Ecological Informatics 48 (2018) 158–170

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf

Seagrass detection in the mediterranean: A supervised learning approach

Dimitrios Effrosynidis[a,*], Avi Arampatzis[a], Georgios Sylaios[b]

[a] Database & Information Retrieval research unit, Department of Electrical & Computer Engineering, Democritus University of Thrace, Xanthi 67100, Greece
[b] Lab of Ecological Engineering & Technology, Department of Environmental Engineering, Democritus University of Thrace, Xanthi 67100, Greece

ARTICLE INFO

Keywords:
Seagrass classification
Dataset integration and fusion
Machine learning
Data mining
Mediterranean Sea

ABSTRACT

We deal with the problem of detecting seagrass presence/absence and distinguishing seagrass families in the Mediterranean via supervised learning methods. By merging datasets about seagrass presence and other external environmental variables, we develop suitable training data, enhanced by seagrass absence data algorithmically produced based on certain hypotheses. Experiments comparing several popular classification algorithms yield up to 93.4% accuracy in detecting seagrass presence. In a feature strength analysis, the most important variables determining presence–absence are found to be Chlorophyll-α levels and Distance-to-Coast. For determining family, variables cannot be easily singled out; several different variables seem to be of importance, with Chlorophyll-α surpassing all others. In both problems, tree-based classification algorithms perform better than others, with Random Forest being the most effective. Hidden preferences reveal that *Cymodocea* and *Posidonia* favor the low, limited-range chlorophyll-α levels ($< 0.5\ mg/m^3$), *Halophila* tolerates higher salinities (> 39), while *Ruppia* prefers euryhaline conditions (37.5–39).

1. Introduction

Environmental systems can rarely be studied adequately with traditional statistical analysis. A great part of the information gathered by environmental scientists often displays non-linearity, unusual distributions, missing values, and complex interactions between data (De'Ath, 2007; Guisan et al., 2002). Machine learning techniques have the capacity to discover hidden linear and non-linear patterns in such datasets, capturing the spatial and temporal peculiarities of each pattern (Kanevski et al., 2004).

The study of the impact of marine environmental conditions to the distribution of biological communities at macroscopic scales (e.g., covering the whole Mediterranean basin) could improve our understanding on the most critical physico-chemical factors controlling species presence-absence. It could also reveal hidden relations to species diversity and distribution, and the underlying community structures existing at particular habitats, serving as a guide to assess climate change effects. Wiley et al. (2003) modeled the hidden relations between marine environmental variables and eighteen marine fish species using a machine learning algorithm (Genetic Algorithm). Tittensor et al. (2009) applied maximum entropy modelling and environmental niche factor analysis methods to identify the environmental conditions favoring the global distribution of deep-sea habitats for stony corals. Similarly, Bentlage et al. (2009) employed the Genetic Algorithm for Rule Set Prediction (GARP) and a maximum entropy approach to describe the presence-only of chirodropid box-jellyfishes by combining their biogeographic distribution with remotely-sensed environmental datasets.

Seagrass beds are considered as highly productive ecosystems strongly related to nutrients biogeochemical cycling, carbon sequestration and food-web structure (Govers et al., 2014). Seagrass meadows serve as nursery grounds supporting coastal fisheries, filtering nutrients and entrapping sediments. The ecological modelling of seagrass distribution is particularly important for ecologists as seagrass species serve as valuable bio-indicators for aquatic ecosystem health assessment. For example, *Halophila minor* and *Halophila ovalis* act as bio-indicator for trace metals pollution and accumulation (Ahmad et al., 2015); *Zostera marina* leaf nitrogen to leaf mass ratio has been found to act as a consistent eutrophication indicator (Lee et al., 2004); *Cystoseira amentacea* and *Cystoseira mediterranea* have also been used as negative sentinel species for pollution (Ferrat et al., 2003), while many authors have noted a regression of *Posidonia oceanica* meadows according to the degree of human impact.

Several research papers have been published recently employing machine learning (ML) to marine environmental data. Some studies about marine ecosystems include the pioneering work of De'ath and Fabricius (2000) using classification and regression trees to analyze complex ecological data, leading to patterns between habitat types and

* Corresponding author.
E-mail addresses: deffrosy@ee.duth.gr (D. Effrosynidis), avi@ee.duth.gr (A. Arampatzis), gsylaios@env.duth.gr (G. Sylaios).

https://doi.org/10.1016/j.ecoinf.2018.09.004
Received 21 June 2018; Received in revised form 29 August 2018; Accepted 3 September 2018
Available online 06 September 2018

environmental variables; the application of ML to derive sponge species richness based on environmental predictors (Li et al., 2017); the application of regression-based ML for short-term prediction of phytoplankton concentration in Adriatic Sea (Volf et al., 2011); the implementation of Bayesian network models to describe the non-linear relationships of chlorophyll-α dynamics to environmental changes Alameddine et al., 2011); the employment of ML techniques to predict fish species richness, biomass, and diversity from a range of habitat variables (Knudby et al., 2010); and the development of ANNs to derive the impact of each environmental variable to the diversity indices of marine nematodes (Merckx et al., 2009).

A very important component of machine learning is model selection (also known as feature selection) and is mandatory in order to reach the best model from other alternative ones. Arthur et al. (2010); Li and Heap (2011) suggest that model selection is important for the popular random forest algorithm and thus, researchers have to focus to the most important variables.

An extensive work on seagrass distribution along the Mediterranean coast was conducted by Giannoulaki et al. (2013). A great number of morphodynamic, environmental and human impact variables were used to predict the presence–absence of *P. oceanica* seagrass species. Comparative tests were performed between the ML results when using the random forest and the maximum entropy algorithms. However, their dataset in terms of presence–absence seemed unbalanced (87.5% of total records signified *Posidonia* absence). Because of that, they modified the natural threshold of 0.5 that discriminates presence–absence incidents using the ROC optimization curve.

In this paper, we employ machine learning (ML) techniques to examine the presence–absence of seagrass meadows in the Mediterranean Sea, and the environmental relationship among seagrasses at family level. To achieve these, we combine data from a broad and diverse range of databases, such as EMODnet, UNEP, and CMEMS, aiming to determine the most appropriate variables affecting the distribution of seagrasses. We used static and temporal variables and chose the most important ones with variable importance method by the random forest algorithm. The temporal variables have additional features such as the values for each month, along with the year min, max and average for surface and seabed, totaling 217 variables. In order to perform binary classification we propose a method to automatically generate an absence dataset based on the presence dataset. For both binary and multi-class classification, 7 different classifiers are compared and their results are discussed.

The rest of this paper is organized as follows. In Section 2 we describe the datasets and variables that were used, as well as the absence dataset that we created. Section 3 briefly presents the machine learning algorithms, model selection technique, and evaluation measures employed. In Section 4 we conduct our experimental work for binary and multi-class classification, and in Section 5 we discuss the results. Finally, Section 6 summarizes our conclusions and gives directions for future work.

2. Materials and methods

2.1. Study site description

The Mediterranean Sea is a mid-latitude, predominantly oligotrophic to ultra-oligotrophic basin considered as the larger semi-enclosed sea on Earth. It is a sea almost completely enclosed by land, north of Africa and south of Europe, with limited connectivity with the Atlantic Ocean, through the narrow Strait of Gibraltar, the man-made connection with the Red Sea via the Suez Canal, and the smaller semi-enclosed Black Sea through the narrow Bosporus Strait. It expands from -17.29° to 36.29° in longitude and from 30.18° to 45.97° in latitude and has a surface of approximately 2,510,000 km^2. It is divided into two basins, the eastern and the western, with a boundary the Strait of Sicily. In this paper we focus on seagrass distribution, therefore at the coastal to continental shelf strip (0–200 m depth).

2.2. Dataset and variables

To understand the environmental, morphodynamic and morphological variables, and patterns governing the seagrass presence–absence and their distribution at family level, we combined data from a broad range of Mediterranean databases. The UNEP-WCMC global biodiversity standardized database (Weatherdon et al., 2015) was used in this study, focusing on the seagrasses of the biogenic habitat category.[1] The database comprises of a geo-referenced shapefile (WCMC-013-014) consisting of polygons and points, illustrating the global distribution of seagrass at species level, from which only the Mediterranean Sea records were retained as a subset (Fig. 1). This shapefile was imported into a Geographic Information System (QGIS). Based on this data, it occurs that seagrass covers most parts of the Mediterranean basin, distributed along the coast of Spain, France, Italy, Tunisia, Greece and Cyprus.

For each point in the dataset, a seagrass species and a seagrass family are reported. Seventeen points were unspecified; these records were removed from the dataset. As some species had limited representation in the dataset (less than 10 records), seagrass species were aggregated into the main seagrass families, as presented in Table 1.

Of all records of the UNEP-WCMC database for the Mediterranean Sea, Zosteraceae (mostly *Zostera noltii*) and Cymodoceaceae (mostly *Cymodocea nodosa*) are the most common and widespread seagrasses along Mediterranean coasts. Following Table 1, Cymodoceaceae is the dominant seagrass family in the Mediterranean Sea. It is a warm water species that prefers the climate of the Mediterranean. For instance, it does not extend further north than the southern coast of Portugal. Cymodoceaceae is capable of living in a range of bathymetry, from shallow waters to depths such as 60 m. *P. oceanica* is also present along most parts of the Western Mediterranean coasts. It is a good biomarker that signals clear waters and it can live up to 50 m. Zosteraceae occurs in almost 10% of the dataset and is a species that is mostly found as small isolated stands, especially in lagoons. It is encountered mostly in the Adriatic Sea, the Tyrrhenian Sea, and Sicily, and lives up to 15 m depth. Another warm water species is *Halophila*, a Red Sea species, which is 'invading' the Mediterranean Sea since the opening of the Suez Channel. It is mostly found in Cyprus, Greece, Italy, and northern Africa. Finally, *Ruppia* has the lowest occurrence in the dataset. It is found in the Aegean Sea, the Ionian Sea, the western part of Sicily, and the Adriatic Sea. These species can be extremely morphologically variable and therefore their identification is often linked to differences in environmental conditions. They are also very euryhaline and can withstand prolonged periods of desiccation.

Selecting the most appropriate environmental variables is considered as an important task in determining the distribution of seagrass taxa under study (Guisan and Zimmermann, 2000). The modelling procedure followed here involved the selection of environmental parameters based on their potential importance in driving seagrass distributions (determined through a literature review and expert opinion). Table 2 summarizes these variables and their attribute type. Environmental variables (predictors) are divided into static (determining the morphologic, morphodynamic and human impact, considered constant over time) and temporal (environmental parameters exhibiting strong temporal change).

The nature of seabed substrate is an important parameter affecting the distribution of seagrass. Although seagrasses inhabit all types of substrates, from mud to rock, the most extensive seagrass beds occur on soft substrates, like sand and mud. The seabed substrate data were retrieved from EMODnet Geology database (EMODnet Consortium et al., 2016) at 1:100,000 scale and contained 12 different substrate types.

[1] http://wcmc.io/seagrass

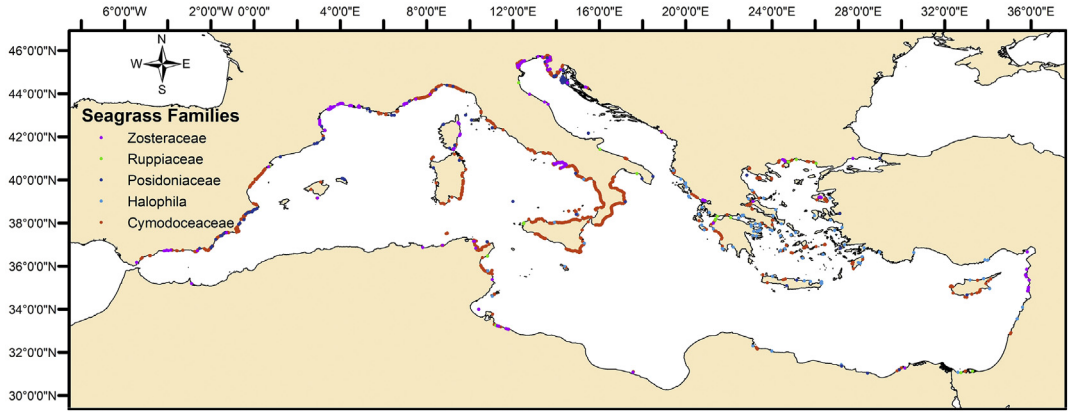

Fig. 1. Seagrass distribution accross the Mediterranean Sea.

Table 1
Seagrass families in the dataset.

Seagrass family	Instances	Percentage
Zosteraceae	187	10.56%
Ruppiaceae	28	1.58%
Halophila	94	5.30%
Cymodoceaceae	1337	75.49%
Posidoniaceae	125	7.07%

Table 2
Variables used.

Name	Type	Layers	Variables
Bathymetry	Static	–	1
Temperature	Temporal	2 (surface, bottom)	38
Salinity	Temporal	2 (surface, bottom)	38
Chlorophyll-α	Temporal	1 (surface)	19
Nitrate	Temporal	2 (surface, bottom)	38
Phosphate	Temporal	2 (surface, bottom)	38
Secchi Disk Depth	Temporal	1 (surface)	19
Wave Height	Temporal	1 (surface)	19
Distance from nearest City	Static	–	2
Distance from nearest River Mouth	Static	–	2
Distance from nearest Port	Static	–	1
Distance to Coast	Static	–	1
Substrate	Static	–	1

These are coarse and mixed sediment, coarse sediment, fine mud, mixed sediment, mud to muddy sand, muddy sand, rock or other hard substrate, sand, sandy mud, sandy mud to muddy sand, and seabed.

Bathymetry is another important parameter linked indirectly to light availability, thus determining the seagrass population structure, the biomass partitioning, and the photosynthetic and respiration rates (Olesen et al., 2002). Bathymetric data at the points/polygons that seagrass exist were retrieved from EMODnet database based on digital elevation models (DTMs) and bathymetric surveys at resolution of 1/8 × 1/8 arc minutes.

As seagrasses are sensitive to human impact, distances of each seagrass point/polygon to the nearest coast, port, city were computed using the haversine distance.

Two datasets for cities were tested: the first consists of all major cities (~4000 cities) that were retrieved from (Desktop, 2011). The cities include national capitals, provincial capitals, major population centers, and landmark cities. The second consists of all communities[2] (~3,800,000 communities).

Apart of light, nutrients also represent energy and matter input to stimulate seagrass growth and total annual production (Elkalay et al., 2003). The distance of seagrass presence points/polygons to the nearest river mouth, as sources of nutrients and suspended matter, were computed using the same strategy with two datasets (Desktop, 2011; Lehner et al., 2006). Finally, one dataset[3] was used to extract the distance from the nearest port and one (Wessel and Smith, 2013) for the distance to coast. When calculating distance to coast, some points were found in land and their distance was set to zero.

Temporal environmental variables included all state variables considered as drivers for seagrass growth, biomass, and distribution, by existing numerical models, such as water temperature, salinity, nitrate and phosphate, chlorophyll-α, significant wave height, and water column transparency (expressed as Secchi Disk Depth). Mean-monthly data for these parameters were extracted from the Copernicus Marine Environmental Service (CMEMS) database (ECJRC, 2018). In Fig. 2, the distribution of four such variables can be seen. Temperature, salinity, and nutrient monthly-mean data were extracted from the surface and bottom of the water column and were imported into our database. Surface chlorophyll-α data were based on remote sensing observations transformed into L4 (MEDSEA_REANALYSIS_PHYS_006_004) datasets at 1 km × 1 km resolution. In absence of satellite data, modeled data were used with a resolution of 0.063° × 0.063°. Surface and bottom temperature, salinity and nutrient data, as well as chlorophyll-α bottom values were extracted from the MedSea Physics Reanalysis dataset, with a horizontal resolution of 0.063° × 0.063°. The mean-monthly values of significant wave height at sea surface were extracted from the Mediterranean Sea wave model hindcasts available at a resolution of 0.042° × 0.042° (See Table 3).

For the temporal variables we derived 1 value for each month of the year, the average for each season of the year, the average for the year, and the maximum and minimum values that were reported for the year, totaling 19 variables. In the cases were 2 layers were used, variables are doubled to 38. Static variables have only one fixed value. The total number of variables that we collected are 217.

When pre-processing the data, substrate was the only categorical variable, and in order for all machine learning algorithms to work, one-hot encoding was used, where a new binary variable is added for each unique categorical value. Also, we wanted to extract the values for the bottom of the sea, where the seagrass meadows live, and check their relative importance compared to the surface values. In this respect, a Python script was developed reading each netCDF pixel and returning the bottom value of the relevant Copernicus variables.

One difficulty that we faced was that our dataset had many coastal observations that were not covered by the Copernicus and EMODNet

[2] https://www.maxmind.com/en/free-world-cities-database

[3] https://msi.nga.mil/NGAPortal/MSI.portal?_nfpb=true_pageLabel=msi_portal_page_62 pubCode=0015

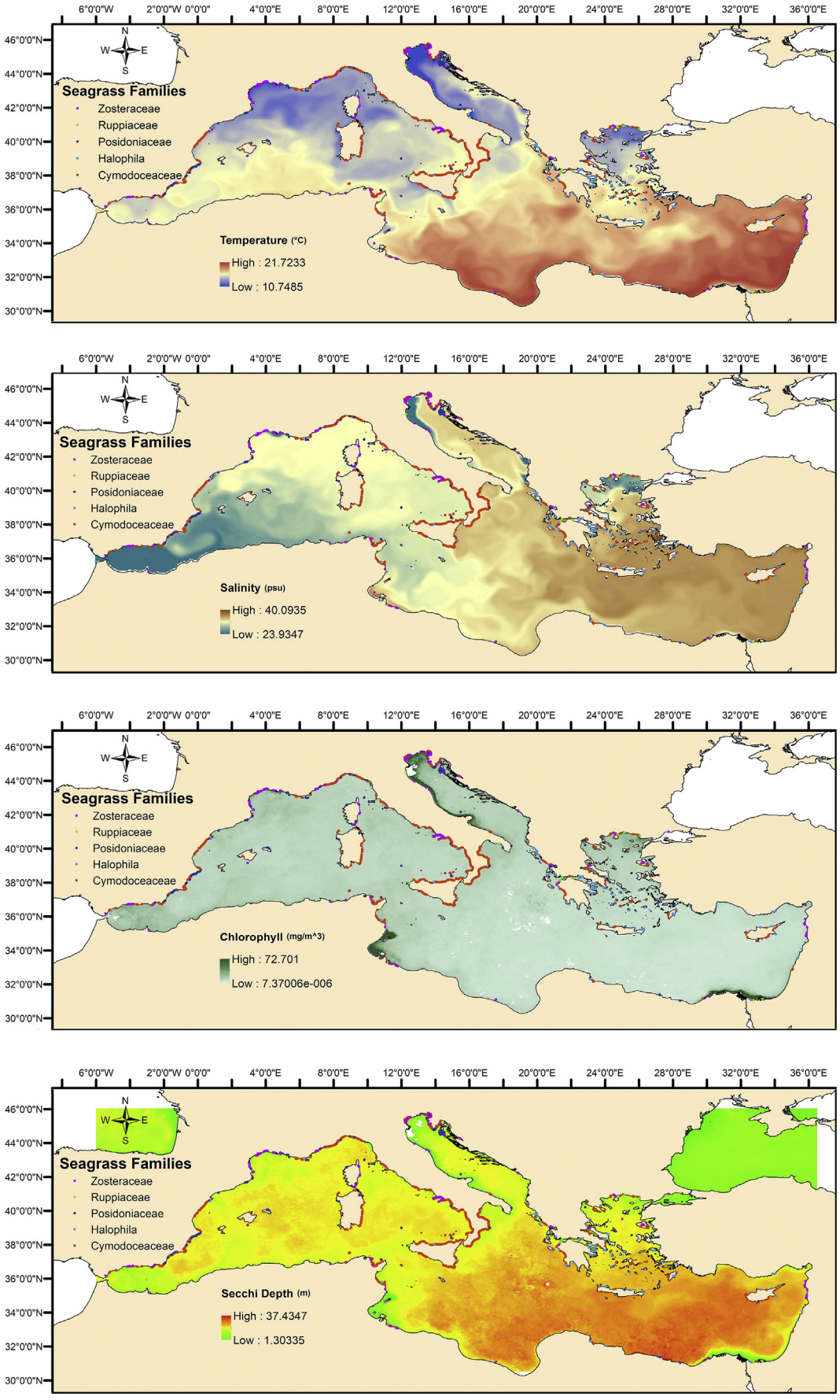

Fig. 2. Temperature, Salinity, Chlorophyll-*a* and Secchi Depth distribution across Mediterranean Sea. Monthly mean values for December.

Table 3
CMEMS datasets used to device the seagrass ML database.

Parameter	CMEMS product	Resolution
Water Temperature	MEDSEA_REANALYSIS_PHYS_006_004	0.063° × 0.063
Salinity	MEDSEA_REANALYSIS_PHYS_006_004	0.063° × 0.063
Nutrients	MEDSEA_REANALYSIS_BIO_006_008	0.063° × 0.063
Chlorophyll-a	MEDSEA_REANALYSIS_BIO_006_008	0.063° × 0.063
Secchi Disk depth	OCEANCOLOUR_GLO_OPTICS_L4_REP_OBSERVATIONS_009_081	1 km × 1 km
Significant wave height	MEDSEA_HINDCAST_WAV_006_012	0.042° × 0.042

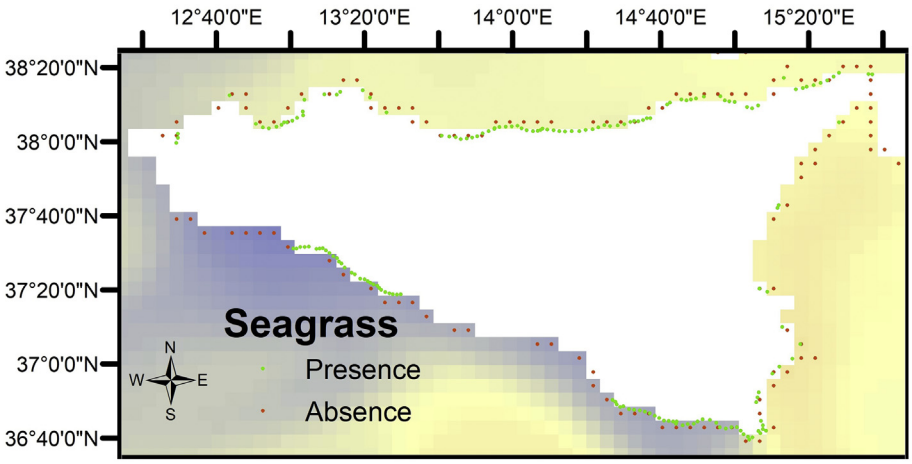

Fig. 3. Coastal points problem in Sicily.

spatial resolution, resulting to many points seemingly 'falling' on land. This problem can be seen in Fig. 3 for Sicily.

For all temporal variables and bathymetry, only a few points of our seagrass dataset were present in the external sources. Table 4 displays these statistics. In order to handle this problem, a script was written that calculates the haversine distance between a point with no value and all the nearby points. It finds the value of the closest point with a value and copies it. So, at this point we have all the values for surface and bottom without any missing values.

2.3. Absence dataset

In order to be able to predict the seagrass presence–absence pattern at an unknown point along the Mediterranean coast, an absence dataset was created. As there is no publicly available absence seagrass dataset, a set of artificial absence records was developed based on certain rules and hypotheses.

With a high probability, we claim that cells next to our initial seagrass presence dataset are in lack of seagrass. We suppose that when observations were made in order to detect seagrass, all adjacent areas were examined, and where no seagrass was found, it was not included in the dataset. If seagrass was found, then it would be available in the dataset.

Table 4
Number of variables that existed for the initial dataset.

Name	Number of values	Percentage
Bathymetry	1116	63.01%
Temperature	598	33.76%
Salinity	598	33.76%
Chlorophyll-α	582	32.86%
Nitrate	597	33.71%
Phosphate	597	33.71%
Secchi Disk Depth	1357	76.62%
Wave Height	825	46.58%

Based on the above assumption, a Python code was developed following a set of well-defined rules to generate points considered exhibiting an absence in seagrass. We used 2 files: the first is the seagrass dataset and the second is the data of any temporal environmental variable obtained from CMEMS. As CMEMS data are gridded, environmental data were assigned at the center point of each CMEMS pixel. For each point of the initial seagrass dataset, we search for the closest point of the shapefile that we just created that has not already found to be absent of seagrass by a previous point. There is also a restriction that forbids a point to be selected if its distance from the closest shore is longer than the fixed value of 10 km. So, if many seagrass presence points are close to each other, there is a chance that there will be generated less absence points than these.

The above methodology tends to follow the coast. A total of 1284 absence points were artificially generated, and our total dataset now consists of 3055 entries. In Fig. 4, the final presence–absence seagrass dataset is depicted, along with an in-depth inspection around Sicily Island. In Section 4.1.1 we will provide a rough estimate on how our experiments are affected by the artificiality of our absence data.

3. Data analysis and machine learning

In this section we investigate several ways to determine the importance of the features selected for classification. We briefly present the classification algorithms that were used for our experiments, as well as the evaluation measures.

3.1. Variable importance

By determining which variables are most important we can simplify the analysis of a dataset, better understand the physical concepts of it, and exclude the ones that confuse the model achieving better accuracy and learning/testing speed.

There exist several ways to determine variable importance, e.g. Decision Trees, Random Forests, Chi-square, and Regression. We deal with tree-based variable importance, like (Arthur et al., 2010) did. It

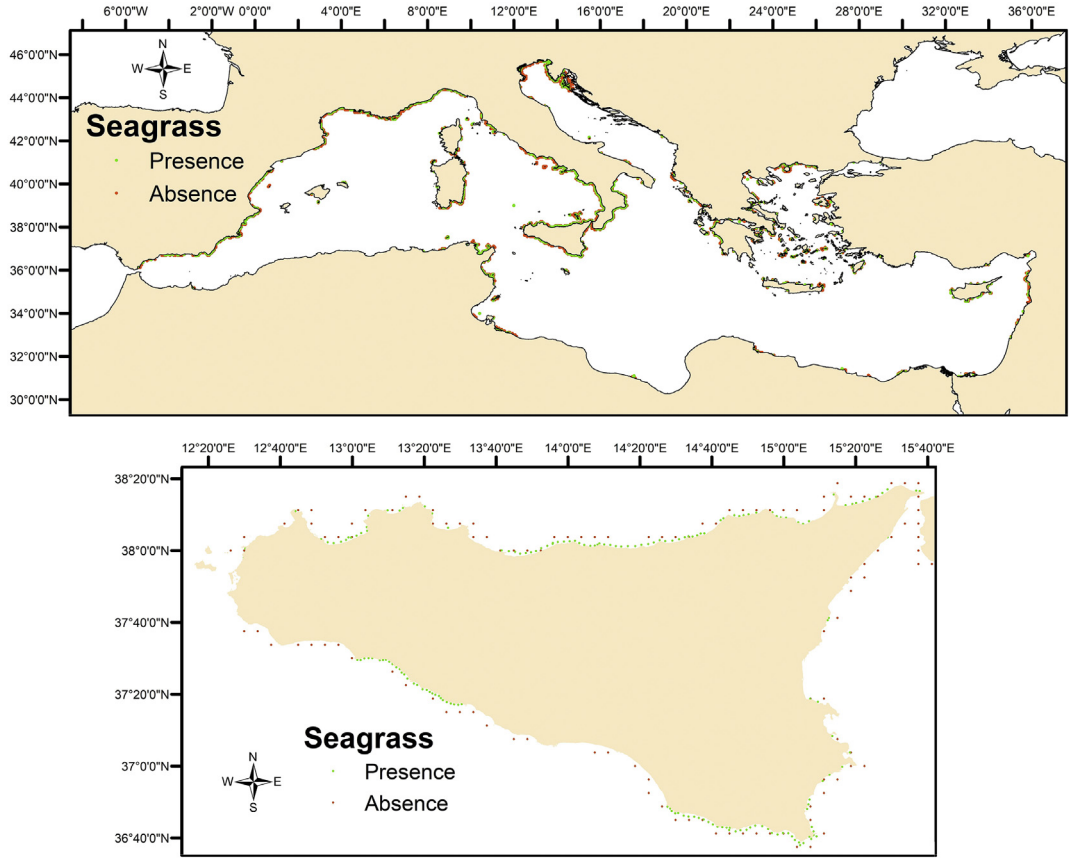

Fig. 4. Presence (green) and absence (red) points of seagrass in Mediterranean (above) and Sicily (below).

attempts to determine how to split the data into smaller, more homogeneous buckets, accessing one variable at a time. When the most important variable is found, it is placed at the top of the tree. However, Decision Trees have some drawbacks. While real world data may not be linear, trees are designed to look at linear separations. Also, a Decision Tree is not able to get past biased data. So, instead of making one tree model, it is best practice to make multiple, e.g. by using Random Forest, which combines many Decision Trees. Now, each model built is looking at a different subset of data and is not always using the same variables.

When we determine the best features, we have to choose which ones to exclude from our final model in order to improve performance. One feature selection method that deals with this need is called Recursive Feature Elimination and is used by Li et al. (2013). Given an external estimator that appoints weights to features, recursive feature elimination aims to choose features by recursively considering fewer and fewer arrangements of features. To start with, the estimator is trained on the underlying arrangement of features and the significance of each feature is acquired through some property. At that point, the less critical features are pruned from the current arrangement of features. This method is recursively rehashed on the pruned set until the coveted number of features to choose is reached.

There is not a perfect feature selection method that works in every case. Researchers have to test them and apply the best depending on their dataset. For our work we found tree-based variable importance to work better, and especially when using the highest performing tree algorithm (Random Forest) to determine the best features.

3.2. Machine learning algorithms

There exist many supervised machine learning algorithms in literature, distinguished into several categories, such as Generalized Linear Models (GLM), Decision Trees (DT), Instance Based (IB), Support Vector Machines (SVM), and others. Seven well-known algorithms from these categories like Passive-Aggressive, Logistic Regression, Ridge, Linear SVC, k-Nearest Neighbors, Decision Tree, and Random Forest were used in this study.

3.2.1. Passive-aggressive

The Passive-Aggressive algorithm belongs to a family of algorithms for large-scale learning. As the name states, this algorithm is passive, meaning that it keeps the model if the classification was correct, and aggressive, meaning that it updates to adjust the misclassified example if the classification was incorrect.

3.2.2. Logistic regression

It is a popular algorithm that belongs to the Generalized Linear Models methods—despite its name—and it is also known as Maximum Entropy. In this model, the probabilities describing the possible outcomes of a single trial are modeled using a logistic function.

3.2.3. Ridge

Ridge (Hoerl and Kennard, 1970) is a classical data modelling method to solve the multicollinearity problem of covariates in samples. It belongs to the Generalized Linear Models, like Linear Regression, but addresses some of the problems of Ordinary Least Squares by imposing a penalty on the size of coefficients. It has a complexity parameter α that controls the amount of shrinkage: the larger the value of α, the greater the amount of shrinkage and thus the coefficients become more robust to collinearity (Pedregosa et al., 2011).

3.2.4. Linear SVC

One of the most popular machine learning methods for classification of linear problems are Support Vector Machines (SVMs) (Cherkassky, 1997) with a linear kernel. They try to find a set of hyperplanes that

separate the space into areas representing classes. These hyperplanes are chosen in a way which maximizes the distance from the nearest data point of each class. The Linear SVC is the simplest and fastest SVM algorithm assuming a linear separation between classes.

3.2.5. K-nearest neighbors

It is an non-parametric 'lazy' learning algorithm. This means that it does not make any assumptions on the underlying data distribution and that it does not use the training data points to do any generalization. It does not attempt to construct a general internal model, but simply stores instances of the training data. Classification is computed from a simple majority vote of the nearest neighbors of each point: a query point is assigned the data class which has the most representatives within the nearest neighbors of the point.

3.2.6. Decision tree

The Decision Tree Classifier is a simple and widely used classification technique. It applies a straightforward idea to solve the classification problem by posing a series of carefully crafted questions about the attributes of the test record. Each time it receives an answer, a follow-up question is asked until a conclusion about the class label of the record is reached.

Random Forest. A Random Forest is an ensemble strategy that joins numerous individual classification trees in the accompanying route: from the first example, numerous bootstrap tests and segments of indicators are drawn, and an unpruned classification tree is fitted to each bootstrap test utilizing the inspected indicators. From the complete forest, the status of the response variable is typically anticipated as the prediction of the forecasts of all trees as the class with majority vote for classification (Breiman, 2001).

3.3. Evaluation measures

A measure commonly used to evaluate classification results is Accuracy, which is the ratio of correct to all classification decisions. Accuracy is a good metric for balanced datasets, like the case studied here for seagrass presence–absence, and is defined as:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FN} + \text{FP}}\ , \tag{1}$$

where TP is the number of true (correct) positive (presence) predictions, TN is the number of correct negative (absence) predictions, FP is the number of false positive predictions and FN is the number of false negative predictions.

Other metrics used are Precision, Recall, and the F-measure. Precision is defined as the fraction of relevant/correct instances among the retrieved instances for a class, while Recall is the fraction of relevant instances that have been retrieved over the total amount of relevant instances. In terms of the same counts used above, they can be expressed as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}\ , \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}\ . \tag{2}$$

The F-measure score is the harmonic mean of Precision and Recall:

$$F = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}\ . \tag{3}$$

Typically, the input dataset is split into two disjoint subsets, the training and the testing set. The training set is used to learn the model, while the test set is used to measure a performance measure. But how confident can we be about the classification performance? The results may be due to accidental characteristics of the specific partitioning. For example, the test set may happen to include points that are easy to be categorized, with the result that the categorizer yields good performance. Consequently, the choice of a fixed, predefined partitioning of the dataset may not be the best way for evaluating classifiers.

Cross-validation is an iterative method for calculating the expected value of a particular measure. It splits the dataset into K equally-sized parts that are called folds. In each iteration, different $(K - 1)$-folds are used for training and the remaining fold for testing. The overall measure of its performance is the average of the measures of the individual iterations. The above method guarantees that every instance will be used both for training and testing. There is an alternative of cross-validation method called stratified cross-validation, where in each fold a balanced number of instances for each class is selected. For our experimentation, we used stratified cross-validation with $K = 10$ folds.

4. Experiments

In this section, we perform two experiments. First, we are trying to predict the existence of seagrass, and then its family. We are also interested on which variables affecting those predictions most.

4.1. Detecting seagrass

First, we try to solve the following problem: 'Given an unknown point, does it have seagrass or not?' This can be seen as a binary classification problem. Before building a classifier, the relative importance of environmental predictors will be determined, using the Random Forest algorithm.

In Fig. 5, we report the relative importance for each predictor of Table 2. For temporal predictors, only the first appearance of the most important month affecting seagrass is displayed. Predictors' importance results indicated that eight out of the top-10 features were the different chlorophyll-α mean-monthly levels. Thus, we conclude that chlorophyll-α is very important parameter in determining the

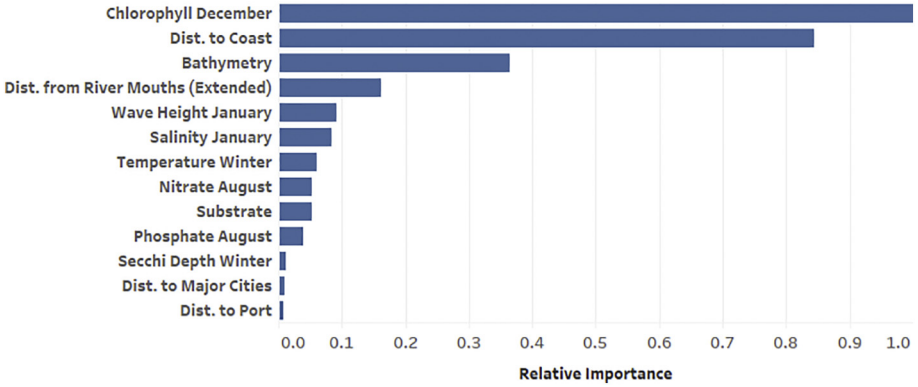

Fig. 5. Relative importance of variables for seagrass absence-presence classification. For temporal variables (e.g. Chlorophyll-α), only their best first period (e.g. month) is displayed.

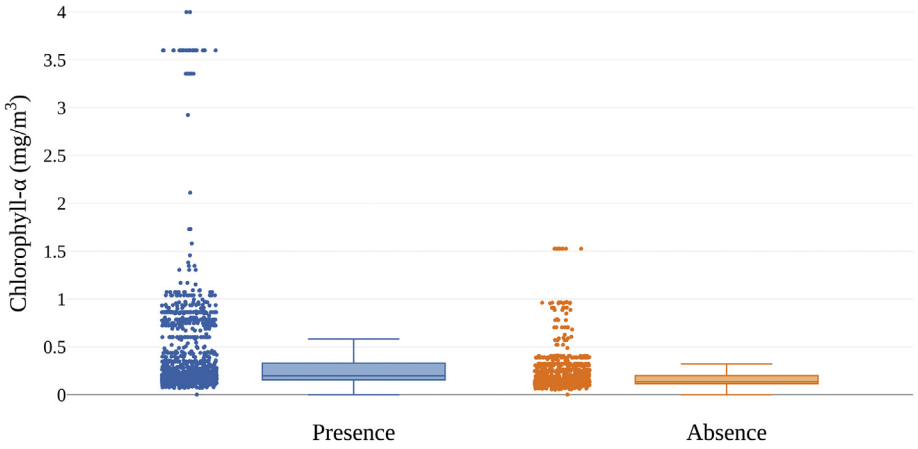

Fig. 6. Distribution of Chlorophyll-*α*-December values per presence–absence class.

presence–absence of seagrass in the Mediterranean Sea. Apart from chlorophyll-*α*, the distance to coast seems to be a strong indicator of seagrass, as well as bathymetry (indirectly related to distance to coast), while all other predictors seemed less important. Another finding is that for most dynamic, temporally-changing variables during the winter months are more important. Chl-*α* winter data exhibited higher values (mean: $0.32 \pm 0.20\, mg\ m^{-3}$) compared to the other seasons, at the areas with seagrass presence. December was the month with mean Chl-*α* concentration ($0.37 \pm 0.53\, mg\, m^{-3}$) at these areas. On the contrary, in points of seagrass absence Chl-*α* concentration peaked in spring (mean: $0.26 \pm 0.35\, mg\ m^{-3}$) with highest values in February and March ($> 0.30\, mg\ m^{-3}$). Finally, the conditions prevailing at the seabed (e.g., substrate type, bottom temperature, salinity, nutrients, etc.) appear to be of lesser significance, as the surface measurements are higher in rank.

Proceeding further to understand better the behavior of predictors, the box plots of the best two variables are displayed in Figs. 6–7. On the left-hand-side of each class, the value of each point is presented, while on the right some statistics like the upper fence (Q3), the median and the lower fence (Q1) are shown. Points outside the lower–upper fence margins are considered as outliers. It can be seen that higher values of chlorophyll-*α* are essential for the growth of seagrasses. Furthermore, from the highly important 'distance-to-coast' predictor, it is clear that seagrass is present at close distances to the coast. While distance-to-coast may be positively correlated to bathymetry, the latter is found to be less than half as important in the variable strength analysis.

In a further step, all machine learning algorithms were trained and tested using the stratified cross-validation technique and the evaluation measures were computed (Table 5). Having a rather balanced dataset (in terms of comparative presence–absence item counts), Accuracy is a suitable measure for model performance evaluation. Nevertheless, all applied algorithms are ranked based on the F-measure (from worst to best), for reasons to be explained in Section 4.1.1 below.

Table 5
Binary classification effectiveness, per classifier, using all variables.

Classifier	Accuracy	Precision	Recall	F-measure
Passive-Aggressive	59.4	45.4	66.6	53.8
k-Nearest Neighbors	68.5	73.3	71.4	72.1
Logistic Regression	61.2	61.9	98.7	75.2
Ridge	61.2	61.6	98.8	75.2
Linear SVC	75.4	76.6	88.5	80.5
Decision Tree	92.5	95.7	91.0	93.2
Random Forest	93.4	98.1	90.4	93.8

Overall, Generalized Linear Models (GLMs), such as Passive-Agressive, Logistic Regression, and Ridge, perform poorly on seagrass presence–absence detection. Linear SVC models produce modest results compared to the strong tree-based models. These findings suggest that the relationship between seagrass presence–absence and environmental variables is non-linear. Tree-based algorithms can deal with that problem better, resulting in F-measure and Accuracy exceeding 90%. Note also that Accuracy gives almost the same ranking of ML algorithms, with the exception of *k*NN which is evaluated higher in Accuracy (better than most linear ones) but lower in F-measure (i.e. bad even among the linear ones). This seems to be due to 'sacrificing' too much Recall for Precision.

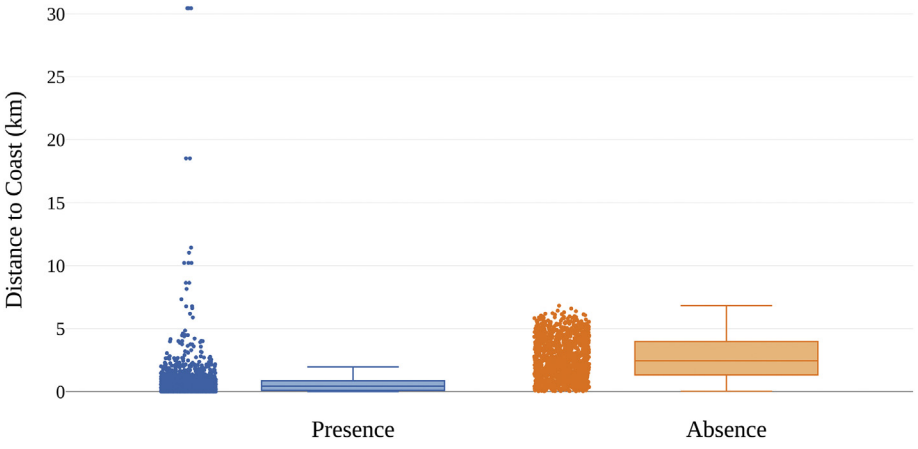

Fig. 7. Distribution of Distance-to-Coast values per presence–absence class.

4.1.1. Estimating uncertainty

Since the absence part of our dataset is generated artificially with the method described in Section 2.3, it may contain errors introducing uncertainty in the absolute numbers of the evaluation measures. Let us do a rough estimation of how large this uncertainty may be.

Our main hypothesis for selecting absence datapoints in Section 2.3 has been that cells next to our presence ones are in lack of seagrass. But what if some of the cells that we marked as absence are in reality presence? Consider a confusion matrix with TP/FP/FN/TN displaying the actual presence/absence observations and the ones that our binary classifiers predicted as presence/absence. The real absence data are split by our classifiers into FP and TN categories. Assuming that our method for generating absence data makes errors in the magnitude of 5% which are split equally (percentage-wise) in FP and TN, then the real values of the evaluation measures can be calculated by moving 5% of the FP and TN counts to TP and FN, respectively, since those are presence data in reality. Taking a specific confusion matrix from one of the folds of the Random Forest run with Accuracy/Precision/Recall/F-measure of 85.9/79.3/89.8/84.2, modifying the counts as described above results to measure numbers of 84.3/80.7/85.4/83.0. Obviously, due to the artificiality of our absence data, we overestimate (mostly—Precision is underestimated) effectiveness by 1.9/−1.8/5.2/1.4 (% changes from real to estimated measures). If we assume 20% errors, the latter differences become 10/−7/21/6.2.

Thus, the least affected measure is the F-measure, that is why we used it as the main measure in this experiment for ranking the ML algorithms. Furthermore, its uncertainty seems small, even for the 20% errors case, which suggests that our ranking of algorithms would be the same with real, error-free absence data. Consequently, our conclusions are not affected, although the absolute effectiveness numbers may be different with error-free data.

In retrospect, the situation may be even better, since we now also train with possibly erroneous data, negatively impacting our classifiers. So, maybe overestimating the measures cancels out this impact a bit. The procedure for estimating uncertainty described in this section, could be generalized, but this is beyond the scope of this paper; a rough estimate/indication is sufficient for our purpose at this point.

4.2. Detecting seagrass family

Another problem we try to solve is: 'If a seagrass exists, which family is it?' This, based on our dataset, is defined as a one-of multi-class classification problem. Note that all experiments in this section do not involve any artificially-generated data, so the absolute numbers of the evaluation measures are not affected. As before, we firstly investigate the variable importance using Random Forest.

Out of the top-10 features, 5 different chlorophyll-α months were again present. So, chlorophyll-α levels are also important for seagrass family classification. Fig. 8 shows the relative importance of the best first occurrence of the variables of Table 2. In contrast to presence–absence classification, here there are not a few variables with higher importance than others, but many of them present a relative high importance. The less important is Distance-to-Port, with a strength of around 1/3. This means that there are a lot of different factors that play a key role when determining which seagrass family exists out of the 5 choices.

Figs. 9–10 present the distribution of the two most important variables per seagrass family. For chlorophyll-α, the mean value for each seagrass family varies. This means that each family prefers specific and different values of chlorophyll-α. For *Cymodocea* and *Posidonia*, chlorophyll-α ranges in a specific range of values, while for the other families this is more spread. For salinity, the mean values also differ among seagrass families, with *Zostera* preferring lower values and *Halophila* higher. The interquanitile ranges also vary, with *Halophila* exhbiting more 'tight' distribution, while *Posidonia*, *Cymodocea* and *Ruppia* appear more spread (See Table 6).

The final step is to train and test the machine learning algorithms in this dataset for 5-class classification. Stratified cross-validation was used, and evaluated with Precision, Recall and F-measure. This dataset is not balanced as can be seen in Table 1. Cymodocea is the dominant class that constitutes the 75% of the dataset. Thus, Accuracy may not be a suitable measure, because it measures how many correct predictions were made overall, and if we predict all the test examples as cymodocea, then Accuracy would be close to 75% without even predicting another class. This is clearly a problem because many machine learning algorithms are designed to maximize overall Accuracy, with the exception of the tree-based algorithms. So, we resort to the F-measure; the algorithms are ranked based on it.

When computing the total F-measure, we use the macro-average. Macro-averaging takes the average of all individual class F-measures, treating small and large classes equally, in contrast to micro-averaging which aggregates the TP/TN/FP/FN counts from all classes and computes a total measure which is biased to the large classes.

We also experimented with using subsets of best features from top-50% down to top-5% with a step of 5%, and the effectiveness maximized at using top-10% of features. When all features are used, the best performing classifiers are the tree-based Decision Tree and Random Forest. For the top-10% of the features, which were calculated using variable importance with Random Forest, all algorithms except the tree-based ones give lower results. However, the best F-measure score (37.1 of Decision Tree) has been improved by 2.6% in comparison with the all-features experiment and is now achieved by Random Forest with 39.7%.

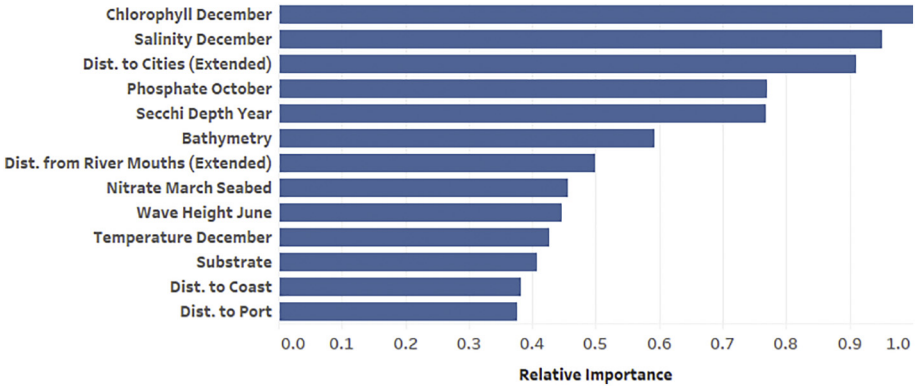

Fig. 8. Relative importance of variables for seagrass family classification. For temporal variables (e.g. Chlorophyll-α), only their best first period (e.g. month) is displayed.

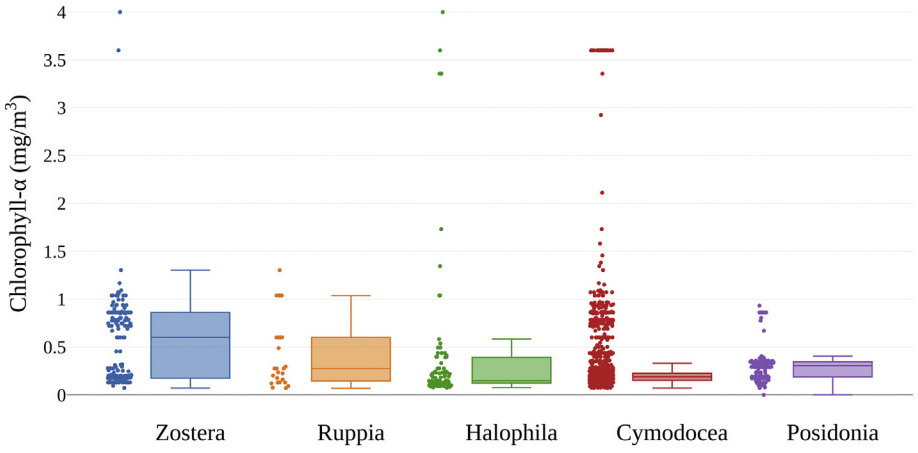

Fig. 9. Distribution of Chlorophyll-*α*-December values per seagrass family.

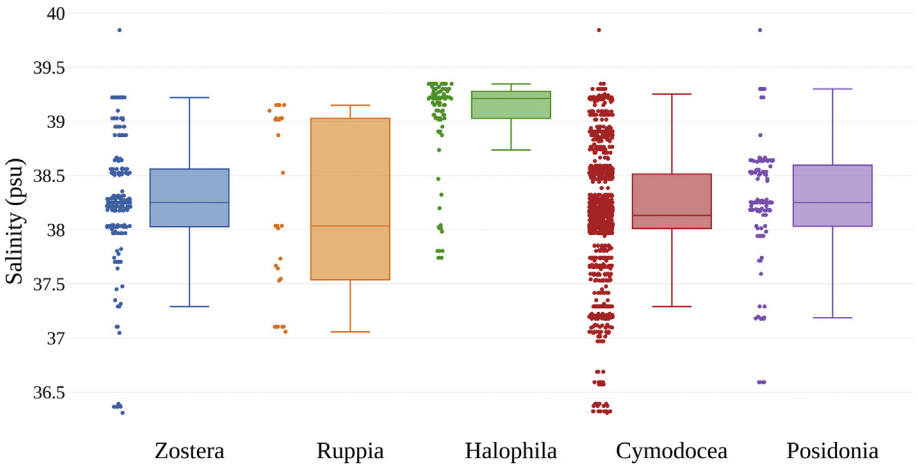

Fig. 10. Distribution of Salinity-December values per seagrass family.

Table 6
Effectiveness (%) of family classification.

Classifier	All features			Top-10% features		
	Precision	Recall	F-measure	Precision	Recall	F-measure
Linear SVC	21.7	30.3	22.4	23.2	24.9	20.2
Logistic Regression	36.6	27.1	26.4	29.3	21.8	21.3
Ridge	33.5	25.6	25.9	31.6	22.2	22.0
Passive-Aggressive	24.1	25.2	22.2	35.0	26.4	25.5
k-Nearest Neighbors	46.8	35.0	35.9	43.5	31.1	32.5
Decision Tree	40.3	42.0	37.1	41.3	42.3	38.9
Random Forest	41.3	38.2	36.6	44.4	41.4	39.7

5. Discussion

Although not entirely understood, it is presently evident that water column quality affects the abundance of seagrass and in many occasions controls the gradual replacement of certain plant species by opportunistic seagrasses capable to adapt under poor ecological conditions. However, the exact processes and dynamics of these changes have not been extensively documented, mostly due to the slow nature of environmental degradation, the even more gradual ecological response of seagrass communities, the non-linearities hidden in marine ecosystem dynamics and the lack of widely-spread systematic environmental datasets (Boudouresque et al., 2009). The relative impacts of environmental drivers, expressed as water temperature, salinity, nutrients, transparency; ecosystem components, as bathymetry and bottom substrate; and the human influence, described by the distance to cities, ports and river outflows have not been previously reported. In this work we attempt to explore the nonlinear dynamics among the environment-ecosystem-human gradient and their impact on seagrass presence and distribution in the Mediterranean Sea. To achieve this, we utilized the power of machine learning algorithms when applied to large and diverse datasets maintained by international organizations, as Copernicus, EMODnet and UNEP. The importance of such work, and others of similar context could be vital, as there is a real need for understanding better coastal benthic processes and human interaction, especially for the Mediterranean Sea, where seagrass loss correlates with the rapid shoreline urbanization and remedial and restoration projects should be undertaken (Green, 2003).

Our results suggest that the main natural parameters affecting the distribution of seagrass at family and genera level are the winter (mainly December) chlorophyll-*α* and salinity levels, the autumn phosphate concentrations and bathymetry, expressing changes in temperature, pressure and light availability. The human impact, expressed as the distance from all coastal communities along the Mediterranean shoreline also determines seagrass classification (Fig. 8). Previous publications support these findings e.g., (Danovaro, 1996; Danovaro and Fabiano, 1995; Olesen et al., 2002), although the relative importance of the above-described factors on seagrass benthic distribution was unknown. This is the strength of data mining and machine learning techniques applied at these large databases, although since results are mostly data-driven, our conclusions could be altered in case a more

explicit or expanded database is being used.

Using machine learning algorithms it was revealed that areas with relatively higher winter Chl-*α* content favor the presence and development of seagrass. On the other hand, areas with absent seagrass species, in the vicinity of these meadows, exhibited higher spring Chl-*α* concentrations, associated to water column eutrophic incidents. The impact of water column deterioration due to spring eutrophication incidents and subsequent light deprivation and bottom anoxia on seabed habitat richness and diversity is eminent and has been reported by many investigators in the past (Bite et al., 2007; Martins et al., 2001). Such conditions may introduce intolerable stress to benthic marine organisms, loss of seagrass biomass and ultimately to habitat deterioration. For species such as *Zostera noltii*, frequent spring eutrophication episodes may lead to their progressive replacement by opportunistic macroalgae (Cardoso et al., 2004).

In parallel, the impact of seawater salinity on seagrass distribution, although well-known (Castriota et al., 2001; Fernández-Torquemada and Sánchez-Lizaso, 2005; Gacia et al., 2007), explicit tolerance levels have never been defined. Our results suggest that in the Mediterranean Sea, winter salinity is the second most important parameter to differentiate the distribution of seagrass species. *Halophila* appears to survive at the most saline environments, within a narrow yearly-averaged salinity range from 39 to 39.3 psu (Fig. 11), characterizing mostly the eastern basin and especially along the northern Africa coastline, Cyprus and the southern parts of Italy and Greece. *Posidonia oceanica* and *Cymodocea nodosa* appear favoring lower, almost similar winter salinity levels, ranging between 37.3 and 39.3 psu. This range lies within the preference level reported by other investigators (Ruíz et al., 2009; Tomasello et al., 2009), although it is shown that *P. oceanica* favors more stable salinity conditions while *C. nodosa* more euryhaline environments with presence along estuaries and near river mouths (Boudouresque et al., 2009). This analysis is consistent to the experimental results of (Sandoval-Gil et al., 2012), showing that *C. nodosa* may grow in diverse coastal environments with variable salinity levels, while *P. oceanica* is limited to more stable in salinity marine environments. Furthermore, our analysis reveals that *Ruppia* exhibits higher relative tolerance, resisting to marked inter-salinity differences (37–39.2 psu). The annually-average values for salinity can be seen in Fig. 11.

Food availability determines seagrass growth, distribution and metabolism, especially as seagrasses can uptake nutrients not only through roots but also through leaves (Brix and Lyngby, 1985). This is particularly important as there exists a diverse variety of nutrient sources along the Mediterranean coastline, providing the appropriate levels to seagrass sustainability, but regularly exceed carrying capacity levels leading to eutrophication incidents. Phosphorus compounds, either as dissolved organic phosphorus (DOP) or particulate organic phosphorus (POP), can be readily usable through hydrolysis by a number of forms of the enzyme alkaline phosphatase (Pérez and Romero, 1993) and then up taken and assimilated by seagrass. Previous uptake studies on several seagrass species showed that both leaf and root tissues exhibited highly variable kinetic parameters at different phosphate levels among species (Fourqurean et al., 1992; Udy and Dennison, 1997), proving that species specific responses to nutrient additions occur, supporting the results of the present study that phosphate levels are important environmental parameter determining seagrass classification.

Finally, the present study results indicate the seasonality determines strongly seagrass species distribution. Cardoso et al. (2004) reported the strong seasonal effect determining the growth of *Z. noltii* in Montego estuary, with higher seagrass above-ground-biomass in spring and summer (due to leaf growth) and the increase in rhizome and roots in autumn and winter. Seasonality is also related to light intensity, indirectly linked to water transparency, expressed by Secchi depth (Fig. 12) showing that *Halophila*, *Cymodocea* and *Posidonia* favor more transparent water column conditions, while *Zostera* and *Ruppia* may grow under limited light availability. Similar results were also drawn by Duarte et al. (2007) utilizing data from 424 reports on seagrass species distribution in relation to light extinction depth.

6. Conclusion and future work

In this work, data from large, systematic and diverse databases were fused to model the factors determining the presence/absence and families' distribution of seagrass in the Mediterranean Sea. Overall, we may conclude that machine learning algorithms and data fusion techniques may support marine ecological studies providing a better understanding of hidden, non-linear processes and interrelations to environmental variables and human impacts. Such models interrelating environmental variations and human impacts on biological subjects (as macrophytes) could be used to back up the implementation of the Marine Strategy Framework and Maritime Spatial Planning Directives in the Mediterranean.

The main findings of the present work suggest that (1) tree-based classification algorithms, and especially the random forest exhibits the best performance on determining seagrass presence/absence and identification at family level; (2) chlorophyll-*α* levels in winter months (mostly in December) is the key parameter determining seagrass presence/absence and family classification; (3) distance from coast, distance from river mouths and bathymetric change are key factors deterining seagrasss presence; (4) salinity and phosphate levels may also identify seagrass family preference; (5) *Cymodocea* and *Posidonia* are more abundant at low, limited-range chlorophyll-*α* levels; (6) *Halophila* seems more tolerant at higher salinities, while *Ruppia* prefers euryhaline conditions, and (7) *Halophila*, *Cymodocea* and *Posidonia* favor more

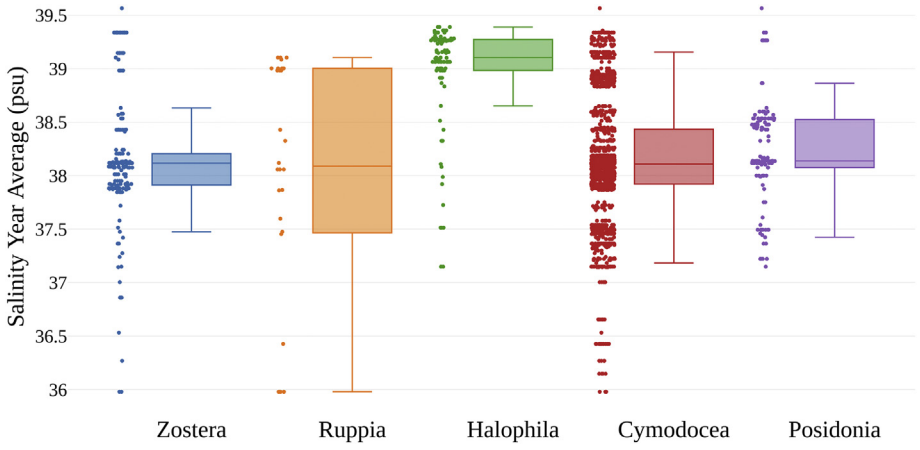

Fig. 11. Distribution of Salinity year average values per seagrass family.

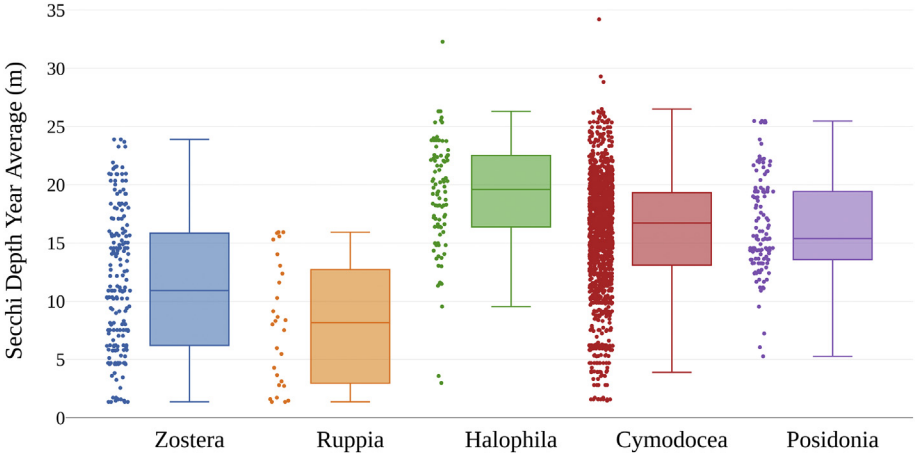

Fig. 12. Distribution of Secchi Depth year average values per seagrass family.

transparent water columns, while *Zostera* and *Ruppia* may grow under limited light and turbid environments. All of the above findings are in-line to conclusions drawn by previous local/regional studies.

We proposed a methodological framework for the development of an absence dataset and an uncertainty assessment analysis. As a future addition to this work, we could add more variables and test their effectiveness in model selection and classification. Candidate variables could be biological like primary production and pH, physical such as sea surface height and currents, and human related activities like dredging through fishing activity and pollution. Other model selection methods except variable importance with forest algorithms could further be tested, like univariate selection and recursive feature elimination. Finally, more algorithms could be investigated, such as the popular neural networks.

Acknowledgements

The research leading to these results received funding from the European Union's Horizon 2020 Research and Innovation Program (H2020-BG-12-2016-2) under grant agreement No. 727277 - ODYSSEA (Towards an integrated Mediterranean Sea Observing System). The article reflects only authors' view and that the Commission is not responsible for any use that may be made of the information it contains. The authors wish to thank Corine Martin (UNEP-WCMC) for providing the database.

References

Ahmad, F., Azman, S., Said, M. I. M., & Lavania-Baloo (2015). Tropical seagrass as a bioindicator of metal accumulation. Sains Malaysiana, 44, 203–210.
Alameddine, I., Cha, Y., Reckhow, K.H., 2011. An evaluation of automated structure learning with Bayesian networks: an application to estuarine chlorophyll dynamics. Environ. Model Softw. 26, 163–172.
Arthur, A.D., Li, J., Henry, S., Cunningham, S.A., 2010. Influence of woody vegetation on pollinator densities in oilseed brassica fields in an Australian temperate landscape. Basic Appl. Ecol. 11, 406–414.
Bentlage, B., Peterson, A.T., Cartwright, P., 2009. Inferring distributions of chirodropid box-jellyfishes (Cnidaria: Cubozoa) in geographic and ecological space using ecological niche modeling. Mar. Ecol. Prog. Ser. 384, 121–133.
Bite, J.S., Campbell, S.J., McKenzie, L.J., Coles, R.G., 2007. Chlorophyll fluorescence measures of seagrasses *Halophila ovalis* and zostera capricorni reveal differences in response to experimental shading. Mar. Biol. 152, 405.
Boudouresque, C.F., Bernard, G., Pergent, G., Shili, A., Verlaque, M., 2009. Regression of Mediterranean seagrasses caused by natural processes and anthropogenic disturbances and stress: a critical review. Bot. Mar. 52, 395–418.
Breiman, L., 2001. Random forests. Machine Learn. 45, 5–32.
Brix, H., Lyngby, J., 1985. Uptake and translocation of phosphorus in eelgrass (*Zostera marina*). Mar. Biol. 90, 111–116.
Cardoso, P., Pardal, M., Lillebø, A., Ferreira, S., Raffaelli, D., Marques, J., 2004. Dynamic changes in seagrass assemblages under eutrophication and implications for recovery. J. Exp. Mar. Biol. Ecol. 302, 233–248.
Castriota, L., Beltrano, A., Giambalvo, O., Vivona, P., Sunseri, G., 2001. A one-year study of the effects of a hyperhaline discharge from a desalination plant on the zoobenthic communities in the Ustica island marine reserve (southern Tyrrhenian Sea). In: 36th CIESM Congress.
Cherkassky, V., 1997. The nature of statistical learning theory. IEEE Trans. Neural Netw. 8, 1564.
Danovaro, R., 1996. Detritus-bacteria-meiofauna interactions in a seagrass bed (*Posidonia oceanica*) of the nw mediterranean. Mar. Biol. 127, 1–13.
Danovaro, R., Fabiano, M., 1995. Seasonal and interannual variation of benthic bacteria in a seagrass bed (*Posidonia oceanica*) of the ligurian sea in relation to the origin, composition and other environmental factors. Aquat. Microb. Ecol. 9, 17–26.
De'Ath, G., 2007. Boosted trees for ecological modeling and prediction. Ecology 88, 243–251.
De'ath, G., Fabricius, K.E., 2000. Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology 81, 3178–3192.
Desktop, E.A., 2011. Release 10. Vol. 437. Environmental Systems Research Institute, Redlands, CA, pp. 438.
Duarte, C.M., Marbà, N., Krause-Jensen, D., Sánchez-Camacho, M., 2007. Testing the predictive power of seagrass depth limit models. Estuar. Coasts 30, 652.
ECJRC (2018). European commission joint reseach centre, directorate space, security and migration, copernicus emergency management service. http://emergency.copernicus.eu, [2018-02-16].
Elkalay, K., Frangoulis, C., Skliris, N., Goffart, A., Gobert, S., Lepoint, G., Hecq, J.-H., 2003. A model of the seasonal dynamics of biomass and production of the seagrass *Posidonia oceanica* in the bay of calvi (northwestern Mediterranean). Ecol. Model. 167, 1–18.
Marine Information Service, et al., 2016. EMODnet Digital Bathymetry (DTM 2016). EMODnet Bathymetry. https://doi.org/10.12770/c7b53704-999d-4721-b1a3-04ec60c87238.
Fernández-Torquemada, Y., Sánchez-Lizaso, J.L., 2005. Effects of salinity on leaf growth and survival of the mediterranean seagrass *Posidonia oceanica* (l.) delile. J. Exp. Mar. Biol. Ecol. 320, 57–63.
Ferrat, L., Pergent-Martini, C., Romeó, M., 2003. Assessment of the use of biomarkers in aquatic plants for the evaluation of environmental quality: application to seagrasses. Aquat. Toxicol. 65, 187–204.
Fourqurean, J.W., Zieman, J.C., Powell, G.V., 1992. Phosphorus limitation of primary production in Florida bay: evidence from c: N: P ratios of the dominant seagrass *Thalassia testudinum*. Limnol. Oceanogr. 37, 162–171.
Gacia, E., Invers, O., Manzanera, M., Ballesteros, E., Romero, J., 2007. Impact of the brine from a desalination plant on a shallow seagrass (*Posidonia oceanica*) meadow. Estuar. Coast. Shelf Sci. 72, 579–590.
Giannoulaki, M., Belluscio, A., Colloca, F., Fraschetti, S., Scardi, M., Smith, C., Panayotidis, P., Valavanis, V., Spedicato, M., 2013. Mediterranean sensitive habitats (mediseh), final project report. DG MARE Specific Contract SI2 600 (741), 557.
Govers, L.L., Lamers, L.P., Bouma, T.J., Eygensteyn, J., de Brouwer, J.H., Hendriks, A.J., Huijbers, C.M., van Katwijk, M.M., 2014. Seagrasses as indicators for coastal trace metal pollution: a global meta-analysis serving as a benchmark, and a Caribbean case study. Environ. Pollut. 195, 210–217.
Green, E.P., 2003. World atlas of seagrasses. Univ of California Press.
Guisan, A., Zimmermann, N.E., 2000. Predictive habitat distribution models in ecology. Ecol. Model. 135, 147–186.
Guisan, A., Edwards Jr., T.C., Hastie, T., 2002. Generalized linear and generalized additive models in studies of species distributions: setting the scene. Ecol. Model. 157, 89–100.
Hoerl, A.E., Kennard, R.W., 1970. Ridge regression: biased estimation for nonorthogonal problems. Technometrics 12, 55–67.
Kanevski, M., Parkin, R., Pozdnukhov, A., Timonin, V., Maignan, M., Demyanov, V., Canu, S., 2004. Environmental data mining and modeling based on machine learning algorithms and geostatistics. Environ. Model Softw. 19, 845–855.
Knudby, A., Brenning, A., LeDrew, E., 2010. New approaches to modelling fish–habitat relationships. Ecol. Model. 221, 503–511.
Lee, K.-S., Short, F.T., Burdick, D.M., 2004. Development of a nutrient pollution indicator

using the seagrass, *Zostera marina*, along nutrient gradients in three new england estuaries. Aquat. Bot. 78, 197–216.
Lehner, B., Verdin, K., Jarvis, A., 2006. Hydrosheds Technical Documentation. version 1.0. World Wildlife Fund US, Washington, DC. pp. 1–27.
Li, J., Heap, A.D., 2011. A review of comparative studies of spatial interpolation methods in environmental sciences: performance and impact factors. Ecol. Inform. 6, 228–241.
Li, J., Siwabessy, J., Tran, M., Huang, Z., Heap, A., 2013. Predicting Seabed Hardness Using Random Forest in R. Data Mining Applications with R. Elsevier, pp. 299–329.
Li, J., Alvarez, B., Siwabessy, J., Tran, M., Huang, Z., Przeslawski, R., Radke, L., Howard, F., Nichol, S., 2017. Application of random forest, generalised linear model and their hybrid methods with geostatistical techniques to count data: predicting sponge species richness. Environ. Model Softw. 97, 112–129.
Martins, I., Pardal, M., Lillebø, A., Flindt, M., Marques, J., 2001. Hydrodynamics as a major factor controlling the occurrence of green macroalgal blooms in a eutrophic estuary: a case study on the influence of precipitation and river management. Estuar. Coast. Shelf Sci. 52, 165–177.
Merckx, B., Goethals, P., Steyaert, M., Vanreusel, A., Vincx, M., Vanaverbeke, J., 2009. Predictability of marine nematode biodiversity. Ecol. Model. 220, 1449–1458.
Olesen, B., Enríquez, S., Duarte, C.M., Sand-Jensen, K., 2002. Depth-acclimation of photosynthesis, morphology and demography of *Posidonia oceanica* and Cymodocea nodosa in the Spanish Mediterranean sea. Mar. Ecol. Prog. Ser. 236, 89–97.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.
Pérez, M., Romero, J., 1993. Preliminary data on alkaline phosphatase activity associated with Mediterranean seagrasses. Bot. Mar. 36, 499–502.
Ruíz, J.M., Marí-Guirao, L., Sandoval-Gil, J.M., 2009. Responses of the mediterranean seagrass *Posidonia oceanica* to in situ simulated salinity increase. Bot. Mar. 52, 459–470.
Sandoval-Gil, J.M., Marín-Guirao, L., Ruiz, J.M., 2012. The effect of salinity increase on the photosynthesis, growth and survival of the Mediterranean seagrass Cymodocea nodosa. Estuar. Coast. Shelf Sci. 115, 260–271.
Tittensor, D.P., Baco, A.R., Brewin, P.E., Clark, M.R., Consalvey, M., Hall-Spencer, J., Rowden, A.A., Schlacher, T., Stocks, K.I., Rogers, A.D., 2009. Predicting global habitat suitability for stony corals on seamounts. J. Biogeogr. 36, 1111–1128.
Tomasello, A., Di Maida, G., Calvo, S., Pirrotta, M., Borra, M., Procaccini, G., 2009. Seagrass meadows at the extreme of environmental tolerance: the case of *Posidonia oceanica* in a semi-enclosed coastal lagoon. Mar. Ecol. 30, 288–300.
Udy, J.W., Dennison, W.C., 1997. Growth and physiological responses of three seagrass species to elevated sediment nutrients in Moreton bay, Australia. J. Exp. Mar. Biol. Ecol. 217, 253–277.
Volf, G., Atanasova, N., Kompare, B., Precali, R., Ožanić, N., 2011. Descriptive and prediction models of phytoplankton in the northern Adriatic. Ecol. Model. 222, 2502–2511.
Weatherdon, L., Fletcher, R., Jones, M., Kaschner, K., Sullivan, E., Tittensor, D., Mcowen, C., Geffert, J., Bochove, J., Thomas, H., et al., 2015. Manual of Marine and Coastal Datasets of Biodiversity Importance. December 2015 edition. UNEP World Conservation Monitoring Centre. Cambridge, UK.
Wessel, P., & Smith, W. (2013). Gshhg—a global self-consistent, hierarchical, high-resolution geography database. Honolulu, Hawaii, Silver Spring, Maryland.(URL: http://www.soest.hawaii.edu/pwessel/gshhg/(Accessed 10 January2013).
Wiley, E. O., McNyset, K. M., Peterson, A. T., Robins, C. R., & Stewart, A. M. (2003). Niche modelin perspective on geographic range predictions in the marine environment using a machine-learning algorithm.