DIPARTIMENTO DI SCIENZE PURE E APPLICATE

CORSO DI DOTTORATO DI RICERCA IN
SCIENZE DI BASE E APPLICAZIONI

Currriculum SCIENZA DELLA COMPLESSITÀ

CICLO XXXII

Analysis and forecasting of the structure of marine phytoplankton assemblages using innovative molecular techniques of NGS (Next Generation Sequencing) and Machine Learning

Settore Scientifico Disciplinare: BIO/07

RELATORE
Chiar.ma Prof.ssa Antonella Penna

CORRELATORE
Chiar.mo Prof. Michele Scardi

DOTTORANDO
Dott.ssa Eleonora Valbi

ANNO ACCADEMICO 2018/2019

Index

Abstract

The work carried out during these three years of Ph.D. followed two different objectives: the development of a predictive model for the occurrence of the toxic algal species *Alexandrium minutum* and the study, through data analysis techniques, of phytoplankton biodiversity at a global and local level, with particular interest into the European area.

Regarding the first of the two objectives, a predictive model was developed using the Random Forest technique. For this purpose, data relating to *A. minutum* occurrence, detected by the PCR technique carried out on samples of water taken in the NE Adriatic Sea area, and data relating to the different predictive variables, were provided to the program. Precisely, two models have been developed, one using 18 predictive variables, including the values of different nutrients, and one using only 12 of the 18 variables, excluding nutrient values, in order to a more easily use of the model in future, without the need for laboratory analysis. Results show that both models have good reliability values and that the 12-variable model works as well as the 18-variable one and can therefore be used without the risk of losing essential information. A work exposing these results has been published in the journal "Scientific Reports".

Regarding the second of the two objectives, we examined data relating to the OSD campaign, an international project that provided for collecting, by different partners all over the world, sea water samples during the day of summer solstice.

Samples were analyzed with metagenomic techniques, which led to the identification at the genus level of the different organisms present within the samples. After a firts exploratory analysis, in which data were analyzed using the Principal Coordinates Analysis technique, starting from a matrix obtained with to Jaccard coefficient between the different stations, the dataset was divided into clusters, thanks to a space-constrained clustering bound by a matrix of geographical connections obtained from a Gabriel graph. Subsequently, we decided to focus our attention on the two Longhurst ecoregions that had the

greatest number of samples, namely Mediterranean Sea and NE Atlantic Shelves Province, looking for associations between *taxa*. Two different association matrices, one for each province, were created using the Fager & McGowan coefficient and were subsequently analyzed using the Mantel test. The test result found a certain correlation between the two matrices. On the same line, also the result of the PROTEST performed on the two ordination originating from the two matrices. This suggests that associations between different *taxa*, more than being linked to the geographical position in which they are located, depend on other issues, such as, for example, the physical characteristics typical for every *taxon*. A manuscript presenting this work and its results is currently being drafted.

Preface

The research project provided, from its initial draft, two objectives: (i) the formulation of a predictive model for the occurrence of the toxic microalgal species, *Alexandrium minutum* Halim, 1960, in the NE Adriatic Sea, and (ii) an exploratory study of metagenomic data, related to the phytoplankton biodiversity at a global level, with particular attention to community structures and associations among the different *taxa*.

The work carried out during the first year of the Ph.D. program focused mainly on the first of the two objectives and concerned the training and the validation phases of the model.

The first months of the second year of the Ph.D. were focused on improving the performance of the previously developed predictive model. Subsequently, a manuscript was drawn up for the dissemination of the obtained results, addressing it to the journal "Scientific Reports". The work was therefore accepted and published, on March 12, 2019, as "Valbi E., Ricci F., Capellacci S., Casabianca S., Scardi M., Penna A. (2019). A model predicting a PSP toxic dinoflagellate occurrence in the coastal waters of the NW Adriatic Sea.".

During the period concerning the last semester of the second year and the whole third year of the Ph.D., the work focused on the second objective of the project, using the sequences of the genes encoding the 18S rRNA of the phytoplanktonic organisms found in sea water samples taken in the Ocean Sampling Day (OSD) campaigns in various areas of the world. A manuscript summarizing the results obtained is currently in the drafting phase.

Therefore, the Results paragraph of this Ph.D. thesis, will be divided into two parts: the first one concerning the developed predictive model, in which the

published article will be attached, the second one which will include all the results of the various exploratory analysis carried out on OSD data.

1. Introduction

1.1.Marine Phytoplankton

Oceans and seas cover 70% of the global earth's surface and constitute the largest continuous ecosystem on earth, accounting for 95% of the volume of the biosphere. The microbial communities that populate them make up about 90% of the marine biomass and play a fundamental role for life, not only in the aquatic environment, but also on land.

Firstly, the micro-organisms are involved in the biogeochemical cycles, oxidation and reduction processes of the chemical elements that allow the recycling of matter within the ecosphere. In particular, the fotoautotrophic part of the microorganisms present in the marine ecosystem, namely the phytoplankton, represented by microalgae and cyanobacteria, living near the surface where there is enough light to allow photosynthesis, is responsible for about the 50% of the global primary production (Falkowski *et al.* 1998; Behrenfeld *et al.* 2001; Falkowski & Raven 2007). Through photosynthesis, phytoplankton uses solar radiation and inorganic nutrients to convert inorganic carbon into organic carbon, releasing oxygen and producing biomass. For this reason, phytoplankton serve as the basis of the marine pelagic food web and, therefore, it is in a central role to life in the oceans (Field *et al.*, 1998, Seymour, 2014). With such a leading role, it seems clear that any alteration to phytoplankton can potentially impact on the whole ecosystem. Moreover, biomass production, being the basis of fish production, strongly influences the fishing sector and therefore a considerable part of the local and global economy.

In addition to indirectly benefit from the phytoplankton ecosystem functions, man uses marine ecosystems even directly, in the blue biotechnologies, for the development of new products with high economic value.

For example, microalgae have the potential to provide a new range of third generation biofuels; they can be used for the treatment of wastewater and industrial waters and in bioremediation programs of contaminated environments. While in the healthcare field marine microorganisms are used for the production of enzymes, drugs (antiviral and anticancer), cosmetics and nutraceutical compounds.

Water masses are contiguous on the planet, but the composition and the abundance of phytoplankton is significantly different in space and time (Gran, 1912).
The variability of phytoplankton depends on many mechanisms, such as the physiology of the organism and its life cycles. Carbon dioxide, sunlight, and nutrients availability are the factors that mainly influence phytoplankton growth, and its distribution may also be affected by the physical environment and the interaction with other organisms (Margalef, 1974). Human activities and global climate changes are now potent new drivers that significantly affect the functioning of coastal and offshore marine ecosystems, too (Hallegraeff, 2010; Huertas *et al.*, 2011; Sunday *et al.*, 2014).

Being the phytoplankton at the base of marine trophic webs, the analysis of its populations has a very important role. Eventual alterations to the community, in fact, can modify the structure and the functioning of an entire ecosystem.
Studying and knowing phytoplankton biodiversity and its community structure, with the dual purpose of being able to benefit from it and preserve it, is today of crucial importance.

Unfortunately, despite their high value, to date there is still little information on most marine microorganisms, their functions and ecological interactions between the different species, since most of these are not, or are difficult to cultivate under laboratory conditions, thus making it impossible to study its physiology.

In studies of microbial communities, this difficulty in cultivating certain phytoplankton strains inevitably turns into a loss of information. Not being able to cultivate it, we lose the information on the occurrence of a specific micro-organism is, obtaining a list of very reduced species, compared to the number naturally present in the reference sample. A further problem could also arise in the case of cultivable strains, since traditional microbiology techniques often involve a first phase of enrichment of the environmental sample and the subsequent isolation of the single species present.

It is therefore evident that already in the choice of the culture medium and the pre-enrichment conditions a selection is made in favor of some microbial species, with physiological needs similar to the selected conditions, but which do not necessarily represent the organisms or the organism that prevails in the environment. Another problem could be given by the probability that many of the microorganisms present in the environmental sample are not able to grow in pure cultures as they need to live in microbial consortia. Furthermore, it must be considered that the maintenance costs of microalgal strains in culture are usually very expensive.

Problems of this kind can be overcome, and in fact have already been overcome in recent decades, thanks to the advent of metagenomics, through which innovative techniques of culture-independent molecular biology have been developed, allowing us to collect a great deal of information on marine biodiversity.

Metagenomics is the study of the genomes of microbial communities directly in their natural environment, through the amplification of the present DNA and the subsequent sequencing of specific target regions. This avoids the need for isolation and cultivation of individual species.

Unlike the first techniques developed in the 1980s and 1990s, which involved the cloning of specific sequences (Sanger et al, 1975), the current techniques of Next Generation Sequencing (NGS) analyze in parallel thousands or millions of DNA sequences or RNA, without having to clone them in bacterial systems. In particular, the Illumina MiSeq® methodic, based on Sequencing By Synthesis (SBS), allows us to analyze millions of sequences in a few hours with maximum

accuracy and yield without reading errors. In this way it is possible to know all the biodiversity present in a given sample, being able to identify, in addition, many species of microorganisms that were once unknown.

These large amounts of information now available (called big data), included in international databases and easily accessible to scientists from around the world, can be used for exploratory studies on biodiversity and community structure.

1.2. Ocean Sampling Day

The idea for the second objective of the Ph.D. project came from the availability of data obtained from the OSD campaign.

The OSD is part of a European project, the EU 7FP Micro B3 (Marine Microbial Biodiversity, Bioinformatics, Biotechnology) based on an interdisciplinary consortium of 32 academic and industrial partners including world experts in bioinformatics, informatics, biology, ecology, oceanography, bioprospecting and biotechnology and legal aspects.

The multiple purposes of this international project are: the study of marine biodiversity, the development of innovative bioinformatics approaches to make data of the genomes of viruses, bacteria, archaea and marine protists and of the metagenomes of marine ecosystems available and the definition of new targets for biotechnological applications.

All samples were collected on a simultaneous sampling campaign of the world's oceans which took place on the summer solstice (June 21st) in 2014 and was repeated in 2015.

191 sampling sites participated in OSD on June 21st, 2014 (Kopf *et al.*, 2015) and most of them joined again in 2015. These sites range from tropical waters to polar environments (Fig. 1).

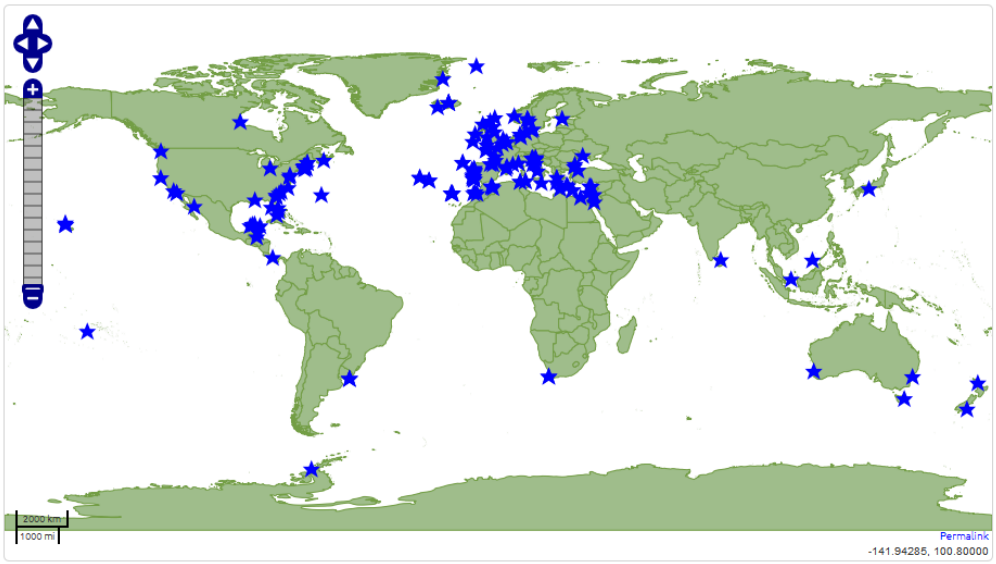

Fig. 1: OSD sampling sites.

The University of Urbino also actively took part in the project, contributing with a sampling carried out by the Environmental Biology laboratory of the Department of Biomolecular Sciences, along the coast of Pesaro.

All the collected samples were sent for metagenomic analysis and data, available on a global scale, were loaded into a database on the ENA-EMBL database.

The dataset includes the list of the different sequences of genes present in each station, the assignment of a taxonomic group of the sequences has been possible up to the genus level for most of the records, while for the remaining ones it has arrived at higher taxonomic levels .

The dataset contains both zoo- and phytoplankton, eukaryotic and prokaryotic species. For this work only the phytoplankton eukaryotic ones were considered, leaving the others aside.

1.3.Harmful Algal Blooms

In recent decades, worldwide, coastal systems have been facing a variety of environmental problems, due to the highly increasing of anthropic pressure. Some

of these problems are degradation or destruction of coral reefs and seagrass meadows (Boudouresque *et al.*, 2012, Bacci *et al.*, 2014), decreasing quality of coastal waters for recreational use, depletion of fish stocks (Pauly *et al.*, 1998) and Harmful Algae Blooms (HABs) events (Smayda and Reynolds, 2001, Bricker *et al.*, 2003, Kudela and Gobler, 2012, Wells *et al.*, 2015) (Fig. 2).

Fig. 2: Seawater affected by microalgal bloom.

HABs are caused by microalgae fast proliferation that has a negative impact on human activities.

When microalgae find suitable physical, biological and chemical conditions for growth, they can quickly reach high concentrations (104–105 cell L^{-1}) in a short period of time (commonly 1–3 weeks).

There are two types of harmful effects and harmful species can be related to one or both characteristics: high-biomass production and toxin production.

High biomass blooms affect the biota, causing fish kills and, therefore, influencing fishing and aquaculture industry (Hoagland and Scatasta, 2006, Berdalet *et al.*, 2015) and causing economical problems connected to the deterioration of the coastal recreational waters.

Marine biotoxins are represented by a heterogeneous group of chemical compounds, structurally different from each other, but with common characteristics. Generally they are stable to heat and acidic environment. A generic division of these compounds can be based on their solubility,

distinguishing them in water-soluble biotoxins and liposoluble ones (Poletti *et al.*, 2003).

Toxic syndromes in humans are caused by either the inhalation of aerosols (Gallitelli *et al.*, 2005, Ciminiello *et al.*, 2015) or the consumption of mussels, claims and oysters contaminated, that, being filter feeders, accumulate high concentration of these toxins in their digestive system.

Harmful species belong to six algal groups (diatoms, dinoflagellates, haptophytes, raphidophytes, cyanophytes, and pelagophytes) each with a specific morphology, physiology and ecology (Zingone and Enevoldsen, 2000; Garcés *et al.*, 2002).

Toxins produced by marine dinoflagellates are the most powerful non-protein poisons known (Steidinger, 1983; Steidinger and Baden, 1984; Anderson and Lobel., 1987). Several studies highlighted the importance of the biotransformation processes of algal toxins by molluscs and fish. In fact it has been shown that the metabolism of these animals can change the chemical structure of the toxin, causing a change in the toxic effect, making it forty times more powerful (Ade *et al.*, 2003). Currently, approximately 2000 cases of intoxication (with a 15% mortality rate) in humans due to consumption of toxic shellfish or fish are registered annually (Hallegraeff *et al.*, 1995).

Human pressure may influence the increasing of HABs events in different ways: transporting resting cysts of toxic species from a place to another, even very far, with ballast water and floating plastic, causing eutrophication, an over-enrichment of nutrient of the water (Smayda, 1989, Hallegraeff, 1993), inducing climate changes, exploiting the coastlines (Vila *et al.*, 2001, Garcés *et al.*, 2002), overfishing.

The HAB monitoring programs recently increased (Anderson *et al.*, 2012a) but, although many affected areas are well monitored, other are less controlled. Moreover, monitoring operations are often very expensive.

A great help about it may come from mathematical models, that are now able, with a high percentage of confidence, to predict these events. The use and

improvement of these techniques will be the next challenge in the upcoming future (Kleindinst *et al.*, 2014).

1.4. General characteristics of dinoflagellates

Dinoflagellates are a group of microscopic algae mostly unicellular and flagellated, which represent one of the most important phytoplankton groups both marine and freshwater with over 2000 living species. The cells are characterized by the presence of an outer membrane below which there is a layer of flattened vesicles (amphiesma), which may contain cellulose plaques. The presence/absence, the number, the layout and the morphology of the plates are a very important character for the classification (Steidinger and Tangen, 1997; Boni *et al.*, 2005).

The cell is divided into two distinct parts (Fig. 3): the epitheca is the upper part, which in some species can be very small and the ipotheca is the lower part (epicone and hypocone, respectively, in athecate species). These two parts are divided by a transverse septum called a cingulum. In the ventral part of the cell there is a longitudinal septum, called sulcus which starts from the cingulum. Some species have expansions, similar to sails, which originate from the two septa, probably to favor floating. Two flagella, which originate from a flagellar pore located where the cingulum and the sulcus converge, give the name to the group.

The life cycle of dinoflagellates has both asexual and sexual reproduction. In unfavorable conditions, the zygote can form a durable and resistant structure to the external environment: the cyst, which can remain in a dormant stage for a long time, and then begin vegetative reproduction once the environmental conditions are return favorable (Spector, 1984).

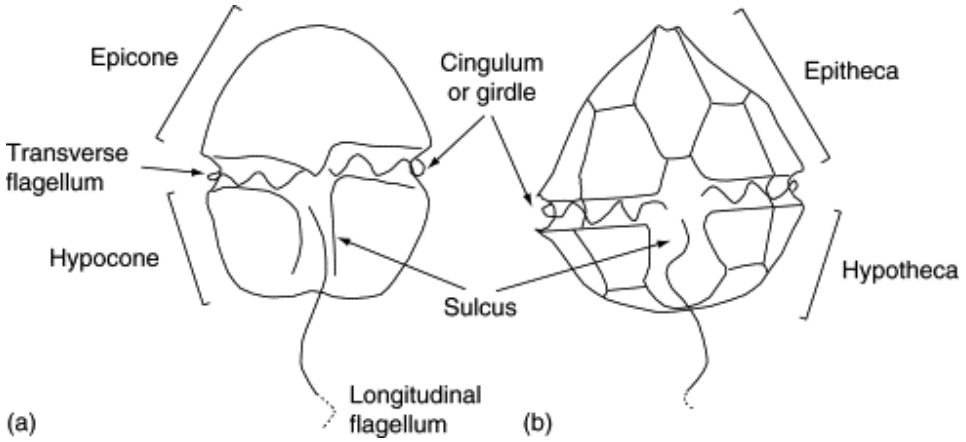

Fig. 3: Typical (a) athecate and (b) thecate dinoflagellate cells in ventral view.
From: Salmaso and Tolotti, 2009.

1.5.*Alexandrium minutum*

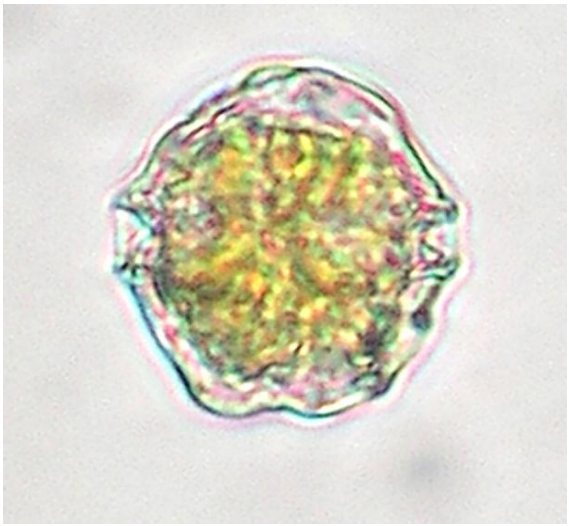

Fig. 4: The dinoflagellate Alexandrium minutum
PSP producing species.

The species studied in this doctoral thesis is the dinoflagellate *Alexandrium minutum* Halim, 1960 (Fig. 4), the most widespread toxic species in the western Mediterranean basin (Giacobbe and Maimone, 1994, Vila *et al.*, 2001). It can produce saxitoxins, GTX1 and 4, that can cause a severe human illness, the Paralytic Shellfish Poisoning (PSP) syndrome (Wiese *et al.*, 2010, Perini *et al.*, 2014), the most widespread HAB-related shellfish poisoning illness (Anderson *et al.*, 2012b), as well as one of the most studied and known syndromes because of the serious consequences that produces in consumers of bivalve molluscs. The onset of symptoms occurs in the thirty minutes following ingestion with paresthesia to the mouth, lips, tongue,

extremity of the limbs, profound muscular asthenia, inability to maintain an upright position. In fatal cases death occurs after 3-12 hours for respiratory paralysis. Patients that survive the first 12 hours, regardless of the amount of toxin ingested, usually recover quickly without other effects (Toyofuku, 2006). The severity of the symptoms depends on the amount of toxin that has been ingested. Symptoms are classified into moderate, severely moderate, and severely extreme. For the moderate the amount of saxitoxin varies from 2 to 30 μg / kg, while in the other cases the quantity varies from 10 to 300 mg / kg. (Tubaro and Hungerford, 2007).

The presence of *A. minutum* has been reported from coastal waters in various locations in the northern hemisphere, including Turkey, ltaly, Spain, Portugal, Ireland, The Netherlands, Germany, and the Atlantic coast of North America (Nehring, 1994), as well as from Australasian waters (Bolch *et al.*, 1991; Chang *et al.*, 1995).

This species has been responsible for toxic blooms along the northwestern coast of the Adriatic Sea (Italy) and Ionian Sea, where mussel farms have been contaminated (Honsell *et al.*, 1996, Penna *et al.*, 2015).

2. Methods

2.1. Mathematical models and Machine Learning

Natural phenomena can be expressed in mathematical language thanks to the use of models.

Their applications are used both for the synthesis and the description of knowledge and for predictive purposes.

The development of a mathematical model involves three fundamental phases (Guisan & Zimmermann, 2000):

- formulation. The starting point is, of course, an ecological concept. In order to carry out an adequate formulation of the model it is fundamental to choose the variables to be analyzed, as well as an adequate space-time scale;
- calibration. The ultimate goal of creating a mathematical model is to obtain the best possible accuracy. To achieve this, during this phase the values of the different parameters involved in the creation of the model are optimized;
- validation. Once the model has been formulated, its accuracy must be verified. In this phase various statistics are used for the purpose.

Given the considerable amount of information available today, it is becoming necessary to use mathematical models that exploit the increasingly efficient calculation tools. In particular, models that fall within the field of Machine Learning have proved to be very useful.

The term Machine Learning focuses the attention on the ability of computers to learn automatically from experience, mimicking what happens in human learning processes. In computer programs, performance is improved using experience, namely the available data that is provided to the algorithm (Alpaydin, 2014). What you get is a generalizable result, which can be used by entering new data, unknown to the machine.

Learning can be:

- supervised, which includes trained algorithms starting from the analysis of one or more response variables. The training is based on the informations provided in input, which are both those relating to the predictive variables, and those relating to the response, which is therefore known. The learning process generates hypotheses that can be used in the future to predict unknown cases, in which only the data relating to the predictive variables are available and the response variable is to be known. (Omary & Mtenzi, 2010). Examples of this type of learning can be Decision Trees, K-Nearest Neighbour (KNN), Support Vector Machines (SVM), Artificial Neural Networks (ANN) e Random Forests;
- unsupervised. This type of learning includes algorithms that analyze data without any predefined response example (Omary & Mtenzi, 2010), as in the case of some types of Neural Networks (Self Organizing Maps) and of different non-hierarchical classification algorithms.

2.2. Main methods used in the present thesis

2.2.1. Random Forest

Regarding the predictive aspect, the technique considered most suitable for the development of the model is the Random Forest (RF) (Breiman, 2001). This technique is based on a set of Classification Trees (CT), which are combined together to obtain predictions, formulated as the observations probability of belonging to the various response classes (Cutler *et al.*, 2007).
In case of predictive models for the occurrence of a given species, such as that examined in this thesis, the response classes are only two, presence and absence.

Each tree is characterized by the presence of splits which represent as many conditions.

In the specific case of a binary classification, there are only two responding alternatives to a given condition. Therefore, at each split a randomly chosen variable is taken into account and data are divided into two groups based on the value of the variable associated with them. Data will be assigned to one or the other class depending on whether the value of the said variable is higher or lower than a given threshold value, suitably defined by the algorithm so as to obtain the best possible division. The first split is followed by others, in which the data is subdivided again based on the value of one of the predictive variables, or even the same, but whose discrimination threshold is different. Finally, the tree ends with leaves, which contain the probability of belonging to each response class.

Variability is not only given by the high number of CTs combined by the RF. In fact, each tree randomly selects only a subset of data from the input set and a subset of predictor variables to be analyzed is selected for each node (Peters *et al.*, 2007).

Every single CT provides an answer, expressed as the probability of belonging to different classes. The various responses are combined together by the RF, whose output will therefore depend on the complex of those of the CTs that form it. The result of each of the different CTs is considered as a "vote". In practice, a RF creates its output based on the majority of the predictions of the trees that constitute it.

The training of the model is represented by a process in which the conditions characterizing the splits, the total number of CTs and the minimum number of cases present within each leaf (size) are defined and optimized.

At the end of the training, the program creates a particular contingency table, the confusion matrix (Tab. 1), in which each cell represents the combinations of observed or predicted data.

As already mentioned above, in this doctoral project a model was developed in which the RF response classes are only two. For this reason, the examples given

in this thesis concern this particular type of model, whose confusion matrix has a 2×2 dimension.

Tab. 1: Confusion matrix. TP: True Positives, TN: True Negatives, FN: False Negatives, FP: False Positives.

	Predicted: present	Predicted: absent
Observed: present	TP	FN
Observed: absent	FP	TN

Thus, four distinct cases are obtained:

- TP: true positives The observed and the predicted data correspond and are both positive, i.e. they are both presence cases;
- TN: true negatives. Even in this case, observed and predicted data correspond, but both are cases of absence of the studied species;
- FP: false positives. There is no concordance between observed and predicted data. The model predicts presence, but in the observed data the species is absent;
- FN: false negatives. Also in this case there is discrepancy. Predicted absent, observed present.

Model validation is based on two different approaches (Guisan & Zimmermann, 2000): il primo prevede l'uso di un singolo set di dati, con cui viene effettuata sia la calibrazione sia la validazione; il secondo prevede l'impiego di due distinti set di dati, il training set, usato durante la calibrazione, e il test set per la validazione (Manel *et al.*, 1999).

The first approach is generally chosen when the initial data set is not large enough and the model is validated by randomly extracting small subsets of data that vary during the training (cross-validation). If, instead, the initial data set is large enough, so that it can be divided into training sets and test sets, then it is possible to carry out a real validation.

Once the model is trained, its predictive ability must be tested. For this purpose are useful Correctly Classified Instances (CCI), Cohen's K coefficient (Cohen, 1960) and the Receiver Operating Characteristic (ROC) curve (Zweig e Campbell, 1993) are useful. In this study there was a subdivision of the original dataset in training and test set and the necessary tests were carried out on the test set results.

2.2.2. Cohen's K coefficient and CCI

Cohen's K coefficient is a statistic used to evaluate the accuracy of the model, measured as the level of agreement between observed and predicted data. It is defined by the author as the proportion of expected disagreements based on the hypothesis of a random association between observed values and expected values. particularly, K is defined as:

$$K = \frac{p_1 - p_0}{1 - p_0}$$

- p_1 is the overall proportion of expected concordance cases;
- p_0 is the overall proportion of cases of agreement expected under the hypothesis of random association.

K value, therefore, ranges from 0 to 1 and each interval of values corresponds to a level of agreement:

- $0 \leq K < 0.2$ the agreement is poor;
- $0.2 \leq K < 0.4$ the agreement is fair;
- $0.4 \leq K < 0.6$ the agreement is moderate;
- $0.6 \leq K < 0.8$ the agreement is good;
- $0.8 \leq K \leq 1$ the agreement is very good.

Correctly Classified Instances (CCI) is defined as the proportion of correctly predicted cases, often also referred to as sensitivity.

2.2.3. Roc curve and AUC

Species is default considered present by RF if at least 50% of the CTs "voted" classifying it as present. However, in cases where, as in ours, the absence data in the original dataset are more than those of presence, the RF learns better to classify the former than the latter.
This results in an imbalance of the results towards false negatives, i.e. cases in which the species is predicted as absent, but it is present, with a consequent decrease in the accuracy of the model. To overcome this problem, we used ROC curve to define the optimal threshold value of discrimination between presences and absences.

This technique relates two important indices, sensitivity and specificity, both calculable starting from the confusion matrix:

- sensitivity is the proportion of presence cases that have been correctly predicted (Smith, 2012), namely true positives;
- specificity is the proportion of absence cases correctly predicted (Manel *et al.*, 1999), namely false negatives.

In an XY graph, all the possible threshold values analyzed are plotted, taking the corresponding specificity values as coordinates on the x-axis and the sensitivity values on y-axis. By joining the points you will get a curve, namely the ROC curve. The optimal threshold value will be given by the cut-off value associated with the ROC curve point which has the maximum distance from the diagonal, which corresponds to the point where the true positives are minimized and the false negatives minimized (Fig. 5).

The area below the ROC curve is defined as Area Under the Curve (AUC), which can be a very useful measure to evaluate the performance of a model based on presence/absence data (Manel *et al.*, 2001). The AUC has a range that varies from a minimum of 0.5 up to a maximum of 1. According to the Swets classification (1988):

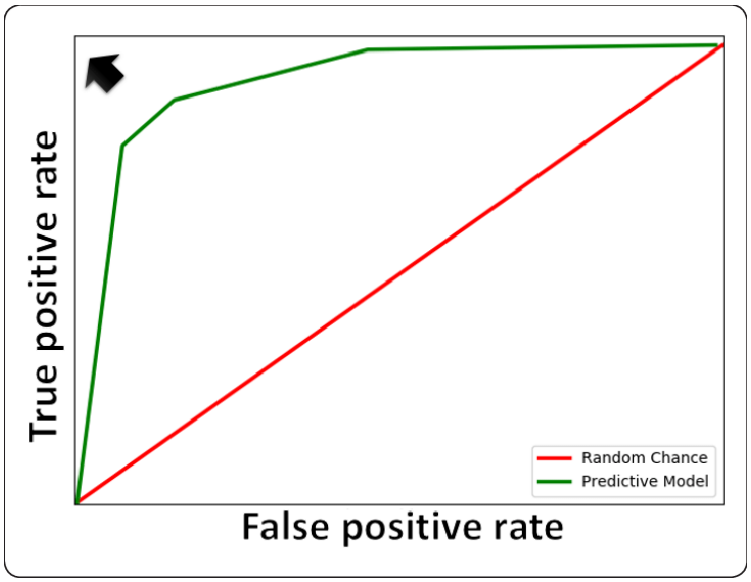

Fig. 5: Representation of a ROC curve. Ideal model (marked by an arrow); hypothetical curve (green) and random chance, diagonal line (red). From: Prieto-Martínez et al.*, 2018.*

- AUC = 0.5: the model is non-informative;
- O.51 < AUC < 0.7: the model has low accuracy;
- 0.71 < AUC < 0.9: the model is accurate;
- 0.91 < AUC < 0.99: the model is highly accurate;
- AUC = 1: the model is perfect.

2.2.4. Similarity, distance and association coefficients

When data matrix is composed of several variables such as, for example, a list of species, in order to evaluate the differences among observations, the original matrix is turned into a similarity or distance matrix, using a similarity or distance index, which calculates the differences between all the different pairs of observations. The choice of the right measure of similarity or distance is of fundamental importance.

Similarity coefficients express the similarity between two samples, assuming values that vary between 0, in the case of completely different observations, and

1, in the case of observations that fully satisfy the criterion used, not necessarily identical to each other.
Distance coefficients measure the differences between two observations and can take values ranging from 0, in the case of identical observations, to a variable value depending on the coefficient, for different observations.
Similarity measures can be transformed into dissimilarity measures, taking their complement to 1. Only some of these, however, own the metric properties typical of true distance coefficients, which allow to order the observations in a Euclidean space, through ordination methods.

Among the available coefficients symmetrical and asymmetrical ones can be distinguished.
This is a very important distinction and the use of one or the other type depends on the nature of the available data.
In ecological studies it can easily happen that for one or more descriptors there are null values. In some cases these correspond to a certain datum, at least within the limits of the error of sampling and determination methods (e.g. the absence of a pollutant). In other cases, instead, especially when analyzing lists of species, zero suggests more the absence of information. This can be due to the previously exposed problems of isolation and laboratory culturing and also identification by specialized operators difficulties. For our data, these circumstances have been avoided thanks to the metasequencing of the samples, which provides results in which it is probably more certain that the null datum indicates a real situation of absence of the species. However, there is another condition that could generate a lack of information, which cannot be prevented with metagenomic techniques: in a random sampling the species could not be taken, just for a stochastic effect. Usually, for this reason, more than one sample is taken, but this only reduces, without eliminating it, the probability of indicating as absent a species that was actually present.
Therefore, since presence data are more certain than absence ones (made more reliable by metasequencing, which avoids identification and classification errors),

in determining the similarity between two samples the former should have a greater weight than the latter.
In these cases, it is, therefore, more appropriate to use an asymmetric coefficient, thanks to which it is avoided to define a high similarity on the basis of information that are not certain.

Another distinction can be made between qualitative (or binary) and quantitative coefficients. While the quantitative ones, as the term, take into account the values of the descriptors, qualitative ones consider only the presence or the absence of the species, without going to investigate the quantity.
Due to the nature of our data, we chose to use a binary coefficient to summarize the relationships among the different samples, the similarity coefficient of Jaccard (1900, 1901, 1908), transformed into dissimilarity by taking its complement from 1.
For the purposes of the description of this coefficient it is useful to define the four possible cases in the comparison between the corresponding elements of two observations. This definition can be represented in a table as follows:

Observation *k* \ Observation *j*	1	0
1	*a*	*b*
0	*c*	*d*

Where *a* indicates the number of elements in common between two observations, *d* the number of null elements (absent) in both and *b* and *c* the number of non-null elements (present) exclusively in one or the other observation.

Jaccard coefficient does not take absences into account and therefore corresponds to the relation between concordances and the number of non-null elements of the observations:

$$S_{jk} = \frac{a}{a + b + c}$$

Association coefficients are used to analyze the relationships existing among the descriptors, in our case among the different *taxa* present in the different samples. In this case data are typically expressed in binary form, since the focus is not on quantitative relationships, but rather on the tendency of several species to occur jointly.
A coefficient developed specifically for the study of species associations is the one proposed by Fager & McGowan (1963):

$$S_{jk} = \frac{a}{\sqrt{(a + b)(a + c)}} - \frac{1}{2 \cdot \sqrt{a + c}} \qquad (c \geq b)$$

The second term represents a correction for preventing rare species from result strongly associated: in fact, the value of the coefficient decreases the most when the most frequent species between the two examined is rare.

Unlike similarity and distance indexes, association coefficients can undergo statistical tests, which usually aim at verifying the null hypothesis of independence between the descriptors.

2.2.5. Hierarchical clustering

Clustering techniques are used for grouping objects and defining subsets as homogeneous as possible.

Classification algorithms are all fairly recent, but, despite this, they constitute a rich and diverse set. They can be divided into two large groups: those of a hierarchical type and those of a non-hierarchical type.

Those of hierarchical type typically proceed by successive aggregation of objects using a matrix of similarity, or distance, between objects as a basis for their aggregation. The choice of the similarity coefficient (or distance) is in many cases even more decisive than that of the clustering algorithm in order to achieve the desired results.

An important category of clustering algorithms is that based on average distance (or similarity) measurements between groups.

In our study, the algorithm used to obtain a hierarchical classification of the sampling stations was the UPGMA (Unweighted Pair Group Method with Arithmetic Mean), which assigns an equal weight to the various groups, regardless of their size, and calculates the distance inter-group based on the average of the distances between the single objects.

2.2.6. Space-constrained clustering and Gabriel graph

When performing a clustering procedure on samples placed in different geographical areas, it is possible to take into account, in groups creation, the distances and geographical connections that exist between the different samples. This procedure is called constrained clustering.

For example, in the study of phytoplankton populations it is likely that geographically neighbour areas tend to have a similar population, due to the species' movement ability.

Constrained clustering differs from unconstrained one. While unconstrained clustering algorithm only uses the information in the dissimilarity or distance matrix, constrained clustering takes into account more information. During clustering procedure, priority is given to the constraint of spatial contiguity.

To take into account geographical contiguity of samples, before clustering, which sites are neighbours in space has to be determined (Legendre and Legendre, 1983).

Sites are usually positioned irregularly in space and they don't follow an ordinate pattern.

For the study of these situations, a very useful geometric connecting scheme is the Gabriel graph (Gabriel & Sokal, 1969).

With this criterion, the line connecting two points is part of the Gabriel graph if and only if no other point C lies inside the circle whose diameter is that line (Fig. 6).

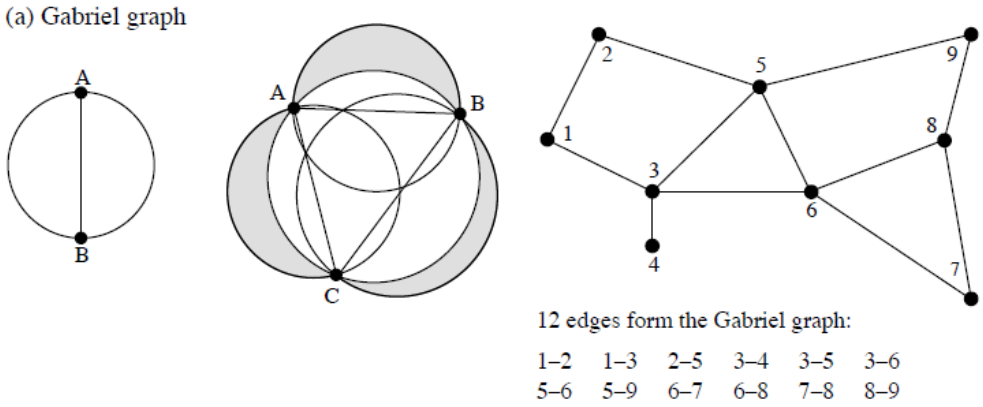

Fig. 6. a) Left: geometric criterion for the Gabriel graph. Centre: the zone of exclusion of the criterion, here for three points (grey zones + white inner circle). Right: graph for the example data, containing 12 edges. From: Legendre and Legendre,1983.

In this way, starting from point A, it is possible to find all the points closest to it, with which it will be connected.

2.2.7. Ordination methods and Principal Coordinates Analysis

Matrices obtained with the different distance, dissimilarity or association indexes can be visully explored to identify the most similar samples. This, however, becomes particularly difficult, if not impossible, with a huge data set, which

therefore generates an extremely large matrix: a typical situation in cases of a list of species observed in a certain number of samples.

In similar situations it is essential to use data analysis techniques, among which ordination methods are very useful. These have a dual purpose: to simplify the data set, reducing its dimensionality, defining linear combinations of its original variables preserving the information and to make a rigid rotation of the axes of the multidimensional space of data in a way to orient them in coherently with patterns of data dispersion.

The result is a graphical output, relatively simple to be interpreted, with the data being projected onto one or more planes, in which the Euclidean distance between the points representing the samples is proportional to the distance, dissimilarity or association value among them.

With this output a general comparison between the sites (or between the species, in the case of association index) is possible, in which the peculiarities of each station (or species) emerge.

As with the case of coefficients, even with ordination methods it is very important to choose the right technique, especially considering the starting data.

In this study Principal Coordinate Analysis (PCooA) (Gower, 1966), also called Metric Muldimensional Scaling (MDS) was used.

The algorithm on which the PCooA relies rotates and rescales the data set so that the distances in the resulting "cloud" are maximally correlated with the distances in the original data set. The optimal solution is found by calculating eigenvalues and eigenvectors through some steps:

- The D matrix of distances, dissimilarities or association between the n objects is transformed into the Δ matrix:

$$\Delta = -\frac{1}{2}\,\mathrm{D}$$

- The Δ matrix is centered so that the origin of the axes system that will be defined is in the centroid of the objects. Thus the matrix C is obtained:

$$C_{ij} = \delta_{ij} - \frac{1}{n}\sum_{h=1}^{n}\delta_{ih} - \frac{1}{n}\sum_{k=1}^{n}\delta_{kj} - \frac{1}{n^2}\sum_{h=1}^{n}\sum_{k=1}^{n}\delta_{hk}$$

where the second and third terms represent the row and column averages of the Δ matrix and the last term represents the general average of the matrix.

- Eigenvalues λ_j (j=1,2,...,m; $m \leq n$-1) and eigenvectors u_{ij} (i=1,2,...,n; j=1,2,...,m) of C matrix are calculated.
- Principal Coordinates f_{ij} of the objects are obtained by multiplying the eigenvectors by the square root of the corresponding eigenvalue:

$$f_{ij} = \sqrt{\lambda_j} \cdot u_{ij}$$

The quality of the ordination obtained for each main axis can be assessed on the basis of the ratio between the corresponding eigenvalue and the sum of the extracted eigenvalues.

2.2.8. PERMANOVA

Permutational Multivariate Analysis of Variance (PERMANOVA) (Anderson, 2001a; Anderson, 2001b; McArdle & Anderson, 2001) is a non-parametric statistical test based on permutations, suitable for dealing with multivariate data. This test verifies whether the differences observed between two or more groups of objects, defined *a priori* by the researcher, are significant, starting from a dissimilarity/distance matrix calculated using any coefficient.

The F statistic is calculated as the ratio between the inter-group variance and the intra-group variance.

The null hypothesis of equal composition is rejected if the variability between the groups is significantly greater than that present within the groups themselves.

To test this, a p-value is assigned to the F statistic value, calculated with a random permutation process of the group elements, carried out for a sufficiently high number of times.
As in any other statistical test, the null hypothesis is formally rejected if $p < 0.05$, or if the probability of randomly obtaining the same value of F or smaller ones is less than 5%.

2.2.9. Mantel Test

This test was originally developed for the study of the spatial distribution of occurrence of cancer cases (Mantel, 1967) and has recently been applied more and more often in the ecological field too.
It allows to obtain a measure of the degree of correlation between two matrices of distance, dissimilarity or association.
The null hypothesis tested is that of independence between the two matrices analyzed, while the probability level relative to the value of the statistic is calculated on the basis of an iterative procedure based on permutations of one of the two matrices.
The Mantel Z statistic, which expresses the degree of correlation between the structure of the two matrices, is calculated as the sum of the products of the corresponding elements of the two matrices, excluding those on the diagonal.
It is also possible to calculate the Mantel statistics in a standardized form and in this case it is indicated with R, as in this work. The probability level associated with the Mantel statistic value is calculated, as already mentioned, on the basis of an iterative procedure that provides for the random permutation of the rows and columns of one of the two matrices and the recalculation of the Mantel statistic for a high number of times. The value of the statistic obtained for the original matrices is compared with the empirical distribution of those obtained by repeating the calculation on randomly permuted matrices: the percentage of iterations in which a value lower than the original one was obtained corresponds to the probability level of the latter. From a practical point of view the null

hypothesis of independence between the matrices will be rejected if less than 5% of the values obtained for the permuted matrices is higher than the original one.

2.2.10. Procrustes test (PROTEST)

Procrustes test is an alternative to Mantel test to relate data matrices.

It is based on orthogonal Procrustes statistic (Hurley & Cattell, 1962), which combines together two ordination, derived from two different matrices with the same objects in rows. Each object has two representations. One of the data sets is kept fixed and the other one is rotated respect to the first, with a rotational-fit algorithm, with the aim of minimizing the sum of squared distances between the corresponding objects.

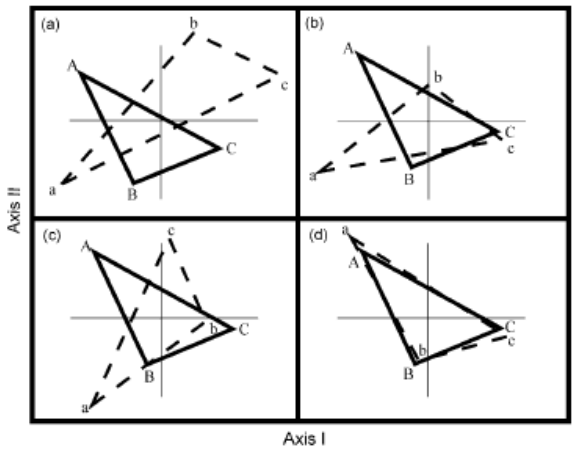

Fig. 7: Procrustes algorithm. For details see the text.

An example is shown in Fig. 7, where there are two triangles (X: A-B-C and Y: a-b-c) with different location, size and orientation.

X and Y are re-scaled and centered together (b). To make their orientation coincide, they are mirror reflected (c). Then. X is kept fixed and Y rotates until the sum of the squared distances between corresponding coordinates is minimized (d). In this way, the optimal superimposition is found.

The value of the squared distance (Δ^2_{12}) can be used as a measure of the concordance between the two datasets. In fact, the lower Δ^2_{12} value is, the greater the relation is.

Superimposition process requires that matrices have the same dimensions. This was not the case of our study, where we had two matrices with the same number

of rows (*taxa*), but a different number of columns (OSD). To overcome this obstacle, we reduced both sets to a two-dimensional space with the use of ordination technique to each matrix, as suggested by Peres-Neto and Jackson (2001).

As for the Mantel one, even Procrustes statistic can be tested to evaluate its significance. The test is called PROTEST (Jackson, 1995). In 2001, Peres-Neto and Jackson showed that PROTEST was more powerful than the Mantel test to identify correlations generated between raw data matrices.

The procedure is the following:

- Procrustes statistic is calculated;
- rows are randomly permuted in relation to each other of one data matrix;
- values for the permuted association are recalculated;
- steps 2 and 3 are repeated a large number of times.

Smaller values of Prurustes statistic indicate higher concordance between data sets.

3. Results

As already said, this paragraph is divided into two subparagraphs, one for each objective of the Ph.D. project.

3.1. Predictive model for *A. minutum* occurrence

The study area is that of the NE Adriatic Sea, the species of interest is, as already mentioned, the dinoflagellates *A. minutum*. Species occurrence data were determined by PCR analysis performed on 187 surface seawater samples collected, monthly, from June 2005 to December 2009 along the transects of the Foglia and Metauro rivers at 500 m and 3000 m from coastland, by the Environmental Biology laboratory of the Department of Biomolecular Sciences of the University of Urbino. Information on environmental variables was obtained from the samples too, with laboratory analysis. The technique used is the RF. To select the combination of the different RF parameters that allowed us to get the best result, several RFs were trained. This multiple training was carried out in parallel for two different models: one trained using all the available environmental variables and one using only 12 out of the 18. The reduced data set excluded information about nutrients to make any future use of the model easier, with no need for water sampling and laboratory analysis to determine nutrients concentrations.

Results showed that 12-variables model was as good as the 18-variables one. In particular, the model is able to correctly predict more than 80% of the instances in the test data set. This underlines the important role that predictive models may play in the study of HABs.

Further information about experimental design and modeling procedure are given in the publication attached below.

3.2. OSD data analysis

For the second objective of the Ph.D. project, we analyzed the sequences of the encoding genes for the 18S rRNA of phytoplankton organisms found in seawater samples collected during the OSD campaigns in various areas of the world.

The dataset includes the list of different gene sequences present in each station. The assignment of a taxonomic identifier to the sequences has been possible up to the genus level for most of the records, while for the rest it has been possible to arrive at higher taxonomic levels.
The dataset contains both zoo- and phytoplankton, eukaryotic and prokaryotic species. For this work only the phytoplankton eukaryotic ones were considered, leaving the others aside.

First, a check was carried out on the original dataset, eliminating the ambiguous *taxa*, or merging them at the higher taxonomic level where identification at a lower level was not possible. Moreover, ubiquitous species (present in more than 80% of the stations) were eliminated. Stations with particular features (for example those of brackish water) were not taken into account, too.
The dataset resulting from these first operations, on which the subsequent analysis were carried out, is composed of 103 stations and 187 different *taxa* (Fig. 8).

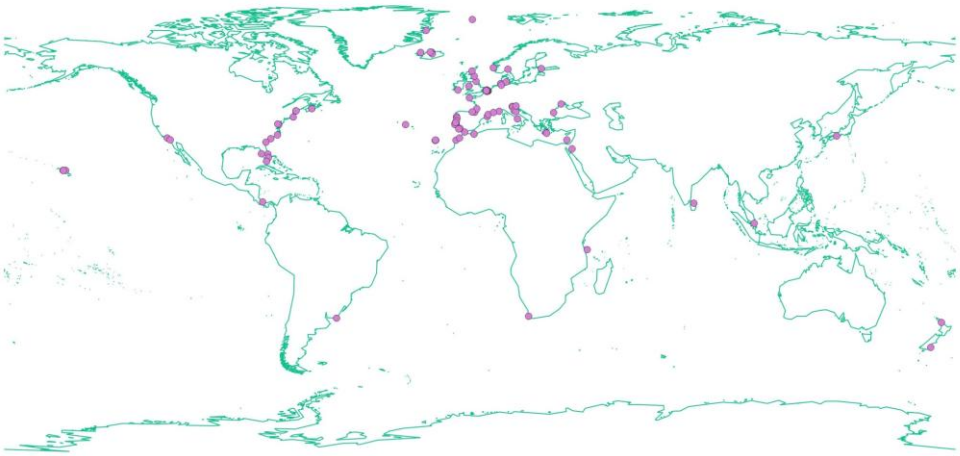

Fig. 8: Examined OSD samples.

From a geographical point of view, marine environment can be divided into several ecoregions. Among the various classifications present in the literature, the one taken in support of our analysis is that of the biogeochemical provinces of Longhurst (1998) (Fig. 9).

Tab. 2: Longhurst biogeochemical provinces and related IDs.

Longhurst Province	**ID**
Coastal - Brazil Current Coastal Province	BRAZ
Coastal - California Upwelling Coastal Province	CCAL
Coastal - Canary Coastal Province (EACB)	CNRY
Coastal - E Africa Coastal Province	EAFR
Coastal - Guianas Coastal Province	GUIA
Coastal - NE Atlantic Shelves Province	NECS
Coastal - New Zealand Coastal Province	NEWZ
Coastal - NW Atlantic Shelves Province	NWCS
Coastal - Red Sea, Persian Gulf Province	REDS
Coastal - Sunda-Arafura Shelves Province	SUND
Polar - Atlantic Arctic Province	ARCT
Polar - Atlantic Subarctic Province	SARC
Polar - Boreal Polar Province (POLR)	BPLR
Trades - Caribbean Province	CARB
Trades - Indian Monsoon Gyres Province	MONS
Trades - N. Pacific Tropical Gyre Province	NPTG
Westerlies - Kuroshio Current Province	KURO
Westerlies - Mediterranean Sea, Black Sea	MEDI

Tab. 2 shows names and identification codes of the different provinces involved in our study.

Province	
Westerlies - N. Atlantic Subtropical Gyral Province (East) (STGE)	NASE
Westerlies - S. Pacific Subtropical Gyre Province	SPSG

Due to the denominations' length, reference will be made to the identification codes, in the text below.

Through PAST (PAleontological Statistics) program, developed by Hammer *et al.* (2001), a dissimilarity matrix was generated, taking the different stations as object and using the complement to 1 of the Jaccard coefficient.

Starting from this matrix, it was possible to explore the data through PCoA. The obtained ordination is shown in Fig. 10.

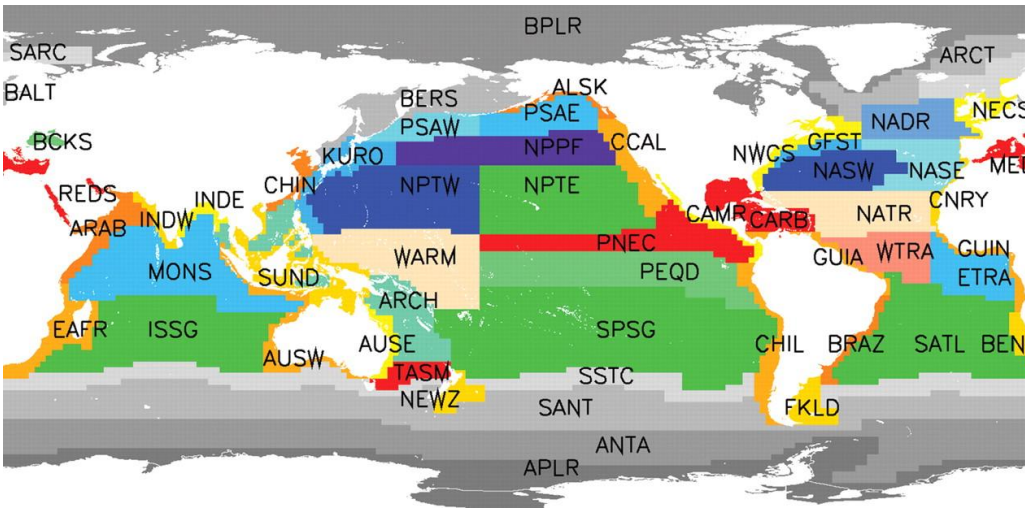

Fig. 9: Longhurst biogeochemical provinces.

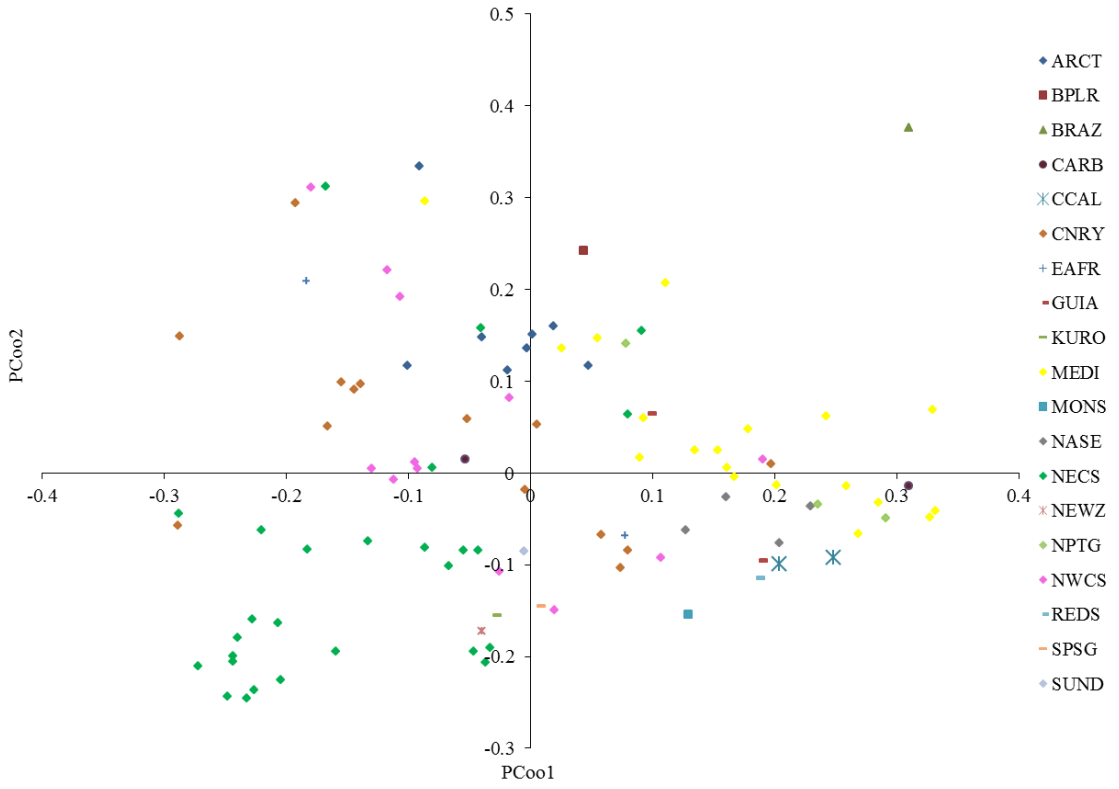

Fig. 10: Ordination obtained through Principal Coordinate Analysis of OSD samples. The ordination is based on Jaccard coefficient and the first two axes explain 19.4% of the total variance. Different colors correspond to different Longhurst ecoregions (legend to the right).

The first two principal axes respectively explain 11.3% and 8.14% of the total variance.

Each point is associated with an OSD and different colors correspond to different Longhurst ecoregions, listed in the legend. Overall, the points are uniformly arranged on the plane. Since some ecoregions are poorly represented, in Fig. 11 only those represented by at least four different stations are shown, with their convex hull.

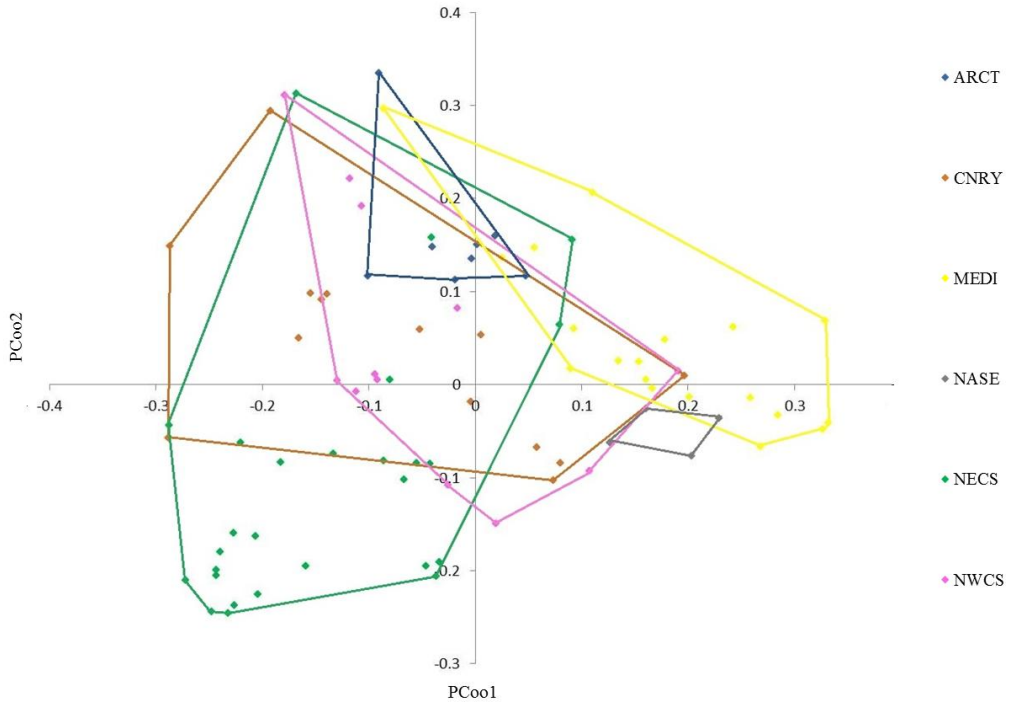

Fig. 11: Ordination of OSD samples belonging to Longhurst regions represented by at least four different stations, with their convex hull.

The latter are almost all partially overlapping. This is true especially for provinces ARCT, CNRY, NECS and NWCS provinces. NASE is the most isolated of the six and this is probably due to the characteristic that distinguishes it from the others: together with NPTG, they are the only two, in the whole dataset, related to non-coastal waters and this could probably influence the phytoplankton composition present. The MEDI group is mainly placed on the first quadrant, with positive X

values and almost all positive Y values, with few exceptions. The ARCT group is mainly located in the second quadrant, with mainly negative X values and always positive Y values. The CNRY group focuses mainly on the second quadrant. The NWCS group is located in the center, covering all the four quadrants. The NECS group mainly occupies the third quadrant, the X values are almost all negative and so are those of the Y, with only a few exceptions. The NASE group is the only one to occupy only the fourth quadrant, with values of X always positive and Y always negative.

Looking at the X-axis, the main distinctions seem to exist between MEDI and CNRY, MEDI and NWCS and, above all, MEDI and NECS. In fact, all the points of the first, with the exception of one, are on the positive quadrants, while all the points of the second, except for two, are on the negative one. Looking at the Y axis, we can find again the distinction between MEDI and NECS and we also note that between NECS and NWCS.

In order to have a response that was not only exclusively visual, PERMANOVA was performed on the same matrix. Bonferroni corrected p-values are significant for almost all the couple of ecoregions considered (Tab. 3).

Tab. 3: Bonferroni corrected p-values of PERMANOVA analysis of the six ecoregions considered.

	ARCT	CNRY	MEDI	NASE	NECS	NWCS
ARCT		0.0136	0.0136	0.34	0.0136	0.1768
CNRY	0.0136		0.0136	0.3808	0.0136	0.8704
MEDI	0.0136	0.0136		0.0272	0.0136	0.0136
NASE	0.34	0.3808	0.0272		0.0136	0.0544
NECS	0.0136	0.0136	0.0136	0.0136		0.0136
NWCS	0.1768	0.8704	0.0136	0.0544	0.0136	

Values highlighted in red do not allow to preserve the null hypothesis of equal composition of *taxa* in the different ecoregions, thus highlighting a difference in phytoplankton populations.

Later, a matrix of geographical connections between the various stations was created through an *ad hoc* program, starting from a Gabriel network. The obtained connections were modified where necessary, eliminating those that seemed unlikely, such as those linking OSDs separated from continental zones, and adding some, where it was known that there could be a connection between the OSDs, as in the case of Hawaiian stations linked to Japanese ones (Fig. 12).

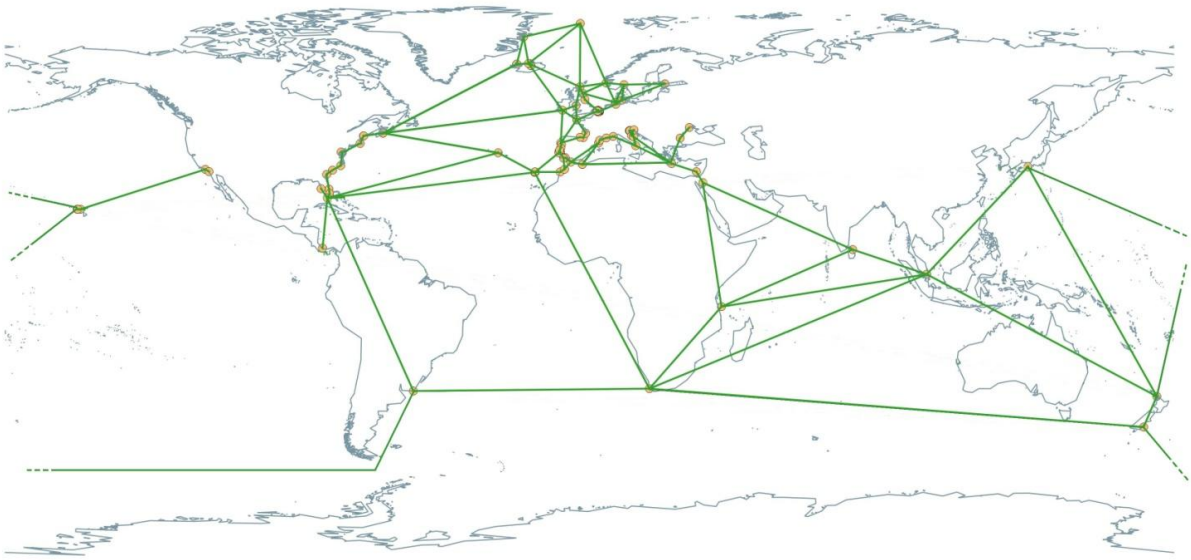

Fig. 13: Gabriel network (modified) of OSD samples.

With another *ad hoc* program, a space-constrained hierarchical clustering was carried out, with respect to the previously created connection matrix.

To determine which was the optimal number of clusters in which to divide the dataset, the mean intracluster distance was taken into account and, looking for the partition for which this value was lower (specifically, equivalent to 0.641), we were able to identify the one in 10 clusters as the most natural subdivision for the available data (Fig. 14 and 15) (legend in Tab. 4).

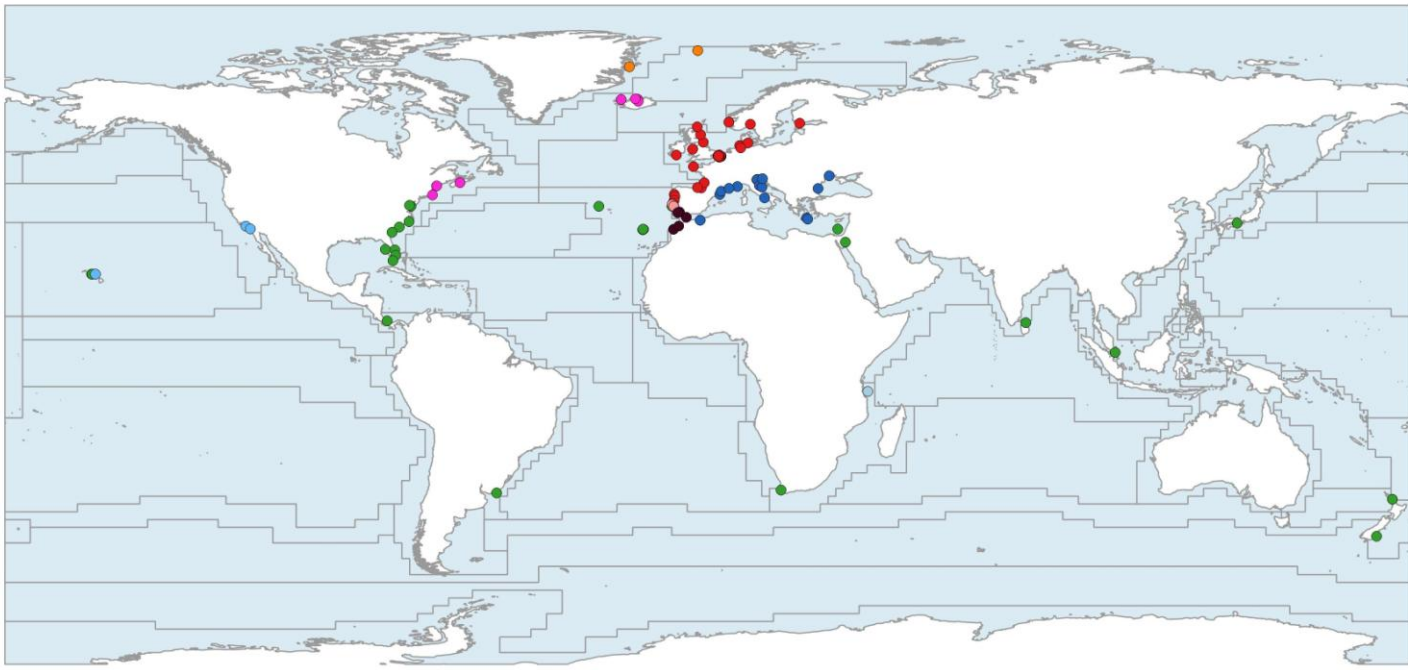

Fig. 14: Space-constrained clustering of OSD samples.

Most abundant clusters are B, which includes 33 stations, located in the North Sea and on the northern coasts of the Iberian peninsula and which fall within the space delimited by two Longhurst ecoregions (CNRY and NECS), A, with 16 stations exclusively present in the Mediterranean Sea (MEDI), C, with 12 stations, distributed on the Icelandic coasts and along the eastern coasts of Canada and northern United States (NWCS and ARCT) and D, the most heterogeneous among the groups, which includes 26 stations, located in various areas of the globe, which fall within 14 different ecoregions. That of group D can be considered a special case. In fact, looking at the map showed in Fig. 14 and 15, it can easily be

seen that these are much more numerous and close together in the northern hemisphere, while only 5 are found in the southern hemisphere.

Fig. 15: Space-constrained clustering: detail of Italian stations.

Tab. 4: Clusters and relative colors on the map.

	Cluster A
	Cluster B
	Cluster C
	Cluster D
	Cluster E
	Cluster F
	Cluster G
	Cluster H
	Cluster I
	Cluster J

These last samplings tend to be single spots, which have little in common with stations that are very distant from them. Therefore, the D group, in addition to the stations present along the eastern coast of the United States, groups all these individual cases, baing not very explanatory for the purposes of our analysis. The remaining 6 clusters are also special cases, which are influenced by specific and localized environmental conditions: E, with 3 stations, two on the west coast of the United States and one on

the Hawaiian Islands (CCAL and NPTG), F, with two stations in the Gulf of Venice (MEDI), G, with 5 stations along the southern coasts of the Iberian peninsula and Morocco (CNRY), H, with 3 stations in the Tagus estuary (CNRY), I, which includes the only station along the East African coast (EAFR) and J, with 2 stations in the Greenland Sea (ARCT and BPLR).

Given the nature of the available data, it was decided to focus the attention on the

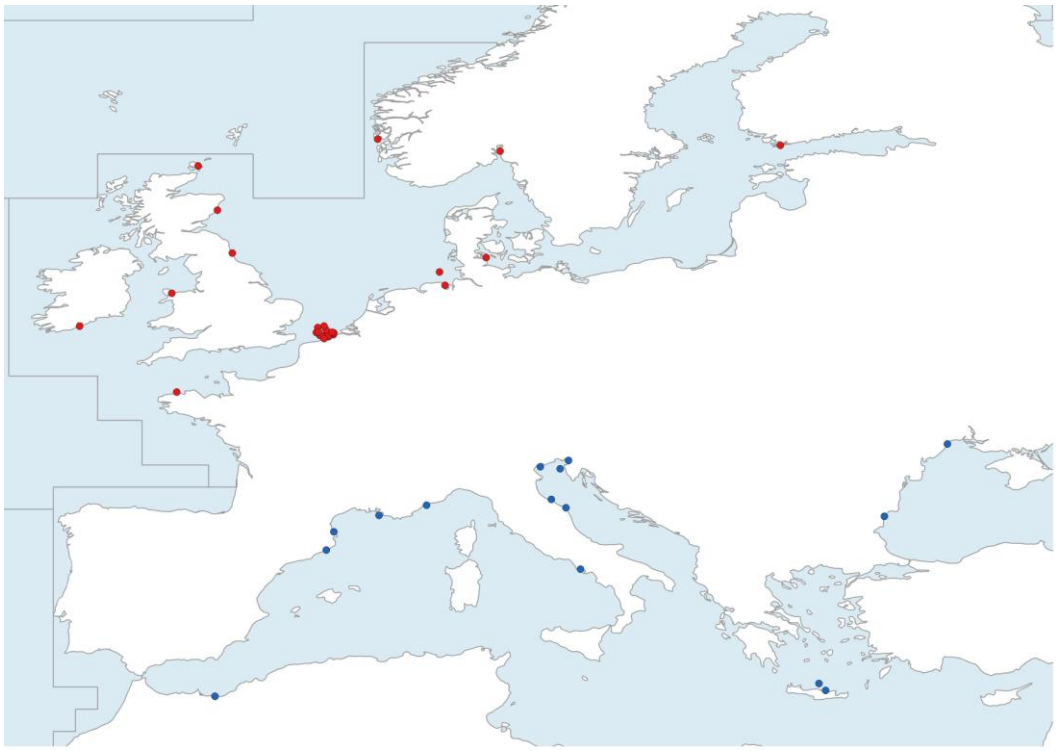

Fig. 16: Clusters of NECS (red) and MEDI (blue) ecoregions.

two NECS and MEDI ecoregions. These include the 45% of the analyzed stations and therefore correspond to the area with the highest sampling density. This choice was also supported by the ordination results previously shown, in which the stations belonging to the two ecoregions are in distinctly separate groups on the plane. Furthermore, within this subset, only the stations belonging to clusters A and B have been taken into account, because the other represent those special

cases previously mentioned. Therefore, for the subsequent analyzes, 43 stations were considered, 16 belonging to group A and 27 to B (Fig. 16) and *taxa* absent in both stations have been eliminated from the dataset, keeping only 157 out of 187.

First, SIMPER analysis was carried out on the two groups of stations, to see which *taxa* contributes the most to determine the differences between the two ecoregions. Those responsible for 95% of these differences are 14. Specifically Guinardia (29.06%), Rhizosolenia (22.09%), Alexandrium (16.85%), Ostreococcus (6.73%), Heterocapsa (6.09 %), Bathycoccus (2%), Gymnodinium (1%), Protoperidinium (1%), Minutocellus (1%), Noctilucales (1%), Minidiscus (1%), with a greater number of reeds in group B, and Cyclotella (3.75%), Tetraselmis (2) and Halostylodinium (1%) with a greater number of reeds in group A.
Overall, 61.8% of *taxa* were present in both ecoregions, 22.9% in group B only and 15.3% in group A only.

Subsequently, the *taxa* frequencies in the two different groups of stations were compared, therefore the *taxa* present only in one of the two groups were eliminated from the dataset, keeping only that 61.8% present in both, so as to have two matrices with the same number of lines (objects) that could be matched. The number of *taxa* taken into account has therefore been reduced to 97.

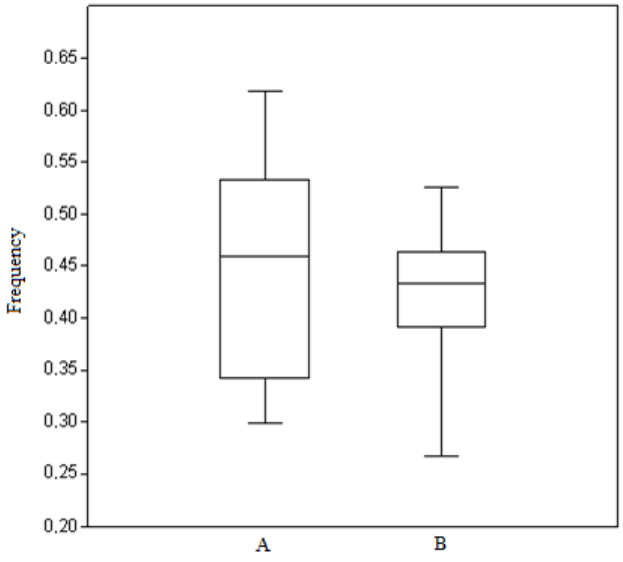

Fig. 17: Box plot: frequency of taxa *in group A and B.*

To estimate the biodiversity of each station, starting from qualitative data, the percentage of *taxa* present compared to the total of those examined was taken into consideration. The result can be displayed in the box plot in Fig. 17.

Median values are close, but interquartile range is higher in group A then in group B.

Frequencies of the different *taxa* for the two different groups have been calculated and reported in the graph in Fig. 18, where for every *taxa* there is a point, whose coordinates are the frequency in group A on the X axis and that in group B on that of Y. Two proportion z-test was applied. Red point on the graph are *taxa* significantly more frequent in group B and blue points are significantly more frequent in group A.

Keeping the two groups separate, two association matrices between the different *taxa* were created, based on the Fager & McGowan coefficient.

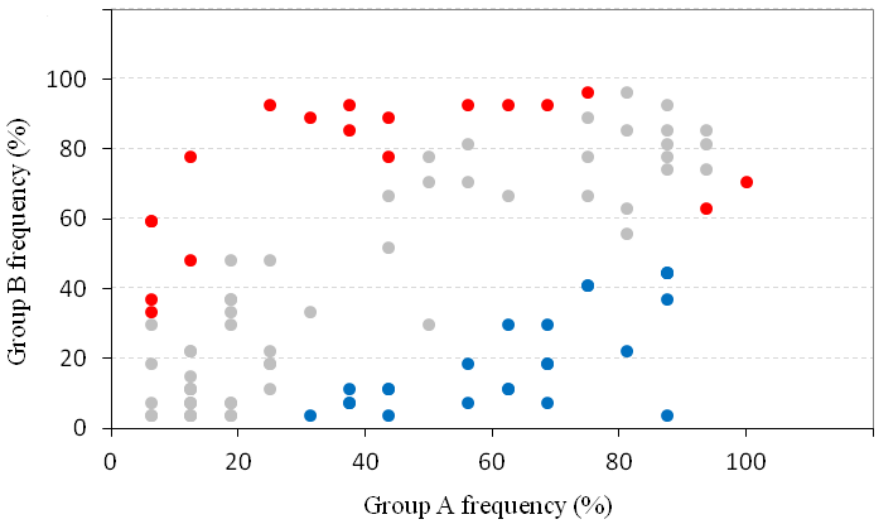

Fig. 18: Frequency of taxa *in group A (X axis) and group B (Y axis). Red point on the graph are taxa significantly more frequent in group B and blue points are significantly more frequent in group A.*

The two matrices were then compared to each other through Mantel test, to verify whether the tendency of certain *taxa* to appear together was linked to the region to which they belonged, or whether it was a more general trend. The result obtained (R = 0.4206 p = 0.0002) does not allow preserving the independence hypothesis,

thus highlighting a certain correlation between the two matrices. This suggests an independence of the association from the geographical area to which it belongs.

Later, the two matrices were used as the basis for the same number of ordination via PCoA (Fig. 19 and 20).

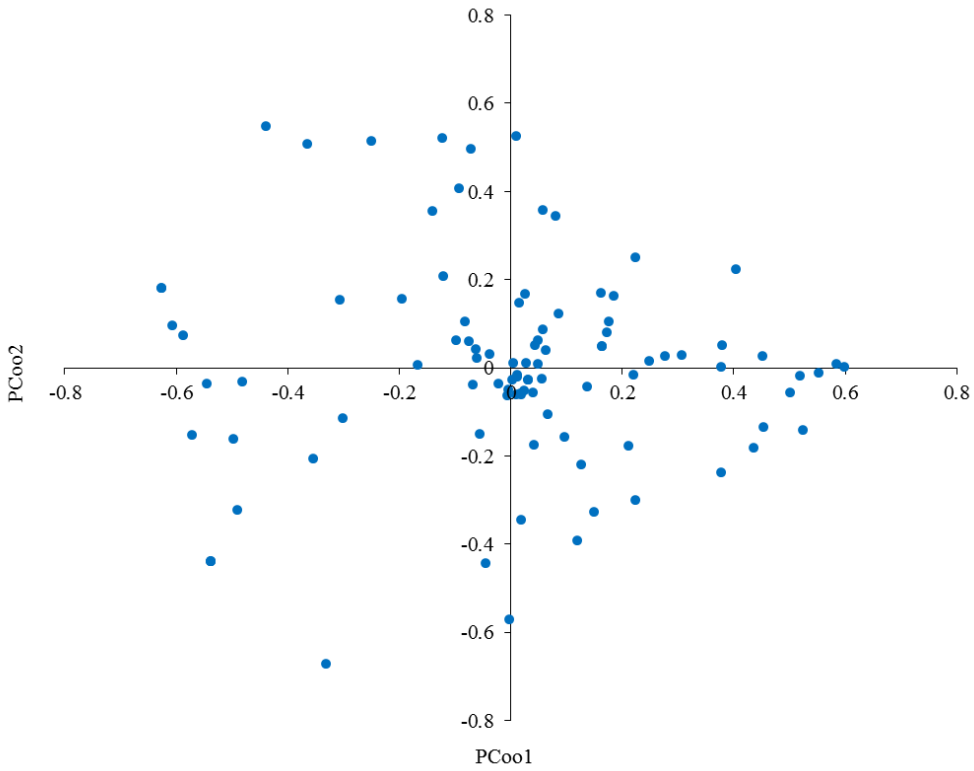

Fig 19: Ordination obtained through Principal Coordinate Analysis of group A taxa. *The ordination is based on the association matrix of Fager & McGowan and the first two axes explain 24.9% of the total variance.*

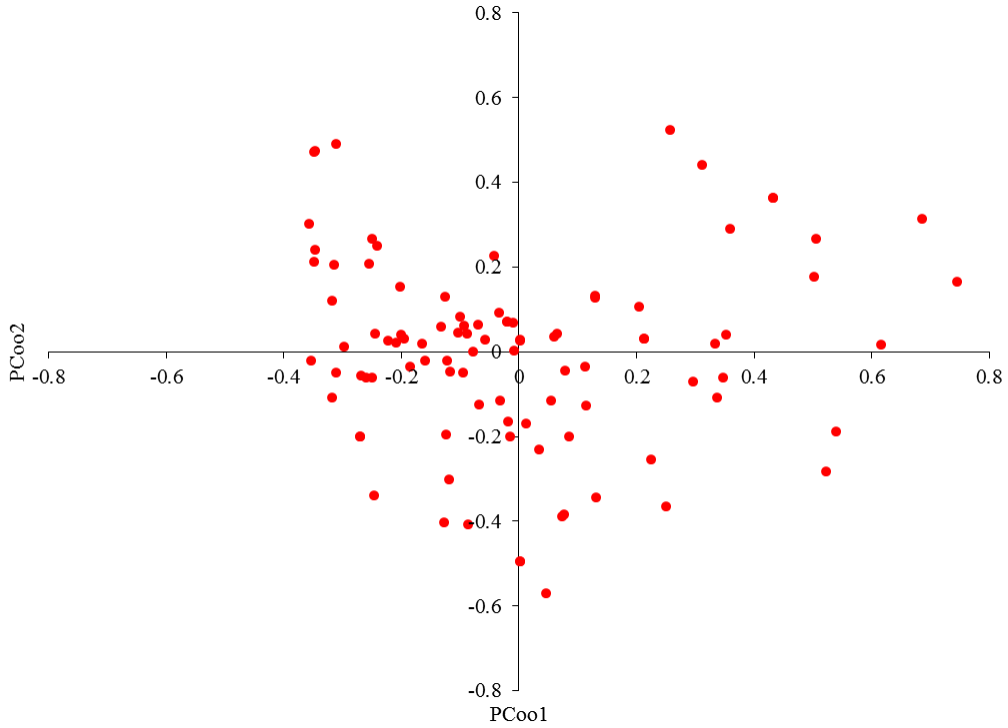

Fig. 20: *Ordination obtained through Principal Coordinate Analysis of group A* taxa. *The ordination is based on the association matrix of Fager & McGowan and the first two axes explain 20% of the total variance.*

PROTEST was used to compare the two ordinations, performed with R.

Results are shown in Fig. 21. Each arrow connects the two representations of each object (derived from the two different matrices). Arrow length provides information about the distance of the same object in the ordinations. Longer arrows correspond to greater distance, namely difference. As can be seen from the graph, values are all very small.

The null hypothesis of independent association matrices cannot be preserved, being the level of significance, $p < 0.0001$. This result confirms what was previously obtained with Mantel test, and suggests that, even if there are differences in *taxa* associations between the two groups, they only concern a relatively small number of *taxa* and that overall relationships remain coherent, therefore independent of the belonging ecoregion.

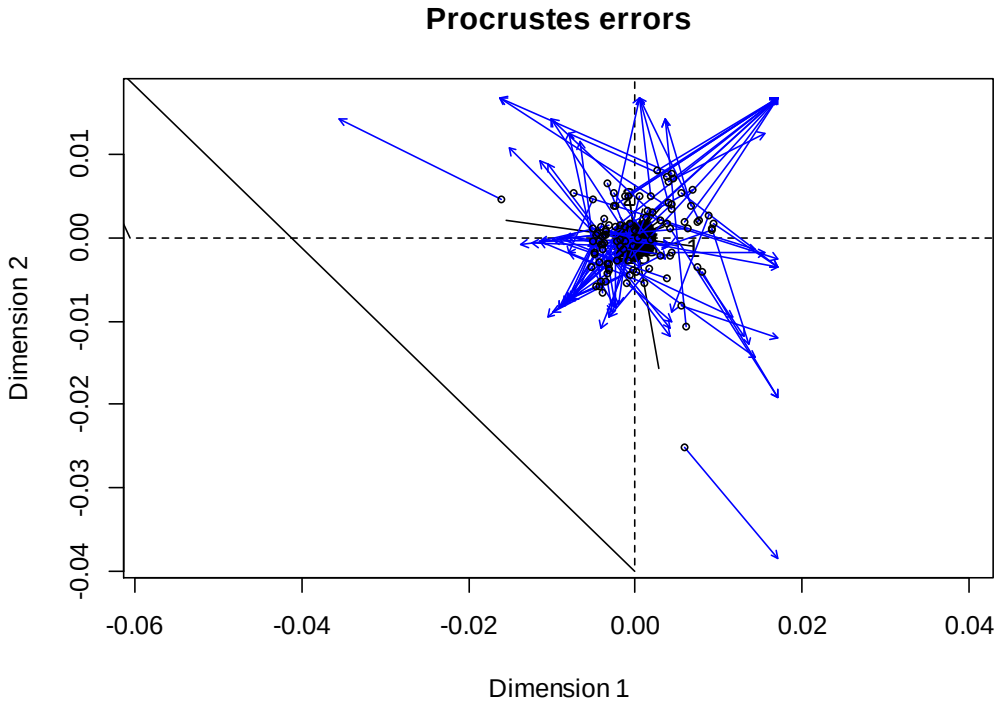

Fig. 21: Procrustes analysis of the two ordinations derived from the two different association matrices of group A and B.

4. Discussion and conclusions

The work carried out in these three years allowed me to venture into the various fields of data analysis and Machine Learning, Random Fortes technique in particular, applying them to the study of phytoplankton species, in an ecological perspective aimed at the study and the conservation of biodiversity and the prevention of HABs, with possible implications also in the aquaculture field.

Regarding the predictive model, the results obtained are very encouraging and, although the model was developed for a limited area, it has proven to respond well, managing to predict more than 80% of cases. An interesting result shows that the 12-variable model works as well as the 18-variable one and can therefore be used in the future without the need to perform laboratory analysis to know the concentrations of the different nutrients present in the water sample. The model

also lends itself to subsequent implementations, adding samples from other locations or taken in periods subsequent to those available so far. It is also interesting to note how the use of PCR is an innovative advantage, which allows a safer and faster recognition of the species, compared to the identification under a microscope and which allows it to be reasonably more confident of absence cases, a fundamental issue for the training of predictive models.

Along this line, there is also the idea of using molecular data to study phytoplankton biodiversity. Microscopy techniques are still very common for this purpose, but the various molecular methods are more and more used in this field. The metabarcode approach is a good alternative but, in order to express all of its potential, the gap between sequence reeds and ecologically meaningful entities must be filled. Grouping two or more *taxa* that are possibly ecologically different just smooths out the differences. This might be the case of our data, where the identification has been possible only at genus level. By doing so, some information may have been lost. Indeed, within some genera, there may be substantial differences between species. Just think of all those genera that contain toxic and non-toxic species.

Our results suggest that, although the phytoplankton populations are qualitatively significantly different between the two studied ecoregions, Mediterranean Sea and NE Atlantic Shelves Province, the association tendency of *taxa* present in both regions is, actually, similar. This could suggest that therefore it is not tied to the geographical area of origin, but rather to physical conditions of the microorganisms belonging to the different *taxa*.

OSD samples were taken in a single day, so it is not possible to do time series studies on them. The number of our samples, moreover, once the dataset has been cleaned of the most ambiguous cases, is not very high. However, an initial exploratory study was possible and, if the OSD campaign continues, more will be possible, with more data available. In any case, the innovation of the study also lies in the use of molecular techniques of species recognition, thanks to the numerous speed and precision advantages involved in the identification.

5. Acknowledgements

Data described herein are available at EBI under the Project ID PRJEB8682: Sequencing of amplicon and metagenome samples from the main OSD event, representing joint effort of marine sampling stations around the world.

We thank Max Planck Institute for MarineMicrobiology in Bremen and Biological Institute of the AlfredWegner Institute in Helgoland where DNA extraction took place; LGC Genomics (LGC Genomics GmbH, Berlin, Germany) for sequencing process including amplification of 18S rRNA gene.

6. References

Ade P., Funari E., Poletti R. (2003). Il rischio sanitario associato alle tossine di alghe marine. Ann. Ist. Super. Sanità. 39, 53-68.

Alpaydin, E. (2014). Introduction to machine learning, eds. (Cambridge: Massachusetts Institute of Technology Press).

Anderson, M.J. (2001a). A new method for non-parametric multivariate analysis of variance. Austral Ecology 26, 32-46.

Anderson, M.J. (2001b). Permutation tests for univariate or multivariate analysis of variance and regression. Canadian Journal of Fisheries and Aquatic Sciences 58, 626-639.

Anderson, D.M., Lobel, P.S. (1987). The continuing enigma of ciguatera. Biol. Bull. 172, 89-107.

Anderson, D. M., Cembella, A. D., Hallegraeff, G. M. (2012a). Progress in understanding harmful algal blooms (HABs): Paradigm shifts and new technologies for research, monitoring and management. Ann. Rev. Mar. Sci. 4, 143–176.

Anderson, D.M., Alpermann, T.J., Cembella, A.D., Collos, Y., Masseret, E., Montresor, M. (2012b). The globally distributed genus Alexandrium: Multifaceted roles in marine ecosystems and impacts on human health. Harmful Algae 14, 10–35.

Bacci, T., La Porta, B., Maggi, C., Nonnis, O., Paganelli, D., Sante Rende, F., Targusi, M. (2014). Manuale e Linee Guida, ISPRA, 106/2014, Conservazione e gestione della naturalità degli ecosistemi marino-costieri. Il trapianto delle praterie di Posidonia oceanica, eds. (Rome, Italy: ISPRA – Settore Editoria).

Behrenfeld, M.J., Randerson, J.T., McClain, C.R., Feldman, G.C., Los, S.O., Tucker, C.J., Falkowski, P.G., Field, C.B., Frouin, R., Esaias, W.E., Kolber, D.D., Pollack, N.H. (2001). Biospheric Primary Production During an ENSO Transition. Science 291, 2594-2597.

Berdalet, E., Fleming, L.E., Gowen, R., Davidson, K., Hess, P., Backer, L.C.,

Moore, S.K., Hoagland, P., Enevoldsen, H.(2015). Marine harmful algal blooms, human health and wellbeing: challenges and opportunities in the 21st century. J. Mar. Biolo. Ass. UK 96, 61–91.

Bolch, C.J., Blackburn, S.l., Cannon, J.A. Hallegraeff, G.M. (1991). The resting cyst of the red tide dinoflagellate *Alexandrium minutum* (Dinophyceae). Phycologia 30, 215-219.

Boni, L., Guerrini, F., Pistocchi, R., Cangini, M., Pompei, M., Cucchiari, E., Romagnoli, T., Totti, C. (2005). Microalghe tossiche del Medio e Alto Adriatico. Guida per acquacoltori e operatori sanitari. Casa editrice Fernandel - Ravenna.

Boudouresque, C. F., Bernard, G., Bonhomme, P., Charbonnel, E., Diviacco, G., Meinesz, A., Pergent, G., Pergent-Martini, C., Ruitton, S., Tunesi, L. (2012). Protection and conservation of Posidonia oceanica meadows. Ramoge and RAC/SPA publisher, Tunis, 1-202.

Breiman, L. (2001). Random Forests. Mach. Learn. 45, 5–32.

Bricker, S.B., Ferreira, J.G., Simas, T. (2003). An integrated methodology for assessment of estuarine trophic status. Ecol. Model. 60, 169–39.

Chang, F.H., MacKenzie, L., Till, D., Hannah, D., Rhodes L.(1995). The first toxic shellfish outbreaks and the associated phytoplankton blooms in early 1993 in New Zealand. In: Lassus P., Arzul G., Erard E., Gentien P. & Marcallou C. (eds.), Harmful Marine Algal Blooms, Lavoisier, 145-150.

Ciminiello, P., Dell'Aversano, C., Dello Iacovo, E., Forino, M., Tartaglione, L. (2015). Liquid chromatography–high-resolution mass spectrometry for palytoxins in mussels. Anal. Bioanal. Chem. 407, 1463–1473.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educ. Psychol. Meas. 20(1), 37-46.

Cutler, D.R., Edwards Jr, T.C., Beard, K.H., Cutler, A., Hess, K.T., Gibson, J., Lawler, J.J. (2007). Random forests for classification in ecology. Ecology, 88(11), 2783-2792.

Fager, E.W., McGowan, J.A. (1963). Zooplankton species groups in the North Pacific. Science (Wash. D.C.) 140, 453-460.

Falkowski, P.G., Raven, J.A. (2007). Aquatic Photosynthesis. New Jersey: Princeton University Press. 484.

Falkowski, P.G., Barber, R.T., Smetacek, V. (1998). Biogeochemical Controls and Feedbacks on Ocean Primary Production. Science 281, 200-206.

Field, C.B., Behrenfeld, M.J., Randerson, J.T., Falkowski, P. (1998). Primary Production of the Biosphere: Integrating Terrestrial and Oceanic Components. Science 281, 237-240.

Gabriel, K.R., Sokal, R.R. (1969). A new statistical approach to geographic variation analysis. Syst. Zool 18, 259–278.

Gallitelli, M., Ungaro, N., Addante, L. M., Gentiloni Silver, N., Sabbà, C. (2005). Respiratory illness as a reaction to tropical algal blooms occurring in a temperate climate. J. Am. Med. Assoc. 293, 2599–2600.

Garcés, E., Zingone, A., Montresor, M., Reguera, B., Dale, B. (eds.), (2002). LIFEHAB: life histories of microalgal species causing harmful blooms. Office for the Official Publications of the European Communities, Luxembourg.

Giacobbe, M.G., Maimone, G. (1994). First report of *Alexandrium minutum* Halim in a Mediterranean Lagoon. Cryptogamie Algol. 15,47–52.

Gower, J.C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika 53, 325-338.

Gran, H.H., (1912). Pelagic plant life, Ch. VI, in: "The Depths of the Ocean," J. Murray, and J. Hjort, eds., McMillan, London.

Guisan, A., Zimmermann, N.E. (2000). Predictive habitat distribution models in ecology. Ecol. Model. 135, 147-186.

Hallegraeff, G.M. (1993). A review of harmful algal blooms and their apparent global increase. Phycologia 32, 79–99.

Hallegraeff, G.M. (2010). Ocean climate change, phytoplankton community responses, and harmful algal blooms: a formidable predictive challenge. J. Phycol. 46, 220–235.

Hallegraeff, G.M., Anderson, D.M., Cembella, A.D. (eds.) (1995). Manual on Harmful Marine Microalgae, vol. 33. UNESCO.

Hammer, Ø., Harper, D.A.T., Ryan, P.D. (2001). PAST: Paleontological Statistics Software Package for Education and Data Analysis. Palaeontologia Electronica 4, 9. Free available at http://palaeo-electronica.org/2001_1/past/issue1_01.htm.

Hoagland, P., Scatasta, S. (2006). The economic effects of harmful algal blooms. Ecology of Harmful Algae. (eds E., Graneli & J. T., Turner) 391–401 (Springer-Verlag, Berlin).

Honsell, G. *et al.* (1996). *Alexandrium minutum* Halim and PSP contamination in the Northern Adriatic Sea (Mediterranean Sea). Harmful and Toxic Algal Blooms. (eds T., Yasumoto, T., Oshima & Y. T., Fukuyo) 77–83 (UNESCO, Paris).

Huertas, I.E., Rouco, M., López-Rodas, V., Costas, E. (2011). Warming will affect phytoplankton differently: evidence through a mechanistic approach. Proc. Biol. Sci. 278, 3534.

Hurley, J.R., Cattell, R.B. (1962). The Procrustes program: Producing direct rotation to test a hypothesized factor structure. Behavioral science 7, 258-262.

Jaccard, P. (1900). Contribution au problème de l'immigration postglaciaire de la flore alpine. Bull. Soc. vaudoise Sci. nat. 36, 87-130.

Jaccard, P. (1901). Etude comparative de la distribution florale dans une portion des Alpes et du Jura. Bull. Soc. vaudoise Sci. nat. 37, 547-579.

Jaccard, P. (1908). Nouvelles recherches sur la distribution florale. Bull.Soc. vaudoise Sci. nat. 44, 223-270.

Jackson, D.A. (1995). PROTEST: a PROcrustean randomization TEST of community environment concordance. Ecoscience 2, 297-303.

Kleindinst, J.L., Anderson, D.M., McGillicuddy D.J.Jr., Stumpf, R.P., Fisher, K.M., Couture, D.A., Hickey, J.M., Nash, C. (2014). Categorizing the severity of paralytic shellfish poisoning outbreaks in the Gulf of Maine for forecasting and management. Deep-Sea Res. II 103, 277–287.

Kopf, A. *et al.* (2015). The ocean sampling day consortium. GigaScience 4, 27.

Kudela, R.M., Gobler, C.J. (2012). Harmful dinoflagellate blooms caused by

Cochlodinium sp.: global expansion and ecological strategies facilitating bloom formation. Harmful Algae 14, 71–86.

Legendre, L., Legendre, P. (1983). Numerical ecology. Elsevier, Amsterdam.

Longhurst, A. (1998). Ecological Geography of the Sea. San Diego, Academic Press.

Manel, S., Dias, J. M., Ormerod, S.J. (1999). Comparing discriminant analysis, neural networks and logistic regression for predicting species distributions: a case study with a Himalayan river bird. Ecol. Model. 120(2), 337-347.

Manel, S., Williams, H.C., Ormerod, S.J. (2001). Evaluating presence–absence models in ecology: the need to account for prevalence. J. Appl. Ecol. 38(5), 921-931.

Mantel, N. (1967). The detection of desease clustering and a generalized regression approach. Cancer Res. 27, 209-220.

Margalef, R. (1974). Ecologia. Barcelona.

McArdle, B.H., Anderson, M.J. (2001). Fitting multivariate models to community data: a comment on distance based redundancy analysis. Ecology 82, 290-297.

Nehring, S. (1994). Spatial distribution of dinoflagellate resting cysts in Recent sediments of Kiel Bight, Germany (Baltic Sea). Ophelia 39, 137-158.

Omary, Z., Mtenzi, F. (2010). Machine learning approach to identifying the dataset threshold for the performance estimators in supervised learning. Int. J. Infon. 3, 314-325.

Pauly, D., Christensen, V., Dalsgaard, J., Froese, R., and Torres, F. Jr. (1998). Fishing down marine food webs. Science 279, 860–863.

Penna, A. *et al.* (2015). The sxt gene and paralytic shellfish poisoning toxins as markers for the monitoring of toxic Alexandrium species blooms. Environ. Sci. Technol. 49, 14230–14238.

Peres-Neto, P.D., Jackson, D.A. (2001). How well do multivariate data sets match? The advantages of a Procrustean superimposition approach over the Mantel test. Oecologia 129, 169–178.

Perini, F., Galluzzi, L., Dell'Aversano, C., Dello Iacovo, E., Tartaglione, L., Ricci,

F., Forino, M., Ciminiello, P., Penna, A. (2014). SxtA and sxtG gene expression and toxin production in the Mediterranean *Alexandrium minutum* (Dinophyceae). Mar. Drugs 12, 5258–5276.

Peters, J., De Baets, B., Verhoest, N.E., Samson, R., Degroeve, S., De Becker, P., Huybrechts, W. (2007). Random forests as a tool for ecohydrological distribution modelling. Ecol. Model. 207(2), 304-318.

Poletti, R., Milandri, A., Pompei, M. (2003). Algal biotoxins of marine origin: new indication from the European Union. Veterinary research communication. 27, 173-182.

Prieto-Martínez, F., Arciniega, M., Medina-Franco, J. (2018). Molecular docking: current advances and challenges. TIP Revista Especializada en Ciencias Químico-Biológicas. 21.

Salmaso, N., Tolotti, M. (2009). Other phytoflagellates and groups of lesser importance. In: Likens GE (ed) Encyclopedia of inland waters. Elsevier, Oxford, pp 174–183.

Sanger, F., Coulson, A.R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol. 25, 441–448.

Seymour,J.R., (2014). A sea of microbes: the diversity and activity of marine microorganisms. Microbiology Australia

Smayda, T.J. (1989). Primary production and the global epidemic of phytoplankton blooms in the sea: a linkage? In: E.M. Cosper, V.M. Bricelj, E.J. Carpenter (eds.), Novel phytoplankton blooms, vol. 35. Berlin: Springer, pp. 449-483.

Smayda, T.J., Reynolds, C.S. (2001). Community assembly in marine phytoplankton: application of recent models to harmful dinoflagellate blooms. J. Plankton Res. 23, 447–461.

Smith, C. J. (2012). Diagnostic tests: sensitivity and specificity. Phlebology, 27, 250-251.

Spector, D.L. (1984). Dinoflagellates. Academic Press, Inc., New York. 545 pp.

Steidinger, K. A. (1983). A re-evaluation of toxic dinoflagellate biology and

ecology, p. 147-188. In F.E. Round and D.J. Chapman (ed.), Progress in phycological research, vol. 2. Elsevier Science Publishing, Inc., New York.

Steidinger, K. A., Baden, D.G. (1984). Toxic marine dinoflagellates. P. 201-261. In D.L. Spector (Ed.), Dinoflagellates. Academic Press, Inc., New York.

Steidinger, K.A., Tangen, K. (1997). Dinoflagellates. In: C.R. Tomas (ed.), Identifying marine phytoplankton. St. Petersburg, FL, USA: Academic, pp.387-584.

Sunday, J.M., Bates, A.E., Kearney, M.R., Colwell, R.K., Dulvy, N.K., Longino, J.T., Huey, R.B. (2014). Thermal-safety margins and the necessity of thermoregulatory behavior across latitude and elevation. PNAS 15, 5610-5615.

Swets, J.A. (1988). Measuring the accuracy of diagnostic systems. Science 240, 1285-1293.

Toyofuku, H. (2006). Joint FAO/WHO/IIOC activities to provide scientific advice on marine biotoxins (research report). Marine Pollution Bulletin. 52, 1735-1745.

Tubaro, A., Hungerford, J. (2007). Toxicology of marine toxins. Veterinary Toxicology 60, 725-752.

Vila, M., Camp, J., Garcés, E., Masó, M., Delgado, M. (2001). High resolution spatio-temporal detection of potentially harmful dinoflagellates in confined waters of the NW Mediterranean. J. Plankton Res. 23, 497–514.

Vila, M., Garcés, E., Masò, M. (2001). Potentially toxic epiphytic dinoflagellate assemblages on macroalgae in the NW Mediterranean. Aquat. Microb. Ecol. 26, 51-60.

Wells, M.L., Trainer, V.L., Smayda, T.J., Karlson, B.S.O., Trick, C.G., Kudela, R.M., Ishikawa, A., Bernard, S., Wulff, A., Anderson, D.M., Cochlank, W.P. (2015). Harmful algal blooms and climate change: learning from the past and present to forecast the future. Harmful Algae 49, 68–93.

Wiese, M., D'Agostino, P.M., Mihali, T.K., Moffitt, M.C., Neilan, B.A. (2010). Neurotoxic alkaloids: Saxitoxin and its analogs. Mar. Drugs 8, 2185–2211.

Zingone, A., Enevoldsen, H.O. (2000). The diversity of harmful algal blooms: a challenge for science and management. Ocean and Coastal Management 43, 725–748.

Zweig, M.H., Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clinical chemistry 39(4), 561-577.

7. Attached manuscript

Title: A model predicting the PSP toxic dinoflagellate *Alexandrium minutum* occurrence in the coastal waters of the NW Adriatic Sea.

Authors: Eleonora Valbi, Fabio Ricci, Samuela Capellacci, Silvia Casabianca, Michele Scardi, Antonella Penna.

Journal: Scientific Reports.

Published: 12 March 2019.

SCIENTIFIC REPORTS

OPEN

A model predicting the PSP toxic dinoflagellate *Alexandrium minutum* occurrence in the coastal waters of the NW Adriatic Sea

Received: 6 September 2018
Accepted: 15 February 2019
Published online: 12 March 2019

Eleonora Valbi[1,3], Fabio Ricci[1,3], Samuela Capellacci[1,3], Silvia Casabianca[1,3], Michele Scardi[2,3] & Antonella Penna[1,3,4]

Increased anthropic pressure on the coastal zones of the Mediterranean Sea caused an enrichment in nutrients, promoting microalgal proliferation. Among those organisms, some species, such as the dinoflagellate *Alexandrium minutum*, can produce neurotoxins. Toxic blooms can cause serious impacts to human health, marine environment and economic maritime activities at coastal sites. A mathematical model predicting the presence of *A. minutum* in coastal waters of the NW Adriatic Sea was developed using a Random Forest (RF), which is a Machine Learning technique, trained with molecular data of *A. minutum* occurrence obtained by molecular PCR assay. The model is able to correctly predict more than 80% of the instances in the test data set. Our results showed that predictive models may play a useful role in the study of Harmful Algal Blooms (HAB).

Anthropic pressures, highly increased in recent decades, have strong impact along the coasts of the Mediterranean Sea. Among the consequences, there are eutrophication, a nutrient over-enrichment of coastal waters (especially due to the massive use of fertilizers in agriculture), transport of phytoplankton species via ballast-water vessels and translocation of shellfish stocks[1–4]. In particular, eutrophication is increasing due to increased population, increased use of fertilizers both for terrestrial and marine animal farm practices and increased fossil fuel use[5]. These phenomena can favor a fast proliferation of microalgal species, known as algal bloom[6,7]. Further, climate change seems having effects on the frequency and abundance of algal blooms due to the complex of altered environmental factors[8,9].

Some microalgal taxa, such as dinoflagellates, can both originate high density biomass proliferation or blooms and produce a variety of toxin compounds that can accumulate along the trophic web through biomagnification process. Such blooms are known as Harmful Algal Blooms (HABs) and they can cause very serious damages to human health and marine organisms[10]. People can be affected either by breathing aerosols[11–13] or by eating vector species, such as mussels, clams and oysters[14,15], which can accumulate high concentrations of toxins in their digestive glands. HABs can cause also fish kills or hypoxia or anoxia events due to algal biomass proliferation. Therefore, HABs phenomena, in addition to human health, are also concerned with fishing and aquaculture industry[16–18].

In recent years, there has been a significant increase of these HABs phenomena worldwide[19–22], including Mediterranean Sea[23,24]. Therefore, the HAB monitoring programs increased[4]. In the future, the next challenge will be the managing and forecasting of HABs[25]. The mathematical models are shown to be useful tools for this purpose and their use has grown in the last decades. The purpose of these models is to describe[26–29] or to forecast HABs providing a survey[30–32], in order to identify environmental, physical and chemical conditions in which the risk of algal blooms is higher and in which it can concentrate efforts, such as sampling frequency to confirm or discharge the predicted bloom. Methods used to build these models are numerical, mathematical, and statistical ones or artificial intelligence techniques, like Artificial Neural Network (ANN)[33,34] and other Machine Learning

[1]Department of Biomolecular Sciences, University of Urbino, Campus E. Mattei, Via Cà le Suore 2/4, 61029, Urbino (PU), Italy. [2]Department of Biology, University of Rome Tor Vergata, Via della Ricerca Scientifica 1, 00133, Rome, Italy. [3]CoNISMa, Consorzio Interuniversitario per le Scienze del Mare, Pz. Flaminio 9, 00196, Rome, Italy. [4]CNR–IRBIM, Largo Fiera della Pesca 1, 60125, Ancona, Italy. Correspondence and requests for materials should be addressed to A.P. (email: antonella.penna@uniurb.it)

(ML) techniques. Recknagel *et al.*[35] used ANN to predict algal blooms in four freshwater systems. In the northern Adriatic Sea, Volf *et al.*[36] used predictive model for the phytoplankton abundance. Only a few studies used predictive models for HABs in coastal waters: Asnaghi *et al.*[29] used a Quantile Regression Forest to predict the concentration of the toxic benthic dinoflagellate *Ostreopsis* cf. *ovata* in Ligurian Sea (North-western Mediterranean) and Kehoe *et al.*[37] used a Random Forest (RF) to build predictive models of benthic PAR (Photosynthetically Active Radiation) at two sites in Moreton Bay affected by *Lyngbia majuscula* blooms.

In order to develop predictive models, it is crucial to have information about the occurrence of the toxic phytoplankton species. Morphological identification and enumeration of toxic phytoplankton species are usually done by using microscopy methods, which are time-consuming and require taxonomic skills and highly-specialized personnel[38,39]. Moreover, in seawater samples, the target species may be present at very low concentrations, representing only a minor component in the phytoplankton assemblage, and it may risk to remain unnoticed, causing the so-called false negative cases. In addition, morphological identification often stops at genus level failing to discriminate between the various species[40].

Molecular PCR-based techniques have proven to be very useful tools for qualitative identification of microalgal species in coastal waters[41,42]. PCR methods can quickly detect even limited very low abundance of cells[43]. The process is also far more precise, because species-specific ribosomal DNA regions are amplified by using taxon-specific primers. This reduces the risk of inaccuracy, a fundamental condition for the activation of direct analysis that can enable more accurate diagnosis[44–46].

In the Mediterranean Sea, most productive areas, due to the nutrient discharged by numerous rivers, are mainly localized at the mouths of big rivers, among them, the Po River in the northern western Adriatic Sea[47,48]. These riverine discharges can generate eutrophication conditions that may lead to bloom events that can be originated by harmful microalgal species or species complex[49].

The dinoflagellate *Alexandrium minutum* Halim, 1960 is the most widespread toxic species in the western Mediterranean basin[50,51]. This species has been responsible for toxic blooms along the northwestern coast of the Adriatic Sea (Italy) and Ionian Sea, where mussel farms have been contaminated[52,53]. *A. minutum* can produce saxitoxins, GTX1 and 4, that can cause a severe human illness, the Paralytic Shellfish Poisoning (PSP) syndrome[15,54], the most widespread HAB-related shellfish poisoning illness[55]. In the Mediterranean Sea, the increase in the frequency of toxic *A. minutum* outbreaks and the number of areas affected has coincided with the overdevelopment of coastlines, which increasingly offer confined nutrient enriched waters suitable for microalgal proliferation[3,56]. Generally, nutrient rich waters are trigger for its blooming along coastal waters and the physical structure of mass water is critically for the bloom initiation, avoiding cell dispersion and assuring high nutrient levels. In shallow areas, such as coastal shoreline, beaches, bays, *A. minutum* occurs during spring in coincidence with higher temperature, enhanced rainfall and freshwater inputs, which could be related to the supply of macro- and micronutrients, and with stabilization of the water column[23,57]. Furthermore, despite the dinoflagellates' preference for settling in confined environments near shore, *A. minutum* has an enormous natural potential for dispersal because of its capacity to grow and produce resting cysts under a wide range of environmental conditions. This feature can be responsible of toxic bloom dispersion[58,59]. Saxitoxin production in *A. minutum* is difficult to be controlled. It is known that the production of STX in some *A. minutum* strains can be influenced by nutritional conditions. In particular, low levels of phosphorus increase it[60–63]. Moreover, grazer-induced toxin production has been shown in *A. minutum* under nutrient replete conditions[64]. Recently, it was found that *A. minutum* responds to pico- to nanomolar concentrations of copepodamides produced by zooplankton with up to a 20-fold increase in production of paralytic shellfish toxins[65]. The *A. minutum* abundance that can determine the toxic levels dangerous for humans and therefore, representing an alert is not known to date, because many variables can influence the contamination of shellfish filter animals (i.e. environmental parameters, cell concentration in the seawater, cellular toxin content); of course, the conditions of pre-bloom and bloom (10^5–10^6 cells/L) are supposed to be critical for an alert. But, anyway, the presence of *A. minutum* cells in the seawater can represent a potential for a bloom formation, and therefore, it is crucial both to predict and control its occurrence.

Furthermore, in the Adriatic Sea, the *Alexandrium* species that occur frequently are the toxic *A. minutum* together with no PSP producing *A. mediterraneum, A. pseudogonyaulax, A. tamutum* and *A. taylori*[66]. In some cases, light microscopy examination, which is the traditional method used in the monitoring activity, can't identify and distinguish exactly the morpho-type species, due to the similarity of morphology. Therefore, it is important having the tools, such as the molecular techniques to identify properly and rapidly the toxic species from the other no PSP producing *Alexandrium* species, and approach analysis to predict its occurrence.

In this study, we developed a model predicting the occurrence of *A. minutum* in the northern western Adriatic coastal water using a Random Forest (RF) (Breiman, 2001), a Machine Learning ensemble technique that combines many Classification Trees (CT). This technique is particularly effective to develop qualitative predictive models, especially when relationships among variables are unknown.

Methods

Study sites and sampling. A total of 187 surface seawater samples were collected, monthly, from June 2005 to December 2009 along the transects of the Foglia (43°56′0.55N; 12°56′0.18E) and Metauro (43°50′0.54N; 13°05′0.9E) rivers at 500 m and 3000 m (NW Adriatic Sea) from coastland. Seawater samples were collected at 0.5 m depth using polyethylene bottles, and frozen at −20 °C after filtration (0.45 μm nitrocellulose filters, Millipore, USA) until chemical analyses, or fixed with pure ethanol and stored at +4 °C for molecular determinations.

Molecular analysis and PCR assay. Molecular PCR analysis was applied both because *A. minutum* is difficult to distinguish from other species within the same genus, as it is characterized by minute details of its thecal plates[67] and because PCR analysis allows us to be fair more certain about the absence data.

Variables
Day
Distance from coastline (m)
Wind maximum speed (Km h^{-1})
Wind direction
Cloud cover (okta)
Water transparency (m)
Sea surface temperature (°C)
Salinity (PSU)
Dissolved oxygen (mg L^{-1})
Oxygen saturation (% sat.)
Chlorophyll *a* (μg L^{-1})
pH
N-NO_3 (μM L^{-1})
N-NO_2 (μM L^{-1})
N-NH_3 (μM L^{-1})
P-PO_4 (μM L^{-1})
Total P (μM L^{-1})
Si-SiO_2 (μM L^{-1})

Table 1. List of environmental parameters used in the training phase.

For DNA extraction a volume of 100 mL of surface seawater samples, was filtered through a 25 mm diameter Isopore membrane filters with a pore size of 3.0 μm (Merck Millipore, Billerica, MA, USA) under gentle vacuum to avoid cell disruption. The filters were placed in Eppendorf with 1.0 mL of 95% ethanol and stored at +4 °C. Cells were washed out from the filters with ethanol and collected by centrifugation at 12,500 rpm for 10 min at room temperature. Pellets were kept frozen at −80 °C until molecular analyses. Total genomic DNA was purified from pellets, using DNeasy Plant Mini Kit (Qiagen, Valencia, CA). DNA concentration and integrity were evaluated on 0.8% (w/v) agarose gel using serially diluted λ DNA standards (Thermo Fisher Scientific, Hanover Park, IL, USA) and a gel-doc apparatus (Bio-rad, Hercules, CA, USA).

Species-specific primers for the amplification of *A. minutum* ITS–5.8S rDNA region and PCR conditions were reported in Penna *et al.*[41]. The PCR products were resolved on 1.8% (w/v) agarose 1x TAE buffer gel and were visualized by standard ethidium bromide staining under UV light in a gel-doc apparatus (Bio-rad, Hercules, CA, USA).

Chemical-physical analysis. Dissolved oxygen, oxygen saturation, salinity, temperature and pH determinations were performed with a CTD probe (Idronaut mod. Ocean Seven 316). The transparency of the seawater column was approached by Secchi depth. Dissolved inorganic nutrients (N-NO_3, N-NO_2, N-NH_4, P-PO_4 and Si-SiO_2) and chlorophyll "*a*" were performed spectrophotometrically (Shimadzu mod. UV- 1700) on filtered water samples following the methods of Strickland and Parsons[68] and APHA AWWA WPCF[69], respectively. Total phosphorus (TP) was determined on unfiltered water samples according to the method of Valderrama[70].

Modelling procedure. Occurrence data (i.e. presence and absence records based on molecular evidence) were associated not only to oceanographic data, but also to other predictive variables, namely day of the year, distance from coastline and three meteorological variables (wind maximum speed, wind direction and cloud cover). Data about the latter variables were retrieved from SYNOP servers.

At first we associated *A. minutum* occurrence data with all the available predictive variables (Table 1) to train RFs. However, at a later stage we also trained a second RF, using only 12 out of the 18 available predictive variables. The reduced data set excluded information about nutrients to make any future use of the model easier, with no need for water sampling and laboratory analysis to determine nutrients concentrations.

Independently of the number of variables used to predict *A. minutum* occurrence, the available records were divided into two different subsets: one third of them was set aside and *a posteriori* used as test set to validate the model. The remaining data were used as a training set, i.e. to provide the information RFs need to grow.

To assign records to the two subsets (training and test), they were first stratified according to *A. minutum* occurrence (presence or absence). Then each resulting subset was sorted according to the day of the year in which samples were collected, as seasonality is a factor that highly influences the presence of *A. minutum*. Then, in each sequence of three records, one was randomly allocated to the test set and the other two to the training set, thus ensuring the homogeneity of the two subsets.

Using both 18 and 12 predictive variables we tested several RFs, each one with different features given by different combinations of three training parameters. These were: the number of trees in the RF (100, 250, 500 or 1000), the number of variables available at each split (3, 4, 5 or 6) and the minimum number of records in each terminal node, i.e. in each "leaf" (1 to 10).

In RFs the overall output is obtained by collecting the output of each tree for each records. In other words, each tree "votes" for one of the possible states of the target variable and the majority wins. In theory, predicting *A. minutum* presence would need 50% + 1 presence predictions from all the trees in the RF. However, especially

Training set		Predicted values	
		presence	absence
Observed values	presence	43	3
	absence	24	55
	CCI% = 78.4		
	K = 0.58		

Test set		Predicted values	
		presence	absence
Observed values:	presence	21	1
	absence	8	32
	CCI% = 85.5		
	K = 0.70		

Table 2. Confusion matrices for 18-variables new RF, after cut- off optimization (t = 0.310).

when the numbers of presence and absence records are not well balanced, the optimal cut-off value for a successful presence prediction can be different. For instance, a RF could be more accurate if it were allowed to predict *A. minutum* presence even when less than 50% of the trees predict that output. In order to optimize the cut-off value to be used instead of 50%, the ROC (Receiver Operating Characteristic) curve[71] was analyzed to look for the best compromise between true positives and false positives in RF predictions. This way the optimal cut-off value, i.e. the minimum number of presence predictions from the trees in the RF that was needed to issue a presence prediction from the whole RF was found for all the RFs we trained. This procedure was especially necessary because the numbers of presence and absence records were not well balanced in our data set (68 presence and 119 absence records, respectively). As absence records were almost twice as much as those of presence of *A. minutum*, the RF training was slightly biased towards the first case, i.e. to the prediction of absence. Therefore, the optimal cut-off was expected to be smaller than 50% of the votes from the trees, i.e. smaller than 0.5. The ROC curve analysis also provided an AUC (Area Under the Curve) value, that can be regarded as a measure of overall model accuracy. However, in order to select the best model among those we developed with different sets of training parameters, we relied upon the Cohen's K statistics[72].

Results and Discussion

Using all the available predictive variables and different combinations of training parameters (number of trees, number of variables per split and minimum number of records per leaf) we trained 160 RFs. The optimal cut-off value for each RF, i.e. the one that maximized the true positive to false positive ratio, was obtained from the ROC curve analysis. After cut-off optimization, Cohen's K values were calculated for the test set. They ranged from 0.54 to 0.7, with a median value of 0.64 and, as expected, they tended to be inversely proportional to the minimum number of records per leaf. As the best candidate for optimal predictive performance we selected the best model out of the 160 we trained, i.e. we chose the one with the largest K value. The optimal RF model was the one with 100 trees, 3 predictive variables selected at each split and fully-grown trees, with only a single record in each leaf. The latter criterion, by the way, is the default option in the original implementation of the RF[73]. The optimized cut-off value for that RF was 0.31 and K values were 0.58 for the training set and 0.7 for the test set, while the ROC curve analysis returned a 0.895 value for the training set and a 0.88 AUC value for the test set. The K values relative to the test set indicated a substantial[74] to good agreement[75], whereas the AUC testified an excellent performance of the RF model according to Hosmer and Lemeshow[76]. Table 2 showed the confusion matrices for training and test sets as well as K values and the percentage of Correctly Classified Instances (CCI%), which is another index of the accuracy of the model, even though not as robust as Cohen's K in the evaluation of unbalanced data set. CCI% ranged from 78.4 to 85.5, respectively for the training and test set.

Nutrient concentrations are often available in coastal monitoring data, but their acquisition requires the collection of water samples and lab analyses, whereas data about all the other predictive variables can be retrieved from meteorological records or from *in situ* measurements obtained from multiparameter probes. Therefore, we trained more RFs using only 12 predictive variables, i.e. excluding nutrient concentrations. As for the previous RF, we tested several combinations of the training parameters, thus obtaining 160 different RFs. After cut-off optimization K values ranged from 0.51 to 0.7, with a median value of 0.62. As for the RF based on 18 predictive variables, K values were mainly influenced by the minimum number of records in RF leaves, although to a larger extent. The model with the best predictive ability was based on 1000 trees, using only 2 candidate variables at each split and fully-grown trees. The optimized cut-off value for the best RF was 0.361 and K values for training and test set were, respectively, 0.59 and 0.7. While the interpretation of K values was exactly the same as in the RF based on 18 predictive variables, the AUC values were 0.891 for the training set and 0.905 for the test set. AUC value for the test set, in this case, was a bit larger than the value for the training set and it was also a bit larger than the value for the test set of the other model, indicating an outstanding accuracy according to Hosmer & Lemeshow[76]. The confusion matrices for both the training and the test set were shown in Table 3, together with K values and CCI%, which, as in the previous case, were higher for the test set.

Comparing the two RFs, the one based on the full set of predictive variables was less dependent than the other one on the optimization of its training parameters, as shown in Fig. 1, where the central quartiles of the K values were narrower than those for the RF based on 12 predictive variables. Moreover, the median K value was larger (0.64 vs. 0.62) in the first case.

However, while using all the predictive variables allowed obtaining less variability depending on the RF training parameters, the best RF model obtained from the reduced set of predictive variables was as good as the best RF model obtained from the full set of predictive variables, if not marginally better (they're slightly better in the AUC value). Therefore, we have to consider nutrient concentrations as not strictly needed. As obtaining information about nutrients requires additional activities, with larger costs in time and money, we regard the model based

Training set		Predicted values: presence	Predicted values: absence	Test set		Predicted values: presence	Predicted values: absence
Observed values:	presence	39	7	Observed values:	presence	20	2
	absence	18	61		absence	7	33
CCI% = 80.0				CCI% = 85.5			
K = 0.59				K = 0.70			

Table 3. Confusion matrices for 12-variables new RF, after cut- off optimization (t = 0.361).

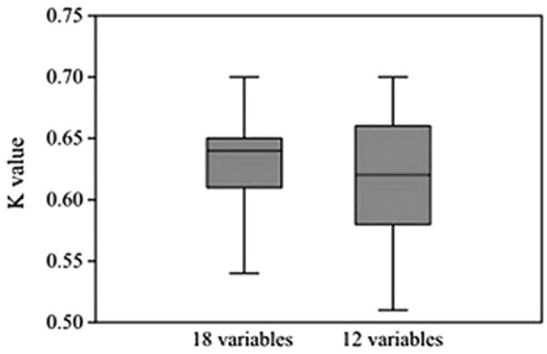

Figure 1. Box plot with K values distribution for all the models tested with different parameters combination. On the left, values of the 18-variables model: minimum value is 0.54, maximum is 0.7. Median value is 0.64. On the right, values of the 12-variables model: minimum value is 0.51, maximum is 0.7 and median value is 0.62.

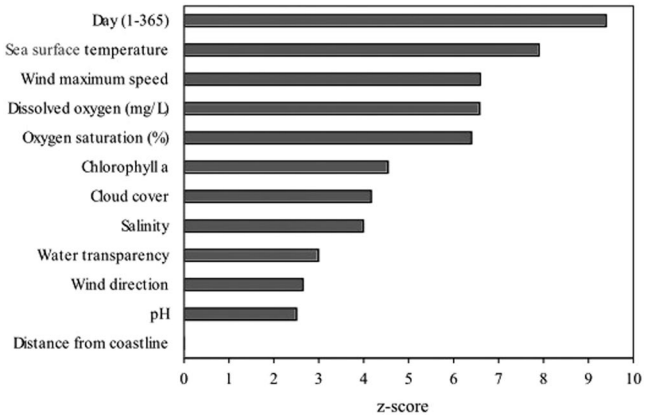

Figure 2. Plot with variable importance for 12-variables RF. The z-scores are obtained by dividing the raw scores by their standard error. All the bars are associated to significant z-scores except the one for distance from coastline, which is non significant and therefore was omitted.

on only 12 predictive variables as the best solution to use for making prediction in the future, not only because of its predictive ability, but also because of practical issues.

While the main drivers of any model can be identified thanks to sensitivity analysis, an interesting property of the RF algorithm is its ability to support an estimate of the relevance of the role played by each predictive variable. Relative importance of the 12 predictive variables used by the reduced RF model was shown in Fig. 2 as z-scores, computed according to the original algorithm proposed by Breiman[73].

As we expected, the day of the year, and therefore the period in which samples were collected, is the variable with the largest importance value, and therefore, the most correlated to *A. minutum* presence. In the studied period, the abundance of *A. minutum* was in the range of 10^3–10^5 cells/L (data not shown). Sea surface temperature (which is obviously not independent of day of the year, i.e. of season) was the second most important predictive variable, followed by wind maximum speed and oxygen concentration and saturation. Interactions between temperature, wind and oxygen concentration were obvious and certainly modulated by seasonal conditions in favoring *A. minutum* presence. The least important variable, according to the z-score obtained from the RF algorithm, was water pH, which was hardly connected, from a theoretical standpoint, to *A. minutum* presence and possibly affected by relatively large measurement errors.

The main goal of HABs management is to provide early warnings to prevent their impacts on public health and economical activities. Microscope identification of target species is a common procedure, although it requires a great deal of taxonomic expertise, in addition to being time consuming and impractical for processing a large number of samples in a monitoring perspective[38,39,77].

Recently, HABs phenomena are increasing in the Mediterranean Sea possibly under the influence of the coastal zone overdevelopment[24]. Climate change and global warming are now the main problems that may increase the risk of reaching critical conditions, especially in the Adriatic Sea. The latter is a very shallow sea and one of the most productive regions in the Mediterranean Sea, with nutrient inputs from riverine discharges[78] and where mussel farms, which play a relevant role in local as well as in Italian mariculture, have already been contaminated[52].

Results obtained in this study suggest that predictive models may be a valid supplementary tool in HABs management. In fact, they could be very useful to gain important information about those events and to identify the particular conditions in which HABs are more likely to occur, thus supporting the implementation of both new research efforts and activities focused on early reaction, whenever the event should occur.

While our models are already able to correctly predict more than 80% of the real-world instances, the RF approach will allow further improvement as soon as more records about *A. minutum* presence or absence will become available. Moreover, while our model was validated only locally, the same procedure can be applied to other sites or to several sites simultaneously. The ultimate goal, obviously, is a general model, trained and validated in a larger region or across the whole Mediterranean basin.

Conclusions

Modelling species distribution, both in space and in time, is usually easier when data about species occurrence are not affected by too many error sources. Undetected occurrences are a very common problem among those that may hinder species distribution models and they are more likely to happen than their positive counterpart, i.e. false occurrences, which may depend on species misidentification. While the first source of error depends on sampling design relative to species distribution, the second source only depends on the taxonomical skills supporting the modeler. As for studies on plankton species or assemblages, using molecular methods for species identification solves both problems, because false negatives and false positives are not likely to occur.

As a consequence, even a relatively small data set can support successful modelling if appropriate methods are selected for species identification. This is certainly the case with our study, because species occurrence data were obtained by molecular PCR analyses, which makes us especially confident about absence records. In fact, the latter can be regarded as real absence rather than as misidentification or undetected presence due to very low density of the target species. Confidence in species detection makes us also confident about the accuracy of our model.

This study was carried out for a single species over a relatively restricted area, but the selected approach can be easily applied elsewhere and at any spatial scale. Moreover, its methodological bases allow an easy application to the prediction of a wide range of different target species and this is the reason why RFs are rapidly becoming one of the most widely applied techniques in species-specific distribution modelling.

Our model allows to correctly classify more than 85% cases of presence or absence of *A. minutum*, with values of the K statistics as high as 0.7 for the test set. This result is certainly adequate for supporting an early warning that can be improved.

While the most common goal of any model is to provide accurate predictions, understanding the underlying ecological relationships is a very common secondary or even alternate objective. In our study, the focus was on the prediction of occurrence, but the importance of the predictive variables was assessed by means of the procedure based on the standardized errors in classification of out-of-bag records obtained from RF training. The assessment of the importance of each predictive variable is obviously based on the available data set only, which can be restricted to a limited number of environmental conditions or to limited sequence of events in a more complex time series. From a purely theoretical viewpoint, however, day of the year, sea surface temperature, wind maximum speed and oxygen concentration and saturation are very likely to be associated to conditions in which *A. minutum* is more frequently found. Needless to say, that association is a fact at local space and time scale and just a hypothesis to be tested at larger scale, as often happens when ecological inferences are based on real data sets.

Our model, however, will certainly play a role in predicting, and possibly better understanding, HABs, although it can only help to identify environmental conditions that might favor HABs, not the actual occurrence of those phenomena. As a matter of fact, we still do not have enough data as to try to understand and possibly modelling what triggers a HAB, but our model is certainly able to point out the conditions that are necessary, although not sufficient, to support that type of event. From this viewpoint, machine learning approaches seem particularly promising because they can be easily updated and optimized as soon as new data become available, thus providing useful support to human experts in HAB risk assessment.

Data Availability

The authors declare the data availability.

References

1. Hamer, J. P., Lucas, I. A. N. & McCollin, T. A. Harmful dinoflagellate resting cysts in ships' ballast tank sediments: potential for introduction into English and Welsh waters. *Phycologia* **40**, 246–255 (2001).
2. Heisler, J. *et al.* Eutrophication and harmful algal blooms: A scientific consensus. *Harmful Algae* **8**, 3–13 (2008).
3. Bravo, I. *et al.* Bloom dynamics and life cycle strategies of two toxic dinoflagellates in a coastal upwelling system (NW Iberian Peninsula). *Deep Sea Res. II* **57**, 222–234 (2010).
4. Anderson, D. M., Cembella, A. D. & Hallegraeff, G. M. Progress in understanding harmful algal blooms (HABs): Paradigm shifts and new technologies for research, monitoring and management. *Ann. Rev. Mar. Sci.* **4**, 143–176 (2012).

5. Glibert, P. M. *et al*. Vulnerability of coastal ecosystems to changes in harmful algal bloom distribution in response to climate change: Projections based on model analysis. *Glob. Chan. Biol.* **20**, 3845–3858 (2014).
6. Smayda, T. J. & Reynolds, C. S. Community assembly in marine phytoplankton: application of recent models to harmful dinoflagellate blooms. *J. Plankton Res.* **23**, 447–461 (2001).
7. Bricker, S. B., Ferreira, J. G. & Simas, T. An integrated methodology for assessment of estuarine trophic status. *Ecol. Model.* **60**, 169–39 (2003).
8. Hallegraeff, G. M. Ocean climate change, phytoplankton community responses, and harmful algal blooms: a formidable predictive challenge. *J. Phycol.* **46**, 220–235 (2010).
9. Fu, F. X., Tatters, A. O. & Hutchins, D. A. Global change and the future of harmful algal blooms in the ocean. *Mar. Ecol. Progr. Ser.* **470**, 207–23 (2012).
10. Hallegraeff, G. M. Harmful algal blooms: a global overview. Manual on Harmful Marine Microalgae. (eds G. M., Hallegraeff, D. M., Anderson & A. D. Cembella) 25–49 (UNESCO, Paris 2003).
11. Gallitelli, M., Ungaro, N., Addante, L. M., Gentiloni Silver, N. & Sabbà, C. Respiratory illness as a reaction to tropical algal blooms occurring in a temperate climate. *J. Am. Med. Assoc.* **293**, 2599–2600 (2005).
12. Casabianca, S. *et al*. Quantification of the toxic dinoflagellate *Ostreopsis spp*. by qPCR assay in marine aerosol. *Environ. Sci. Technol.* **47**, 3788–3795 (2013).
13. Ciminiello, P., Dell'Aversano, C., Dello Iacovo, E., Forino, M. & Tartaglione, L. Liquid chromatography–high-resolution mass spectrometry for palytoxins in mussels. *Anal. Bioanal. Chem.* **407**, 1463–1473 (2015).
14. Deeds, J. R., Landsberg, J. H., Etheridge, S. M., Pitcher, G. C. & Longan, S. W. Non-Traditional vectors for paralytic shellfish poisoning. *Mar. Drugs* **6**, 308–348 (2008).
15. Wiese, M., D'Agostino, P. M., Mihali, T. K., Moffitt, M. C. & Neilan, B. A. Neurotoxic alkaloids: Saxitoxin and its analogs. *Mar. Drugs* **8**, 2185–2211 (2010).
16. Hoagland, P. & Scatasta, S. The economic effects of harmful algal blooms. Ecology of Harmful Algae. (eds E., Graneli & J. T., Turner) 391–401 (Springer-Verlag, Berlin 2006).
17. Morgan, K. L., Larkin, S. L. & Adams, C. M. Firm-level economic effects of HABS: A tool for business loss assessment. *Harmful Algae* **8**, 212–218 (2009).
18. Berdalet, E. *et al*. Marine harmful algal blooms, human health and wellbeing: challenges and opportunities in the 21st century. *J. Mar. Biolo. Ass. UK* **96**, 61–91 (2015).
19. Kudela, R. M. & Gobler, C. J. Harmful dinoflagellate blooms caused by *Cochlodinium* sp.: global expansion and ecological strategies facilitating bloom formation. *Harmful Algae* **14**, 71–86 (2012).
20. Lewitus, A. J. *et al*. Harmful algal blooms along the North American west coast region: history trends, causes, and impacts. *Harmful Algae* **19**, 133–159 (2012).
21. Paerl, H. W. Mitigating harmful cyanobacterial blooms in a human-and climatically-impacted world. *Life* **4**, 988–1012 (2014).
22. Wells, M. L. *et al*. Harmful algal blooms and climate change: learning from the past and present to forecast the future. *Harmful Algae* **49**, 68–93 (2015).
23. Vila, M. *et al*. A comparative study on recurrent blooms of *Alexandrium minutum* in two Mediterranean coastal areas. *Harmful Algae* **4**, 673–695 (2005).
24. Garcés, E. & Camp, J. Habitat changes in the Mediterranean Sea and the consequences for Harmful Algal Blooms formation. Life in the Mediterranean Sea: A Look at Habitat Changes. (ed. Noga Stambler Israel) 519–541 (2012).
25. Kleindinst, J. L. *et al*. Categorizing the severity of paralytic shellfish poisoning outbreaks in the Gulf of Maine for forecasting and management. *Deep-Sea Res. II* **103**, 277–287 (2014).
26. Jeong, K. S., Kim, D. K., Whigham, P. & Joo, G. J. Modelling *Microcystis aeruginosa* bloom dynamics in the Nakdong River by means of evolutionary computation and statistical approach. *Ecol. Model.* **161**, 67–78 (2003).
27. Lee, J. H. W., Huang, Y., Dickman, M. & Jayawardena, A. W. Neural network modeling of coastal algal blooms. *Ecol. Model.* **159**, 179–201 (2003).
28. Wang, J., Tang, D. & Sui, Y. Winter phytoplankton bloom induced by subsurface upwelling and mixed layer entrainment southwest of Luzon Strait. *J. Mar. Syst.* **83**, 141–149 (2010).
29. Asnaghi, V. *et al*. A novel application of an adaptable modeling approach to the management of toxic microalgal bloom events in coastal areas. *Harmful Algae* **63**, 184–192 (2017).
30. Hamilton, G., McVinish, R. & Mengersen, K. Bayesian model averaging for harmful algal bloom prediction. *Ecol. Appl.* **19**, 1805–1814 (2009).
31. Anderson, C. R. *et al*. Predicting potentially toxigenic *Pseudo-nitzschia* blooms in the Chesapeake Bay. *J. Mar. Syst.* **83**, 127–140 (2010).
32. Blauw, A. N., Los, F. J., Huisman, J. & Peperzak, L. Nuisance foam events and *Phaeocystis globosa* blooms in Dutch coastal waters analyzed with fuzzy logic. *J. Mar. Syst.* **83**, 115–126 (2010).
33. Colasanti, R. L. Discussions of the possible use of neural network algorithms in ecological modelling. *Binary* **3**, 13–15 (1991).
34. Edwards, M. & Morse, D. R. The potential for computer-aided identification in biodiversity research. *Trends Ecol. Evol.* **10**, 153–158 (1995).
35. Recknagel, F., French, M., Harkonen, P. & Yabunaka, K. I. Artificial neural network approach for modelling and prediction of algal blooms. *Ecol. Model.* **96**, 11–28 (1997).
36. Volf, G., Atanasova, N., Kompare, B., Precali, R. & Oani, N. Descriptive and prediction models of phytoplankton in the northern Adriatic. *Ecol. Model.* **222**, 2502–2511 (2011).
37. Kehoe, M. *et al*. Random forest algorithm yields accurate quantitative prediction models of benthic light at intertidal sites affected by toxic *Lyngbya majuscula* blooms. *Harmful Algae* **19**, 46–52 (2012).
38. Smayda, T. J. Harmful algal blooms: their ecophysiology and general relevance to phytoplankton blooms in the sea. *Limnol. Oceanogr.* **42**, 1137–1153 (1997).
39. Penna, A. & Galluzzi, L. The quantitative real-time PCR applications in the monitoring of marine harmful algal bloom (HAB) species. *Environ. Sci. Poll. Res.* **20**, 6851–6862 (2013).
40. Godhe, A. *et al*. Intercalibration of classical and molecular techniques for identification of *Alexandrium fundyense* (Dinophyceae) and estimation of cell densities. *Harmful Algae* **6**, 56–72 (2007).
41. Penna, A. *et al*. Monitoring of HAB species in the Mediterranean Sea through molecular methods. *J. Plankton Res.* **29**, 19–38 (2007).
42. Battocchi, C. *et al*. Monitoring toxic microalgae *Ostreopsis* (dinoflagellate) species in coastal waters of the Mediterranean Sea using molecular PCR-based assay combined with light microscopy. *Mar. Pollut. Bull.* **60**, 1074–84 (2010).
43. Perini, F. *et al*. New approach using the real-time PCR method for estimation of the toxic marine dinoflagellate *Ostreopsis* cf. *ovata* in marine environment. *PLoS One* **6**(3), e17699 (2011).
44. Murray, S. A. *et al*. Differential accumulation of paralytic shellfish toxins from *Alexandrium minutum* in the pearl oyster, *Pinctada imbricata*. *Toxicon* **54**, 217–223 (2009).
45. Delaney, J. A., Ulrich, R. M. & Paul, J. H. Detection of the toxic marine diatom *Pseudo-nitzschia multiseries* using the RuBisCO small subunit (rbcS) gene in two real-time RNA amplification formats. *Harmful Algae* **11**, 54–64 (2011).
46. Pugliese, L., Casabianca, S., Perini, F., Andreoni, F. & Penna, A. A high-resolution melting method for the molecular identification of the potentially toxic diatom *Pseudo-nitzschia* spp. in the Mediterranean Sea. *Sci. Rep.* **7**, 4259 (2017).

47. Raicich, F. On the fresh water balance of the Adriatic Sea. *J. Mar. Syst.* **9**, 305–319 (1996).
48. Degobbis, D. *et al*. Long-term changes in the northern Adriatic ecosystem related to anthropogenic eutrophication. *Int. J. Environ. Poll.* **13**, 495e533 (2000).
49. Marić, D. *et al*. Phytoplankton response to climatic and anthropogenic influences in the north-eastern Adriatic during the last four decades. *Estuar. Coast. Shelf Sci.* **115**, 98–112 (2012).
50. Giacobbe, M. G. & Maimone, G. First report of *Alexandrium minutum* Halim in a Mediterranean Lagoon. *Cryptogamie Algol.* **15**, 47–52 (1994).
51. Vila, M., Camp, J., Garcés, E., Masó, M. & Delgado, M. High resolution spatio-temporal detection of potentially harmful dinoflagellates in confined waters of the NW Mediterranean. *J. Plankton Res.* **23**, 497–514 (2001).
52. Honsell, G. *et al*. *Alexandrium minutum* Halim and PSP contamination in the Northern Adriatic Sea (Mediterranean Sea). Harmful and Toxic Algal Blooms. (eds T., Yasumoto, T., Oshima & Y. T., Fukuyo) 77–83 (UNESCO, Paris 1996).
53. Penna, A. *et al*. The *sxt* gene and paralytic shellfish poisoning toxins as markers for the monitoring of toxic *Alexandrium* species blooms. *Environ. Sci. Technol.* **49**, 14230–14238 (2015).
54. Perini, F. *et al*. *SxtA* and *sxtG* gene expression and toxin production in the Mediterranean *Alexandrium minutum* (Dinophyceae). *Mar. Drugs* **12**, 5258–5276 (2014).
55. Anderson, D. M. *et al*. The globally distributed genus *Alexandrium*: Multifaceted roles in marine ecosystems and impacts on human health. *Harmful Algae* **14**, 10–35 (2012).
56. Bravo, I., Vila, M., Maso, M., Ramilo, I. & Figueroa, R. I. *Alexandrium catenella* and *Alexandrium minutum* blooms in the Mediterranean Sea: toward the identification of ecological niches. *Harmful Algae* **7**, 515–522 (2008).
57. Giacobbe, M. G., Oliva, F. D. & Maimone, G. Environmental factors and seasonal occurrence of the dinoflagellate *Alexandrium minutum*, a PSP potential producer in a Mediterranean lagoon. *Estuar. Coast. Shelf Sci.* **42**, 539–549 (1996).
58. Anglés, S., Garcés, E., René, A. & Sampedro, N. Life-cycle alternations in Alexandrium minutum natural populations from the NW Mediterranean Sea. *Harmful Algae* **16**, 1–11 (2012).
59. Anderson, D. M. *et al*. *Alexandrium fundyense* cysts in the Gulf of Maine: Long-term time series of abundance and distribution, and linkages to past and future blooms. *Deep Sea Res. II* **103**, 6–26 (2014).
60. Guisande, C., Frangópulos, M., Maneiro, I., Vergara, A. R. & Riveiro, I. Ecological advantages of toxin production by the dinoflagellate *Alexandrium minutum* under phosphorus limitation. *Mar Ecol Prog Ser* **225**, 169–176 (2002).
61. Lippemeier, S., Frampton, D. M. F., Blackburn, S. I., Geier, S. C. & Negri, A. P. Influence of phosphorus limitation on toxicity and photosynthesis of *Alexandrium minutum* (dinophyceae) monitored by in-line detection of variable chlorophyll fluorescence. *J. Phycol* **38**, 320–331 (2003).
62. Frangópulos, M., Guisande, C., deBlas, E. & Maneiro, I. Toxin production and competitive abilities under phosphorus limitation of *Alexandrium* species. *Harmful Algae* **3**, 131–139 (2004).
63. Touzet, N., Franco, J. M. & Raine, R. Influence of inorganic nutrition on growth and PSP toxin production of *Alexandrium minutum* (Dinophyceae) from Cork Harbour, Ireland. *Toxicon* **50**, 106–119 (2007).
64. Selander, E., Thor, P., Toth, G. B. & Pavia, H. Copepods induce paralytic shellfish toxin production in marine dinoflagellates. *Proc. R. Soc. Lond. Ser. B-Biol. Sci.* **273**, 1673–1680 (2006).
65. Selander, E. *et al*. Predator lipids induce paralytic shellfish toxins in bloom-forming algae. *Proc. Nat. Acad. Sci.* **112**, 6395–6400 (2015).
66. Penna, A. *et al*. Phylogenetic relationships among the Mediterranean *Alexandrium* (Dinophyceae) species based on sequences of 5.8 S gene and Internal Transcript Spacers of the rRNA operon. *Eur. J. Phycol.* **43**, 163–178 (2008).
67. Taylor, F. J. R. & Fukuyo, Y., Larsen, J. Taxonomy of harmful dinoflagellates. Manual of Harmful Microalgae. (eds G. M., Hallegraeff, D. M., Anderson & A. D. Cembella) 283–317 (IOC UNESCO, Paris 1995).
68. Strickland, J. D. H. & Parsons, T. R. A practical handbook of seawater analysis. *J. Fish. Res. Bd.* **167**, 49–89 (1972).
69. American Public Health Association, American Water Works Association, and Water Pollution Control Federation (APHA/AWWA/WPCF). Standard Methods for Water and Wastewater Treatment. (ed.16th APHA) 1067–1072 (Washington 1985).
70. Valderrama, J. C. The simultaneous analysis of total nitrogen and total phosphorus in natural waters. *Mar. Chem.* **10**, 109–122 (1981).
71. Zweig, M. H. & Campbell, G. Receiver-Operating Characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin. Chem.* **39**, 561–577 (1993).
72. Cohen, J. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* **20**, 37–46 (1960).
73. Breiman, L. Random Forests. *Mach. Learn.* **45**, 5–32 (2001).
74. Landis, J. R. & Koch, G. G. The measurement of observer agreement for categorical data. *Biometrics* **33**, 159–174 (1977).
75. Fleiss, J. L. Statistical methods for rates and proportions. (Wiley, New York 1981).
76. Hosmer, D. W. & Lemeshow, S. L. Applied Logistic Regression. (Wiley, New York 2000).
77. Sellner, K. G., Doucette, G. J. & Kirkpatrick, G. J. Harmful algal blooms: causes, impacts and detection. *J. Ind. Microbiol. Biotechnol.* **30**, 383–406 (2003).
78. Giani, M. *et al*. Recent changes in the marine ecosystems of the northern Adriatic Sea. *Estuar. Coast. Shelf Sci.* **115**, 1–13 (2012).

Acknowledgements

This research was supported by Regione Marche Project Coastal Monitoring n. 49 of 23/12/2013 of Table C. The monitoring and sampling carried out with Athena Vessel were also funded by the Department of Biomolecular Sciences and University of Urbino "Carlo Bo".

Author Contributions

E.V., M.S., A.P. contributed to the conception and design of the study; E.V. carried out the study. E.V. performed the statistical analyses. F.R. and S.C. carried out the chemical physical analysis. S.C. performed the molecular analyses. All authors were involved in the manuscript preparation and revision approval of the final version of the manuscript.

Additional Information

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.