidota*	*Bacteroidia*	*Flavobacteriales*	*Flavobacteriaceae*	*Polaribacter*
asv0681	*Bacteria*	*Bacteroidota*	*Bacteroidia*	*Flavobacteriales*	*Flavobacteriaceae*	*Formosa*
asv1027	*Bacteria*	*Bacteroidota*	*Bacteroidia*	*Flavobacteriales*	*Flavobacteriaceae*	*Polaribacter*
asv2231	*Bacteria*	*Bacteroidota*	*Bacteroidia*	*Flavobacteriales*	*Flavobacteriaceae*	*Aurantivirga*

* *Planktomarina temperata*

4 FUTURE WORK AND CONCLUSIONS

4.1 CONCLUSIONS

The use of ML methods in the study of microbial communities is largely limited by the number of samples and this project is a good example of it. It is a particular problem for the study of microbiomes not related to human health where the number of samples is even lower as sampling that requires fieldwork is usually very costly, and budgets are usually tight. To overcome those problems, scientific community should make a community effort. Most scientific endeavours directed to environmental microbiomes include the deposition of sequences in public databases. But that is not often the case for other kind of data within the same projects. The lack of associated metadata hiders the use of the large amounts of available environmental sequences. This issue has been arisen before, and methods to carefully annotate sequences with their corresponding metadata were proposed more than a decade ago (Yilmaz et al. 2011). Still, it is rare to find an environmental dataset compiling the specifications to make those sequences useful for other researchers, especially in ecology where environmental data is essential.

Despite of that, the available ML tools have allowed to overcome these difficulties. Even though the number of samples included in the study was very low, and the dataset was quite imbalanced, a classification model with high performance was obtained demonstrating, one more time, the applicability of ML methods to standard microbial ecology studies. The future analysis of the biomarkers found and its comparison with the results obtained with classical statistical tools would determine the advantage of this kind of approaches.

This work has been an example of the importance of developing a good plan of data management in the first steps of a research endeavour. This project inherited the data from two oceanographic campaigns. Although some initial data management was expected to integrate both projects, the time and effort required to get and organize the datasets, even within the same campaign, was totally unexpected. The lack of unique and common systems to name de samples, between and within datasets, and the difficulties to locate the datasets themselves delayed this project from its very beginning. That negatively affected the development of the project, making difficult to reach the proposed objectives.

Achieving the proposed goal was expected because few similar studies searching biomarkers on microbial communities with ML methods and low number of samples have been published (see section 1.1). However, the initial problems with the datasets negatively impacted the temporal plan that had to be adjusted through the development of the TFM. Also, the problem of imbalanced datasets was not anticipated.

Therefore, the methodology had to be adjusted as well to introduce the SMOTE method. Fortunately, the mitigation actions worked out and the project ended on time.

4.2 Future Work

Even though the all the planned goals were achieved, the are some areas that should be explored in more detail.

First, other methods to classify samples between normal or bloom levels should be analysed. For instance, other environmental variables such as primary productivity or microalgal biomass could be used instead of or in combination with chlorophyll to determine when a bloom is happening. Coastal blooms are complex events controlled by many factors and not well understood (Deng, Vallet, Pohnert 2022), and only using one variable for its determination can be limiting. In the data used in the project, those variables contained many missing values or were not measured at all in the coastal station. Therefore, finding other useful datasets would be crucial to do not reduce the number of samples even further if this approach is explored. Another possibility is the use of unsupervised machine learning techniques to group samples according to environmental variables. The analysis should be extended beyond the one already developed as the variables considered on it were limited.

Due to temporal constraints, the SMOTE method was applied to filtered ASVs only. It would be interesting to apply the same analysis to the other two datasets, clusters and unfiltered ASVs, and compare the performance results and biomarkers found.

As all classifiers failed until synthetic data was generated, a regression approach should also be explored. On it, the problem of an imbalanced dataset will not be present. This was the original idea when the project was planned. In the light of the problems faced, it would be worth to explore the use of generalize linear models or regression random forests. The linear models offer an extra advantage over random forest, which is of particular interest to interpret the biomarkers. In that case, the strength, and the kind of relationship (positive or negative) between the biomarker and the variable or group of variables to predict can be inferred from the biomarker´s weight in the model. That would facilitate finding the role of those biomarkers in the environment.

Another method to determine feature importance in the classifiers should be explored as well. In particular, the method based on the total decrease in node impurity or Gini index. Then, the biomarkers found should be compared with the current results to check their consistency.

In any case, the biomarkers found need to be analysed in detail to try to understand their relationship with the blooms. Although these microorganisms are the most important for the classifier, it is still unknown if they are more relevant during the blooms or during the normal situation of the water column. For that, their spatio-temporal trends need to be studied. Also, the results obtained with RF models should be compared with the results obtained with standard techniques such as differential analysis. This comparison would allow to determine if ML approaches are more capable than classical tools to find complex relationships on these datasets.

Finally, it would be quite interesting to use this approach to metagenomic data instead of marker genes. That would allow us to find biomarkers related to metabolic and other cellular processes linked to phytoplankton blooms.

4.3 Impacts on Sustainability, Ethical Behaviour, Social Responsibility and Diversity.

The foreseen positive impacts on sustainability and social responsibility are expected to occur in the future. And these impacts will take place only if the knowledge generated serves to advance in the management of coastal areas and fisheries and social stakeholders develop policies to implement changes. Therefore, achieving those impacts is far beyond the development of this project.

The impact of intense computing in the greenhouse gas emissions could not be mitigated during the development of this project because Random Forest algorithms stand out in performance compared to other less intensive classifiers. Planned future work includes the exploration of other less intensive methods that, if their performance is good enough, might be used for future biomarker search on microbial communities.

Ethical behaviour and diversity impacts were faced during the progress of the TFM and mitigated with the methods already mentioned in section 1.3. In particular, the use of a citation style that includes the full name of authors in the bibliography has been applied. No other issues have been detected.

5 GLOSSARY

ASV: Amplicon Sequence Variant.

AUC: Area Under the Curve.

Bacterioplankton: Prokaryotic microorganisms living in the water column.

Bloom: an event of rapid phytoplankton proliferation in aquatic ecosystems that results in dense assemblages.

EBCZ: Easter Boundary Coastal Zones.

k-fold RCV: k-fold Repeated Cross Validation.

Microbiome: a microbial community.

ML: Machine learning.

mtry: hyperparameter of Random Forest models. The number of features used to build each individual tree.

Phytoplankton: Microalgae that are the base of aquatic food webs.

rclr: robust centered log-ration transformation.

rRNA: ribosomal ribonucleic acid.

RF: Random Forest.

SMOTE: Synthetic Minority Oversampling Technique.

6 BIBLIOGRAPHY

BREIMAN, Leo, 2001. Random forests. *Machine Learning*. Vol. 45, no. 1, pp. 5–32. DOI 10.1023/A:1010933404324/METRICS.

BUNSE, Carina and PINHASSI, Jarone, 2017. *Marine Bacterioplankton Seasonal Succession Dynamics*. Elsevier Ltd. Trends in Microbiology 25. DOI 10.1016/j.tim.2016.12.013.

BUSATO, Sebastiano et al., 2023. Compositionality, sparsity, spurious heterogeneity, and other data-driven challenges for machine learning algorithms within plant microbiome studies. *Current Opinion in Plant Biology*. Vol. 71, p. 102326. DOI 10.1016/j.pbi.2022.102326.

CALLAHAN, Benjamin J., MCMURDIE, Paul J. and HOLMES, Susan P., 2017. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME Journal*. Vol. 11, no. 12, pp. 2639–2643. DOI 10.1038/ismej.2017.119.

CALLAHAN, Benjamin J. et al., 2016. DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*. Vol. 13, no. 7, pp. 581–583. DOI 10.1038/nmeth.3869.

COSTAS-SELAS, Cecilia et al., 2022. Role of Bacterial Community Composition as a Driver of the Small-Sized Phytoplankton Community Structure in a Productive Coastal System. *Microbial Ecology*. DOI 10.1007/s00248-022-02125-2.

CRAN TEAM, 2023. The Comprehensive R Archive Network. Online. 2023. Retrieved from: https://cran.r-project.org/index.html [accessed 17 June 2023].

DENG, Yun, VALLET, Marine and POHNERT, Georg, 2022. Temporal and Spatial Signalling Mediating the Balance of the Plankton Microbiome. *Annual Review of Marine Science*. Vol. 14, pp. 239–260. DOI 10.1146/annurev-marine-042021.

DOE-JGI, 2006. Integrated Microbial Genomes and Microbiomes (IMG/M). Online. 2006. Retrieved from: https://img.jgi.doe.gov/ [accessed 8 March 2023].

ELSEVIER, 2004. Scopus. Online. 2004. Retrieved from: https://www.scopus.com/home.uri [accessed 7 March 2023].

EMBL-EBI, 2023. European Nucleotide Archive (ENA). Online. 2023. Retrieved from: https://www.ebi.ac.uk/ena/browser/home [accessed 8 March 2023].

FREE SOFTWARE FOUNDATION, Inc., 2023. *Bash*. Online. Retrieved from: https://www.gnu.org/software/bash/

GHANNAM, Ryan B. and TECHTMANN, Stephen M., 2021. *Machine learning applications in microbial ecology, human microbiome studies, and environmental monitoring*. Elsevier B.V. Computational and Structural Biotechnology Journal 19. DOI 10.1016/j.csbj.2021.01.028.

GLASSMAN, Sydney I and MARTINY, Jennifer B H, 2018. Broadscale Ecological Patterns Are Robust to Use of Exact Sequence Variants versus Operational Taxonomic Units. *mSphere*. Vol. 3, no. 4, pp. e00148-18. DOI 10.1128/mSphere.

GONZALEZ, Antonio et al., 2018. Qiita: rapid, web-enabled microbiome meta-analysis. *Nature Methods*. Vol. 15, no. 10, pp. 796–798. DOI 10.1038/s41592-018-0141-9.

GOOGLE, 2004. Google Scholar. Online. 2004. Retrieved from: https://scholar.google.es [accessed 7 March 2023].

GRONNIGER, Jessica L. et al., 2022. Rapid changes in coastal ocean microbiomes uncoupled with shifts in environmental variables. *Environmental Microbiology*. Vol. 24, no. 9, pp. 4167–4177. DOI 10.1111/1462-2920.16086.

HE, Haibo and GARCIA, Edwardo A., 2009. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*. Vol. 21, no. 9, pp. 1263–1284. DOI 10.1109/TKDE.2008.239.

HERNÁNDEZ-RUIZ, Marta et al., 2018. Seasonal succession of small planktonic eukaryotes inhabiting surface waters of a coastal upwelling system. *Environmental Microbiology*. Vol. 20, no. 8, pp. 2955–2973. DOI 10.1111/1462-2920.14313.

JANSSEN, René et al., 2019. An artificial neural network and Random Forest identify glyphosate-impacted brackish communities based on 16S rRNA amplicon MiSeq read counts. *Marine Pollution Bulletin*. Vol. 149. DOI 10.1016/j.marpolbul.2019.110530.

JOGLAR, Vanessa et al., 2020. Spatial and temporal variability in the response of phytoplankton and prokaryotes to B-vitamin amendments in an upwelling system. *Biogeosciences*. Vol. 17, no. 10, pp. 2807–2823. DOI 10.5194/bg-17-2807-2020.

JOGLAR, Vanessa et al., 2021. Cobalamin and microbial plankton dynamics along a coastal to offshore transect in the Eastern North Atlantic Ocean. *Environmental Microbiology*. Vol. 23, no. 3, pp. 1559–1583. DOI 10.1111/1462-2920.15367.

KOVÁCS, György, 2019. An empirical comparison and evaluation of minority oversampling techniques on a large number of imbalanced datasets. *Applied Soft Computing*. Vol. 83, p. 105662. DOI 10.1016/J.ASOC.2019.105662.

LI, Peishun et al., 2022. Machine learning for data integration in human gut microbiome. *Microbial Cell Factories*. Vol. 21, no. 1, pp. 1–16. DOI 10.1186/s12934-022-01973-4.

LIU, Bin et al., 2022. Machine learning-assisted identification of bioindicators predicts medium-chain carboxylate production performance of an anaerobic mixed culture. *Microbiome*. Vol. 10, no. 1. DOI 10.1186/s40168-021-01219-2.

LLORET, Josep et al., 2018. Small-scale coastal fisheries in European Seas are not what they were: Ecological, social and economic changes. *Marine Policy*. Vol. 98, pp. 176–186. DOI 10.1016/j.marpol.2016.11.007.

MARCOS-ZAMBRANO, Laura Judith et al., 2021. *Applications of Machine Learning in Human Microbiome Studies: A Review on Feature Selection, Biomarker Identification, Disease Prediction and Treatment*. Frontiers Media S.A. Frontiers in Microbiology 12. DOI 10.3389/fmicb.2021.634511.

MARKOWITZ, Victor M. et al., 2006. An experimental metagenome data management and analysis system. In: *Bioinformatics*. Oxford University Press. 15 July 2006. DOI 10.1093/bioinformatics/btl217.

PARADA, Alma E., NEEDHAM, David M. and FUHRMAN, Jed A., 2016. Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. *Environmental Microbiology*. Vol. 18, no. 5, pp. 1403–1414. DOI 10.1111/1462-2920.13023.

PAUL, Allanah Joy et al., 2022. Upwelled plankton community modulates surface bloom succession and nutrient availability in a natural plankton assemblage. *Biogeosciences*. Vol. 19, no. 24, pp. 5911–5926. DOI 10.5194/bg-19-5911-2022.

PEDRÓS-ALIÓ, Carlos, 2012. The Rare Bacterial Biosphere. *Annual Review of Marine Science*. Vol. 4, no. 1, pp. 449–466. DOI 10.1146/annurev-marine-120710-100948.

PHIL EWELS, 2014. SRA-explorer. Online. 27 March 2014. Retrieved from: https://sra-explorer.info [accessed 27 March 2023].

PONTILLER, Benjamin et al., 2022. Rapid bacterioplankton transcription cascades regulate organic matter utilization during phytoplankton bloom progression in a coastal upwelling system. *ISME Journal*. Vol. 16, no. 10, pp. 2360–2372. DOI 10.1038/s41396-022-01273-0.

PORETSKY, Rachel et al., 2014. Strengths and limitations of 16S rRNA gene amplicon sequencing in revealing temporal microbial community dynamics. *PloS one*. Vol. 9, no. 4, p. e93827. DOI 10.1371/journal.pone.0093827.

PUENTE-SÁNCHEZ, Fernando, AGUIRRE, Jacobo and PARRO, Víctor, 2016. A novel conceptual approach to read-filtering in high-throughput amplicon sequencing studies. *Nucleic Acids Research*. Vol. 44, no. 4, pp. e40–e40. DOI 10.1093/NAR/GKV1113.

PYTHON SOFTWARE FOUNDATION, 2023. *Python Language Reference*. Online. 3.8. 3.8. Retrieved from: http://www.python.org

QUAST, Christian et al., 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic acids research*. Vol. 41, no. Database issue, pp. D590–D596. DOI 10.1093/nar/gks1219.

R CORE TEAM, 2023. *R: A language and environment for statistical computing*. Online. Vienna, Austria, Austria: R Foundation for Statistical Computing. Retrieved from: http://www.r-project.org/

SALAZAR, Guillem et al., 2019. Gene Expression Changes and Community Turnover Differentially Shape the Global Ocean Metatranscriptome. *Cell*. Vol. 179, no. 5, pp. 1068-1083.e21. DOI 10.1016/j.cell.2019.10.014.

SCHLOSS, Patrick D, 2020. Reintroducing mothur: 10 Years Later. *Applied and Environmental Microbiology*. Vol. 86, no. 2, pp. e02343-19. DOI 10.1128/AEM.

SCHMIEDER, Robert et al., 2010. TagCleaner: Identification and removal of tag sequences from genomic and metagenomic datasets. *BMC Bioinformatics*. Vol. 11, no. 1, p. 341. DOI 10.1186/1471-2105-11-341.

SUNAGAWA, Shinichi et al., 2015. Structure and function of the global ocean microbiome. *Science*. Vol. 348, no. 6237, pp. 1–10. DOI 10.1126/science.1261359.

TATARU, Christine A. and DAVID, Maude M., 2020. Decoding the language of microbiomes using word-embedding techniques, and applications in inflammatory bowel disease. *PLoS Computational Biology*. Vol. 16, no. 5, pp. 1–25. DOI 10.1371/journal.pcbi.1007859.

TEELING, Hanno et al., 2012. Substrate-controlled succession of marine bacterioplankton populations induced by a phytoplankton bloom. *Science*. Vol. 336, no. 6081, pp. 608–611. DOI 10.1126/science.1218344.

THOMPSON, Luke R et al., 2017. A communal catalogue reveals Earth's multiscale microbial diversity. *Nature*. Vol. 551, no. 7681, pp. 457–463. DOI 10.1038/nature24621.

TOPÇUOĞLU, Begüm D. et al., 2020. A framework for effective application of machine learning to microbiome-based classification problems. *mBio*. Vol. 11, no. 3. DOI 10.1128/mBio.00434-20.

TOPÇUOĞLU, Begüm et al., 2021. mikropml: User-Friendly R Package for Supervised Machine Learning Pipelines. *Journal of Open Source Software*. Vol. 6, no. 61, p. 3073. DOI 10.21105/joss.03073.

UBUNTU COMMUNITY, 2020. *UBUNTU*. Online. 20.04. 20.04. Retrieved from: https://help.ubuntu.com/20.04/ubuntu-help/index.html

WANG, Qiong et al., 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and environmental microbiology*. Vol. 73, no. 16, pp. 5261–5267. DOI 10.1128/AEM.00062-07.

WILHELM, Roland C., VAN ES, Harold M. and BUCKLEY, Daniel H., 2022. Predicting measures of soil health using the microbiome and supervised machine learning. *Soil Biology and Biochemistry*. Vol. 164. DOI 10.1016/j.soilbio.2021.108472.

YILMAZ, Pelin et al., 2011. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nature biotechnology*. Vol. 29, no. 5, pp. 415–420. DOI 10.1038/nbt.1823.

YUAN, Xuelian et al., 2022. Bacterial biomarkers capable of identifying recurrence or metastasis carry disease severity information for lung cancer. *Frontiers in Microbiology*. Vol. 13. DOI 10.3389/fmicb.2022.1007831.

7 APPENDICES

7.1 OTHER PCAS OF DIMENSION CAMPAIGN.

PCAs of Dimension campaign coloured according to variables not showed in the main text.

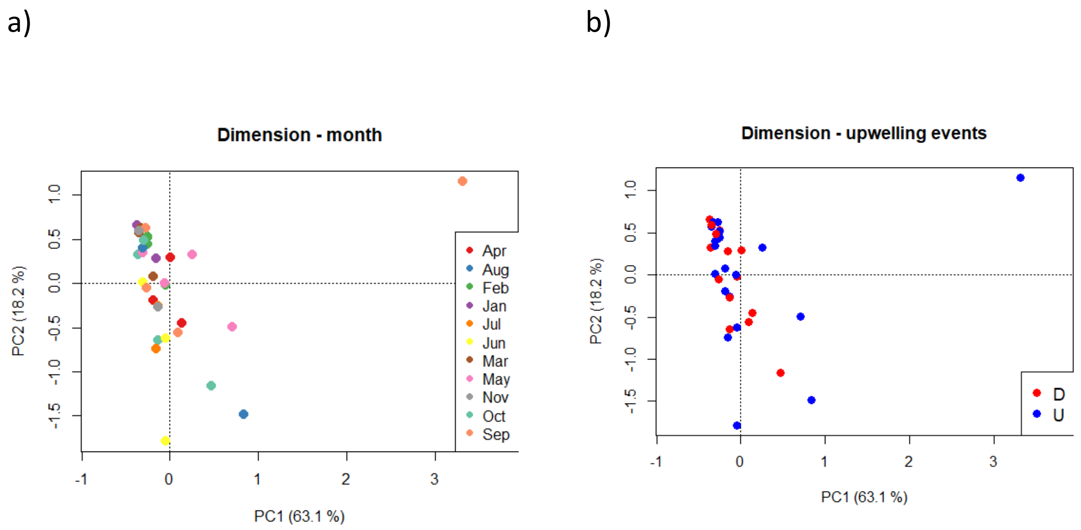


Figure A.1: *PCAs of environmental variables according to month (a) and Upwelling events (b). In Dimension the upwelling index that determines if an upwelling or downwelling event is occurring was calculated (D – Downwelling, U – Upwelling).*

7.2 TOTAL CHLOROPHYL DATA EXPLORATION

To help collaborators, the sample frequencies among different months, seasons, and depths when data was divided according to different percentiles of total chlorophyll concentration.

Division of data based on 50 and 75 percentiles.

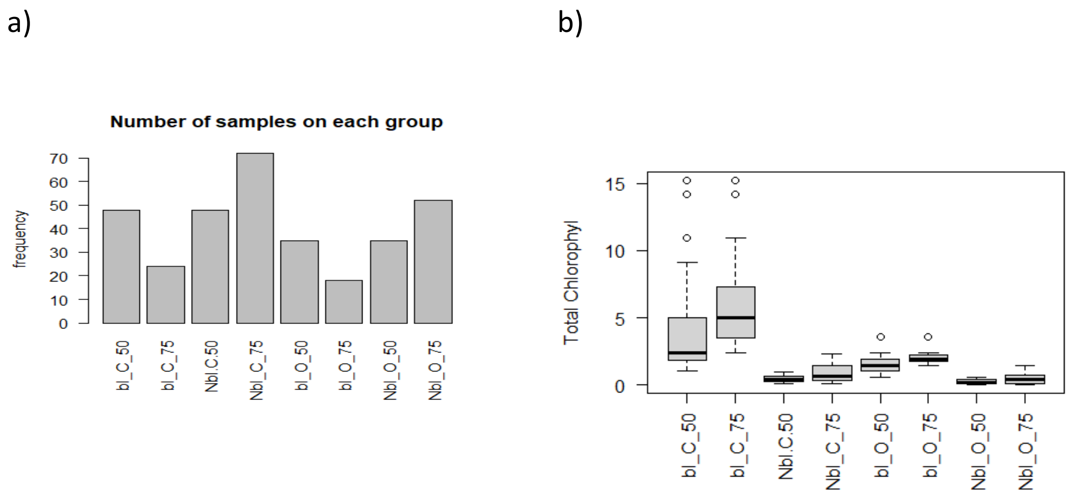


Figure A.2. *a) Number of samples on each group, b) boxplot of total chlorophyll concentrations on each group. Group name encoding: Bl: bloom, Nbl: no-bloom, C: coastal station, O: Ocean station, 50: division using 50 percentile, 75: division using 75 percentile.*

Details on the 75 percentile groups

Check the sample distributions through months, depths and other variables for each campaign using 75 percentile cutoff.

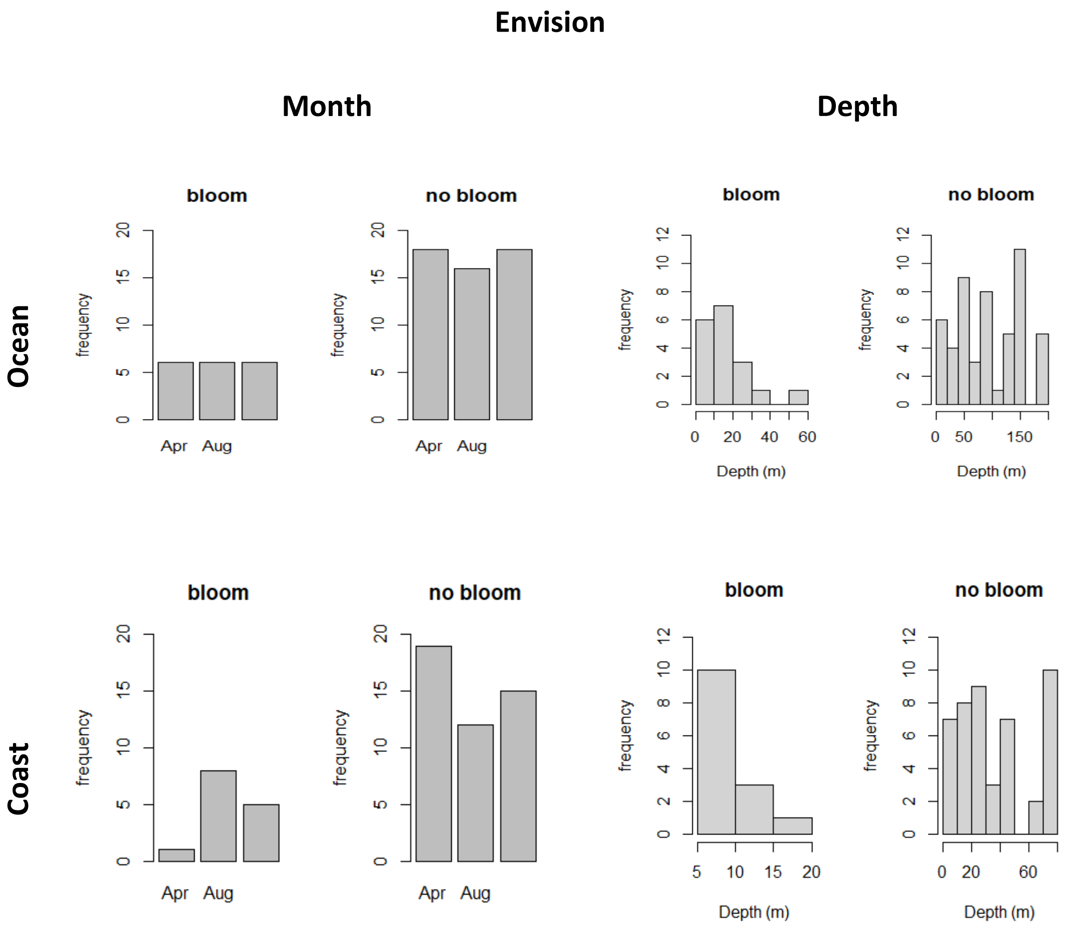


Figure A.3. *75 percentile division. Number of samples on bloom and no-bloom or normal groups depending on month and depth for coastal and oceanic stations for Envision campaign.*

Dimension – Coast only

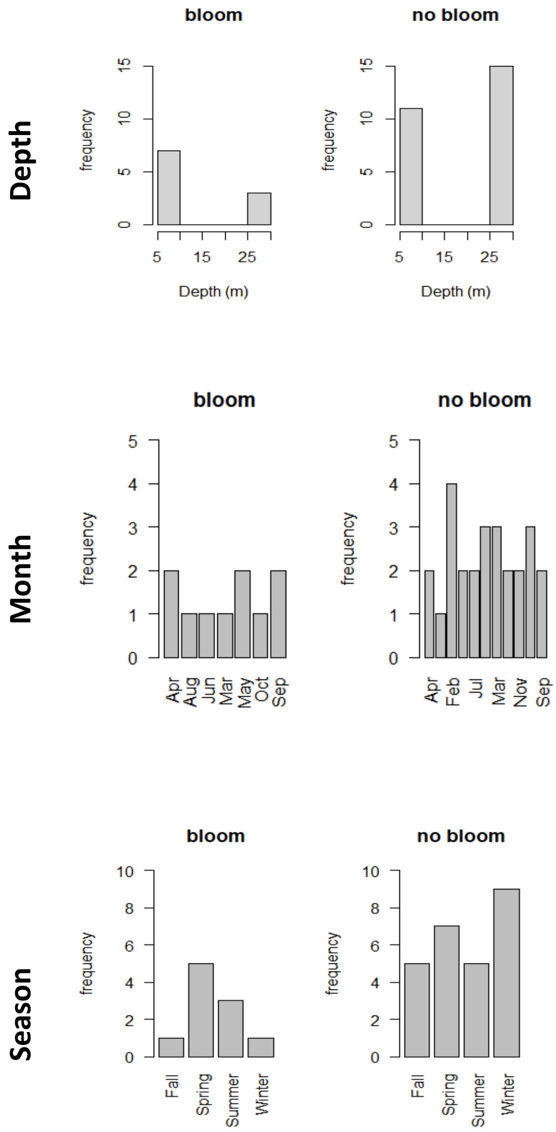
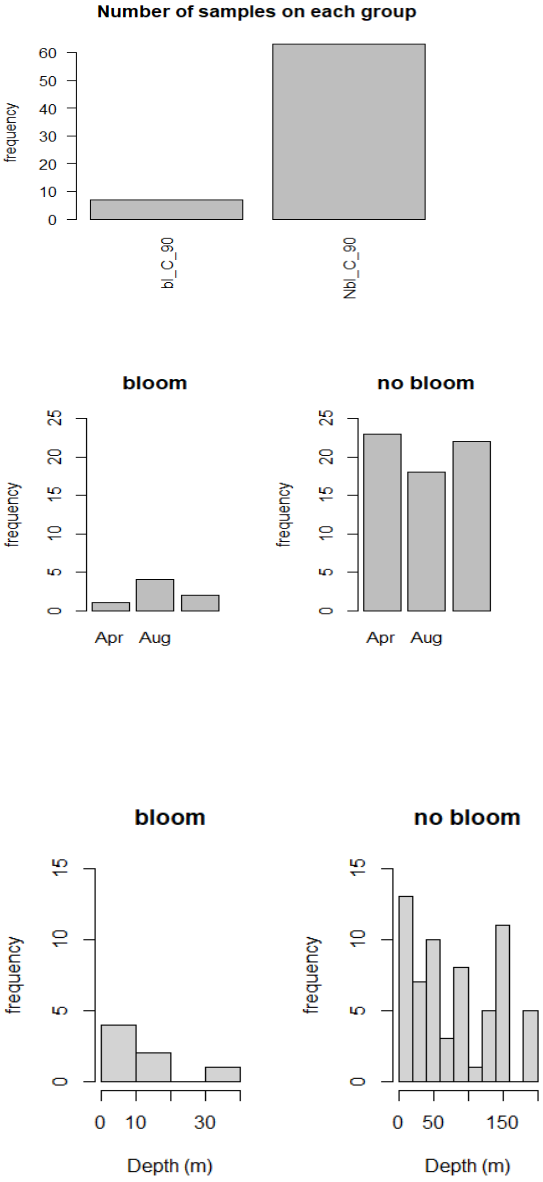


Figure A.4. *75 percentile division. Number of samples on bloom and no-bloom or normal groups depending on depth, month and season for Dimension campaign.*

90 percentiles for oceanic station.

Exploration of sample distributions using the 90 percentiles for the oceanic station.



Bloom - Normal

Month

Season

Figure A.5. *90 percentile division. Number of samples between samples on bloom and no-bloom or normal groups for the oceanic station, and according to depth, month and season.*

7.3 SEQUENCE COUNTS PER SAMPLE

Table A.1. *Number of sequences on each sample before and after chimera removal by DADA2*

Sample	Input	No chimera
DIMprok1	37507	31599
DIMprok10	21116	17873
DIMprok13	32583	25683
DIMprok16	28769	22188
DIMprok19	39469	32029
DIMprok22	23695	18697
DIMprok25	27114	20159
DIMprok28	17863	15994
DIMprok31	26951	18930
DIMprok34	23902	19975
DIMprok37	28029	20050
DIMprok4	20274	18856
DIMprok40	23054	20571
DIMprok43	29509	21106
DIMprok46	28232	24484
DIMprok50	221488	139305
DIMprok51	40291	30254
DIMprok52	20091	12477
DIMprok53	16864	11834
DIMprok54	17711	10585
DIMprok55	41982	26019
DIMprok56	36264	25207
DIMprok57	67335	48580
DIMprok58	87601	56605
DIMprok59	53048	40274
DIMprok60	53573	40983
DIMprok61	37475	29482
DIMprok62	44787	31639
DIMprok63	47069	36294
DIMprok64	24429	16535
DIMprok65	26568	21180
DIMprok66	21470	12314
DIMprok67	45914	37960
DIMprok68	35401	27193
DIMprok69	19800	16572
DIMprok7	29132	23210
ENVProk001	7141	5409
ENVProk002	6357	5597
ENVProk003	7637	6820
ENVProk004	7093	5287

ENVProk005	7169	5936
ENVProk006	9182	6923
ENVProk007	7479	4914
ENVProk008	8507	7399
ENVProk009	5405	4233
ENVProk010	6703	5510
ENVProk011	5919	5046
ENVProk012	6025	4500
ENVProk013	6950	5061
ENVProk014	6665	5385
ENVProk015	7539	7075
ENVProk016	7478	6326
ENVProk017	9359	7917
ENVProk018	6601	4843
ENVProk019	8653	5985
ENVProk020	6853	6196
ENVProk021	7250	5900
ENVProk022	9907	9341
ENVProk023	6451	4172
ENVProk024	6594	4628
ENVProk025	5149	3530
ENVProk026	8148	7149
ENVProk027	8036	7294
ENVProk028	6275	4777
ENVProk029	9347	7254
ENVProk030	8700	7739
ENVProk031	9003	7047
ENVProk032	8564	7487
ENVProk033	6664	4377
ENVProk034	7968	5770
ENVProk035	7603	5648
ENVProk036	6462	5316
ENVProk037	6579	5912
ENVProk038	7488	5525
ENVProk039	7437	5327
ENVProk040	8311	7211
ENVProk041	6315	4604
ENVProk042	8551	7499
ENVProk043	6092	5564
ENVProk044	7069	6366
ENVProk071	6811	4425
ENVProk072	6267	4800
ENVProk073	6543	4021
ENVProk074	8931	6034

ENVProk075	7586	4962
ENVProk076	6775	4666
ENVProk077	9944	7706
ENVProk078	8490	6506
ENVProk079	8443	7285
ENVProk080	6095	4654
ENVProk081	9462	5658
ENVProk082	7769	5692
ENVProk083	7141	4528
ENVProk084	8858	6930
ENVProk085	11391	7604
ENVProk086	9901	6672
ENVProk087	8526	7364
ENVProk088	13648	10271
ENVProk089	13077	9814
ENVProk090	7702	5522
ENVProk091	9696	8903
ENVProk092	10464	9460
ENVProk093	11422	8597
ENVProk094	7772	4377
ENVProk095	7914	5330
ENVProk096	10808	8735
ENVProk097	8234	5700
ENVProk098	7292	4844
ENVProk099	5066	3041
ENVProk100	7950	5935
ENVProk101	8959	6320
ENVProk102	7020	5774
ENVProk103	7015	4526
ENVProk104	7898	5404
ENVProk105	8341	5828
ENVProk106	7557	5376
ENVProk107	6458	4699
ENVProk108	7185	6201
ENVProk109	5907	5467
ENVProk110	7938	5471
ENVProk111	8597	6652
ENVProk112	6634	5884
ENVProk113	7308	5816
ENVProk114	7954	7232
ENVProk136	10742	6778
ENVProk137	8852	6214
ENVProk138	9453	7153
ENVProk139	10552	8567

ENVProk140	8276	6578
ENVProk141	3462	2173
ENVProk142	7193	6059
ENVProk143	7711	6040
ENVProk144	4488	4089
ENVProk145	10455	9123
ENVProk146	5950	5109
ENVProk147	9425	6489
ENVProk148	5314	4565
ENVProk149	9636	7410
ENVProk150	11625	9397
ENVProk151	9439	7298
ENVProk152	7442	5937
ENVProk153	8208	7140
ENVProk154	8325	7838
ENVProk155	5944	5176
ENVProk156	9327	8063
ENVProk157	5947	5091
ENVProk158	6013	4211
ENVProk159	4844	4211
ENVProk160	4553	4269
ENVProk161	8682	7447
ENVProk162	8977	7617
ENVProk163	4793	4095
ENVProk164	4003	3523
ENVProk165	8666	7392
ENVProk166	10723	9187
ENVProk167	6107	5679
ENVProk168	8501	6230
ENVProk169	11363	8750
ENVProk170	11461	8959
ENVProk171	9136	7528
ENVProk172	6092	5745
ENVProk173	4938	3663
ENVProk174	5694	4651
ENVProk175	11101	9277
ENVProk176	6661	6063
ENVProk177	9056	7831
ENVProk178	9901	8985

7.4 Exploratory analysis of unfiltered ASVs

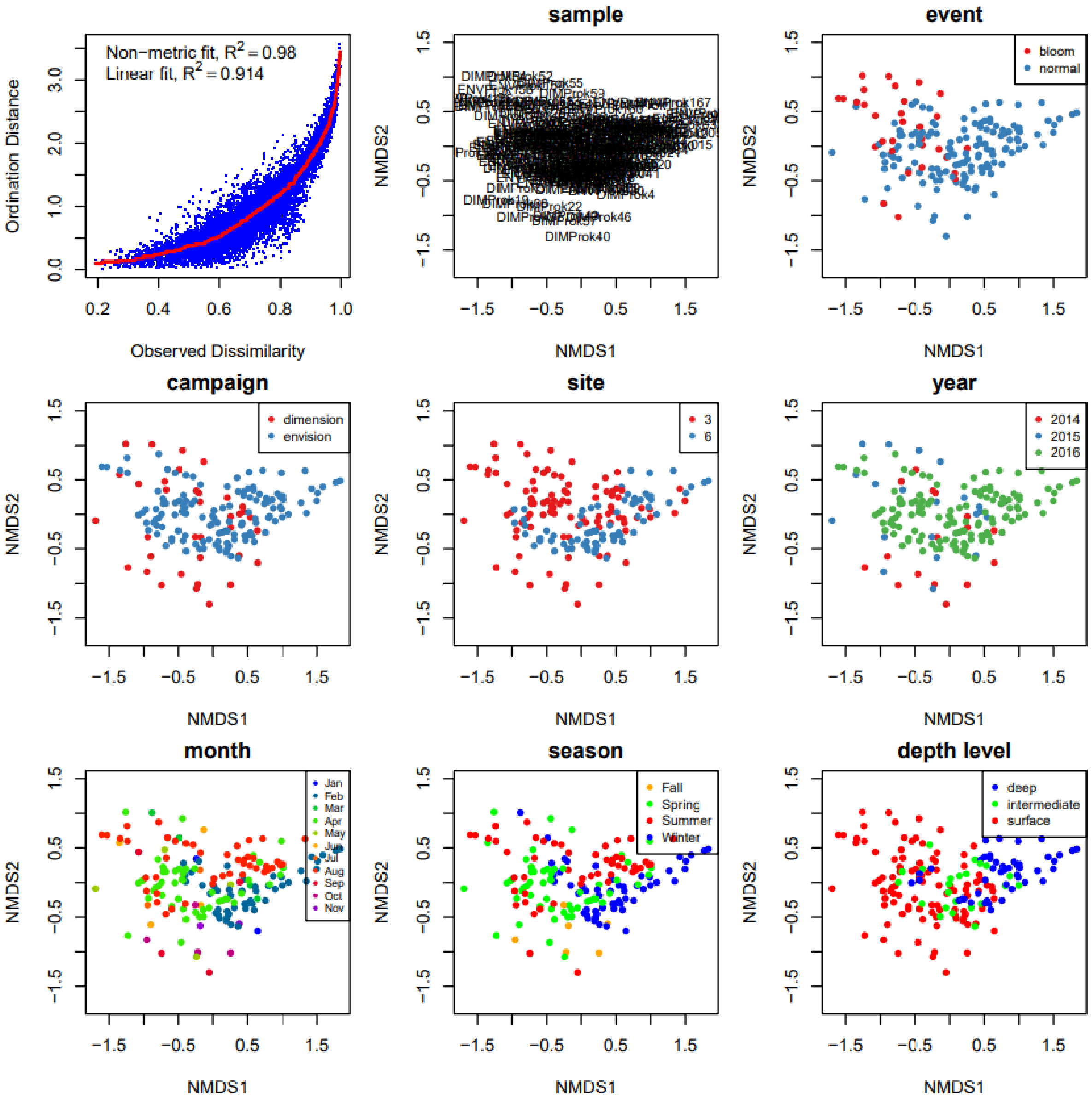


Figure A.6. *NMDS calculated with unfiltered ASVs with data transformed by the square-root of proportions. First panel contains the stress plot. The other panels contain the NMDS plot in which samples are labelled with their names or coloured by the categorical variable indicated.*

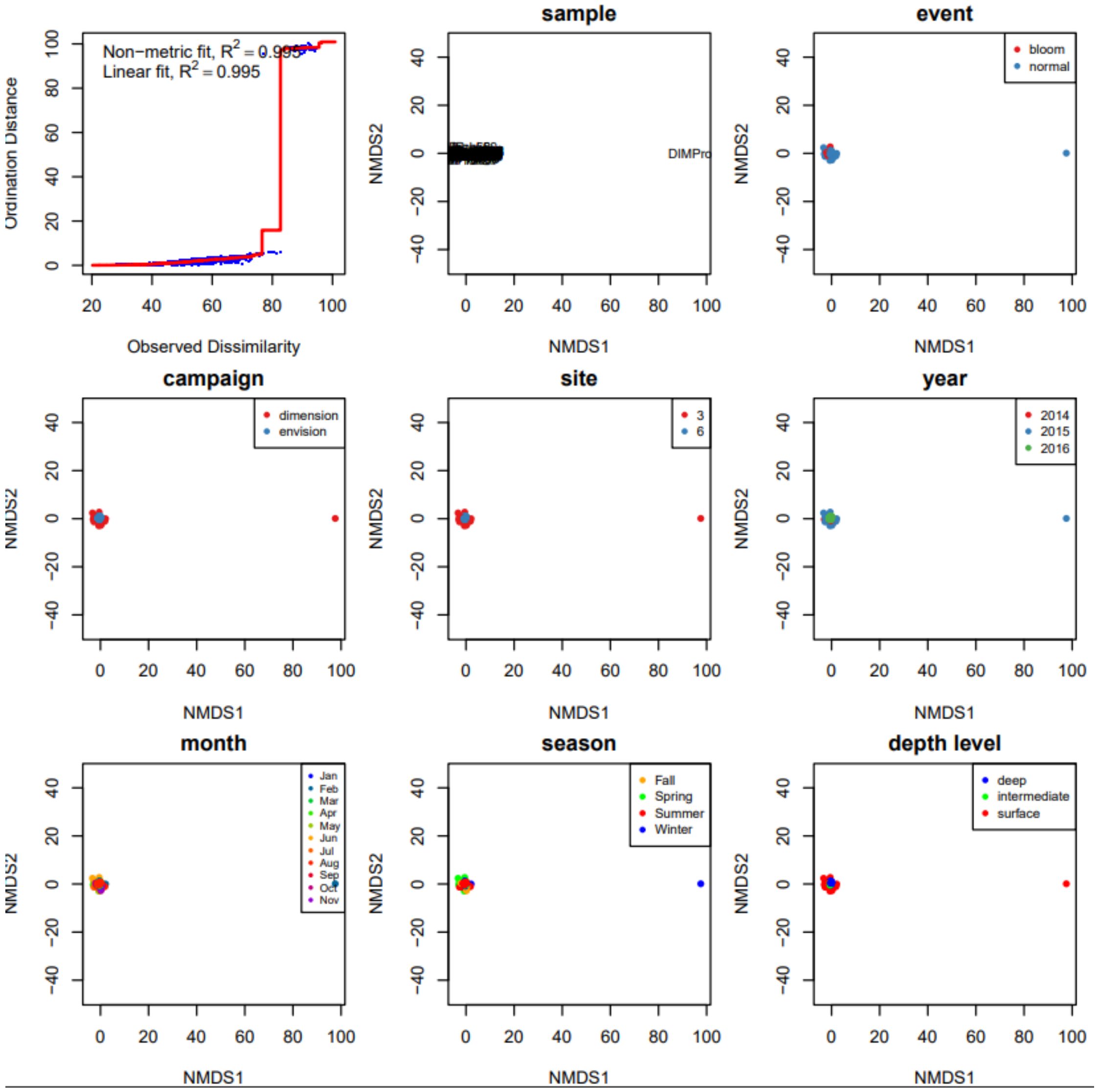


Figure A.7. *NMDS of unfiltered ASVs transformed with rclr. First panel contains the stress plot. The other panels contain the NMDS plot in which samples are labelled with their names or coloured by the categorical variable indicated.*

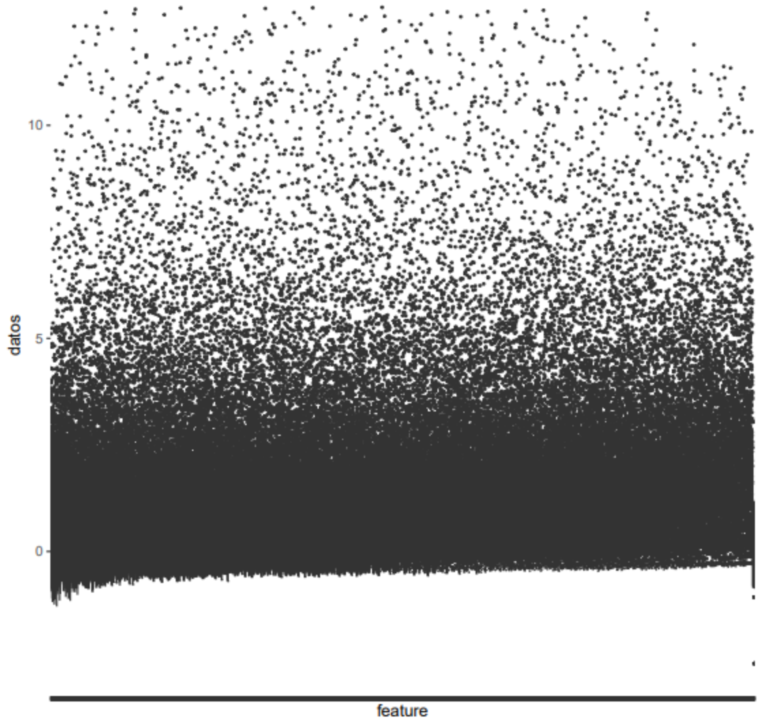


Figure A.8. *Boxplots of feature values from the unfiltered ASVs dataset after preprocessing and selection*

7.5 Features with mean importance values over 10

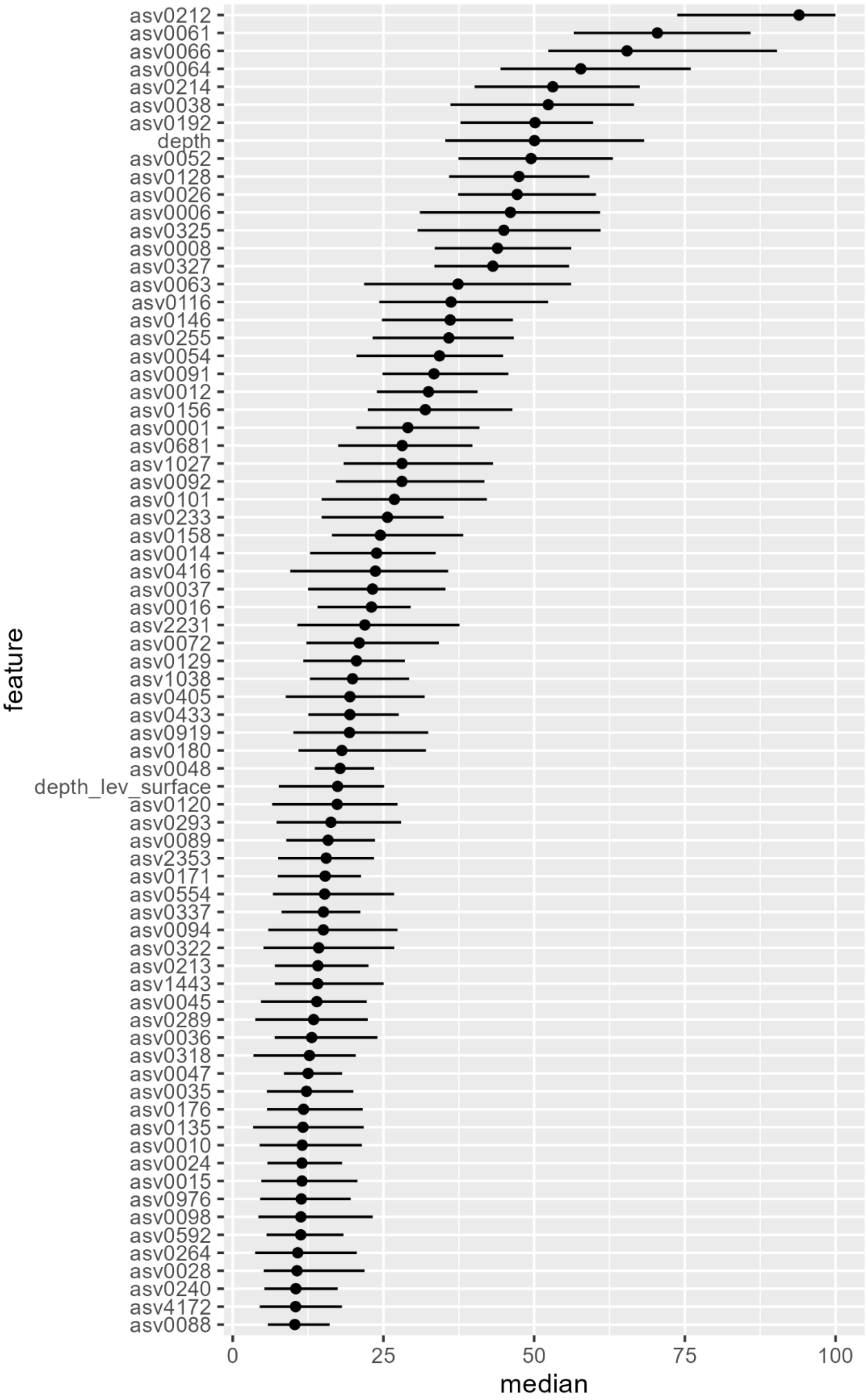


Figure A.8. *Biomarkers. Features with median scaled importance value across all 100 RF models over 10. Bars indicate the 25 and 75 percentiles.*

7.6 REPOSITORY CONTENTS

Scripts are located in the code folder that has the following structure:

Folder Original_files_preprocess:

Data download from SRA: `Data_collection.md, sra_explorer_download.sh`

Environmental metadata and sequence files preparation: `Metadata_files_preprocess.md`

Raw sequences preprocess: `Sequences_preprocess.md`

ASVs calculation: `ASVs_calculation.md`

Folder Chlorophyll_groups:

PCAs of environmental variables: `PCA_environmental_data.R`

Data distribution according to different percentiles of total chlorophyll concentrations: `Percentiles.R`

Main code folder scripts:

Data preparation:

Preparation of DADA2 results: `DADA2_tables.R`

Merge of Envision and Dimension environmental metadata and sample labelling: `Initial_preprocess.R`

Data exploration:

ASVs without filtering exploration: `ASVs_no_filter_exploration.R`

ASVs filtering and exploration: `ASVs_filter_exploration.R`

Clusters generation and exploratory analyses: `Clusters_grouping_exploration.R`

Train, validation with 5 and 10 k-fold repeated cross-validation and performance test of Random Forest and SVM models:

Random Forest with Clusters: `Clusters_model_1.R`

Random Forest with Filtered ASVs: `ASVs_model_1.R`

Random Forest with Unfiltered ASVs: ASVs_no_filter_model_1.R

SVM Radial with Filtered ASVs and validation results: `ASVs_SVM_model.R`

Validation and performance metrics for RF and SVM models: `RF_SVM_models_results.R`

Synthetic data:

Data generation and models training: `ASVs_synthetic_RF.R`

Performance results of RF model with synthetic data: `ASVs_synthetic_RF_results.R`

Features importance: `feature_importance.R`