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