UNIVERSITÀ
DI PARMA

UNIVERSITA' DEGLI STUDI DI PARMA

DOTTORATO DI RICERCA IN
Biologia Evoluzionistica ed Ecologia
CICLO XXXV

Applications of Machine Learning techniques in Ecology

Coordinatore:
Chiar.mo Prof. *Pierluigi Viaroli*

Tutori:
Chiar.mo Prof. *Pierluigi Viaroli*
Chiar.ma Prof. *Valeria Rossi*

Dottorando: *Nicolò Bellin*

Anni accademici 2019/2020 – 2021/2022

Index

Abstract

The size and diversity of ecological data are growing in exponential ways due to the modern advances in informatics applications, web services, and cloud systems that yield a great flux of information available to scientists, stakeholders, and the public.

The great global challenges at the level of nature conservation, biodiversity loss due to anthropogenic effects, global changes, vector epidemiological monitoring, and sustainability are complex problems that require fast and accurate real-time analysis with suitable statistical tools. Machine Learning (ML) is a subfield of Artificial Intelligence (AI) that includes a heterogeneous set of theories and data-driven algorithms that allow computers to capture relationships and hidden patterns "not explicitly given by humans" better than traditional statistical methods.

Here, we trained and evaluated using different approaches a set of ML algorithms, to identify environmental drivers that shape the realized niche of different species and to evaluate the effects of climate change in organisms from freshwater and marine ecosystems. Moreover, ML was used to study wings' shape variation of sibling malaric vectors in a contest of epidemiological surveillance and to identify the influence of chemical and physical environmental features on the assemblage patterns of different freshwater zooplankton communities. A particular branch of ML that acquired importance in the last years, deep learning, was applied to ecoacustics, to demonstrate how deep learning captures different aspects of the marine environment using large marine Passive Acoustic Monitoring (PAM) data.

We demonstrated how the flexibility of the ML algorithms address successfully different ecological problems across taxa and different environments. Finally, data sharing and free AI programs might improve the use of ML in ecology to speed up the process that leads to new scientific discoveries.

Riassunto

Le dimensioni e la diversità dei dati ecologici stanno crescendo in modo esponenziale grazie ai moderni progressi nelle applicazioni informatiche, nei servizi web e cloud. Questi sistemi producono un grande flusso di informazioni disponibili per scienziati, enti legislativi e pubblico.

Le grandi sfide a livello di conservazione della natura, perdita di biodiversità dovuta ad effetti antropici, cambiamenti globali, monitoraggio epidemiologico e sostenibilità ambientale richiedono analisi rapide in tempo reale con metodi statistici adeguati. Il Machine Learning (ML) è un sottocampo dell'Intelligenza Artificiale (AI) che include un insieme eterogeneo di teorie e algoritmi guidati dai dati che consentono ai computer di estrapolare relazioni e ricorrenze nascoste, "non forniti esplicitamente dagli esseri umani", in modo superiore rispetto ai tradizionali metodi statistici.

In questa tesi, sono stati addestrati e valutati una serie di algoritmi ML utilizzando diversi approcci, per identificare i fattori ambientali che modellano la nicchia realizzata di diverse specie, per valutare gli effetti del cambiamento climatico in organismi di acqua dolce e marina e per identificare l'influenza delle variabili ambientali chimico e fisiche dell'acqua nelle diverse comunità di zooplancton d'acqua dolce. Inoltre, il ML è stato utilizzato per studiare la variazione della forma delle ali dei vettori malarici in un contesto di sorveglianza epidemiologica. Un campo particolare del ML che ha acquisito importanza negli ultimi anni, il deep learning, è stato applicato all'ecoacustica, per dimostrare come il deep learning catturi diversi aspetti dell'ambiente marino utilizzando grandi dati di monitoraggio acustico passivo (PAM).

È stato dimostrato come la flessibilità degli algoritmi di ML risolva con successo diversi problemi ecologici considerando taxa e ambienti diversi. Infine, una maggior condivisione dei dati unita a programmi di intelligenza artificiale gratuiti potrebbero migliorare l'uso del ML in ecologia per accelerare il processo che porta a nuove scoperte scientifiche.

1. Introduction

1.1 What is Machine Learning?

Machine Learning (ML) is a subfield of Artificial Intelligence (AI) that includes a heterogenous set of theories and algorithms originating from statistical learning, pattern recognition, and knowledge discovery theories (Harrington, 2012, Shai and Shai, 2014). In 1950, Alan Turing brought to light an important question "*Can machines think?"*, a deeper question that in recent years, due to modern advanced in technologies based on silicon and processing performances, has prompted a swirling amount of new research (Figure 1) (Turing, 1950; Rhys, 2020). In the early days of AI's conceptualization, many researchers and experts hypothesized that with enough knowledge of a system, a set of explicit rules imposed by humans, and good programming skills, computers could be achieved human-level performances (Zhou, 2021). This paradigm was known as symbolic AI. Although symbolic AI was useful in solving a well-defined logical problem, such as playing chess, it has many limitations to solve more complex real problems, where noise, complex relationships, and hidden patterns give rise to reality. ML bears to overcome the limitation of symbolic AI, under the hypothesis that a computer was able to capture relationships not explicitly given by humans (Burger, 2018).

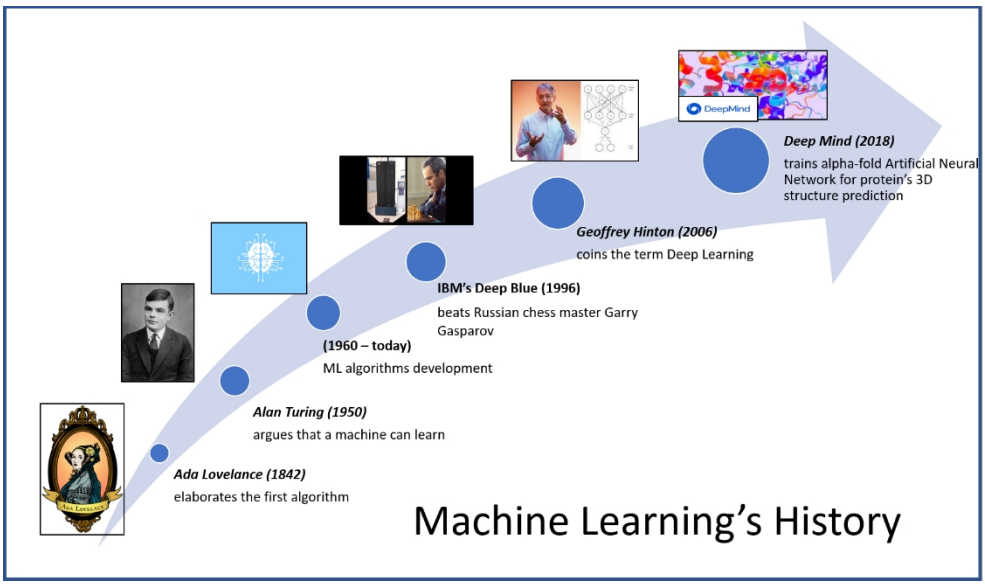

Figure 1 Machine Learning's history: since the XIX century countess Ada Lovelace, an English mathematician, elaborated the first algorithm. In 1950, Alan Turing, the father of AI, published a manuscript (Turing, 1950) in which a philosophical question was made: "Can machines think?". In early 1960, the first ML algorithms were developed. At the beginning of the XX century, the ML discipline experienced a development period interrupted by a period of abandonment of research (AI winters). In 1996, IBM trained an artificial neural network, called Deep Blue, in playing chess and in a famous game beat the chess master player Garry Gasparov. During the first and second decades of the XXI century, with the improvement of specific hardware components (GPUs), new classes of deep neural networks rose. Deep neural networks can solve complex problems such as speech recognition and image processing. In the present day, a plethora of algorithms were trained and made available for the public domain. In 2018, the DeepMind group (Google)

trained an artificial neural network, called alpha-fold, to predict the tertiary structure of proteins from amino acid sequences.

This new paradigm has triggered the ML field, giving rise to methods that automatically learn features and patterns from complex data. ML framework relies on seeing a particular system, possibly under different states, using data collected from that system (Gèron, 2019). The main goal is to make predictions with an algorithm. The algorithm, during the processing data phase, works with an optimization procedure to reduce the predictions' errors related to given task. In simpler words, the ML field depends strictly on three components: the availability of data, a computer with a programming language, and an algorithm that drives the computer to solve real-world problems (Chollet and Allaire, 2018).

1.1.1 Deep Learning

In the last few years, an important branch of ML, Deep Learning (DL), has grown in application in various scientific fields and everyday tasks (Goodfellow et al., 2016) (Figure 2). DL algorithms are become popular considering their performances and the high flexibility. DL is based on deep neural networks, a class of algorithms that achieved the highest accuracy records in image classification (Krizhevsky et al., 2012) and speech recognition (Hinton et al., 2012). Deep neural networks mimic how the biological nervous systems process information (Sejnowski, 2018). The basic elements of interconnected units are called neurons. The neurons are embedded inside layers: an input layer that accepts the predictive variables, one or many hidden layers, and an output layer where a target variable is predicted. DL's algorithms are revolutionizing the automatic feature extraction and systems' knowledge with the computer machine. DL high performant networks arise due to the development of improved algorithms that optimize well the connection weights among neurons and the modern steepness of available computing power and training data (Goodfellow et al., 2016). Classes of deep neural networks that prompted the application of DL in many fields, especially in computer vision tasks, are the Convolutional Neural Networks (CNNs). CNNs are well suited for the automatic classification and features extraction from images and video frames. The CNN architecture consists of convolutional layers and pooling layers that mimic biological visual systems (Wäldchen and Mäder, 2018). CNNs have been applied successfully to several ecological problems, and their use in ecology is growing (Christin et al., 2019, Borowiec et al., 2022). For example, CNNs have processed camera trap images to identify species, age classes, numbers of animals, to classify behaviors patterns (Lumini et al., 2019; Norouzzadeh et al., 2018; Tabak et al., 2019) and in the recognition of mosquitoes borne disease (Goodwin et al., 2021). CNNs have achieved high performances also in the analysis of sounds, for example in the project carried out by the collaboration between the National Oceanic and Atmospheric Administration (NOAA) and Google for the automatic recognition and monitoring of the Humpback Whales, using a network of hydrophones (https://www.fisheries.noaa.gov/science-blog/ok-google-find-humpback-whales). Another special class of deep neural networks are the Recurrent Neural Networks (RNNs) adapted for time series analysis (Kraft et al., 2021).

.

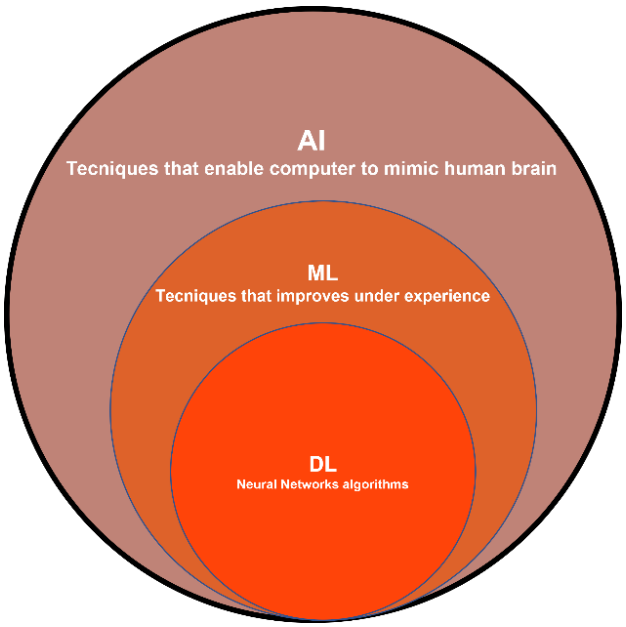

Figure 2 Schematic representation of Artificial Intelligence (AI) with the subfields Machine Learning (ML) and Deep Learning (DL) depicted as inclusive sets.

1.1.2 Big Data and Machine Learning

Nowadays enormous amounts of information are available in the hands of scientists and stakeholders. The rise of Internet, the technological advances that allow capturing and storing massive quantities of data, the variety and speed of collection of raw information, have hinted at the rise of new methods of analysis (Zhou et al., 2017). Moreover, many real-world problems such as conservation priorities, biodiversity loss and climate change effects need fast and correct solutions. The term "Big Data" is applied to datasets that grow so large that traditional database management systems become ineffective. The Big Data sizes are beyond the ability of commonly used software and storage systems. Big Data's sizes are constantly increasing, currently ranging from terabytes (TB) to many petabytes (PB). Some of the major difficulties related to big data include capture, storage, search, sharing, analytics, and visualizing (Elgendy and Erlang, 2014). The classical statistical approaches have become inappropriate or have less statistical explanatory power to extract insight into this great amount of data (Qiu et al., 2016). Normally in the learning phase, Machine Learning algorithms requires many training data due to the large number of parameters that must be tuned during some gradient based optimization to converge and found an optimal solution. Another important requirement to train ML with Big Data is the computational capacity of the machine. Computer performances have also risen exponentially following the famous Moore's Law due to a mix of improvements in hardware technology, architectural innovations, and compiler optimizations (Gustafson, 2011). Modern laptops can process a large quantity of data and make greater multitask processing and parallel computation combined with the lower cost of the electronic components than before. Moreover, the

advent of free and open-source programming languages, kept developed by a community of informatics and scientists, has helped the widespread use and development of ML algorithms (Al-Jarrah et al., 2021).

In the Big Data era, ML has undergone a process of improvement with new techniques and algorithm discoveries. The novel set of ML tools that rely on learning from experiences have optimized the synthesis and automation of the information's flow carried by multidimensional data. Considering the task or the problem to solve, three main categories of ML approaches were developed: supervised learning, unsupervised learning, and reinforcement learning (Figure 3).

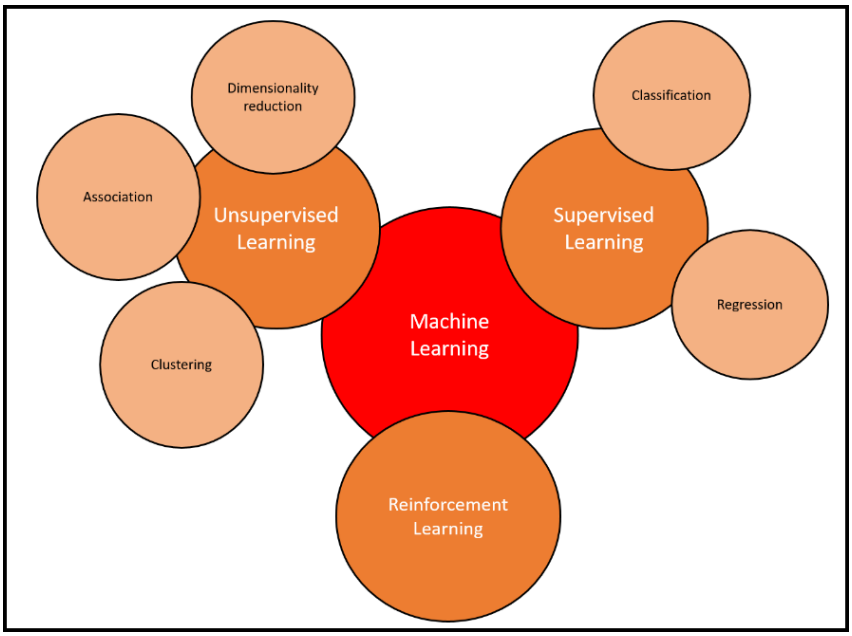

Figure 3 Representations of the three main approaches in ML: supervised learning, unsupervised learning, and reinforcement learning. The supervised learning approach is used to solve a problem related to classification and regression. Unsupervised learning deals with a problem related the dimensionality reduction and clustering. Reinforcement learning is a relative ML approach where an agent learns to solve problems in a complex environment.

1.1.3 Supervised Learning

Supervised learning is an ML approach that relies on training an algorithm with human expert supervision (James et al., 2014). The goal is to correctly predict a dependent variable with the lowest error rate. Four different steps make supervised learning: 1) split of the observations that belong to a given dataset into different sets, 2) training phase, 3) evaluation, and 4) prediction (Figure 4) (Kotsiantis, 2007).

In the first phase of the supervised approach, the main data set is split into three different subsets: training set, validation set, and testing set. The training set is used to train the model (Rhus, 2020). Usually, the training set is composed of 70-80 % of the total observations. Each observation is labeled according to prior knowledge of a human expert, including inputs and correct outputs, which drive the model to learn rules and patterns. The validation set is composed of less observation than the training set and it is used as an independent set during the training procedure. The validation set allows us to make evaluations of instances

not previously seen from the algorithm. If sufficient data are available, a further independent set might be used: the testing set. The testing set is used after the training procedures and the operations of hyperparameters' tuning, to evaluate the improvement of the algorithm prediction's capability (Chollet and Allaire, 2018).

During the training phase, the algorithm measures its accuracy using a loss function or some performance metrics. The goal of the algorithm is to adjust its parameters until the loss (or a performance metric) has been minimized (or maximized) in a process of optimization. After the training and evaluation process, the model can infer predictions of new data. As new observations increase, the model might be updated to further improve the accuracy (Suthaharan, 2016).

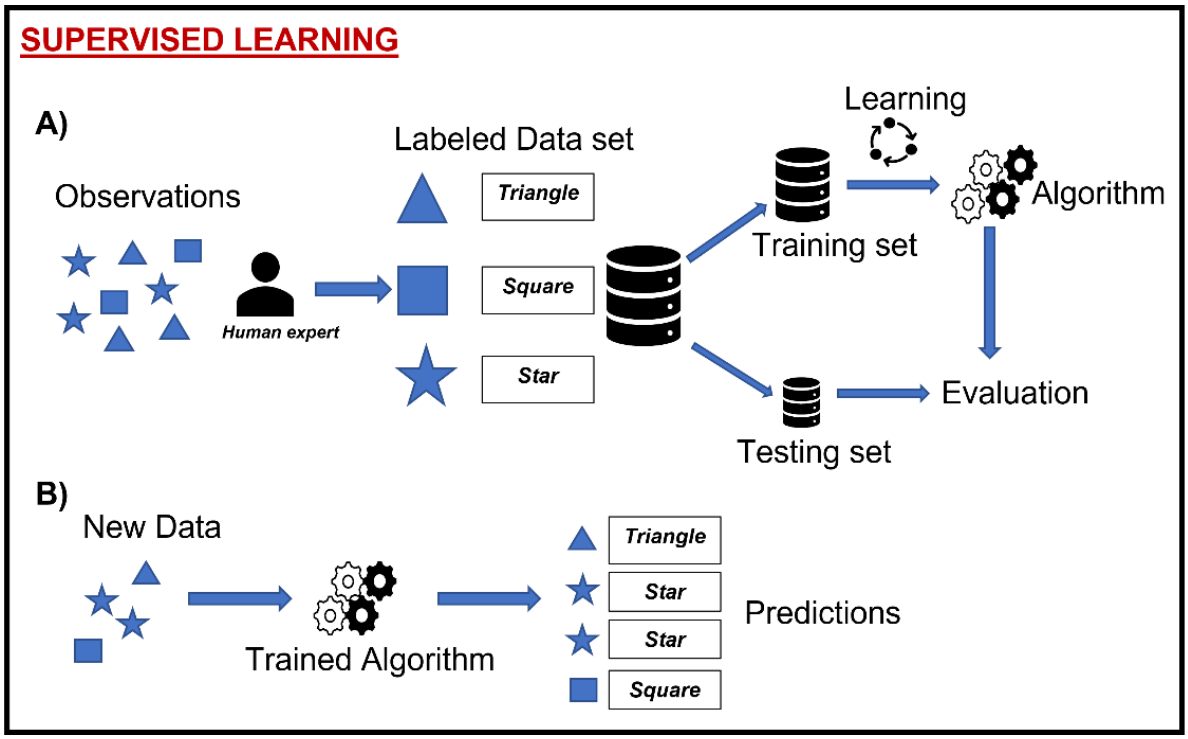

Figure 4 Schematic view of the Supervised Learning approach. In panel A) a human expert labels the raw observations. The labeled data set is split into training and testing sets. The algorithm learns features from the observations in the training set. A final independent evaluation is made on the testing set. In the panel, B) new data are analyzed by the trained algorithm (model) to obtain predictions.

Supervised Learning is used to solve classification or regression tasks (Figure 5). The classification task is defined when the independent variable is categorical (Borowiec et al., 2022). The categorical variable might show several levels (classes) with a number greater or equal to two. Given a set of explanatory variables, the scope of the algorithm is to find a function that correctly maps the values of the explanatory variable's space toward the classes of the independent variable. When the classes are two, for example in presence/absence data, the classification problem is defined as binary. When the goal is to classify more than two classes, the problem is defined as multiclass classification. In the binary and multiclass classification approach, the classes are considered mutually exclusive; there are no overlaps of different classes in a single instance

(observation). When the classes are not mutually exclusive, the classification task is defined as multilabel classification.

The regression task is performed when the independent variable is continuous (numerical). The goal of the algorithm is to combine numerically a set of explanatory variables to obtain a continuous output of the dependent variable (Crisci et al., 2012). During the training procedure, an error term is computed between the prediction and the reference value for a particular instance. The information gained after the computation of the error term is backpropagated to reduce the error of future predictions in an iterative way.

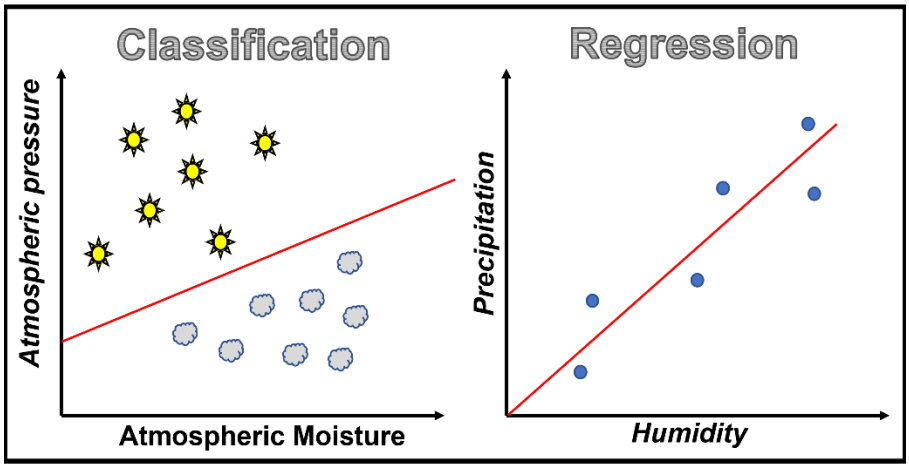

Figure 5 The left panel reports an example of a classification approach. The classification refers to two different classes: sunny and cloudy days. The algorithm fit a decision boundary (red line) that separates regions of the variables' space to correctly map two classes (sunny and cloudy days) using the information from two independent variables (atmospheric moisture and atmospheric pressure). In the right panel, a simple regression problem was depicted as the relationship between an independent variable (precipitation) and a dependent variable (humidity). The algorithm found a function (red line) that predicts the values of the dependent variable with the minimum error.

Several algorithms that belong to the supervised framework were developed, and four main categories are recognized: logic-based, kernel-based, statistics-based, and lazy algorithms (Mandal and Bhattacharya, 2020).

Logic-based algorithms deal with the tasks of classification or regression with a step-by-step procedure and logic rules are applied in each step. For example, one of the most popular and fundamental logic-based algorithms is the decision tree, used for both classification and regression (Breiman et al., 1984). Decision trees generate a set of decision sequences with a recursive partition method that lead to a particular prediction. The decision tree consists of nodes connected to a root with no incoming edges. The node that has outgoing edges is called internal node. The rest of terminal nodes are the leaves. For example, in classification problem, each leaf is related to one class and represents the prediction. The leaf may hold a probability vector that indicates the probability of the target class having a certain value. The single observations are classified by following the path from the root of the tree down the leaf. Decision tree is the basic unit of many complex logic ensemble learning algorithms built with bagging or boosting methods

(Strobl et al., 2009). Bagging, known as bootstrap aggregation, is made by many weak models (decision trees) trained independently on data samples generated by bootstrap. Usually, the result of each decision tree is considered, and the majority vote is selected as prediction. Bagging is a powerful method that reduce the variance within noisy datasets. A powerful algorithm, the random forest, is an example of bagging method that generate a forest of hundreds/thousands of decision trees (Oshiro et al., 2012). Boosting methods aim to produce a series of weak learners using a sequential learning process, and they can predict more accurate outcomes than a single weaker classifier. Boosting methods, in many cases, shown the best classification performance in different applications (Li et al., 2019; Osman et al., 2021; Kumar and Kumar 2021; Pandeyz et al., 2021) and the most popular algorithms are the AdaBoost and XGBoost.

Kernels-based methods use linear classifiers to solve complex non-linear problem by projecting the data into higher dimensions to facilitate a linear separable task (Mandal and Bhattacharya, 2020). The support vector machines (SVMs) are the most important algorithms used to analyze data for both classification and regression tasks. For example, in classification, SVM maps training observation toward a higher dimensional space and drawn a hyperplane (support vector) that relate different region of the space to specific classes (James et al., 2014). New observations are then projected into that same space and predicted to belong to a given class based on which side of the hyperplane they fall (Qiu et al., 2016).

Statistics-based algorithms generalize problems with probability density functions to predict or solve different tasks (Mandal and Bhattacharya, 2020). For example, Naïve Bayes is a popular and simple statistics-based algorithm for predictive modeling that rely on the Bayesian theorem of probability.

The last category, lazy algorithms, known also as “instance-based”, delay the process of generalization until the classification task is performed (Mandal and Bhattacharya, 2020). For example, the KNN or K-Nearest Neighbor is an instance-based learning algorithm that stores all available records and predicts the class of a new observation giving attention to similarity measurements from the nearest neighbors of that observation (James et al., 2014). KNN is a simple classification algorithm but despite the simplicity, it can produce highly competitive results. It can deal with both classification and regression types of predictive problems. However, it is more used to perform and execute classification task.

1.1.4 Unsupervised Learning

In Unsupervised learning ML approach, a given algorithm uses unlabeled data to find hidden patterns and groups without prior human information (James et al., 2014; Celebi and Aydin, 2016). The unsupervised approach works to find similarities or differences among data and consists of three main different techniques: dimensionality reduction, clustering, and associations (Alloghani et al., 2020).

The dimensionality reduction is used when a large amount of information is present in a data set (Velliangiri et al., 2019). The entropy, the complexity, and the great number of recorded variables might make it difficult to highlight relationships and hidden patterns. The dimensionality reduction aims to reduce the complexity of

the information by reducing the number of variables in a dataset. The data dimensions' compression is made by minimizing the loss of information and maximizing the great amount of variance explained.

Clustering is a set of operations performed by specific algorithms, clustering algorithms, used to group raw and unlabeled data (Omran et al., 2007). Clustering algorithms can be separated into different types: exclusive clustering (hard clustering), overlapping clustering (soft or fuzzy clustering), hierarchical and probabilistic. The exclusive clustering groups the observations into different clusters and a particular observation belongs to only one cluster. Overlapping clustering instead allows a given observation to belong to many clusters simultaneously with a degree of membership. Hierarchical clustering works with a sequential fusion of similar clusters using two different approaches: agglomerative (bottom–ups) or divisive (ups–bottom). Probabilistic clustering uses the density and the estimated distributions learned from data to group the observations with the likelihoods.

Association is a set of methods that use rules to find relationships among different variables (Diaz-Garcia et al., 2022). The association rules can pinpoint frequent patterns or collections of observations that frequently recur together, estimating the likelihood of recurrences. The association rules are defined as data mining procedures, due to the high frequency of application in the first phase of the data science pipeline.

1.1.5 Reinforcement Learning

Reinforcement learning (RL) is an ML approach that considers agents and dynamic environments (Frankenhuis et al., 2019) (Figure 6). Unlike supervised and unsupervised ML, RL does not rely on a static dataset, but works in a dynamic environment and learns from experiences (Wang et al., 2022). The agent tries looking for solutions considering a given problem, throughout trial-and-error attempts. In each time step, the agent shows a given state and opts for an action, defining a coupled state-action system. The agent interacts with the environment for thousands of time steps. The amount of time needed for an agent to learn and found the optimal policy cannot be anticipated or predetermined and depends on many factors, including both the complexity of the agent and environment. The set of actions performed by a given agent is called action space. The action space can be described as discrete or continuous. In a discrete action space, the agent interacts with the environment throughout quantized and finite actions, for example the directions of the movement allowed on a plane environment (top, bottom, right and left). In a continuous action space, the actions allowed are not quantized, such as the velocity or different angles of movement. In the RL framework a wide variety of environments can be implemented: deterministic, stochastic, sequential, or episodic. In a deterministic environment, the next state is determined based on the current state, and it is predictable, while in a stochastic environment the future state is aleatory and cannot be always predicted. In a sequential environment the agent's actions relate to the previous actions it made, while in episodic, actions are not time related. In each environment can be introduced a single agent or many agents with its own policy and actions to take. The agent received a reward or a penalty from the environment after performing an action. At each time step, the agent adapts its actions following a particular policy. The policy is a set of

rules that updates and suggests new future actions as function of the past agent's rewards. For example, deep neural networks are used as agents trained with RL framework to mimic and encode complex behaviors (Hirakawa et al., 2018). This approach is flexible in non-linear systems when the traditional methods might fail. The agent did not receive prior training or human knowledge for a specific action to take it find out the action will yield to the greatest reward, exploring the set of possible actions. RL approach is used to solve problems in the control of nonlinear systems, autonomous driving, robotics, and planning problems. In ecology, RL framework was applied successfully to unsolved management scenarios in fisheries stock conservation and ecological tipping points (Lapeyrolerie et al., 2022), in behavioral ecology (Frankenhuis et al., 2019), to understand predator – prey systems (Wang et al., 2020) and to predict animal movement (Hirakawa et al., 2018).

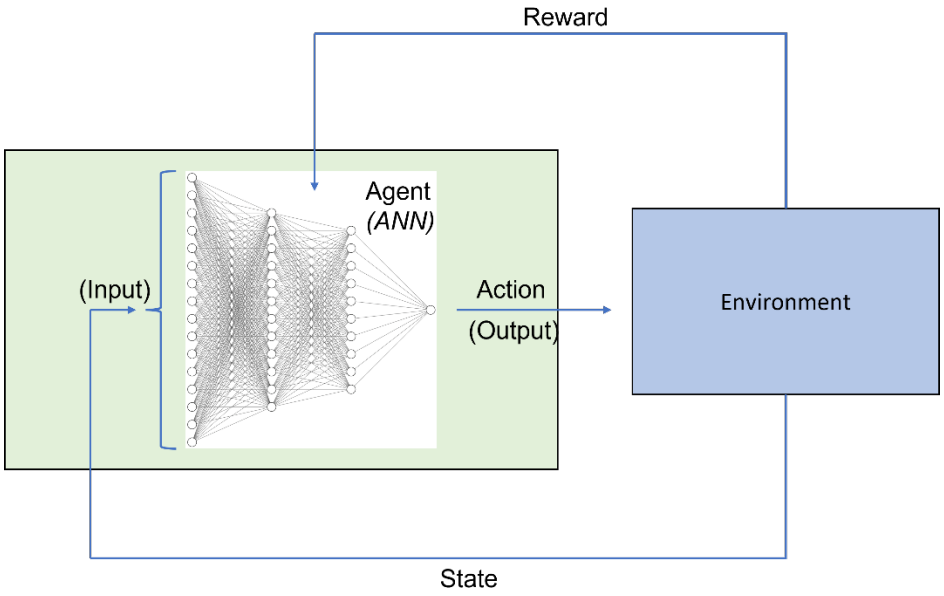

Figure 6 RL framework is defined by an agent (artificial neural network (ANN) algorithm) that perform actions as outputs in a complex environment. The environment influences the state of the agent (input) and gives a reward for the previous action. In an iterative way, the agent learns to maximize the rewards modifying its hyperparameters and consequently adjusting the sequence of actions to take.

1.2 Applications of Machine Learning in Ecology

Systems and mechanisms underlying ecological processes exhibit complex and nonlinear relationships and a wide variety of instruments and technologies generate a great size of variable data (Farley et al., 2018) (Figure 7).

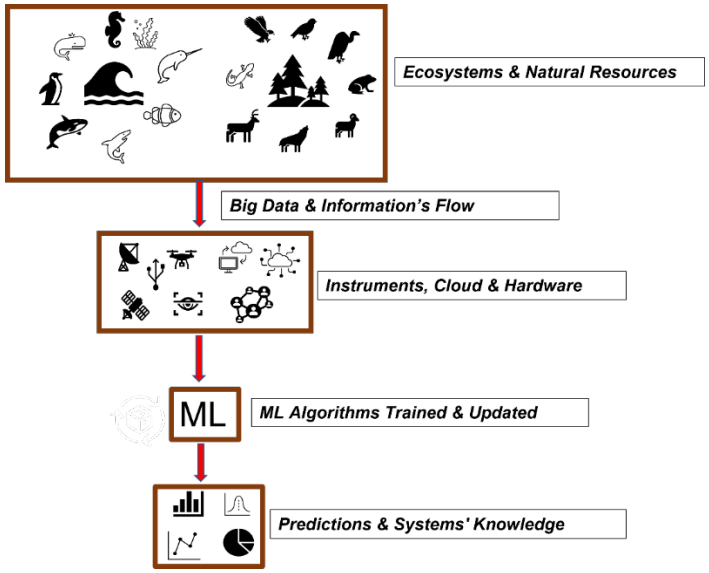

Figure 7 The ML framework was depicted considering ecological systems. The data are acquired from different types of sensors, giving a great variety of collected information (Big Data). The ML algorithms are trained in a cycle of improvement and the result is a deep knowledge of the system thanks to automatic processing.

Data diversity is a key component of the modern advances in ecology, especially in monitoring the population dynamics and community systems, in study involving functional traits (Vasseour et al., 2022), ecological networks (Pichler et al., 2019), morphological diversity (Lailvaux et al., 2022), global changes impacts (Humphries et al., 2019) and phylogeny (Azouri et al., 2021). The size of ecological data is still growing in an exponential way due to the modern advances in informatics application, web services and cloud systems that yield a great flux of information available to the public, scientists and stakeholders, also from remote and non-accessible areas (Christin et al., 2019). This amount of information is interconnected in a complex growing network. As the complexity of ecological data increases, many statistical problems might arise: different types of data's noise, confounding drivers that hide information and numeric biases. As reported in Crisci et al. (2012), another problem that modern ecological datasets are facing is the Bellman's curse of dimensionality, defined as the increment of complexity carried by dataset. Classical statistical methods often model linear relationships and rely on the use parametric statistics. These methods are often not appropriate due to prediction biases and strong prior assumptions. Several studies show that ML techniques outperform traditional statistical methods in ecology (Elith et al., 2006; Cutler et al., 2007; Olden et al., 2008; Zhao et al. 2011; Lucas, 2020), although not ever systematically (Keller and Dzerosky, 2011). Moreover, in recent years, ML predictions were constructed as black boxes, that is as systems that hide their internal logic to the user (Guidotti et al., 2018). This lack of explanation constitutes both a practical and an

ethical issue. However, the literature reports many approaches aimed at overcoming this crucial weakness (Saltelli et al., 2020; Lundberg and Lee, 2017).

ML applied to environmental sciences, including ecology, increased in an exponential way, with an acceleration in the last 5 years (Figure 8). Moreover, many real-world problems deserve priorities and need fast and accurate solutions. Considering the rate at which biodiversity is declining, the lack of information related to many species (up to 17000 data deficient species by IUCN Red List Threatened Species) and the rapid shifting of the earth systems, ML tools might automatize, predict and help human experts in assessment on different spatiotemporal resolutions.

This thesis will focus on a series of ML techniques with ecological applications. The first and second chapters describe the integration of the Species Distribution Models (SDM) and the Geometric Morphometrics (GM) frameworks, respectively, with ML tools. In the third chapter, the application of ML to community ecology is reported. In the last chapter a comprehensive view of ML algorithms applied to ecoacustics discipline is described.

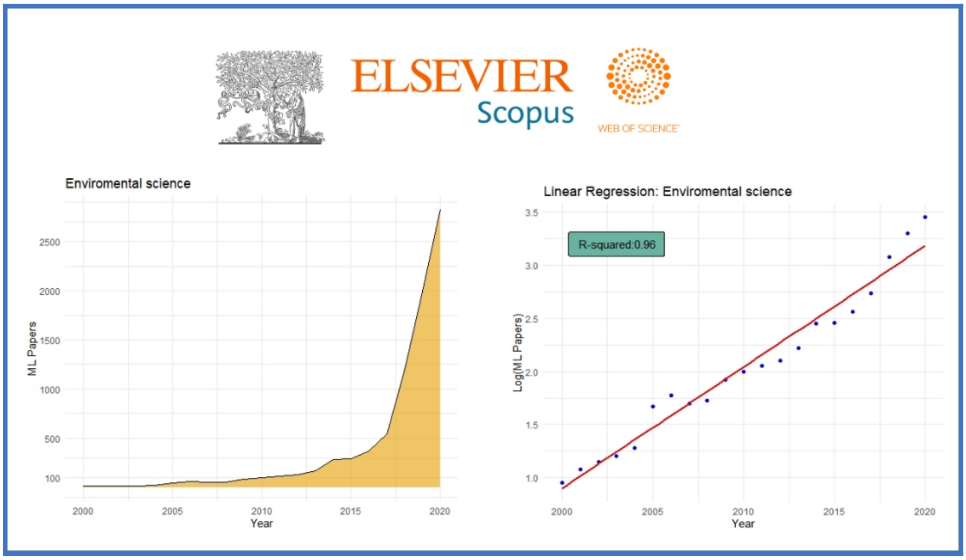

Figure 8 Published literatures on the web of science, the details of the research were: a time interval of 20 years, the key word Machine Learning, and the environmental science as research domain. In the left panel, the number of papers that report ML techniques was reported as function of years. In the right panel, the same relationship was reported in log base 10 scale with a linear regression fit (red line).

1.2.1 Species distribution models

In ecology, an important application of ML methods involves studies that correlate the relationship between organism and the habitat characteristics, especially in the framework of Species Distribution Models (SDM). Species Distribution Modelling (SDMs) are quantitative empirical methods predicting the probability of occurrence of species or habitats using a suite of environmental predictor variables (Elith et al., 2006; Elith and Leathwick, 2009; Melo-Merino et al., 2020; Charney et al., 2021; Guisan and Zimmermann, 2000; Zurell et al., 2020). The species–environment relationships were developed using species location data

(abundances or occurrences) and those environmental variables thought to influence species distributions. Models describing the distribution of species and their niche have been called in many ways: “bioclimatic envelope models”, “habitat suitability models”, “species distribution models”, and “ecological niche models” (Melo-Merino et al., 2020). These different terms led to confusion of model results and interpretability (Peterson and Soberón, 2012; Soberón et al., 2017). Although the correct use of terms is under debate, two main terms referring to different type of models have acquired some acceptance in scientific literature: ecological niche models (ENM) and species distribution models (SDM) as model’s frameworks that target to answer to different questions (Peterson and Soberón, 2012). ENMs model the fundamental niches of species while SDMs aim to predict occurrences in geographic space (Peterson and Soberón, 2012; Soberón et al., 2017). SDMs and ENM may predict species’ distributions in unknown locations or in different time scales, as well as niche shifts under processes of disturbance, invasion, or speciation. The preparation of spatial distribution maps could play an important role in studying the current and future impacts of global changes to species habitat and understanding the extinction rate in natural ecosystems (Franklin, 2010; Van Echelpoel et al., 2015).

As remote sensing and weather monitoring systems have improved, the inclusion of a wide number of environmental variables that carry information for both terrestrial and marine ecosystems are available at the hand of scientist. The wide set of covariates is a powerful source of information to describe the land cover types, the temperatures, the precipitation regime, the soil composition, and many other physisco-chemical and topological drivers that might shape the species’ niches. Another fundamental source of information derives from geolocalized spatial observations of a species or for a set of taxa that compose an ecological community. Species records can be available as presence/absence or presence only data. Presence/absence are binary information of species occurrence, containing notifications of sites of the target species sampled from field campaigns/studies. Presence-only data are presence records of the target species through a random encounter within a particular region and nowadays are primarily collected as citizen science data or by camera traps. Many georeferenced, historical, or real-time observations of species occurrences are becoming available on internet open-access platforms. For example, a famous platform is the Global Biodiversity Information Facility (GBIF), a dense network of data that provides open access to data about all types of life on Earth (https://www.gbif.org/what-is-gbif). However, all distributional platforms such as GBIF could retain spatial biases due to unbalanced sampling effort, data acquisition and sharing (Beck et al., 2014). Such differences at the level of nations with different funding and networks, might lead to model distortions and bias in predictions. These aspects if not taken into account with appropriated methods, such as subsampling and spatial thinning, will limit the automatization of SDMs to large distributional databases (Flemons et al., 2007; Guralnick and Hill, 2009).

In the framework of species distribution models, many statistical methods and algorithms are available to model the spatial distribution of a target species (Williams et al., 2009; Elith et al., 2008; 2011; Assis et al., 2014; Rocchini et al., 2019; McKenna and Kocovsky, 2020). Regression based methods widely used are Generalized Linear Models (GLM), Generalized Additive Models (GAM) and regression splines (MARS)

(Becker et al., 2020; Mateo et al., 2010). Other methods of the ML field are growing in literature: conquer and divide approaches (classification and regression tree); artificial neural network (ANN) and maximum entropy (MAXENT) (Elith et al., 2011; Hill et al., 2017; McKenna and Kocovsky, 2020; Williams et al., 2009).

SDM are successfully used to predict habitat suitability of species over space and time, tracing the relationship between specific environmental conditions and the probability of occurrence of the organisms and produced response curves that reflect the species' realized environmental niche (Guisan and Zimmermann, 2000; Zurell et al., 2020). For example, SDMs have been used to identify climate refugia in marine environments, which is crucial for conservation, as they are likely to facilitate the persistence of ecologically and economically relevant species (Davis et al., 2021). SDM was used also to define niche breadth of specific organisms (Murphy and Smith, 2021).

The main applications of SDM's integrated with ML tools are in the control of invasive species (Garcia et al., 2022; Konowalik et al., 2017), in epidemiology of vector – borne disease (Akpan et al., 2019), in paleobiology (Svenning et al., 2011), in monitoring strategies (Khwarahm et al., 2021; McKenna and Kocovsky, 2020), and for the identification of sites with high likelihood to observe rare/cryptic species (Fois et al., 2018). A recent project is an interactive app of mosquito's distribution to assist vector-borne disease management in Western Europe (http://arset.gsfc.nasa.gov/) combining information that derives by the NASA's satellites and the Global Mosquito Alert Consortium's citizen science.

1.2.2 Geometric Morphometric

The geometric morphometric (GM) is a useful tool to investigate shape variation among and within species that are difficult to discriminate with standard taxonomic approaches (Petrarca et al., 1998; Ayala et al., 2011; Marquez et al., 2011; Lorenz et al., 2015b; Gomez and Correa, 2017). The landmarks coordinates identify points located on the body of the individual subjected to morphometric analysis; these landmarks are homologous in evolutionary terms. Their configuration retains information about the size and shape of single individuals. The superimposition with the generalized procrustes analysis removed the effect of orientation, translation, and scaling, allowing the analysis and the comparison of the shapes expressed as procrustes coordinates (Bookstein, 1991). Furthermore, the shape coordinates are processed with statistical multivariate methods for shape visualization and classification. In vector epidemiological studies Wilke et al. (2016) used geometric morphometric and discriminant analysis to identify 12 species of three different genera of mosquito with a correct reclassification of 99% on different genera and 96% on different subgenera. Lorenz et al. (2012) recognized three malaric vector species of the genera *Anopheles* (*An. bellator*, *An. cruzii* and *An. homunculus*) in the Brazilian Atlantic Forest with the geometric morphometric applied to wing shape. Another approach involves the use of ML algorithms. Lorenz et al. (2015a) combined geometric morphometric with an artificial neural network (ANN) to classify 17 species of the genera *Anopheles, Aedes* and *Culex* that are vectors of different pathogens and ANN reached a higher classification accuracy of

species than traditional multivariate methods. GM and ML are combined also in studies that address a great variety of biological questions (Mapp et al., 2017; Nattier et al., 2017; Soda et al., 2017; Fang et al., 2018). For example, Quenu et al. (2020) investigate size and shape variations of *Placostylus* snail's shells to search phenotypic clusters and delimit snail species using both supervised (ANN) and unsupervised approach (Gaussian Mixture Models). Lloyld et al. (2019) used the GM framework and 9 common ML algorithms in the analysis of carnivore tooth marks, obtaining classification scores of 100%.

1.2.3 Community ecology

An ecological community is a group of interacting species located in the same geographical area. Communities are connected in the same environment where a network of relationships among species arise and change through time and space (Mittelbach et al., 2019). Community ecology is a complex subfield of ecology and ecologists study the factors that shape biodiversity and community structure. The abiotic dimension is an important driver that filters and shapes the species occurrence, abundance, and community composition. Moreover, the biotic component (competition, predation, commensalism, etc.) and the connectiveness of different habitats are other key factors that might act from finer to regional scales contributing on species co-occurrences patterns and distributions (Ovaskainen and Abrego, 2020).
In community ecology, data might show biases and stochasticity, correlated variables and many environmental predictors compared to the number of samples available (Crisci et al., 2012). In community ecology, a lot of techniques are used to explore environmental and biological relationships such as multivariate analyses and classical clustering algorithms, but these techniques require many statistical assumptions and are less powerful than ML to deal with noise and non-parametric distributions (Viana et al., 2022). ML unsupervised algorithms were used to reveal temporal variations in communities (Chon et al., 2000), to classify ecological associations in marine ecological communities (Fiorentino et al., 2017) and to identify cryptic spawning sites for a fish species in combination with supervised learning (Brownscombe et al., 2020). In the study of microbial communities, Sperlea et al. (2021) used a machine learning-based framework for the quantification of the covariation between microbiomes and 27 environmental variables of lake ecosystems. Other studies demonstrated that ML could predict distribution models of a target species based on ecological interactions with other species (Chen, et al., 2016). ML methods could become the avenue for studying ecological interactions (Desjardins-Proulx et al., 2017). Recurrent networks have also been shown to successfully predict abundance and community dynamics based on environmental variables for phytoplankton (Jeong et al., 2001) and benthic communities (Chon et al., 2001).

1.2.4 Ecoacustics and sounds analysis

Data might be collected by microphones or passive acoustic technologies, where real time records of sounds represent different soundscapes or the complex calls of species along time and space (Dufourq et al., 2022).

Environmental sounds provide a proxy to investigate ecological processes (Gibb et al., 2019; Rycyk et al., 2020), including exploring complex interactions between anthropogenic activity and biota (Erbe et al., 2019; Kunc et al., 2016). Sound provides useful information on environmental conditions and ecosystem health, allowing, for example, the rapid identification of disturbance (Elise et al., 2019). In concert, numerous species (i.e., birds, mammals, fish, and invertebrates) rely on acoustic communication for foraging, mating, reproduction, habitat use and other ecological functions (Eftestl et al., 2019; Kunc and Schmidt, 2019; Luo et al., 2015; Schmidt et al., 2014). Noise produced by anthropogenic activities (e.g., vehicles, stationary machinery, explosions) can interfere with animal communication, affecting the health and reproductive success of several taxa (Kunc and Schmidt, 2019). In response to concerns about noise pollution, increasing effort is being invested in developing, testing, and implementing noise management measures in both terrestrial and marine environments. Consequently, Passive Acoustic Monitoring (PAM) has become a mainstream tool in biological monitoring (Gibb et al., 2019). PAM represents a set of techniques that are used for the systematic collection of acoustic recordings for environmental monitoring. It allows collecting large amounts of environmental information at multiple locations and over extended periods of time. One of PAM's most common applications is in marine mammal monitoring and conservation. Marine mammals produce complex vocalizations that are species-specific (if not individually unique), and such vocalizations can be used in estimating species' distributions and habitat use (Durette-Morin et al., 2019; Kowarski and Moors-Murphy, 2020). PAM applications in marine mammal research span from the study of their vocalizations and behaviors (Madhusudhana et al., 2019; Vester et al., 2017) to assessing anthropogenic disturbance (Nguyen Hong Duc et al., 2021). PAM datasets can reach considerable sizes, particularly when recorded at high sampling rates, and projects often rely on experts to manually inspect the acoustic recordings for the identification of sounds of interest (Nguyen Hong Duc et al., 2021). For projects involving recordings collected over multiple months at different locations, conducting a manual analysis of the entire dataset can be prohibitive, and often only a relatively small portion of the acoustic recordings is subsampled for analysis. Passive Acoustic Monitoring (PAM) datasets can reach considerable sizes, particularly when recorded at high sampling rates and often rely on experts to manually inspect the acoustic recordings to identify the sounds of interest (Nguyen Hong Duc et al., 2021). Deep Learning algorithms are suitable to deal with analysis of audio files. In particular, CNNs algorithms have been applied successfully to several ecological problems, and their use in ecology has been growing (Christin et al., 2019), such as to process camera trap images to identify species, age classes, numbers of animals, and to classify behavior patterns (Lumini et al., 2019; Norouzzadeh et al., 2018; Tabak et al., 2019). CNN's algorithms perform well also for acoustic classification (Hershey et al., 2017), including the identification of a growing number of species vocalizations such as crickets, cicadas and mosquitoes (Dong et al., 2018; Fernandes et al., 2020; Kiskin et al., 2020), birds and frogs (LeBien et al., 2020), fish (Mishachandar and Vairamuthu, 2021), and marine mammals (Usman et al., 2020). The latter include training neural networks for detecting North Atlantic right whale calls using a mix of real and synthetic data (Padovese et al., 2021), and the classification of Sperm

Whale clicks (Bermant et al., 2019). Currently, most CNN applications focus on species detection rather than a broader characterization of the acoustic environment.

1.3 Aims

The aim of the thesis was to apply ML techniques to a diversified set of ecological data and problems.

In ***Chapter 1***, three different cases of study were shown to demonstrate that the combination of the SDM framework with ML is a useful tool to investigate species distribution and niche and to highlight conservation priorities. A set of different ML algorithms were trained and evaluated using the supervised learning approach, to understand environmental drivers that shape the realized niche of different species and to evaluate the effects of climate change in freshwater and marine ecosystems.

The first case of study reported the investigation of ecological drivers that influences the distributions of two different freshwater zooplankton species: *Eucyclops serrulatus* (Copepoda) and *Daphnia longispina* (Cladocera) in a system of 283 shallow and ephemeral freshwater habitats in the Northern Italian Appennines. For each species, we model the habitat suitability by comparing one regression-based model, one generalized linear model (GLM) and two ML algorithms: random forest (RF) and artificial neural network (ANN) with one hidden layer. We used a total of 27 predictor variables. The modeling framework was also used considering a scenario of future climate change to evaluate potential shifts in spatial distribution of the zooplankton species.

In the second case of study, the SDM's framework was combined with different ML algorithms to predict the distribution of three gorgonian species (soft corals): *Paramuricea clavata*, *Eunicella cavolini* and *Eunicella singularis*. The study was performed in a marine area along the North-West Mediterranean Sea. The niche and the spatial distribution were modelled considering present and a future worst emission scenario of climate change (RCP8.5).

In the last case of study, ML algorithms were used to investigate the threats on the conservation status of two Mediterranean solitary corals (Scleractinia): *Balanophyllia. europaea* (endemic and zooxanthellate) and *Leptopsammia pruvoti* (non-endemic and azooxanthellate) in a marine area that extend to the entire Mediterranean Sea. A total of 13 environmental variables and four different machine learning algorithms were tested to obtain present-day potential habitat suitability for the species and future environmental change scenarios (2040-2050).

In ***Chapter 2*** a series of ML approaches, both supervised and unsupervised, was combined with the GM framework, for the automatic recognition of four malaric sibling mosquito's species (Maculipennis complex) and to study inter-intraspecific diversity. The study of wings' shape variation with a supervised approach was performed in a contest of epidemiological surveillance. Moreover, the inter-intraspecific diversity of the wing shape was investigated and described using the unsupervised approach with dimensionality reduction (UMAP) and clustering algorithms (HDBSCAN).

In ***Chapter 3,*** unsupervised ML techniques and Data Mining procedures were applied to assess the factors influencing the assemblage composition and distribution patterns of 12 zooplankton taxa in 24 shallow ponds located in Northern Italy. Fuzzy sets are suitable descriptors of ecological communities as compared to other standard algorithms and allow the description of decisions that include elements of uncertainty and vagueness. However, fuzzy sets are scarcely applied in ecology. The fuzzy c-means algorithm was implemented to classify the ponds in terms of taxa they support, and to identify the influence of chemical and physical environmental features on the assemblage patterns. Moreover, association rules were used to summarize and disentangle the associations among taxa within the zooplankton community.

In ***Chapter 4*** an alternative to the use of ecoacoustics indices with the application of multiple machine learning techniques for soundscape and vocalizations of marine mammals' analysis was reported. A combination of pre-trained acoustic classification model (CNN), dimensionality reduction, and random forest algorithms were used to demonstrate how machine-learned acoustic features capture different aspects of the marine environment using large PAM data. Two different datasets were analyzed showing how acoustic features extracted by ML algorithms can be used to discriminate between the vocalizations of marine mammals, beginning with high-level taxonomic groups, and extending to detecting differences among conspecifics belonging to distinct populations. Discrimination amongst different marine environments and monitoring of anthropogenic and biological sound sources were also performed.

2. Chapter 1: Species distribution models

2.1 Species distribution modeling and machine learning in assessing the potential distribution of freshwater zooplankton in Northern Italy

Freshwater zooplankton inhabits a wide range of aquatic ecosystems, from temporary pools to lakes, and has an important role in structuring food webs and in sustaining ecosystem health (Schindler and Scheuerell, 2002). The direct response of freshwater zooplankton to many biological, physical and chemical factors represents a signal correlated to the alteration or shifting in aquatic ecosystems such as climate shift. Climate change has been shown to affect freshwater biodiversity by altering species extinction rates, causing range shifts in species distribution, and improving the invasion success of non-native species (Havel et al., 2015; Manickam et al., 2018). Plankton monitoring programs have been recognized as sentinels of global change and the response to environmental parameters varies according to the species (Hays et al., 2005; Vadadi-Fülop et al., 2012; Vadadi-Fülop and Hufnagel, 2014). Temperature is one of the most important factors accounting for variation in the zooplankton biological cycle, metabolism, phenology, and population dynamics (Gillooly et al., 2001, 2002; Savage et al., 2004). Zooplankton organisms are poikilothermic and their physiological processes (e.g. ingestion, respiration, and reproductive development) are sensitive to temperature, with physiological rates doubling or tripling with an increment of 10 ◦C (Mauchline, 1998). Climatic variables such as mean temperature and the drought-precipitation regime have a direct influence on zooplankton occurrence, reproduction rate, persistence, and abundance in freshwater environments (Chaparro et al., 2011; Ji et al., 2017; Rasconi et al., 2015). Due to their short life cycle, zooplanktonic species are good bioindicators of climate change (Azani et al., 2021; Hays et al., 2005; Richardson and Richardson, 2008). Variations in precipitation, droughts and altered mixing regimes all represent major forces on the abiotic and biotic template and the fundamental niche where the population growth rate is assumed positive (Kearney, 2006; Pulliam, 2000). The submerged soil has an important role in the mineralization and nutrients exchange with the water phase and variables such as water chemistry in ponds and ephemeral habitats are strongly affected by soil physical–chemical properties (Das et al., 2005; Lemaire et al., 2017; Ponnamperuma, 1972; Yang et al., 2017). Such environmental variables have a direct influence on species' physiology and phenology and may affect dispersal and biotic interactions that shape the species' occurrence (Chaparro et al., 2011; Ji et al., 2017; Rasconi et al., 2015).

Here we explore a species distribution model framework combined with machine learning algorithms and global sensitivity and uncertainty analysis (GSUA) to assess the potential environmental drivers that shape the distribution of two freshwater zooplankton species, *Daphnia longispina* (Cladocera) and *Eucyclops serrulatus* (Copepods), in systems of shallow and ephemeral freshwater habitats at a regional level. The modeling framework was used considering a scenario of future climate change to evaluate the potential shift in the spatial distribution of the two zooplankton species. We hypothesized that direct and indirect anthropogenic pressures may affect the predicted potential shift.

Material and Methods

Sampling area

The sampling area was an East-West transect located in the North of Italy along the Northern slope of the Appennine mountains, an area that extends from the Liguria region (Ligurian sea) to the Marche region (Adriatic Sea) with a total area of 47,329 Km2 and altitude from 60 m to 1971 m asl (mean altitude = 855 m asl) (Figure 9). A total of 283 freshwater habitats, including small shallow lakes and ephemeral ponds were sampled in the period 1960/1970 (Moroni and Bellavere, 2004). A total of 60 Cladocera and Copepoda species were identified. To reduce the uncertainty of the 283 pairs of coordinates, the locations were checked and corrected using the Google Earth engine. The study area was delimited using a convex hull that encompassed all the sampled locations points with a buffer distance of 5 Km. The convex hull was drawn using the R package `rangemap` (Cobos et al., 2021). Two zooplankton species, *D. longispina* (Cladocera) and *E. serrulatus* (Copepoda, Cyclopoida), were chosen due to their prevalence. *D. longispina* was found in 97 habitats (presences) and was not found in 181 habitats (absences) with a prevalence of 34%. *E. serrulatus* showed a prevalence that approximately 50% (144 presences and 139 absences).

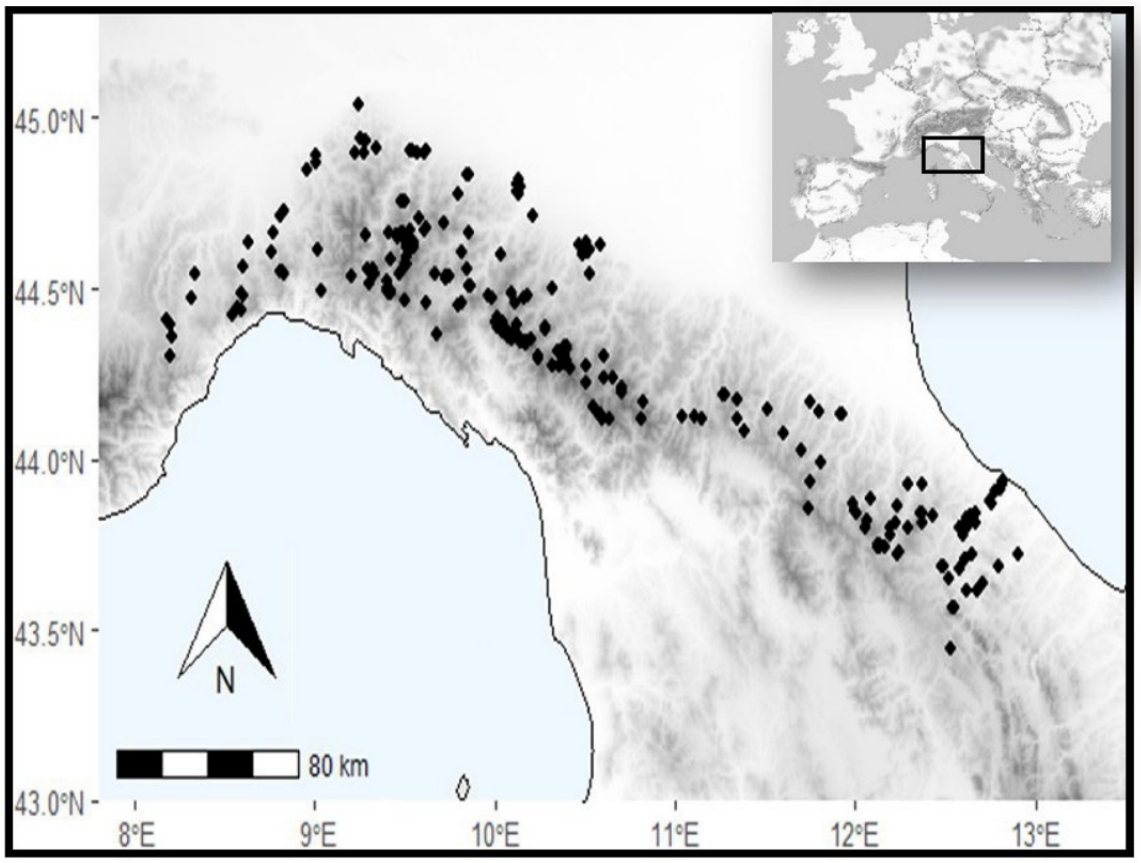

Figure 9 Sampling area and spatial distribution of the selected 283 freshwater habitats in the Northern Apennine (Italy).

Spatial thinning

To reduce sampling bias, spatial thinning on the occurrences data was applied (Steen et al., 2021). The presences and absences of *D. longispina* were thinned with a minimum distance of 2 Km and 3.2 Km, respectively. The presences and absences of *E. serrulatus* were thinned with a minimum distance of 5 Km and 8 Km, respectively. Differences in the minimum distance of presences and absences (thinning parameter) were set to maintain the original prevalence of each species in the thinned dataset (34% for *D. longispina* and 50% for *E. serrulatus*). In both species, the values of the presence minimum distances were lower than the values of the absence minimum distances because absences were more frequent (in *D. longispina*) or more dispersed (in *E. serrulatus*) than presences. For each species, the thinning algorithm was repeated 50 times, and the repetition with the highest standard deviation of the longitude was selected. The thinning procedure was performed with the R package `spThin` (Aiello-Lammens et al., 2015).

Conceptualization and environmental variable selection

Environmental variables used in this study represented four types of environmental constraints on zooplankton species distributions: climatic, hydrological, and soil properties (physico-chemical). All environmental variables were projected into the WGS84 coordinate reference system with a resolution of 30 s. Twelve bioclimatic variables considering precipitations/temperature extremes and seasonality referred to a past period (1970–2000) (Fick and Hijmans, 2017) (Table 1). A set of further six expanded climatic variables referred to a past period (1960–1990), that have a direct influence on water availability in the soil and that might affect the occurrence of zooplankton species in ephemeral freshwater habitats was 3 considered (Title and Bemmels, 2018). The future climatic condition was considered for the period 2040 to 2060. Seventeen global circulation models (GCMs) and one socioeconomic pathway regarding the worst emission scenario (SPP 8.5) were used (CMIP 6; Eyring et al., 2016). The 12 bioclimatic variables (Table 1) were obtained from WorldClim v.2.0 (https://www.worldclim.org/) for each of the 17 GCMs. The six expanded climatic variables (Envirem in Table 1) were computed with the R package envirem (Title and Bemmels, 2018), using mean temperatures and precipitation projections for each of the 17 GCMs. To constrain the geographic environmental suitability to past and future freshwater habitats, two hydrological variables, the drainage direction and the flow accumulation were included (Lehner et al., 2008) (Table 1). The drainage direction defines the direction of water flow in the conditioned digital elevation model toward neighboring regions with higher steepness. The flow accumulation is a measure of the upstream catchment area, where low values refer to high topographic points and high values are in proximity to the outlet of primary rivers. Seven topsoil physical-chemicals properties were used to investigate how the soil characteristics might affect the occurrences of both zooplankton species in the past and future. For two top-soil physical-chemicals properties (sand and clay content in Table 1), a data transformation was applied. Sand and clay content expressed as a percentage of soil weight were mapped into a new categorical variable: soil texture. In a

particular habitat, the mapping procedure was made using the highest percentage between sand and clay content to classify the soil texture as sandy (class 0) or clay (class 1).

Table 1 Environmental variables considered in the species distribution model for both zooplankton species.

Type	***Environmental Variable***	***Source***
Climatic	Isothermality (Bio 3)	Fick and Hijmans, 2017
	Temperature seasonality (Bio 4)	
	Mean temperature of the Wettest quarter (Bio 8)	
	Mean temperature of the Driest quarter (Bio 9)	
	Mean temperature of the Warmest quarter (Bio 10)	
	Mean temperature of the Coldest quarter (Bio 11)	
	Annual Precipitation (Bio 12)	
	Precipitation seasonality (Bio 15)	
	Precipitation of the Wettest quarter (Bio 16)	
	Precipitation of the Driest quarter (Bio 17)	
	Precipitation of the Warmest quarter (Bio 18)	
	Precipitation of the Coldest quarter (Bio 19)	
	Annual Potential Evapotranspiration (PET) (Envirem 1)	Title and Bemmels, 2018
	Aridity Index Thornthwaite (Envirem 2)	
	PET of the Wettest quarter (Envirem 11)	
	PET of the Driest quarter (Envirem 12)	
	PET of the Warmest quarter (Envirem 14)	
	PET of the Coldest quarter (Envirem 15)	
Hydrological	Drainage direction (Hydroshed)	Lehner et al., 2008
	Flow accumulation (Hydroshed)	
Soil Properties	Carbon (total)	Wieder et al., 2014 Poggio et al., 2021
	Bulk Density	
	pH	
	Sand content	
	Clay Content	
	Available water storage capacity	
	Nitrogen (total)	

To evaluate multicollinearity the set of environmental variables was processed with the variance inflation factor analysis (VIF) (James et al., 2014). A threshold of VIF = 10 was used and only environmental variables with VIF values <10 were retained in the modeling framework.

Modeling framework

For each species, we model the habitat suitability by considering one regression-based model, the generalized linear model (GLM), and two machine learning algorithms: random forest (RF) and artificial neural network (ANN) with one hidden layer. The GLM was used as a benchmark with additive terms and a binomial family function to improve the comparison and evaluation of the two machine learning algorithms. The model that showed the best performance (see below) was selected and fine-tuned to describe the species' presences/absences in the study area in the past (1960–1970) and to predict the species distribution in the future (2040–2060). The GLM was computed using the `glm` function of the R package `stat` (R Core Team, 2017). The RF was fitted using the R package `randomForest` (Liaw and Wiener, 2002) and the ANN was fitted with the R packages `tensorflow` (Allaire and Tang, 2020) and `keras` (Allaire and Chollet, 2021).

Model selection: Block cross validation

Block cross-validation was used to evaluate the predictive models (Roberts et al., 2017). We split the data (presence/absence) into squared spatial blocks. The dimension of each spatial block was computed considering the mean spatial autocorrelation of each environmental variable to obtain 5 folds of blocks (Figure 1SM). For each continuous environmental variable, autocorrelation was obtained by an empirical variogram that was estimated using 5000 random points. The block cross-validation was performed with the R package `BlockCV` (Valavi et al., 2019). All continuous environmental variables were standardized. Each model was run on each fold and evaluated by four performance metrics: the percent classified correctly (Pcc), Kappa, true skill statistic (TSS), and area under the receiver operating characteristic curve (Auc) (Konowalik and Nosol, 2021; Liu et al., 2011). Three performance metrics were threshold-dependent: Pcc, kappa and Tss. We used a standard threshold (0.5). Auc was threshold-independent. The model that maximizes the mean of most metrics on five-folds was selected and fine-tuned.

Fine tuning and threshold optimization

The selected model was trained and fine-tuned. For each species, the presence/absence data were split into two-fold: the training set (80%) and the validation set (20%). In the training phase, using the training set, a grid search procedure was used: the model's hyperparameters were varied in combinations and for each combination, the four-performance metrics (Pcc, Kappa, Tss and Auc) were computed. In the validation phase, using the validation set, the hyperparameter combination that maximizes the most performance metrics was selected. The control of differences between training and validation metrics allowed us to avoid overfitting. To describe or predict in the best way the species' presences/absences the model threshold was optimized according to the mean threshold computed by 12 different criteria of threshold selection (Freeman

and Moisen, 2008). The threshold optimization was carried out with the R package `PresenceAbsence` (Freeman and Moisen, 2008).

Variable importance and response curves

To rank the relative contribution of each environmental variable to the probability of occurrence of the species, the SHAP analysis was used (SHapley Additive exPlanations) (Lundberg et al., 2017; Lundberg et al., 2018; Lundberg et al., 2020). SHAP values were computed using a game theoretic method that improves the understanding of machine learning algorithms. Each environmental variable was ranked according to the SHAP mean absolute values. The SHAP analysis was performed in Python programming language (Van Rossum and Drake, 1995) with the library `shap` (Lundberg et al., 2017). For each species, the partial response curves of the model were computed for the three most important environmental variables along the environmental gradient of the study area.

Global sensitivity and uncertainty analysis (GSUA)

The global sensitivity and uncertainty analysis (GSUA) was computed to identify key determinants of model outputs and the model's interaction terms among environmental variables (Convertino et al., 2014; Pianosi et al., 2016). For each continuous environmental variable, a standardized normal distribution was considered; for soil texture, a categorical variable, a Bernoulli distribution (considering $p = 0.75$ the probability of clay soil and $q = 0.25$ the probability of sandy soil) was used. For each environmental variable, 1000 observations were sampled from the probability distributions using a sample design based on quasi-random numbers. To predict the ANN model output the sampled matrix was used. To quantify the first-order and total-order indices of each environmental variable, the Sobol method was computed (Saltelli et al., 2008). To quantify the second and third-order interaction indices, the environmental variable with the highest total interaction value (difference between total order and first-order indices) was considered. GSUA was computed using the R package `sensobol` (Puy et al., 2021).

Model prediction: Past and future climatic condition

For each species, the final model was used to draw a study area map with presences and absences considering the past and future climatic conditions. The shift in the spatial occurrence of both zooplankton species was quantified by computing the differences between future and past maps. To evaluate how the anthropogenic pressure may affect the potential predicted shift, the difference map was superimposed on the land use data from the Copernicus Global Land Service of 2019 (Buchhorn et al., 2020). The 23 land use classes in the Copernicus Global Land Service were mapped into 3 main land use classes: Urban,

Agricultural, and Natural. The spatial shift was quantified for each land use class. Differences in the spatial shift between species were tested with the Chi-squared test.

Spatial shift: Probability distribution functions (pdfs) over space and time

To understand site-specific and ecosystem shifts, spatial expansion and contraction were analyzed as a function of the environmental variables considering joint probability distribution functions (pdfs). For each species, 2000 points relative to spatial contraction and spatial expansion (dependent variable) were randomly sampled with a balanced design. A logistic regression model was fitted using the differences between future and past environmental conditions of the two most important variables as explanatory variables. To highlight attractor regions, a decision boundary fitted by the logistic regression model in the environmental variable space was predicted (Sharp et al., 2013).

Result

The spatial thinning algorithm produced a map with 57 presences and 113 absences for *D. longispina* and with 63 presences and 63 absences for *E. serrulatus*, retaining for both species the original prevalence.
The VIF analysis identified 8 climatic variables with multicollinearity problems (5 bioclimatic and 3 envirem) that were removed from the modeling framework (Table 1SM). The size of the squared blocks that defined the five spatial folds were 38,4 Km (Figure 1SM).
For both species, the 5-fold spatial cross validation showed that the artificial neural networks (ANN) produced the highest mean values for all the performance metrics (Figure 10 and Table 2SM).

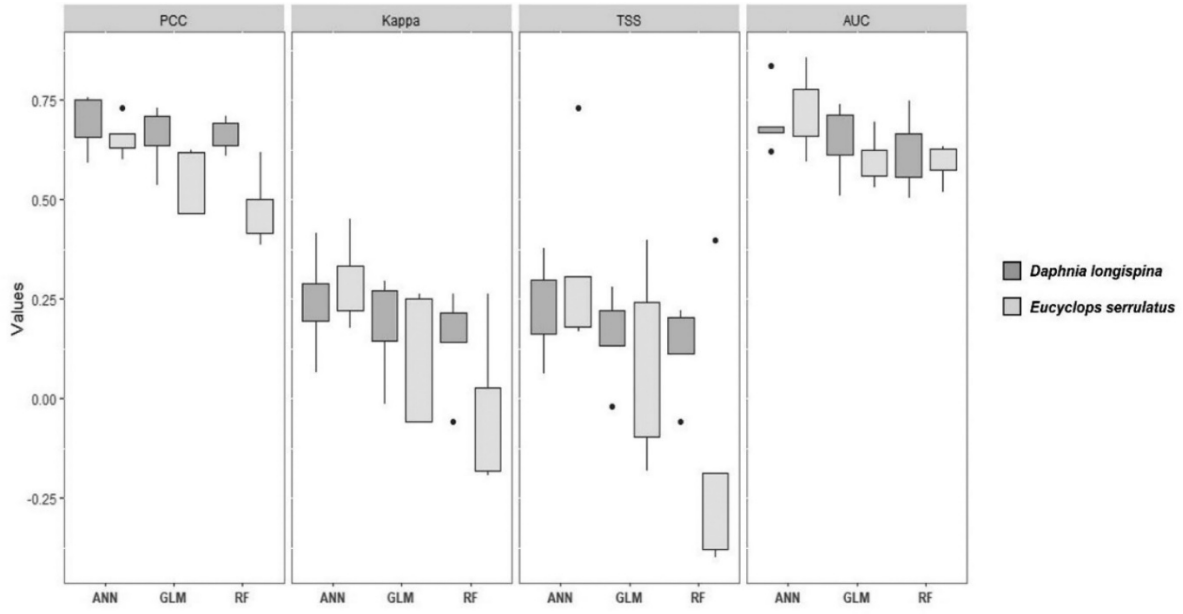

Figure 10 For each species (*D. longispina* and *E. serrulatus*) and each model (GLM, RF and ANN) the values of the four-performance metrics Pcc, Kappa, Tss and Auc after the block cross validation were reported.

For both species, the artificial neural networks were fine-tuned on the training set (80%) and evaluated on the validation set (20%) considering three hyperparameters in combination (number of combinations = 120): number of neurons in the hidden layer (16, 12, 8, 4), the learning rate (0.0005, 0.001) and the batch size (32, 16, 8, 4). For *D. longispina* and *E. serrulatus* the best combination of hyperparameters was: 12 neurons in the hidden layer, a batch size of 32 and a learning rate of 0.001 and 0.0005, respectively (Table 2).

Table 2 For each species, the values of the performance metrics after fine tuning of the ANN algorithms were reported.

Species	Pcc	Kappa	Tss	Auc
Daphnia longispina	0.72	0.42	0.45	0.68
Eucyclops serrulatus	0.71	0.41	0.42	0.64

For *D. longispina* and *E. serrulatus* the values of the optimized threshold were 0.38 and 0.50. The three most important variables ranked by the SHAP analysis were different in different species (Figure 11). For *D. longispina* the most important variables were temperature seasonality (Bio 4), precipitation of the warmest quarter (Bio 18) and nitrogen. For *E. serrulatus* the most important variables were precipitation of the warmest quarter (Bio 18), temperature seasonality (Bio 4) and mean temperature of the wettest quarter (Bio 8).

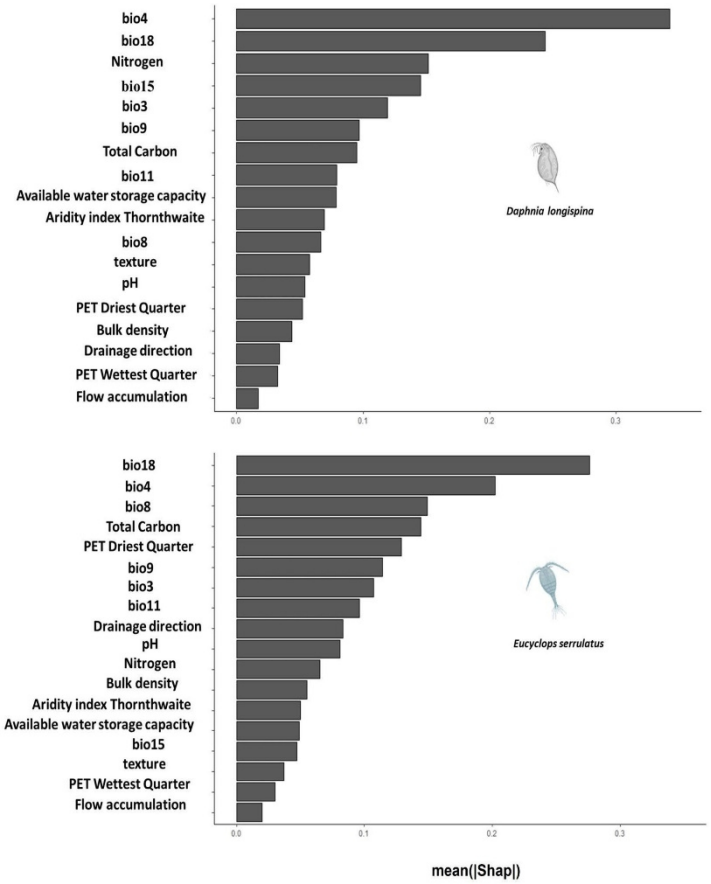

Figure 11 SHAP means absolute values of each environmental variable for *D. longispina* (top panel) and *E. serrulatus* (bottom panel) indicated the variable importance.

In both species, the probability of occurrence increased with values of temperature seasonality (Bio 4) (Figure 12). In *D. longispina*, the probability of occurrence increased monotonically with the total nitrogen and reached an optimum when the precipitation of the warmest quarter (Bio 18) was between 200 and 300 mm. For *E. serrulatus*, the probability of occurrence increased with precipitation of the warmest quarter (Bio 18) and decreased with the mean temperature of the wettest quarter (Bio 8). According to GSUA, for both species, Bio 18 and Bio 4 showed the highest first and total effect indices (Figure 13). The aridity index of Thornthwaite and soil texture were the least influential variables in the sensitivity of ANN. Bio 4 showed the highest interaction term (Figure 13).

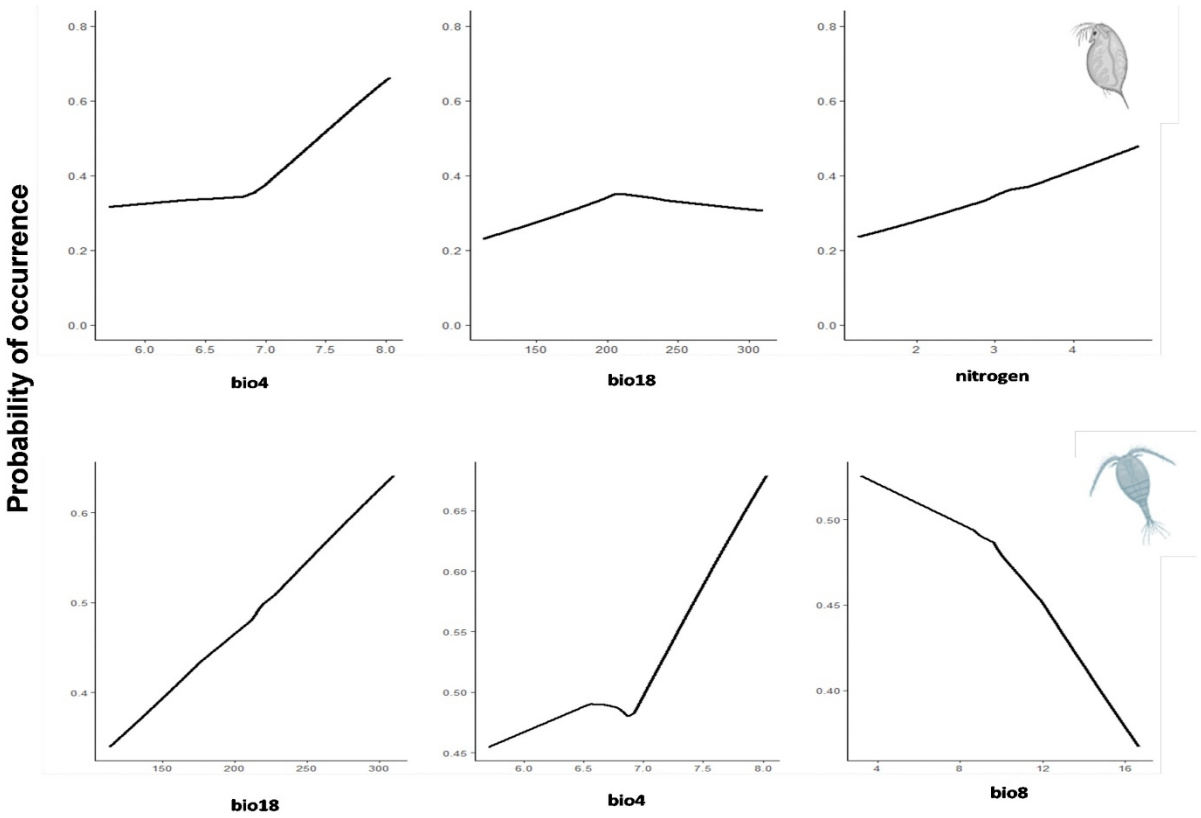

Figure 12 Partial response curves of three most important environmental variables: temperature seasonality (Bio 4), precipitation of the warmest quarter (Bio 18) and nitrogen for *D. longispina* (top panels) and precipitation of the warmest quarter (Bio 18), temperature seasonality (Bio 4) and mean temperature of the wettest quarter (Bio 8) for *E. serrulatus* (bottom panels).

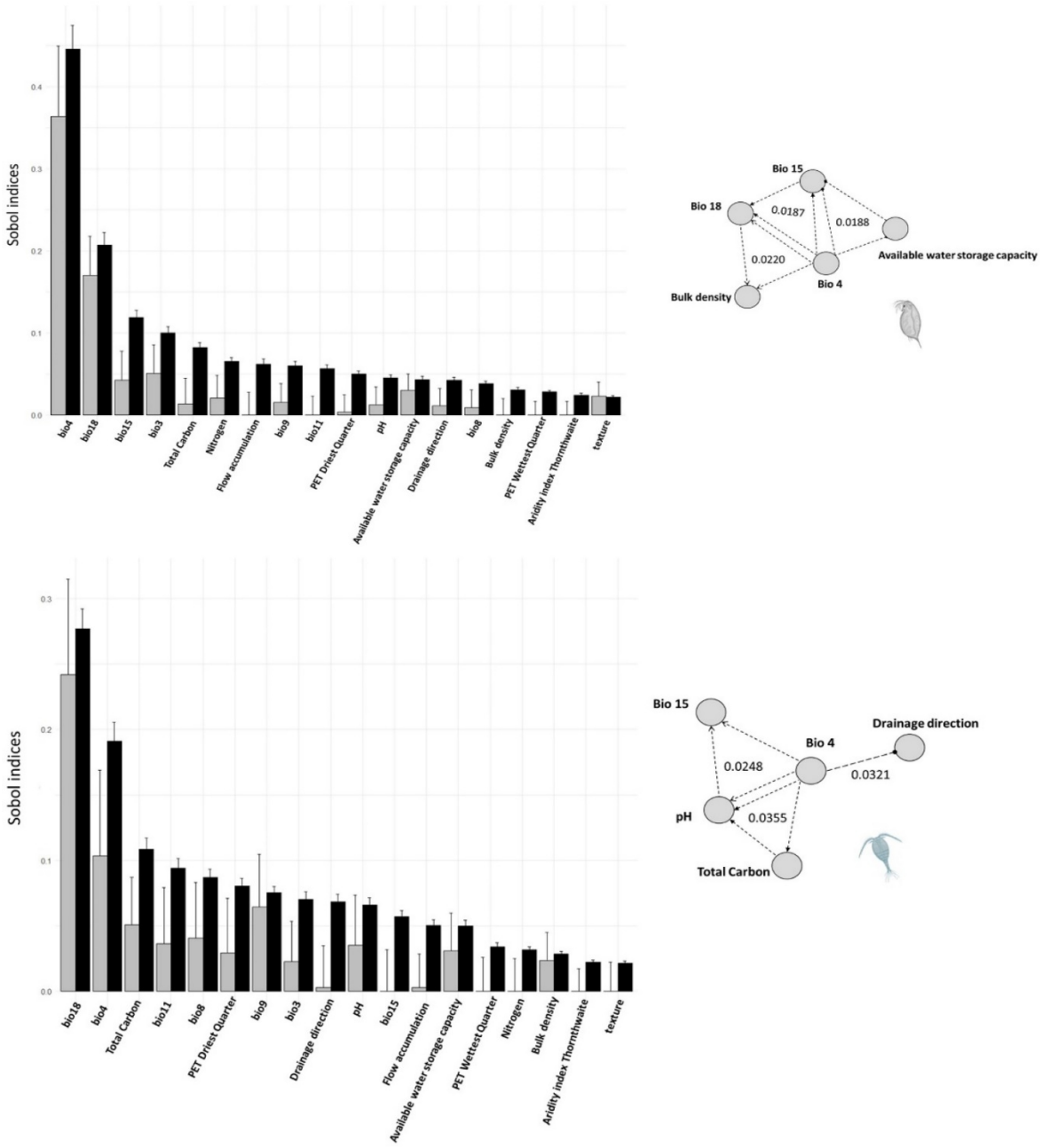

Figure 13 Sobol' sensitivity indices (first order indices in grey and total order indices in black) for each environmental variable for *D. longispina* (top panel) and *E. serrulatus* (bottom panel). On the right side, the three most important second and third order interaction terms of Bio 4 were reported.

For *D. longispina*, the three most important third order interaction terms with Bio 4 were: Bio 18 and bulk density, Bio 15 and Bio 18, Bio 15 and available water storage capacity. For *E. serrulatus*, the two most important third order interaction terms with Bio 4 were: pH and total carbon, pH and Bio 15. A second order interaction term was detected between Bio 4 and drainage direction. In the past, the spatial distribution of *D. longispina* was predominantly in the North-West part of the study area, while *E. serrulatus* showed a scattered distribution along an East - West gradient (Figure 2SM). Both species, in future, are expected to

shift their distribution toward lower altitude habitats with an overall expansion of 7% with respect to the past climatic condition (Figure 14).

Figure 14 Spatial shift from past to future climatic conditions for *D. longispina* (left panel) and for *E. serrulatus* (right panel).

D. longispina is expected to expand the distribution range from 12,602 Km^2 to 16,064 Km^2, while *E. serrulatus* is expected to expand the occupancy range from 15,987 Km^2 to 18,887 Km^2. However, the spatial expansion of *D. longispina* and *E. serrulatus* was qualitatively different. In agricultural and natural areas, the expansion of *E. serrulatus* was greater than that of *D. longispina* (agricultural area: $\chi^2 = 370.86$ and $p < 0.0001$, natural area: $\chi^2 = 159.49$ and $p < 0.0001$) but, in natural areas the expansion of *E. serrulatus* was counterbalanced by a greater spatial contraction with respect to that of *D. longispina* ($\chi^2 = 1305.3$ and $p < 0.0001$) (Figure 15). Both species increased their spatial expansion in urban areas with no statistical difference ($X^2 = 3.26$ and $p = 0.07$). Considering the environmental space defined by Bio 4 and Bio 18 change in time, the decision boundary split into two distinct and different regions of contraction and expansion of both species (Figure 16). For *D. longispina*, the probability density function (pdf) of spatial contraction showed a bimodal distribution in a region of the environmental space characterized by a contemporary reduction of Bio 4 (temperature seasonality) and Bio 18 (precipitation of the warmest quarter). The spatial expansion distribution was in a region of the environmental space characterized by increase of the temperature seasonality. For *E. serrulatus*, the probability density function (pdf) of spatial contraction was observed in a region of the environmental space characterized by an increase of Bio 4 and a reduction of Bio 18.

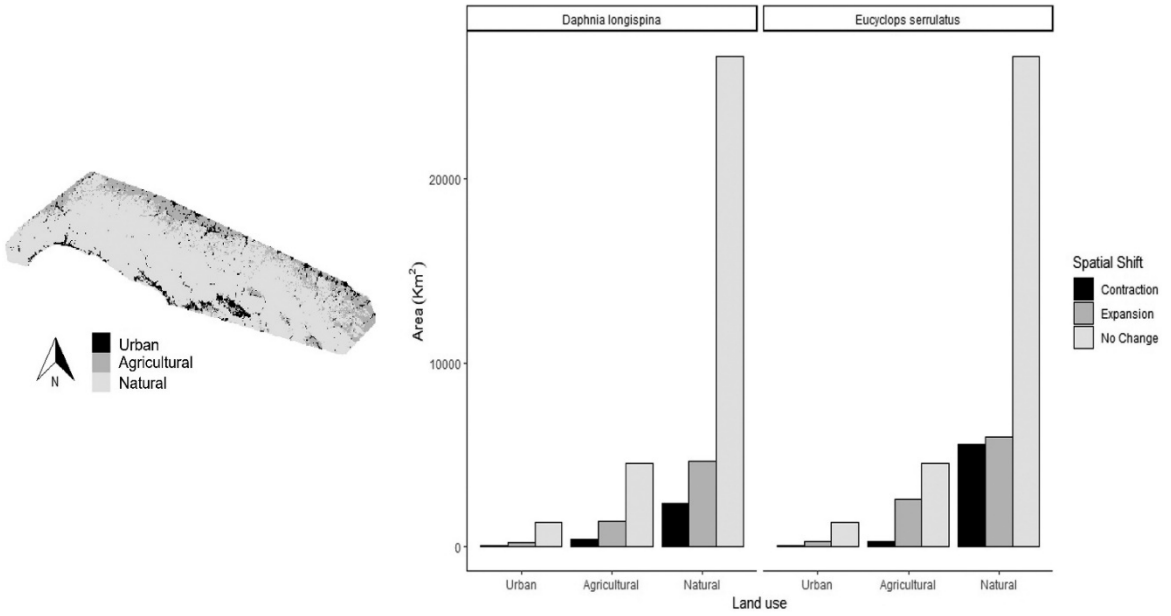

Figure 15 Land use of the study area (left panel) and the future range shift (right panel) considering three land use classes: urban, agricultural and natural.

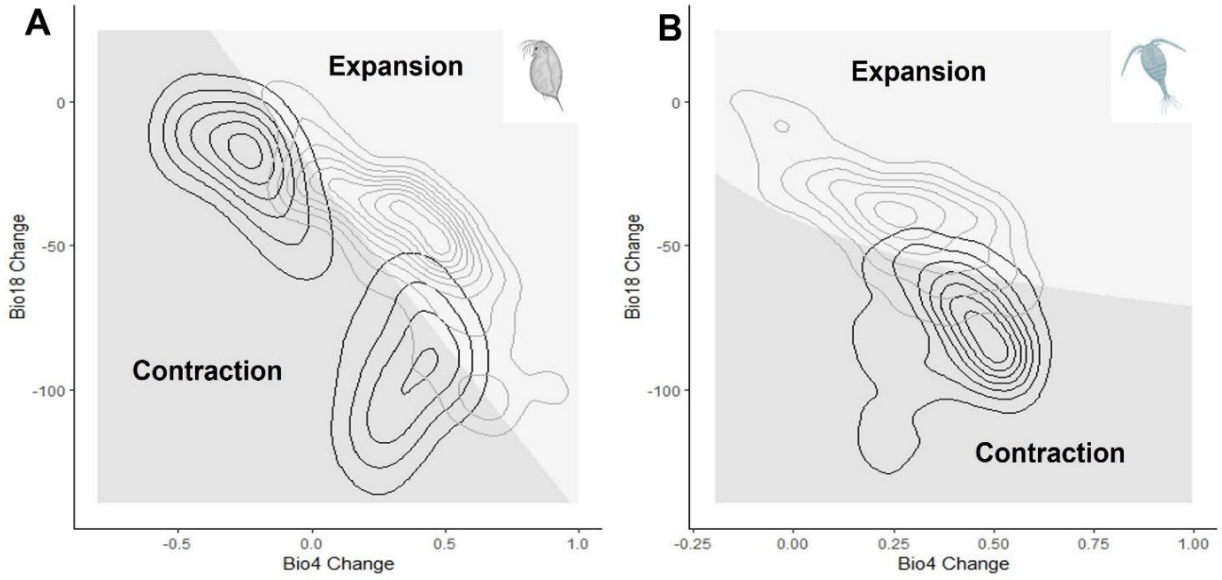

Figure 16 The decision boundary identified the regions of the environmental variables Bio 4 and Bio 18 corresponding to spatial contraction (grey) and spatial expansion (light grey) for *D. longispina* (A) and *E. serrulatus* (B). For each species, probability density functions (pdfs) in contraction and expansion areas were reported.

Discussion

In this study, we modeled the past and future spatial distribution of two freshwater zooplankton species on the Northern slopes of the Apennine mountains (Italy). A cladocera, *D. longispina*, and a copepod *E. serrulatus* were considered to highlight zooplankton responses to climate change at a regional level. Both *D. longispina* and *E. serrulatus* are expected to expand their range distribution to the North and lower altitudes. The percentage of expansion is 7% but in agricultural and natural areas, the expansion of *E. serrulatus* was

greater than that of *D. longispina*. In natural areas, the expansion of *E. serrulatus* was counterbalanced by a greater spatial contraction than that of *D. longispina*. By the analysis of variable importance, response curves, and GSUA, we showed that both species respond to similar environmental conditions. Not surprising, temperature seasonality (Bio 4) and precipitation of the warmest quarter (Bio 18) were the most important climatic variables that drive spatial distribution and shift of occupancy with a putative range expansion in climate change. *D. longispina* and *E. serrulatus* showed a high probability of occurring in habitats characterized by high levels of summer precipitations. The availability of summer rainfall increases the occurrence of ephemeral freshwater habitats in favorable warm season when the growth and reproduction of poikilothermic organisms, sensitive to environmental temperature, may accelerate (Maier, 1990; Gerten and Adrian, 2002; Gillooly et al., 2001, 2002; Savage et al., 2004). Temperature seasonality was another important factor in determining the spatial distribution of the zooplankton species due to the generally small size and the elevation of most of the habitats considered in this study. *D. longispina* copes with the temperature and temporary habitat seasonality with dormant resting eggs that remain available in the sediments until favorable environmental conditions for growth restore (Brendonck and De Meester, 2003; Caceres, 1997). The higher variation in temperature among seasons is a strong environmental cue that increases the population's long-run growth rate and synchronizes directly the life cycle of zooplankton and indirectly the whole aquatic food web by predation, competition, and nutrients cycle (Drake, 2005; Toyota et al., 2019). According to our results, the increasing level of total nitrogen increased the probability of the occurrence of *D. longispina*. This pattern highlights the fundamental role of the nitrogen cycle across the food webs. Nitrogen, as well as phosphorous, composes of dead organic matter deposited in the soil or sediments and it is remineralized or dissolved before it can be absorbed by primary producers (Guignard et al., 2017; Sanders et al., 2015). Especially in shallow habitats, due to their overall small size and the small ratio between water volume and sediment surface, phytoplankton communities have the potential to control inorganic nutrients, regulate dissolved oxygen, inorganic carbon concentrations, and water pH and, represent the main resource for primary consumers such as *Daphnia* (Bennion and Smith, 2000; Lisheid et al., 2018; Marlene et al., 2020). Although some chemical forms of nitrogen, such as ammonia, in higher concentrations could lead to the inactivation of the filtration in Cladocera, the stoichiometric ratio of carbon:nitrogen:phosphorus in phytoplankton might follow the pattern of hydrography (Serra et al., 2019). The energy flow and the transfer of nutrients through trophic levels affect the whole aquatic ecosystem and the life history traits of zooplankton (growth, reproduction, and survival rate) vary according to the quality and quantity of the filtered phytoplankton. Xu et al. (2021) examined the transcriptome response and phenotypic shift to a nitrogen or phosphorus-limited diet in D. magna. They highlighted that, under nitrogen limitation, the element assumed from the diet should be assigned to body growth rather than reproduction in agreement with reported data on marine copepods (Kuijper et al., 2004). Hydrological connectivity has an important role in water quantity and quality and may affect the causality of variables for freshwater zooplankton distribution (Zhang et al., 2021). It increases the ability of organisms to colonize new habitats or exploit new resources, but also results in the removal of organisms or introduces competitors or predators

(Napiorkowski et al., 2019; Reid et al., 2016). However, hydrological connectivity may have controversial effects according to the taxa, the environment, the landscape, and the scale: the future trends, particularly assessing the influences of climate changes, should be discussed (Zhang et al., 2021). In our study, flow accumulation and drainage direction, which are hydrological variables correlated to hydrologic connectivity, showed a very low relative contribution to the probability of occurrence of *D. longispina* and *E. serrulatus*. Flow accumulation was ranked 18/18 for both species while drainage direction was ranked 16/18 for *D. longispina*. For *E. serrulatus*, drainage direction was ranked 9/18 and a second-order interaction term was detected with Bio 4 using GSUA. For both species, different regions of contraction and expansion in climate change conditions were estimated. For *D. longispina*, with the same precipitation of the warmest quarter (Bio 18), the pdf of spatial contraction was observed with a reduction of temperature seasonality (Bio 4). For *E. serrulatus*, the pdf of spatial contraction was observed in a region of the environmental space characterized by an increase in temperature seasonality (Bio 4) and a reduction of precipitation in the warmest quarter (Bio 18). ANNs are deep learning algorithms made by artificial neurons that combine an input layer (explanatory variables) with an output layer (Recknagel, 2001). The signal is processed and transformed by neurons; the computing elements are interconnected with synapsis. During the training phase in the supervised learning framework, the algorithm iteratively modifies the strength of the synapsis to minimize an output error. ANNs are capable of learning more complex geometries and non-linearity than other classes of algorithms (Olden and Jackson, 2002). The performance of the ANN algorithms reached sufficient accuracy considering all performance metrics. TSS, Kappa, and Auc showed values that were greater than a random classifier (Allouche et al., 2006; Elith et al., 2006). Although the performance values we found were not among the highest in the literature, a rigorous process of fine-tuning was made to control the model's overfitting and to produce reliable results. Further studies that consider biological interactions, community composition, functional traits, and the introduction of all possible classes of environmental variables that might act at regional scales, could improve the performances of this class of algorithms. Future climatic changes are expected to increase the mean temperature and a reduction of mean precipitations across the study area. This might shift the zooplankton species distribution toward habitats located on the northern slope of the Apennine mountains at lower altitudes (Po plain). Here, the suitable habitats are in proximity to agricultural and urban landscapes characterized by different anthropic drivers that could negatively affect and reduce the realized niche of both *D. longispina* and *E. serrulatus*. Both species increased their spatial expansion in urban areas. Anthropogenic pressures such as high pollution levels, in synergy with temperature increases, might reduce the predicted habitat suitability and expansion of zooplankton species (Gianuca et al., 2017; Leitao et al., 2013). The functioning of small water bodies may vary in space and time due to disturbance or the absence of a stable steady state (Bellin et al., 2021). Such instability is favored by the vulnerability of ponds to a large set of pressures and has important implications for ecosystem restoration, especially in heavily impacted agricultural areas (Bennion and Smith, 2000; Lischeid et al., 2018). A high potential for endemisms and biodiversity in urban areas has been revealed (De Bie et al., 2008; Ejsmont-Karabin and Kuczynska- Kippen, 2001; Langley et al., 1995; Maier et al., 1998; Mimouni et al., 2015). However,

according to Shen et al. (2021) the diversity of zooplankton decreased with increasing urbanization levels due to the quality of wetland deterioration.

1.2 Modelling climate change impacts on the habitat suitability of Mediterranean gorgonians

Species distributions are determined by different environmental factors, resources, and conditions, and field observations can be related to environmental predictor variables (Guisan and Thuiller 2005). Ecosystem engineer species, directly or indirectly, modulate the availability of resources, for themselves and to other species by causing physical state changes in biotic or abiotic factors (Jones et al., 1994). Autogenic engineers, such as trees or corals modify, maintain, and/or create habitats via their own structures. Gorgonian soft corals are characterised by high level of biological diversity and play an important role as habitat providers (Garrabou and Harmelin 2002; Coma et al., 2004; Linares et al., 2007, Linares et al., 2008). They supply interlinked ecological services such as nutrient cycling, carbon sequestration, sediment stabilization, current deviation, and services to human populations through support of fisheries, tourism, and broader benefits such as the protection of coastal areas from waves and storm impacts and delay the spread of invasive algae (Costanza et al., 1998; Ballesteros, 2003; Piazzi and Balata 2009; Cerrano et al., 2010; Casas-Güell et al., 2015; de Ville d'Avray et al., 2019). Modifications of the edaphic conditions caused by gorgonians forests influences larval settlement and recruitment processes of the benthic assemblages, while supporting diverse food webs, nursery grounds for numerous associated species and biodiversity of marine communities (Reaka-Kudla, 1997; Thomsen et al., 2010; Ponti et al., 2014; Liconti et al., 2022). The conservation of gorgonian forests is crucial to avoid depletion or degradation of coralligenous ecosystem functioning, especially in the early-stage recruitment (Cerrano et al., 2000; Cerrano and Bavestrello, 2008; Coma et al., 2004; Linares et al., 2005; Ponti et al., 2014). The local disappearance of gorgonians may cause a shift of the epibenthic assemblages from crustose coralline algae to filamentous algae dominated and reduce the resilience of coralligenous bioconstructions (Ponti et. al., 2014). Spatial-temporal distribution of gorgonians is determined by the combined effects of biological and environmental factors that can affect the recruitment, growth, and death rates of individuals in singles species populations (Gori et al., 2011). In sessile marine organisms such as gorgonians, the interaction between the recruitment and survival of individuals results in patchy distribution patterns and spatially structured populations (Sebens, 1991; Karlson, 2006; Gori et al., 2011). Such patterns will have a fundamental influence on ecological processes in the short term and on the spatial structure in the long term (Illian et al., 2008).

The distribution and abundance of gorgonians are increasingly threatened by exposure to multiple stressors including global warming, carbonate chemistry of seawater, the spread of alien species and local anthropogenic activities (Cerrano et al., 2000; Milazzo et al., 2002; Coma et al., 2009; Huete-Stauffer et al., 2011; Verdura et al., 2019; Cebrian et al., 2018; Galil, 2019). Mediterranean species generally have cold affinity and are particularly sensitive to increasing temperatures. In the North-Western Mediterranean Sea, mass mortality events (MME) among gorgonian forests have been increasing in frequency and intensity since the end of the last century (Martin et al., 2002; Rivetti et al., 2014; Chimienti, 2021; Iborra et al., 2022). The Mediterranean Sea is a semi-enclosed basin where global warming is causing substantial impacts and it is considered a climate change hot spot (Tuel and Eltahir, 2020). Thermal stress, water stratification

and associate diseases are the most likely causes of gorgonians MMEs coinciding with high water temperature below more than 40 meters in depth (Coma et al., 2009; Vezzulli et al., 2013).

The red gorgonian *Paramuricea clavata* (Risso, 1826) is a long-lived, slow-growing species characterized by colonies that can exceed 1.5 m in height and live for over a century. It exhibits a bathymetric range that goes from 5 to 200 m (Mokhtar-Jamai et al., 2011). *Eunicella cavolinii* (Koch, 1887) is of Mediterranean coastal waters very common in the western Mediterranean Sea and in the Adriatic Sea (Sini et al., 2015). Its distribution range is wide, but patchy in terms of abundance and it has a high depth range distribution (5-150 m) (Russo, 1985; Sini et al., 2015). *E. cavolinii* lives mainly on rocky hard substrate in the coralligenous and pre-coralligenous habitat, where the light irradiance is not too low, often associated with colonies of *P. clavata*. The white gorgonian *E. singularis* (Esper, 1791) is one of the most representative habitat-forming species of the rocky bottoms and Mediterranean coralligenous assemblages (Pey et al., 2013). It was one of the most impacted during past mortality events and, since it has a very low recovery capacity, the future of this sessile invertebrate is in danger: it is mentioned among the vulnerable species of the IUCN Red List. *E. singularis* is the sole symbiont gorgonian with autotrophic dinoflagellates in the Mediterranean. From the very first records in 1999 up to the most recently published and ongoing research activities conducted on marine environments, increasing evidence of thermal-related stress events have been recorded (Cerrano et al., 2005; Cerrano and Bavestrello, 2008; Gambi et al., 2018).

E. singularis and *P. clavata* are most abundant in the Western Mediterranean Sea while *E. cavolinii* has been found to be very common only in the eastern part of the Western Mediterranean Sea: it should be absent or very rare along the coasts located west of Marseille (Gori et al., 2011). According to Fava et al. (2010) the species of the genus *Eunicella* are more resistant and resilient than *P. clavata* and assemblages dominated by *E. cavolinii* tolerate values of irradiance higher than those tolerated by assemblages with *P. clavata.* The thermotolerance of *P. clavata* and *E. cavolinii* might be mediated by the bacterial communities of the two species (Tignat-Perrier et al., 2022).

Here we explore a species distribution model framework combined with machine learning algorithms and global sensitivity and uncertainty analysis (GSUA), to assess the potential environmental drivers that shape the distribution and habitat suitability at regional level of three gorgonian species *P. clavata, E. cavolinii* and *E. singularis* in the Mediterranean Sea. The modelling framework was used to predict their future habitat suitability under the worst IPCC scenario RCP8.5 (Bellin et al., 2022). Understanding of the distribution patterns of species in space and time is crucial for the identification of sites of special interest both inside and outside of existing marine protect areas and in the establishment of management and conservation actions and strategies (Fortin and Dale, 2005). We hypothesized that global warming may negatively affect the potential predicted habitat suitability of these gorgonians' species in the Mediterranean Sea. We compared results of different genera and, within *Eunicella* genera, between symbiotic, *E. singularis,* and the non-symbiotic *E. cavolinii*. According to the literature, we hypothesized that *P. clavata* would be at the highest risk than species of the genus *Eunicella* and that the sensitivity of the symbiotic algae (zooxanthellae) may

affect the susceptibility of the holobiont *E. singularis* to thermal stress (Fitt et al., 2001; Baker et al., 2004; Fava et al., 2010).

Materials and Methods

Species presence data and environmental variables

For all species, presence points within the Mediterranean Sea (Figure 17) were obtained from the Global Biodiversity Information Facility database (GBIF) (www.gbif.org) using the R package `rgbif` (Chamberlain, et al., 2022) and from Liconti et al. (2021). Although *P. clavata* was also recorded in the Atlantic Ocean and in the Aegean Sea (Boavida et al., 2016) and *E. cavolinii* was sampled from the Tunisian and Algerian costs, Aegean Sea, and Marmara Sea (Sini et al. 2015; Masmoudi et al. 2016), our collection of occurrence points was limited to North-Western Mediterranean Sea, Ligurian Sea, Tyrrenian Sea, Ionian Sea and Adriatic Sea.

Presence data was collected as spatial points whose longitudinal and latitudinal coordinates referred to the Coordinate Reference System (CRS) WGS84. The dataset downloaded from GBIF included information on the coordinate uncertainty in meters related to each occurrence. To reduce the geo-localization error, each occurrence with an uncertainty higher than 250 meters was discarded. The `duplicate` function in the R package (R core Team, 2021) was used to delete the duplicated data to avoid data redundancy. Geo-localization mismatches were also checked and excluded when found. The final cleaned dataset contained a total of 2474 data points, 843 for *P. clavata*, 938 for *E. cavolinii* and 693 for *E. singularis*.

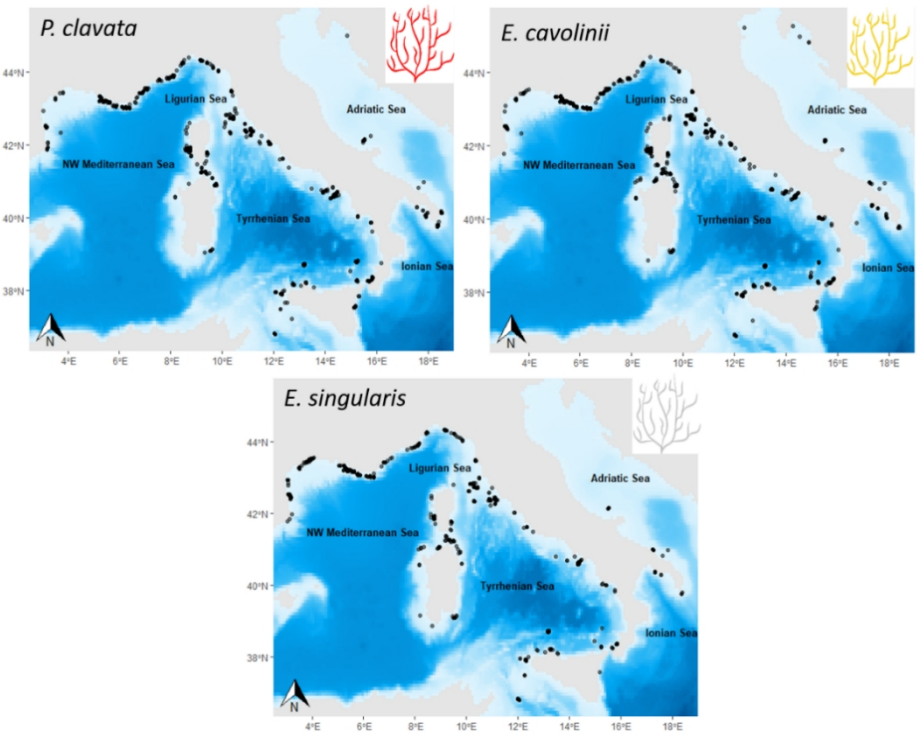

Figure 17 Occurrences of *P. clavata*, *E. cavolinii* and *E. singularis* reported as black points within study area in the Mediterranean Sea.

To model present-day (2000-2014) habitat suitability for the selected species, 21 physico-chemical environmental variables were retrieved from Bio-Oracle (Assis et al., 2018), and 4 geophysical variables were retrieved from MARSPEC (Sbrocco and Barber, 2013) (Table 3) with the R package `sdmpredictors` (Bosch and Fernandez, 2022). All environmental variables were at 30 arc-second (~ 1 km^2) of resolution.

For the predicted future climatic conditions (2040-2050), the RCP8.5 emission scenario (Schwalm et al., 2020) was selected alongside three available environmental variables: current velocity, temperature and salinity for surface and benthic layers were obtained from Bio-Oracle (Assis et al., 2017). The future pH of the surface layer was estimated using the annual trend of pH reduction (−0.0044 units per year) calculated by high frequency observational data in the Mediterranean Sea (Flecha et al., 2015) considering a period from 2014 to 2045. The other environmental variables were kept constant at present values. All environmental variables were at 30 seconds of resolution in the CRS WGS84.

Table 3 List of the selected physico-chemical and geophysical environmental variables.

Type	**Variables**	**Source**
Geophysical	Bathymetry, Plan Curvature, Concavity, E-W aspect	MARSPEC (Sbrocco and Barber, 2013)
Chemico-physical surface layers	Temperature, Current velocity, Photosynthetic available radiation (PAR), Diffuse attenuation, Phytoplankton, pH, Phosphate, Nitrate, Silicate, Dissolved oxygen, Calcite, Salinity	Bio-Oracle (Assis et al., 2018)
Chemico-physical benthic layers (average depth)	Temperature, Current velocity, Light at bottom, Phytoplankton, pH, Phosphate, Nitrate, Silicate, Dissolved oxygen, Calcite, Salinity	Bio-Oracle (Assis et al., 2018)

Modelling approach

Spatial thinning and environmental variable filtering

The spatial thinning was applied with a minimum distance of 3 Km among points (Steen et al., 2020). This procedure was carried out with the R package `spThin` (Aiello-Lammens et al., 2015).

A conservative threshold of VIF = 4 was used and only environmental variables with VIF values ≤ 4 were kept within the modelling framework. The VIF analysis was carried out with R package `usdm` (Naimi et al., 2014).

To identify the most informative environmental variables and to reduce the model complexity, we applied the lasso regression (Tibshirani, 1996). Lasso is a regression technique based on penalized maximum likelihood, and it is based on a shrinkage parameter (λ). When λ=0, no shrinkage of the coefficient is

performed, and as λ increases, the model's coefficients shrinkage become higher. The lasso regression was run using a binomial family with 100 different values of λ. The 10-fold cross validation was performed, and the best set of predictors was selected considering the minimum average value of the mean absolute error (MAE). MAE is a measure of error: it is the average over the test sample of the absolute differences between predictions and actual observations where all individual differences have equal weight. The Lasso regression was performed with R package `glmnet` (Friedman et al., 2010).

Pseudo-absences generation

For both species, the occurrences were presence-only data consisting of the locations of species observations but lacking absence points (Renner et al., 2015). VanDerWal et al., (2009) carried out modelling experiments with MAXENT on 12 species and they found that the relationship between the geographic extent used to sampling the pseudoabsences, model performance and environmental factors importance was maximized using 200 Km. Moreover, Barbet-Massin et al., (2012) recommended that when machine learning algorithms are integrated in species distribution models (SDM) and the number of occurrences was <1000, pseudo-absences should be drawn with a balanced design using a geographical exclusion of 2 latitudinal degrees from presences. Pseudo-absences sampled from small areas might yield to spurious results, meanwhile sampling carried out in a broad area might produce model overfitting and more simplified relationship controlled by few environmental variables. In this study, a random sampling with a balanced design and a buffer radius distance of 150 km was used as trade-off. The pseudo-absences sampling was carried out with the R package `ENMTools` (Warren and Dinnage, 2022).

Model selection

To model habitat suitability of the three species, three Machine Learning (ML) models were tested: XGBoost, Random Forest (RF) and the K-nearest neighbour (KNN) (Bellin et al., 2022; Konowalik and Nosol, 2021; Valavi et al., 2021). The XGBoost, Random Forest (RF) and KNN models were fitted by R packages `xgboost`, `ranger` and `caret` (Chen et al., 2022; Wright and Ziegler 2017; Kuhn, 2021).

Cross-validation is one of the most widely used data resampling methods to estimate the true prediction error of models and to tune model parameters (Berrar, 2018). To select the best algorithm to model habitat suitability, the 10-fold cross validation was repeated 10 times (100 model runs).

To quantify the performance of each machine learning model, the Area Under the receiving operating Curve (AUC) performance metric was used. The AUC varies between 0-1 and measures the two-dimensional area under the entire Receiving Operating Curve (ROC). Values of AUC equal to 1 represent a perfect predictor, while 0.5 a random predictor. Differences in the performance of models were tested with pairwise Wilcoxon rank sum test applying the Holm correction.

Global sensitivity and uncertainty analysis (GSUA)

For each environmental variable, a uniform distribution was used, and 10000 observations were sampled using a sample design based on quasi random numbers. To quantify the first order and total order indices (S_i and T_i) of each environmental variable, the Sobol method was computed using 1000 permutations (Saltelli et al., 2008). The second order interactions (S_{i+j}) were quantified for the environmental variable with the highest importance. GSUA was carried out using the R package `sensobol` (Puy et al., 2021).

Response curves and habitat suitability (present and future)

For each of the three gorgonian species, the partial response curve of the most important environmental variable was estimated considering the others as fixed at their mean value and the environmental variable that showed the highest second order interaction term with the most important one was fixed at the extremes and at the mean value of the environmental gradient (Bellin et al., 2022).
To obtain the present (2000-2014) and future (2040-2050) predictions of habitat suitability within the study area, the selected algorithm was trained, and the model output (habitat suitability) was predicted across the study area. To estimate the spatial shift of occupancy between present and future condition, a thresholding procedure was applied according to Liu et al. (2013). The threshold that maximizes the sum between sensitivity and specificity was selected, and the model output (habitat suitability) was converted into a binary representation (present or absent). The spatial shift of occupancy was computed as difference between present and future conditions considering the occupied areas. The threshold selection was performed with the package `PresenceAbsence` (Freeman et al., 2008).

Results

After the spatial thinning procedure, a total of 173 presences for *P. clavata,* 189 presences for *E. cavolinii,* and 147 presences for *E. singularis* were obtained.
According to the VIF analysis, several environmental variables showed multicollinearity problems (VIF > 4); benthic layers: temperature, nitrate, phosphate, dissolved oxygen, phytoplankton; surface layers: salinity, diffuse attenuation, and PAR. For this reason, all these variables were removed from the modelling framework (Table 3SM). For the three species, the lasso regression showed that the minimum average value of MAE was reached retaining all the remaining environmental variables in the modelling framework. The computed average MAE values for *P. clavata*, *E. cavolinii* and *E. singularis* were: 0.40, 0.38, 0.41, respectively.

For all the species, the repeated cross validation showed that Random Forest (RF) and XGBoost produced higher median values for AUC (Figure 18) than KNN (Wilcoxon rank sum test $p < .0001$). Difference in model performance between RF and XGBoost was not significant ($p = 0.43$). XGBoost was selected for further analyses for comparison with other bagging approaches used to model the habitat suitability of *P. clavata* (Boavida et al., 2016).

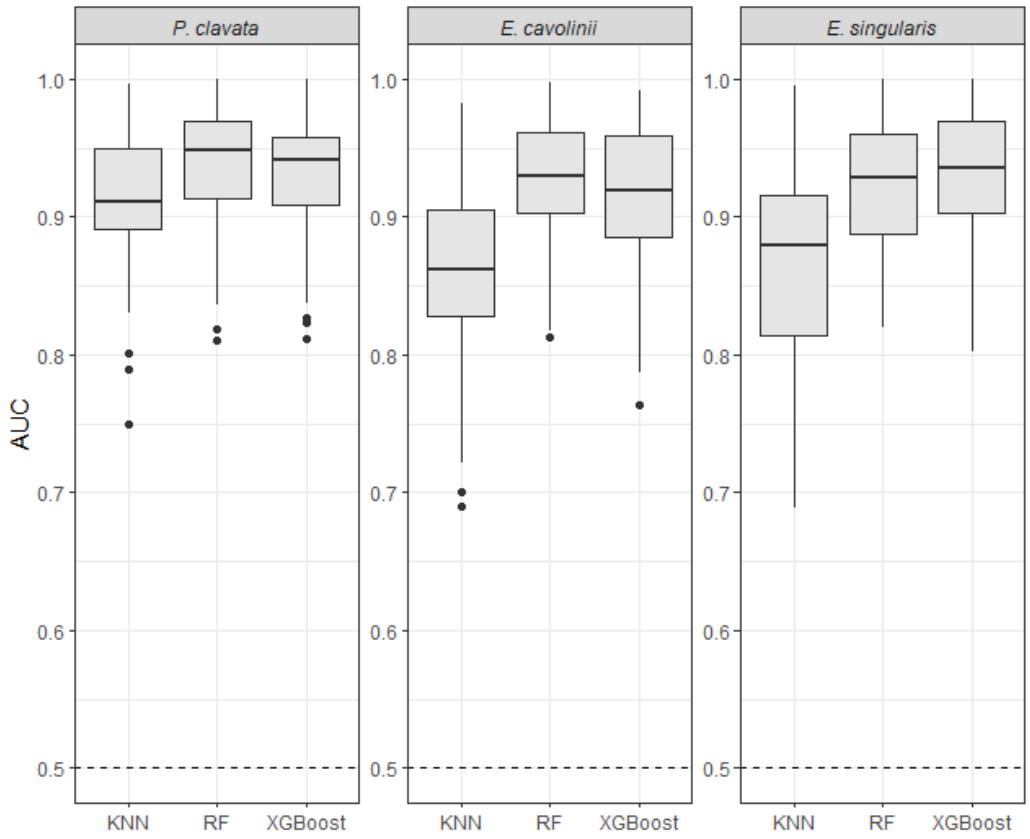

Figure 18 The performance metric AUC was reported to assess the model selection. The horizontal dashed line represented the baseline of a random model.

For all the three species, the most important variable ranked by GSUA was bathymetry, with the highest first and total effect indices (*P. clavata*: $S_{bathymetry}$ = 0.57 and $T_{bathymetry}$= 0.82, *E. cavolinii*: $S_{bathymetry}$ = 0.44 and $T_{bathymetry}$= 0.61, *E. singularis*: $S_{bathymetry}$ = 0.35 and $T_{bathymetry}$= 0.46) (Figure 19). Another important geophysical environmental variable was concavity, especially for *E. singularis* (*P. clavata*: $S_{concavity}$ = 0.06 and $T_{concavity}$ = 0.15, *E. cavolinii*: $S_{concavity}$ = 0.12 and $T_{concavity}$ = 0.20, *E. singularis*: $S_{concavity}$ = 0.31 and $T_{concavity}$ = 0.45). Other environmental variables appeared not important to shape the habitat suitability of the three species with first and total effects values equal or near to zero (surface layers: nitrate and calcite; benthic layer: silicate and bottom light).

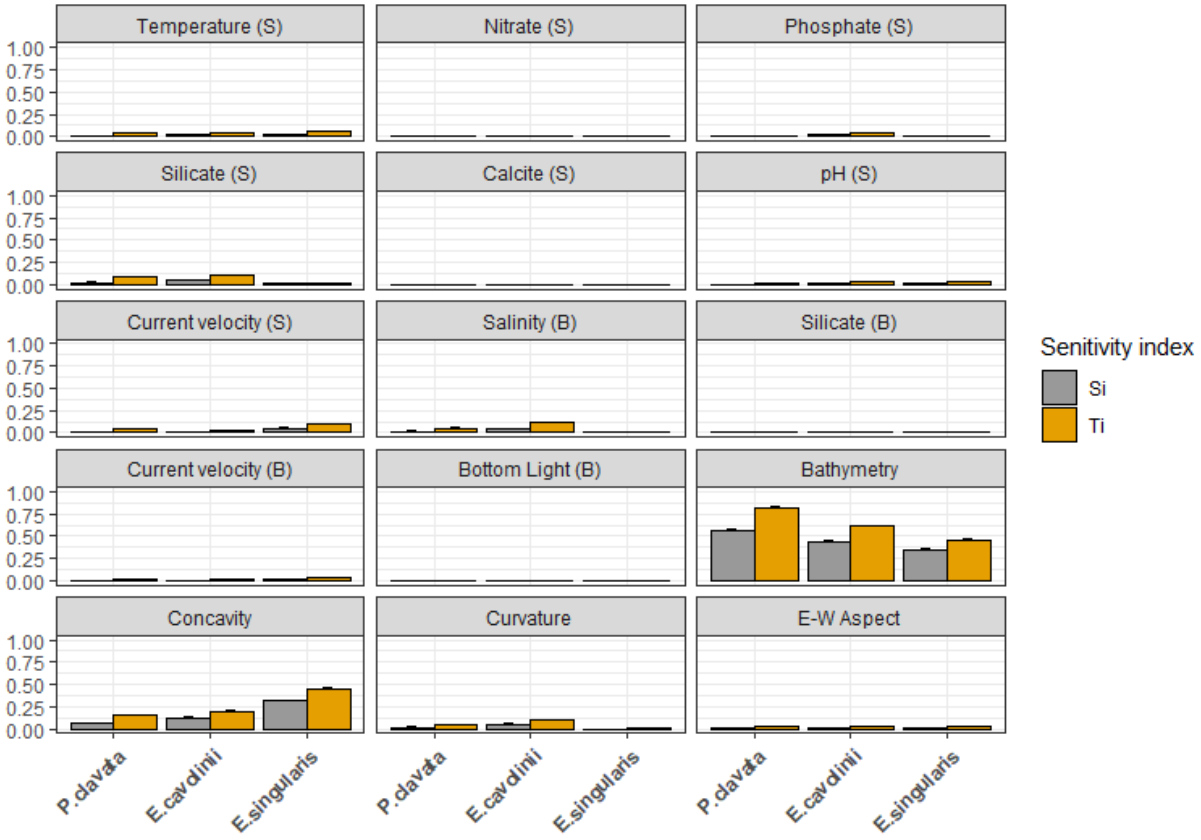

Figure 19 For each environmental variable and species, the sensitivity indices (Sobol indices) were reported in different color: first order (S_i) as grey and total order (T_i) as orange. The error bars represented the standard errors (1000 permutations). The nomenclature (S) and (B) referred to the surface and benthic layer, respectively.

For *P. clavata* the most important second order interactions with bathymetry were silicate in the surface layer followed by concavity ($S_{bathymetry+silicate} = 0.078$ and $S_{bathymetry+concavity} = 0.059$) and for *E. cavolinii* were salinity in the benthic layer followed by concavity ($S_{bathymetry+salinity} = 0.050$ and $S_{bathymetry+concavity} = 0.040$). For *E. singularis* the most important second order interaction with bathymetry was the concavity ($S_{bathymetry+concavity} = 0.054$) while the other environmental predictors showed weak interactions (Figure 20).

According to the response curves, the habitat suitability decreased with bathymetry: deeper than 1000 m for. *P. clavata* and *E. cavolinii*, and deeper than 250 m for the zooxanthellate *E. singularis*. For *P. clavata,* the habitat suitability between 0 and 1000 m was higher at the lowest silicate concentration (1.83 $mol.m^{-3}$). For *E. cavolinii,* the habitat suitability between 0 and 1000 m was higher at low and medium salinity (36.9 PSS). For *E. singularis,* the habitat suitability was the highest between 0 and 250 m; between 250 and 750 m was higher in valleys than in hills (Figure 21).

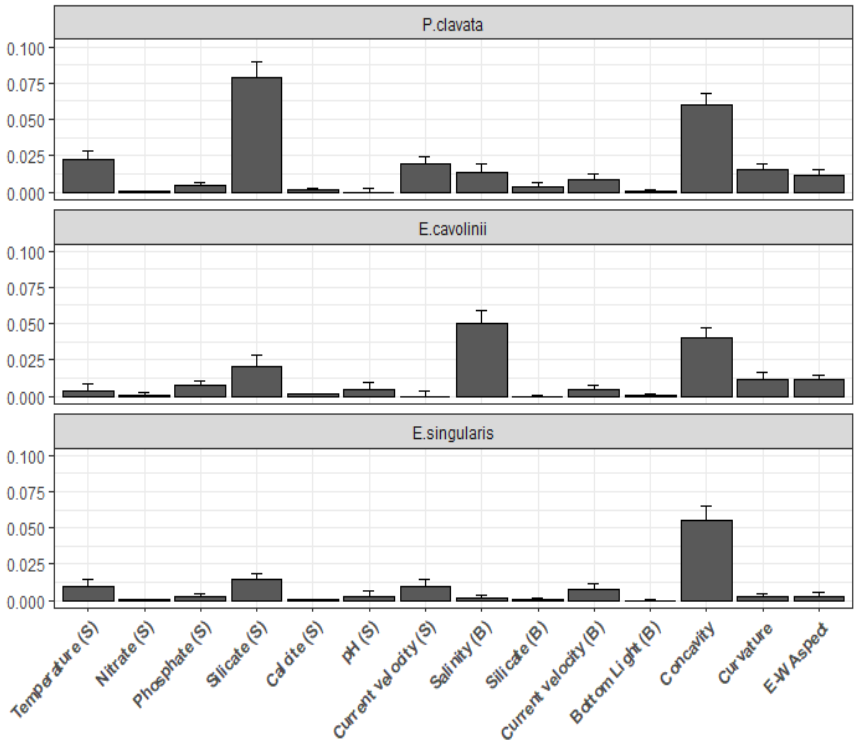

Figure 20 For each species, the second order interaction terms between the most important variable (bathymetry) and the other environmental variables. The error bars represented the standard errors (1000 permutations). The nomenclature (S) and (B) referred to the surface and benthic layer, respectively.

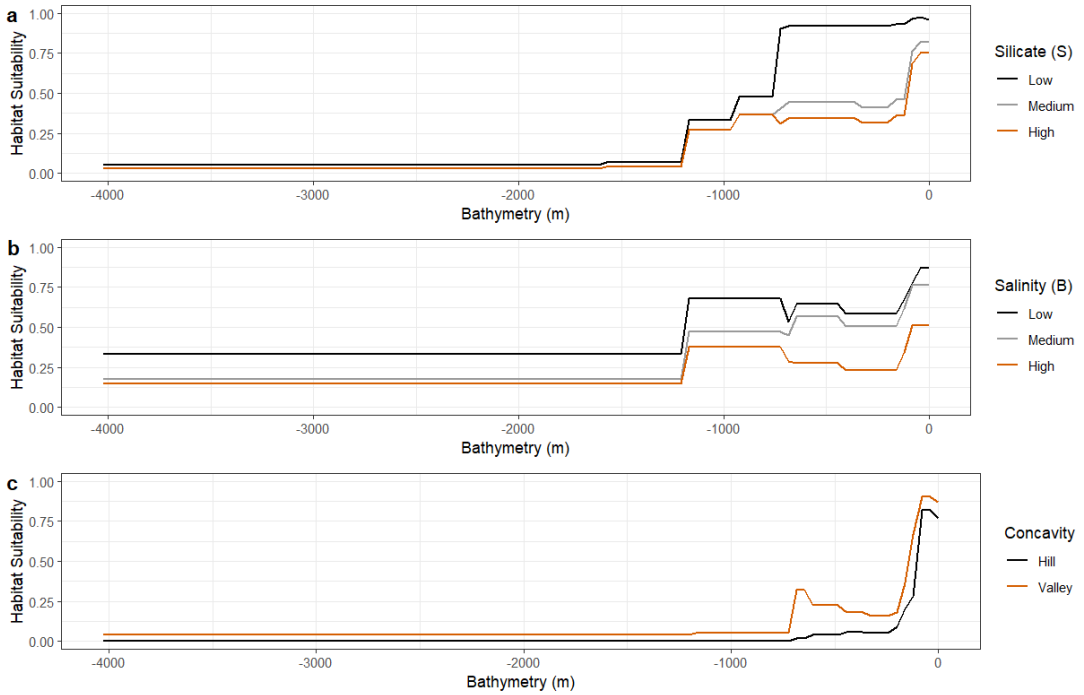

Figure 21 For each species (*P. clavata* (a), *E. cavolinii* (b) and *E. singularis* (c)), response curves of the most important variables (bathymetry) were computed considering the extremes and the mean value of the environmental variables with the highest interaction while the other variables were kept fixed at the mean values. For the concavity, only the extremes values, hill or valley, were considered. The nomenclature (S) and (B) referred to the surface and benthic layer, respectively.

For the three gorgonian species, the highest values of habitat suitability were concentrated in proximity of the coasts. For *E. singularis,* high values of habitat suitability were recorded in North and Central Adriatic Sea, where the bathymetry is lower compared to other areas of the Mediterranean Sea (Figure 22). In future climatic conditions, *P. clavata* was expected to shift the habitat suitability from lower to higher latitudes, mainly in the Adriatic Sea. For *E. cavolinii*, the main pattern of variation between present and future climatic conditions were along the Tyrrenian coast where the main reduction in habitat suitability was observed. In future climatic conditions, the habitat suitability reduction of *E. singularis* was expected mainly along the coasts of the Adriatic Sea.

To estimate the spatial shift of occupancy between present and future condition, a thresholding procedure was applied: the thresholds were 0.5 for *P. clavata,* 0.38 for *E. cavolinii* and 0.39 for *E. singularis*.

In the future, *P. clavata* was expected to increase the occupancy area of 757 Km^2 with respect to the present (+0.6%) (Figure 23) while *E. cavolinii* and *E. singularis* were expected to reduce the occupancy area of 59335 Km^2 and 23341 Km^2 (-49 % and -15%), respectively (Figure 23).

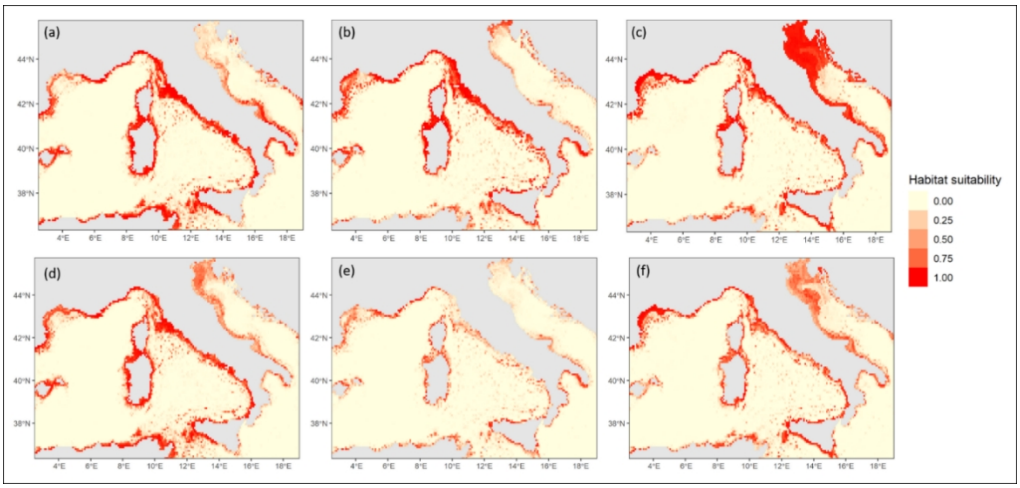

Figure 22 Present and future (2040-2050 RCP8.5) habitat suitability of the three gorgonian species: *P. clavata* (a-d), *E. cavolinii* (b-e) and *E. singularis* (c-f).

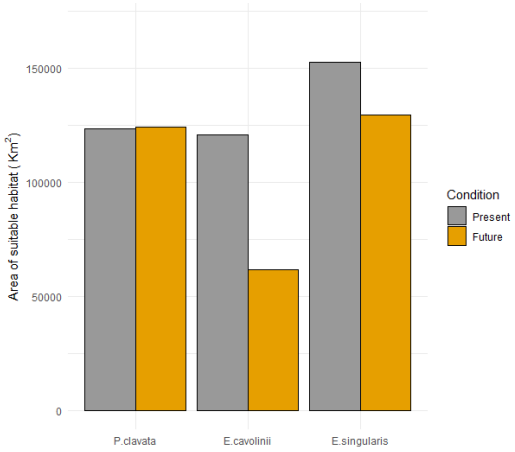

Figure 23 Area of suitable habitat (Km^2) in present and future (2040-2050 RCP8.5) climatic conditions of the three gorgonian species.

Discussion

In this study, we explore a species distribution model framework combined with machine learning algorithms, to assess the potential environmental drivers that shape the distribution at regional level of three Mediterranean gorgonian species *P. clavata, E. cavolinii* and *E. singularis*. The modelling framework was also used to predict their future habitat suitability under the worst climate change IPCC scenario RCP8.5. For all species, the supervised machine learning algorithms XGBoost and RF, reached the highest values of the AUC performance metric. XGBoost was chosen to model the habitat and ecological niche of the three species, to assess the factors that influenced their distribution pattern, and their projection in response to climate change for comparison with Boavida et al. (2016) which used a similar algorithm to model the habitat suitability of *P. clavata*. They showed that temperature (11.5–25.5 °C) and a geophysical variable, the slope, are the most important predictors to define the niche of *P. clavata*. The prediction from these variables modelled a wider distribution than previously known. According to the observations by Sini et al. (2015) and by Masmoudi et al. (2016), our modelling framework predicted the habitat suitability for *P. clavata* and *E. cavolinii* along habitats of Algerian and Tunisian coasts. The methods might be used to identify new populations along poorly sampled area and considering different bathymetries. In fact, Gori et al. (2013) highlight the importance to explore the ecological and evolutionary features of the deep sublittoral gorgonian populations and the connectivity with shallow populations exposed to more frequent perturbations.

By the analysis of variable importance, response curves and GSUA, we showed that the studied gorgonians species respond to similar environmental conditions and the spatial distribution of all the three species in the studied area is influenced by bathymetry. Several studies found that gorgonian habitat suitability was strongly related with geophysical factors: bathymetry relates with topography of the seabed and with hydrography (Yesson et al., 2012; Kinlan et al., 2020). In northwestern Mediterranean, the occurrence of *P. clavata* and *E. singularis* is reported in areas characterized by intense benthic currents (Gori et al., 2013). Bathymetry interacts with topography, temperature and water movement in concentrating nutrients, particulate organic matter and microzooplankton (Mortensen and Buhl-Mortensen 2004; Jenkins and Steven 2021). These complex interactions affect the physiology and the ecological interactions of the gorgonians (Coma and Ribes, 2003; Coma et al., 2004; Ezzat et al., 2013). According to our result, for *P. clavata* and *E. cavolinii* the critical depth is around 1000 meters but the bathymetric occupancy of the two species ranged from 5 to 200 m and 5 to 150 m (Russo 1985; Mokhtar-Jamai et al., 2011; Sini et al., 2015; Gori et al., 2019). These discrepancies might arise from the scale of resolution of the bathymetric layer (1 Km x 1 Km). In raster cells that includes rocky shores, the bathymetry can be very deep hiding the actual slope of the shoreline; however, these results are embedded in the typical range of deep-waters gorgonians in the Mediterranean Sea ranging from 200 m to 1,000 m (Mortensen and Buhl-Mortensen, 2005). Gori et al., (2013) found deep sublittoral populations of *E. singularis* and *P. clavata,* highlighting the importance of survey the distribution of gorgonian species at great bathymetrical range. For *E. singularis*, the zooxanthellate species, the critical bathymetry was lower, around 250 m, than the other two species,

suggesting a dependency with light due to algal endosymbiosis. The endosymbionts can transfer photosynthesized carbon to *E. singularis*, additionally providing autotrophic nutrition.

For all the three species, the most important variable ranked by GSUA was bathymetry, with the highest first and total effect indices. For each species, the bathymetry showed a second order interaction with different environmental variables.

For *P. clavata*, the habitat suitability between 0 and 1000 m was the highest at low silicate concentration. The Mediterranean Sea is an oligotrophic sea with a low silicon concentration (from 1 to 4 μM) (Bergamasco and Malanotte-Rizzoli, 2010; Schroeder et al., 2010; Sospedra et al., 2018). Our result is in accordance with previous studies that identified a negative relationship between coral richness and suitability and silicate concentration (Reyes Bonilla and Cruz Piñón, 2002; Barbosa et al., 2020).

For *E. cavolinii*, the habitat suitability between 0 and 1000 m was high at low and medium salinity. In general, gorgonians seem to withstand hypersaline conditions more easily than reduced salinities, with an optimal range of 29.5–42.5 (Kupfner Johnson and Hallock, 2020).

For *E. singularis*, the habitat suitability was the highest between 0 and 250 m. Between 250 and 750 m the suitability was higher in valleys than in hills. In fact, *E. singularis* is a species commonly found on horizontal or sloping sediment-covered bottoms subjected to irradiance conditions that range between 3 and 44% of surface values. This observation might explain the importance of concavity in relation to bathymetry, where a slight difference in habitat suitability was found between valley and hills.

Under the present-day conditions, the modelling approach properly addressed the expected habitat distribution of the species mainly on rocky coasts along the Ligurian and Tyrrenian Sea (Italy), Corsica (France), around the Elba Island (Italy) and within the Gulf of Lyon (France). Furthermore, it highlighted the presence of populations of the three species around Sardinia and Corsica islands (Italy and France, respectively). For *E. singularis* higher values of habitat suitability were predicted in the North Adriatic Sea. The future projection, in a climate change scenario under the worst IPCC scenario RCP8.5 for the next 30 years, showed a reduction in the habitat suitability of the two *Eunicella* species, especially for *E. cavolinii*. This species is expected to reduce the occupancy range of 49 % from the study area. Contrary to our initial hypothesis, *P. clavata* was expected to increase the occupancy range across the study area shifting the habitat suitability from lower to higher latitudes, mainly in the Adriatic Sea. *E. singularis* was expected to reduce the occupancy area of 15% suggesting that the sensitivity of the symbiotic algae (zooxanthellae) is not the main responsible of the corresponding susceptibility of the *Eunicella* to thermal stress (Kupfner Johnson and Hallock, 2020). *E. cavolinii*, seems the most vulnerable species. This result might be due to the habitat suitability of the species that, according to our result, would be higher at low and medium salinity (36.9 PSS). Increasing sea temperatures should result in more evaporation, leading to an increasing water salinity, especially in the Atlantic Ocean and Mediterranean Sea. In our prediction, the range of salinity in the worst emission scenario was between 37.16 and 39.21 PSS and hence higher than the optimum values of the response curve in relation to bathymetry.

The simulation highlights the importance of bathymetry in the habitat suitability of gorgonians. As stressed by Gori et al. (2011) and Pivotto et al. (2015), new studies are important to explore the ecological characteristics, differences and adaptations of the deep sublittoral populations, the possible connectivity with shallow populations, whether they are exposed to more stable environmental conditions and whether they might play a role in the re-colonization of the shallower areas in which gorgonians are exposed more frequently to fewer stable conditions and to frequent perturbations.

Although variation in environmental factors spatial are key components in determining distribution and abundance patterns of species can also arise due to interactions among organisms (Tignat-Perrier et al., 2022) and to increase tolerance to rapid environmental changes through acclimatization, genetic adaptation, and migration (Hoegh-Guldberg, 2014). For example, the susceptibility of *E. singularis* to heat has been recently investigated in controlled heat stress experiments, demonstrating a differential sensitivity of shallow water populations, with intrinsic physiological mechanisms of acclimatization (Pey et al., 2014). Ultimately, the reproductive cycle of the species, recruitment, the dispersal abilities of the larvae and stochasticity determine the survival growth and reproduction of new individuals. (Chiappone and Sullivan, 1996; Edmunds, 2000; Baird et al., 2003, Gori et al., 2011).

The approach involved only three gorgonian species, but further research might include all the coral species recorded in the Mediterranean area as well as other site in the world and might be significantly improved by considering biological interactions, community composition, functional traits, and the introduction of all possible classes of environmental variables that might act at regional scales.

Measures and modelling of water temperature in the last decade allowed to correlate the mortality event to the rising of water temperature at mesophotic depths, including the displacement in depth of the lower limit of the thermocline. Heatwaves and global warming, together with massive mucilaginous aggregates and macroalgal overgrowth on living corals, represent a combination of stressors that are threatening coastal coral forests in an unprecedented way. When nature is left alone, it has a tremendous ability to care for itself; however, anthropogenic activities play a key role in global environmental change, both driving biodiversity loss and altering ecosystem functioning such as the Mediterranean Sea and corralligenous habitat (Bramanti et al., 2017; Ponti et al., 2018; Sini et al., 2019; Coppari et al., 2019).

2.3 Assessing climate change's impacts on the habitat suitability of two coral species in the Mediterranean Sea

Coral reefs are the most biodiverse marine ecosystem, providing breeding areas, nurseries and food for many economically important marine species, and forming an important link in nutrient cycling from land to the open ocean (Cabral and Geronimo, 2018; Froelich, 2002). Thus, their protection constitutes a key element for the conservation of oceanic fauna. However, their survival is now critically threatened by climate change. Increasing greenhouse gas emissions into the atmosphere are inducing water warming, acidification and deoxygenation (Gruber, 2011; Bijma et al., 2013; Bindoff et al., 2019). Consequently, oceanographic parameters are shifted towards new equilibrium points, making it hard for marine species to adapt to such rapid changes. Variations in ocean chemistry and currents, sea-level rise, increased storm intensity and altered nutrient availabilities are some of the main factors responsible for the marine biodiversity loss of the last century (Bindoff et al., 2019). A warming ocean causes thermal stress that contributes to coral bleaching and infectious diseases (Rosenberg and Ben-Haim, 2002; Hughes et al., 2017). Simultaneously, sea level rise may lead to an increasing sedimentation for reefs located near land-based sources of sediment, leading to the smothering of coral (Jones et al., 2019). Ocean acidification is causing a reduction in pH levels which decreases coral growth and structural integrity (Fantazzini et al., 2015; Movilla et al., 2016; Caroselli et al., 2019; Kline et al., 2019). Extreme weather events can produce stronger and more frequent storms causing the destruction of coral reefs (Knutson, 2021; Harmelin-Vivien, 1994). Altered ocean currents may lead to changes in connectivity and temperature regimes contributing to the lack of food for corals and hampering the dispersal of coral larvae (Munday et al., 2019). Eventually, water pollution is responsible for a strong stress response in most marine animals (Stoliar and Lushchak, 2019; Lushchak, 2011).

The Mediterranean Sea has long stood out in successive generations of global climate models as being particularly sensitive to rising concentrations of greenhouse gases (Tuel and Eltahir, 2020). With temperatures increasing much faster than the global average, and sea level rise expected to exceed one meter by 2100, the Mediterranean is becoming the fastest-warming sea on Earth, and it is considered a climate change hot spot (Tuel and Eltahir, 2020). The conservation status of most Mediterranean coral species is currently classified as least concern (Otero et al., 2017). Nevertheless, considering the lack of data but the strong impacts that climate change can have on marine ecosystems, predicting coral resilience to such changes of Mediterranean coastal environments deserves urgent investigation.

The predictive potential makes SDMs a great tool to study climate change's impacts on global biodiversity (Ramirez-Vollegas et al. 2014). Couce et al. (2012) showed that in shallow-water corals the dominant environmental predictors are temperature-related variables, such as annual mean sea surface temperature (SST), monthly and weekly minimum SST, followed by regional patterns of nutrient concentrations, light availability and aragonite saturation state. In the Gulf of Mexico, geomorphology of the sea bottom, along with temperature, salinity, depth, acidity, dissolved oxygen, and chlorophyll-a, was a key variable in the

determination of the coral distribution (Hu et al., 2020). Similar results are useful for the creation of conservation plans and help support conservation prioritization and management.

Among the most common corals inhabiting the Mediterranean Sea, in this study *Balanophyllia europaea* and *Leptopsammia pruvoti* were selected as model species to predict present-day and future habitat suitability. Both species are Scleractinia solitary corals, native of the Mediterranean Sea. However, while *B. europaea* is endemic, *L. pruvoti* expands its distribution range in the Atlantic Ocean along the coasts of Portugal, Brittany, the Channel Islands and southwestern England. Besides sharing the same growth mode (solitary polyp), these species differ for their trophic strategy (*B. europaea* is mixotrophic, mostly relying its organic carbon source on symbiosis with zooxanthellae unicellular algae, while *L. pruvoti is* heterotrophic, acquiring organic carbon and nutrients from external predation on plankton and particulate organic matter), and distinct susceptibilities towards environmental stress (Franzellitti et al., 2018). Several biological parameters related to coral growth, calcification, and abundance have been showed as negatively correlated with sea surface temperature in *B. europaea* Goffredo et al., 2007, 2008, 2009; Airi et al. 2014), while almost unchanged in *L. pruvoti* (Goffredo et al., 2007; Caroselli et al., 2012a, b; Airi et al., 2017). This leads to the assumption that *L. pruvoti* could be more tolerant to temperature changes, and that lack of symbiosis and heterotrophic predation may play a role in prompting stress tolerance of this species (Caroselli et al., 2015).

The International Union for Conservation of Nature (IUCN) classified *B. europaea* as at Least Concern (LC), with a last assessment being produced in 2014 (Bavestrello et al., 2015), while for *L. pruvoti* only a regional assessment is available (IUCN Italian Committee, 2014), which means that at present neither species is under any special conservation measure.

Here we try to estimate potential present-day spatial distribution patterns of the coral species within the Mediterranean Sea, which are currently pending. We project habitat suitability under the SPP5-8.5 IPCC scenario (IPCC, 2021) unraveling the main environmental parameters related to distribution changes in the coral species and potential linkages with their respective physiological needs and life history traits. Determining the spatial distribution of these coral species and their present day and future dynamics is a critical first step towards supporting regional management plans in the Mediterranean Sea.

Materials and Methods

Species presence data and environmental variables

For both *B. europaea* and *L. pruvoti*, presence points within the Mediterranean Sea (Figure 24) were obtained from the Global Biodiversity Information Facility database (GBIF) (www.gbif.org) using the R package dismo (Hijmans, et al., 2011). Further occurrences were obtained from scientific literature. Particularly, presence points of *B. europaea* were obtained from the studies of Goffredo et al. (2004, 2008, 2015), Kruzic´ and Popijac (2015), Purser et al. (2014), Terrón-Sigler and López-González (2005), Rodolfo-Metalpa et al. (2001), Fenner et al. (2013) and Zibrowius (1980). Occurrences of *L. pruvoti* were obtained

from Caroselli et al. (2012), Zibrowius (1980), Goffredo et al. (2009), Fenner et al. (2013), Boscari et al. (2019), Gerovasileiou et al. (2015).

Although *L. pruvoti* also inhabits colder waters around the United Kingdom, both these coral species are widespread in the Mediterranean Sea and the collection of occurrence points was limited to this basin only (Jackson, 2008). Moreover, even though the presence of *L. pruvoti* and *B. europaea* along the African coasts as well as in Israel is reported (Caroselli and Goffredo, 2014), coordinate points for all the southern and eastern portion of the Mediterranean Sea were unavailable. The presence data were collected as spatial points whose longitudinal and latitudinal coordinates referred to the Coordinate Reference System (CRS) WGS84. The dataset included information on the coordinate uncertainty in meters related to each occurrence. To reduce the geo-localization error, each occurrence with an uncertainty higher than 250 meters was discarded. The `duplicate` function in the R package (R core Team, 2021) was used to delete the duplicate data thus avoiding data redundancy. Geo localization mismatches were also checked and excluded when found. The final cleaned dataset contained a total of 1058 data points, 482 for *B. europaea* and 576 for *L. pruvoti*.

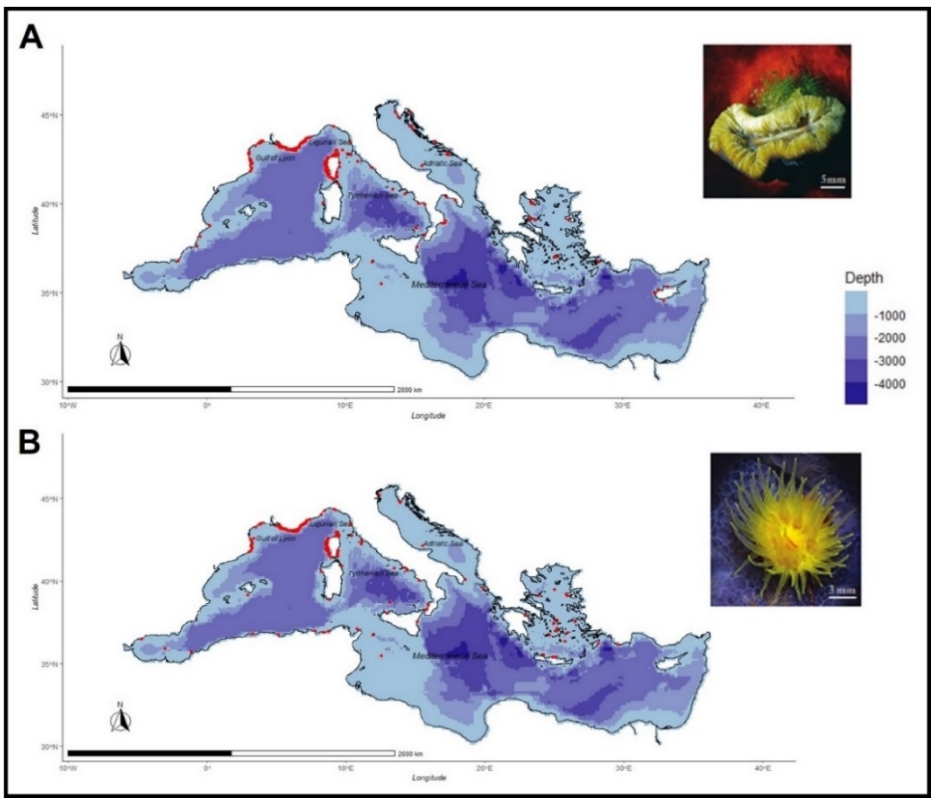

Figure 24 Occurrences of *B. europaea* (A) and *L. pruvoti* (B) reported as red points within the Mediterranean Sea. The bathymetric layer is reported in blue colour gradient.

To model present-day (2000-2014) habitat suitability for the selected species, 12 physico-chemical environmental variables were retrieved from the BioOracle database (bio-oracle.org), and bathymetry was retrieved from the GEBCO global bathymetric grids (gebco.net) (Table 4). Spatial patterns of the variables in the study area are reported in Figure 25.

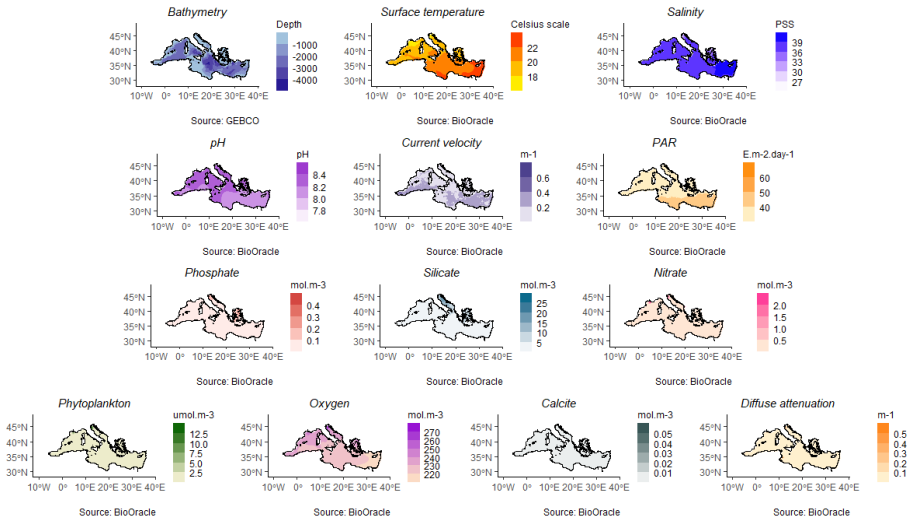

Figure 25 Raster layers of the 13 selected environmental variables.

For the predicted future climatic conditions (2040-2050), the SPP5-8.5 (IPCC, 2021) emission scenario was selected, and change patterns of current velocity, sea surface temperature and salinity were retrieved from BioOracle (bio-oracle.org). The future average pH of the Mediterranean Sea was estimated using the annual trend of pH reduction (−0.0044 units per year) calculated by high frequency observational data (Flecha et al., 2015) from 2014 to 2045. The other environmental variables were kept constant at present values. All environmental variables were at 30 seconds of resolution in the CRS WGS84.

Table 4 List of the selected physico-chemical environmental variables.

Type	Variable	Units	Source database
Geomorphological	Bathymetry	m	GEBCO (gebco.net)
Chemico-physical	Sea surface temperature (SST)	°C	BioOracle (Assis et al., 2017; bio-oracle.org)
	Current velocity	m/s	
	Photosynthetic available radiation (PAR)	$E.m^{-2}.day^{-1}$	
	Diffuse attenuation	m^{-1}	
	Phytoplankton Calcite (*Expressed as carbon in sea water*)	$\mu mol.m^{-3}$	
	pH	-	
	Phosphate	$\mu mol.m^{-3}$	

	Nitrate	µmol.m^{-3}	
	Silicate	µmol.m^{-3}	
	Dissolved oxygen	µmol.m^{-3}	
	Calcite	mol.m^{-3}	
	Salinity	psu	

Modelling approach

Spatial thinning and VIF analysis

Spatial thinning was applied with a minimum distance of 1 Km among points (Steen et al., 2020). The thinning algorithm was repeated 50 times. The repetition with the highest number of occurrences and with the highest mean geographic distance between points was selected. This procedure was carried out with the R package `spThin` (Aiello-Lammens et al., 2015).

A threshold of VIF = 10 was used and only environmental variables with VIF values < 10 were kept within the modelling framework.

Pseudo-absences generation

For both species, the occurrences were presence-only data consisting of the locations of species observations but lacking absence points (Renner et al., 2015). The pseudo-absences data were randomly sampled at depths higher than 100 m using a balanced number of points with respect to the presence data (Barbet-Massin et al., 2012). Every pseudoabsence generated along shallow rocky shores was discarded and substituted. The cycle was repeated 50 times in the Mediterranean basin and 50 pseudo-absence datasets were obtained. The repetition with the greater mean geographic distance among pseudo-absence points was selected. The minimum distance between each presence and pseudo-absence point was also checked using a threshold of 5 km: no absence point was closer than 5 km to any presence point. This ensured there was no overlapping between presences and pseudo-absences nor between raster cells.

Model selection

To model habitat suitability of the two species, five Machine Learning (ML) models were tested: XGBoost, Random Forest (RF), Support Vector Machine (SVM), Artificial Neural Network (ANN) and an ensemble (EN) of the previous models (Bellin et al., 2022; Konowalik and Nosol, 2021; Valavi et al., 2021). The XGBoost, Random Forest (RF), Support Vector Machine (SVM), Artificial Neural Network (ANN) models were fitted by R packages `xgboost`, `ranger`, `e1071`, `keras` and `tensorflow`, respectively (Chollet et al., 2015; Abadi et. al., 2015; Chen and Guestrin, 2016; Wright et al., 2017; Meyer et al., 2021). Ensemble learning allows the merging of several different model algorithms together (Opitz and Maclin, 1999). The

ensemble model was fitted considering the mean prediction of each single model allows the merging of several different model algorithms together (Opitz and Maclin, 1999). The ensemble model was fitted considering the mean prediction of each single model.

To select the best algorithm to model habitat suitability, an 8-fold cross validation was used. Cross-validation is one of the most widely used data resampling methods to estimate the true prediction error of models and to tune model parameters (Berrar, 2018). To define the folds, we used a random assignment of each observation to a particular fold retaining the original balanced number of presences and pseudo-absences. The Kolmogorov-Smirnov test was applied on each environmental variable in a specific fold, to test if all generated folds were statistically representative of the whole environmental variable dataset.

To quantify the performance of each machine learning model, two performance metrics were used: the Area Under the receiving operating Curve (AUC), the Mean Absolute Error (MAE). MAE is a measure of error: it is the average over the test sample of the absolute differences between predictions and actual observations where all individual differences have equal weight. Therefore, the lower the value, the better the model prediction. Following the results of Konowalik and Nosol (2021), AUC and MAE were graphically combined, being respectively placed on the y-axis and x-axis: the best-performing algorithms should be placed in the upper left region of the plot.

Variable importance, response curves and predictions

To quantify the variable importance for the habitat suitability of both species, a random shuffling (500 times) of each environmental variable was made. MAE was used to monitor the reduction in algorithm performance. The variable importance was estimated using a permutation procedure and MAE was computed after permuting each environmental variable. Whenever MAE increased, the specific environmental variable was considered important.

For both species, the partial response curves for the three most important environmental variables were estimated considering the others as fixed at their mean value (Bellin et al., 2022).

To obtain the present (2000-2014) and future (2040-2050) predictions of habitat suitability within the Mediterranean Sea, the selected algorithm was run 100 times considering 100 different sets of randomly sampled pseudo-absences and the mean prediction of habitat suitability was considered. To obtain a presence-absence model output, a binarization procedure was carried out. The threshold was computed considering the mathematical mean of 5 different criteria (Table 5) (Freeman and Moisen, 2008).

The overall interaction terms of all environmental variables and the two-way interactions of the environmental variables that revealed the highest interactions strength were computed using the H-statistic (Friedman and Popescu, 2008). The variable importance and the environmental variable interactions were computed using the R package `iml` (Molnar et al., 2018). The 5 different criteria of threshold selection were computed using the R package `PresenceAbsence` (Freeman and Moisen, 2008).

Table 5 List and descriptions of the 5 criteria of threshold selection

Threshold criteria	Description
MaxSens+Spec	Maximization of the sum of sensitivity and specificity metrics
MaxKappa	Maximization of the value of Kappa metric
MaxPCC	Maximization of the value of percent of correctly classified metric
MaxROC	Minimization of the distance between the ROC curve and the top left corner of the ROC space
Cost	Costs between false positive and false negative metrics

Results

Spatial thinning and variables selection

The spatial thinning algorithm and the random pseudo-absence sampling produced a total of 285 presences and 285 pseudo-absences for *B. europaea* and a total of 235 presences and 235 pseudo-absences for *L. pruvoti*.

The VIF analysis identified dissolved oxygen as the only environmental variable with multicollinearity issues (VIF > 10); therefore, it was removed from the modelling framework (Table 4SM).

For both species, the random fold generation produced folds with environmental variables that were not statistically different from the original sample distribution after Kolmogorov-Smirnov test ($p > 0.05$; Table 5SM).

Model evaluation

For *B. europaea,* the 8-fold spatial cross validation showed that Ensemble (EN) and Random Forest (RF) produced the highest mean values for AUC (0.997) and XGBoost produced the lowest value for MAE (0.0412). Moreover, XGBoost was occupying the most top-left side of the graph when plotting together AUC and MAE values (Figure 26). It was thus selected as the best model. For *L. pruvoti* the 8-fold cross validation showed that the random forest (RF) reached the highest AUC value (0.994), while the lowest MAE value (0.0543) was reached by XGBoost. The biplot combined metrics showed that two models were placed in the top-left side of the graph: XGBoost (Figure 26).

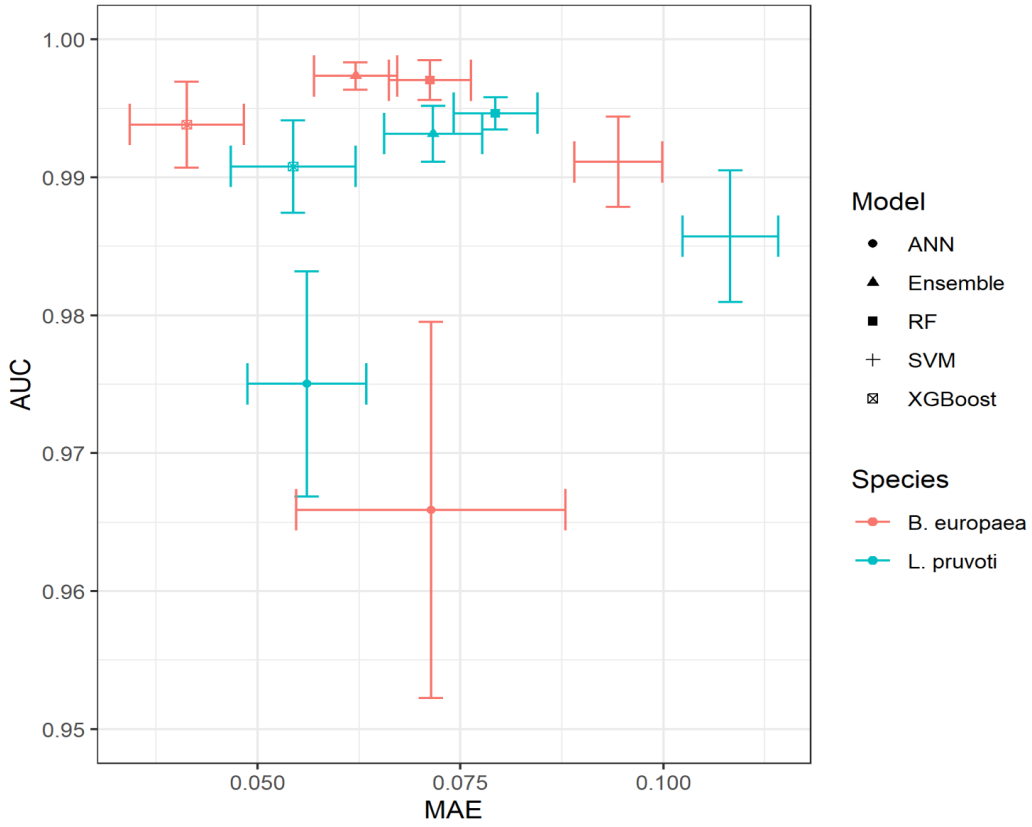

Figure 26 Biplot of the performance metrics. The Area Under the receiving operating Curve (ROC) (AUC) and the mean absolute error (MAE) are combined to assess the model selection. Models that showed the highest performance (high AUC and low MAE) are placed in the top-left region of the graph.

Variable importance, interactions, and response curves

For both *B. europaea* and *L. pruvoti* the most important variable was the current velocity, with a mean increase in MAE values of 55.22 (permutation error = 0.20) and 35.15 (permutation error =0.17), respectively (Figure 27). For *B. europaea,* the response curve of current velocity showed an optimum of habitat suitability in the range of 0 and 0.1 m/s, with an abrupt reduction registered for higher values (Figure 27). For *L. pruvoti*, response curve for current velocity showed an abrupt reduction in habitat suitability at values higher than 0.1 m/s (Figure 28).

The habitat suitability decreased at deep bathymetric levels (Figure 28). A change in silicate values corresponds to a very moderate change in the habitat suitability of the animal. Therefore, even though listed as among the three most important variables by permutation analyses, silicate showed a consistently negligible relevance in shaping habitat suitability of *B. europaea* compared to current velocity and bathymetry.

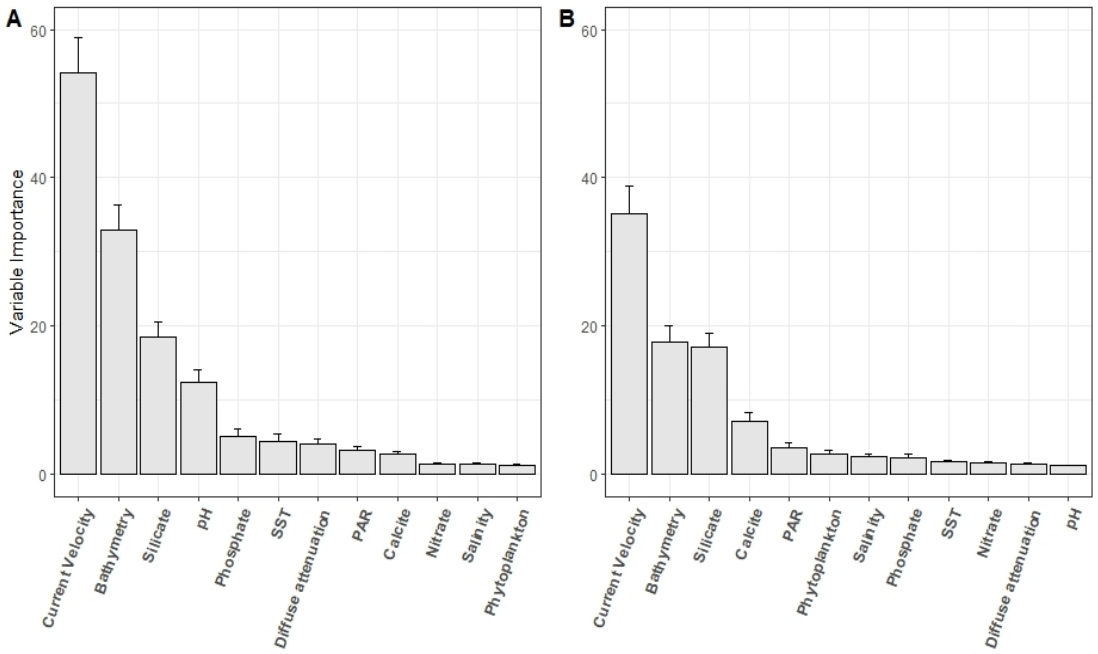

Figure 27 Ranking plot of variable importance (assessed as mean increase of MAE) for *B. europaea* (A) and *L. pruvoti* (B). Error bars indicate 95% confidence intervals (500 bootstrap replicates).

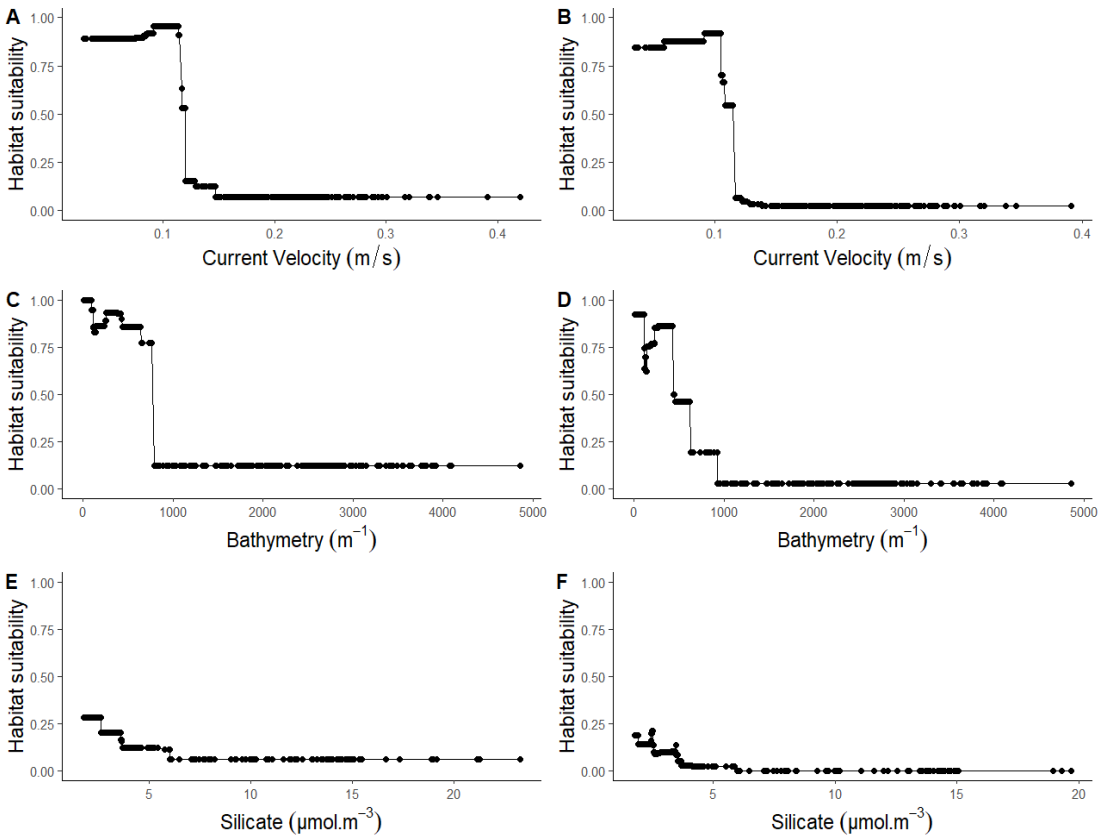

Figure 28 Response curves of the three most important variable generated by the XGBoost model for *B. europaea*: current velocity (A), bathymetry (C), silicate (E); and for *L. pruvoti*: current velocity (B), bathymetry (D), silicate (F).

For *B. europaea*, the variable that showed the highest interaction term was the current velocity (H statistic = 0.29) (Figure 29). This variable showed the highest second order interaction with sea surface temperature (H statistic = 0.83) and the lowest second order interaction with phytoplankton (H statistic = 0.02). For *L. pruvoti*, the variable that showed the highest interaction terms was the silicate (H statistic = 0.41) (Figure 29). This variable showed the highest second order interactions term with calcite (H statistic = 0.38) and the lowest second order interaction term with pH (H statistic = 0.05) (Figure 29).

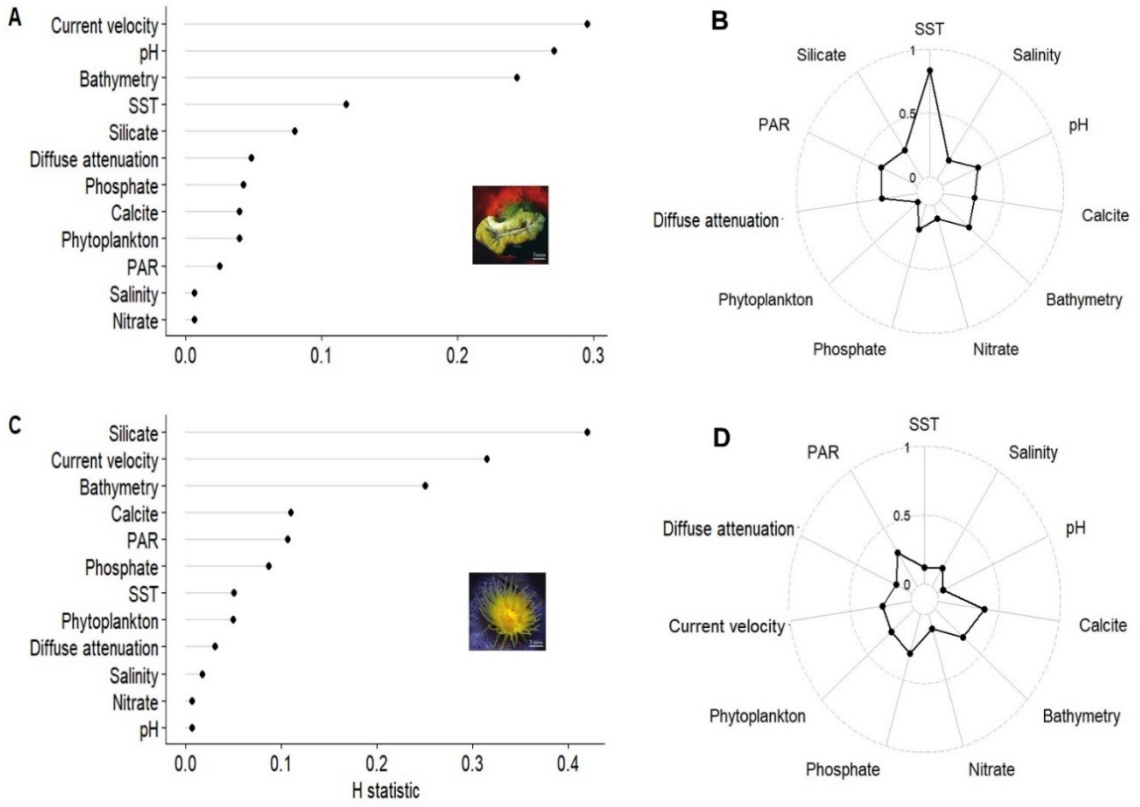

Figure 29 Total interaction terms (H statistic) computed for *B. europaea* (A) and *L. pruvoti* (C). Polar plots reporting the two-way interactions (H statistic) for the variables with the highest total interaction terms for *B. europaea* (B) and *L. pruvoti* (D).

Present and future habitat suitability

The calculated threshold selection was 0.69 for *B. europaea* and 0.66 for *L. pruvoti*. In present-day conditions, the occurrence of both species seems concentrated around Corsica and Sardinia, along the Ligurian and Tyrrenian shorelines, around the Elba Island (Italy) and within the gulf of Lyon from Marseille to Barcelona and in Greece (Figure 30). Furthermore, SDM for *B. europaea* showed occurrences along the Croatian coasts, while for *L. pruvoti* good habitat suitability is predicted in the area around Gibraltar and along the Algerian coasts (Figure 30). Neither of the species is projected to have suitable habitats in the central Adriatic Sea as well as in the deeper areas of the whole Mediterranean (Figure 30).

The extension of suitable habitats for both species is expected to undergo a remarkable reduction by the end of the century and depending on the geographical areas. Such a reduction appears more severe for *B. europaea* than for *L. pruvoti*. Both species are expected to have a strong reduction in habitat suitability along the Southern Tyrrhenian coasts as well as along the South Adriatic coastal areas and in Greece. Nevertheless, *L. pruvoti* was projected to retain suitable habitats along North-Weststern Mediterranean coasts (Ligurian shorelines, in the last corner of the Gulf of Lyon and at the border with Spain), around the Elba Island (Italy), but also along some traits of the Algerian coasts and near Gibraltar. A slight potential presence of *L. pruvoti* is also predicted to persist in Sardinia and Puglia (Italy).

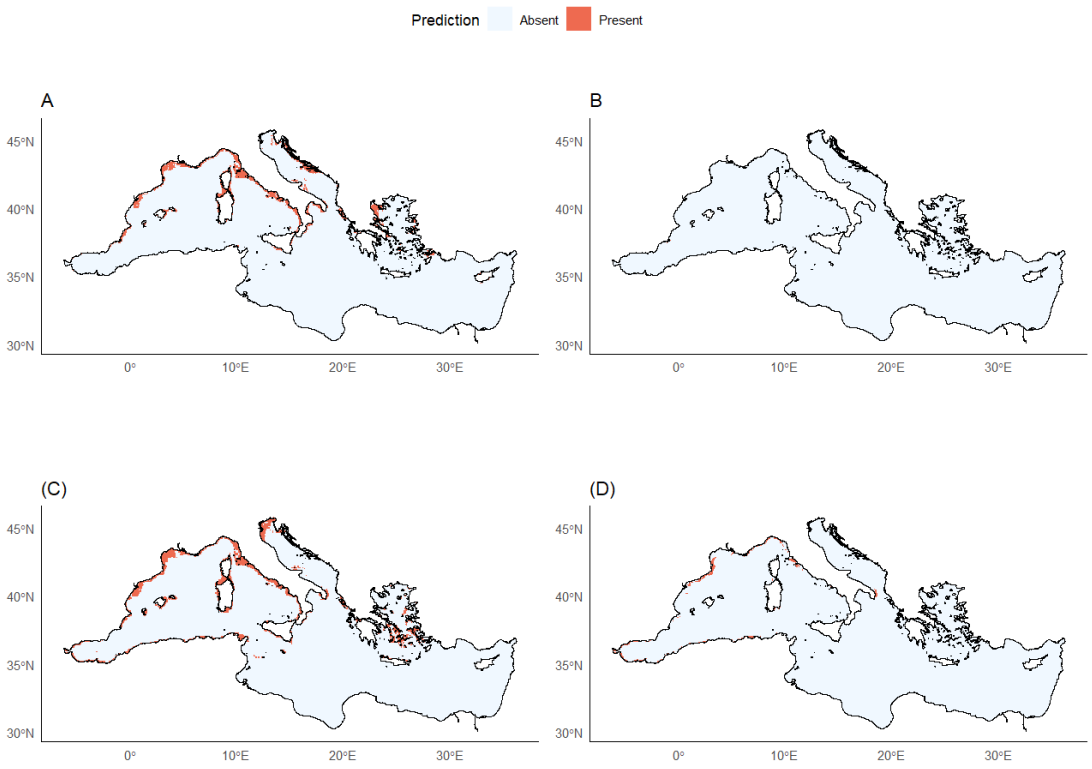

Figure 30 Habitat suitability maps for *B. europaea* (A-B) and *L. pruvoti* (C-D) predicted by the XGBoost model under (A - C) present-day conditions and (B - D) projected SPP5-8.5 IPCC scenario.

Discussion

Changes in environmental parameters are currently impacting the natural habitats of numerous marine species and ocean warming is just one of the multiple shifts which are now affecting the survival of sea animals and plants. In this study, the habitat suitability of *B. europaea* and *L. pruvoti* was modelled in the whole Mediterranean basin based on the presence data currently available in public databases and in literature. Five different models in the framework of a machine learning approach were tested. All the models performed well in producing habitat suitability predictions for both species. This may be explained by the rigorous process of variable selection during the earliest conceptualization stage and the later

statistical filtering together with the representativeness of the occurrence points in describing the realized niches of both species (Irving et al., 2019; Van Eupen et al., 2021). Moreover, the process of uncertainty control, which limited the spatial error, as well as the balanced sampling of pseudo absences contributed to the goodness of fit (Steen et al., 2020).

Under the present-day conditions, the modelling approach properly addressed the expected habitat distribution of the species mainly on rocky bottoms along the Ligurian and Tyrrenian coasts of Italy, Corsica, around the Elba Island, within the Gulf of Lyon, in agreement with the real presence dataset. Furthermore, it highlighted the presence of populations of both species around Sardinia, and in Greece, where habitat suitability appeared consistently high. *L. pruvoti* showed a potential distribution along the Algerian coasts. *B. europaea*, instead, showed a remarkable presence along the Croatian coasts, again in agreement with the real presence dataset, although real data for the Eastern Adriatic border are spatially scattered, while the modelled output give a consistently continuous distribution.

In general, low current velocities and shallow bathymetries characterize the suitable present-day habitat conditions for both species in the Mediterranean Sea, as suggested by MAE and response curve analyses. Whether bathymetry is well known to constrain vertical distribution of corals as a function of temperature and light profiles, current velocity can shape distribution and abundance directly by affecting settlement, growth, survival of individuals, and food source delivery (Poff et al., 1997), and indirectly by altering environmental cues and affecting coral sensing its surrounding environment (Lenihan et al., 2015).

These results are in line with the species distribution and the prior knowledge of the physiology of the animals. Caroselli et al. (2020) compared population dynamics of *B. europaea* populations from the Dardanelles with populations of the north-western Mediterranean Sea. Results showed a positive correlation between population density and depth in the Dardanelles. This was explained by considering lower current and wave action, and higher salinity at higher depth. In comparison with Italian populations, age structures presented a higher frequency of young individuals and were more stable in the Dardanelles, likely due to the less intense wave action. These findings envisage the relevance of currents and salinity in shaping species niche and a strong correlation of these two variables with the ocean bathymetry.

As to bathymetry, we should consider that response curves showed a prevalence of the species within the first 100 m, while we know that these corals are generally staying at shallower depths *B. europaea* is inhabiting waters up to 50 m while *L. pruvoti* up to 70 m. This bias may be related to resolution of raster cells available for the modelling approach (1 km x 1 km).

Future predictions under the SPP5-8.5 (IPCC, 2021) showed remarkable reductions of habitat suitability for both species. According to the employed threshold selection, loss of habitat suitability is expected to be dramatically severe for *B. europaea* with apparently no clear spatial patterning. On the other hand, despite the relevant predicted habitat loss, *L. pruvoti* is expected to shift its distribution towards the Western Mediterranean Sea. Such projections are coherent with historical data and scientific literature on these species. Indeed, Goffredo et al. (2008) hypothesized that *B. europaea* may be close to its thermal limits by 2100, showing relevant changes of its physiological features in response to temperature. On the contrary, *L.*

pruvoti seems more tolerant to temperature increase (Caroselli et al., 2012; Franzellitti et al., 2018). Changes of further key environmental variables show opposite responses between the two species. For example, *B. europaea* seems more tolerant to reductions of seawater pH compared to *L. pruvoti*, at least in the short term (Goffredo et al., 2015).

Overall, the modelling approach employed in this study successfully estimated present-day distribution of the temperate corals *B. europaea* and *L. pruvoti* across the Mediterranean Sea. This result is *per se* a relevant contribution for planning future monitoring efforts that should be undertaken in the field of marine biodiversity conservation strategies in the area and deserved to the Eastern Mediterranean region (EU Green Deal; EU Biodiversity Strategy to 2020).

The conservation status of *B. europaea* was classified as least concern according to the International Union for the Conservation of Nature (IUCN Red List, 2022; https://www.iucnredlist.org/). *L. pruvoti* does not appear within the global list of threatened species but was also registered as least concern by the Italian Committee (IUCN Comitato Italiano, http://www.iucn.it/). In warmer regions, *B. europaea* showed a decrease in population density and biological fitness. IUCN reported that the population drop will not reach the thresholds for a vulnerable status considering three generation lengths (approximately 30 years). In our study we assess that the habitat suitability of these species might be reduced considering the future climatic condition under the worst emission scenario. Other sources of threat might act to reduce the fitness and population growth rate such as competition with alien species, diseases, anthropic effluents and pollution of the surface waters and human intrusion/recreational activities in proximity of the shore (EPA, 2022). Measures of conservation for *B. europaea* and *L. pruvoti* are not currently available. Our study brings new information for the ecology and habitat suitability in the Mediterranean Sea that should be considered. Novel study will be carried out to improve the knowledge of environmental stress that might act on the realized habitat niche of both species. The establishment of monitoring and management planning strategies for these temperate coral species are recommended.

Chapter 2: Integration of Geometric Morphometric with Machine Learning

2.1 Supervised and Unsupervised machine learning combined with geometric morphometrics as tools for the identification of inter and intraspecific variations in the Anopheles Maculipennis complex

The *Anopheles* genus includes more than 480 species, of which 70 are known to transmit malaria (Manguin et al., 2008). The genus includes several complexes of species, often indistinguishable at the morphological level, and with different vectorial capacity (Manguin et al., 2010). The Maculipennis complex is one of these groups, which shows a Holarctic distribution. Species of this complex may be differentiated by the egg morphology and the decoration of the exochorion; some species were defined on a cytogenetic basis. Barcoding techniques, especially the use of ITS2 marker, are very useful in identifying the species of the complex. The application of these techniques led to the definition of the new taxon *Anopheles (An.) daciae* based on ITS2 polymorphism (Lilja et al., 2020; Nicolescu et al., 2004). This taxon is strictly related to *An. messeae* and had a debated rank; here we refer to it as *An. daciae* sp. inq. (species inquirenda). Italy was declared malaria-free by the WHO in 1970; the last endemic cases of malaria, due to *Plasmodium vivax*, were recorded in Sicily in the 1962. Several cryptic (e.g. with an unknown mode of acquisition) cases of malaria were then reported in Italy, suggesting the possibility of local transmission of the disease if a carrier arrives when potential vectors are still present. In Italy, the *Anopheles maculipennis* complex comprises 5 sibling species not morphologically recognizable: *An. atroparvus*, *An. labranchiae*, *An. maculipennis* s. s., *An. messeae* and *An. subalpinus* (Boccolini et al., 2020), the last synonymized with *An. melanoon* (Linton et al., 2002). An extensive study conducted in Northern Italy found that all the mosquitoes referable to the *An. messeae/An. daciae* taxon bear the ITS2 polymorphic basis referable to *An. daciae* sp. inq. (Calzolari et al., 2021). A sixth species *An. sacharovi* disappeared or became very rare in Italy (Bietolini et al., 2006). The main Italian historic malaria vectors were *An. sacharovi* and *An. labranchiae,* the last present in Central and Southern Italy. *An. atroparvus* and *An. melanoon* were considered occasional vectors in some area of the Po plain. *An. messeae* is predominantly zoophilic (Severini et al., 2009) but its malariogenic potential was recently reported in Russia (Mironova et al., 2020). The distribution in Northern Italy was recently updated by an extensive field sampling identifying four species of the Maculipennis complex by means of the barcoding technique: *An. maculipennis s. s.*, *An. daciae* sp. inq., *An. atroparvus* and *An. melanoon* (Calzolari et al., 2021).

The occurrence of cryptic sibling-species (morphologically similar but genetically distinct species) is far more common than previously thought (Pfenninger and Schwenk, 2007). On the other hand, due to phenotypic plasticity, i.e. the ability of a genotype to produce different phenotypes in response to environmental stimuli, conspecific specimens may be assigned to different taxa (DeWitt and Scheiner, 2004; West-Eberhard, 2005; Sommer, 2020). Moreover, populations adapted to local conditions, which are ecotypes, show specialization and geographic variation within species, responsible for generating a range of

phenotypes in response to different environmental cues (Begon et al., 2006). Ecotypes are the result of the strict interaction between genetic heritage and specific environments. The distinction between local ecotypes and polymorphic populations is not always clear and easy to identify. Molecular methods greatly improve our ability to recognize cryptic species, phenotypic plasticity and ecotypes but the results can in some cases be biased due to, for instance, incomplete sampling (in time and space) or the markers used (Vrijenhoek et al., 2009; Magoga et al., 2021). The occurrence of cryptic sibling-species, phenotypic plasticity and ecotypes may lead to significant problems in surveillance and control when morphologically similar species differ in vector capacity due to differences in their ecology, ethology and thus in the propensity to bite humans (Gildenhard et al., 2019; Francuski et al., 2019; Kareemi et al., 2021).

In this study, the geometric morphometric approach was used to evaluate differences in wing shape, size, and allometric effects among the four sibling species of the Maculipennis complex previously identified by DNA barcoding in two region of Northern Italy (Emilia-Romagna and Lombardy). The geometric morphometrics was combined with four supervised machine learning algorithms for the discrimination of two sibling species of the Maculipennis complex and the algorithms' performance was evaluated. Furthermore, this combined approach was compared with classical multivariate statistics used in geometric morphometrics. To investigate the variation in wing shape within species of the Maculipennis complex the geometric morphometric analysis was combined also with unsupervised machine learning techniques (UMAP and HDBSCAN). Our aim was to distinguish between phenotypic plasticity and ecotypes by evaluating: (1) wing shape variation among and within species; (2) the morphometric analytic support of inter group consistencies of *An. maculipennis* s. s. and *An. daciae* sp. inq. identified based on genetic information (Calzolari et al. under revision) and the variability of wing shape; and (3) the spatial and temporal distribution of different morphotypes of *An. maculipennis* s. s. and *An. daciae* sp. inq.

Materials and methods

Study area

The surveyed area included the plain areas in Emilia-Romagna and Lombardy, two densely populated regions of Northern Italy, with 14.5 million people (Figure 31). We sampled mainly in the Po Valley area of the two regions, the most suitable environment for mosquitoes, featured by vast areas of agricultural land, often interspersed with industrial-urban areas. The agricultural land is predominantly cropland with fields sometimes bordered by green strips, few and scattered trees and a dense irrigation network. Natural areas are rare, consisting mainly of river banks, characterized by riparian vegetation, or protected and re-naturalized areas. The surveyed area features a wide variety of breeding sites suitable for *Anopheles* mosquitoes, such as rice fields (e.g. Lomellina area) or the wetlands near the Po river delta, one of the largest wetland areas in Europe (Valli di Comacchio and Po River Delta).

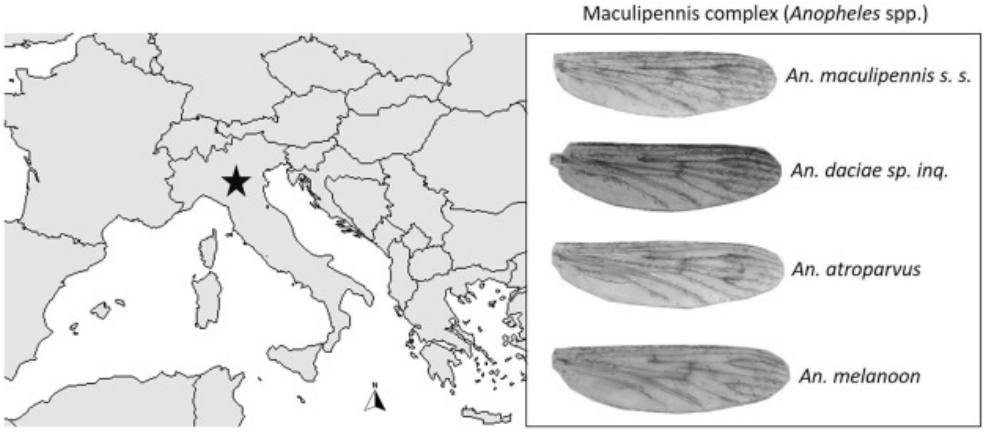

Figure 31 The study area (black star) located in the Po plain (Northern Italy) and the right wing of the four sibling species (Maculipennis complex) sampled during the surveillance campaign.

Mosquitoes sampling and genetic data generation

In 2017 and 2018, mosquitoes were collected using manual aspirators in farms or adult overwintering sites at 43 sites and using carbon dioxide-baited traps at 103 sites included in the WNV surveillance plans (Calzolari et al., 2021). Manual aspirations were performed on farms with a variety of animals (cattle, horses, goats and poultry), suitable for the collection of engorged and host-seeking mosquitoes, and in uninhabited buildings, suitable for the collection of overwintering mosquitoes. From each collected specimen, the internal transcribed spacer 2 (ITS2) was PCR amplified using as a template the DNA extracted from a single leg; the PCR amplicons were then sequenced to identify the individual species (data from Calzolari et al., 2021). To investigate the intraspecific genetic variability and its congruence with intraspecific wing shape variation, a fragment of the mitochondrial COI was PCR amplified (Calzolari et al., 2021) for a subset of randomly selected individuals belonging to *An. maculipennis* s. s. and *An. daciae* sp. inq. The COI sequences obtained were aligned using MUSCLE (Edgar, 2004) with the default parameters and then the different haplotypes were identified using R version 3.6.2 (R Core Team, 2019) and the library `haplotypes` (https://biolsystematics.wordpress.com/r/). The morphometric analyses considered only haplotypes consisting of more than five individuals. K2P (Kimura, 1980) nucleotide distances were estimated between the selected haplotypes using the library `ape` (Popescu et al., 2012), as in Magoga et al. (2018). According to the ITS2 sequences, the four following species were identified: *An. daciae* sp. inq. (322), *An. maculipennis* s. s. (124), *An. atroparvus* (10), and *An. melanoon* (4) (Calzolari et al., 2021; Bellin et al., 2021). Haplotype diversity within species was computed using the Shannon-Wiener diversity index (Shannon and Wiener, 1963).

Geometric morphometric

A subsample of 460 mosquito females was morphologically analyzed (*An. maculipennis s. s.* = 124; *An. daciae* sp. inq. = 322, *An. atroparvus* = 10, *An. melanoon* = 4). The right wing of each female was dissected

and mounted on a stereomicroscope slide with cover slip in Hoyer's medium. Each wing was photographed under 40× magnification with the software NIS-Elements V.4.0 connected to a digital camera. Each picture was rescaled to 200 pixels = 1 mm. For each rescaled wing, 19 landmarks were digitalized with the software CLIC V.8.6.3 (Figure 32) by a single operator to reduce the error due to the user effect (Dujardin et al., 2010; Garros and Dujardin, 2013).

To remove the effect of orientation, translation, and scale, the raw landmark coordinates of all 460 photos belonging to the four different sibling species were superimposed with a generalized procrustes analysis (GPA) (Bookstein, 1991; Goodal, 1991; Bookstein, 1996; Klingenberg and Marugan-Lobon, 2013; Tatsuta et al., 2018). In order to evaluate the user error level, for the two most abundant species (*An. maculipennis s. s.* and *An. daciae* sp. inq.), potential outliers were identified by procrustes distances distribution from the mean shape. The size of each wing was estimated considering the centroid size (CS) or the square root of the sum of the squared distances between all landmarks and their centroid. Due to non-normal data distribution, differences in centroid size values among species were tested with the Kruskal-Wallis test. The pairwise comparison was performed with the Wilcoxon rank sum test with p- value adjustment (Benjamini and Hochberg, 1995). In order to evaluate differences in shapes among species and to quantify the allometric effect or the relationship between shape and size, a multivariate analysis of covariance (MANCOVA) was performed (Klingenberg, 2016). In MANCOVA, procrustes coordinates obtained from GPA (landmark coordinates or shape variables) were considered as a dependent variable while the logarithm of CS and species were considered as independent variables. The interaction between the logarithm of CS and species was also considered. Regression coefficients were tested with a permutation procedure (n = 1000) (Anderson, 2001). Differences in shape between species were evaluated with a post hoc test based on Euclidean distances with a permutation test (n = 1000). In order to identify shape patterns for the four sibling species, the morphological space was visualized with principal component analysis (PCA). All the described analyses were performed with the R packages `FactoMineR` and `geomorph` (Le et al., 2008; Adams et al., 2020).

Classification by supervised machine learning approach

Different algorithms were trained to recognize *Anopheles'* sibling species: support vector machine (SVM), random forest (RF), fully connected neural network (ANN) and an ensemble model (Lek and Gu´egan, 1999; Crisci et al., 2012). The dataset consisted of unbalanced classes with a low number of individuals of the species *An. atroparvus* (n = 10) and *An. melanoon* (n = 4). For this reason, only two species, *An. maculipennis s. s.* and *An. daciae* sp. inq., were considered for this analysis. The landmarks (19 × 2 = 38 coordinates) obtained by GPA were used as predictors and the dataset was split into two sets: - the training set composed of the coordinates of 100 individuals of *An. maculipennis s. s.* and 100 individuals of *An. daciae* sp. inq.*;* - the testing set with the coordinates of the remaining individuals: 24 of *An. maculipennis s. s.* and 222 of *An. daciae* sp. inq. In order to train the machine learning algorithms, a binary code to the two

species was assigned: the most abundant species in the dataset, *An. daciae* sp. inq., was coded as zero, while the less abundant species, *An. maculipennis s. s.,* was coded as one. The support vector machine (SVM) builds a decision boundary (hyperplane) in the multidimensional space of the input data (landmarks) to maximize the distances between two different classes (species) (Noble, 2006). The radial basis function of SVM was used as kernel: it augments the space dimensions in order to produce a better separation of different species. During the training phase, a grid search method was used. The algorithm considered a combination of two different hyperparameter values: the cost (C) and sigma. Each combination was tuned with 10-fold cross validation with validation split equal

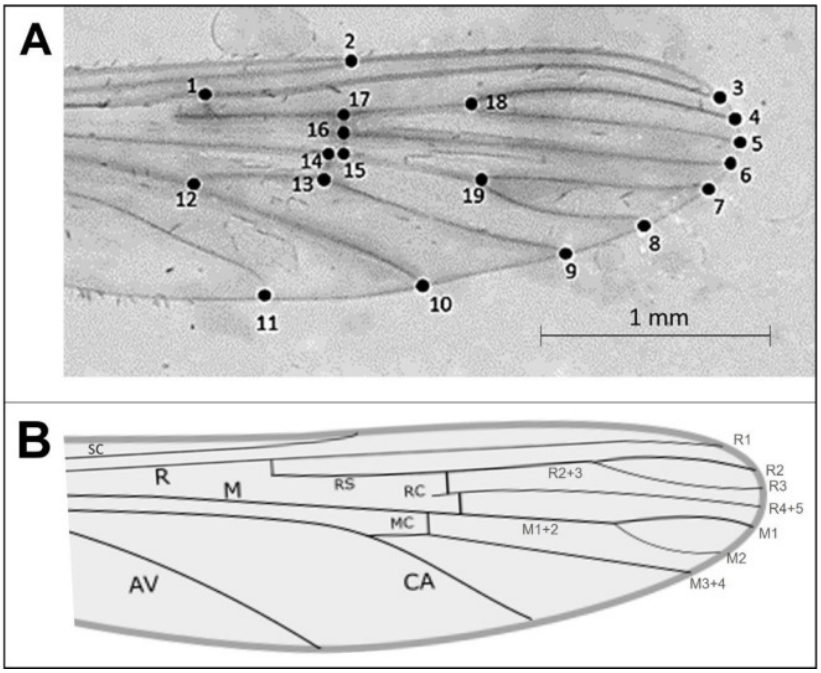

Figure 32 In panel A right wing of an *An. daciae* sp. inq. female with the 19 landmarks. In panel B right wing representation with depicted the principal veins: subcostal (SC), radius (R), radius 1 (R1), radius 2 (R2), radius 3 (R3), radius 2 + 3 (R2+3), radius 4 + 5 (R4+5), radial sector (RS), radiomedial (RC), media (M), media 1 (M1), media 2 (M2), media 2 + 3 (M2+3), media 3 + 4 (M3+4) mediocubital (MC), cubitus anterior (CA), anal (AV).

to 25%. The best combination was selected based on the highest mean validation accuracy over the 10 folds. The random forest (RF) algorithm consists of a forest of 2000 unpruned classification trees (Breiman, 2001). Each tree built a set of rules to discriminate the classes (species) based on a set of features (landmarks) randomly selected. Input data are processed, and each tree made a prediction: the species with the majority of votes represented the selected prediction. During the training phase, a grid search method was used. The algorithm considered the combination of the number of settled trees (2000) (Oshiro et al., 2012) and the maximum number of features extracted randomly from the data set (mtry). Each combination was tuned with 10-fold cross validation with validation split equal to 25%. The best combination was selected based on the highest mean validation accuracy over the 10 folds. Artificial neural network (ANN) is a deep learning algorithm that consists of an input layer, a variable number of hidden layers, and an output layer (Recknagel,

2001). In each layer type, neurons are the base computing elements and are interconnected with synapses (weights) that are optimized during the training phase. The back-propagation algorithm iteratively adjusts the strength of the interconnections to minimize the output error. The artificial neural network was trained using three different architectures (Figure 33). For each architecture, the Adam optimizer, the activation function ReLU in each neuron of the hidden layers, and the sigmoid activation function of the neuron in the output layer were used (Aggarwal, 2018). The output error was computed using the binary cross entropy equation (Murphy, 2012). In order to reduce the overfitting, the loss of generalization of the model during the training phase, two regularization procedures, the dropout and the batch normalization, were used (Figure 33) (Baldi and Sadowski, 2014; Aggarwal, 2018). To select the best ANN architecture, 10-fold cross validation with 500 epochs, a batch size of 32, and a validation split equal to 25% were used. The best model was selected based on the highest mean validation accuracy over the 10 folds. Support vector machine (SVM), random forest (RF) and the artificial neural network (ANN) that achieved the highest validation accuracy were selected and combined to build an ensemble model (EM) (Dietterich, 2000). In order to quantify the performance of each algorithm, SVM, RF, ANN and EM were compared. We used the SVM with linear kernel not tuned as a benchmark. For all algorithms, we computed four different metrics for unbalanced classes on the test set:

- specificity: the correct number of individuals classified as *An. maculipennis s. s.* on the total number of specimens of *An. maculipennis s. s.;*

- sensitivity: the correct number of individuals classified as *An. daciae* sp. inq. on the total number of specimens of *An. daciae* sp. inq.; - G-mean: the geometric mean between specificity and sensitivity (Ri and Kim, 2020);

- balanced classification accuracy: the arithmetic mean between specificity and sensitivity.

The final evaluation of each algorithm relied on the estimation of the receiving operators curve (ROC) and the precision-recall curve (PRC). The PRC curve shows the algorithm performance as precision (true positive / (true positive + false positive)) and recall (true positive / (true positive + false negative)), with a different output probabilistic threshold that varies over a range of different values. In our binary codification a true positive is a correct classification of an *An. maculipennis s. s.* female, a false positive is an uncorrected classification of an *An. daciae* sp. inq. Female and a false negative is an uncorrected classification of an *An. maculipennis s. s.* female. The PRC-AUC quantified the algorithm performance in consideration of the unbalanced dataset when the individuals within species are not equiripartite (Saito and Rehmsmeier, 2015; Sofaer et al., 2019). A random classifier has a PRC-AUC value equal to the proportion of individuals belonging to the less abundant class. The algorithm that showed the best performance considering all metrics was selected and the importance of each landmark for species classification was evaluated by recomputing a ROC-AUC for each procrustes coordinate. The importance of each coordinate scaled between 0 and 100 was qualitatively compared by the superimposition of the mean shape of each species visualized with wireframe graphs. A further comparison between the best machine learning algorithm selected and a classical multivariate method used in geometric morphometrics, the linear discriminant analysis (DA), was performed

(Viscosi and Cardini, 2011). DA was 10-fold cross-validated on the training set and the four metrics for unbalanced classes (see above) were computed on the test set. The SVMs and RF algorithms, the procrustes coordinate importance, and DA were computed by the `caret` package (Kuhn, 2008). The artificial neural network (ANN) was computed by the `keras` package (Allaire and Chollet, 2020) and `tensorflow` package (Allaire and Tang, 2020). The ROC-AUC and PRC-AUC were computed with the R package `precrec` (Saito and Rehmsmeier, 2017).

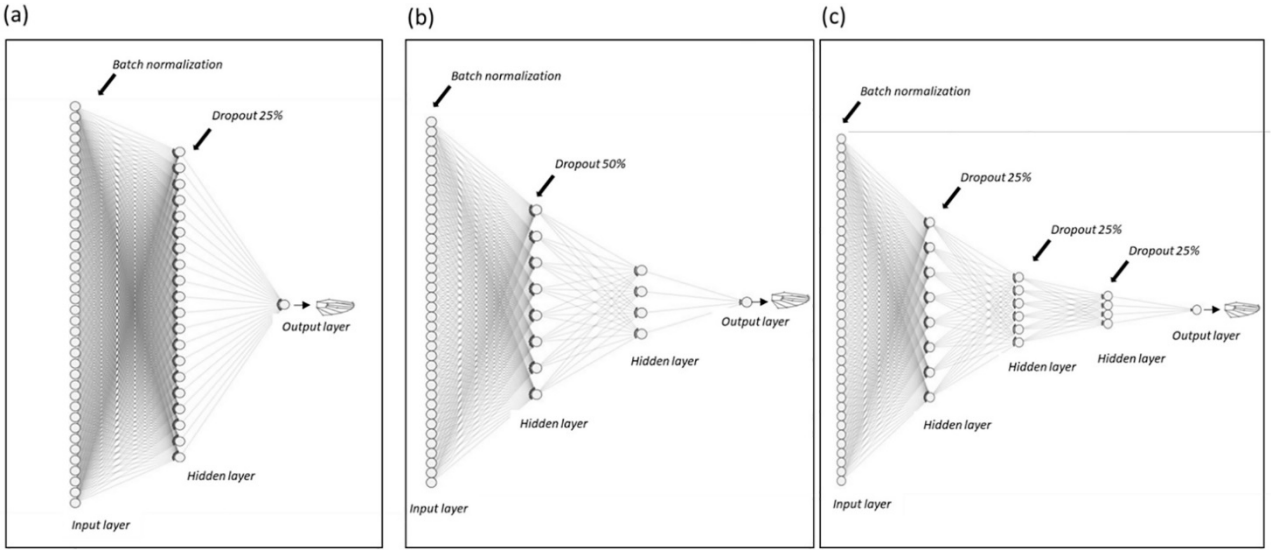

Figure 33 Three different architectures of the artificial neural networks designed for the study: (a) one hidden layer with 20 neurons; (b) two hidden layers with 8 and 4 neurons, respectively; (c) three hidden layers with 8, 6 and 4 neurons, respectively. In each panel, arrows referred to the regularization method used: dropout and batch normalization (see text).

Inter-specific diversity of wing shape in embedding space (UMAP)

The data set with procruster coordinates was further processed with a dimensionality reduction algorithm (UMAP) (McInnes et al., 2018) that reduced the dimension of the dataset from 38 dimensions (19 pairs of landmarks coordinates, x and y) to two dimensions (UMAP 1 and UMAP 2). The UMAP algorithm is driven by two important hyperparameters: the first is the number of neighbors, which evaluates how the algorithm balances the local versus global structure of the data. Low neighbor values force UMAP to capture local structure, while high values capture global structures, losing finer and local relationships. To obtain a good global interspecific representation of the shape of the Maculipennis complex's, a neighbor value of 70 was set. The second hyperparameter is the minimum distance among neighbors: this evaluated how tightly similar points are grouped in the embedding. Low values result in clumpier embedding. To highlight differences among groups of species a distance value of 0 was set. The species identity information obtained from DNA barcoding analysis was superimposed on UMAP. To get a general idea of the wing shape variation represented by UMAP embedding, an inverse transformation approach was used. A convex hull that encompasses the embedding space was drawn and a grid of 13 points equispaced in the convex hull area

were sampled. The sampled points were inverse transformed to obtain the representation of 13 wing shapes. Using wireframe graphs, the sub sample of wing shapes obtained by the inverse transformation was superimposed and compared with the mean shape of the Maculipennis complex. The Generalized Procrustes Analysis (GPA) was performed with the R package `geomorph` (Adams et al., 2020). UMAP analysis was performed in the Python `umap` library (https://github. com/lmcinnes/umap).

Intra-specific diversity of wing shape in *An. maculipennis s.s.* and *An. daciae* sp. inq.

To find possible relationship between shapes and genetic markers and to capture local representations of the data, UMAP embedding was computed for *An. maculipennis* s. s. and *An. daciae* sp. inq. For this purpose, the first hyperparameter (number of neighbors) was set to a value of 15 while the second hyperparameter (minimum distance) was maintained at 0. For each species, the genetic information of the COI groups obtained from DNA barcoding analysis was superimposed on each UMAP embedding. To test differences in wing shapes of individuals belonging to the two different groups identified based on the COI gene tree, a PERMANOVA test was performed on shapes coordinates with 999 permutations (Anderson, 2001). The correlation between K2P nucleotide distance (Kimura, 1980) among the identified haplotypes and the mean wing shape of each haplotype was computed by Mantel test using Spearman's method and 999 permutations. To identify different intraspecific wing morphotypes, an unsupervised clustering method was used that considered hierarchical estimates, Hierarchical Density-Based Spatial Clustering (HDBSCAN) algorithm (Campello et al., 2013). Clustering organized the data into a finite set of categories. In the density-based clustering paradigm, clusters are defined as dense areas separated by sparse regions. HDBSCAN outperforms others density clustering algorithms as it separates points that belong to clusters with outliers. The algorithm also assigns a soft partition value expressed as probability; for each observation, the probability is proportional to its membership (probability of belonging a particular cluster). HDBSCAN is driven by two main hyper parameters: min_cluster_size and min_samples. To find the best combination of HDBSCAN hyper parameters, a randomized grid search procedure was used. Along the two hyperparameters ranges, different values were randomly sampled. The best couple was selected according to the maximization of a validity measure (DBCV) proposed by a clustering density approach (Moulavi et al., 2014). HDBSCAN clustering analysis was performed in the Python `hdbscan` library (https://github.com/sciki t-learn-contrib/hdbscan) and scikit-learn library (https://scikit-learn.or g/stable/about.html). For both *Anopheles* species, the mean shape of different morphotypes was compared by PERMANOVA test with 999 permutations and considering the residual randomization. The mean shapes of different morphotypes were compared by a pairwise post-hoc test based on Euclidean distance. To visualize landmark pattern variation among intraspecific morphotypes, the cluster's mean shape was superimposed on the mean shape of each taxon by wireframe graph. To test intraspecific spatial-temporal differences in morphotype abundance, a GLMM model was used with Poisson family function for count data and with a maximum likelihood method (Laplace approximation). The model accounted for random and fixed effects. The random effects included a nested temporal structure for sampling dates (day in month) and a nested spatial structure for sampling sites

(locality in province). The estimated intercept varied between crossed random effects (month and province). The fixed effect included the type of trap used to capture the specimens (CO2 or manual), the morphotype and the interaction among factors. GLMM models were computed by R package `lme4` (Bates et al., 2015).

Results

The difference in centroid size (CS) among species was significant (χ^2 = 1.17, p = 0.01). The pairwise comparison between species showed that the wings of *An. maculipenniss s. s.* and *An. daciae* sp. inq. Were significantly larger than the wing of *An. atroparvus* (Figure 34 and Table 6SM). The user error level for the two most abundant species was 3.2% for *An. maculipennis s. s*. and 1.5% for *An. daciae* sp. inq. The allometric relationship between log(CS) and shape variables was statistically significant (F = 9.21, p-value = 0.001) (Table 7SM and Figure 3SM). The shape variables were different among species (F = 14.89, p-value = 0.001). The interaction between log(CS) and species was not significant (F = 1.09, p-value = 0.32). Pairwise differences in shape between species were statistically significant for all pairs (Table 8SM). By principal component analysis (PCA), the two first principal components explained only 33% of the total variance (PC1 = 19.2 and PC2 = 13.8) (Figure 35).

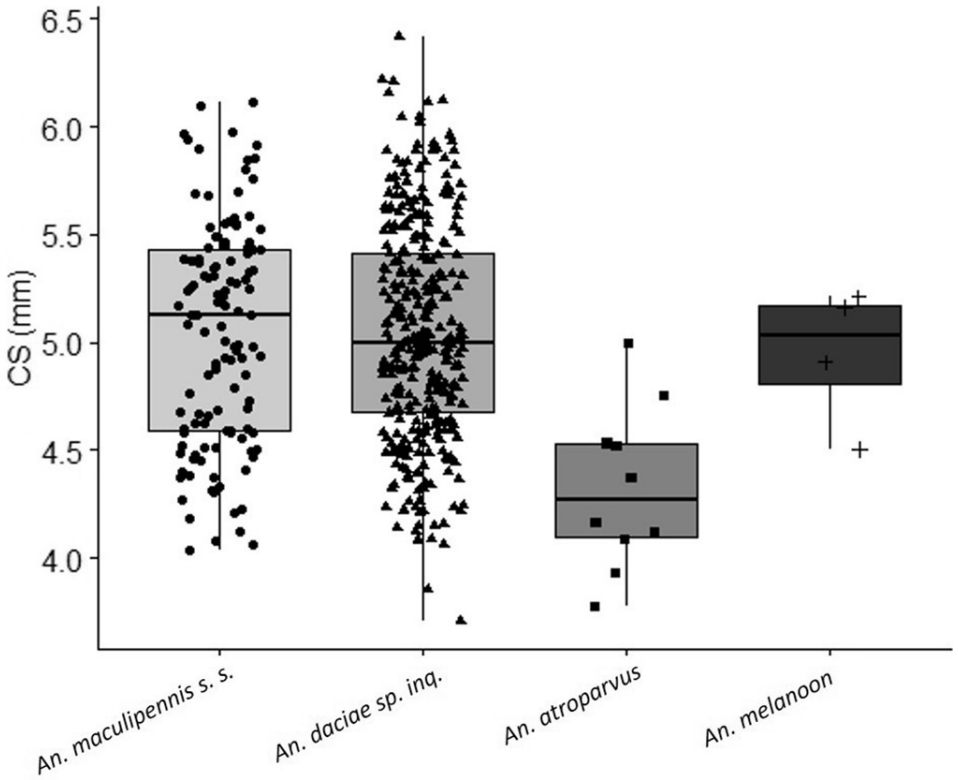

Figure 34 For each species the distributions of the centroid size were reported.

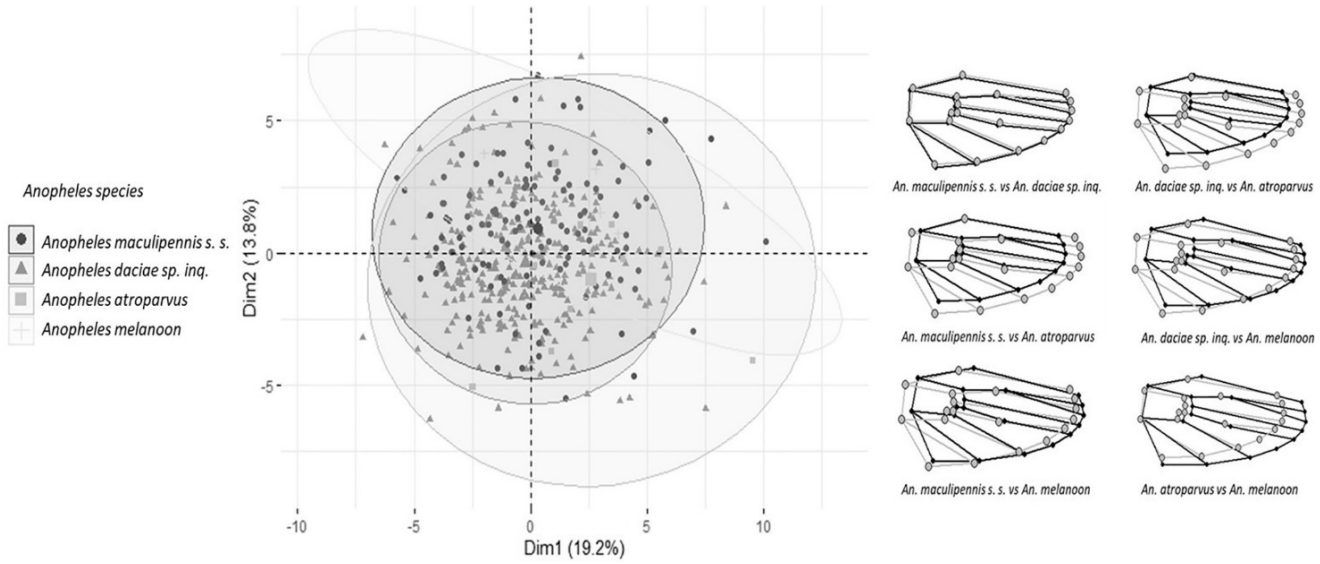

Figure 35 Results of PCA are reported. For each species, the ellipses were built with a confidence level of 95%. On the right side, the mean shape of each pair of species superimposed after GPA are reported. The wing shapes are represented as wireframe graphs with two different colors: first species (grey) vs second species (black).

Classification by supervised machine learning approach

For the classification of *An. daciae* sp. inq. and *An. maculipennis* s. s., the SVM algorithm reached the highest mean validation accuracy of 80% for C = 5 and sigma = 0.02 (Figure 4SM). The RF algorithm showed the highest mean validation accuracy of 77% for mtry = 11. The best ANN architecture was obtained with one hidden layer and 20 neurons, and it reached a mean validation accuracy of 80% (Figure 4SM). The highest sensitivity was recorded for SVM (83%) and the lowest for RF (62%), while the highest specificity was recorded for EM (84%) and the lowest for SVM (79%) (Figure 36). The highest values of G-mean and balanced classification accuracy were recorded for SVM (81%). Considering all metrics, the RF algorithm showed lower performances than the benchmark. SVM's specificity indicated that 83% of *An. maculipennis s. s.* individuals were correctly classified as *An. maculipennis s. s.* and SVM's sensitivity indicated that 79% of *An. daciae* sp. inq. Individuals were correctly classified as *An. daciae* sp. inq. All algorithms performed better than a random classifier (ROC-AUC > 0.50) with the highest value of ROC-AUC shown by SVM (0.81) (Figure 37). The highest PRC-AUC was recorded for EM (0.29). SVM showed the highest sensitivity, balanced classification accuracy, G-mean, and ROC-AUC value; moreover, it showed the lowest difference between specificity and sensitivity (4%) (Figure 36). The SVM algorithm performed better than the classical multivariate method DA in all metrics but specificity (Table 6). 11 x was the most important procrustes coordinate for the classification of *An. daciae* sp. inq. and *An. maculipennis* s. s. (Figure 38).

Table 6 Metrics' values computed on the test set for two different classificators: Support Vector Machine with radial basis function (SVM) and linear discriminant analysis (DA). The reported values were expressed as percentages.

Method	Sensitivity	Specificity	Balanced classification accuracy	G-mean
SVM	83	79	81	81
DA	70	82	76	75

For comparison, differences in shape between *An. daciae* sp. inq. and *An. maculipennis* s. s. were evaluated by the three first principal components (PCA) considering all procrustes coordinates and by a three-dimensional plot of the three most important procrustes coordinates (11×, 16× and 15×) (Figure 39). The three first principal components explained only 41% of the total variance (PC1 = 19, PC2 = 14 and PC3 = 8) (Figure 39 (a)). The pattern in the morphological space revealed a less clear differentiation between the two species than the three dimensional space of the most important procrustes coordinates (Figure 39 (a)).

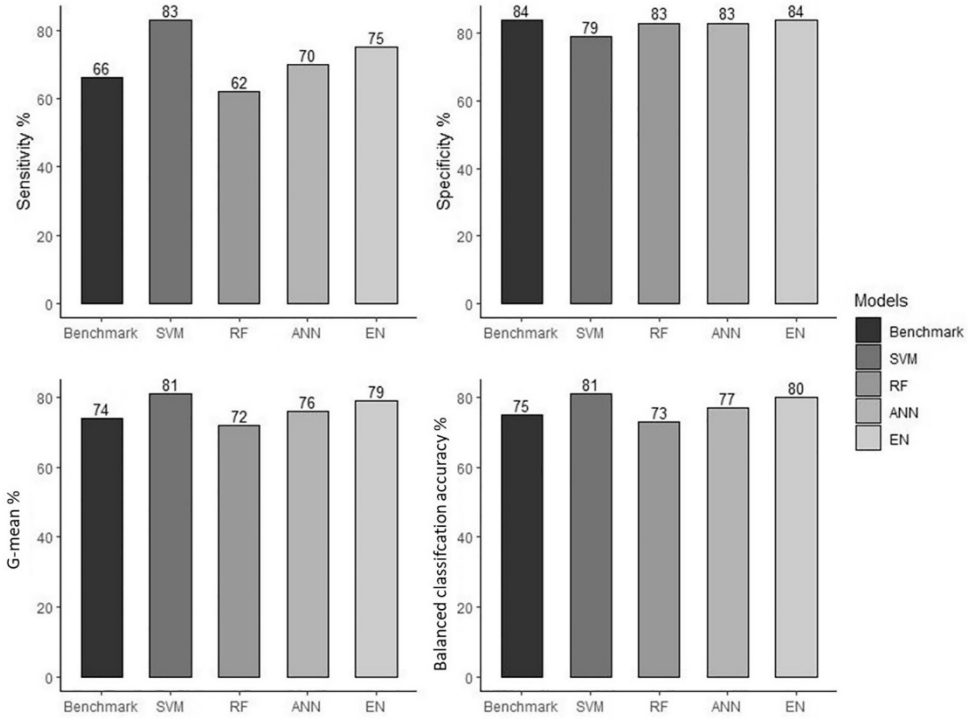

Figure 36 Metrics recorded after the evaluation procedure on the test set. The benchmark was relative to the support vector machine with linear kernel without tuning procedure.

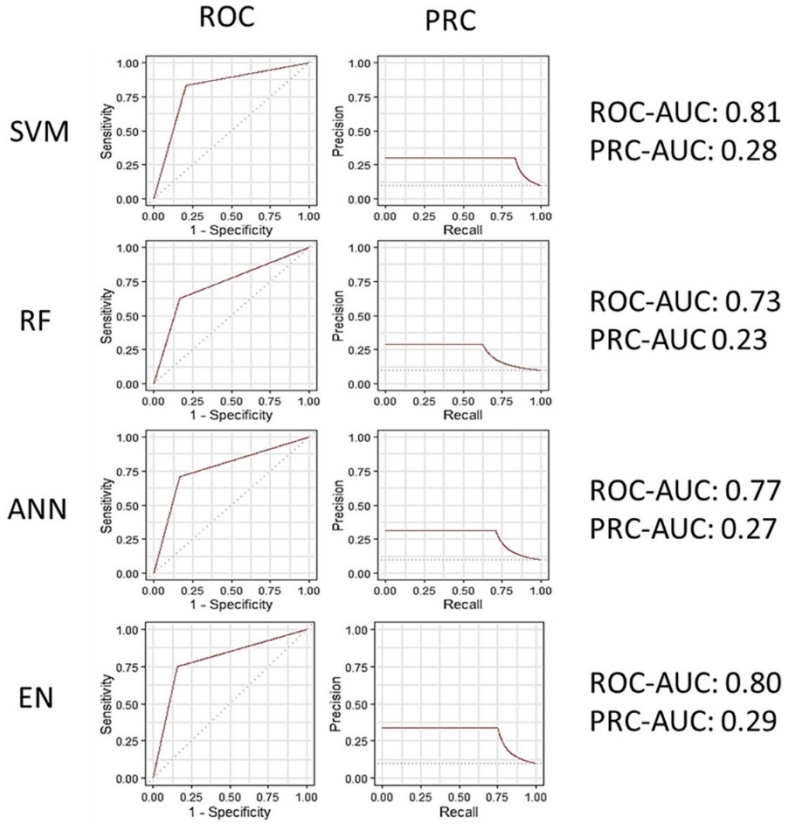

Figure 37 For each model receiving operator curves, ROC and the precision-recall curve (PRC) are reported as a continuous line; the dashed line represents the random classifier of ROC and PRC. For each model, on the right side, the value of the area under the ROC and PRC curves (AUC) are reported.

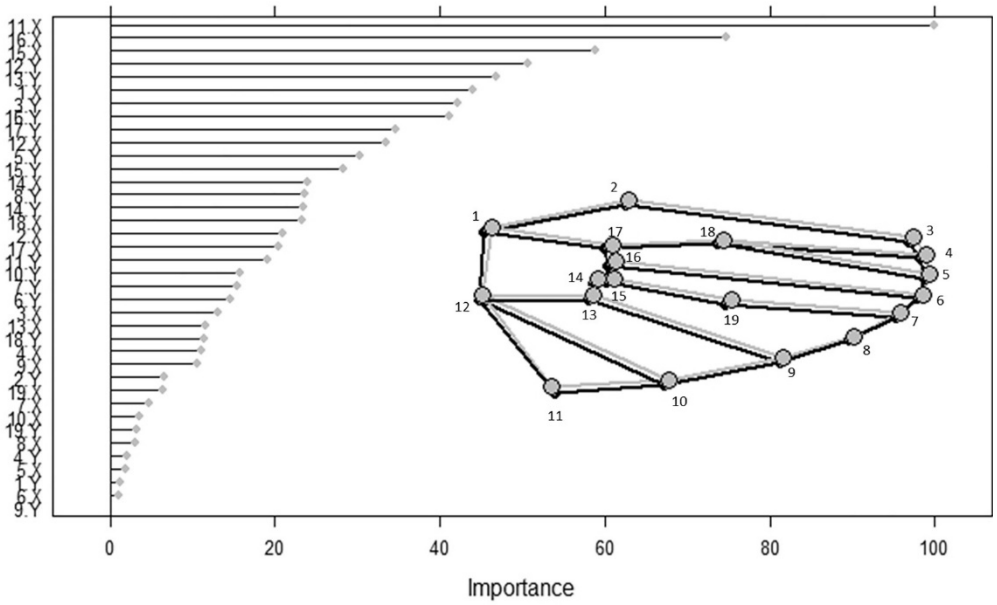

Figure 38 For each landmark (1–19), the percentage of importance of x and y coordinates in the wing classification is reported (see Material and Methods). For a comparison and the landmark coordinates legend, the wing shapes of *An. maculipennis* s. s. (grey) vs *An. daciae* sp. inq. (black) are represented as wireframe graphs after GPA.

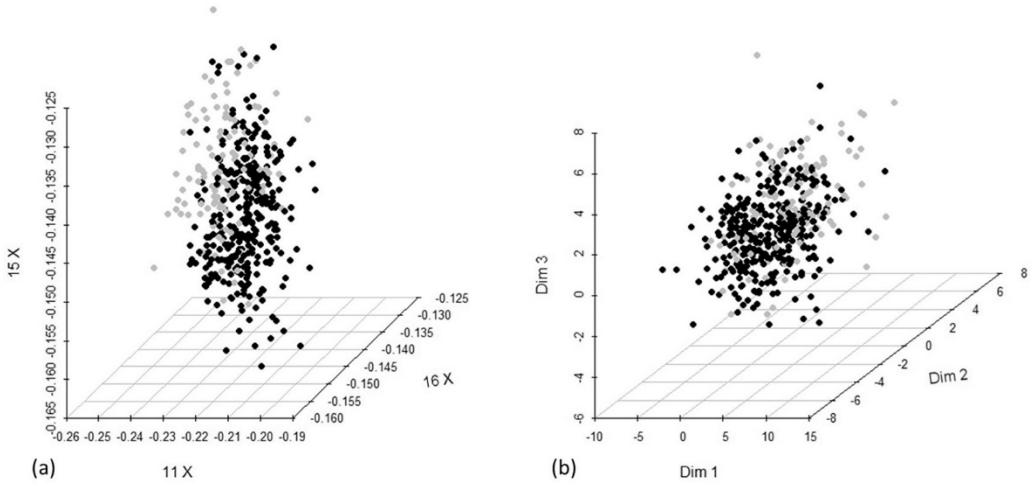

Figure 39 The three most important procrustes coordinates (a) compared with the first three principal components computed by PCA (b) (variance explained 41%). The individuals of each species were reported with two different colors: *An. daciae* sp. inq. (black) and *An. maculipennis* s. s. (grey).

Inter-specific diversity in embedding space (UMAP)

In the embedding space generated by UMAP, *An. melanoon* clustered separately from *An. atroparvus* specimens along UMAP 2: the first at the bottom and the second at the top (Figure 40 panel A). Most *An. maculipennis* s. s. specimens were arranged in the top-left of the plot, with other specimens spread along the maps. *An. daciae* sp. inq. specimens showed the highest dispersion, with a major concentration of specimens in the bottom-right region of the maps. The wireframe graphs obtained by UMAP inverse transformation showed the 13 main pattern variations in wing shape among species within the Maculipennis complex (Figure 40 panel B). Differences in *An. atroparvus* and *An. melanoon* were clearly shown by the distance between the centroids (Figure 40 panel A) and the wing shapes 3 and 13, respectively (Figure 40 panel B). The pattern of differentiation is less clear between *An. daciae* sp. inq. and *An. maculipennis* s. s. considering the centroids and, most of all, the continuum of different shapes.

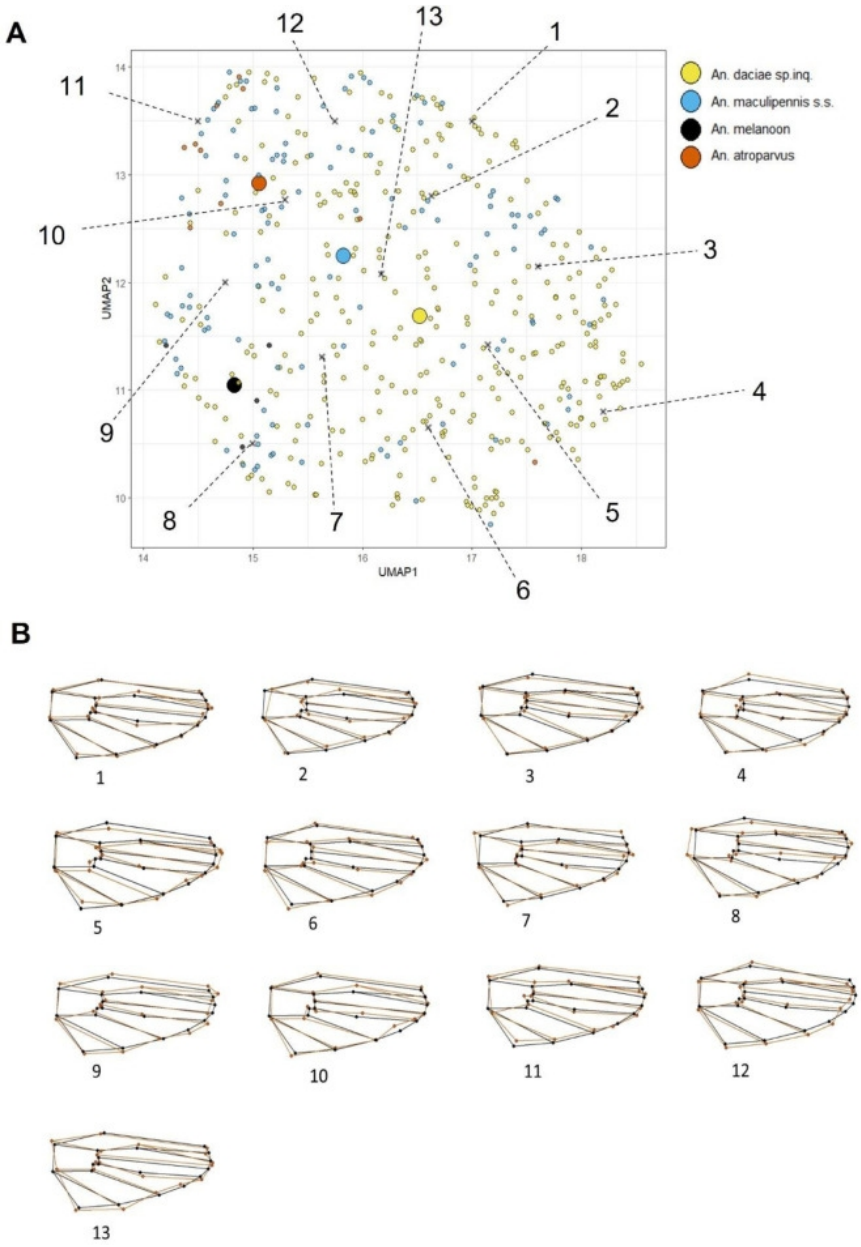

Figure 40 A. UMAP embedding space relative to the four species wing shapes (*An. daciae* sp. inq., *An. maculipennis* s. s., *An. melanoon* and *An. atroparvus*) of the Maculipennis complex; the species centroid (greater size points) and the position of equispaced sampled points in the convex hull were reported; dashed lines and numbers indicate the wing shape of the complex reported in panel B. **B.** wings shapes obtained by inverse transformation (black color) superimposed on the mean shape of the complex (red color).

Intra-specific diversity of *An. daciae* sp. inq. and *An. maculipennis* s. s.

The 166 *An. daciae* sp. inq. and 80 *An. maculipennis* s. s. individuals analyzed in this study were found to belong to 77 and 45 haplotypes, respectively, based on COI gene (Table 7). The three most abundant haplotypes were identified within *An. daciae* sp. inq. COI group A (Calzolari et al. under revision), namely h2, h27 and h11, which included 23, 14, 10 individuals, respectively. Within *An. daciae* sp. inq. COI group B, the two most abundant haplotypes were h20 and h15, including 7 and 6 individuals, respectively. For *An.*

maculipennis s. s. COI group 1 (Calzolari et al. under revision) only one abundant haplotype (h1) was detected, including 8 individuals, while in COI group 2, no haplotype including more than 5 individuals was found. Haplotype diversity (Shannon-Wiener index) was very similar in *An. daciae* sp. inq. (3.80) and *in An. maculipennis* s. s. (3.59). For both the most abundant taxon (*An. daciae* sp. inq. and *An. maculipennis* s. s.), the superimposition of COI groups and haplotypes did not support the correlation between COI information and shape ordination (Figure 41). This result was confirmed by PERMANOVA test (Tables 8 and 9SM) and the correlation between haplotype COI nucleotide distance and haplotype shape difference was not significant (Mantel statistic: 0.049 and *p*-value: 0.44).

Table 7 Summary of statistics for the analyzed *An. daciae* sp. inq. and *An. maculipennis* s. s. (a: COI group identified by Calzolari et al. (under review), b: number of identified haplotypes, c: number of individuals)

Species	**COI group**[a]	**Haplotypes**[b]	**N**[c]
An. daciae sp. inq.	A	46	107
	B	31	59
An. maculipennis s. s.	1	25	47
	2	20	33

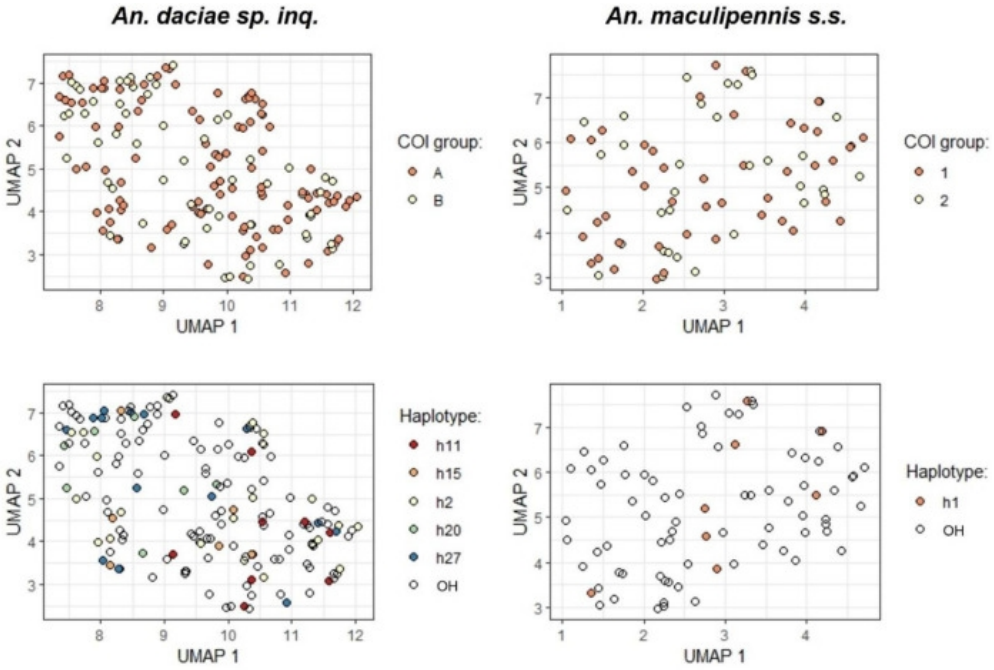

Figure 41 Superimposition of the COI genetic groups (top panels) and the COI haplotypes (low panels) on the embedding space of UMAP. In low panels, OH (others) referred to haplotypes including less than six individuals.

Table 8 PERMANOVA test of shape differences among genetic COI groups with 999 random permutations. For each taxon, the degree of freedom (Df), the sum of squares (SS), the means squared error (MS), the coefficient of determination of the test (Rsq), the F statistic, the effect sizes (Z) and the p-value of the test were reported.

***An. daciae* sp. inq.**		**Df**	**SS**	**MS**	**Rsq**	**F**	**Z**	***p*-value**
	COI groups	1	0.0014	0.0014	0.0090	1.4	1.1	0.13
	Residual	164	0.16	0.0009	0.99			
	Total	165	0.16					
***An. maculipennis* s. s.**	COI groups	1	0.00070	0.00070	0.0084	0.6	−0.73	0.75
	Residual	78	0.083	0.0010	0.99			
	Total	79	0.083					

The best set of HDBSCAN's hyperparameters was min_cluster_size = 5 and min_samples = 5 for both *An. daciae* sp. inq. and *An. maculipennis* s. s. with a cluster validity metric (DBCV) values of 0.12 and 0.20, respectively. HDBSCAN identifies 12 morphotypes for *An. daciae* sp. inq. and 4 morphotypes for *An. maculipennis* s. s. (Figure 42).

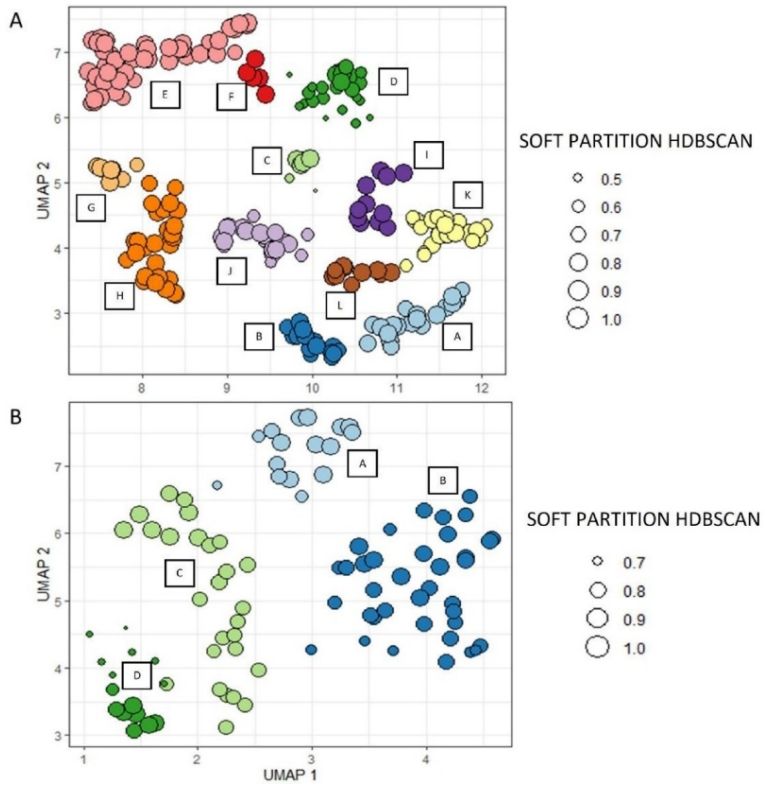

Figure 42 Panel A (*An. daciae* sp. inq.) and panel B (*An. maculipennis* s. s.) reported the result of HDBSCAN algorithm. The colors represented the morphological clusters identified by the algorithm. The size of each point was proportional to the membership values estimated by HDBSCAN. Outliers that are not assigned to the clusters were removed and considered as noise.

In both taxa, the PERMANOVA test revealed a significant difference in wing shape among morphotypes. All pairwise comparisons between morphotypes were significant (Table 10SM).

In *An. daciae* sp. inq., the most frequent pattern of variation from the mean shape involved landmark 2 (in 7 out 12 clusters), located in the subcostal vein, and landmark 11, located in the anal vein (in 7 out of 12 clusters) (Figures 43). Other frequent patterns of variation observed involved the radio medial cross veins (landmarks 15, 16 and 17) and the cubitus veins (landmark 10) (Figures 43). Cluster C did not diverge from the mean shape of the species. The highest difference in variation was observed for cluster D. In *An. maculipennis* s. s., the most frequent pattern of variation from the mean shape involved landmark 2, 10, 11 and 16 (in 2 out of 4 clusters). Another pattern of variation was observed in the radio medial cross veins (landmarks 15 and 17), in the medio cubital cross vein (landmark 12) and in the bifurcation of the radius vein (R_{2+3}) (landmark 18). The landmarks from 4 to 8, located in the marginal part of the wings, did not change in any clusters for both taxa. Considering the capture techniques, morphotype abundance did not show significant intraspecific spatial-temporal differences (Table 11SM).

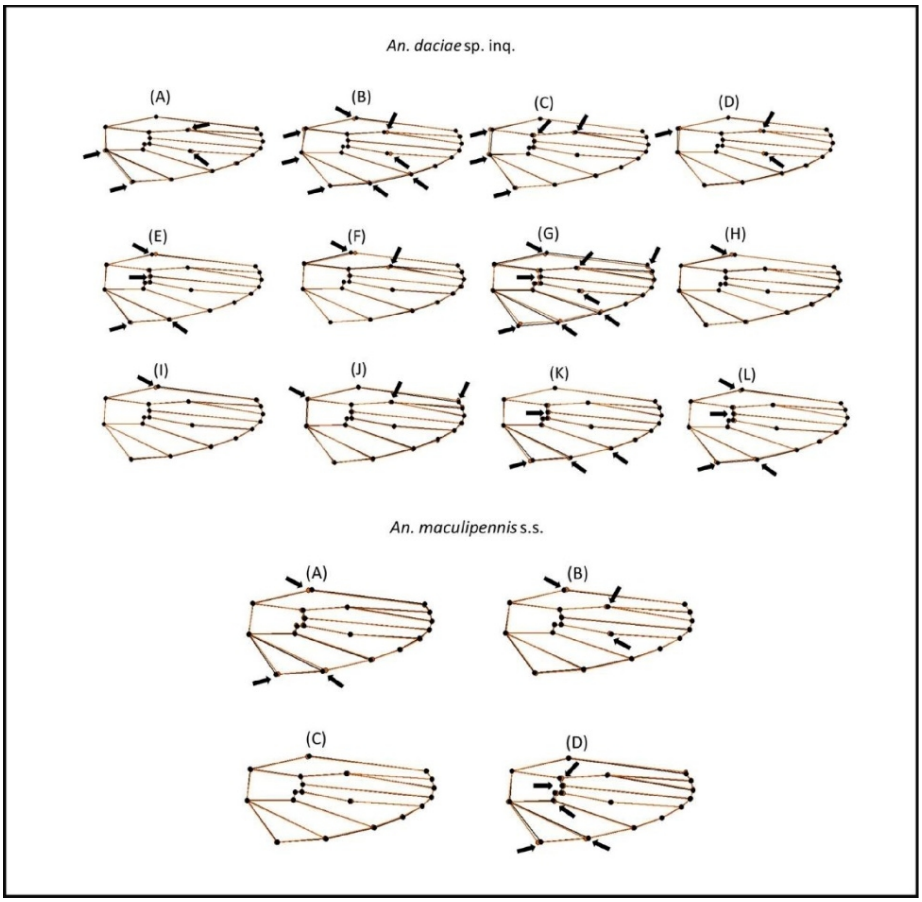

Figure 43 For each species the mean wing shapes obtained by HDBSCAN clustering (black) was superimposed on the mean shape of the species (red). For each wing, the number referred to the cluster reported in Figure 42. The arrows indicated the main landmarks variation with the respect to the mean shape of the species.

Discussion

A combined machine learning (both supervised and unsupervised) and geometric morphometric approach was useful to investigate shape variation when taxa are difficult to discern using standard taxonomic methods (Lorenz et al., 2012, 2015a, 2015b; Wilke et al., 2016). In this study, the first two PCs explained only 33% of the total variance and appeared not very useful for discriminating among all species of the complex as well as *An. maculipennis* s. s. from *An. daciae* sp. inq. The correct classification of 83% *An. maculipennis* s. s. and 79% of *An. daciae* sp. inq. was obtained by the integration of geometric morphometric analysis and a supervised machine learning algorithm that reach the highest performance (support-vector machine). In order to investigate the relationship between wing shape and genetic markers, and to capture intraspecific differentiation of the four sibling species of the Maculipennis complex, we combined the geometric morphometric with the unsupervised machine learning algorithms UMAP and HDBSCAN. The use of machine learning improved the geometric morphometrics framework and allowed to describe and recognize variability patterns among and within sibling species. In the analysis of shape, especially in sibling species, the combined approach of UMAP and geometric morphometrics is unusual. Unlike PCA and most common eigenvector analysis, UMAP was able to capture data nonlinearity (Yang et al., 2021). Then, UMAP, an unsupervised machine learning algorithm, allowed us to describe the wing shape variation patterns among the four sibling species of Maculipennis complex, namely *An. atroparvus, An. melanoon, An. maculipennis* s. s.*,* and *An. daciae* sp. inq. In addition, it mapped the morphological variation within species. UMAP dimensionality reduction did not allow a clear distinction between morphotypes of *An. maculipennis* s. s. and *An. daciae* sp. inq. and confirmed that several specimens of both taxa were not completely split in the UMAP embedding. However, the centroids position might indicate evolutionary trajectories that have differentiated the species. UMAP algorithm was used in *Saccharomyces cerevisiae* to identify groups of genes related to protein structures, protein complexes and pathways (Dorrity et al., 2020) and to find fine-scale relationships and cryptic structures in the geography, genotypes and phenotypes in human populations (Diaz-Papkovich et al., 2019). This procedure should be tested in other complex or cryptic species to verify its effectiveness and generalizability. In this study, UMAP allowed us to describe the occurrence of discontinuous wing shape morphotypes in the four analyzed species and highlighted the great inter and intra specific variability of the Maculipennis complex. COI mtDNA region is often used as barcoding region for species identification but also for a first assessment of the genetic population structure (e.g., Brunetti et al., 2019; Zheng et al., 2019; Doorenweerd et al., 2020). Due to the intraspecific variability of the COI we found in *An. daciae* sp. inq and in *An. maculipennis* s. s., it was interesting to investigate the morphological variability of COI haplotypes also considering that this mitochondrial marker is mostly used, sometimes in association with others, in integrated taxonomic studies. Within the two most abundant taxa (*An. daciae* sp. inq. and *An. maculipennis* s. s.)*,* two different groups and several haplotypes were described based on COI sequences. The number of haplotypes is higher in *An. daciae* sp. inq. than in *An. maculipennis* s. s. but the diversity index is very similar in the two species. However, UMAP ordination and statistical tests indicated that the correlation

between COI variation and shape ordination/variation was not significant. This result is not surprising because as well as other possible factors, wing shape is a multigenic trait with high hereditability and selective pressures acting on the underline genes may be different from those of the COI (Gilchrist and Partridge, 2001; Hoffmann and Shirriffs, 2002; Moraes et al., 2004; Patterson and Klingenberg, 2007; Henry et al., 2010). Morphological variation described within haplotype gives interesting results in the framework of phenotypic plasticity, i.e. the ability of a genotype to produce different phenotypes in response to stimuli or inputs from the environment (DeWitt and Scheiner, 2004; West-Eberhard, 2005; Sommer, 2020). Phenotypic plasticity may account for population responses to rapid environmental change or fluctuation and to adaptive tracking on an ecological time scale (Rudman et al., 2022). Within species, regardless of haplotypes, the HDBSCAN unsupervised ML algorithm clustered different morphotypes: 12 in *An. daciae* sp. inq. and 4 in *An. maculipennis* s. s. Each morphotype shared a similar pattern of variation in the subcostal vein, in the anal vein and in the radio medial cross veins of the wing. Interestingly, in the two species *An. daciae* sp. inq. and *An. maculipennis* s. s., there were several similar morphotypes and patterns of variation. At the same time, in the marginal part of the wings, no variation was detected in both species. According to our previous results (Bellin et al., 2021), two coordinates relative to variation in radio medial cross veins (landmarks 15 and 16; Figure 2.1.2), are important in the discrimination between sibling species (Severini et al., 2009; Becker et al., 2010). In many species of Culicidae, landmarks located on the center of the wing showed higher variability (Beriotto et al., 2021). In contrast, landmarks with lower variability were found on the margin of the wing, suggesting that landmarks with aerodynamic restrictions are evolutionarily preserved (Bomphrey et al., 2017). Interestingly, several morphotypes, pattern of variation and morphological stasis were similar in the two species. Morphological might be related to various functional roles and responses to selective pressures, or different ontogenetic processes (Zelditch et al., 2006; Aytekin et al., 2007; De Morais et al., 2010). The stasis or variability of a landmark is probably regulated by phylogenetic and functional constraints. As in other insects, the mosquito wings are complex three-dimensional structures that are mainly evolved for locomotion but have several functions under selective pressures (Krishna et al., 2020). The structure and architecture of the veins are crucial for the biomechanical properties of the wings and determine wing deformation during flight (Combes and Daniel, 2003; Appel et al., 2015; Rajabi et al., 2016a; Sun et al., 2021). The veins also enhance the fracture toughness of heavily stressed wings, mitigate collision damage and the tapered shape improves span efficiency during root-flapping (Dirks and Taylor, 2012; Mountcastle and Combes, 2014; Rajabi et al., 2015). The current shape, however, is probably from the sole result of an evolutionary selection process towards maximum aerodynamic performance (Ray et al., 2016). Insect wings generally serve for more than flight; wing-beat frequency, for instance, is important in male and species recognition, territorial or sexual signaling that are fundamental evolutionary requirements affecting the organism's fitness and reproductive isolation in sympatric populations of closely-related mosquito species (Gibson et al., 2010; Chapman et al., 2003). Moreover, insect wings may be involved in other biological functions, such as protection and defense, thermoregulation, self-cleaning and have super-hydrophobic and antimicrobial properties (Byun et al., 2009; Ivanova et al., 2013; Pogodin et al., 2013;

Kuitunen et al., 2014; Nguyen et al., 2014; Pass, 2018). In both *An. daciae* sp. inq. and *An. maculipennis* s. s., the lack of correlation between COI genotype and wing shape and the same spatial-temporal distribution among different morphotypes indicated that they cannot be considered ecotypes (Gildenhard et al., 2019). The recurrent variations or stasis observed among species and within species may have a phylogenetic and functional origin. Variability among and within sympatric species could be related to environmental factors (e.g. temperature, water scarcity, anthropic action, land use and chemicals). Such factors not only determine the species distribution, habitat suitability and niche dimension but may affect developmental plasticity by altering gene-expression patterns and give rise to polyphenisms (Gilbert, 2001; Rodriguez and Beldade, 2020). The occurrence of genotypes that differ in the amount and direction of plasticity that they are able to express is major mechanism of rapid adaptation and response to environmental and global change (Behera and Nanjundiah, 2004; Fox et al., 2019). Looking ahead, the effect of temperature during egg development on different morphotypes of *An. daciae* sp. inq. and in *An. maculipennis* s. s. could be evaluated (Kingsolver and Buckley, 2017; Rodriguez and Beldade, 2020; Bertola et al., 2022). The use of an instrument to capture images and wingbeat frequency and the analysis of such data by artificial intelligence and deep learning are innovative approaches in biology and ecology (Christin et al., 2019). Convolutional Neural Networks (CNNs) have demonstrated high accuracy in performing image classification tasks, including spectrogram classification (Hershey et al., 2017; Dong et al., 2018). Advances in automated mosquito identification could provide critical tools to monitor mosquito populations and surveillance in real-time (Kim et al., 2021).

4. Chapter 3: Community Ecology

4.1 Unsupervised Machine Learning and Data Mining Procedures Reveal Short Term, Climate Driven Patterns Linking Physico-Chemical Features and Zooplankton Diversity in Small Ponds

Data in ecology often present high stochasticity, correlated features and many predictors compared to the sample size of the dataset. In community analysis, useful techniques to explore environmental and biological datasets include multivariate analyses and classical clustering algorithms. The rise of machine learning algorithms in ecology in recent decades has become accessible thanks to the advance in computation power, large amounts of data and software availability (Rammer et al., 2019). These algorithms are well suited to deal with complex and large ecological datasets and with nonlinearity (Christin et al., 2019). Some machine learning algorithms are useful with datasets composed by a higher number of features as compared to the number of observations (Brownscombe et al., 2020). Generally, ML algorithms are divided into two groups: supervised and unsupervised (Crisci et al., 2020; Lek and Guégan, 1999; Olden et al., 2008; Recknagel, 2001). In supervised learning, the algorithms learn from labelled data during a training phase and extract features to solve classification (Armitage and Ober, 2010; Lumini and Nanni; 2019; Mellios et al., 2020) or regression problems (Lee et al., 2016) when many classes or a response variable are involved in model prediction. In unsupervised learning, the algorithms identify patterns in data without considering target variables to identify clusters and structures. The fuzzy c-means is a standard method of unsupervised learning. The fuzzy-set theory provides a mathematical approach that can cope with imprecision. The fuzz classification is a set of rules that allows one to cluster a set of objects without defining discrete boundaries between clusters (Zadeh, 1965). The classical clustering procedure does not consider the incompleteness of information and the randomness of ecological data (Zimmermann, 1999; Salski, 2007). Equihua (1990) provided a demonstration that fuzzy sets are a suitable description of ecological communities as compared to other standard algorithms, but the former was scarcely applied in ecology. The main advantage of the fuzzy approach over hierarchical and partitioning clustering techniques is the ability to produce a graded membership of data (Marsili-Libelli, 1991). The fuzzy set theory has achieved good results in unsupervised classification; it was used in the identification of fuzzy soil classes (Odeh et al., 1992) and to classify existing chemicals according to their ecotoxicological properties (Friederichs et al., 1996). An approach of pattern extraction from data, widely used in market basket analysis (Zhang and Wu, 2011) but not applied in ecology, is association rules mining. The discovery of association rule is a fundamental procedure in data mining in which many algorithms are suited to identify interesting relationships among features in a dataset (Nasreen et al., 2014) and correlations among them (Geng and Hamilton, 2006). Association rules might be used to explore patterns and taxa co-occurrence in community ecology in order to disentangle and highlight which drivers acted in shaping the community structure. Several algorithms are proposed such as Apriori, Frequent Pattern Growth (FP-growth), Rapid Association Rule Mining (RARM) and equivalence class

clustering along with bottom-up lattice traversal (ECLAT) (Han et al., 2007). These algorithms show different levels of efficiency during the operation of data mining.

Small permanent or temporary water bodies generally host large biodiversity, play a major role in biogeochemical processing and global cycles and represent model sites for studies in ecology and conservation biology (Céréghino et al., 2014; De Meester et al., 2005; Downing and Leibold, 2010; Verdonschot et al.,2011). Ponds contribute a great deal to biodiversity at a regional level as networks of habitat patches that also act as 'stepping stones' to facilitate the movement of species through the landscape (Hassall, 2014). These ecosystems are widely distributed in agricultural areas and are generally considered marginal due to their isolation, unpredictable duration and natural or anthropic disturbance with greater biotic and environmental temporal amplitudes than rivers and lakes. The factors affecting crustacean zooplankton community structure and the comparison between different water bodies have been described and their effect may be blurred by historic or geographic reasons (Céréghino et al., 2008; Søndergaard et al., 2010; Dodson et al., 2000; Dzialowski, 2013; Kruk et al., 2009; Meerhoff et al., 2007; Pinto-Coelho et al., 2005; Wei et al., 2017). Among others, climate change, land use, irrigation strategies and contamination by heavy metals and pesticides may cause different adverse effects on species diversity. Sensitive species may be eliminated or replaced, food-web or predator– prey interactions may be altered, and species or strains may acclimate or be selected by stress (Belfiore, 2001; Bossuyt and Janssen, 2005; Guan and Wang, 2006; Hanazato, 1991; Hunter and Pyle, 2004; Schindler, 2001; Schindler, 2009; Riessen, 2012; Vadadi-Fülöp, 2012). Factors acting upon species diversity in shallow ponds may produce local effects, contrasting among ponds located in the same geographical area. Water supply, for example, may be different among ponds, either from groundwater or from surface runoff

and precipitation, producing quite different temperatures and hydrochemical regimes, in turn affecting primary and secondary production. Climatic anomalies and the increasing use of water for irrigation purposes, both from surface and from the aquifer, add complexity to this topic and can produce diverging paths of shallow ponds. Irrigation and climate change are demonstrated to produce large inter-annual and intra-annual vertical

migrations of the aquifer, which are expected to produce large differences in the chemistry and biology of small-volume water bodies (Rotiroti et al., 2019). Fuzzy c-means and association rules mining were applied to assess the factors influencing pond assemblage composition and biodiversity. The distribution patterns of zooplankton taxa in 24 ponds located in an agricultural landscape in the core of Po river Basin (Northern Italy) were studied in relation to various habitats and environmental variables. Data recorded in 2014 and in 2015 were compared as, in the study area, mean temperatures in this 2-year period were very different. In 2014, the winter was much warmer while late spring and summer were much colder than the average recorded in the past and in 2015 (Rossi et al., 2015). In most of 2015, mean monthly temperatures were much warmer than the average recorded during the past. It was hypothesized that in small and shallow aquatic ecosystems, water temperature, chemistry and the build-up of short-living biological communities

(e.g., planktonic organisms) follow completely different trajectories depending on climatic anomalies affecting the amount and source of the water supply.

Shallow ponds are weakly buffered against perturbations due to their small water volume and limited thermal and dilution capacity. Colder temperatures, associated with groundwater inputs, may delay algal blooms. Stagnation may produce anoxia and accumulation of solutes from sediments whereas diffuse inputs from their watersheds, especially in agricultural areas, may increase nutrient concentrations favoring algal growth. Overall, these sometimes contrasting effects prevent a clear understanding of shallow pond diversity and functioning (e.g., the two-way interactions between physico-chemical features and biological communities) (Cèreghino et al., 2008). Fuzzy c-means algorithms were applied as an analytical tool to classify the 24 farm ponds in terms of the 12 zooplankton taxa they supported, and to specify the influence of environmental variables related to land-use and to pond characteristics on the assemblage patterns. Data recorded in 2014 and 2015 were compared taking into account that interannual temperature variations might explain apparently erratic community-wide responses. Besides this main objective, the present work represents a methodological contribution to environmental sciences research, and in particular an application of machine learning in a case study that is generally analyzed by multivariate statistical analysis.

Materials and Methods

Data Collection

In this study we focused on the occurrence of the main zooplankton taxa in 24 pools and ponds that were randomly selected in a 200 km^2 area located in the Cremona province (central part of northern Italy) (Marková et al., 2016) (Figure 44).

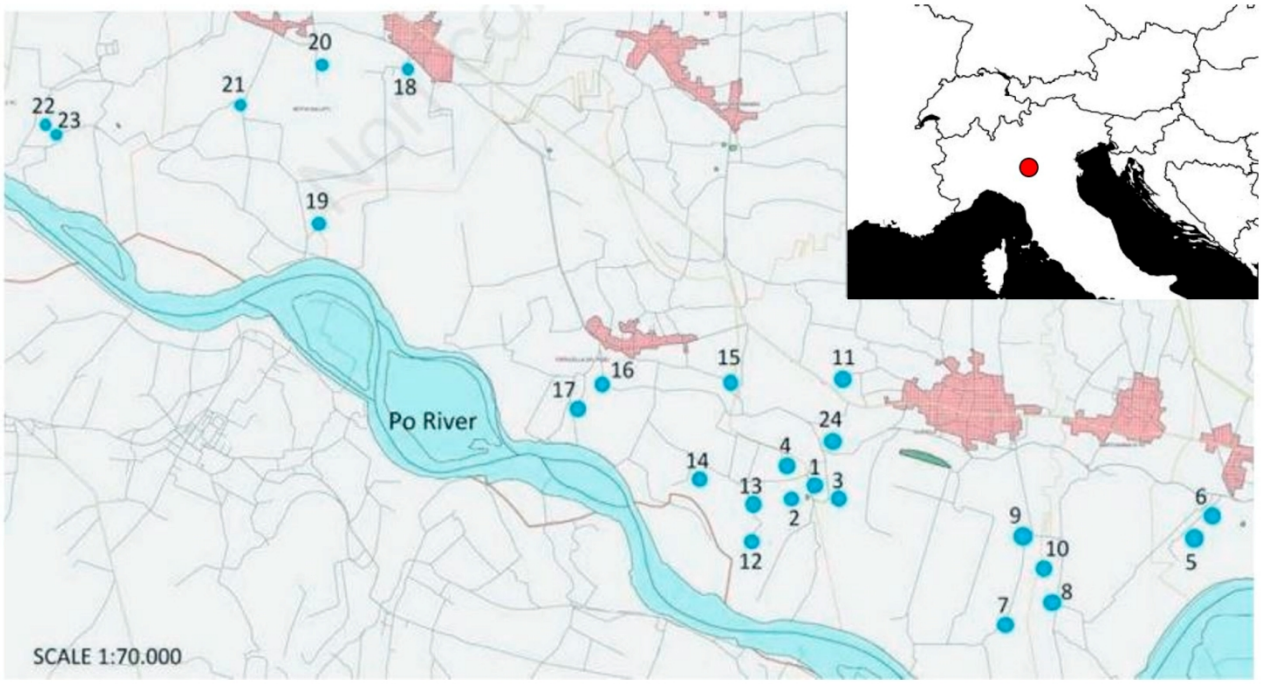

Figure 44 The map reports the position of the 24 ponds (Table 12SM), located along the left hydrographyc bank of the Po River, in the Cremona province (North Italy).

Analyzed temporary pools and ponds, locally named bodri, have originated by flooding events of the Po River: erosive processes dug cone-shaped holes with depths up to 6–10 m and size varying between 1529 and 7070 m^2. Bodri generally display pronounced water level fluctuations, regulated by the Po river hydrometric level, precipitations, runoff, vertical migration of the aquifer, also due to irrigation, and summer evaporation. They represent spots of naturality within heavily exploited agricultural contexts and are vulnerable to diffuse pollution due to their small size. Many of the studied water bodies originated before 1723 (AAVV, 1999) (for details see Table 12SM). At present, most bodri are eutrophic, undergo rapid infilling and display pronounced seasonal and daily variation of physico-chemical features. During surveys, they were characterised by the dominant form of primary producers (i.e., phytoplankton, submersed, floating leaves or emerged plants), for the level of saturation of dissolved gas of biological interest (i.e., O_2, CO_2, N_2 and CH_4), for dissolved nutrients (the inorganic forms of N, P and Si) and for sedimentary features (i.e., organic matter content).

Each pond was sampled twice: the first time between May and June 2014 and the second time between June and July 2015. Qualitative zooplankton samples were collected by 105 μm-mesh size plankton nets. Two to sixteen litres of water were filtered for each sample according to the estimated water volume and depth. All samples were preserved in 95% ethanol. All organisms present in the sample were sorted under a stereomicroscope and cladocerans were identified to genus level whereas copepods were distinguished in Calanoida and Cyclopoida. For each pond, 2 litres of water were sampled with a PE bottle. Nine chemical and three physical environmental descriptors were determined for each pond (Tables 13SM and 14SM). Water temperature (wT), dissolved oxygen concentration, pH and electrical conductivity (EC) were measured in situ with a multiparameter probe (YSI model 566 MPS). In the laboratory, the water collected was filtered with Whatman GF/F filters (0.45 μm) and stored in glass vials (Labco Exetainer®, Lampeter, Wales, UK) for the determination of dissolved inorganic carbon (DIC) (Anderson et al., 1986) and soluble reactive phosphorus (SRP) (Valderrama, 1977), and in PE vials for the determination of dissolved reactive silica (SiO_2), ammonium (NH_4^+) (Water Environmental Federation, 1981) and nitrate (NO_3^-) (Rodier, 1987) (Tables 13SM and 14SM). Chlorophyll-a (Chla) concentration was determined spectrophotometrically after filtration of 100–500 mL of water (0.45 μm Whatman GF/F filters) and extraction of pigments with 90% acetone. Besides physico-chemical and biological parameters, the ponds perimeter and main depth were also considered in the study as proxies of size (D'Auria and Zavagno, 1999) (Tables 13SM and 14SM).

Environmental Features Selection

To avoid multicollinearity and to reduce redundant information from the set of environmental features, a score called variance inflation factor (VIF) was computed (Bruce and Bruce, 2017; James et al., 2014). For a given predictor (p), the variance inflation factor measures how much the variance of a regression coefficient

is inflated due to multicollinearity in the model. The smallest possible value of VIF is one (absence of multicollinearity). A VIF value that exceeds 5 or 10 indicates a problematic amount of collinearity (James et al., 2014). In this work, as a conservative rule of thumb, a threshold equal to 4 was set. VIF was computed for all the environmental features with a stepwise procedure. The environmental features with VIF values > 4 were then excluded and the procedure was repeated until no environmental features with VIF greater than threshold remained.

Fuzzy Clustering

Fuzzy clustering aims at defining a membership value of an object that can be split between different clusters. The most common clustering method, the fuzzy c-means (Tilson et al., 2005), bases the clustering procedure on the minimization of an objective function as reported in Equation (1):

$$J(c) = \sum_{i=1}^{p} \sum_{j=1}^{c} \left(\mu_{ij}\right)^{m} d_{ij}^{2} \qquad (1)$$

where d_{ij} is the distance between the ith observation and jth centroid, p is the number of observations, c is the number of clusters ($2 \leq c \leq n$), μij is the membership degree of the ith observation to the jth cluster and satisfying the following conditions:

$\mu_{ij} \in [0,1]$ for $1 \leq i \leq p$, $1 \leq j \leq c$
$\Sigma\, \mu_{ij} = 1$ for $1 \leq i \leq p$
$\Sigma\, \mu_{ij} > 0$ for $1 \leq j \leq c$

The exponent fuzzifier m defines the degree of fuzziness of the partition, when m approximates the value of 1, it operates as the k-means algorithm. Meanwhile, when m increases in value, the degree of fuzziness increases, and the fuzzy c-means leads to a solution where the memberships of each observation approximate 1/c (Tilson et al., 2005). The evaluation of the quality of the cluster procedure is made with a particular function that will be maximized or minimized according to the number of clusters c (Roubens, 1982). These procedures allow one to know how well the algorithm fitted the data structure (cluster validity problem). The most common measures for this task are the partition coefficient (PC) and the partition entropy (PE). In this study, the fuzzy c-means on the environmental dataset was used. The environmental features were standardized and a search grid procedure was used: the fuzzy c-means was run multiple times. For each run, a combination of the parameter c (number of cluster) and the fuzzifier exponent m, were set. The best partition was selected according to the maximum value of PC, or, in alternative, to the minimum value of PE. To improve the clustering procedure for each run, the algorithm was randomly initialized 50 times. The Principal Component Analysis (PCA) was performed as an operation of dimensionality reduction, to

improve the visualization of the ponds with the membership values estimated from fuzzy c-means. The prototypes or the centroids of the estimated clusters, that are the values of the environmental features that characterized each cluster, were compared and the Trophic State Index (TSI) based on the values of prototype of the Chla was computed according to Carlson (1977) using the Equation (2):

$$TSI = 10\left(6 - \frac{2.04 - 0.68 \ln Chla}{\ln 2}\right) \qquad (2)$$

To quantify the habitat heterogeneity of the environmental features, the Euclidean distances matrix between each observation (pond) and the median of each cluster were standardized and computed. Habitat heterogeneity was estimated by the average distance from the clusters' medians (Heino et al., 2013). The Permutational analysis of multivariate dispersions (PERMDISP) test for the analysis of multivariate homogeneity of groups was used (Anderson, 2005). PERMDISP compares within-group variance among clusters using the mean distance from individual observations to their cluster median. Bosco Bodini pond was excluded as it was dry in 2015 whereas two chemical parameters ($NH4^+$ and DIC) were excluded due to multicollinearity in 2014. Square root corrections were applied for groups of unequal size (Stier, 2013) and to test differences in habitat heterogeneity among clusters, a permutation procedure (n = 999) was used. Average Euclidean distances from clusters' medians were visualized in a reduced space with Principal Coordinate Analysis (PCoA). The analysis was carried out with the R package `vegan` (Oksanen et al., 2020).

Richness and Beta Diversity

For each year and for each cluster the taxa richness and the community structure were computed. The richness of the number of taxa observed was compared between clusters of the same year and by Mann-Whitney U Test. Differences in Richness were tested also between years with Wilcoxon Signed-Rank Test for paired samples. The alpha diversity (α) or the mean number of taxa were computed between years. The Sorensen index (βSOR) for presence/absence data was used as a measure of beta diversity for multiple sites (Baselga, 2013). The beta diversity was partitioned in two components: nestedness (βSNE) and turnover (βSIM). The overall beta diversity (βSOR) and its components were computed considering different years and different clusters within years. To compare βSOR between and within years, a resampling procedure was applied. An equal number of sites sampled (n = 5) and a total number of samples (n = 500) were set. This procedure allowed to estimate the distributions of βSOR and the relative components, nestedness and turnover, for multiple sites with equal number of ponds. The estimated distributions were compared with the Kolgomorov-Smirnov test. In order to highlight differences in zooplankton community diversity across time, pairwise measures of βSOR of each pond in two different years (2014 and 2015) were compared (Baselga

and Orme, 2012). The βSOR analysis was carried out with the R package `betapart` (Baselga and Orme, 2020).

Community Structure and Association Rules

Community structure was described by characteristic taxa, association rules mined from frequent pattern tree growth (FP-Growth) and visualized by frequent pattern tree (FP-tree). For each cluster, the characteristics taxa were computed using the indices of presence (Pi) (Rachor et al., 2007). For each taxon, Pi was expressed as Pi = Pic/Nstc, where Pic is the i taxon belonging to a particular cluster, and Nstc is the number of ponds in a particular cluster. A taxon was identified as characteristic if its indices of presence were higher than the threshold Pi, set at 0.6 (Rachor et al., 2007).

Considering the whole dataset of presence/absence data, association rules were extracted using frequent pattern growth algorithm (FP-growth), to highlight and evaluate correlations among co-occurrence of different taxa. An association rule is an im-plication X→Y that describes the existence of a relationship between X and Y species or group of species (Hoppner, 2009). FP-growth is based on a divide and conquer approach, the algorithm identifies small patterns by decomposing the mining problem into a set of smaller ones represented by conditional database, extracted on a compressed data representation, the FP tree. This approach reduces the search space and the computational effort (Nasreen et al., 2014).

To select an association rule from the set of all possible rules, constraints of various quantitative measures of interestingness and significance were applied, using objective measures (Freitas, 1998; Silberschatz and Tuzhilin, 1995). Interestingness measures the strength of the relationship between X and Y. As first step of association rule mining, the threshold values of support-confidence framework were used (Agrawal et al., 1993). Support measures the probability to observe a particular group of species X in the dataset, while the confidence is the conditional probability to observe the species Y given the presence of the species X. A threshold value of minimum support equal to 0.1 and minimum confidence equal to 0.80 were set. A second step relied on an interestingness measure called lift. Lift quantifies the statistical dependence of two or more taxa in a particular rule; it is a positive real number, with a value equal to 1 under statistical independence (Geng and Hamilton, 2006). Association rules were sorted in descending order of lift and association rules with lift value lower or equal to 1 were not considered (Chiu et al., 2006). The zooplankton community structure of each cluster was visualized in a compact way using the frequent pattern tree (FP tree) (Nasreen et al., 2014; Geng and Hamilton, 2006). Association rules were mined by Weka software version 3.8.4 (Frank et al., 2016) and visualized with the R packages `Arules` and `ArulesViz` (Hashler et al., 2019; Hashler et al., 2020).

Results

Environmental Features Selection

The analyzed shallow ponds exhibited pronounced variations of physico-chemical and biological parameters, reflecting different, site-specific equilibria between assimilative (e.g., oxygen-producing algal blooms, controlling nutrients) and dissimilative processes (e.g., heterotrophic microbial oxygen consumption recycling nutrients). During 2014 and 2015, most environmental features such as water temperature (wT), pH, conductivity (EC), soluble reactive phosphorus (SRP), nitrate (NO_3^-), chlorophyll-a (Chla), dissolved reactive silica (SiO_2), depth and perimeter showed values of VIF < 4 (Table 9). The selected variables were 9 in 2014 and 11 in 2015 according to VIF value > 4. Dissolved inorganic carbon (DIC) and ammonium (NH_4^+), showed multicollinearity in 2014 whereas dissolved oxygen (O_2), showed multicollinearity in both years, and were removed from further analyses. Dissolved inorganic carbon (DIC) was positively correlated with conductivity (EC), ammonia (NH_4^+), soluble active phosphorus (SRP) and reactive silica (SiO_2). Ammonium (NH_4^+) was positively correlated with soluble reactive phosphorus (SRP) and negatively correlated with chlorophyll-a (Chla), pH and oxygen (O_2). In 2014, dissolved oxygen (O_2) was positively correlated with water temperature (wT), pH, chlorophyll-a (Chla), and negatively correlated with dissolved inorganic carbon (DIC), soluble reactive phosphorus (SRP) and ammonium (NH_4^+) (Figure 5SM). In 2015, oxygen (O_2) was positively correlated with pH and nitrate (NO_3-) whereas ammonium (NH_4^+) and dissolved inorganic carbon (DIC) showed values of VIF < 4.

Table 9 VIF values for each chemical and physical environmental features in 2014 and 2015. Environmental features with values of VIF > 4 were reported in bold.

Environmental Features	VIF (2014)	VIF (2015)
Water temperature (wT)	1.40	1.69
pH	3.90	2.15
Oxygen (O_2)	**>4**	**>4**
Conductivity (EC)	1.96	2.78
Ammonia (NH_4^+)	**>4**	3.88
Dissolved inorganic carbon (DIC)	**>4**	1.81
Soluble reactive phosphorus (SRP)	1.99	1.90
Nitrate (NO_3^-)	1.22	2.25
Chlorophyll-a (Chla)	2.38	1.29
Silica ($Si0_2$)	2.06	2.60
Depth	1.33	1.97
Perimeter	1.32	1.74

Fuzzy c-Means

In both years, the number of clusters was c = 2, corresponding to a value of the fuzzifier m = 1.5 (Figure 6SM). The evaluation of the quality of the clustering was made considering the best partition of the maximum value of partition coefficient (PC) and the minimum value of partition entropy (PE). In 2014 and 2015, the maximum value of PC, 0.68 and 0.63, respectively, and the minimum value of PE, 0.49 and 0.56, respectively, were obtained for c = 2. In 2014, the highest membership associated to the cluster 1 were observed for Pavarini, Pescaroli West and Santa Maria Maddalena, and to the cluster 2 for Bosco Piazza, San

Giorgio and Cascina Tavernelle (Table 10). In 2015, the highest memberships associated to cluster 1 and 2 were observed for Motta, Bicocca and Pastore 1 and for Pastore 4, Bosco Braca and Pescaroli West, respectively. In 2014, the prototypes showed that ponds in cluster 1 were characterized by higher values of wT, pH, Chla, SiO_2, depth and perimeter and lower values of EC, SRP and $NO3^-$ than ponds in cluster 2 (Figures 45). In 2015, the ponds grouped in cluster 1 were characterized by lower values of pH, Chla and higher values of EC, NH_4^+, DIC, SRP, NO_3^-, reactive silica (SiO_2), depth and perimeter than ponds in cluster 2. The wT of cluster 1 was similar to that of cluster 2 (Table 10). In 2015, both clusters estimated by fuzzy c-means showed higher values of the prototypes relative to wT, Chla and SiO_2, compared to the prototypes of clusters estimated in 2014. In 2014, the difference of trophic status between clusters was higher than in 2015. In 2014, the TSI was 40.30 for cluster 1 and 29.56 for cluster 2 while, in 2015, it was 42.70 for cluster 1 and 45.23 for cluster 2. Ponds that, in both years, remained grouped in the same cluster were Pastore 3, Vecchio, Bazzi and Motta, grouped in cluster 1 and Temporanea, Bosco Valloni, San Giorgio, Forche, Martignana, Bosco Piazza and Cascina Tavernelle, grouped in cluster 2. Habitat heterogeneity, estimated by the average distance from clusters' median, was 2.10 (cluster 1) and 2.30 (cluster 2) in 2014, while it raised to 3.10 (cluster 1) and 2.54 (cluster 2) in 2015 (Figure 7SM). The permutation test showed that habitat heterogeneity was not significantly different among clusters ($F_3 = 1.3103$, p-value = 0.273).

Table 10 Freshwater pond's membership for each cluster in the year 2014 and 2015. Values underlined and in bold were referred to the higher membership associated to a particular cluster.

Ponds	2014		2015	
	Cluster 1	Cluster 2	Cluster 1	Cluster 2
Pastore 3	**0.72**	0.28	**0.63**	0.37
Pastore 1	0.14	**0.86**	**0.77**	0.23
Pastore 4	**0.56**	0.432	0.11	**0.89**
Temporanea	0.30	**0.70**	0.41	**0.59**
Bosco Braca	**0.88**	0.12	0.13	**0.87**
Pavarini	**0.93**	0.07	0.26	**0.74**
Bosco Valloni	0.41	**0.59**	0.31	**0.69**
San Giorgio	0.08	**0.92**	0.33	**0.67**
Forche	0.39	**0.61**	0.1	**0.9**
Martignana	0.29	**0.71**	0.29	**0.71**
Santa Maria Maddalena	**0.88**	0.12	0.34	**0.66**
Bosco Bodini	**0.65**	0.35	-	-
Cacina Mortara	0.18	**0.82**	**0.72**	0.28
Bosco Piazza	0.03	**0.97**	0.3	**0.7**
Cascina Tavernelle	0.09	**0.91**	0.25	**0.75**
Vecchio	**0.63**	0.37	**0.79**	0.21
Bazzi	**0.72**	0.28	**0.57**	0.43
Motta	**0.75**	0.25	**0.81**	0.19
Ronchetto	**0.91**	0.09	0.22	**0.77**
Rita	**0.54**	0.46	0.42	**0.58**
Bicocca	0.26	**0.74**	**0.77**	0.23
Pescaroli West	**0.93**	0.072	0.12	**0.88**
Pescaroli East	0.17	**0.83**	**0.70**	0.30

Sabbie	**0.61**	0.39	0.28	**0.72**

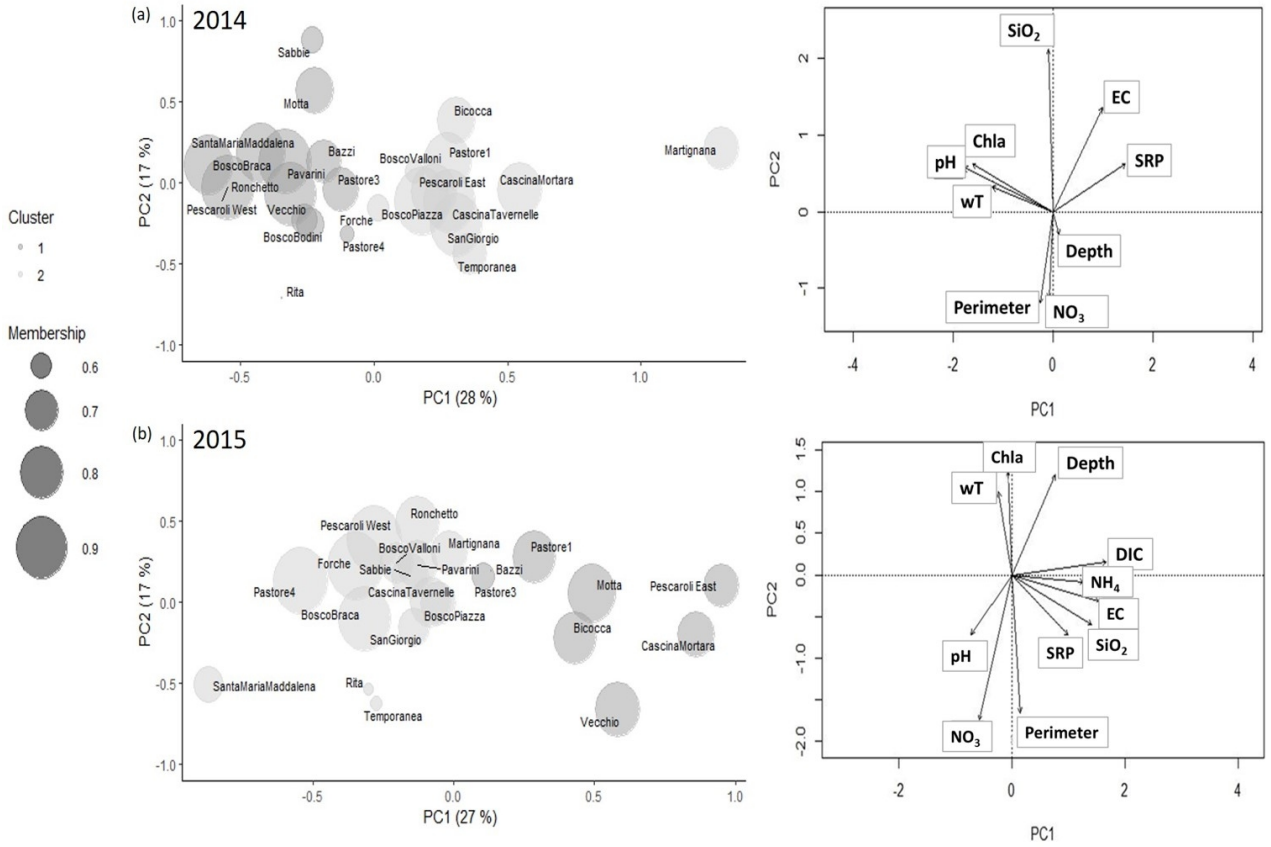

Figure 45 Principal component analysis (PCA) for the 2014 (a) and 2015 data (b). In 2014, 9 environmental features were considered, and the first two principal components explained 45% of the variance. In 2015, 11 environmental features were considered, and the first two principal components explained 44% of the variance. Each pond was represented by a bubble with size proportional to the membership value of the pond to a cluster. In both years, the fuzzy c-means algorithm identified two clusters: cluster 1 (dark grey) and cluster 2 (light grey). On the right side, the loadings of each variable were reported. The arrow lengths provided the degree of correlation among each original variable and the principal components.

Table 11 Cluster's prototypes of each environmental features in the 2014 and 2015.

Environmental Features	**Prototypes 2014**		**Prototypes 2015**	
	Cluster 1	**Cluster 2**	**Cluster 1**	**Cluster 2**
Water temperature (T)	20.29	17.69	23.49	23.92
pH	8.00	7.63	7.67	7.71
Conductivity (EC)	542.18	673.30	584.32	364.03
Ammonia (NH_4)	-	-	3.58	2.35
Dissolved inorganic carbon (DIC)	-	-	0.65	0.24
Soluble reactive phosphorus (SRP)	0.058	0.18	0.098	0.044
Nitrate (NO_3)	0.11	0.17	0.17	0.15
Chlorophyll-a (Chla)	2.70	0.90	3.44	4.46
Silica ($Si0_2$)	2.44	2.24	14.30	7.47

Depth	4.28	4.18	4.50	4.02
Perimeter	209.08	205.97	221.58	204.73

Richness and Beta Diversity

In 2014, the taxa richness was significantly lower in cluster 1 than in cluster 2 (W = 35.5 and p-value = 0.036), whereas in 2015 it was not different between clusters (W = 69 and p-value = 0.5651) (Figure 46). Richness was higher in 2014 than in 2015 (V = 152.5, p-value = 0.02) (Figure 46), with values of alpha diversity (α) equal to 3.61 in 2014 and to 2.56 in 2015. The distributions of beta diversity index (βSOR) and the relative components nestedness (βSNE) and turnover (βSIM), estimated by resampling, were statistically different considering clusters and years (Table 12 and Figure 47). In 2014 the overall beta diversity (βSOR) and the turnover (βSIM) were higher than in 2015, but the nestedness (βSNE) was lower in 2014 than in 2015 (Table 15SM and Figure 47).

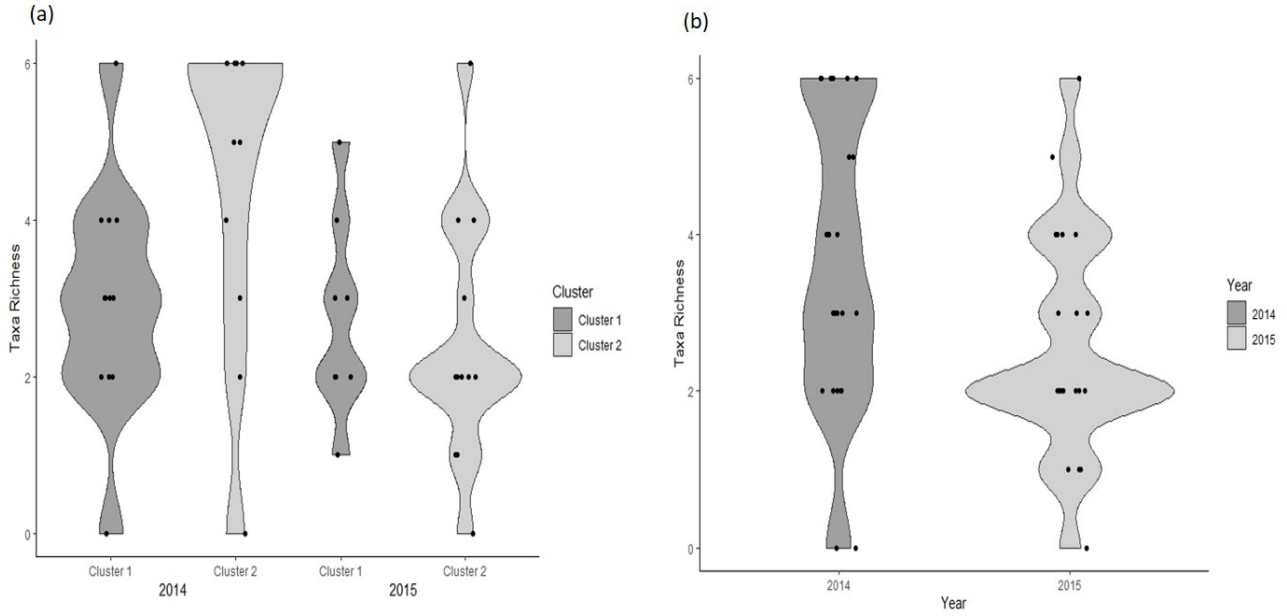

Figure 46 The panel (a) reports the violin plot of cluster 1 (dark grey) and cluster 2 (light gray) taxa richness relative to the years 2014 and 2015. The panel (b) reports the violin plot of taxa richness in the pooled clusters in 2014 (dark grey).

Table 12 Beta diversity and the relative components nestedness (βSNE) and turnover (βSIM), computed for different cluster and years; α was the p-value of the results obtained with Kolmorov-Smirnov test after the permutation procedure (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ and **** $p < 0.0001$).

Beta	2014			2015			Overall	Overall	α
Diversity	Cluster 1	Cluster 2	α	Cluster 1	Cluster 2	α	2014	2015	
β_{SOR}	0.85	0.76	****	0.71	0.84	****	0.89	0.87	****
β_{SNE}	0.07	0.14	****	0.19	0.15	****	0.07	0.12	****
β_{SIM}	0.78	0.61	****	0.52	0.69	****	0.82	0.75	****

In 2014, beta diversity (βSOR) and turnover (βSIM) were higher in cluster 1 than in cluster 2, but nestedness (βSNE) was lower in cluster 1 than in cluster 2. In 2015, beta diversity (βSOR) and turnover (βSIM) were higher in cluster 2 than in cluster 1, but the nestedness (βSNE) component was higher in cluster 1 than in cluster 2 (Table 15SM and Figure 47). The pairwise comparison of beta diversity (βsor) between ponds in different years was maximum (βsor =1) for Vecchio, Santa Maria Maddalena, Rita, Pescaroli East, Pavarini, Motta, Martignana and Bosco Braca while the minimum value was recorded for Bosco Valloni (βsor = 0.33) (Figure 47).

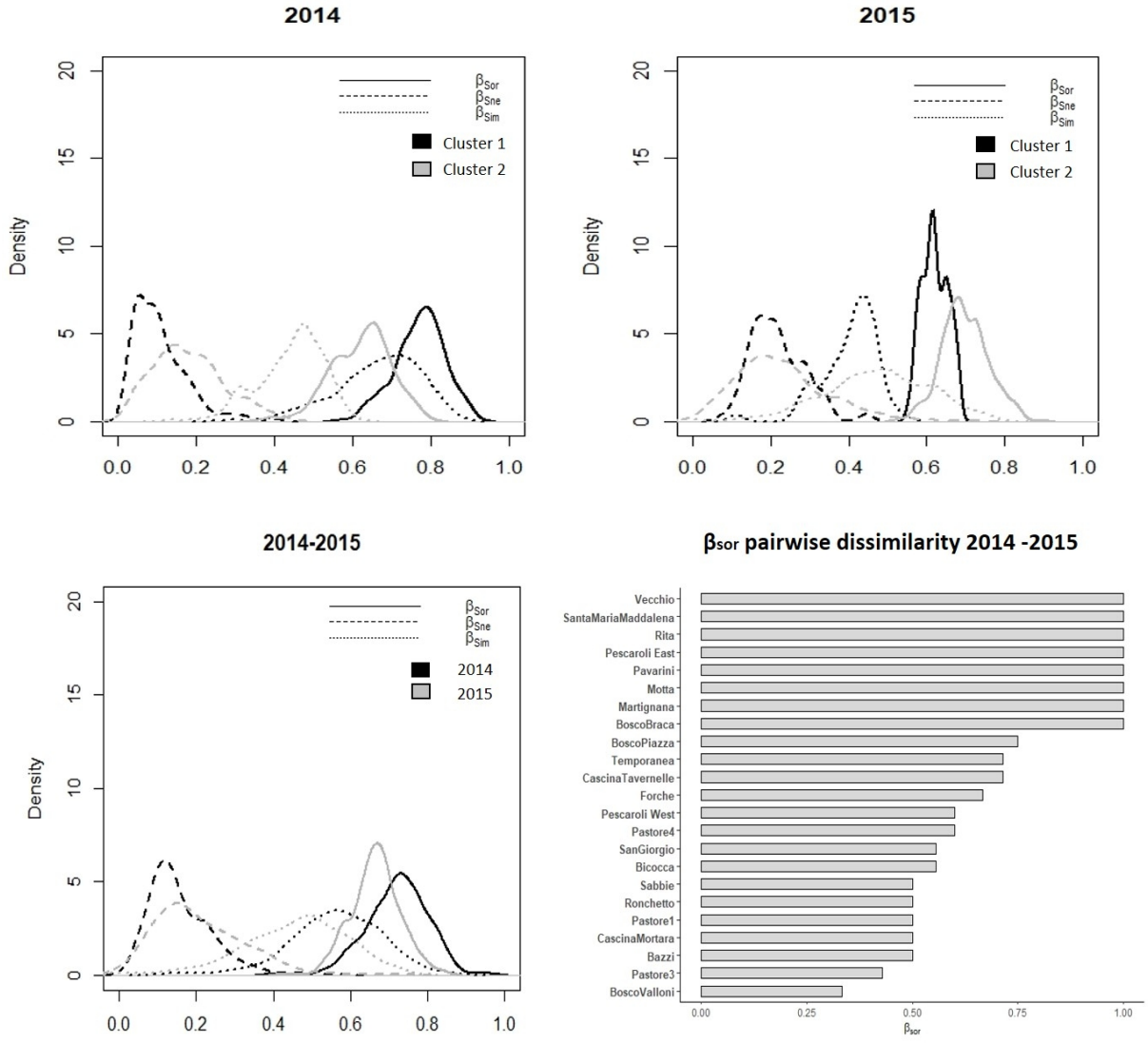

Figure 47 The panels (a, b) report the distributions of the beta diversity index (βSOR, continuous line), beta nestedness (βSNE, coarse dashed line) and beta turnover (βSIM, tiny, dashed line) for the 2 clusters and for the 2 years of study. The distributions were estimated with a boostrapping procedure (n = 500). The panel (c) reports the overall distribution of beta diversity and the relative component, nestedness and turnover, in 2014 and 2015. The panel (d) reports the histograms of the pairwise beta diversity (βsor) between the same pond in two different years.

Community Structure and Association Rules

In 2014, the characteristic taxa were Calanoida (Pi = 0.69) for cluster 1 and *Simochephalus* (0.72), Calanoida (0.72), *Pleuroxus* (0.64), and Cyclopoida (0.63) for cluster 2. In 2015, the characteristic taxa of both clusters were represented by Calanoida (cluster 1 = 0.87 and cluster 2 = 0.73) and *Daphnia* (cluster 1 = 0.62 and cluster 2 = 0.63) (Figure 48). Considering the whole dataset, 9 association rules were found (Figure 49 and Table 16SM). One association rule showed the higher values of lift (3.92) and indicated to the co-occurrence of Calanoida, Cyclopoida, and *Pleuroxus* with *Simochephalus*. This association was found in ponds of cluster 2 in 2014. Three association rules with lift values equal to 1.52 indicated the co-occurrence of *Daphnia*, *Simocephalus*, *Chydorus* and *Pleuroxus* with Calanoida (Table 16SM).

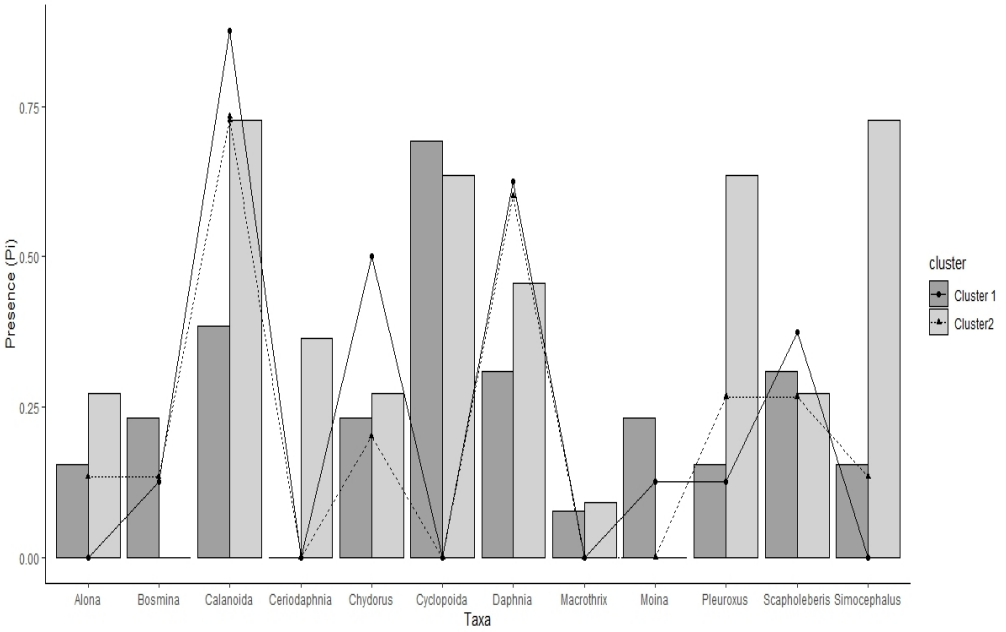

Figure 48 The barplot was relative to the years 2014, each values of the taxa presence (Pi) was relative to a particular cluster, reported in different colors (dark grey for cluster 1 and light grey for cluster 2). The lines and points showed the values of taxa presence (Pi) for the year 2015, where each cluster were reported with different shape of points and line types (circle with continuous line for cluster 1 and triangle with dotted line for cluster 2).

These associations were found in the clusters 2 of both years 2014 and 2015 and in cluster 1 of 2015. The association rule with the lowest lift value (lift = 1.26) indicated the co-occurrence of *Chydorus*, Cyclopoida with Calanoida and was found in both clusters 1 and 2, in 2014. Frequent pattern tree (FPt) revealed different patterns in the community structure between clusters and years. In 2014, most of the ponds of cluster 1 were characterized by the presence of Calanoida and Cyclopoida, whereas the community structure of cluster 2 was characterized by the co-occurrence of Simocephalus, Calanoida, *Pleuroxus* and Cyclopoida (Figure 50 and Figure 8SM) as expressed by the first association rule with the highest value of lift. In 2015, the community structure of cluster 1 was characterized by the presence of Calanoida, *Daphnia* and *Chydorus*

while cluster 2 was characterized by the presence of Calanoida, *Daphnia* and *Scapholeberis* (Figure 50 and Figure 8SM).

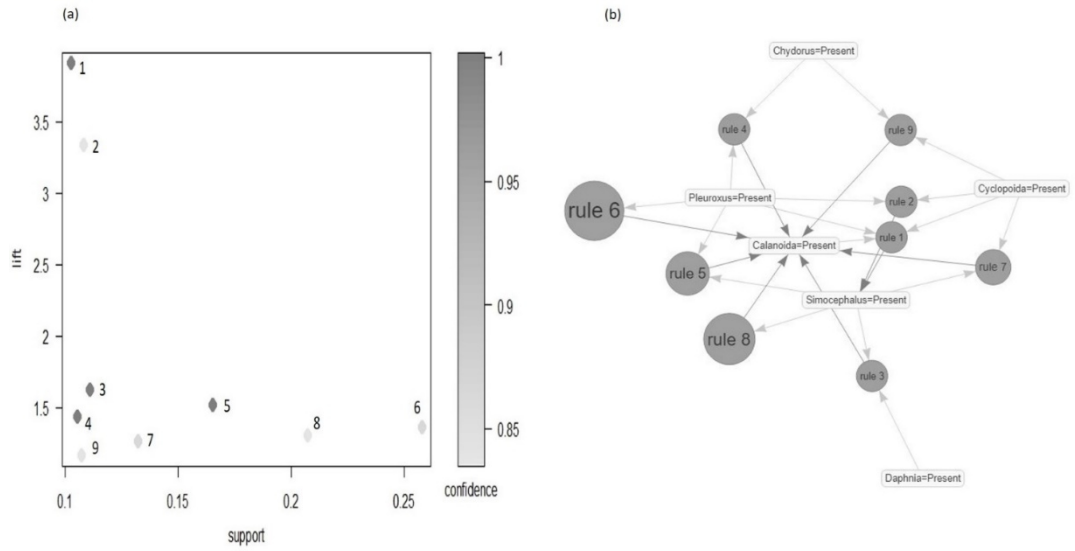

Figure 49 The panel (a) shows the scatterplot of the interestigness measure support and lift for each association rules after the pruning procedure. The gray scale color is proportional to the confidence value of each rule. The labels' number refers to the descending order by lift. The panel (b) reports the taxa co-occurrence as network structure. Each taxon relates to each other by association rule estimated with FP-growth algorithm. The labels' number of each rule are related to the descending order by lift.

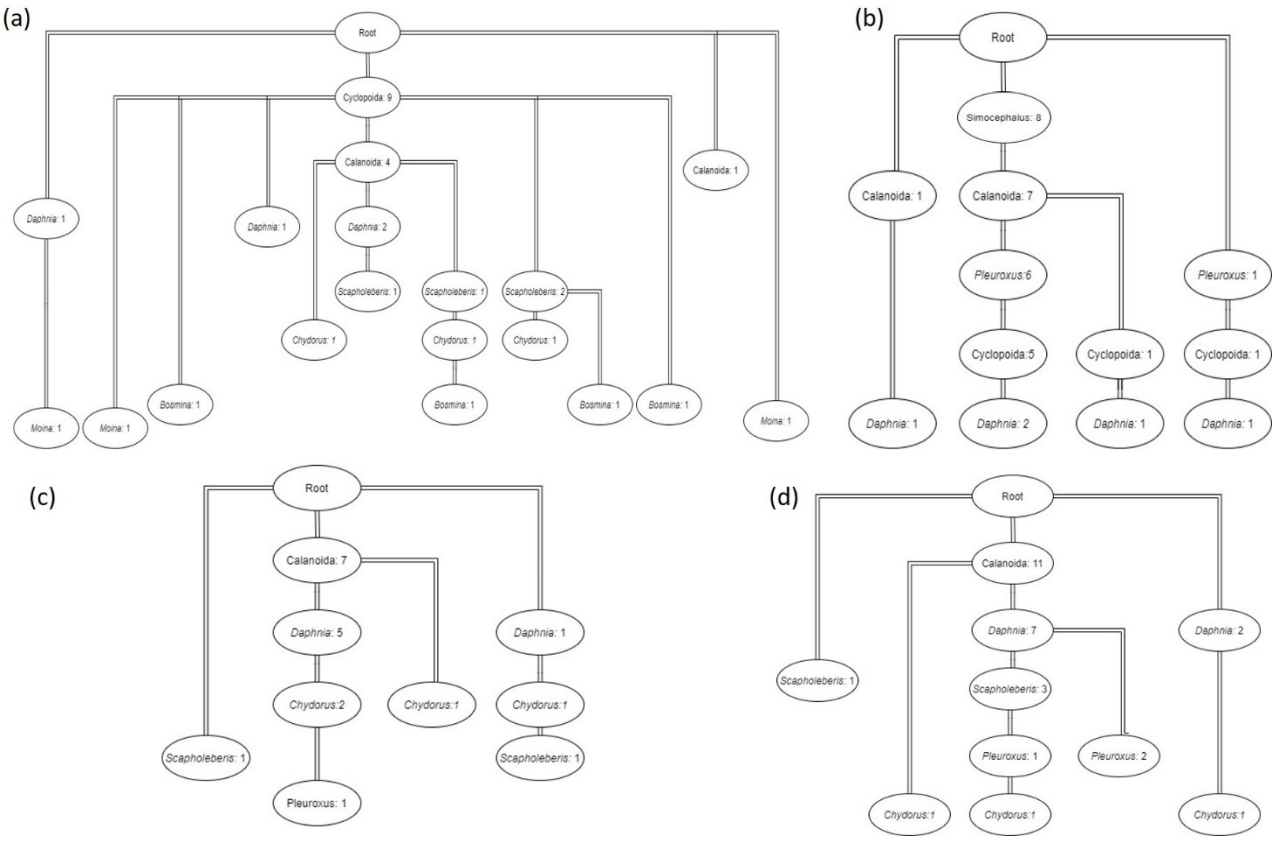

Figure 50 Frequent pattern trees (FPt) for the community structure in cluster 1 (a) and 2 (b) in 2014 and in cluster 1 (c) and 2 (d) in 2015. Each node represents a specific taxon and its absolute frequency (number of

ponds where the taxon was found). The branches join the co-occurrence of taxa. Only the taxa with frequency > 20% were reported (see also Figure 3.1.4SM).

Discussion

Results from this study suggest that shallow ponds may undergo completely different trajectories in the same geographical area and may display pronounced differences in terms of water physico-chemical and biological parameters. This is not surprising due to their overall small size and to the small ratio between their water volume and sediment surface. The small water volume has limited buffer capacity against climatic anomalies or water ingression from the aquifer or from the watershed, resulting in local, sharp changes of physico-chemical parameters and, as a consequence, of biological communities (Bennion and Smith, 2000; Lischeid et al., 2018; Marlene et al., 2020).

Phytoplankton communities have in turn the potential to control inorganic nutrients and regulate dissolved oxygen, inorganic carbon concentrations and water pH. Assimilative processes are contrasted by nutrient regeneration from sediments, that together with the low ratio between water volume and sediment surface amplify the effects produced by microbial dissimulative pathways (e.g., oxygen shortage) (Lischeid et al., 2018). However, the most interesting result of this study does not deal with different solutes or chlorophyll concentrations in the analyzed shallow water ecosystems. Ponds are intrinsically heterogeneous, they can be net autotrophic or net heterotrophic and these extremes correlate with high chlorophyll and oxygen and low nutrient concentrations or the opposite, respectively (Recknagel and Michene, 2018). What is novel here is that the functioning of ponds, exemplified by snapshots showing phytoplankton and nutrient concentrations and zooplankton community composition, may diverge from year to year due to some sort of continuous disturbance or to the absence of a stable steady state. Such instability is favored by the vulnerability of ponds to a large set of pressures and may contribute to the paradox of their diversity, which is a cen-tral topic in recent freshwater research and has important implications for ecosystems restoration, in particular in heavily impacted agricultural areas (Bennion and Smith, 2000; Lischeid et al., 2018).

The main findings of this study were extracted from the dataset via the application of an unsupervised machine learning and data mining algorithm, fuzzy c-means and frequent pattern growth. Such approach allowed to assess the factors that influenced assemblage composition and the apparently erratic distribution patterns of zooplankton taxa in 24 ponds and in two consecutive years. Data in ecology are characterized by high uncertainty, bias and hierarchical level of complexity. Machine learning tools were used according to the level of complexity of ecological systems in order to understand environmental and biological dynamics (Humphries and Huettmann, 2018) shift in species assemblages along time (Chon et al., 2000) and to plan conservation actions for ecological communities threatened by anthropic pressure and climate change (Senent-Aparicio et al., 2017). Classical clustering methods generally define sharp boundaries between groups and each object belong only to a particular cluster. These classical procedures do not consider the continuous realm of ecological features. The assumptions based on boolean rules might lead to

misclassifications and might fail to detect outliers. A set of ecological objects, in our cases the freshwater ponds, can be partitioned using the fuzzy logic, where a probabilistic approach helps to capture the continuous nature of ecological data.

Brownscombe et al. (2020) applied supervised machine learning techniques combined with unsupervised fuzzy c-means to a wide range of informational sources in order to identify potential spawning aggregation sites for a marine fish species. The flexibility and the probabilistic output of the fuzzy logic with respect to the classical partition procedure based on crispy clustering, was used to describe species association in marine ecosystems, that consist of communities or cohesive units of not random taxa groups and assemblages of taxa that are randomly associated (Fiorentino et al., 2017). Fuzzy clustering algorithms were used also for the identification and partition of similar regions and hydrologically homogeneous watersheds (Allen et al., 1999).

In this study, for both years, the fuzzy c-means algorithm allowed to identify two different clusters. In 2014, ponds in Cluster 1 were characterized by high concentrations of chlorophyll-a, high pH and water temperature while ponds in cluster 2 were characterized by high concentrations of chemical species, with silica as only exception. Cluster 1 was more autotrophic and showed a higher TSI than cluster 2. In 2015, temperature was comparable between clusters and the situation of 2014 was reversed, with Cluster 1 more heterotrophic and characterized by higher concentration of chemical species (SRP, DIC, $NH4^{+}$, $NO3^{-}$), higher conductivity, depth and perimeter, and lower values of pH than cluster 2. In different years, in many ponds the chemistry of water changed, and large differences were recorded in Pastore 3, Vecchio, Bazzi and Motta, where the concentrations of all chemical parameters increased and a reduction of pH was recorded. In opposite, Temporanea, Bosco Valloni, San Giorgio, Forche, Martignana, Bosco Piazza and Cascina Tavernelle showed a reduction in the concentration of nutrients and of electrical conduc-tivity. In 2015, all ponds showed higher concentrations of chlorophyll-a than in 2014, confirmed by the increase of the trophic state index. This result was probably due to the effect of temperature, that was higher in 2015 than in 2014. Differences between clusters community might be related to different factors, among which the amount and the quality of water inputs from runoff or from the aquifer or the thermal regime in winter and late spring. A general increase in the concentration of reactive silica was observed in 2015 as compared to 2014, likely due to regeneration from sediments uncoupled to uptake. Fuzzy c- means identified a pond (Rita) with the lowest value of cluster's membership in both years and the highest value of nitrate, likely due to diffuse inputs. The zooplankton com-munity structure (species richness and species composition) is potentially affected by both water chemistry and site morphology, and by anthropogenic pressures in lakes and watersheds (Dodson et al., 2000; Allen et al., 1999; Belyea et al., 2012). At geographical level, the species pool is driven by dispersal constraint, whereas the habitat species pool is due to environmental constrains (Gyllström et al., 2005). The high taxonomic diversity in zooplankton communities is only partially expressed in individual freshwater habitats and the differences in zooplankton community structures among systems is largely associated with specific environmental conditions (Havens and Hazanato, 1993; Wellborn et al., 1996; Arnott and Vanni, 1993). Abiotic factors (e.g., pH, temperature, light intensity) can also influence the

zooplankton community structure of fishless aquatic habitats directly by sorting species based on differential physiological tolerances, or indirectly by interacting with biotic conditions such as primary production and invertebrate predation (Steiner, 2004; Weidman et al., 2014; Wright and Reeves, 1992). The observed taxa are also deter-mined by internal dynamics due to biotic factors, such as predation, intraspecific and interspecific competition. A particular pattern of taxa was the result of internal and external process that defined the community structure.

In this work, a data mining algorithm was used to evaluate the co-occurrence of taxa on presence-absence data in a system of freshwater ponds. This method allowed to quantify possible correlations among taxa in frequent pattern extracted from data and to highlight differences in the community structure between consecutive years. In 2014, the taxa richness was higher in cluster 2 than in cluster 1 and the community structure was different. Cluster 2 was dominated by the presence of *Simocephalus* in many ponds, while in cluster 1 the taxa with higher presence were Cyclopoida and Calanoida. Furthermore, *Simocephalus* was not present in ponds characterized by higher pH in both years. In 2014, Pastore 4 showed the lowest value of membership for cluster 1: all environmental features were similar to prototypes of cluster 1 except for chloropyll-a and reactive silica, that were more similar to cluster 2. However, the community structure in Pastore 4 was composed of Cyclopoida, Calanoida and *Daphnia* that was the most frequent pattern of cluster 1. In 2014, the community structure was characterized by not common taxa association composed by *Pleuroxus*, *Alona*, *Moina* and *Macrothrix*, while, in 2015, the presence of these taxa was not recorded. The smallest membership associated to a particular cluster was found for ponds belonging to the group with lowest nutrient concentrations. In both years, a low beta diversity was observed for clusters with higher concentrations of chemical species, high conductivity, and pH.

In general, nestedness of species assemblages occurs when the biota of sites with smaller numbers of species are subsets of the biota at richer sites (Ulrich and Gotelli, 2007; Gaston and Blackburn, 2000), reflecting a non-random process of species loss as a consequence of any factor that promotes the orderly disaggregation of assemblages (Qian, et al., 2004). Turnover implies the replacement of some species by other, as a consequence, of environmental sorting or spatial and historical constraints (Gianuca et al., 2017). The environmental features working as driver at local scale might have shaped the community assemblage, decreasing the species replacement between the ponds as shown by a lower turnover. As reported in Gianuca et al. (2017) a higher heterogeneity usually produces turnover patterns but in our study, from PERMDISP analysis, a difference in heterogeneity among clusters was not recorded. In this work, the ponds characterized by lower nutrient concentrations showed higher turnover. In cluster with higher value of trophic status, the component of nestedness increased, that is poorest ponds in taxa richness were subsets of the richest ponds. A nestedness pattern may highlight an internal cluster gradient of environmental features that might drive the community assemblage. Margalef (1958) gave rise to the widespread concept that the lower the level of lake eutrophication, the more complex the structure of aquatic animal communities. A general reduction in taxa richness was observed in 2015 compared to 2014, suggesting a tendency to a higher trophic status of both clusters. The higher temperatures of the water in 2015 may have favored the increase

of phytoplankton production. In 2015, a general reduction in the complexity of the community structure was observed than in 2014, this condition was highlighted by the FP-tree. In 2015, Calanoida and *Daphnia* characterized the community structure. Calanoid copepods generally appear to be best adapted to oligotrophic conditions whilst cyclopoid copepods and cladocerans are relatively more abundant in eutrophic waters (Gannon and Stemberger, 1978; Mauchline, 1998). In our study, the presence of Cladocera decrease between year, and Cyclopoida disappeared with the highest trophic status observed in 2015. Variation in community structure alters ecosystem functioning and biodiversity metrics can indicate how communities influence ecosystems (Burns et al., 2001; Iii et al., 2000; Doubek et al., 2019; Hèbert et al., 2017). Seasonal and interannual increases in *Daphnia* abundance have been associated with P limitation due to higher requirements in *Daphnia* than in other taxa (Sterner and Elser, 2002). Moreover, filter *Daphnia* species fed on the smallest food particles with a low selectivity while many species of cyclopoids show a raptorial feeding type and high selectivity preferences on much larger food items (Barnett et al., 2007). Chydorids are more successful in very productive habitats feeding by scraping algal particles from periphyton. On the contrary, *Bosmina* shows the lowest clearance rate, declines with increasing food concentrations and does not co-occur with *Daphnia* (Barnett et al., 2007). However, information about food quality and quantity as well as adaptive life his-tory strategy need to understand the mechanistic role of association rules in ecosystem functioning. In perspective, information on taxa's functional traits might be included in analysis by unsupervised machine learning.

Results from physico-chemical and biological (e.g., chlorophyll a and zooplankton communities) analysis of shallow ponds reveal large variability of all single (e.g., nitrate, pH) and aggregated (e.g., trophic status, biodiversity indexes) parameters over short temporal scales. Under such conditions, traditional statistical approach may fail to extract significant patterns or aggregations and consider them as erratic. The fuzzy logic allowed grouping ponds in clusters that differed in two consecutive years likely due to small difference in external stressors, affecting the unstable equilibrium between autotrophic and heterotrophic processes and their dominance. The latter, in turn, affect nutrient concentrations and the intensity of algal blooms, producing cascade consequences on zooplankton diversity and community composition. Year to year slight variations in water temperatures, different timing or absence of diffuse nutrient input via runoff and variable interactions with the aquifer may drive timing of blooms but also the intensity of heterotrophic microbial activities in sediments of shallow ponds. Such variations result in dynamic re-arrangement of ponds in clusters, might end up in excess nutrients sustaining algal growth or in nutrient limitation, stimulating zooplankton richness.

5. Chapter 4: Ecoacustic and sounds analysis

5.1 Make the CPUs do the hard work - Automated acoustic feature extraction and visualization for marine ecoacoustics applications illustrated using marine mammal Passive Acoustic Monitoring datasets

Abrupt changes in the ocean environment are increasing in frequency as climate change accelerates (Ainsworth et al., 2020), resulting in loss of key ecosystems (Sully et al., 2019), and shifts in endangered species' distributions (Plourde et al., 2019). Detecting such changes requires both historical and real-time (or near-real time) data made readily available to managers and decision-makers. Scientists and practitioners are being tasked with finding efficient solutions for monitoring environmental health and detecting incipient change (Gibb et al., 2019; Kowarski and Moors-Murphy, 2020). This challenge includes monitoring for changes in species' presence, abundance, distribution, and behaviour (Durette-Morin et al., 2019; Fleming et al., 2018; Root-Gutteridge et al., 2018), monitoring anthropogenic activity and disturbance levels (Gómez et al., 2018), monitoring the physical environment (Almeira and Guecha, 2019), and detecting harmful events (Rycyk et al., 2020), among others.

Environmental sounds provide a proxy to investigate ecological processes (Gibb et al., 2019; Rycyk et al., 2020), including exploring complex interactions between anthropogenic activity and biota (Erbe et al., 2019; Kunc et al., 2016). Sound provides useful information on environmental conditions and ecosystem health, allowing, for example, the rapid identification of disturbed coral reefs (Elise et al., 2019). In concert, numerous species (i.e., birds, mammals, fish, and invertebrates) rely on acoustic communication for foraging, mating and reproduction, habitat use and other ecological functions (Eftestøl et al., 2019; Kunc and Schmidt, 2019; Luo et al., 2015; Schmidt et al., 2014). Noise produced by anthropogenic activities (e.g., vehicles, stationary machinery, explosions) can interfere with such processes, affecting the health and reproductive success of multiple marine taxa (Kunc and Schmidt, 2019). In response to concerns about noise pollution, increasing effort is being invested in developing, testing, and implementing noise management measures in both terrestrial and marine environments. Consequently, Passive Acoustic Monitoring (PAM) has become a mainstream tool in biological monitoring (Gibb et al., 2019). PAM represents a set of techniques that are used for the systematic collection of acoustic recordings for environmental monitoring. It allows collecting large amounts of environmental information at multiple locations and over extended periods.

One of PAM's most common applications is in marine mammal monitoring and conservation. Marine mammals produce complex vocalizations that are species-specific (if not individually unique), and such vocalizations can be used when estimating species' distributions and habitat use (Durette-Morin et al., 2019; Kowarski and Moors-Murphy, 2020). PAM applications in marine mammal research span from the study of their vocalizations and behaviors (Madhusudhana et al., 2019; Vester et al., 2017) to assessing anthropogenic

disturbance (Nguyen Hong Duc et al., 2021). PAM datasets can reach considerable sizes, particularly when recorded at high sampling rates, and projects often rely on experts to manually inspect the acoustic recordings for the identification of sounds of interest (Nguyen Hong Duc et al., 2021). For projects involving recordings collected over multiple months at different locations, conducting a manual analysis of the entire dataset can be prohibitive, and often only a relatively small portion of the acoustic recordings is subsampled for analysis.

At its core, studying acoustic environments is a signal detection and classification problem in which a large number of spatially and temporally overlapping acoustic energy sources need to be differentiated to better understand their relative contributions to the soundscape. Such an analytical process, termed acoustic scene classification (Geiger et al., 2013), is a key step in analyzing environmental information collected by PAM recorders. Acoustic scenes can contain multiple overlapping sound sources, which generate complex combinations of acoustic events (Geiger et al., 2013). This definition overlaps with the ecoacoustics definition of soundscape (Farina and Gage, 2017), providing a bridge between the two fields, where a soundscape represents the total acoustic energy contained within an environment and consists of three intersecting sound sources: geological (i.e., geophony), biological (i.e., biophony), and anthropogenic (i.e., anthrophony). A goal of ecoacoustics is to understand how these sources interact and influence each other, with a particular focus on biological-anthropogenic acoustic interactions.

In this study, two different PAM dataset were analyzed by ML algorithms to discriminate between the vocalizations of marine mammals, beginning with high-level taxonomic groups, and extending to detecting differences among conspecifics belonging to distinct populations. Discrimination amongst different marine environments and monitoring of anthropogenic and biological sound sources were performed.

Material and Methods

Data acquisition and preparation

We collected all records available in the Watkins Marine Mammal Database website listed under the “all cuts" page. We limited the analysis to 37 marine mammal species by discarding data for species with a low number of audio samples; we processed 17.1 hours of audio. For each audio file in the WMD the associated metadata included a label for the sound sources present in the recording (biological, anthropogenic, and environmental), as well as information related to the location and date of recording. We selected audio clips that contained a marine mammal as the main and only sound source present in the recording and labelled the vocalizations according to taxonomic group (*Odontocetae, Mysticetae, Otariidae,* and *Phocidae*), order, family, and species.

We created an additional label defining the population of origin for the orca (*Orcinus orca*) samples, which split them into five groups. The first three, *EN Atlantic, WN Atlantic* and *EN Pacific*, are recordings of orcas in the wild. *EN Atlantic* samples include orcas recorded in the Norwegian Sea and in a Norwegian fjord. *WN Atlanti*c samples include orcas recorded outside St. John’s Harbour (Newfoundland, Canada) and orcas recorded approximately 130 km south of Martha’s Vineyard (Massachusetts, U.S.). The *EN Atlantic* and *WN*

Atlantic samples most likely contain a mix of two orca ecotypes (T1 and T2). *EN Pacific* samples included whales recorded in Saanich Inlet (British Columbia, Canada) and Dabob Bay (Washington, U.S.). These recordings could belong to three orca ecotypes (i.e., resident, offshore, and transient). The last two abels, *EN Atlantic – captive* and *EN Pacific - captive*, indicate recordings of captive whales Moby Doll, a resident orca captured in British Columbia, and Keiko, captured in Iceland (either a T1 or a T2 ecotype).

The Placentia Bay Database includes recordings collected by Fisheries and Oceans Canada at multiple stations within Placentia Bay (Newfoundland, Canada), from 2017 to 2020. From the PBD, we selected three days of recordings from summer 2019. The first two days (2019/08/10 and 2019/10/10) were collected by an AMAR G4 hydrophone (sensitivity: -165.02 dB re 1V/µPa at 250 Hz) deployed at 65 m of depth, approximately 13 km south of the town of Burin. The third day of recordings was collected by an AMAR G4 hydrophone (sensitivity: -164.92 dB re 1V/µPa at 250 Hz) deployed at 100 m of depth, approximately 2 km south of Red Island. Both hydrophones were set to operate following 15 min cycles, with the first 60 s sampled at 512 kHz, and the remaining 14 min sampled at 64 kHz.

For the purpose of this study, we limited the analysis to the 64 kHz recordings. From the Burin deployment, we selected the 10th of August as it contained seismic airgun noise from oil and gas exploration activity happening in the Grand Banks, approximately 170 km south of the hydrophone deployment location. From the Red Island deployment, we selected the 26th of July, which contained ship transits and humpback whale vocalizations. Before proceeding with the analysis, all recordings were labelled by time stamp and location. All days contained humpback whale vocalizations.

Acoustic feature extraction

The audio files from the WMD and PBD databases were used as input for VGGish (Abu-El-Haija et al., 2016; Simonyan and Zisserman, 2014), a CNN developed and trained to perform general acoustic classification. VGGish was trained on the Youtube8M dataset, containing more than two million user-labelled audio-video files. Rather than focusing on the final output of the model (i.e., the assigned labels), here the model was used as a feature extractor (Sethi et al., 2020). VGGish converts audio input into a semantically meaningful vector consisting of 128 features. The model returns features at multiple resolution: ~1 s (960 ms); ~1 min (59’520 ms); ~5 min (299’520 ms). All of the visualizations and results pertaining to the WMD were prepared using the finest feature resolution of ~1 s. The visualizations and results pertaining to the PBD were prepared using the ~1 min features, except for the humpback whale detection test, which was conducted on the ~1 s features.

UMAP ordination and visualization

To allow for data visualization and to reduce the 128 features to two dimensions for further analysis, we applied Uniform Manifold Approximation and Projection (UMAP) to both datasets in full and inspected the

resulting plots. UMAP is a non-linear dimensionality reduction algorithm based on the concept of topological data analysis which, unlike other dimensionality reduction techniques (e.g., tSNE), preserves both the local and global structure of multivariate datasets (McInnes et al., 2018).

The UMAP algorithm generates a low-dimensional representation of a multivariate dataset while maintaining the relationships between points in the global dataset structure (i.e., the 128 features extracted from VGGish). Each point in a UMAP plot in this paper represents an audio sample with duration of either ~ 1 sec or ~ 1 min. Each point in the two-dimensional UMAP space also represents a vector of 128 VGGish features. The nearer two points are in the plot space, the nearer the two points are in the 128-dimensional space, and thus the distance between two points in UMAP reflects the degree of similarity between two audio samples in our datasets. Areas with a high density of samples in UMAP space should, therefore, contain sounds with similar characteristics, and such similarity should decrease with increasing point distance. The visualizations and classification trials presented here illustrate how the two techniques (VGGish and UMAP) can be used together for marine ecoacoustics analysis.

Labelling sound sources

Sample labels were obtained with a mix of techniques: labels for the WMD records were obtained from the database metadata; for the PBD recordings, the start and end of seismic exploration was identified through manual inspection, ship presence was inferred from sound pressure levels (SPL) in the ship noise band (40-315 Hz)(Baldwin et al., 2021), and Humpback Whale presence was inferred using an acoustic detection model (Allen et al., 2021).

To label anthropogenic noise sources in the PBD, we first used PAM Guide (Merchant et al., 2015) to process the acoustic recordings. We computed broadband SPL (dB re 1 μPa) between 50 and 4,000 Hz (1 min resolution) as a global measure of sound pressure level in the dataset. As an indicator of ship noise, we computed the SPL between 40 and 315 Hz (i.e., ship band hereafter) at 1 min resolution. The ship band encompasses the 63,150, and 250 Hz 1/3 octave bands (Baldwin et al., 2021), which are indicators of low-frequency ship noise levels (Merchant, et al., 2014). Samples that satisfied the following two conditions were considered as ship presences: 1) the ship band SPL was within 12 dB of the broadband SPL; 2) the 5 min mean ship band SPL was 3 dB above the global median SPL (i.e., computed on the full dataset). PBD samples collected near Burin on 08/10/2019 were inspected to identify the start and end of seismic airgun activity. All 1-min samples with a time stamp falling within the period of seismic exploration were marked as airgun noise present and contained multiple blasts.

Biological noise sources in the PBD recordings were processed using the humpback whale acoustic detector created by NOAA and Google (Allen et al., 2021), providing a model score for every ~1s sample. The model returns scores ranging from 0 to 1 indicating the confidence in the predicted humpback whale presence. We used the results of this detection model to label the PBD samples according to presence of humpback whale

vocalizations. We selected 0.8 as the minimum model score needed to declare a humpback present, while every sample with a score lower than 0.8 was labelled as an absence.

Label prediction performance

To predict labels from the acoustic features for both the WMD and the PBD datasets, we applied nested k-fold cross validation to a random forest model, with ten-folds in the outer loop, and five-folds in the inner loop. We selected nested cross validation as it allows model optimization of hyperparameters and performance evaluation in a single step. Models were trained either on the two UMAP dimensions, or on the full set of 128 acoustic features, depending on model performance. Model performance was evaluated using two metrics: F1 and balanced accuracy scores, both on a scale from 0 to 1. The F1 score combines model recall and precision, favoring models with a high score in both metrics (Chinchor, 1992). Balanced accuracy is suited for measuring model performance when samples are highly imbalanced, and represents the average recall obtained for each model class (Brodersen et al., 2010). When the F1 and balanced accuracy scores indicated poor performance of the classifier, we repeated the trial using the 128 acoustic features instead of the two UMAP dimensions.

In total, we conducted 13 trials on the two databases, six on the WMD, and seven on the PBD . The first WMD trial included building a classifier for *Mysticete*, *Odontocete*, and *Pinniped*. For the remaining five trials, we created subsets of the WMD and ran classifiers for three *Mysticete* (*Balaenidae*, *Balaenopteridae*, and *schrichtiidae*) and four *Odontocete* families (*Delphinidae*, *Monodontidae*, *Phocoenidae*, and *Physteridae*); three *Balaenopteridae* species (minke, fin, and humpback whales), 14 *Delphinidae* species; and three distinct orca populations. Species with less than 100 samples were removed from the analysis.

Trials on the PBD labels proceeded as follows: i) classification of hydrophone locations (i.e., Burin and Red Island); classification of anthropogenic noise sources, including ii-iii) seismic airguns and iv-v) ships; and presence of humpback whales using vi) the two UMAP dimensions and vii) the 128 acoustic features, respectively. Presences represented a very small fraction of the PBD (<0.003 %), leading to high-class imbalance. We used two strategies to reduce class imbalance: we selected a subset of the PBD containing only hours with at least ten presences (this reduced the PBD dataset to 19 hours of PAM recordings); and then implemented a balanced random forest classifier (Lemaître et al., 2017) in place of the model used for the previous trials (Figure 51).

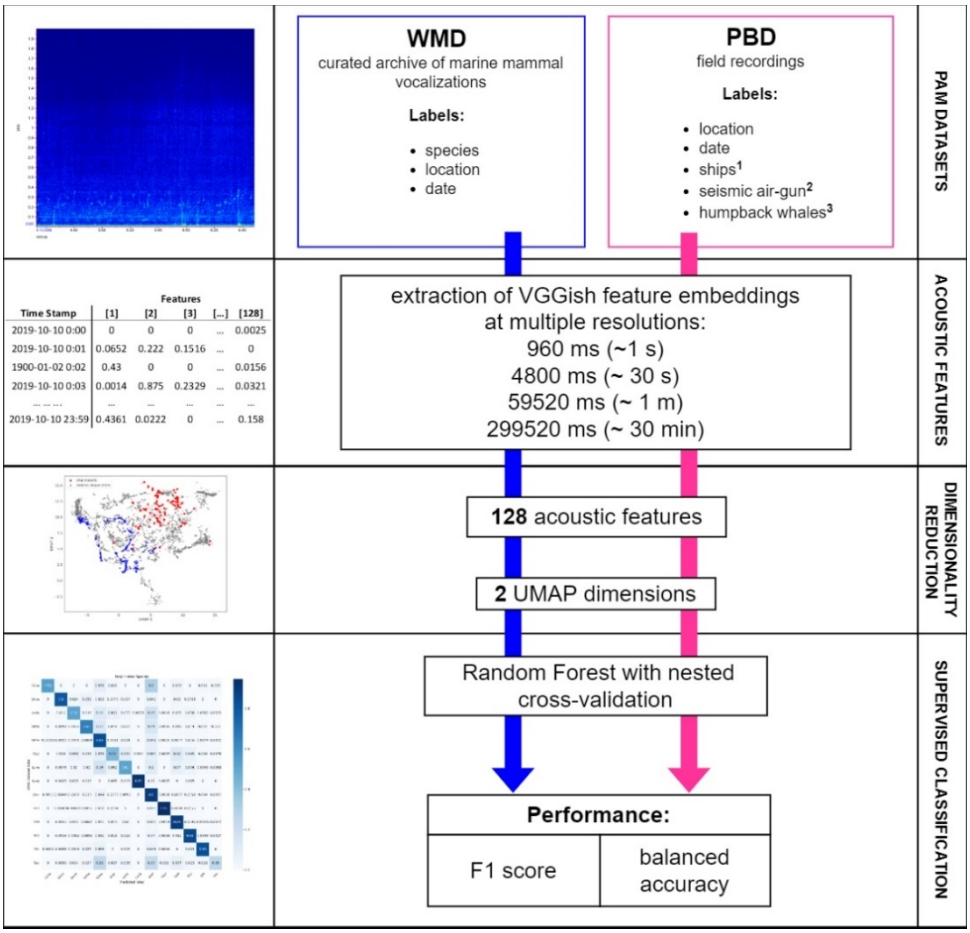

Figure 51 Analytical framework showing the different steps outlined in the Materials and Methods section. [1]labelled using ship band noise statistics; [2]labelled through visual inspection of spectrograms; [3]labelled using Google and NOAA humpback whale detector (Allen et al., 2021).

Result

Watkins Marine Mammals Sounds Database

UMAP Visualizations

Our inspection of the UMAP 2D ordination plot of three large marine mammal taxonomic groups, *Mysticete, Odontocete*, and *Pinniped*, revealed a separation between *Mysticete* and *Odontocete* sounds (Figure 52). However, the two groups overlapped in some areas of the plot, and *Pinniped* vocalization clustered close to the center of the plot, scattered between the first two groups. Within the *Mysticete* group, only three families contained enough samples to be considered for further analysis: *Balaenopteridae, Balaenidae,* and *Eschrichtiidae.* In the subsequent UMAP ordination, *Balaenidae* samples were almost completely overlapped with *Balaenopteridae* vocalizations, close to the plot centre (Figure 53).

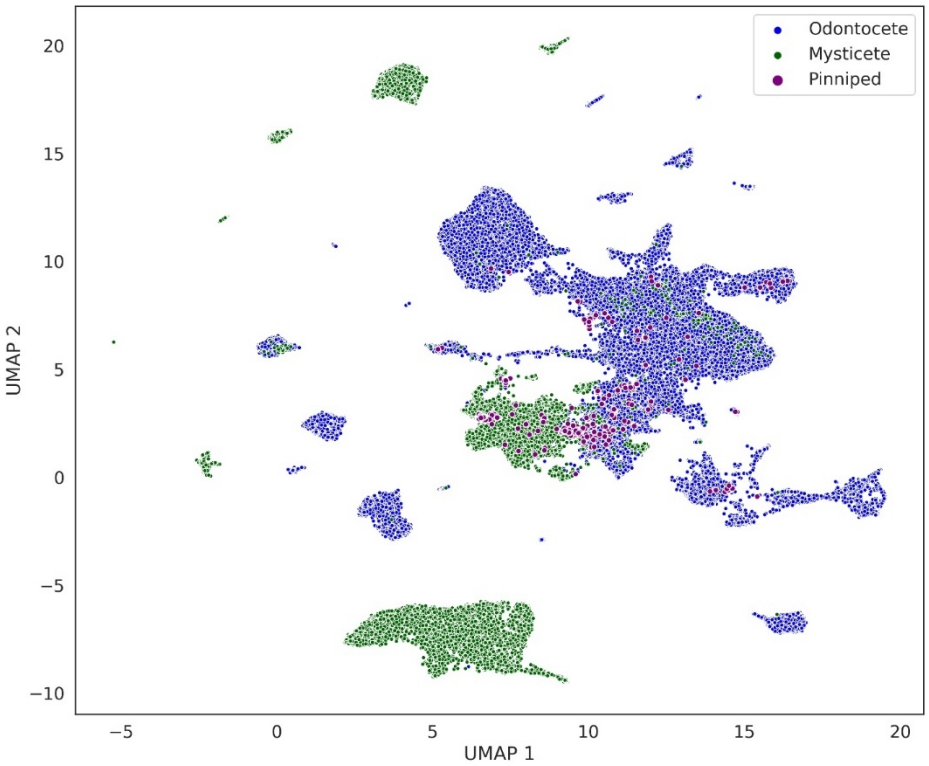

Figure 52 UMAP ordination of the WMD dataset with samples colored according to three large taxonomic groups (Mysticete, Odontocete, and Pinniped). Pinniped sample points were plotted at double size to improve visualization.

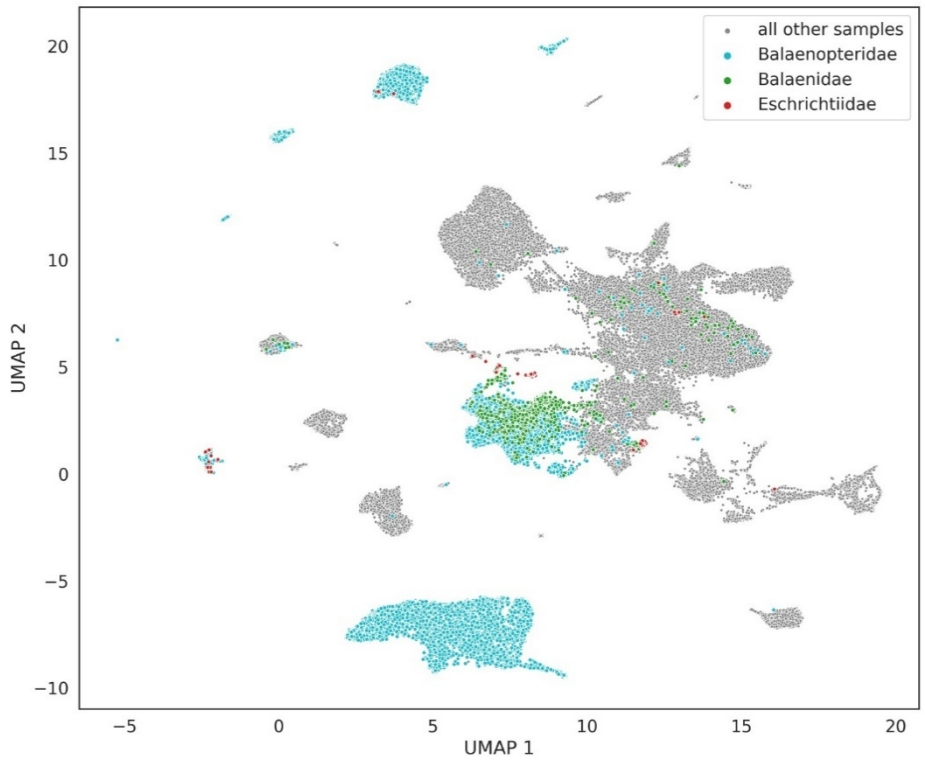

Figure 53 UMAP ordination of the WMD dataset with samples belonging to the Mysticete group colored according to three families. All other samples (Odontocete and Pinniped) are marked in grey.

Eschrichtiidae samples, the least represented label (i.e., the minority label) for the *Mysticete*, clustered in four distinct areas of the UMAP plot. The *Odontocete* group was dominated by the *Physteridae* family, which represented the majority label for the subset, followed by *Delphinidae* and *Monodontidae* (Figure 54). *Phocoenidae* vocalizations were the minority label, and, similarly to *Eschrichtiidae,* samples belonging to this family formed small clusters scattered across the UMAP plot area.

The labelled orca vocalizations (Figure 55) showed separation between four of the five population labels, apart from NE Pacific orcas, the minority class of the group. EW Atlantic was the only label whose samples formed one large and distinct cluster. Samples from the two captive orcas (EN Atlantic – captive and EN Pacific – captive), formed two distinct clusters, while the EW Atlantic samples were scattered across a large area of the UMAP plot.

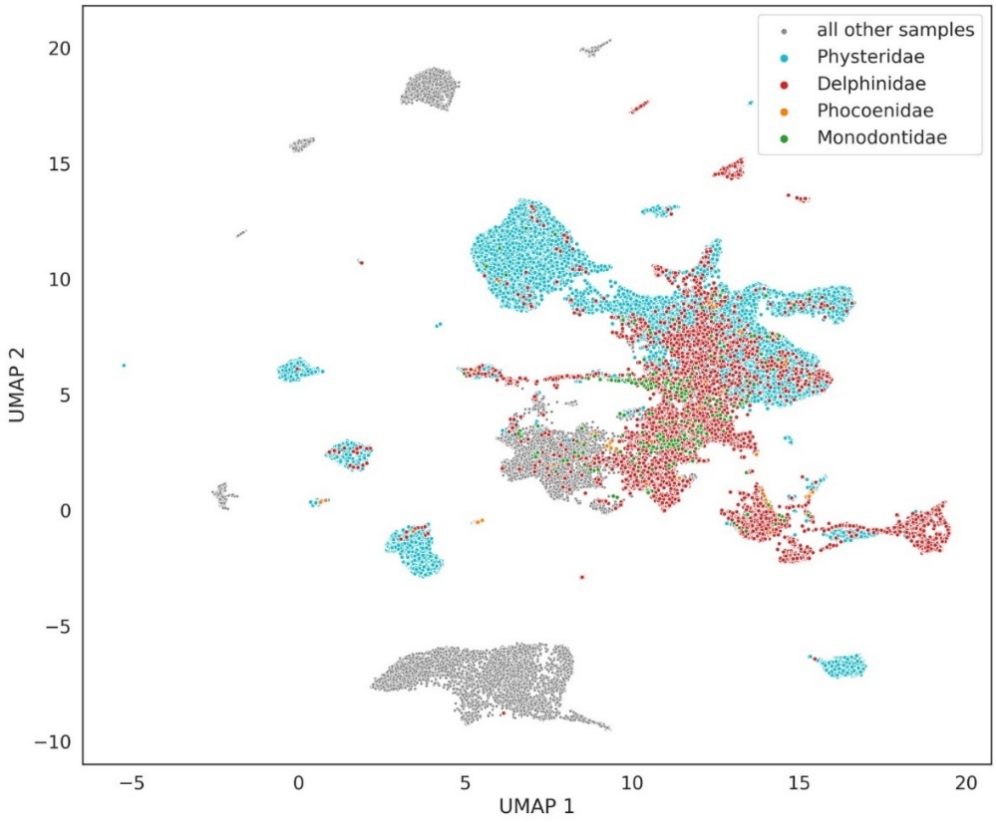

Figure 54 UMAP ordination of the WMD dataset with samples belonging to the Odontocete group colored according to four families. All other samples (Mysticete and Pinniped) are marked in grey.

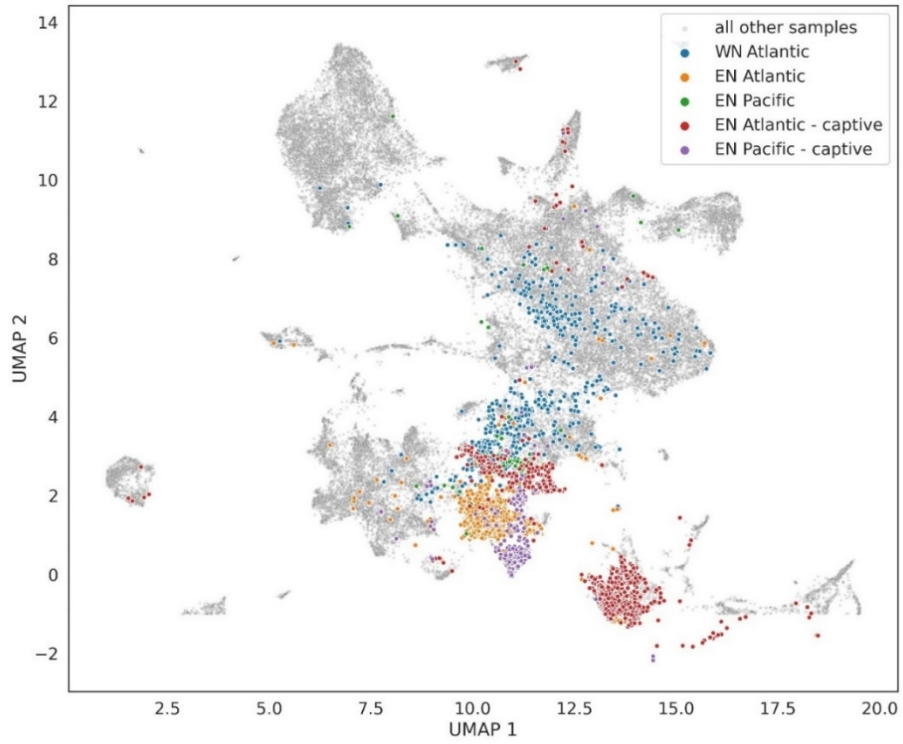

Figure 55 Detail of the WMD dataset UMAP ordination with samples belonging to *Orcinus orca*, colored according to their population of origin and wild versus captive status, when recorded. All other samples are marked in grey.

UMAP label prediction performance

Model evaluation scores were above 0.7 for all the WMD trials (Table 13), but with varying results depending on the specific label. The best classification results were obtained for Balanopteridae species (F1 = 0.998; balanced accuracy = 0.987), while the classifier built for Delphinidae species had the lowest performance (F1 = 0.829; balanced accuracy = 0.703). Classification accuracy varied across trials. For example, in the first trial, most Mysticete and Odontocete samples were correctly labelled, while 59% of the Pinniped samples were mislabeled. In the second trial, 99%, 74%, and 71% of the Balaenopteridae, Eschrichtiidae, and Balaenidae samples were correctly classified. Of the four Odontocede families, Physteridae, Delphinidae, and Phocoenidae, 99%, 90%, and 78% of the samples were correctly classified, respectively. Only 56% of the testing samples for the family Monodontidae were classified correctly.

All of the three Balaenoptera species considered in the study were correctly classified in the vast majority of cases, with scores equal or above 98% of correct predictions. Eight of the 14 Delphinidae species had 80% or more correct label predictions. Of the four labels tested for orcas, correct labels ranged from 87% (WN Atlantic) to 92% (EN Atlantic), except for the EN Pacific labels, with only 33% of the labels guessed correctly. Both model performance metrics reflected such class imbalances, with lower scores for models containing a mix of labels with low and high prediction accuracy. Balanced-accuracy scores provided a more conservative metric and were more sensitive to class imbalance than the F1 scores.

Table 13 k-fold nested cross-validation input and results. The table reports model features (X), labels (Y), and evaluation metrics (F1 score, Balanced Accuracy score). Best models, model hyperparameters, and scores per run can be found in supplementary material.

Trial	X (features)	Y (labels)	F1 score	Balanced Accuracy
1	UMAP dim	Taxonomic group	0.989	0.8
2	UMAP dim	*Mysticete* families	0.962	0.806
3	UMAP dim	Balaenopteridae species	0.998	0.987
4	UMAP dim	*Odontocete* families	0.961	0.726
5	UMAP dim	*Delphinidae* species	0.829	0.703
6	UMAP dim	*Orcinus orca* populations	0.897	0.791
7	UMAP dim	Location	0.961	0.957
8	UMAP dim	Airgun noise P/A	0.968	0.858
9	Audioset Features	Airgun noise P/A	0.987	0.917
10	UMAP dim	Ship noise P/A	0.979	0.698
11	Audioset Features	Ship noise P/A	0.99	0.859
12	UMAP dim*	Humpback whale P/A	0.594	0.623
13	Audioset Features*	Humpback whale P/A	0.999	1†

Notes: * indicates test results obtained from balanced random forest classifiers; † Only three absences were missclassified as presences, and no presences (46 in total) were misclassified as absences (See Appendix S1.J).

Placentia Bay Dataset

UMAP Visualizations

Our inspection of the UMAP ordination of the ~1 min acoustic features of the two deployment locations: Burin and Red Island revealed two overlapping clusters, with samples from Burin predominantly distributed around the edges of the Red Island cluster (Figure 56).

Samples labelled as seismic airgun noise and ship noise separated and occupied two distinct portions of the UMAP ordination plot (Figure 56). A small number of samples from the two sources overlapped, indicating ship transits occurring during seismic exploration. However, we could not observe a clear distinction between presences and absences of the sources.

Lastly, we inspected how UMAP ordinated the ~1 s acoustic features labelled by their chance of containing a humpback whale vocalization (Figure 57). Detections per hour peaked at 1:00 and 13:00 and 15:00 for the Burin samples, while the Red Island samples showed a single distinct peak at 12:00. The ~1 s resolution UMAP ordination showed a concentration of humpback whale detection scores (> 0.8) towards the right end of the plot, with samples densely aggregated along the second UMAP dimension. However, and similarly to the anthropogenic noise sources, we could not observe a clear separation between presences and absences.

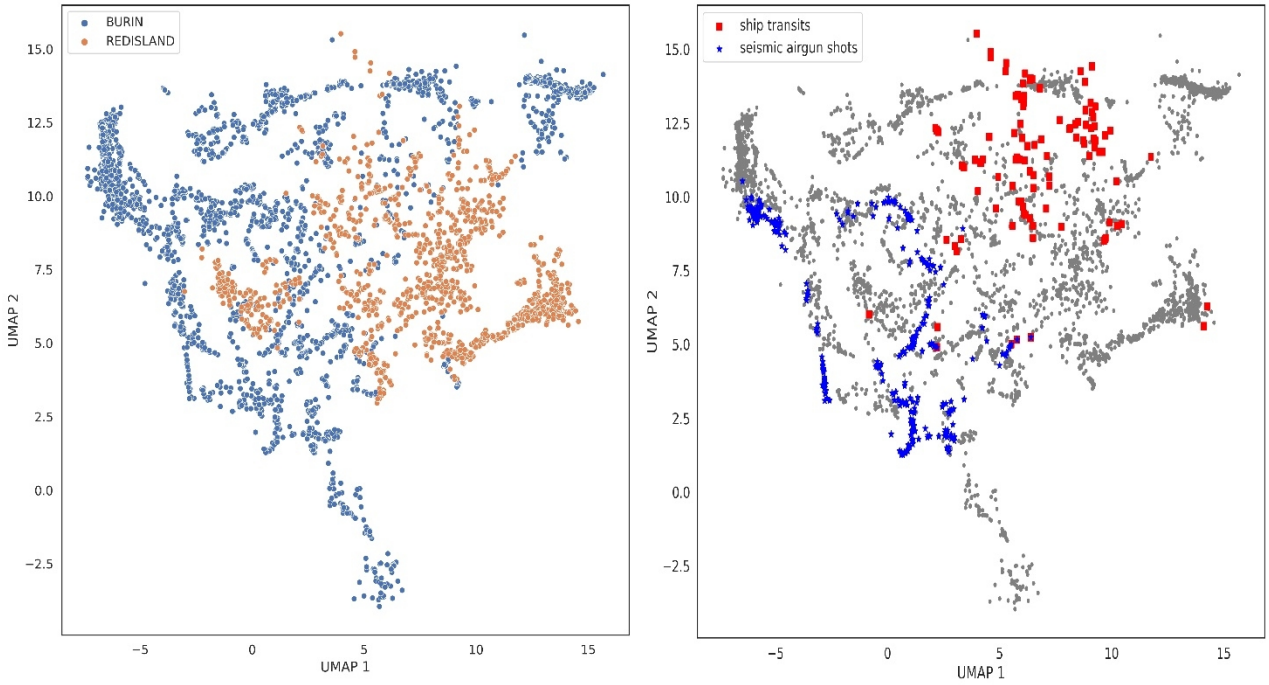

Figure 56 PBD dataset UMAP ordination at ~1 min resolution. Samples grouped by hydrophone deployment location (left). Samples grouped by sound source (right). All other samples are shown in grey.

Label prediction performance

Balanced accuracy scores for the 1-min UMAP dimensions were high (> 0.85) for the location label (Table 18SM). Of the samples labelled as 'Burin' and 'Red Island', 94% and 95% were correctly identified using the UMAP dimensions, respectively. Scores for seismic airgun presence were also high; however, model sensitivity was poor (58.3%), meaning that true positive and false negative predictions occurred with almost equal frequency. Repeating model training using the 128 acoustic features improved performance and resulted in a drop of both false negatives and false positives. The ship presence classifier trained on the two UMAP dimensions showed a balanced accuracy score of 0.7, with only 33% of samples being correctly identified as presences. The acoustic features classifier displayed a higher balanced accuracy score (0.86), and the number of correctly predicted presences, although still low, increased to 58%.

The random forest classifiers for humpback whale presence trained on the two UMAP dimensions showed the lowest F1 and balanced accuracy score (0.59 and 0.62, respectively), resulting in many mislabeled samples. Once again, repeating model fitting using the acoustic features improved model performance. Training the classifier on the 128 dimensions resulted in increased balanced accuracy score, mainly due to a dramatic increase in classifier sensitivity (93.9%) when compared to the performance of the classifier trained on UMAP dimensions (<0.001%). Confusion matrices for the WMD and PBD cross validation runs are reported in Supplementary material (Figures 9SM – 21SM, Table 17SM).

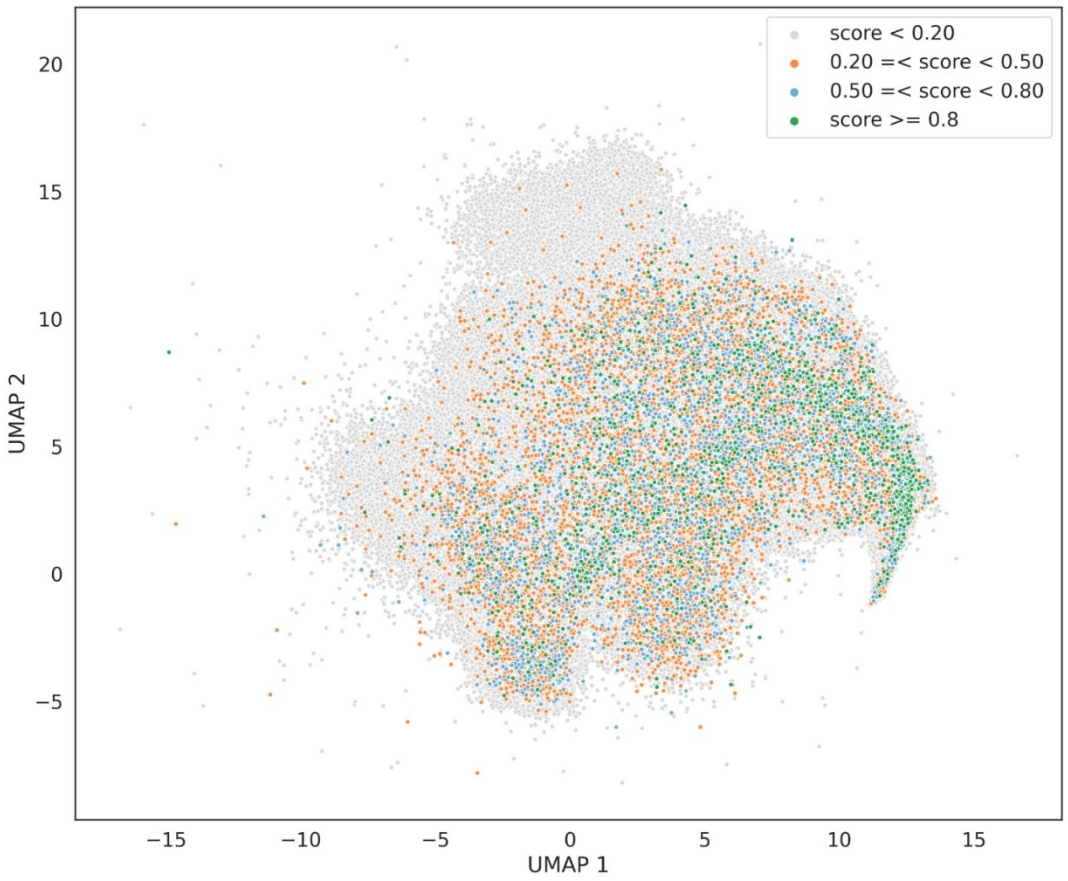

Figure 57 UMAP ordination at ~1 s resolution. Samples are coloured according to humpback whale detection probability (model scores). Scores above or equal to 0.8 were considered as presences.

Discussion

Managing the wellbeing of ecosystems requires identifying when and where human activities are affecting species' occurrence, movement, and behaviour. PAM is a useful approach for the detection of both large- and small-scale changes in urban and wild environments, as it allows for continuous and prolonged ecosystem monitoring. Challenges in employing PAM as a standard monitoring tool arise after data collection, when researchers and practitioners need to quickly extract useful information from large acoustic datasets, to understand when and where management actions are needed to preserve the well-being of ecosystems. The relatively new field of ecoacoustics provides the theoretical background for linking specific characteristics of the acoustic environment to biodiversity and ecosystem health. However, identifying a common analytical approach has been an obstacle to the broad application of ecoacoustics theory so far, and most studies employing ecoacoustics indices are not suited for replicability and comparison.

We addressed these problems by linking marine ecoacoustics assessment to the realms of machine learning and dimensionality reduction. We applied a deep-learning approach to characterize the biological and anthropogenic components of marine acoustic environments, and we illustrated how acoustic features

derived from a pre-trained Convolutional Neural Network capture both the coarse and fine-grained structure of large PAM datasets. These methods can be applied to a broad range of marine and terrestrial systems.

Our analyses revealed several applications for inferring population- and location-specific information from acoustic datasets. When datasets are already labelled and focused on a specific taxon, such as the WMD, we found that acoustic features were particularly suited for the discrimination of marine mammal vocalizations. Understanding the evolution of vocal diversity and the role of vocalizations in the ecology of a species is one of the key objectives of bioacoustics research (Luís et al., 2021). Full acoustic repertoires are not available for most species, as building comprehensive lists of vocalizations requires considerable research effort. Here we show how a general acoustic classification model (VGGish) used as a feature extractor allows us to detect differences and similarities among marine mammal species, without requiring prior knowledge on the species' vocal repertoires. Our results for orcas are of particular interest, as they provide insights on the vocal similarities and differences between distinct populations of the same species. Many orca call samples labelled as EN Pacific were classified as WN Atlantic whales using the methodology in this study. Orcas show both genetic divergence and differences in call frequency that are more pronounced for sympatric ecotypes than whales found in different ocean basins (Filatova et al., 2015). Although we cannot consider the artefactual conflation of EN Pacific orcas with NW Atlantic orcas in the WMD as definitive evidence of convergence in vocal behavior, we suggest that this aspect should be further investigated, perhaps using more recent recordings of these different orca populations.

More than 60 different ecoacoustic indices are being employed as descriptors of terrestrial soundscapes (Bradfer-Lawrence et al., 2019), making the search for indices that are successfully measuring biodiversity across widely variable environments very challenging (Minello et al., 2021). So far, ecoacoustic indices have been applied to marine environments with little success (Bohnenstiehl et al., 2018). Due to higher sound propagation efficiency, marine acoustic environments can receive acoustic energy from many sources with some that are hundreds of kilometers distant, making them more complex to study than terrestrial environments. Accordingly, the biases shown by acoustic indices measuring terrestrial species diversity (Eldridge et al., 2018; Fairbrass et al., 2017; Heath et al., 2021) are amplified when transferred to the study of marine environments (Bohnenstiehl et al., 2018; Dimoff et al., 2021; Minello et al., 2021).

Machine learned acoustic features are a promising alternative to the use of ecoacoustics indices for monitoring terrestrial biodiversity (Heath et al., 2021; Sethi et al., 2020). In this study, we show how this approach can also be extended to the study of marine soundscapes. The derived acoustic features were successful in discriminating between two different marine environments that differed in type and intensity of anthropic activity: distant seismic airgun pulses in the low frequency range dominated recordings collected in Burin, and the Red Island hydrophone recordings were characterized by frequent ship noise. Both sites yielded recordings of humpback whale vocalizations, and our results show that machine-learned acoustic features can be employed for detecting marine mammal sounds across different acoustic contexts. Machine-learned acoustic features respond to multiple marine sound sources and can be employed successfully for investigating both the biological and anthropic components of marine soundscapes.

Reducing acoustic features to two UMAP dimensions, however, resulted in poorly performing classifiers for three sets of labels: airgun noise presence, ship presence, and humpback whale presence. In all three cases, repeating the analysis on a larger set of 128 features improved model performance at the cost of increased processing time. The best models used as little as two features, and as many as 64, whereas classifiers based on the full 128 features were selected as best models for all iterations of the humpback whale classifier. This indicates that the number of acoustic features could be significantly reduced in some instances, thus reducing processing time and virtual memory requirements. The poor performance observed in the UMAP ship presence classifiers could be partly due to the approach adopted for labelling presences and to the fact that ship noise was almost ubiquitous in the Red Island recordings. Most samples collected at the Red Island deployment location were more than 3 dB higher than the full dataset median, but only a fraction of such samples contributed to the broadband SPL, indicating that ship presence may have been underestimated. As an alternative, using records of vessel positions obtained from the Automatic Identification System (AIS) as an indicator of ship presence may improve model performance, at the cost of underestimating the presence of small vessels, which are rarely equipped with AIS.

Acoustic features have been shown to overcome many of the limitations of ecoacoustics indices; for example, acoustic features outperform common ecoacoustic indices in discriminating different environmental characteristics (Sethi et al., 2020). Furthermore, acoustic features are resilient to audio file compression and reduction of Nyquist frequency and provide results that are independent from type of recorders deployed and choices relative to the temporal fragmentation of acoustic datasets (Heath et al., 2021; Sethi et al., 2020). Here, we show that acoustic features and UMAP dimensions allow for the comprehensive exploration of marine PAM datasets. Features can be used to train classification models focusing on biological and anthropogenic sound sources and allow for fine-grain comparison of marine mammal vocalizations.

Two limitations persist. VGGish, the CNN used to extract the acoustic features, is pre-trained on audio files with a sampling rate of 16 kHz, resulting in a Nyquist frequency of 8 kHz. This is sufficient to capture low frequency vocalizations but reduces its ability to discriminate high-frequency sounds. Nonetheless, we were able to correctly classify both high- and low-frequency vocalizations in the WMD examples, including Phocoenidae sounds, a family that includes species that can produce sounds up to 150 kHz. A second limitation is that acoustic features are not a plug and play product, as establishing links between features and relevant ecological variables requires additional analyses, while ecoacoustic indices are designed as measures of specific environmental characteristics.

By presenting a set of examples focused on marine mammals, we have demonstrated the benefits and challenges of implementing acoustic features as descriptors of marine acoustic environments. Our future research will extend feature extraction and testing to full PAM datasets spanning several years and inclusive of multiple hydrophone deployment locations. Other aspects warranting further investigation are how acoustic features perform when the objective is discriminating vocalizations of individuals belonging to the

same species or population, as well as their performance in identifying samples with multiple active sound sources.

Acoustic features are abstract representations of PAM recordings, which preserve the original structure and underlying relationships between the original samples, and, at the same time, are a broadly applicable set of metrices that can be used to answer ecoacoustics, ecology, and conservation questions. As such, they can help us understand how natural systems interact with, and respond to, anthropogenic pressures across multiple environments. Lastly, the universal nature of acoustic features analysis could help bridge the gap between terrestrial and marine soundscape research. This approach could deepen our understanding of natural systems by enabling multi-system environmental assessments, allowing researchers to investigate and monitor, for example, how stressor-induced changes in one system may manifest in another. In addition, these benefits accrue from an approach that is more objective than manual analyses and requires far less human effort.

6. Discussion and Conclusion

Machine Learning (ML) is an increasingly accessible discipline in computer science that develops dynamic algorithms capable of data-driven decisions. ML enables useful inferences using data collected automatically i.e. via remote sensing or other autonomous sensors or without experimental design (e.g. recording of species sightings by the public) (Lucas, 2020). ML is also used to analyze environmental data collected via social media platforms (Wäldchen and Mäder, 2018) or that has been generated synthetically via the modeling process (Chen et al., 2018). ML approaches can deal with many predictors, are robust to correlations in explanatory variables, and can allow for varying functional relationships between predictor and response variables (Hochachka et al., 2007). These features make ML well-suited to the analysis of high-dimensionality ecological complex systems. Currently, ML in ecology is mostly applied to species distribution modeling (SDM) (Elith et al. 2006) and in studies involving automatic species recognition (Tuia et al., 2022). In Chapter 1, SDM and ML were combined to describe the habitat suitability of different aquatic species: *D. longispina* and *E. serrulatus,* two out of 60 zooplankton taxa in 283 water bodies in Northern Apennines, three Mediterranean gorgonian species (soft corals) *P. clavata*, *E. cavolinii*, and *E. singularis* and two solitary corals (Scleractinia) *B. europaea* and *L. pruvotii*. A set of different ML algorithms were tested to select the best model to predict habitat suitability and species distribution in time and space. The supervised ML algorithms used were Random Forest (RF), XGboost, Artificial Neural Network (ANN), Support Vector Machine (SVM) and K-nearest neighbor (KNN). All these algorithms rely on a different strategy to solve binary classification tasks or to correctly identify suitable spatial locations for a particular species. The tested algorithms draw decision boundaries with different geometry to discriminate a binary response, and the performance of each algorithm might vary depending on the specific problem. RF and XGboost are recursive-partitioning methods (Strobl et al., 2009) and were first introduced by Breiman et al. (1984) as Classification and Regression Trees (CART). These methods use a simplified building block called decision tree, that is a numerical procedure in which several split nodes (decisions) are made using the explanatory variables as drivers and a cost function as meaures of correct partitioning. The RF and XGboost are both ensemble models of many decision trees. The RF uses the bagging strategy while the XGboost uses the boosting strategy. The SVM is a kernel-based algorithm that transforms data into a high-dimensional space and constructs a hyperplane that maximizes the distance to the nearest data point of any of the input classes, while the ANNs are based on networks of computing units called neurons, connected with synapsis. Finally, the KNN classifies a new data point into the target class, depending on the features of its neighboring data points, and differently from the previous algorithms is known as “lazy learner” because it does not learn a discriminative function from the training data but “memorizes” the training dataset instead. For the gorgonian and coral species, although all tested algorithms showed high performances, the XGBoost was selected as the best to model distributions of *P. clavata*, *E. cavolinii*, *E. singularis*, *B. europaea*, and *L. pruvotii*. This result agrees with studies that report the highest performance reached by the gradient boosting

method in different applications (Li et al., 2019; Osman et al., 2021; Kumar and Kumar 2021; Pandeyz et al., 2021).

ML algorithms improve data inferences with respect to several traditional statistical methods (Lucas, 2020). In our study, Chapter 1, the supervised ML algorithm ANN reached the highest values of the performance metrics in assessing the factors that influenced the distribution patterns, presence or absence for *D. longispina* and *E. serrulatus* and outperforms the classical generalized linear model (GLM) based on maximum likelihood estimation. ANN's achieved also good results in studies that model the habitat suitability of spawning European grayling (*Thymallus thymallus*) and the species distributional range of a Carpathian endemic plant (*Leucanthemum rotundifolium*) (Fukuda et al., 2013; McKenna and Kocovsky, 2020). Another example is reported in Chapter 2 and refers to the recognition of sibling malaric mosquiotes species (*Anopheles*): the SVM algorithms overcome the classical discriminant analysis in the correct classification of *An. maculipennis* s. s. and *An. daciae* sp. inq. The improvement of data representation was reached with the UMAP unsupervised approach, which separates better than the PCA the four *Anopheles* sibling species using wing morphology information. In Chapter 2, we demonstrated that geometric morphometrics combined with ML algorithms are a useful tool to deepen the analysis of inter and intra-specific shape variability and to evaluate evolutionary constraints related to wing functionality. An alternative to the use of geometric morphometrics is the application of computer vision algorithms (CNN), which might speed up the process of recognition and monitoring. Kittichai et al. (2021) used a deep learning algorithm (YOLO) to localize and classify simultaneously the images of 13 different species and gender of mosquitoes, reaching a mean precision and sensitivity of 99% and 92.4%, respectively. Ong et al. (2021) developed a device equipped with a microcomputer and a camera module to classify two species of the genus *Aedes* spp. with CNN and achieved an accuracy of more than 98%, which is not statistically different from human experts' recognition.

The unsupervised machine learning approach was applied in Chapter 3, in a case study involving community ecology. Unsupervised machine learning algorithms, fuzzy c-means, and association rules mining were applied to assess the factors influencing the assemblage composition and distribution patterns of 12 zooplankton taxa in 24 shallow ponds in Northern Italy. Data retrieved during 2014 and 2015, were compared, taking into account that 2014 late spring and summer air temperatures were much lower than historical records, whereas 2015 mean monthly air temperatures were much warmer than historical averages. In both years, fuzzy c-means show a strong clustering of ponds in two groups, contrasting sites characterized by different physico-chemical and biological features. Climatic anomalies, affecting the temperature regime, together with the main water supply to shallow ponds (e.g., surface runoff vs. groundwater), represent disturbance factors producing large interannual differences in the chemistry, biology, and short-term dynamic of small aquatic ecosystems. Unsupervised machine learning algorithms and fuzzy sets allowed to catch such apparently erratic differences as well as in the case of mosquito wing morphology captured by HDBSCAN (Chapter 2). The explained morphological variation within *Anopheles* species gives interesting

results in the framework of phenotypic plasticity which is the ability of a genotype to produce different phenotypes in response to stimuli or input from the environment (Sommer, 2020).

ML algorithms are considered black boxes because their mechanisms of making decisions are not explicitly accessible to human cognition and their results are less easy to interpret than those obtained by traditional statistical models (e.g. regression-based methods) (Guidotti et al., 2018; Sheu, 2020). The great predictive accuracy of ML algorithms allows for discovering relationships not hypothesized a priori and represents a new way to make inferences. The interpretation of ML results needs to be done carefully and requires user interpretation that is not generally required in standard statistical models. ML tend to explore and generate hypothesis while more robust statistical methods are needed to formally test most hypotheses (Lucas et al., 2020). However, to "open the ML black box" and to understand the importance of variables and the whole result's interpretability several methods are available (Cha et al., 2021). Global sensitivity and uncertainty analysis (GSUA) we used to understand the behavior and the relationships among explanatory variables in the spatial distribution models of freshwater zooplankton and gorgonians (Chapter 2), is a set of statistical tools that make perturbations of a particular model, using Monte Carlo experiment and variance decomposition (Saltelli et al., 2000; Bellin et al., 2020; 2021). The SHAP analysis used in the case of zooplankton is a general method applicable to all type of ML algorithms and allow us to rank the importance of variables and their interactions (Lundberg and Lee, 2017).

At present, supervised approaches are mainly used in ecology, due to their high performance and the simple training procedures (LeCun et al., 2015). Future applications and research might rely on the use of novel unsupervised methods, removing the need for an annotated dataset, or using the reinforcement learning approach to study dynamically the ecosystems and ecological networks at different spatio-temporal scales. DL algorithms show a lot of promise for ecological and evolutionary data improvement and might be able to cover a large set of ecological questions. In Chapter 4, an example of a DL application based on a pre-trained deep neural network (VGGish) combined with dimensionality reduction and RF algorithm was used in the analysis of marine acoustic data. Acoustic features and their UMAP projections exhibited good performance in the classification of marine mammal vocalizations. Most of the taxonomic levels investigated here could be classified using the UMAP projections, apart from underrepresented species. Both anthropogenic (ships and airguns) and biological (humpback whales) sound sources could be identified in field recordings. The acoustic feature extraction, visualization, and analysis allow the retention of most of the environmental information contained in PAM recordings, overcoming the limitations encountered when using ecoacoustics indices. Acoustic features extracted from a pre-trained CNN (VGGish) are universal, permitting comparisons of results collected from multiple environments. This approach can be used to simultaneously investigate the macro and micro characteristics of marine soundscapes, with a more objective method and with far less human effort.

Ecologists in the new era of data science should have access to good programming skills and mathematical tools to deal with complexity. Stronger collaboration between computer scientists, informatics, and ecologists could provide new tools and methods in both applied and theoretical research. I strongly

encourage data sharing and free AI programs for scientists, to improve and speed up scientific discoveries. At present, the great global challenges at the level of nature conservation, biodiversity loss due to anthropogenic effects, global changes, vector epidemiological monitoring, and sustainability are complex problems that require fast and accurate real-time analysis with suitable statistical tools. ML has the potential to address many of these requirements.

7. Acknowledgements

I would like to thank Nicola Marchesani and Giacomo Tesi from University of Parma for the first study reported in Chapter 1 and published in Bellin et al. (2022a), specifically to make the data digitization and for the help given during the work development.

I would like to thank prof. Silvia Franzellitti, prof. Erik Caroselli, prof. Stefano Goffredo and Dr. Rachele Spezzano from University of Bologna for the revision of discussion and the advice regarding the species distribution and physiology of *Balanophyllia europaea* and *Leptopsammia pruvoti*.

I would like to thank prof. Marco Bartoli, Dr. Erica Racchetti and Dr. Catia Maurone from University of Parma for the data collection and expertise contribution carried out in Chapter 2 and published in Bellin et al. (2021a).

Many thanks to Dr. Mattia Calzolari, Dr. Emanuele Callegari, Dr. Davide Lelli, Dr. Paolo Bonilauri and Dr. Michele Dottori from Istituto Zooprofilattico Sperimentale della Lombardia e dell'Emila Romagna (Bruno Ubertini) by involving me in the data analysis of samples obtained from West Nile Virus monitoring campaigns, in the study reported in Chapter 3 and published in Bellin et al. (2021b) and Bellin et al. (2022b).

I would like to thank prof. Matteo Montagna and Dr. Giulia Magoga from University of Milano for the genetics analysis and for the advice in the evolutionary biology and genetics of Diptera.

Finally, I would like to thank Dr. Simone Cominelli, proff. Carissa Brown and Dr. Jack Lawson from Memorial University of Newfoundland (Canada) for the data collection and the precious advice in ecoacustic, in the case of study reported in Chapter 4, this collaboration led to the production of a manuscript (Cominelli et al., under review).

Many thanks to Dr. Marco Corradi and Dr. Luca Dallacasagrande for the valuable cooperation respect to the design of new future projects in Machine Learning applied to ecology.

Many thanks to prof. Stefano Allesina from University of Chicago (US) and prof. Duccio Rocchini from University of Bologna for agreeing the role of outside-examiners of my PhD thesis.

Published Literature

Bellin, N., Racchetti, E., Maurone, C., Bartoli, M., Rossi, V., 2021a. Unsupervised Machine Learning and Data Mining Procedures Reveal Short Term, Climate Driven Patterns Linking Physico-Chemical Features and Zooplankton Diversity in Small Ponds. Water, 13(9), 1217.

Bellin, N., Calzolari, M., Callegari, E., Bonilauri, P., Grisendi, A., Dottori, M., Rossi, V., 2021b. Geometric morphometrics and machine learning as tools for the identification of sibling mosquito species of the Maculipennis complex (Anopheles). Infection, Genetics and Evolution, 95(1), 105034.

Bellin, N., Calzolari, M., Magoga, G., Callegari, E., Bonilauri, P., Lelli, D., Dottori, M., Montagna, M., Rossi, V., 2022b. Unsupervised machine learning and geometric morphometrics as tools for the identification of inter and intraspecific variations in the Anopheles Maculipennis complex. Acta tropica, 233, 106585.

Bellin, N., Tesi, G., Marchesani, N., and Rossi, V., 2022a. Species distribution modeling and machine learning in assessing the potential distribution of freshwater zooplankton in Northern Italy. Ecological Informatics, 69, 101682.

Cominelli, S., Bellin, N., Brown, C., Rossi, V., Lawson, J. Make the CPUs do the hard work - Automated acoustic feature extraction and visualization for marine ecoacoustics applications illustrated using marine mammal Passive Acoustic Monitoring datasets. Ecology and Evolution (Under Review).

Bellin, N., Rossi, V. Modelling climate change impacts on the habitat suitability of Mediterranean gorgonians. Marine Ecology Progress Series (Under Review).

Other Publications

Bellin, N., Groppi, M., Rossi, V., 2020. A model of egg bank dynamics in ephemeral ponds. Ecological Modelling, 430, 109126.

Bellin, N., Spezzano, R., Rossi, V., 2021c. Assessing the Extinction Risk of *Heterocypris incongruens* (Crustacea: Ostracoda) in Climate Change with Sensitivity and Uncertainty Analysis. Water, 13(13), 1828.

8. Supplementary Material

Chapter 1: Species distribution models

Species distribution modeling and machine learning in assessing the potential distribution of freshwater zooplankton in Northern Italy

Table 1SM The VIF values of the continuous environmental variables. Environmental variables with VIF values >=10 were reported in bold and removed from the modeling framework.

Type of Variable	Variable	VIF
Climatic	Isothermality (Bio 3)	1.59
	Temperature seasonality (Bio 4)	4.23
	Mean temperature of the Wettest quarter (Bio 8)	4.19
	Mean temperature of the Driest quarter (Bio 9)	1.84
	Mean temperature of the Warmest quarter (Bio 10)	**>10**
	Mean temperature of the Coldest quarter (Bio 11)	5.13
	Annual Precipitation (Bio 12)	**>10**
	Precipitation seasonality (Bio 15)	3.05
	Precipitation of the Wettest quarter (Bio 16)	**>10**
	Precipitation of the Driest quarter (Bio 17)	**>10**
	Precipitation of the Warmest quarter (Bio 18)	3.42
	Precipitation of the Coldest quarter (Bio 19)	**>10**
	Annual Potential Evapotranspiration (PET) (Envirem 1)	**>10**
	Aridity Index Thornthwaite (Envirem 2)	3.40
	PET of the Wettest quarter (Envirem 11)	1.62
	PET of the Driest quarter (Envirem 12)	1.80

	PET of the Warmest quarter (Envirem 14)	**>10**
	PET of the Coldest quarter (Envirem 15)	**>10**
Hydrological	Drainage direction	1.07
	Flow accumulation	1.02
Soil Properties	Carbon content	3.35
	Bulk Density	2.11
	pH	1.96
	Nitrogen	3.69

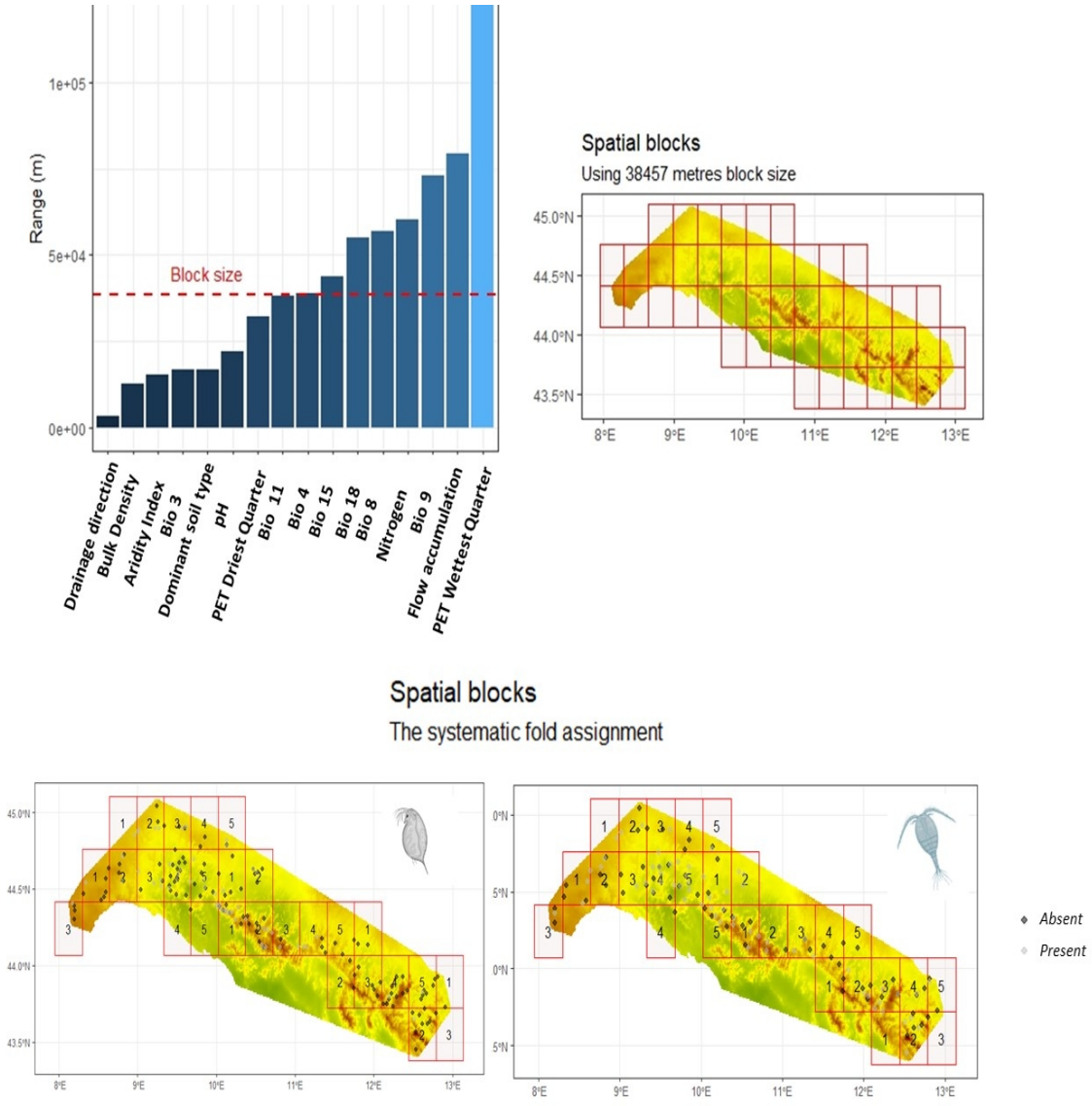

Figure 1SM. Top left panel showed the spatial autocorrelation of the continuous environmental variables and the block size in meters (horizontal red line) considering the mean of the spatial autocorrelation values. The top right panel showed the spatial distribution of the blocks in the study area. The bottom left - right panels showed the systematic assignment of the fold using the information of the spatial blocks and the presence/absence values of both species (*Daphnia longispina* on the left and *Eucyclops serrulatus* on the right).

Table 2SM. For each species (*D. longispina* and *E. serrulatus*) and each model (GLM, RF and ANN) the mean values of the four performance metrics (Pcc, Tss, Kappa and Auc) were reported after the block cross validation. The highest mean value of each performance metric was reported in bold and the standard deviation in brackets.

Model	*Daphnia longispina*				*Eucyclops serrulatus*			
	Pcc	**Tss**	**Kappa**	**Auc**	**Pcc**	**Tss**	**Kappa**	**Auc**
GLM	0.66 (0.08)	0.17 (0.12)	0.19 (0.13)	0.65 (0.09)	0.55 (0.083)	0.10 (0.24)	0.12 (0.16)	0.60 (0.06)
RF	0.66 (0.04)	0.12 (0.11)	0.14 (0.12)	0.61 (0.09)	0.47 (0,092)	-0.15 (0.32)	-0.05 (0.20)	0.59 (0.05)
ANN	**0.69** (0.07)	**0.22** (0.12)	**0.24** (0.13)	**0.70** (0.08)	**0.65** (0.049)	**0.32** (0.23)	**0.29** (0.11)	**0.72** (0.10)

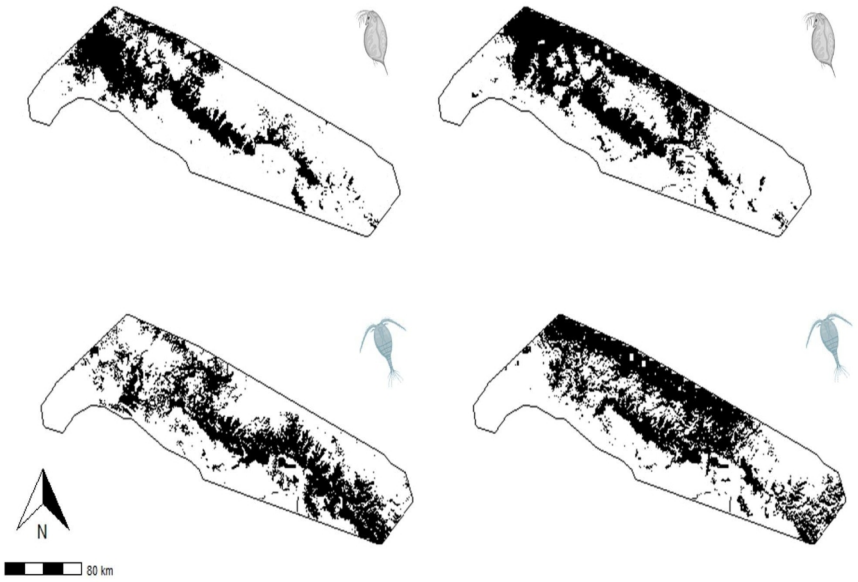

Figure 2SM. Spatial distribution for *Daphnia longispina* and *Eucyclops serrulatus*. The left panels referred to the past climatic conditions, while the right panels to the future climatic condition.

Table 3SM Environmental variables with VIF values <= 4 were retained in the modelling framework. The environmental variables that showed multicollinearity (VIF > 4) for the benthic layers were: temperature, nitrate, phosphate, dissolved oxygen, phytoplankton; and for the surface layers: salinity, diffuse attenuation, and PAR.

Type	**Variables**	**VIF values <= 4**
Chemico-physical surface layers	Temperature	2.003443
	Nitrate	1.326016
	Phosphate	2.607566
	Silicate	2.439113
	Current velocity	2.542855
	pH	3.452158
	Calcite	1.410136
Chemico-physical benthic layers (average depth)	Salinity	2.397499
	Silicate	2.515781
	Current velocity	1.648047
	Light at bottom	1.706335
Geophysical	Bathymetry	2.348911
	Concavity	1.438825
	Curvature	1.468548
	E-W Aspect	1.040252

Table 4SM For each environmental variable the result of the VIF analysis was reported. Environmental variables with VIF values greater than the threshold of 10 were reported in bold.

Environmental Variable	***VIF***
Bathymetry	1.703871
Calcite	2.356242
Current Velocity	2.125708
Diffuse attenuation	2.856209
Dissolved oxygen	**>10**
Nitrate	2.436228
Photosynthetic available radiation	8.121243
pH	3.847405
Phosphate	5.312007
Phytoplankton	2.830844
Salinity	6.795321
Silicate	1.802115
Sea Surface Temperature	6.559845

Table 5SM For each species and environmental variable, the p-value of Kolgomorov-Smirnov test relative to each fold was reported.

Species	***Environmental Variables***	***Kolgomorov-Smirnov p-value***							
		FOLD							
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
B. europaea	Sea Surface Temperature	0.854	0.399	0.459	0.969	0.436	0.857	0.934	0.924
	Salinity	0.934	0.822	0.244	0.973	0.862	0.441	0.728	0.920
	pH	0.761	0.866	0.695	0.929	0.894	0.806	0.932	0.941
	Calcite	0.917	0.634	0.585	0.475	0.875	0.701	0.632	0.999
	Bathymetry	0.994	0.844	0.768	0.728	0.763	0.844	0.831	0.887
	Nitrate	0.849	0.656	0.682	0.703	0.993	0.910	0.996	0.946
	Phospate	0.798	0.997	0.971	0.595	0.494	0.732	0.306	1.000
	Phytoplankton	0.994	0.913	0.664	0.346	0.987	0.640	0.935	0.999
	Current velocity	0.894	0.950	0.914	0.826	0.672	0.872	0.852	0.831
	Diffuse attenuation	0.776	0.753	0.378	0.448	0.916	0.611	0.961	1.000
	Photosynthetic available radiation	0.748	0.478	0.487	0.751	0.946	0.829	0.851	0.940
	Silicate	0.340	0.749	0.820	0.619	0.998	0.134	0.841	0.917
	Sea Surface Temperature	0.951	0.483	0.754	0.183	0.800	0.744	0.906	0.993
	Salinity	0.883	0.998	0.737	0.992	0.281	0.945	0.836	0.797
	pH	0.871	0.153	0.236	0.780	0.563	0.229	0.992	0.701
	Calcite	0.985	0.556	0.410	0.707	0.670	0.052	0.975	0.653

L. pruvoti	Bathymetry	0.992	0.902	0.972	0.704	0.841	0.549	0.907	0.901
	Nitrate	0.672	0.532	0.807	0.413	0.214	0.143	0.888	0.837
	Phospate	0.919	0.671	0.778	0.965	0.607	0.951	0.922	0.998
	Phytoplankton	0.543	0.168	0.519	0.461	0.100	0.235	0.838	0.861
	Current velocity	0.975	0.586	0.597	0.539	0.670	0.602	0.990	0.808
	Diffuse attenuation	0.997	0.517	0.556	0.239	0.454	0.113	0.758	0.945
	Photosynthetic available radiation	0.971	0.406	0.220	0.761	0.879	0.573	0.839	0.661
	Silicate	0.991	0.137	0.960	0.637	0.742	0.491	0.922	0.969

Chapter 2: Geometric Morphometric

2.1 Supervised and Unsupervised machine learning combined with geometric morphometrics as tools for the identification of inter and intraspecific variations in the Anopheles Maculipennis complex

Table 6SM. Showed the pairwise comparison between species (p- value adjusted) using Wilcoxon rank sum test. The asterisks were referred to the significant levels (* $p < 0.05$, ** $p < 0.001$, *** $p < 0.0001$).

Pairwise comparison			
	An. maculipennis s. s.	*An. daciae* sp. inq.	*An. atroparvus*
An. daciae sp. inq.	0.85	-	-
An. atroparvus	0.00045 **	0.00037 **	-
An. melanoon	0.84	0.85	0.072

Table 7SM. Showed the results of the MANCOVA analysis. It were reported the variables, the interaction terms and the residuals, the sum of squares (SS), the degree of freedom (df), the R-squared (R^2), Statistics (F) and the p-value obtained after the permutation procedure (n = 1000). The asterisks were referred to the significant levels (* $p < 0.05$, ** $p < 0.001$, *** $p < 0.0001$).

Variables	**SS**	**df**	**R^2**	**F**	**p-value**
Species	0.015	1	0.029	14.89	0.001***
log(CS)	0.028	3	0.055	9.21	0.001***
Species X Log(CS)	0.0033	3	0.006	1.09	0.32
Residuals	0.46	452	0.90	-	-

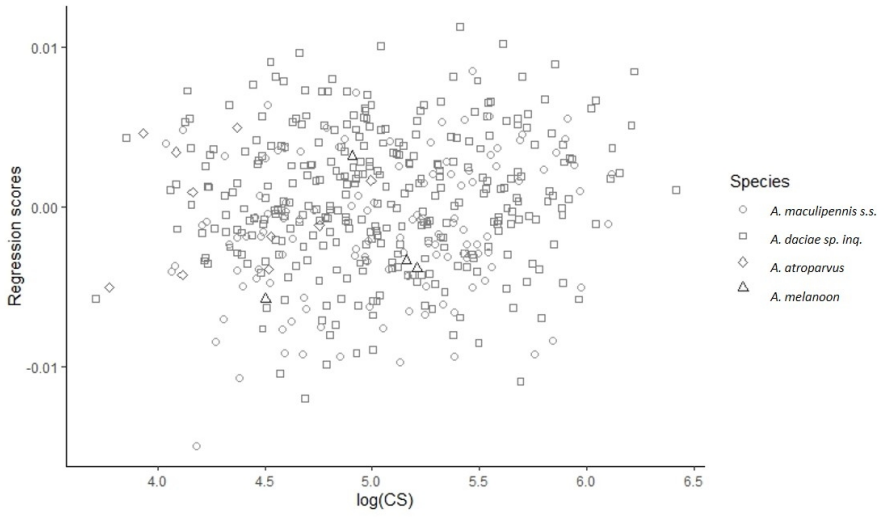

Figure 3SM Scatterplot of the regression scores obteined by MANCOVA analysis. Regression scores are standardized projected shape scores, along the axis defined by the regression of shape on size.

Table 8SM Showed the results of the permutation test considering the Euclidean distances between pairs of species. The asterisks were referred to the significant levels (* $p < 0.05$, ** $p < 0.001$, *** $p < 0.0001$).

Species	Euclidean distance	p -value
An. maculipennis s. s. - *An. daciae* sp. inq.	0.013	0.001 ***
An. maculipennis s. s – *An. atroparvus*	0.023	0.001 ***
An. maculipennis s. s – *An. melanoon*	0.023	0.033 *
An. daciae sp. inq. – *An. atroparvus*	0.031	0.001 **
An. daciae sp. inq. – *An. melanoon*	0.030	0.002 *
An. atroparvus – *An. melanoon*	0.038	0.001 **

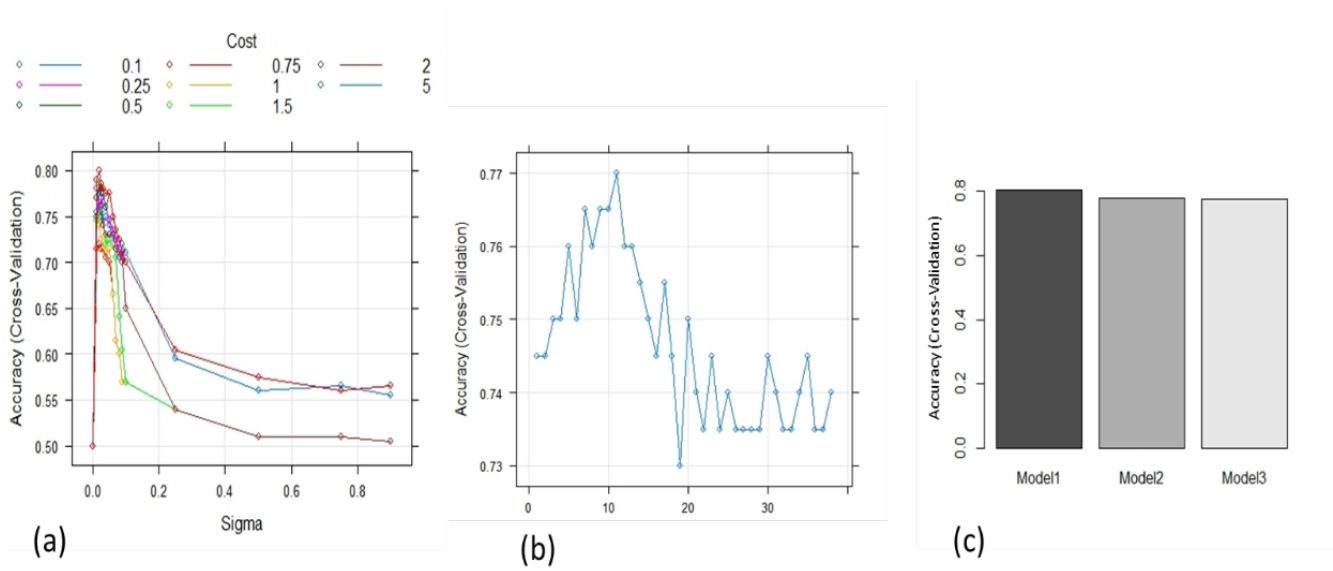

Figure 4SM. Mean validation accuracy computed for each combination of the hyperparameters of SVM (a), RF (b) and for three different ANN architectures (c).

Table 9SM PERMANOVA test of shapes differences among haplotypes with 999 random permutations. For the species *An. daciae* sp. inq. the degree of freedom (Df), the sum of squares (SS), the means squared error (MS), the coefficient of determination of the test (Rsq), the F statistic, the effect sizes (Z) and the p-value of the test.

		Df	SS	MS	Rsq	F	Z	p-value
***An. daciae* sp. inq.**	Haplotype	4	0.0041 8	0.00105	0.0768	1.144 8	0.633	0.263
	Residual	55	0.0502	0.000913	0.923			
	Total	59	0.0544					

Table 10SM PERMANOVA test of shapes differences among morphological clusters with 999 random permutations. For each species the degree of freedom (Df), the sum of squares (SS), the means squared error (MS), the coefficient of determination of the test (Rsq), the F statistic, the effect sizes (Z) and the p-value of the test were reported.

		Df	SS	MS	Rsq	F	Z	p-value
***An. daciae* sp. inq.**	Groups	11	0.0915	0.00832	0.415	13.5	18.5	0.001 **
	Residual	210	0.128	0.000613	0.584			
	Total	221	0.220					
***An. maculipennis* s. s.**	Groups	3	0.0264	0.00880	0.257	10.5	8.49	0.001 **
	Residual	91	0.0760	0.000835	0.742			
	Total	94	0.102					

Table 11SM. GLMM model result to test intraspecific spatial-temporal differences in morphotype abundance. For each species the degree of freedom (Df), the sum of squares (SS), the means squared error (MS) and the p-value were reported.

	Factors	Df	SS	MS	p value
***An. daciae* sp. inq.**	morphotype	11	6.37	0.58	0.58
	capture techniques	1	0.40	0.40	0.40
	morphotype: capture techniques	11	4.17	0.38	0.38
***An. maculipennis* s. s.**	morphotype	3	1.48	0.49	0.49
	capture techniques	1	0.91	0.91	0.91
	morphotype: capture techniques	3	0.19	0.06	0.06

Chapter 3: Community Ecology

Unsupervised Machine Learning and Data Mining Procedures Reveal Short Term, Climate Driven Patterns Linking Physico-Chemical Features and Zooplankton Diversity in Small Ponds

Table 12SM Reported the latitude and longitude in WGS84 coordinate reference system of the 24 ponds under study, the number of years since origin.

Id	Ponds	Age	Latitude	Longitude
1	Pastore3	218	10.324003	45.000660
2	Pastore1	69	10.320438	45.000116
3	Pastore4	218	10.328615	45.000381
4	Temporanea	5	10.322755	45.003484
5	Bosco Braca	69	10.388445	44.994899
6	Pavarini	47	10.391577	44.997471
7	Bosco Valloni	218	10.355710	44.985057
8	San Giorgio	219	10.365307	44.987526
9	Forche	218	10.358832	45.000701
10	Martignana	219	10.363053	44.991424
11	Santa Maria Maddalena	300	10.329035	45.014149
12	Bosco Bodini	46	10.313195	44.995203
13	Cascina Mortara	35	10.313565	44.999700
14	Bosco Piazza	218	10.305016	45.002634
15	Cascina Tavernelle	300	10.309694	45.014147
16	Vecchio	300	10.288462	45.014119
17	Bazzi	64	10.284595	45.011462
18	Motta	219	10.256552	45.052146
19	Ronchetto	218	10.240978	45.033728
20	Rita	300	10.242026	45.052845
21	Bicocca	300	10.228081	45.048045
22	Pescaroli West	218	10.194924	45.046146
23	Pescaroli East	218	10.196969	45.044955
24	Sabbie	300	10.327130	45.006925

Table 13SM Showed the physical and chemical environmental features measured for each pond, the unity of measure, the symbols adopted in the study and the laboratory assay or the method of estimation.

Environmental feature	Type	Unity of measure	Symbol	Laboratory Assay – Method of estimation
Water temperature	Physical	Celsius (°C)	wT	Multiparameters probe (YSI model 566 MPS)
pH	Chemical	dimensionless	pH	Multiparameters probe (YSI model 566 MPS)
Oxygen	Chemical	milligrams per liter ($mg\ l^{-1}$)	O_2	Multiparameters probe (YSI model 566 MPS)
Conductivity	Chemical	microsiemens per meter ($\mu S.m^{-1}$)	ES	Multiparameters probe (YSI model 566 MPS)
Dissolved inorganic carbon	Chemical	millimolars (mM)	DIC	Anderson et al., 1986
Ammonia	Chemical	milligrams per liter ($mg.l^{-1}$)	NH_4+	A.P.H.A, 1981
Soluble reactive phosphorus	Chemical	milligrams per liter ($mg.l^{-1}$)	SRP	Valderrama et al., 1977
Nitrate	Chemical	milligrams per liter ($mg.l^{-1}$)	NO_3-	Rodier, 1987
Chlorophyll a	Chemical	micrograms per liter ($\mu g.\ l^{-1}$)	Chla	A.P.H.A, 1981
Reactive silica	Chemical	milligrams per liter ($mg.l^{-1}$)	SiO_2	A.P.H.A, 1981
Depth	Physical	meters (m)	Depth	D'Auria and Zavagno, 1999
Perimeter	Physical	meters (m)	Perimeter	D'Auria and Zavagno, 1999

Table 14SM Descriptive statical parameters: Range, Mean, Median and Standard deviation (SD) of the chemico-physical enviromental features, in 2014 and 2015. Water temperature (wT) was expressed in Celsius (°C); Oxygen (O_2), ammonia (NH_4^+), soluble active phosphorus (SRP), nitrate ($N0_3^-$) and soluble reactive silica (SiO_2) were expressed in $mg.l^{-1}$; dissolved inorganic carbon (DIC) was expressed in mM and clorophyll a (Chla) in $\mu g.l^{-1}$. Depth and perimeter were expressed in meters (m).

Environmental feature	Range	Mean	Median	SD
wT	10 – 27.7	21.31	21.40	4.18
pH	6.9-9.6	7.75	7.70	0.45
O_2	0-31.6	7.07	6.20	4.84
ES	180-961	532.6	494	235.65
DIC	1.28 – 11.73	4.71	4.23	2.78
NH_4^+	0-3.44	0.27	0.09	0.57
SRP	0-1.59	0.09	0.02	0.25
NO_3^-	1-1.09	0.16	0.07	0.23
Chla	0-11.20	2.86	1.95	2.78
SiO_2	0.01-24.41	6.21	3.31	6.39
Depth	0.70 – 7.50	4.20	4.40	1.64
Perimeter	16-408	210.1	196	75.68

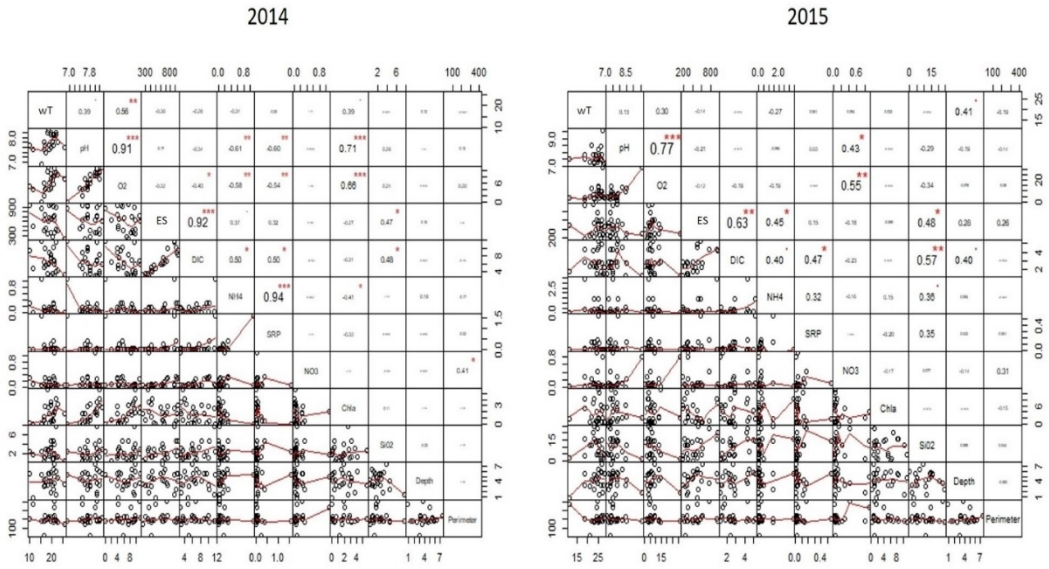

Figure 5SM The panels showed the correlation matrix of the environmental features for the years 2014 and 2015. The diagonal showed the labels of the eleven environmental features: water temperature (wT), pH, Oxygen (O_2), conductivity (EC), dissolved inorganic carbon (DIC), ammonia (NH_4), soluble reactive phosphorus (SRP), nitrate (NO_3), chlorophyll a (Chla), silica (SiO_2), depth and perimeter. In the upper diagonal part was present the Pearson correlation coefficient between pairs of variables proportional in size to the magnitude of the value, the stars showed the significance level of the correlation test (* $p < 0.05$, ** $p < 0.001$, *** $p < 0.0001$). The lower diagonal part showed the scatterplot between pairs of variables. The red line reported the lowess smoother.

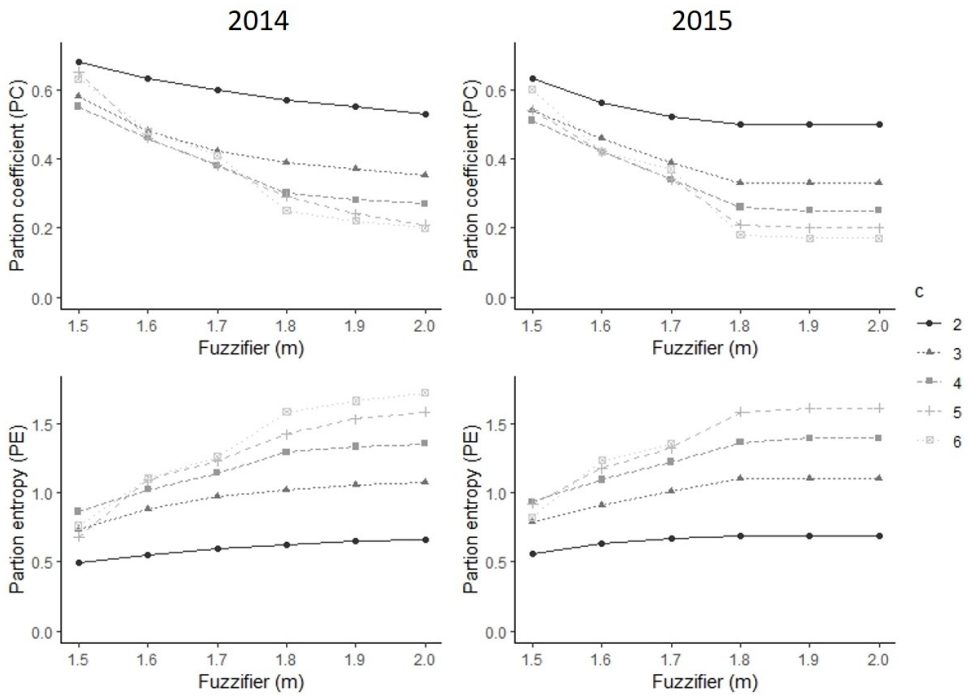

Figure 6SM Values of partition coefficient (PC) and partition entropy (PC) for the years 2014 (left panel) and 2015 (right panel). For each plot, the grey scale colors, the point's shape and the geometry of the lines were relative to the different number of clusters (c) in the range 2 - 6.

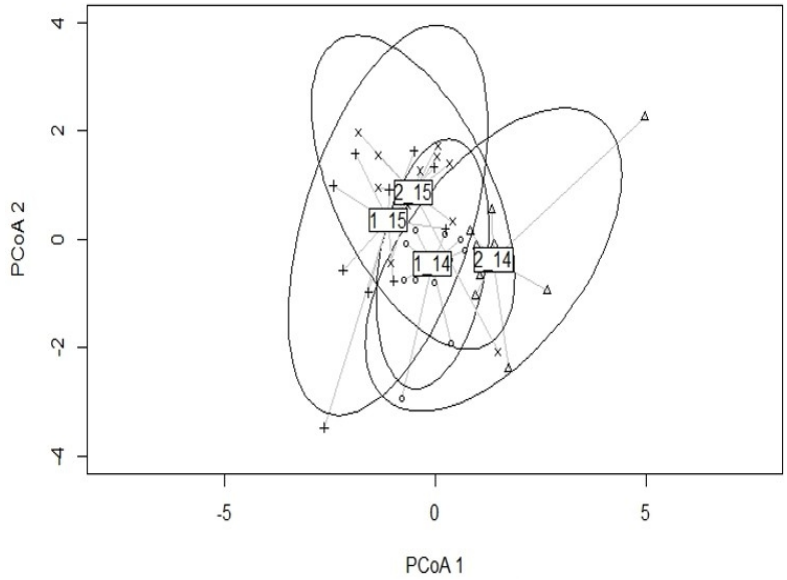

Figure 7SM Showed the ordination plot of the average euclidean distance of the scaled environmental features of the water chemistry, from the median of each cluster found by fuzzy c-means. Each pond was reported on the first two principal coordinate axis and the symbols were relative to the ponds of cluster 1 (○) and cluster 2 (Δ) in 2014, cluster 1 (+) and cluster 2 (x) in 2015. The ellypses represent 1 standard deviation of the euclidean distances from the median of the clusters. The PERMIDISP analysis after permutation test, revealed not significant difference between groups in habitat heterogeneity.

Table 15SM Showed the results of Kolmogorov-Smirnov test between distributions of beta diversity indices after the resampling procedure. The stars represent the p-value (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ and **** $p < 0.0001$).

Beta diversity index	2014	2015	2014 -2015
	Cluster 1 - Cluster 2	Cluster 1 - Cluster 2	
β_{SOR}	D = 0.75****	D = 0.60****	D = 0.39****
β_{SNE}	D = 0.41****	D = 0.16****	D = 0.31****
β_{SIM}	D = 0.81****	D = 0.43****	D = 0.37****

Table 16SM Showed the association rules from presence/absence data mined with frequent pattern growth algorithm, in 2014 and 2015. The association rules highlight the frequency and the correlations between taxa co-occurrences. For each rule were reported the quantitative measures of interestingness: support, confidence and lift.

N°	Association rules	Support	Confidence	Lift
1	*Pleuroxus*, Cyclopoida, Calanoida => *Simocephalus*	0.10	1.00	3.92
2	*Pleuroxus*, Cyclopoida => *Simocephalus*	0.10	0.83	3.26
3	*Daphnia*, *Simocephalus* => Calanoida	0.10	1.00	1.52
4	*Chydorus, Pleuroxus* => Calanoida	0.10	1.00	1.52
5	*Simocephalus*, *Pleuroxus* => Calanoida	0.17	1.00	1.52
6	*Pleuroxus* => Calanoida	0.25	0.85	1.29
7	*Simocephalus*, Cyclopoida => Calanoida	0.12	0.85	1.29
8	*Simocephalus* => Calanoida	0.21	0.83	1.26
9	*Chydorus*, Cyclopoida => Calanoida	0.10	0.83	1.26

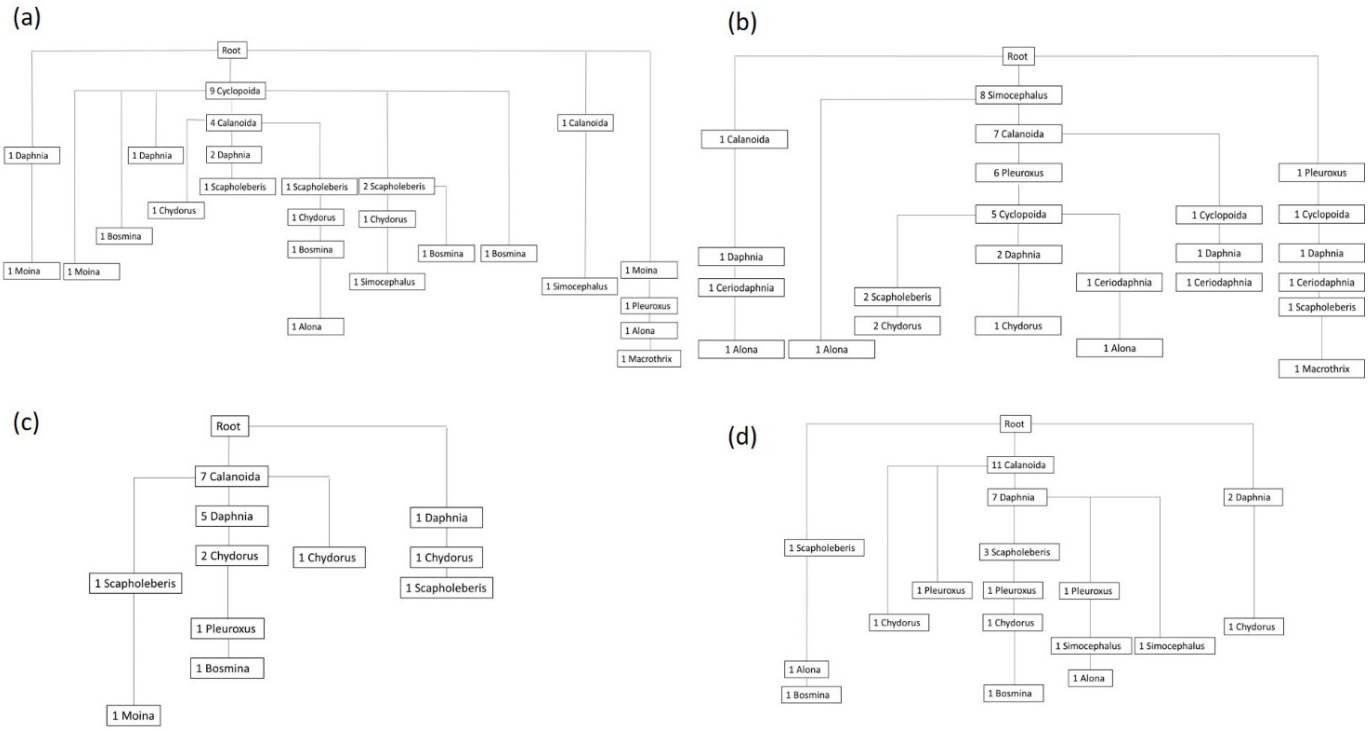

Figure 8SM Frequent pattern trees (FP_t) for the community structure in cluster 1 (a) and 2 (b) in 2014 and in cluster 1 (c) and 2 (d) in 2015. Each node represents a specific taxon and its absolute frequency (number of ponds where the taxon was found). The branches join the co-occurrence of taxa.

Chapter 4: Ecoacustic and sounds analysis

Make the CPUs do the hard work - Automated acoustic feature extraction and visualization for marine ecoacoustics applications illustrated using marine mammal Passive Acoustic Monitoring datasets

It was reported a set of reports of the results of 10-fold nested-cross validation for both the Watkins Marine Mammal Sounds Database (WMD) and the Placentia Bay Database (PBD).

Y and X indicate the labels and features being tested, respectively; acc_f1 and acc_bal report the accuracy and balanced accuracy scores, respectively; max features indicate the number of features selected by the inner 5-fold cross-validation loop; n estimators indicate the number of trees selected to generate predictions using random forest.

Most runs executed on the WMD dataset labels selected random forest classifiers based on only one of the two UMAP dimensions. Runs executed on the PBD dataset with X = UMAP dim 1, UMAP dim 2 occasionally retained both dimensions, while runs with X = [128 acoustic features] selected models with as few as 2 features and as many as 128 features. Confusion matrices with scores for each class are also included. Scores either range between 0 and 1, indicating % of correct predictions, or report actual sample sizes.

Taxonomic Groups (WMD)

Nested cross-validation results

Y = Mysticete, Odontocete, Pinniped
X = UMAP dim 1, UMAP dim 2

```
>acc_f1=0.989,>acc_bal=0.771, cfg={'max_features': 1, 'n_estimators': 200}
>acc_f1=0.989,>acc_bal=0.766, cfg={'max_features': 1, 'n_estimators': 200}
>acc_f1=0.991,>acc_bal=0.836, cfg={'max_features': 1, 'n_estimators': 150}
>acc_f1=0.987,>acc_bal=0.755, cfg={'max_features': 1, 'n_estimators': 200}
>acc_f1=0.991,>acc_bal=0.811, cfg={'max_features': 1, 'n_estimators': 200}
>acc_f1=0.988,>acc_bal=0.802, cfg={'max_features': 1, 'n_estimators': 200}
>acc_f1=0.988,>acc_bal=0.786, cfg={'max_features': 1, 'n_estimators': 200}
>acc_f1=0.988,>acc_bal=0.792, cfg={'max_features': 1, 'n_estimators': 150}
>acc_f1=0.991,>acc_bal=0.848, cfg={'max_features': 1, 'n_estimators': 200}
>acc_f1=0.989,>acc_bal=0.832, cfg={'max_features': 1, 'n_estimators': 200}

F1 score: 0.989 (0.001)
Balanced accuracy score: 0.800 (0.030)
```

Confusion matrix

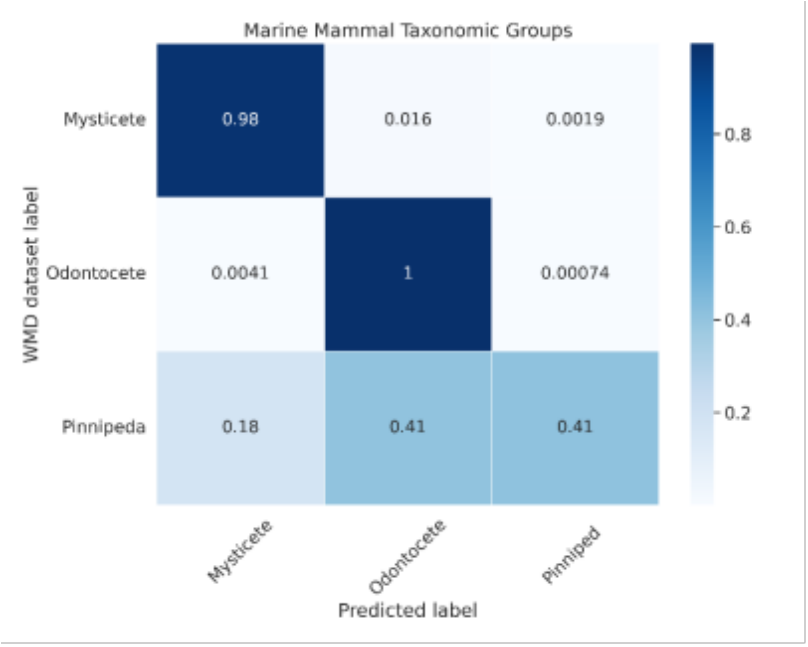

Figure 9SM Confusion matrix for the taxonomic groups' labels.

Mysticete Families (WMD)
Nested cross-validation results

Y = Balaenidae, Balaenopteridae, Eschrichtiidae;
X = UMAP dim 1, UMAP dim 2

```
>acc_f1=0.959,>acc_bal=0.822, cfg={'max_features': 1, 'n_estimators': 100}
>acc_f1=0.961,>acc_bal=0.804, cfg={'max_features': 1, 'n_estimators': 200}
>acc_f1=0.962,>acc_bal=0.842, cfg={'max_features': 1, 'n_estimators': 100}
>acc_f1=0.971,>acc_bal=0.862, cfg={'max_features': 1, 'n_estimators': 200}
>acc_f1=0.964,>acc_bal=0.764, cfg={'max_features': 1, 'n_estimators': 50}
>acc_f1=0.960,>acc_bal=0.829, cfg={'max_features': 1, 'n_estimators': 150}
>acc_f1=0.966,>acc_bal=0.771, cfg={'max_features': 1, 'n_estimators': 50}
>acc_f1=0.960,>acc_bal=0.748, cfg={'max_features': 1, 'n_estimators': 200}
>acc_f1=0.957,>acc_bal=0.740, cfg={'max_features': 1, 'n_estimators': 150}
>acc_f1=0.964,>acc_bal=0.881, cfg={'max_features': 1, 'n_estimators': 200}

F1 score: 0.962 (0.004)
Balanced accuracy score: 0.806 (0.046)
```

Confusion matrix

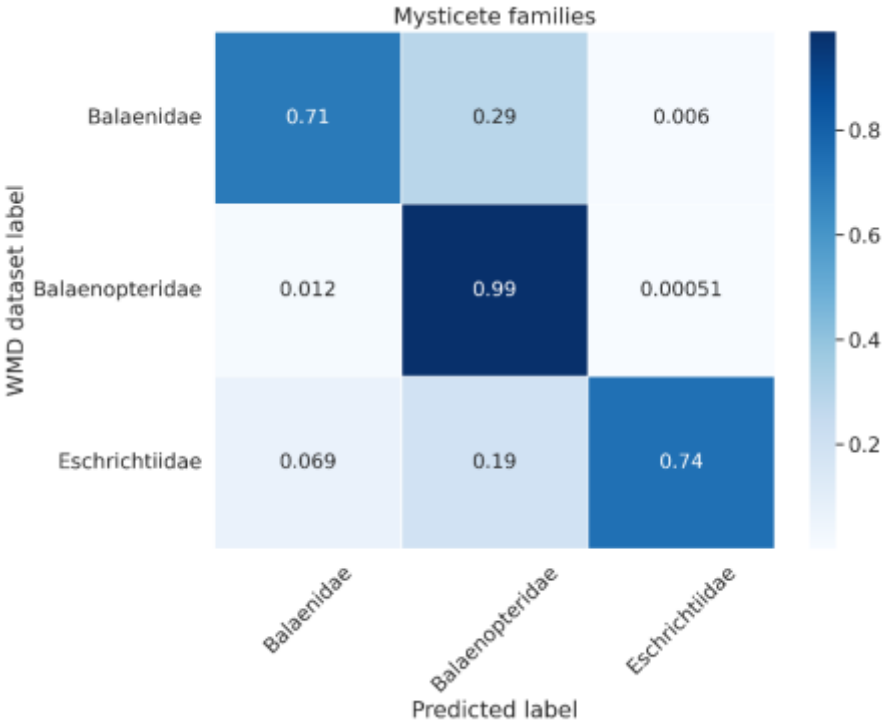

Figure 10SM Confusion matrix for mysticete families' labels.

Balaenopteridae species (WMD)
Nested cross-validation results

Y = minke whale, fin whale, humpback whale
X = UMAP dim 1, UMAP dim 2

```
>acc_f1=0.999,>acc_bal=1.000, est=0.998, cfg={'max_features': 1, 'n_estimators': 50}
>acc_f1=0.998,>acc_bal=0.999, est=0.999, cfg={'max_features': 1, 'n_estimators': 100}
>acc_f1=0.997,>acc_bal=0.962, est=0.998, cfg={'max_features': 1, 'n_estimators': 200}
>acc_f1=0.997,>acc_bal=0.985, est=0.999, cfg={'max_features': 1, 'n_estimators': 100}
>acc_f1=0.999,>acc_bal=0.980, est=0.998, cfg={'max_features': 1, 'n_estimators': 50}
>acc_f1=0.999,>acc_bal=0.999, est=0.998, cfg={'max_features': 1, 'n_estimators': 50}
>acc_f1=0.999,>acc_bal=0.979, est=0.999, cfg={'max_features': 1, 'n_estimators': 50}
>acc_f1=0.999,>acc_bal=0.980, est=0.998, cfg={'max_features': 1, 'n_estimators': 150}
>acc_f1=0.999,>acc_bal=1.000, est=0.998, cfg={'max_features': 1, 'n_estimators': 50}
>acc_f1=0.997,>acc_bal=0.983, est=0.999, cfg={'max_features': 1, 'n_estimators': 200}
F1 score: 0.998 (0.001)
Balanced accuracy score: 0.987 (0.012)
```

Confusion matrix

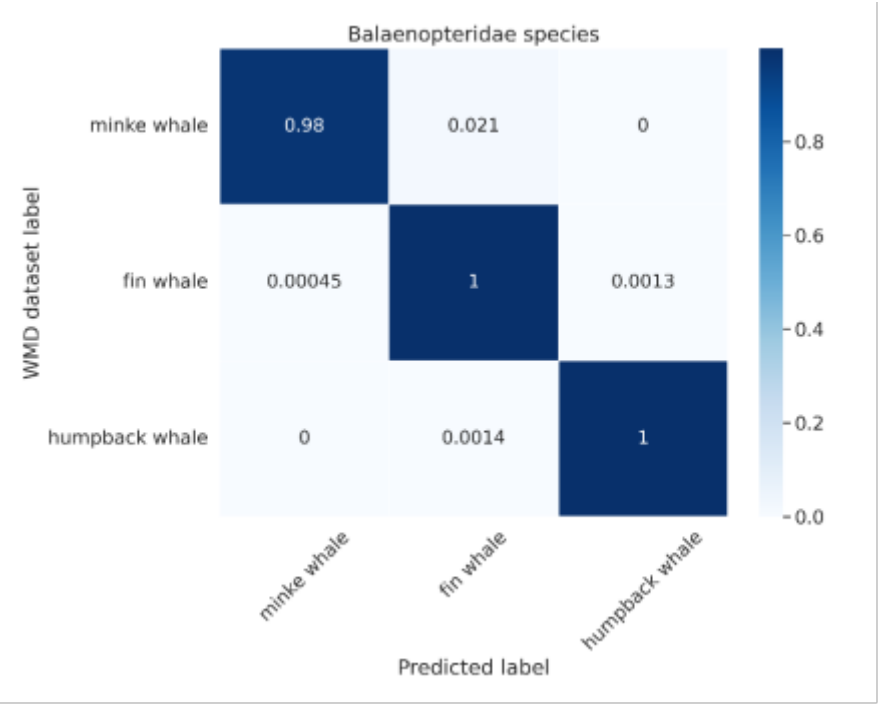

Figure 11SM Confusion matrix for the Balaenopteridae species' labels.

Odontocete families (WMD)

Nested cross-validation results

Y = Delphinidae, Monodontidae, Phocoenidae, Physteridae

X = UMAP dim 1, UMAP dim 2

```
>acc_f1=0.960,>acc_bal=0.693, est=0.961, cfg={'max_features': 1, 'n_estimators': 100}
>acc_f1=0.959,>acc_bal=0.695, est=0.961, cfg={'max_features': 1, 'n_estimators': 150}
>acc_f1=0.957,>acc_bal=0.750, est=0.961, cfg={'max_features': 1, 'n_estimators': 200}
>acc_f1=0.965,>acc_bal=0.684, est=0.960, cfg={'max_features': 1, 'n_estimators': 200}
>acc_f1=0.960,>acc_bal=0.729, est=0.961, cfg={'max_features': 1, 'n_estimators': 100}
>acc_f1=0.962,>acc_bal=0.739, est=0.960, cfg={'max_features': 1, 'n_estimators': 200}
>acc_f1=0.966,>acc_bal=0.781, est=0.960, cfg={'max_features': 1, 'n_estimators': 200}
>acc_f1=0.960,>acc_bal=0.707, est=0.961, cfg={'max_features': 1, 'n_estimators': 200}
>acc_f1=0.961,>acc_bal=0.784, est=0.960, cfg={'max_features': 1, 'n_estimators': 150}
>acc_f1=0.963,>acc_bal=0.697, est=0.960, cfg={'max_features': 1, 'n_estimators': 150}
F1 score: 0.961 (0.003)
Balanced accuracy score: 0.726 (0.035)
```

Confusion matrix

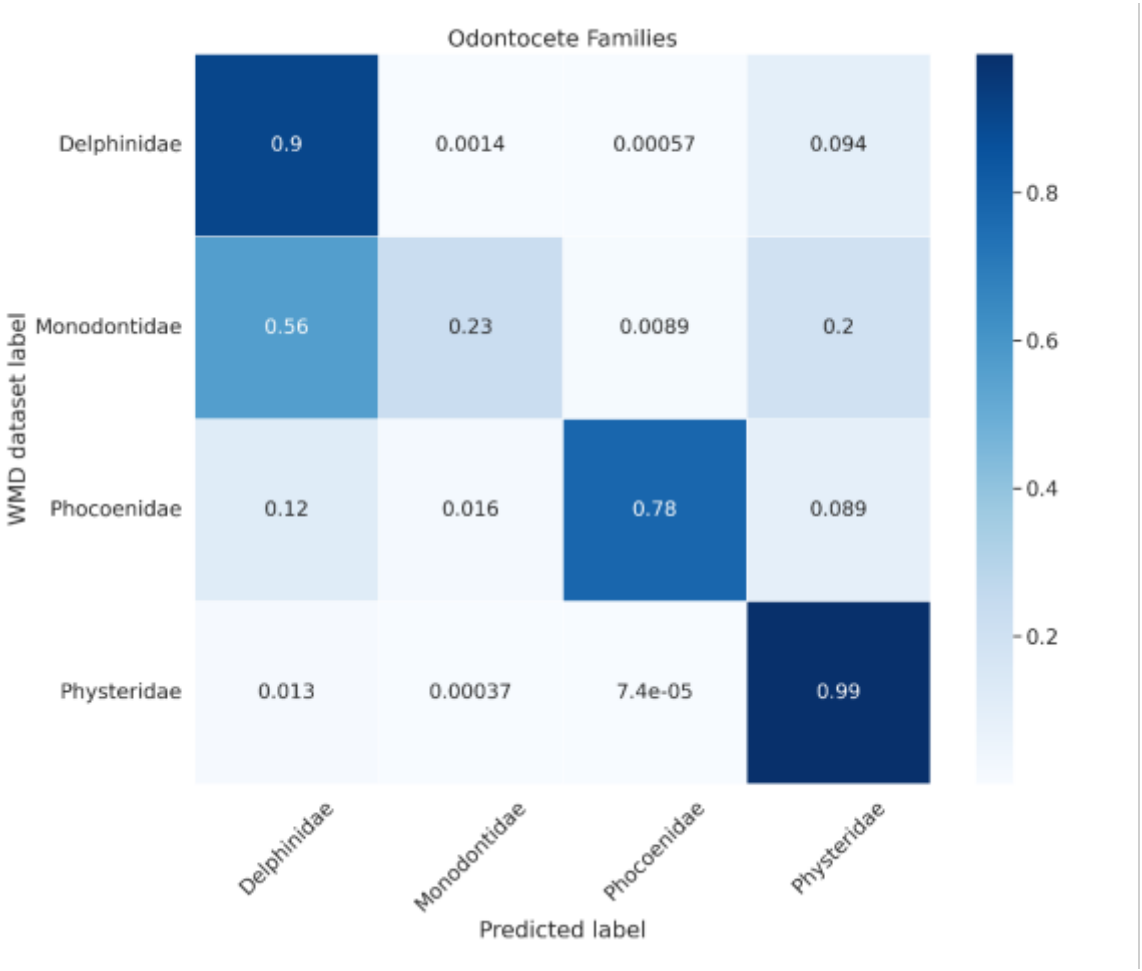

Figure 12SM Confusion matrix for the odontocete families' labels.

Delphinidae species (WMD)
Nested cross-validation results

Y = 14 species (see species codes below): Cehe; Deba; Dede; Glma; Glme; Grgr; Laho; Laob; Oror; Pscr; Stat; Stcl; Stfr; Tutr
X = UMAP dim 1, UMAP dim 2

```
>acc_f1=0.824,>acc_bal=0.677, est=0.827, cfg={'max_features': 1, 'n_estimators': 150}
>acc_f1=0.823,>acc_bal=0.708, est=0.828, cfg={'max_features': 1, 'n_estimators': 200}
>acc_f1=0.833,>acc_bal=0.746, est=0.829, cfg={'max_features': 1, 'n_estimators': 200}
>acc_f1=0.805,>acc_bal=0.687, est=0.832, cfg={'max_features': 1, 'n_estimators': 50}
>acc_f1=0.836,>acc_bal=0.701, est=0.827, cfg={'max_features': 1, 'n_estimators': 100}
>acc_f1=0.815,>acc_bal=0.665, est=0.828, cfg={'max_features': 1, 'n_estimators': 150}
>acc_f1=0.830,>acc_bal=0.672, est=0.826, cfg={'max_features': 1, 'n_estimators': 200}
>acc_f1=0.841,>acc_bal=0.724, est=0.829, cfg={'max_features': 1, 'n_estimators': 150}
>acc_f1=0.836,>acc_bal=0.711, est=0.826, cfg={'max_features': 1, 'n_estimators': 200}
```

```
>acc_f1=0.844,>acc_bal=0.735, est=0.823, cfg={'max_features': 1, 'n_estimators': 100}
F1 score: 0.829 (0.012)
Balanced accuracy score: 0.703 (0.026)
```

Confusion matrix

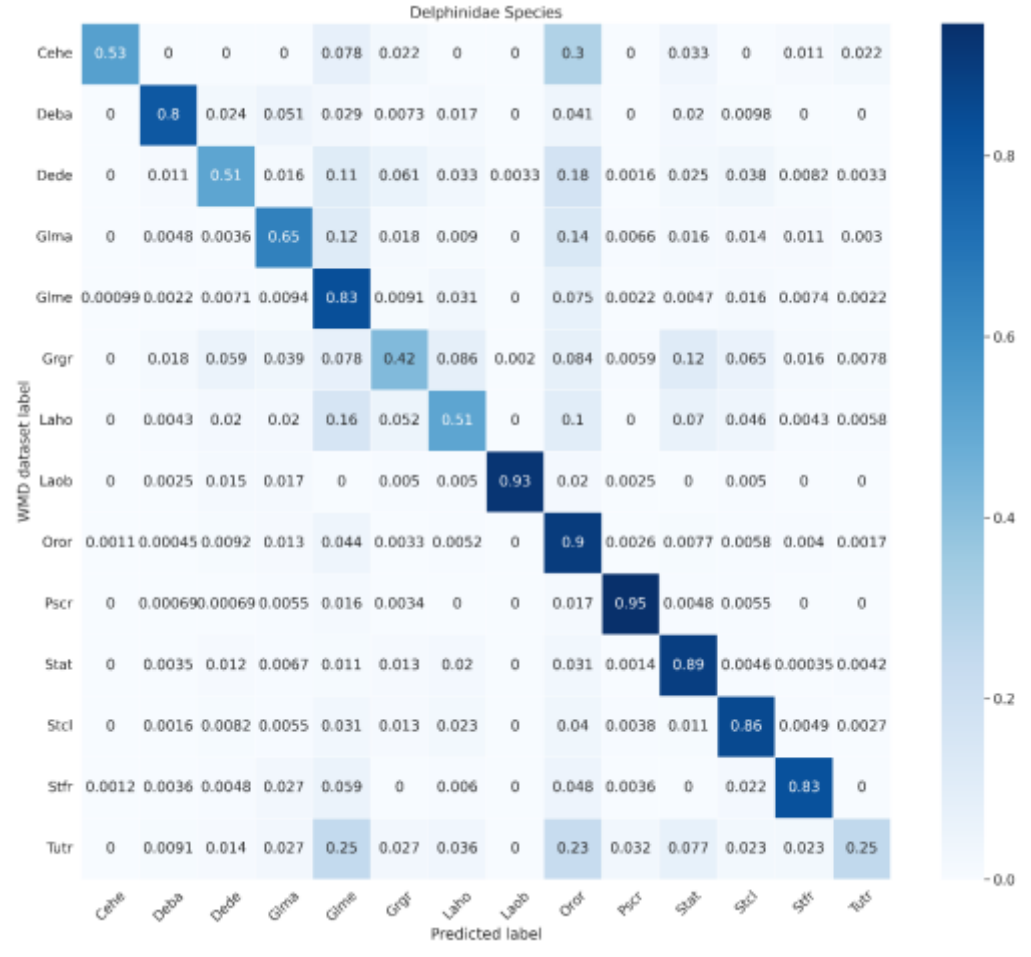

Figure 13SM Confusion matrix for the Delphinidae species' labels

Species Codes:

Table 17SM Species codes, scientific names, and common names for the Delphinidae species confusion matrix.

CODE	SCIENTIFIC NAME	COMMON NAME
Cehe	*Cephalorhynchus heavisidii*	Heaviside's dolphin
Deba	*Delphinus capensis* (formerly *D. bairdii*)	Long-beaked common dolphin
Dede	*Delphinus delphis*	Short-beaked common dolphin
Glma	Globicephala macrorhynchus	Short-finned pilot whale
Glme	Globicephala melas	Long-finned pilot whale
Grgr	Grampus griseus	Risso's dolphin
Laho	Lagenodelphis hosei	Fraser's dolphin
Laob	*Lagenorhynchus obliquidens*	Pacific white-sided dolphin
Oror	*Orcinus orca*	Orca / Killer whale
Pscr	*Pseudorca crassidens*	False killer whale
Stat	*Stenella attenuata*	Pantropical spotted dolphin
Stcl	*Stenella ceruleoalba*	Striped dolphin
Stfr	*Stenella frontalis*	Atlantic spotted dolphin
Tutr	*Tursiops truncatus*	Bottlenose dolphin

Orcinus orca populations (WMD)
Nested cross-validation results

Y = EN Atlantic, EN Atlantic - Captive, EN Pacific, EN Pacific - Captive, WN Atlantic
X = UMAP dim 1, UMAP dim 2

```
>acc_f1=0.912,>acc_bal=0.829, est=0.897, cfg={'max_features': 1, 'n_estimators': 200}
>acc_f1=0.896,>acc_bal=0.790, est=0.894, cfg={'max_features': 1, 'n_estimators': 200}
>acc_f1=0.907,>acc_bal=0.782, est=0.896, cfg={'max_features': 1, 'n_estimators': 150}
>acc_f1=0.887,>acc_bal=0.810, est=0.896, cfg={'max_features': 1, 'n_estimators': 150}
>acc_f1=0.887,>acc_bal=0.836, est=0.892, cfg={'max_features': 1, 'n_estimators': 50}
>acc_f1=0.896,>acc_bal=0.753, est=0.895, cfg={'max_features': 1, 'n_estimators': 100}
>acc_f1=0.900,>acc_bal=0.752, est=0.892, cfg={'max_features': 1, 'n_estimators': 100}
>acc_f1=0.880,>acc_bal=0.768, est=0.896, cfg={'max_features': 1, 'n_estimators': 150}
>acc_f1=0.916,>acc_bal=0.768, est=0.895, cfg={'max_features': 1, 'n_estimators': 200}
>acc_f1=0.895,>acc_bal=0.820, est=0.896, cfg={'max_features': 1, 'n_estimators': 200}
F1 score: 0.897 (0.011)
Balanced accuracy score: 0.791 (0.030)
```

Confusion matrix

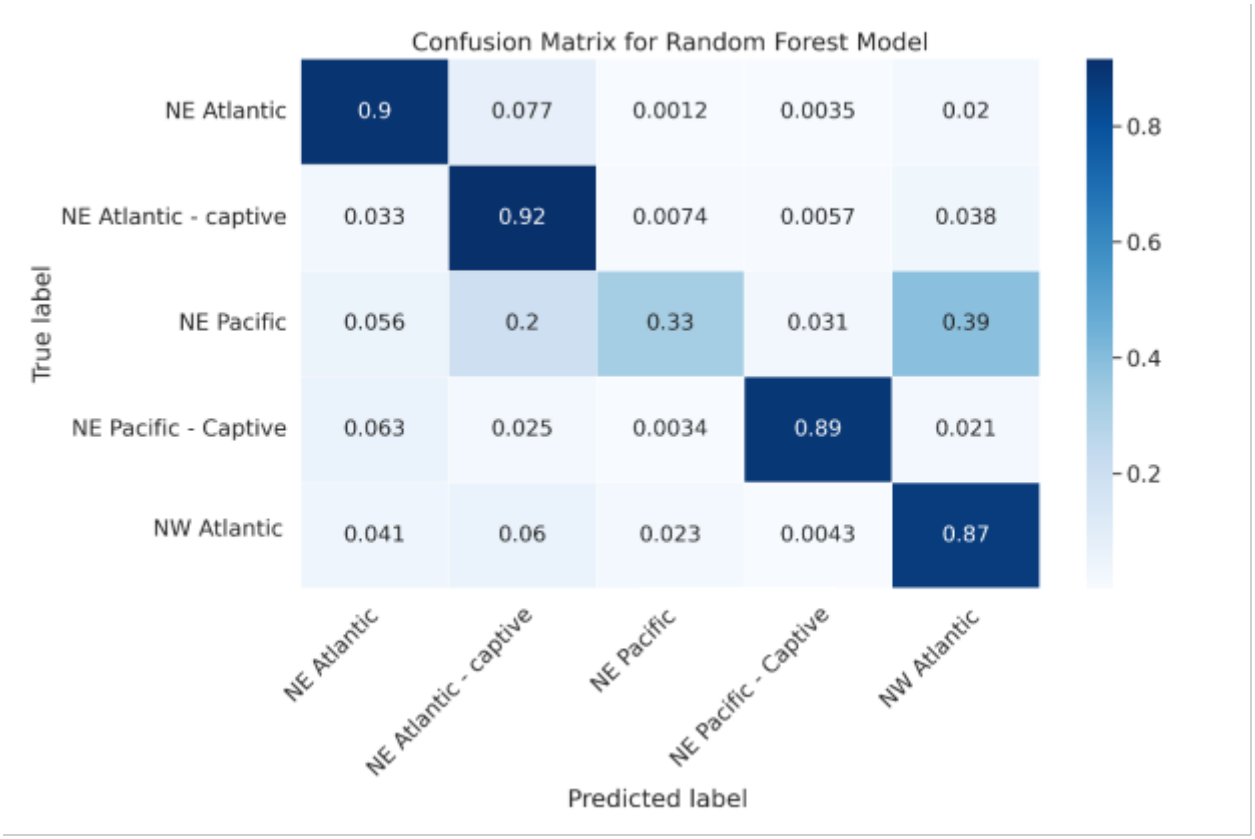

Figure 14SM Confusion matrix for the Orcinus orca populations' labels

Locations (PBD)
Nested cross-validation results

Y = Burin, Red Island
X = UMAP dim 1, UMAP dim 2

```
>acc_f1=0.978,>acc_bal=0.974, est=0.957, cfg={'max_features': 1, 'n_estimators': 50}
>acc_f1=0.963,>acc_bal=0.952, est=0.957, cfg={'max_features': 1, 'n_estimators': 200}
>acc_f1=0.965,>acc_bal=0.961, est=0.952, cfg={'max_features': 1, 'n_estimators': 100}
>acc_f1=0.956,>acc_bal=0.956, est=0.957, cfg={'max_features': 1, 'n_estimators': 150}
>acc_f1=0.965,>acc_bal=0.954, est=0.959, cfg={'max_features': 1, 'n_estimators': 100}
>acc_f1=0.968,>acc_bal=0.968, est=0.958, cfg={'max_features': 1, 'n_estimators': 150}
>acc_f1=0.965,>acc_bal=0.961, est=0.954, cfg={'max_features': 1, 'n_estimators': 200}
>acc_f1=0.950,>acc_bal=0.944, est=0.957, cfg={'max_features': 1, 'n_estimators': 100}
>acc_f1=0.955,>acc_bal=0.957, est=0.957, cfg={'max_features': 1, 'n_estimators': 150}
>acc_f1=0.946,>acc_bal=0.947, est=0.961, cfg={'max_features': 1, 'n_estimators': 150}
F1 score: 0.961 (0.009)
Balanced accuracy score: 0.957 (0.009)
```

Confusion matrix

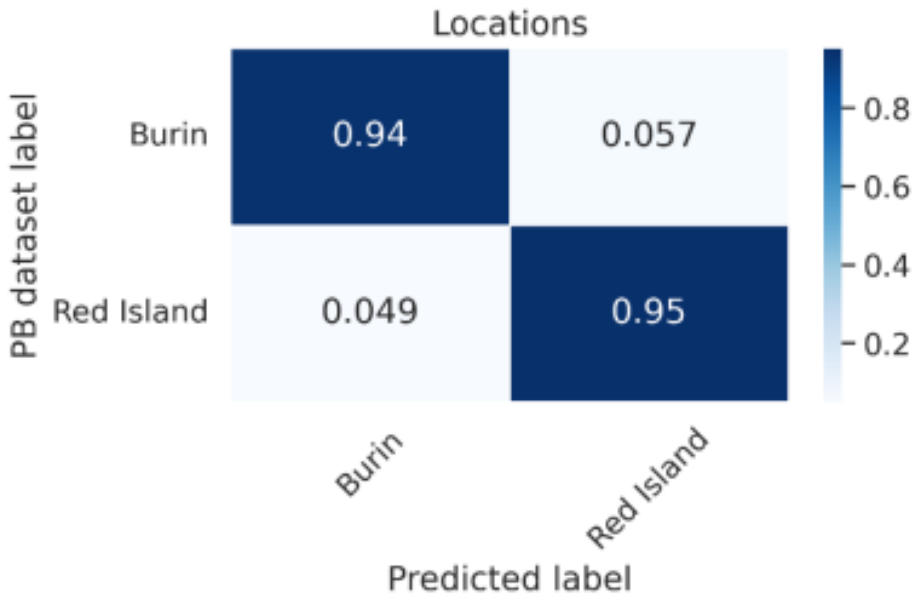

Figure 15SM Confusion matrix for the location labels.

Seismic airgun presence/absence
Nested cross-validation results – UMAP

Y = Burin, Red Island
X = UMAP dim 1, UMAP dim 2

```
>acc_f1=0.965,>acc_bal=0.843, est=0.968, cfg={'max_features': 2, 'n_estimators': 200}
>acc_f1=0.956,>acc_bal=0.842, est=0.964, cfg={'max_features': 1, 'n_estimators': 100}
>acc_f1=0.983,>acc_bal=0.945, est=0.964, cfg={'max_features': 1, 'n_estimators': 50}
>acc_f1=0.956,>acc_bal=0.820, est=0.965, cfg={'max_features': 1, 'n_estimators': 50}
>acc_f1=0.980,>acc_bal=0.900, est=0.967, cfg={'max_features': 1, 'n_estimators': 200}
>acc_f1=0.978,>acc_bal=0.913, est=0.965, cfg={'max_features': 1, 'n_estimators': 150}
>acc_f1=0.983,>acc_bal=0.887, est=0.967, cfg={'max_features': 1, 'n_estimators': 100}
>acc_f1=0.960,>acc_bal=0.808, est=0.967, cfg={'max_features': 1, 'n_estimators': 100}
>acc_f1=0.968,>acc_bal=0.840, est=0.966, cfg={'max_features': 2, 'n_estimators': 150}
>acc_f1=0.955,>acc_bal=0.781, est=0.967, cfg={'max_features': 2, 'n_estimators': 50}
F1 score: 0.968 (0.011)
Balanced accuracy score: 0.858 (0.049)
```

Nested cross-validation results – Acoustic Features

Y = Burin, Red Island
X = 128 acoustic features

```
>acc_f1=0.980,>acc_bal=0.867, est=0.985, cfg={'max_features': 16, 'n_estimators': 100}
>acc_f1=0.980,>acc_bal=0.892, est=0.983, cfg={'max_features': 32, 'n_estimators': 200}
>acc_f1=0.993,>acc_bal=0.950, est=0.982, cfg={'max_features': 16, 'n_estimators': 200}
>acc_f1=0.988,>acc_bal=0.942, est=0.983, cfg={'max_features': 16, 'n_estimators': 200}
>acc_f1=0.990,>acc_bal=0.941, est=0.985, cfg={'max_features': 32, 'n_estimators': 150}
>acc_f1=0.993,>acc_bal=0.958, est=0.982, cfg={'max_features': 64, 'n_estimators': 150}
>acc_f1=0.995,>acc_bal=0.968, est=0.983, cfg={'max_features': 32, 'n_estimators': 50}
>acc_f1=0.980,>acc_bal=0.892, est=0.984, cfg={'max_features': 32, 'n_estimators': 200}
>acc_f1=0.985,>acc_bal=0.870, est=0.985, cfg={'max_features': 32, 'n_estimators': 200}
>acc_f1=0.988,>acc_bal=0.896, est=0.982, cfg={'max_features': 32, 'n_estimators': 50}
F1 score: 0.987 (0.005)
Balanced accuracy score: 0.917 (0.036)
```

Confusion matrices: **UMAP dimensions**

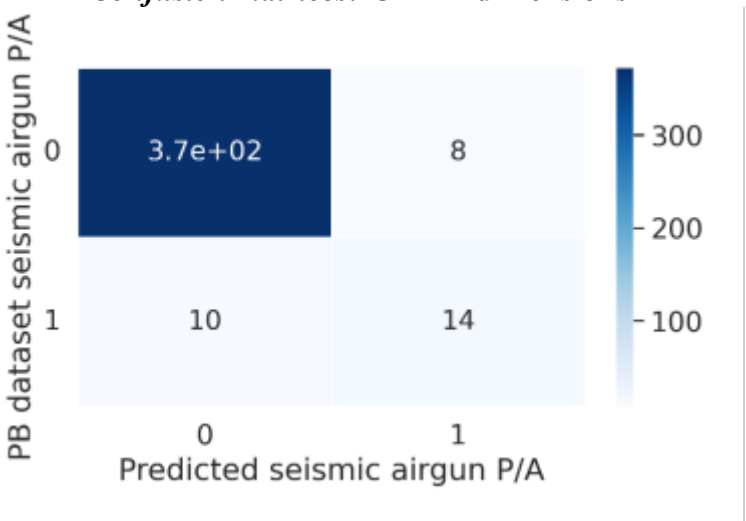

Figure 16SM Confusion matrix for the seismic airgun presence labels predicted by two UMAP dimensions.

***Confusion matrices:* Acoustic Features**

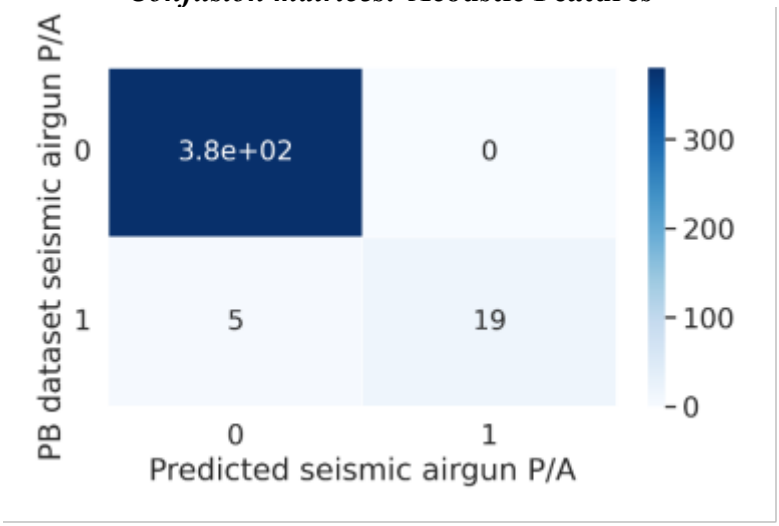

Figure 17SM Confusion matrix for the seismic airgun presence labels predicted by 128 acoustic features.

Ship presence/absence
Nested cross-validation results – UMAP

Y = Presence / Absence
X = UMAP dim 1, UMAP dim 2

```
>acc_f1=0.983,>acc_bal=0.770, est=0.979, cfg={'max_features': 1, 'n_estimators': 200}
>acc_f1=0.978,>acc_bal=0.723, est=0.978, cfg={'max_features': 1, 'n_estimators': 200}
>acc_f1=0.985,>acc_bal=0.831, est=0.977, cfg={'max_features': 1, 'n_estimators': 100}
>acc_f1=0.968,>acc_bal=0.618, est=0.980, cfg={'max_features': 1, 'n_estimators': 200}
>acc_f1=0.985,>acc_bal=0.712, est=0.979, cfg={'max_features': 1, 'n_estimators': 150}
>acc_f1=0.975,>acc_bal=0.678, est=0.977, cfg={'max_features': 1, 'n_estimators': 50}
>acc_f1=0.978,>acc_bal=0.679, est=0.979, cfg={'max_features': 1, 'n_estimators': 150}
>acc_f1=0.985,>acc_bal=0.642, est=0.978, cfg={'max_features': 1, 'n_estimators': 150}
>acc_f1=0.980,>acc_bal=0.667, est=0.979, cfg={'max_features': 1, 'n_estimators': 200}
>acc_f1=0.975,>acc_bal=0.664, est=0.979, cfg={'max_features': 1, 'n_estimators': 50}
F1 score: 0.979 (0.005)
Balanced accuracy score: 0.698 (0.060)
```

Nested cross-validation results – Acoustic Features

Y = Presence / Absence
X = 128 acoustic features

```
>acc_f1=0.990,>acc_bal=0.951, est=0.992, cfg={'max_features': 4, 'n_estimators': 200}
>acc_f1=0.988,>acc_bal=0.817, est=0.993, cfg={'max_features': 8, 'n_estimators': 200}
>acc_f1=0.990,>acc_bal=0.874, est=0.992, cfg={'max_features': 8, 'n_estimators': 100}
>acc_f1=0.988,>acc_bal=0.853, est=0.991, cfg={'max_features': 2, 'n_estimators': 200}
>acc_f1=0.993,>acc_bal=0.786, est=0.991, cfg={'max_features': 16, 'n_estimators': 50}
>acc_f1=0.990,>acc_bal=0.907, est=0.992, cfg={'max_features': 32, 'n_estimators': 50}
>acc_f1=0.998,>acc_bal=0.955, est=0.992, cfg={'max_features': 16, 'n_estimators': 100}
>acc_f1=0.993,>acc_bal=0.786, est=0.991, cfg={'max_features': 8, 'n_estimators': 100}
>acc_f1=0.985,>acc_bal=0.871, est=0.992, cfg={'max_features': 8, 'n_estimators': 200}
>acc_f1=0.988,>acc_bal=0.792, est=0.991, cfg={'max_features': 2, 'n_estimators': 200}
F1 score: 0.990 (0.003)
Balanced accuracy score: 0.859 (0.061)
```

Confusion matrices: UMAP dimensions

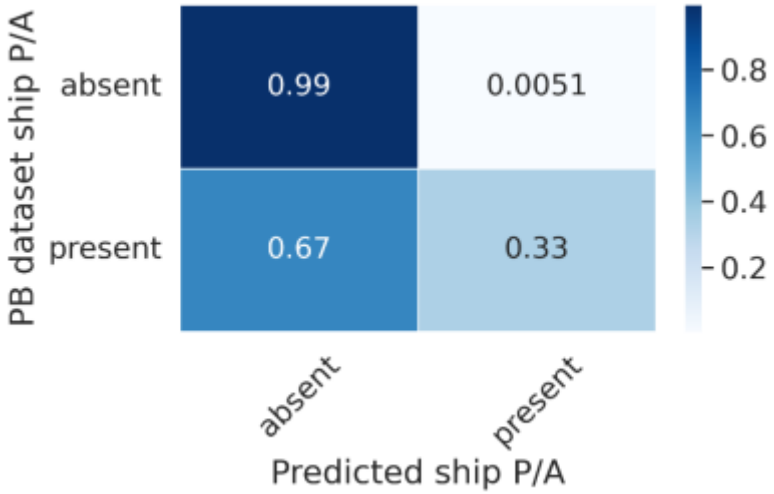

Figure 18SM Confusion matrix for the ship presence labels predicted by two UMAP dimensions.

Confusion matrices: Acoustic Features

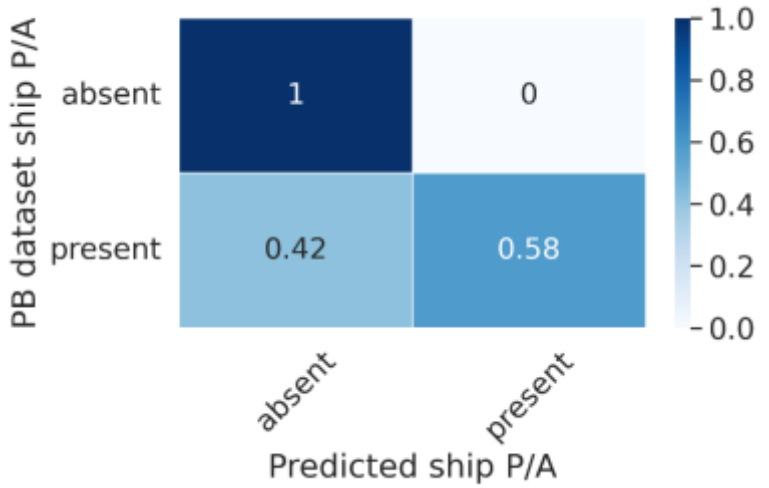

Figure 19SM Confusion matrix for the ship presence labels predicted by 128 acoustic features.

Humpback whale presence/absence
Nested cross-validation results – UMAP

Y = Presence / Absence
X = UMAP dim 1, UMAP dim 2

```
>acc_f1=0.542,>acc_bal=0.591, est=0.564, cfg={'max_features': 1, 'n_estimators': 50}
>acc_f1=0.586,>acc_bal=0.583, est=0.577, cfg={'max_features': 1, 'n_estimators': 50}
>acc_f1=0.546,>acc_bal=0.584, est=0.568, cfg={'max_features': 1, 'n_estimators': 50}
>acc_f1=0.584,>acc_bal=0.542, est=0.581, cfg={'max_features': 1, 'n_estimators': 50}
>acc_f1=0.579,>acc_bal=0.547, est=0.573, cfg={'max_features': 1, 'n_estimators': 50}
>acc_f1=0.575,>acc_bal=0.527, est=0.575, cfg={'max_features': 1, 'n_estimators': 50}
>acc_f1=0.595,>acc_bal=0.534, est=0.572, cfg={'max_features': 1, 'n_estimators': 50}
>acc_f1=0.587,>acc_bal=0.530, est=0.581, cfg={'max_features': 1, 'n_estimators': 50}
```

```
>acc_f1=0.567,>acc_bal=0.580, est=0.565, cfg={'max_features': 1, 'n_estimators': 50}
>acc_f1=0.579,>acc_bal=0.453, est=0.582, cfg={'max_features': 1, 'n_estimators': 50}
F1 score: 0.574 (0.017)
Balanced accuracy score: 0.547 (0.039)
```

Nested cross-validation results – Acoustic Features

Y = Presence / Absence

X = 128 acoustic features

```
>acc_f1=0.999,>acc_bal=1.000, est=0.999, cfg={'max_features': 128, 'n_estimators': 200}
>acc_f1=0.999,>acc_bal=1.000, est=0.999, cfg={'max_features': 128, 'n_estimators': 150}
>acc_f1=0.998,>acc_bal=0.999, est=0.999, cfg={'max_features': 128, 'n_estimators': 50}
>acc_f1=0.999,>acc_bal=0.999, est=0.999, cfg={'max_features': 128, 'n_estimators': 100}
>acc_f1=0.999,>acc_bal=1.000, est=0.999, cfg={'max_features': 128, 'n_estimators': 200}
>acc_f1=0.999,>acc_bal=1.000, est=0.999, cfg={'max_features': 128, 'n_estimators': 50}
>acc_f1=0.999,>acc_bal=1.000, est=0.999, cfg={'max_features': 128, 'n_estimators': 100}
>acc_f1=0.999,>acc_bal=1.000, est=0.999, cfg={'max_features': 128, 'n_estimators': 50}
>acc_f1=0.999,>acc_bal=1.000, est=0.999, cfg={'max_features': 128, 'n_estimators': 200}
>acc_f1=1.000,>acc_bal=1.000, est=0.999, cfg={'max_features': 128, 'n_estimators': 50}
F1 score: 0.999 (0.000)
Balanced accuracy score: 1.000 (0.000)
```

***Confusion matrices:* UMAP dimensions**

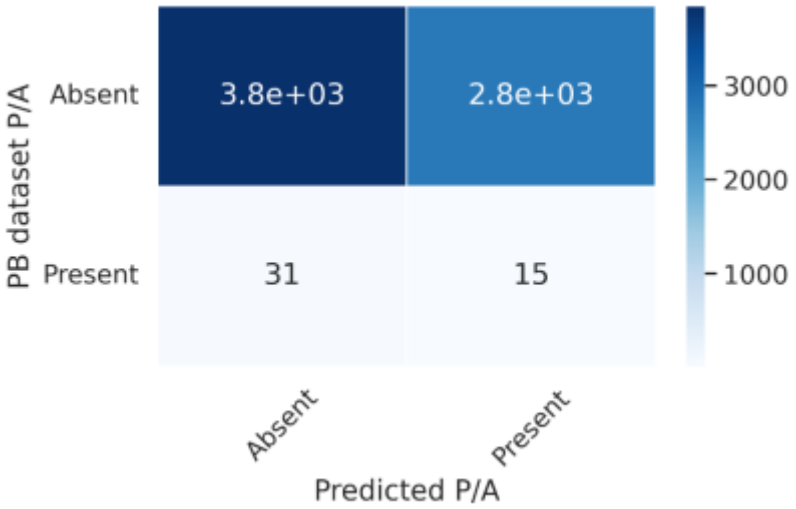

Figure 20SM Confusion matrix for the humpback whale presence labels predicted by 2 UMAP dimensions.

***Confusion matrices*: Acoustic Features**

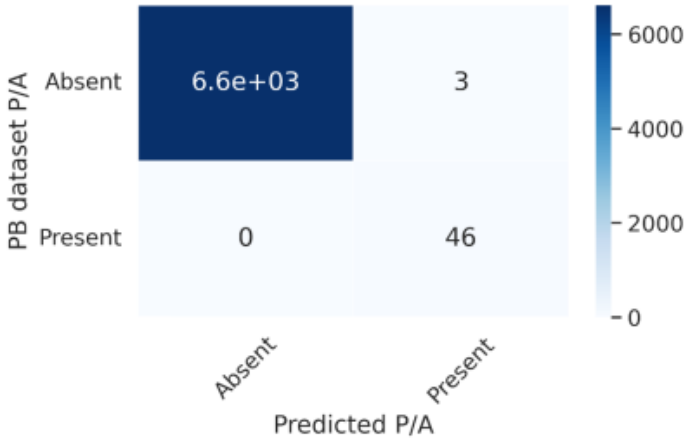

Figure 21SM Confusion matrix for the humpback whale presence labels predicted by 128 acoustic features.

9. References

AAVV. 1999. Appunti Sulla Golena del Po. Le Lanche di Motta e Torricella del Pizzo, Comune di Cremona: Cremona, Italy.

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mane, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viegas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X., 2015, TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, P., Toderici, G., Varadarajan, B., Vijayanarasimhan, S., 2016. YouTube-8M a large-scale video classification benchmark. arXiv:1609.08675.

Adams, D.C., Collyer, M.L., Kaliontzopoulou, A., 2020. Geomorph: software for geometric morphometric analyses. R package version 3.2.1.

Aggarwal, C.C., 2018. Neural networks and deep learning. In: Neural Networks and Deep Learning.

Agrawal, R., Imielinski, T., Swami, A., 1993. Mining association rules between sets of items in large databases. Acm Sigmod Rec., 22, 207–216.

Ahmedbahaaaldin, I., Ahmed, O., A.N. Ahmed, Ming F.C., Yuk, F.H., Ahmed, E.S. 2021. Extreme gradient boosting (XGBoost) model to predict the groundwater levels in Selangor Malaysia. Ain Shams Engineering Journal, 12(2), 1545-1556.

Aiello-Lammens, M.E., Boria, R.A., Radosavljevic, A., Vilela, B., and Anderson, R.P., 2010. spThin: An R package for spatial thinning of species occurrence records for use in ecological niche models. Ecography, 38(5), 541–545.

Ainsworth, T.D., Hurd, CL., Gates, R. D., Boyd, P.W., 2020. How do we overcome abrupt degradation of marine ecosystems and meet the challenge of heat waves and climate extremes? Glob Chang Biol, 26(2), 343–354.

Al-Jarrah, O.Y., Yoo, P.D., Muhaidat, S., Karagiannidis, G.K., Taha, K., 2015. Efficient Machine Learning for Big Data: A Review. Big Data Research, 2(3), 87-93.

Allaire, J., Chollet, F., 2020. keras: R Interface to “Keras”.

Allaire, J., Tang, Y., 2020. tensorflow: R Interface to “TensorFlow”.

Allen, A. N., Harvey, M., Harrell, L., Jansen, A., Merkens, K. P., Wall, C. C., Cattiau, J., and Oleson, E. M., 2021. A convolutional neural network for automated detection of humpback whale song in a diverse, long-term passive acoustic dataset. Front. Mar. Sci., 8.

Allen, A.P., Whittier, T.R., Kaufmann, P.R., Larsen, D.P., O’Connor, R.J., Hughes, R.M., Stemberger, R.S., Dixit, S.S., Brinkhurst, R.O., Herlihy, A.T., et al., 1999. Concordance of taxonomic richness patterns across multiple assemblages in lakes of the northeastern United States.Can. J. Fish. Aquat. Sci., 56, 739–747.

Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A., Aljaaf, A.J., 2020. A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science. In: Berry, M., Mohamed, A., Yap, B. (eds) Supervised and Unsupervised Learning for Data Science. Unsupervised and Semi-Supervised Learning. Springer, Cham.

Almeira, J., Guecha, S., 2019. Dominant power spectrums as a tool to establish an ecoacoustic baseline in a premontane moist forest. Landsc., 15(1), 121–130.

Anderson, L.G.; Hall, P.O.J., Iverfeldt, A., Van Der Loejf, M.M.R., Sundby, B., Westerlund, S.F.G., 1986. Benthic respiration measured bytotal carbonate production. Limnol. Oceanogr, 31, 319–329.

Anderson, M.J., 2005. Distance-Based Tests for Homogeneity of Multivariate Dispersions. Biometrics, 62, 245-253.

Anderson, M.J., 2001. A new method for non-parametric multivariate analysis of variance. Aust. Ecol. 26(1), 32-46.

Appel, E., Heepe, L., Lin, C.P., Gorb, S.N., 2015. Ultrastructure of dragonfly wing veins: composite structure of fibrous material supplemented by resilin. J. Anat. 227, 561–582.

Armitage, D.W., Ober, H.K. A, 2010. Comparison of supervised learning techniques in the classification of bat echolocation calls. Ecol. Inform., 5, 465–473.

Arnott, S.E.; Vanni, M.J., 1993. Zooplankton Assemblages in Fishless Bog Lakes: Influence of Biotic and Abiotic Factors. Ecology, 74, 2361–2380.

Artemov, G.N., Fedorova, V.S., Karagodin, D.A., Brusentsov, I.I., Baricheva, E.M., Sharakhov, I.V., Gordeev, M.I., Sharakhova, M.V., 2021. New Cytogenetic Photomap and Molecular Diagnostics for the Cryptic Species of the Malaria Mosquitoes Anopheles messeae and Anopheles daciae from Eurasia. Insects 12(9), 835.

Assis, J., Serrão, E.A., Claro, B., Perrin, C., Pearson, G.A., 2014. Climate-driven range shifts explain the distribution of extant gene pools and predict future loss of unique lineages in a marine brown alga. Mol Ecol, 23, 2797-2810.

Assis, J., Tyberghein, L., Bosch, S., Verbruggen, H., Serrão, E.A., De Clerck, O., 2018. Bio-ORACLE v2.0: Extending marine data layers for bioclimatic modelling. Global Ecology Biogeography, 27, 277– 284.

Atkinson, D., 1994. Temperature and organism size a biological law for ectotherms? Adv. Ecol. Res. 25, 1–58.

Ayala, D., CaroRiano, H., Dujardin, J.P., Rahola, N., Simard, F., Fontenille, D., 2011. Chromosomal and environmental determinants of morphometric variation in natural populations of the malaria vector Anopheles funestus in Cameroon. Infect. Genet. Evol., 11, 940–947.

Aytekin, A.M., Alten, B., Caglar, S.S., Ozbel, Y., Kaynas, S., Simsek, F.M., Kasap, O.E., Belen, A., 2007. Phenotypic variation among local populations of phlebotomine sand flies (Diptera: psychodidae) in southern Turkey. J. Vector Ecol., 32 (2), 226.

Baird, A.H., Babcock, R.C., Mundy, C.P., 2003. Habitat selection by larvae influences the depth distribution of six common coral species. MEPS, 252, 289-293.

Baker, A., Starger, C., McClanahan, T., Glynn, P. W., 2004. Corals' adaptive response to climate change. Nature, 430, 741.

Baker, D.J., Maclean, I.M.D., Goodall, M., Gaston, K.J., 2021. Species distribution modelling is needed to support ecological impact assessments. J. Appl. Ecol., 58, 21-26.

Baldi, P., Sadowski, P., 2014. The dropout learning algorithm. Artif. Intell., 210, 78-122.

Ballesteros, E., 2003. The coralligenous in the Mediterranean Sea: Definition of the coralligenous assemblage in the Mediterranean, its main builders, its richness and key role in benthic ecology as well as its threats. Project for the preparation of a Strategic Action Plan for the Conservation of the Biodiversity in the Mediterranean Region (SAP BIO). UNEP-MAP-RAC/SPA, 87.

Barbet-Massin, M., Jiguet, F., Albert, C.H., Thuiller, W., 2012. Selecting pseudo-absences for species distribution models: how, where and how many? Methods Ecol. Evol., 3(2), 327–338.

Barbosa, R.V., Davies, A.J., Sumida, P.Y.G., 2020. Habitat suitability and environmental niche comparison of cold-water coral species along the Brazilian continental margin, Deep Sea Research Part I: Oceanographic Research Papers, 155, 103147.

Barnett, A.J., Finlay, K., Beisner, B.E., 2007. Functional diversity of crustacean zooplankton communities: Towards a trait-based classification. Freshw. Biol., 52, 796–813.

Baselga, A. 2013. Multiple site dissimilarity quantifies compositional heterogeneity among several sites, while average pairwise dissimilarity may be misleading. Ecography, 36, 124–128.

Baselga, A., Orme, C.D.L., 2012. Betapart: An R package for the study of beta diversity. Methods Ecol. Evol., 3, 808–812.

Baselga, A.; Orme, D.; Villeger, S.; de Bortoli, J.; Leprieur, F.; Logez, M., 2020. Betapart: Partitioning Beta Diversity into Turnover and Nestedness Components. R Package Version 1.5.2.

Bates, D., M¨achler, M., Bolker, B.M., Walker, S.C., 2015. Fitting linear mixed-effects models using lme4. J. Stat. Softw., 67 (1), 1-48.

Beck, J., Böller, M., Erhardt, A., Schwanghart, W., 2014. Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. Ecol. Inform., 19, 10-15.

Becker, N., Petric, D., Zgomba, M., Boase, C., Madon, M.B., Dahl, C., Kaiser, A., 2010. Mosquitoes and their control, 2nd ed., Springer-Verlag, Berlin Heidelberg.

Begon, M., Townsend, C.R., Harper, J.L., 2006. Ecology: From Individuals to Ecosystems, 4th (Ed.). Blackwell Publishing.

Behera, N., Nanjundiah, V., 2004. Phenotypic plasticity can potentiate rapid evolutionary change. J. Theor. Biol., 226 (2), 177-184.

Belfiore, N.M. 2001. Effects of contaminants on genetic patterns in aquatic organisms. Mutat. Res., 489, 97–122.

Belyea, L.R.; Lancaster, J. Assembly within a contingent rules ecology. Oikos 2012, 86, 402–416.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. 57, 289-300.

Bennion, H.; Smith, M.A. 2000. Variability in the water chemistry of shallow ponds in southeast England, with special reference to the seasonality of nutrients and implications for modelling trophic status. Hydrobiologia, 436, 145-158.

Bergamasco, A., Malanotte-Rizzoli, P., 2010. The circulation of the Mediterranean Sea: a historical review of experimental investigations. Adv. Oceanogr. Limnol., 1(1), 11-28.

Beriotto, A.C., Garz´on, M.J., Schweigmann, N., 2021. Is there a minimum number of landmarks that optimizes the geometric morphometric analysis of mosquito (Diptera, Culicidae) wings? J. Med. Entomol. 58(2), 576–587.

Bermant, P. C., Bronstein, M. M., Wood, R. J., Gero, S., and Gruber, D. F. (2019). Deep machine learning techniques for the detection and classification of sperm whale bioacoustics. Sci. Rep., 9(1), 1–10.

Berrar, D., 2018. Cross-validation. Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics, 1-3, 542–545.

Bertola, M., Mazzucato, M., Pombi, M., Montarsi, F., 2022. Updated occurrence and bionomics of potential malaria vectors in Europe: a systematic review (2000–2021). Parasites Vectors 15, 88.

Bietolini, S., Candura, F., Coluzzi, M., 2006. Spatial and long-term temporal distribution of the Anopheles maculipennis complex species in Italy. Parassitologia, 48, 581–608.

Bindoff, N.L., L Cheung, W.W., Kairo, J.G., Arístegui, J., Guinder, V.A., Hallberg, R., Hilmi Monaco, N., Jiao, N., saiful Karim, M., Levin, L., Acar, S., Jose Alava Ecuador, J., Allison, E., 2019. Changing Ocean, Marine Ecosystems, and Dependent Communities. IPCC.

Bittle, M., and Duncan, A., 2013. A review of current marine mammal detection and classification algorithms for use in automated passive acoustic monitoring. Proceedings of Acoustics, 1–8.

Boavida, J., Assis, J., Silva, I., and Serrão, E.A. (2016). Overlooked habitat of a vulnerable gorgonian revealed in the Mediterranean and Eastern Atlantic by ecological niche modelling. Scientific Reports, 6, 36460.

Boccolini, D., Menegon, M., Di Luca, M., Toma, L., Severini, F., Marucci, G., D'Amato, S., Caraglia, A., Maraglino, F.P., Rezza, G., Romi, R., Gradoni, L., Severini, C., 2020. Non-imported malaria in Italy: paradigmatic approaches and public health implications following an unusual cluster of cases in 2017. BMC Public Health, 20, 1–12.

Bohnenstiehl, D. R., Lyon, R. P., Caretti, O. N., Ricci, S. W., and Eggleston, D. B., 2018. Investigating the utility of ecoacoustic metrics in marine soundscapes. Journal of Ecoacoustics, *2*(2), 1.

Bomphrey, R.J., Nakata, T., Phillips, N., Walker, S.M., 2017. Smart wing rotation and trailing edge vortices enable high frequency mosquito flight. Nature, 544, 92–95.

Bookstein, F.L., 1991. Morphometric Tools for Landmark Data. Cambridge University Press, New York.

Bookstein, F.L., 1996. Combining the tools of geometric morphometrics. Adv. Morphometrics, 131–151.

Borowiec, M.L., Dikow, R.B., Frandsen, P.B., McKeeken, A., Valentini, G., White, A.E., 2022. Deep learning as a tool for ecology and evolution. Methods Ecol. Evol., 13, 1640– 1660.

Boscari, E., Abbiati, M., Badalamenti, F., Bavestrello, G., Benedetti-Cecchi, L., Cannas, R., Cau, A., Cerrano, C., Chimienti, G., Costantini, F., Fraschetti, S., Ingrosso, G., Marino, I.A.M., Mastrototaro, F., Papetti, C., Paterno, M., Ponti, M., Zane, L., Congiu, L., 2018. A population genomics insight by 2b-RAD reveals populations' uniqueness along the Italian coastline in *Leptopsammia pruvoti* (Scleractinia, Dendrophylliidae). Divers Distrib., 25, 1101-1117.

Bosch, S., and Fernandez, S., 2022. sdmpredictors: Species Distribution Modelling Predictor Datasets. R package version 0.2.12.

Bossuyt, B.T.; Janssen, C. R., 2005. Copper toxicity to different field-collected cladoceran species: Intra- and inter-species sensitivity. Environ. Pollut. 136, 145–154.

Bradfer-Lawrence, T., Gardner, N., Bunnefeld, L., Bunnefeld, N., Willis, S. G., Dent, D. H., 2019. Guidelines for the use of acoustic indices in environmental research. Methods in Ecology and Evolution, 10(10), 1796–1807.

Bramanti, L., Benedetti, M. C., Cupido, R., Cocito, S., Priori, C., Erra, F., Iannelli, M., and Santangelo, G., 2017. Demography of Animal Forests: The Example of Mediterranean Gorgonians. In Rossi, S., Bramanti, L., Gori, A., Orejas, C. (Eds.) Marine Animal Forests, Springer, Cham.

Breiman, L., 2001. Random forests. Mach. Learn., 45, 5–32.

Brodersen, K. H., Ong, C. S., Stephan, K. E., Buhmann, J. M., 2010. The balanced accuracy and its posterior distribution. 20th International Conference on Pattern Recognition, 3121–3124.

Brownscombe, J.W.; Griffin, L.P.; Morley, D.; Acosta, A.; Hunt, J.; Lowerre-Barbieri, S.K.; Adams, A.J.; Danylchuk, A.J.; Cooke, S.J. Application of machine learning algorithms to identify cryptic reproductive habitats using diverse information sources, Oecologia, 194, 283–298.

Bruce, P. 2014. A Practical Statistics for Data Scientists. O'Reilly Media: Sebastobol, CA, USA.

Brunetti, M., Magoga, G., Iannella, M., Biondi, M., Montagna, M., 2019. Phylogeography and species distribution modelling of Cryptocephalusbarii (Coleoptera: chrysomelidae): is this alpine endemic species close to extinction? Zookeys, 856, 3–25.

Burger, S.V., 2018. Introduction to Machine Learning With R: Rigorous Mathematical Analysis. Oreilly and Associates Inc.

Burns, C.W.; Schallenberg, M. Calanoid copepods versus cladocerans: Consumer effects on protozoa in lakes of different trophicstatus. Limnol. Oceanogr. 2001, 46, 1558–1565.

Byun, D., Hong, J., Saputra, Ko, J.H., Lee, Y.J., Park, H.C., Byun, B.K., Lukes, J.R, 2009. Wetting characteristics of insect wing surfaces. J. Bionic Eng., 6 (1), 63–70.

Cabral, R.B., Geronimo, R.C., 2018. How important are coral reefs to food security in the Philippines? Diving deeper than national aggregates and averages, Marine Policy,

Calzolari, M., Desiato, R., Albieri, A., Bellavia, V., Bertola, M., Bonilauri, P., Callegari, E., Canziani, S., Lelli, D., Mosca, A., Mulatti, P., Peletto, S., Ravagnan, S., Roberto, P., Torri, D., Pombi, M., Di Luca, M., Montarsi, F., 2021. Mosquitoes of the maculipennis complex in Northern Italy. Sci. Rep., 11(1), 1–12.

Campello, R.J.G.B., Moulavi, D., Sander, J, 2013. Density-based clustering based on hierarchical density estimates. In: Pei, J., Tseng, V.S., Cao, L., Motoda, H., Xu, G. (Eds.), Advances in Knowledge Discovery and Data Mining. PAKDD 2013. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg.

Carlson, R.E., 1977. A trophic state index for lakes1. Limnol. Oceanogr. 22, 361–369.

Caroselli, E., Gizzi, F., Prada, F., Marchini, C., Airi, V., Kaandorp, J., Falini, G., Dubinsky, Z., Goffredo, S., 2018. Low and variable pH decreases recruitment efficiency in populations of a temperate coral naturally present at a CO2 vent. ASLO, Limnol. Oceanogr, 64(3), 1059-1069.

Caroselli, E., Goffredo, S., 2014. Mediterranean Coral Population Dynamics: A Tale of 20 Years of Field Studies. In: Goffredo, S., Dubinsky, Z. (eds) The Mediterranean Sea. Springer, Dordrecht.

Caroselli, E., Zaccanti, F., Mattioli, G., Falini, G., Levy, O., Dubinsky, Z., Goffredo, S., 2012. Growth and demography of the solitary scleractinian coral Leptopsammia pruvoti along a sea surface temperature gradient in the Mediterranean Sea. PLoS ONE, 7(6).

Casas-Güell, E., Teixidó, N., Garrabou, J., Cebrian, E., 2015. Structure and biodiversity of coralligenous assemblages over broad spatial and temporal scales. Marine Biology, 162, 901–912.

Cebrian, E., Tomas, F., López-Sendino, P., Montserrat, V., and Ballesteros, E., 2018. Biodiversity influences invasion success of a facultative epiphytic seaweed in a marine forest. Biological Invasions 20, 2839–2848.

Celebi, M.E., Aydin, K., 2016. Unsupervised Learning Algorithms. Springer Cham.

Céréghino, R.; Boix, D.; Cauchie, H.-M.; Martens, K.; Oertli, B., 2014. The ecological role of ponds in a changing world. Hydrobiologia, 723, 1–6.

Cerrano, C., Bavestrello, G., 2008. Medium-term effects of die-off of rocky benthos in the Ligurian Sea. What can we learn from gorgonians? Chem. Ecol., 24(1), 73-82.

Cerrano, C., Arillo, A., Azzini, F., Calcinai, B., Castellano, L., Muti, C., Valisano, L., Zega, G., and Bavestrello, G., 2005. Gorgonian population recovery after a mass mortality event. Aquat. Conserv., 15(2), 147-157.

Cerrano, C., Bavestrello, G., Bianchi, C., Cattaneo-vietti, R., Bava, S., Morganti, C., Morri, C., Picco, P., Sara, G., Schiaparelli, S., Siccardi, A., Sponga, F., 2000. A catastrophic mass-mortality episode of gorgonians and other organisms in the Ligurian Sea (North-western Mediterranean), summer 1999. Ecol. Lett., 3, 284-293.

Céréghino, R., Biggs, J., Oertli, B.; Declerck, S., 2008. The ecology of European ponds: Defining the characteristics of a neglected freshwater habitat. Hydrobiologia, 597, 1-6.

Cerrano, C., R. Danovaro, R., Gambi C., Pusceddu, A., Riva, A., Schiaparelli, S., 2010. Gold coral (*Savalia savaglia*) and gorgonian forests enhance benthic biodiversity and ecosystem functioning in the mesophotic zone. Biodivers. Conserv., 19, 153-167.

Chamberlain, S., Barve, V., Mcglinn, D., Oldoni, D., Desmet, P., Geffert, L., Ram, K., 2022. rgbif: Interface to the Global Biodiversity Information Facility. R package version 3.7.2.

Chapman, H.F., Hughes, J.M., Ritchie, S.A., Kay, B.H., 2003. Population structure and dispersal of the freshwater mosquitoes *Culex annulirostris* and *Culex palpalis* (Diptera: culicidae) in Papua New Guinea and Northern Australia. J. Med. Entomol., 40(2), 165–169.

Charney, N. D., Record, S., Gerstner, B. E., Merow, C., Zarnetske, P. L., and Enquist, B. J. (2021). A Test of Species Distribution Model Transferability Across Environmental and Geographic Space for 108 Western North American Tree Species. Front. Ecol. Evol.

Chatpiyaphat, K., Sumruayphol, S., Dujardin, J.P., Samung, Y., Phayakkaphon, A., Cui, L., Ruangsittichai, J., Sungvornyothin, S., Sattabongkot, J., Sriwichai, P., 2020. Geometric morphometrics to distinguish the cryptic species Anopheles minimus and An. harrisoni in malaria hot spot villages, western Thailand. Med. Vet. Entomol, 35(3), 293-301.

Chen, T., Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., Li, Y., Yuan, J., 2022. xgboost: Extreme Gradient Boosting. R package version 1.5.2.1.

Chiappone, M., Sullivan, K. M., 1996. Distribution, abundance, and species composition of juvenile scleractinian corals in the Florida reef tract. Bull. Mar. Sci., 58(2), 555-569.

Chinchor, N., 1992. MUC-4 evaluation metrics. Proceedings of the 4th Conference on Message Understanding - MUC4, 22.

Chiu, S.H., Chen, C.C., Yuan, G.F., Lin, T.H., 2006. Association algorithm to mine the rules that govern enzyme definition and toclassify protein sequences. BMC Bioinform., 7, 304.

Chollet, F., Allaire, J. J. Deep Learning with R. Manning, 2018

Chon, T.-S.; Park, Y. S., Park, J. H. 2000. Determining temporal pattern of community dynamics by using unsupervised learning algorithms. Ecol. Model. 132, 151–166.

Christin, S., Hervet, E., Lecomte, N., 2019. Applications for deep learning in ecology. Methods Ecol. Evol. 10 (10), 1632–1644.

Clink, D. J., Klinck, H., 2020. Unsupervised acoustic classification of individual gibbon females and the implications for passive acoustic monitoring. Methods in Ecology and Evolution, 12(2), 328-341.

Coma, R., Ribes, M., 2003. Seasonal energetic constraints in Mediterranean benthic suspension feeders: effects at different levels of ecological organization. Oikos, 101, 205-215.

Coma, R., Pola, E., Ribes, M., Zabala, M., 2004. Long-term assessment of the patterns of mortality of a temperate octocoral in protected and unprotected areas: a contribution to conservation and management needs. Ecol. Appl., 14, 1466–1478.

Coma, R., Ribes, M., Serrano, E., Jiménez, E., Salat, J., Pascual, J., 2009. Global warming-enhanced stratification and mass mortality events in the Mediterranean. Proc. Natl. Acad. Sci. U.S.A, 106(15), 6176-618.

Combes, S.A., Daniel, T.L., 2003. Flexural stiffness in insect wings II. Spatial distribution and dynamic wing bending. J. Exp. Biol., 206(17), 2989–2997.

Convertino, M., Muñoz-Carpena, R., Chu-Agor, M. L., Kiker, G. A., Linkov, I., 2014. Untangling drivers of species distributions: Global sensitivity and uncertainty analyses of MaxEnt. Environ. Model. Softw., 51, 296-309.

Coppari, M., Zanella, C., Rossi, S., 2019. The importance of coastal gorgonians in the blue carbon budget. Sci. Rep., 9, 13550.

Costanza, R., D'Arge, R., De Groot, R., Farber, S., Grasso, M., Hannon, B., Limburg, K., Naeem, S., O'Neill, R. V., Paruelo, J., Raskin, R. G., Sutton, P., Van den Belt, M., 1998. The value of the world's ecosystem services and natural capital. Nature, 387, 253–260.

Couce, E., Ridgwell, A., Hendy, E. J., 2012. Environmental controls on the global distribution of shallow-water coral reefs. J. Biogeogr., 39(8), 1508–1523.

Couret, J., Moreira, D.C., Bernier, D., Loberti, A.M., Dotson, E.M., Alvarez, M., 2020. Delimiting cryptic morphological variation among human malaria vector species using convolutional neural networks. PLoS Negl. Trop. Dis. 14, e0008904.

Crisci, C.; Ghattas, B.; Perera, G. A., 2012. Review of supervised machine learning algorithms and their applications to ecological data. Ecol. Model. 240, 113–122.

D'Auria, G., Zavagno, F. Indagine sui Bodri della Provincia di Cremona. Monogr. Pianura, 3, 5–229.

Davis, T. R., Champion, C., Coleman, M. A., 2021. Climate refugia for kelp within an ocean warming hotspot revealed by stacked species distribution modelling. Marine Environmental Research, 166.

De Meester, L., Declerck, S., Stoks, R., Louette, G., Van de Meutter, F., De Bie, T., Michels, E., Brendonck, L., 2005. Ponds and pools asmodel systems in conservation biology, ecology and evolutionary biology.Aquat. Conserv. Mar. Freshw. Ecosyst., 15, 715–725.

De Morais, S.A., Moratore, C., Suesdek, L., Marrelli, M.T., 2010. Genetic-morphometric variation in Culex quinquefasciatus from Brazil and La Plata, Argentina. Mem. Inst. Oswaldo Cruz 105(5), 672–676.

De Ville d'Avray, L. T., Ami, D., Chenuil, A., David, R., Féral, J.P., 2019. Application of the ecosystem service concept at a small-scale: The cases of coralligenous habitats in the North-Western Mediterranean Sea. Mar. Pollut., 138, 160-170.

DeWitt, T.J., Scheiner, S.M., 2004. Phenotypic Plasticity: Functional and Conceptual Approaches. Oxford University Press, New York.

Diallo, M., Sangare, D., Levashina, E.A., 2019. Mosquito microevolution drives Plasmodium falciparum dynamics. Nat. Microbiol., 4(6), 941–947.

Diaz-Garcia, J.A., Ruiz, M.D. and Martin-Bautista, M.J., 2022. A survey on the use of association rules mining techniques in textual social media. Artif Intell Rev.

Diaz-Papkovich, A., Anderson-Trocme, L., Ben-Eghan, C., Gravel, S., 2019. UMAP reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts. PLoS Genet. 15 (11), e1008432

Dietterich, T.G., 2000. Ensemble methods in machine learning. In: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Springer Verlag, 1-15.

Dimoff, S. A., Halliday, W. D., Pine, M. K., Tietjen, K. L., Juanes, F., Baum, J.K., 2021. The utility of different acoustic indicators to describe biological sounds of a coral reef soundscape. Ecol. Indic., 124.

Dirks, J.H., Taylor, D., 2012. Veins improve fracture toughness of insect wings. PLoS One 7(8), e43411.

Dodson, S., Arnott, S., Cottingham, K. 2000. The relationship in lake communities between primary productivity and species richness. Ecology, 81, 2662–2679.

Dong, X., Yan, N., Wei, Y., 2018. Insect sound recognition based on convolutional neural network. IEEE 3rd International Conference on Image, Vision and Computing (ICIVC), 855–859.

Doorenweerd, C., San Jose, M., Barr, N., Leblanc, L., Rubinoff, D., 2020. Highly variable COI haplotype diversity between three species of invasive pest fruit fly reflects remarkably incongruent demographic histories. Sci. Rep., 10, 688.

Dorrity, M.W., Saunders, L.M., Queitsch, C., Fields, S., Trapnell, C., 2020. Dimensionality reduction by UMAP to visualize physical and genetic interactions. Nat. Commun., 11(1), 1–6.

Doubek, J.P., Campbell, K.L., Lofton, M.E., McClure, R.P., Carey, C.C., 2019. Hypolimnetic Hypoxia Increases the Biomass Variabilityand Compositional Variability of Crustacean Zooplankton Communities. Water, 11, 2179.

Downing, A.L.; Leibold, M.A., 2010. Species richness facilitates ecosystem resilience in aquatic food webs. Freshw. Biol. 55,2123–2137.

Dujardin, J.P., Bermudez, H., Casini, C., Schofield, I., Tibayrenct, M., 1997. Metric differences between silvatic and domestic Triatoma infestans (Heteroptera: Reduviidae) in Bolivia. J. Med. Entomol., 34, 544–551.

Dujardin, J.P., Kaba, D., Henry, A.B., 2010. The exchangeability of shape. BMC Res. Not. 3, 1–7.

Dunne, R.P., Brown, B.E., 2001. The influence of solar radiation on bleaching of shallow water reef corals in the Andaman Sea, 1993-1998. Coral Reefs, 20(3), 201–210.

Durette-Morin, D., Davies, K.T.A., Johnson, H. D., Brown, M. W., Moors-Murphy, H., Martin, B., and Taggart, C.T., 2019. Passive acoustic monitoring predicts daily variation in North Atlantic right whale presence and relative abundance in Roseway Basin, Canada. Mar. Mamm. Sci., 35(4), 1280–1303.

Dzialowski, A.R., 2013. Invasive zebra mussels alter zooplankton responses to nutrient enrichment. Freshw. Sci., 32, 462–470.

Edgar, R.C., 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinform. 5, 113.

Edmunds, P.J., 2000. Patterns in the distribution of juvenile corals and coral reef community structure in St John, US Virgin Islands. MEPS, 202, 113–124.

Eftestøl, S., Flydal, K., Tsegaye, D., Colman, J.E., 2019. Mining activity disturbs habitat use of reindeer in Finnmark, Northern Norway. Polar Biology, 42(10), 1849–1858.

Eldridge, A., Guyot, P., Moscoso, P., Johnston, A., Eyre-Walker, Y., Peck, M., 2018. Sounding out ecoacoustic metrics: Avian species richness is predicted by acoustic indices in temperate but not tropical habitats. Ecol. Indic., 95(1), 939–952.

Elise, S., Bailly, A., Urbina-Barreto, I., Mou-Tham, G., Chiroleu, F., Vigliola, L., Robbins, W. D., Bruggemann, J.H., 2019. An optimised passive acoustic sampling scheme to discriminate among coral reefs' ecological states. Ecological Indicators,107(February), 105627.

Elith, J., Leathwick, J.R., 2009. Species distribution models: Ecological explanation and prediction across space and time. Annu. Rev. Ecol. Evol. Syst., 40.

Elith, J., Graham, C.H., Anderson, P.R., Dudík, M., Ferrier, S., Guisan, A., Hijmans, J.R., Huettmann, F., Leathwick, R.J., Lehmann, A., Li, J., Lohmann, G.L., Loiselle, A.B., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J. McC. M., Townsend Peterson, A., Phillips, J.S., Richardson, K., Scachetti Pereira, R., Schapire, E.R., Soberón, J., Williams, S., Wisz, S. M., Zimmermann, E.N., 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography, 29, 129-151.

Elith, J., Leathwick, J. R., Hastie, T., 2008. A working guide to boosted regression trees. J. Anim. Ecol., 77, 802-813.

Elgendy, N., Elragal, A., 2014. Big Data Analytics: A Literature Review Paper. P. Perner (Ed.): ICDM 2014, LNAI 8557, pp. 214–227, Springer International Publishing, Switzerland.

Equihua, M. Fuzzy Clustering of Ecological Data. J. Ecol. 1990,78, 519.

Erbe, C., Marley, S.A., Schoeman, R.P., Smith, J.N., Trigg, L.E., Embling, C.B. (2019). The effects of ship noise on marine mammals-A review. Front. Mar. Sc., 6, 606.

Ezzat, L., Merle, P. L., Furla, P., Buttler, A., Ferrier-Pagès, C., 2013. The Response of the Mediterranean Gorgonian *Eunicella singularis* to Thermal Stress Is Independent of Its Nutritional Regime. PLOS ONE, 8(5), e64370.

Fairbrass, A. J., Rennert, P., Williams, C., Titheridge, H., Jones, K. E., 2017. Biases of acoustic indices measuring biodiversity in urban areas. Ecological Indicators, 83, 169–177.

Fantazzini, P., Mengoli, S., Pasquini, L. et al., 2015. Gains and losses of coral skeletal porosity changes with ocean acidification acclimation. Nat. Commun., 6, 7785.

Farley, S. S., Dawson, A., Goring, S. J., Williams, J.W., 2018. Situating Ecology as a Big-Data Science: Current Advances, Challenges, and Solutions, BioScience, 68(8), 563–576.

Farina, A., Gage, S.H., 2017. Ecoacoustics: the ecological role of sounds. In A. Farina and S. H. Gage (Eds.), Ecoacoustics: The ecological role of sounds, 1-366. John Wiley and Sons, Ltd.

Fava, F., Bavestrello, G., Valisano, L., and Cerrano, C, 2010. Survival, growth, and regeneration in explants of four temperate gorgonian species in the Mediterranean Sea. Ital. J. Zool., 77(1), 44-52.

Fenner, D., Riolo, F., Vittorio, M., 2013. New records of scleractinian corals from shallow waters of the Ionian coast of Italy. Mar. Biodivers. Rec., 6, E136.

Filatova, O. A., Miller, P. J. O., Yurk, H., Samarra, F. I. P., Hoyt, E., Ford, J. K. B., Matkin, C. O., and Barrett-Lennard, L.G., 2015. Killer whale call frequency is similar across the oceans but varies across sympatric ecotypes. J. Acoust. Soc. Am., 138(1), 251–257.

Fiorentino, D., Pesch, R.; Guenther, C. P., Gutow, L., Holstein, J., Dannheim, J., Ebbe, B., Bildstein, T., Schroeder, W., Schuchardt, B., et al., 2017. A 'fuzzy clustering' approach to conceptual confusion: How to classify natural ecological associations. MEPS, 584, 17–30.

Fitt, W. K., Brown, B. E., Warner, M. E., Dunne, R.P., 2001. Coral bleaching: interpretation of thermal tolerance limits and thermal thresholds in tropical corals. Coral Reefs, 20, 51–65.

Flecha, S., Pérez, F.F., García-Lafuente, J., Sammartino, S., Ríos, A. F., Huertas, I.E., 2015. Trends of pH decrease in the Mediterranean Sea through high frequency observational data: indication of ocean acidification in the basin. Sci. Rep., 5(1), 1–8.

Fleming, A.H., Yack, T., Redfern, J.V., Becker, E. A., Moore, T. J., Barlow, J., 2018. Combining acoustic and visual detections in habitat models of Dall's porpoise. Ecol. Model., 384, 198–208.

Flemons, P., Guralnick, R., Krieger, J., Ranipeta, A., Neufeld, D., 2007. A web-based GIS tool for exploring the world's biodiversity: the global biodiversity information facility mapping and analysis portal application (GBIF-MAPA). Ecol. Inform., 2, 49-60.

Fortin, M. J., Dale, M.R.T., 2005. Spatial Analysis: A Guide for Ecologists. Cambridge University Press, Cambridge, 1-30.

Fox, R.J., Donelson, J.M., Schunter, C., Ravasi, T., Gaitan-Espitia, J.D., 2019. Beyond buying time: the role of plasticity in phenotypic adaptation to rapid environmental change. Philos. Trans. R. Soc. Lond. B Biol. Sci., 374 (1768).

Francuski, L., Gojkovic, N., Krtinic, B., Milankov, V., 2019. The diagnostic utility of sequence-based assays for the molecular delimitation of the epidemiologically relevant Culex pipiens pipiens taxa (Diptera: culicidae). Bull. Entomol. Res., 109(6), 752–761.

Frank, E.; Hal, L.M.A.; Witten, I.H. The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools andTechniques, 4th ed.; Morgan Kaufmann: Burlington, MA, USA, 2016.

Frankenhuis, W. E., Panchanathan, K., and Barto, A. G., 2019. Enriching behavioral ecology with reinforcement learning methods. Behav. Processes, 161, 94-100.

Franklin, J., 2010. Moving beyond static species distribution models in support of conservation biogeography. Divers. Distrib., 16, 321-330.

Freeman, E. A., Moisen, G., 2008. PresenceAbsence: An R Package for Presence-Absence Model Analysis. J. Stat. Softw., 23(11), 1-31.

Freitas, A.A., 1998. On objective measures of rule surprisingness. InTransactions on Petri Nets and Other Models of Concurrency XV; Springer Science and Business Media LLC: Berlin/Heidelberg, Germany, Vol. 1510.

Friederichs, M., Fränzle, O., Salski, A. 1996. Fuzzy clustering of existing chemicals according to their ecotoxicological properties. Ecol. Model., 85, 27–40.

Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent. J. Stat. Softw., 33(1), 1-22.

Friedman, J.H., Popescu, B.E., 2008. Predictive learning via rule ensembles, The Annals of Applied Statistics,

Froelich A.S., 2002, Functional aspects of nutrient cycling on coral reefs, Rosenstiel School of Marine and Atmospheric Science, NOAA Miami Regional Library.

Gomez, G.F., Correa, M.M., 2017. Discrimination of Neotropical Anopheles species based on molecular and wing geometric morphometric traits. Infect. Genet. Evol., 54, 379–386.

Galil, B.S., Danovaro, R., Rothman, S.B.S., Gevili, R., Goren, M., 2019. Invasive biota in the deep-sea Mediterranean: an emerging issue in marine conservation and management. Biol. Invasions, 21, 281–288.

Gambi, M.C., Sorvino, P., Tiberti, L., Gaglioti, M., Teixido, N., 2018. Mortality events of benthic organisms along the coast of Ischia in summer 2017. Biol. Mar. Mediterr, 25 (1). 212-213.

Garrabou, J., and Harmelin, J.G., 2002. A 20-year study on life-history traits of a harvested long-lived temperate coral in the NW Mediterranean: insights into conservation and management needs. J. Anim. Ecol., 71, 966-978.

Garros, C., Dujardin, J.P., 2013. Genetic and phenetic approaches to Anopheles systematics. In: Manguin, S. (Ed.), Anopheles Mosquitoes - New Insights into Malaria Vectors. InTech, Rijeka, 81–105.

Gaston, K.J., Blackburn, T.M., 2000. Pattern and Process in Macroecology; Gaston, K.J., Eds: Blackburn, T.M., Wiley: Hoboaken, NY, USA.

GBIF.org, 2021, *B. europaea*, GBIF Occurrence Download, https://doi.org/10.15468/dl.ysmfdk, (accessed 28 October 2021)

GBIF.org, 2021, GBIF Home Page. Available from: https://www.gbif.org

GBIF.org, 2021, L. pruvoti, GBIF Occurrence Download https://doi.org/10.15468/dl.un4386, (accessed 28 October 2021)

Geiger, J.T., Schuller, B., Rigoll, G., 2013. Large-scale audio feature extraction and SVM for acoustic scene classification. 2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 1–4.

General Bathymetric Charts of the Oceans (GEBCO), 2021. https://www.gebco.net/data_and_products/gridded_bathymetry_data/, (accessed in October 2021)

Geng, L., Hamilton, H.J., 2006. Interestingness measures for data mining. ACM Comput. Surv., 38(3), 24.

Géron, A., 2019. Hands-On Machine Learning with Scikit-Learn, Keras, and Tensorflow: Concepts, Tools, and Techniques to Build Intelligent Systems. Oreilly and Associates Inc.

Gerovasileiou, V., Chintiroglou, C., Vafidis, D., Koutsoubas, D., Sini, M., Dailianis, T., Issaris, Y., Akritopoulou, E., Dimarchopoulou, D., Voutsiadou, E., 2015. Census of biodiversity in marine caves of the eastern Mediterranean Sea. Mediterranean Marine Science, 16(1), 245–265.

Gianuca, A.T., Declerck, S.A.J., Lemmens, P., De Meester, L., 2017. Effects of dispersal and environmental heterogeneity on the replacement and nestedness components of β-diversity. Ecology, 98, 525-533.

Gibb, R., Browning, E., Glover-Kapfer, P., Jones, K.E., 2019. Emerging opportunities and challenges for passive acoustics in ecological assessment and monitoring. Methods Ecol. Evol., 10(2), 169-185.

Gibson, G., Warren, B., Russell, I.J., 2010. Humming in Tune: Sex and Species Recognition by Mosquitoes on the Wing. JARO, 11, 527–540.

Gilchrist, A.S., Partridge, L., 2001. The contrasting genetic architecture of wing size and shape in Drosophila melanogaster. Heredity, 86 (2), 144–152.

Gildenhard, M., Rono, E.K., Diarra, A., Boissi`ere, A., Bascunan, P., Carrillo-Bustamante, P., Camara, D., Krüger, H., Mariko, M., Mariko, R., Mireji, P., Nsango, S.E., Pompon, J., Reis, Y., Rono, M.K., Seda, P.B., Thailayil, J., Traore, A., Yapto, C.V., Awono-Ambene, P., Dabire, R.K., Diabat´e, A., Masiga, D., Catteruccia, F., Morlais, I., 2019. Mosquito microevolution drives Plasmodium falciparum dynamics. Nat Microbiol., 4, 941–947.

Goffredo, S., Caroselli, E., Mattioli, G., Pignotti, E., Dubinsky, Z., Zaccanti, F., 2009. Inferred level of calcification decreases along an increasing temperature gradient in a Mediterranean endemic coral. Limnol. Oceanogr., 54, 930-937.

Goffredo, S., Caroselli, E., Mattioli, G., Pignotti, E., Zaccanti, F., 2008. Relationships between growth, population structure and sea surface temperature in the temperate solitary coral Balanophyllia europaea (Scleractinia, Dendrophylliidae). Coral Reefs, 27(3).

Goffredo, S., Caroselli, E., Pignotti, E., Mattioli, G., Zaccanti, F., 2007. Variation in biometry and demography of solitary corals with environmental factors in the Mediterranean Sea. Mar Biol. 152:351–361.

Goffredo, S., Mancuso, A., Caroselli, E., Prada, F., Dubinsky, Z., Falini, G., Levy, O., Fantazzini, P., Pasquini, L., 2015. Skeletal mechanical properties of Mediterranean corals along a wide latitudinal gradient. Coral Reefs, 34(1), 121–132.

Goffredo, S., Mattioli, G., Zaccanti, F., 2004. Growth and population dynamics model of the Mediterranean solitary coral *Balanophyllia europaea* (Scleractinia, Dendrophylliidae). Coral Reefs, 23, 433-443.

Gómez, W.E., Isaza, C.V., Daza, J.M., 2018. Identifying disturbed habitats: A new method from acoustic indices. Ecol. Inform., 45, 16–25.

Goodal, C., 1991. Procrustes methods in the statistical analysis of shape. J. R. Stat. Soc. Ser. C 53 (2), 285–339.

Goodal, C., 1991. Procrustes methods in the statistical analysis of shape. J. R. Stat. Soc., 53, 285–339.

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. The MIT Press.

Goodwin, A., Padmanabhan, S., Hira, S., Glancey, M., Slinowsky, M., Immidisetti, R., Scavo, L., Brey, J., Murali, B., Sai, M., Ford, T., Heier, C., Linton, Y.M., Pecor, D.B., Quiroga, L.C., Acharya, S., 2021. Mosquito species identification using convolutional neural networks with a multitiered ensemble model for novel species detection. Sci.Rep., 1–15.

Gori, A., Grinyó, J., Dominguez-Carrió, C., Ambroso, S., López-González, P. J., Gili, J.-M., Bavestrello, G., and Bo, M., 2019. Gorgonian and Black Coral Assemblages in Deep Coastal Bottoms and Continental

Shelves of the Mediterranean Sea. In Orejas, C., Jiménez, C., (Eds.), Mediterranean Cold-Water Corals: Past, Present and Future. Coral Reefs of the World, (Vol 9.). Springer, Cham.

Gori, A., Rossi, S., Berganzo, E., Pretus, J.L., Dale, M.R.T., Gili, J.M., 2011. Spatial distribution patterns of the gorgonians *Eunicella singularis*, *Paramuricea clavata*, and *Leptogorgia sarmentosa* (Cape of Creus, Northwestern Mediterranean Sea). Marine Biology, 158, 143–158.

Guan, R., Wang, W.X., 2006. Multigenerational cadmium acclimation and biokinetics in Daphnia magna. Environ. Pollut., 141,343–352.

Guisan, A., Thuiller, W., 2005. Predicting species distribution: offering more than simple habitat models. Ecol. Lett., 8, 993– 1009.

Guisan, A., Zimmermann, N.E., 2000. Predictive habitat distribution models in ecology. Ecol. Modell., 135(2–3), 147-186.

Guralnick, R., Hill, A., 2009. Biodiversity informatics: automated approaches for documenting global biodiversity patterns and processes. Bioinformatics, 25, 421-428.

Gustafson, J.L., 2011. Moore's Law. In: Padua, D. (eds) Encyclopedia of Parallel Computing. Springer, Boston, MA.

Gyllström, M., Hansson, L.A., Jeppesen, E., Criado, F.G., Gross, E., Irvine, K., Kairesalo, T., Kornijow, R., Miracle, M.R., Nykänen, M., et al., 2005. The role of climate in shaping zooplankton communities of shallow lakes. Limnol. Oceanogr., 50, 2008–2021.

Hahsler, M., 2019. ArulesViz: Visualizing Association Rules and Frequent Itemsets. R Package Version 1.3-3.

Hahsler, M., Buchta, C., Gruen, B., Hornik, K., 2020. Arules: Mining Association Rules and Frequent Itemsets. R Package Version 1.6-6.

Han, J.; Cheng, H.; Xin, D., Yan, X., 2007. Frequent pattern mining: Current status and future directions. Data Min. Knowl. Discov., 15, 55–86.

Hanazato, T., 1991. Influence of food density on the effects of a *Chaoborus*-released chemical on *Daphnia ambigua*. Freshw. Biol., 25, 477–483.

Harmelin-Vivien, M.L., 1994. The Effects of Storms and Cyclones on Coral Reefs: A Review, Journal of Coastal Research, Special Issue No. 12. Coastal Hazards: Perception, Susceptibility and Mitigation (1994), pp. 211-231; Published by: Coastal Education and Research Foundation, Inc.

Harrington, P., 2012. Machine Learning in Action. Manning Publications Co., Shelter Island, New York.

Hassall, C., 2014. The ecology and biodiversity of urban ponds. Wiley Interdiscip. Rev. Water, 1, 187–206.

Havens, K.E.; Hanazato, T., 1993. Zooplankton community responses to chemical stressors: A comparison of results from acidification and pesticide contamination research. Environ. Pollut., 82, 277–288.

He, H., Ma, Y., 2013. Imbalanced Learning: Foundations, Algorithms, and Applications. John Wiley and Sons, Inc.

Heath, B. E., Sethi, S. S., Orme, C. D. L., Ewers, R. M., Picinali, L., 2021. How index selection, compression, and recording schedule impact the description of ecological soundscapes. Ecol. Evol., 11(19), 13206–13217.

Hébert, M.P., Beisner, B.E., Maranger, R., 2017. Linking zooplankton communities to ecosystem functioning: Toward an effect-trait framework. J. Plankton Res., 39, 3–12.

Heino, J.; Grönroos, M.; Ilmonen, J.; Karhu, T.; Niva, M.; Paasivirta, L. 2013. Environmental heterogeneity and β diversity of streammacro invertebrate communities at intermediate spatial scales. Freshw. Sci.,32, 142–154.

Henry, A., Thongsripong, P., Fonseca-Gonzalez, I., Jaramillo-Ocampo, N., Dujardin, J.P., 2010. Wing shape of dengue vectors from around the world. Infect. Genet. Evol., 10 (2), 207–214.

Hershey, S., Chaudhuri, S., Ellis, D. P. W., Gemmeke, J. F., Jansen, A., Moore, R. C., Plakal, M., Platt, D., Saurous, R.A., Seybold, B., Slaney, M., Weiss, R. J., Wilson, K., 2017. CNN architectures for large-scale audio classification. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing *(ICASSP)*, 131–135.

Hershey, S., Chaudhuri, S., Ellis, D.P.W., Gemmeke, J.F., Jansen, A., Moore, R.C., Plakal, M., Platt, D., Saurous, R.A., Seybold, B., Slaney, M., Weiss, R.J., Wilson, K., 2017. CNN architectures for large-scale audio classification. In: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 131–135.

Hirakawa, T., Yamashita, T., Tamaki, T., Fujiyoshi, H., Umezu, Y., Takeuchi, I., Matsumoto, S., Yoda, K., 2018. Can AI predict animal movements? Filling gaps in animal trajectories using inverse reinforcement learning. Ecosphere, 9(10).

Hirzel, A.H., Le Lay, G., Helfer, V., Randin, C., Guisan, A., 2006. Evaluating the ability of habitat suitability models to predict species presences. Ecol. Modell., 199(2), 142-152.

Hoegh-Guldberg, O., 2014. Coral reef sustainability through adaptation: glimmer of hope or persistent mirage? Curr Opin Environ Sustain, 7, 127-133.

Hoffmann, A.A., Shirriffs, J., 2002. Geographic variation for wing shape in Drosophila serrata. Evolution, 56 (5), 1068–1073.

Höppner, F. Association Rules. In Data Mining and Knowledge Discovery Handbook; Springer: Cham, Switzerland, 2009.

Houegnigan, L., Safari, P., Nadeu, C., Van Der Schaar, M., Andre, M., 2017. A novel approach to real-time range estimation of underwater acoustic sources using supervised machine learning. OCEANS, 1–5.

Hu, Z., Hu, J., Hu, H., Zhou, Y., 2020. Predictive habitat suitability modeling of deep-sea framework-forming scleractinian corals in the Gulf of Mexico. Science of the Total Environment, 742.

Huete-Stauffer, C., Vielmini, I., Palma, M., Navone, A., Panzalis, P., Vezzulli, L., Misic, C., Cerrano, C., 2011. *Paramuricea clavata* (Anthozoa, Octocorallia) loss in the Marine Protected Area of Tavolara (Sardinia, Italy) due to a mass mortality event. Marine Ecology, 32, 107-116.

Hughes, T., Kerry, J., Álvarez-Noriega, M. et al., 2017. Global warming and recurrent mass bleaching of corals. Nature, 543, 373–377.

Humphries, G.R.W., Huettmann, F., 2018. Machine Learning in Wildlife Biology: Algorithms, Data Issues and Availability, Workflows, CitizenScience, Code Sharing, Metadata and a Brief Historical Perspective; J.B. Metzler, pp. 3–26, Stuttgart, Germany.

Hunter, K.; Pyle, G. 2004. Morphological Responses of *Daphnia pulex* to *Chaoborus americanus* Kairomone in the Presence and Absence of Metals. Environ. Toxicol. Chem., 23, 1311–1316.

Huston, M. A., 1985. Patterns of Species Diversity on Coral Reefs. Annual Review of Ecology and Systematics, 16, 149–77.

Iborra, L., Leduc, M., Fullgrabe, L., Cuny, P., Gobert, S., 2022. Temporal trends of two iconic Mediterranean gorgonians (*Paramuricea clavata* and *Eunicella cavolini*) in the climate change context. J. Sea Res., 186, 102241.

Iii, F.S.C., Zavaleta, E.S., Eviner, V.T., Naylor, R.L., Vitousek, P.M., Reynolds, H.L., Hooper, D.U., Lavorel, S., Sala, O.E., Hobbie, S.E., et al., 2000. Consequences of changing biodiversity. Nat. Cell Biol., 405, 234-242.

Illian, J., Penttinen, A., Stoyan, H., Stoyan, D., 2008. Statistical analysis and modelling of spatial point patterns. pp. 560, Wiley, Chichester,.

International Union for Conservation of Nature (IUCN), 2021. The IUCN Red List of Threatened Species. Version 2021-3. https://www.iucnredlist.org. (accessed on 3 April 2022).

Ipcc Expert Meeting Report, 2007. Towards New Scenarios for Analysis of Emissions, Climate Change, Impacts, And Response Strategies.

Ivanova, E.P., Hasan, J., Webb, H.K., Gervinskas, G., Juodkazis, S., Truong, V.K., Wu, A. H.F., Lamb, R.N., Baulin, V.A., Watson, G.S., Watson, J.A., Mainwaring, D.E., Crawford, R.J., 2013. Bactericidal activity of black silicon. Nat. Commun., 4(1), 1–7.

Jackson, A., 2008. *Leptopsammia pruvoti* Sunset cup coral. In Tyler-Walters H. and Hiscock K. Marine Life Information Network: Biology and Sensitivity Key Information Reviews. Plymouth: Marine Biological Association of the United Kingdom.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2014. An Introduction to Statistical learning. Second Edition, Springer Texts in Statistics.

Jaramillo-O, N., Dujardin, J.P., Calle-Londono, D., Fonseca-Gonz´alez, I., 2014. Geometric morphometrics for the taxonomy of 11 species of Anopheles (Nyssorhynchus) mosquitoes. Med. Vet. Entomol., 29, 26–36.

Jenkins, T. L., Stevens, J. R., 2022. Predicting habitat suitability and range shifts under projected climate change for two octocorals in the north-east Atlantic. PeerJ, 10, e13509.

Jian, Y., Silvestri, S., Belluco, E., Saltarin, A., Chillemi, G., Marani, M., 2014. Environmental forcing and density-dependent controls of Culex pipiens abundance in a temperate climate (Northeastern Italy). Ecol. Model., 272, 301–310.

Jimenez-Valverde, A., 2012. Insights into the area under the receiver operating characteristic curve (AUC) as a discrimination measure in species distribution modelling. Glob. Ecol. Biogeogr., 21, 498–507.

Jones R., Fisher, R., Bessell-Browne, P., 2019. Sediment deposition and coral smothering. PLOS ONE, 14(6): e0216248.

Jones, C. G., Lawton, J. H., and Shachak, M., 1994. Organisms as Ecosystem Engineers. Oikos, 69(3), 373-386.

Kareemi, T.I., Nirankar, J.K., Mishra, A.K., Chand, S.K., Chand, G., Vishwakarma, A.K., Tiwari, A., Bharti, P.K., 2021. Population dynamics and insecticide susceptibility of *Anopheles culicifacies* in Malaria endemic districts of Chhattisgarh, India. Insects, 12(4), 284.

Karlson, R. H., 2006. Metapopulation dynamics and community ecology of marine systems. In J. P. Kritzer and P.F. Sale (Eds.), Marine metapopulations (pp. 457–515), Academic Press, Burlington.

Kim, D., DeBriere, T.J., Cherukumalli, S., White, G.S., Burkett-Cadena, N.D., 2021. Infrared light sensors permit rapid recording of wingbeat frequency and bioacoustic species identification of mosquitoes. Sci. Rep., 11(1), 1–9.

Kimura, M.A., 1980. Simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol., 16, 111–120.

Kingsolver, J.G., Buckley, L.B., 2017. Quantifying thermal extremes and biological variation to predict evolutionary responses to changing climate. Philos. Trans. Royal Soc. B., 372(1723).

Kinlan, B.P., Poti, M., Drohan, A.F., Packer, D.B., Dorfman, D.S., Nizinski, M.S., 2020. Predictive modeling of suitable habitat for deep-sea corals offshore the Northeast United States, Deep Sea Research Part I: Oceanographic Research Papers, 158, 103229.

Kline, D.I., Teneva, L., Okamoto, D.K., Schneider, K., Caldeira, K., Miard, T., Chai, A., Marker, M., Dunbar, R.B., Mitchell, B.G., Dove, S., Hoegh-Guldberg, O., 2019. Living coral tissue slows skeletal dissolution related to ocean acidification. Nat Ecol Evol, 3, 1438–1444.

Klingenberg, C.P., 2013. Visualizations in geometric morphometrics: how to read and how to make graphs showing shape changes. Hystrix, Italian J. Mammal., 24(1), 15–24.

Klingenberg, C.P., 2016. Size, shape, and form: concepts of allometry in geometric morphometrics. Dev. Genes Evol., 226, 113–137.

Klingenberg, C.P., Marugan Lobon, J., 2013. Evolutionary covariation in geometric morphometric data: analyzing integration, modularity, and allometry in a phylogenetic context. Syst. Biol. 62, 591–610.

Knutson, T.R., Chung, M.V., Vecchi, G., Sun, J., Hsieh, T.L., Smith, A.J., 2021. Climate change is probably increasing the intensity of tropical cyclones. In Critical issues in climate change science. Science Brief Review.

Konowalik, K., Nosol, A., 2015. Evaluation metrics and validation of presence-only species distribution models based on distributional maps with varying coverage. Sci. Rep., 11(1), 1482.

Kotsiantis, S., 2007. Supervised machine learning: A review of classification techniques. Informatica Journal, 31, 249–268.

Kowarski, K.A., Moors-Murphy, H., 2020. A review of big data analysis methods for baleen whale passive acoustic monitoring. Mar. Mamm. Sci., 37(2), 652–673.

Kowarski, K.A., Wilson, C.C., Delarue, J.J.Y., Martin, S.B., 2021. Flemish Pass acoustic monitoring: ambient characterization and marine mammal monitoring near a mobile offshore drilling unit. Technical report by JASCO Applied Sciences for Wood Environment and Infrastructure Solutions.

Kraft, B., Besnard, S. Koirala, S., 2021. Emulating Ecological Memory with Recurrent Neural Networks. In Deep Learning for the Earth Sciences. Eds Camps-Valls, G., Tuia, D., Zhu, X.X., Reichstein, M.

Krishna, S., Cho, M., Wehmann, H.N., Engels, T., Lehmann, F.O., 2020. Wing design in flies: properties and aerodynamic function. Insects, 11(8), 466.

Kruk, C., Rodríguez-Gallego, L., Meerhoff, M., Quintans, F., Lacerot, G., Mazzeo, N., Scasso, F., Paggi, J.C., Peeters, E.T.H.M., Marten, S., 2009. Determinants of biodiversity in subtropical shallow lakes (Atlantic coast, Uruguay). Freshw. Biol., 54, 2628–2641.

Kruzic, P., Popijac, A., 2015. Mass mortality events of the coral Balanophyllia europaea (Scleractinia, Dendrophylliidae) in the Mljet National Park (Eastern Adriatic Sea) caused by sea temperature anomalies. Coral Reefs, 34, 109–118.

Kuhn, M., 2008. Building predictive models in R using the caret package. J. Stat. Softw., 28, 1–26.

Kuitunen, K., Kovalev, A., Gorb, S.N., 2014. Sex-related effects in the superhydrophobic properties of damselfly wings in young and old Calopteryx splendens. PLoS One, 9(2), e88627.

Kunc, H.P., Schmidt, R., 2019. The effects of anthropogenic noise on animals: A meta-analysis. Biol. Lett., 15(11), 20190649.

Kunc, H.P., McLaughlin, K.E., Schmidt, R., 2016. Aquatic noise pollution: implications for individuals, populations, and ecosystems. Proceedings of the Royal Society B: Biological Sciences, 283(1836), 20160839.

Kupfner Johnson, S., Hallock, P., 2020. A review of symbiotic gorgonian research in the western Atlantic and Caribbean with recommendations for future work. Coral Reefs, 39, 239–258.

Laffoley, D., Baxter, J. M., 2016. Explaining Ocean Warming: Causes, scale, effects and consequences. IUCN, Gland, Switzerland.

Lapeyrolerie, M., Chapman, M. S., Norman, K. E. A., Boettiger, C., 2022. Deep reinforcement learning for conservation decisions. Methods Ecol. Evol., 13, 2649– 2662.

Lauria, V., Garofalo, G., Fiorentino, F., Massi, D., Milisenda, G., Piraino, S., Russo, T., Gristina, M., 2017. Species distribution models of two critically endangered deep-sea octocorals reveal fishing impacts on vulnerable marine ecosystems in central Mediterranean Sea. Sci. Rep., 7, 8049.

Le, S., Josse, J., Husson, F., 2008. FactoMineR: an R package for multivariate analysis. J. Stat. Softw., 25(1), 1–18.

LeBien, J., Zhong, M., Campos-Cerqueira, M., Velev, J. P., Dodhia, R., Ferres, J. L., Aide, T. M., 2020. A pipeline for identification of bird and frog species in tropical soundscape recordings using a convolutional neural network. Ecol. Inform., 59, 101113.

Lee, K.Y.; Chung, N.; Hwang, S. 2016. Application of an artificial neural network (ANN) model for predicting mosquito abundances inurban areas. Ecol. Inform., 36, 172–180.

Lek, S., Guegan, J.F., 1999. Artificial neural networks as a tool in ecological modelling, an introduction. Ecol. Model. 120, 65–73.

Lemaître, G., Nogueira, F., Aridas, C. K., 2017. Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning. J. Mach. Learn. Res., 18(1), 559-563.

Li, W., Yin, Y., Quan, X., Zhang, H., 2019. Gene Expression Value Prediction Based on XGBoost Algorithm. Front. Genet. 10, 1077.

Liconti, A., Pittman, S. J., Rees, S.E., Mieszkowska, N., 2022. Identifying conservation priorities for gorgonian forests in Italian coastal waters with multiple methods including citizen science and social media content analysis. Divers. Distrib., 28, 1430–1444.

Lilja, T., Eklof, D., Jaenson, T.G.T., Lindstrom, A., Terenius, O., 2020. Single nucleotide polymorphism analysis of the ITS2 region of two sympatric malaria mosquito species in Sweden: anopheles daciae and Anopheles messeae. Med. Vet. Entomol. 34(3), 364–368.

Linares, C., Coma, R., Diaz, D., Zabala, M., Hereu, B., Dantart, L., 2005. Immediate and delayed effects of a mass mortality event on gorgonian population dynamics and benthic community structure in the NW Mediterranean Sea. MEPS, 305, 127– 137.

Linares, C., Coma, R., Garrabou, J., Díaz, D., Zabala, M., 2008. Size distribution, density and disturbance in two Mediterranean gorgonians: *Paramuricea clavata* and *Eunicella singularis*. J. Appl. Ecol., 45, 688-699.

Linares, C., Doak, D.F., Coma, R., Díaz, D., Zabala, M., 2007. Life history and viability of a long-lived marine invertebrate: the octocoral *Paramuricea clavata*. Ecology, 88, 918-928.

Linton, Y.M., Smith, L., Harbach, R., 2002. Observations on the taxonomic status of *Anopheles subalpinus* Hackett and Lewis and *An. melanoon* Hackett. Eur. Mosq. Bull. 13.

Lischeid, G., Kalettka, T., Holländer, M., Steidl, J., Merz, C., Dannowski, R., Hohenbrink, T., Lehr, C., Onandia, G., Reverey, F., Patzig, M., 2018. Natural ponds in an agricultural landscape: External drivers, internal processes, and the role of the terrestrial-aquaticinterface. Limnologica, 68, 5–16.

Liu, C., White, M., Newell, G., 2013. Selecting thresholds for the prediction of species occurrence with presence-only Data. J. Biogeogr., 40(4), 778-789.

Lloyd A. Courtenay, José Yravedra, Rosa Huguet, Julia Aramendi, Miguel Ángel Maté-González, Diego González-Aguilera, Mari Carmen Arriaza, 2019. Combining machine learning algorithms and geometric morphometrics: A study of carnivore tooth marks. Palaeogeography, Palaeoclimatology, Palaeoecology. 522, 28-39.

Lorenz, C., Almeida, F., Almeida-Lopes, F., Louise, C., Pereira, S.N., Petersen, V., Vidal, P.O., Virginio, F., Suesdek, L., 2017. Geometric morphometrics in mosquitoes: what has been measured? Infect. Genet. Evol., 54, 205–215.

Lorenz, C., Ferraudo, A.S., Suesdek, L., 2015a. Artificial neural network applied as a methodology of mosquito species identification. Acta Trop., 152, 165–169.

Lorenz, C., Marques, T.C., Sallum, M.A.M., Suesdek, L., 2012. Morphometrical diagnosis of the malaria vectors Anopheles cruzii, An. homunculus and An. bellator. Parasites Vectors, 5(1), 2–8.

Lorenz, C., Marques, T.C., Sallum, M.A.M., Suesdek, L., 2012. Morphometrical diagnosis of the malaria vectors Anopheles cruzii, An. homunculus and An. bellator. Parasites Vectors, 5, 2–8.

Lorenz, C., Patane, J.S.L., Suesdek, L., 2015b. Morphogenetic characterization, date of divergence, and evolutionary relationships of malaria vectors *Anopheles cruzii* and *Anopheles homunculus*. Infect. Genet. Evol., 35, 144–152.

Luís, A.R., Collado, L.J.M., Gospić, N.R., Gridley, T., Papale, E., Azevedo, A., 2021. Vocal universals and geographic variations in the acoustic repertoire of the common bottlenose dolphin. Sci. Rep., 11(1), 1–9.

Lumini, A., Nanni, L., Maguolo, G., 2019. Deep learning for plankton and coral classification. Applied Computing and Informatics.

Lumini, A.; Nanni, L 2019. Deep learning and transfer learning features for plankton classification. Ecol. Inform., 51, 33–43.

Luo, J., Siemers, B. M., Koselj, K., 2015. How anthropogenic noise affects foraging. Glob. Chang. Biol., 21(9), 3278–3289.

Lushchak, V.I., 2011. Environmentally induced oxidative stress in aquatic animals, Aquatic Toxicology, ISSN 0166-445X.

Madhusudhana, S. K., Chakraborty, B., Latha, G., 2019. Humpback whale singing activity off the Goan coast in the Eastern Arabian Sea. Bioacoustics, 28(4), 329–344.

Magoga, G., Fontaneto, D., Montagna, M., 2021. Factors affecting the efficiency of molecular species delimitation in a species-rich insect family. Mol. Ecol. Resour., 21(5), 1475–1489.

Magoga, G., Sahin, D.C., Fontaneto, D., Montagna, M., 2018. Barcoding of Chrysomelidae of Euro-Mediterranean area: efficiency and problematic species. Sci. Rep., 8(1), 13398.

Mandal, J.K., Bhattacharya, D. 2020. Emerging Technology in Modelling and Graphics, Proceedings of IEM Graph 2018. Advances in Intelligent Systems and Computing. Springer Nature Singapore.

Manguin, S., Bangs, M.J., Pothikasikorn, J., Chareonviriyaphap, T., 2010. Review on global co-transmission of human Plasmodium species and *Wuchereria bancrofti* by *Anopheles* mosquitoes. Infection. Genet. Evol. 10(2), 159–177.

Manguin, S., Garros, C., Dusfour, I., Harbach, R.E., Coosemans, M., 2008. Bionomics, taxonomy, and distribution of the major malaria vector taxa of Anopheles subgenus Cellia in Southeast Asia: an updated review. Infection, Genet. Evol., 8(4), 489–503.

Marcondes, C.B., Borges, P.S.S., 2000. Distinction of males of the Lutzomyia intermedia (Lutz and Neiva, 1912) species complex by ratios between dimensions and by an artificial neural network (Diptera: Psychodidae, Phlebotominae). Mem. Inst. Oswaldo Cruz 95, 685–688.

Margalef, R. 1958. Information Theory in Ecology. Gen. Syst. 3, 36–71.

Gannon, J.E., Stemberger, R.S., 1978. Zooplankton (Especially Crustaceans and Rotifers) as Indicators of Water Quality. Trans. Am. Microsc. Soc., 97.

Marková, S., Maurone, C., Racchetti, E., Bartoli, M.; Rossi, V. 2016. Daphnia diversity in water bodies of the Po River Basin. J. Limnol., 76, 261–271.

Marlene, P., Kalettka, T., Onandia, G., Balla, D., Lischeid, G., Pätzig, M. 2020. How much information do we gain from multiple-year sampling in natural pond research? Limnologica, 80, 125728.

Marsili-Libelli, S. 1991. Computer assisted vegetation analysis. In Handbook of Vegetation Science, 1st ed., Feoli, E., Orloci, L., (Eds.), Springer: Berlin/Heidelberg, Germany, Volume 11.

Martin, Y., Bonnefont, J.L., and Chancerelle, L., 2002. Gorgonians mass mortality during the 1999 late summer in French Mediterranean coastal waters: the bacterial hypothesis. Water Res., 36(3), 779-782.

Marquez, E., Jaramillo-O, N., Gomez-Palacio, A., Dujardin, J.P., 2011. Morphometric and molecular differentiation of a Rhodnius robustus-like form from R. robustus Larousse, 1927 and R. prolixus Stal, 1859 (Hemiptera, Reduviidae). Acta Trop. 120, 103–109.

Masmoudi, M.B., Chaoui, L., Topçu, N.E., Hammami, P., Kara, M.H., Aurelle, D., 2016. Contrasted levels of genetic diversity in a benthic Mediterranean octocoral: Consequences of different demographic histories? Ecol. Evol, 6, 8665– 8678.

Mauchline, J. Advances in Marine Biology; The Biology of Calanoid Copepods, 1st ed.; Academic Press: San Diego, CA, USA, 1998.

McInnes, L., Healy, J., and Melville, J., 2018. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv:1802.03426

McKenna, J.E., Kocovsky, P.M., 2020. Habitat characterization and species distribution model of the only large-lake population of the endangered Silver Chub (*Macrhybopsis storeriana*, Kirtland 1844). Ecol. Evol, 10, 12076– 12090.

Meerhoff, M., Clemente, J.M.; de Mello, F.T., Iglesias, C., Pedersen, A.R., Jeppesen, E., 2007. Can warm climate-related structure of littoral predator assemblies weaken the clear water state in shallow lakes? Glob. Chang. Biol., 13, 1888–1897.

Mellios, N., Moe, S.J., Laspidou, C., 2020. Machine Learning Approaches for Predicting Health Risk of Cyanobacterial Blooms in Northern European Lakes. Water, 12, 1191.

Melo-Merino, S.M., Reyes-Bonilla, H., and Lira-Noriega, A., 2020. Ecological niche models and species distribution models in marine environments: A literature review and spatial analysis of evidence. Ecol. Modell., 415, 108837.

Memon, L., Patel, S.B, Patel, D.P., 2019. Comparative Analysis of Artificial Neural Network and XGBoost Algorithm for PolSAR Image Classification, Pattern Recognition and Machine Intelligence.

Merchant, N.D., Fristrup, K.M., Johnson, M. P., Tyack, P.L., Witt, M.J., Blondel, P., Parks, S.E., 2015. Measuring acoustic habitats. Methods Ecol. Evol, 6(3), 257–265.

Merchant, N.D., Pirotta, E., Barton, T. R., Thompson, P. M., 2014. Monitoring ship noise to assess the impact of coastal developments on marine mammals. Marine Pollution Bulletin, 78(1–2), 85–95.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C., Lin., C., 2021. Package "e1071."

Milazzo, M., Chemello, R., Badalamenti, F., Camarda, R., Riggio, S., 2002. The Impact of Human Recreational Activities in Marine Protected Areas: What Lessons Should Be Learnt in the Mediterranean Sea? Mar. Ecol., 23, 280-290.

Minello, M., Calado, L., Xavier, F.C., 2021. Ecoacoustic indices in marine ecosystems: a review on recent developments, challenges, and future directions. ICES J. Mar. Sci., 78(9), 3066–3074.

Mironova, V.A., Shartova, N.V., Beljaev, A.E., Varentsov, M.I., Korennoy, F.I., Grishchenko, M.Y., 2020. Re-introduction of vivax malaria in a temperate area (Moscow region, Russia): a geographic investigation. Malar. J., 19, 1–20.

Mishachandar, B., Vairamuthu, S., 2021. Diverse ocean noise classification using deep learning. Appl. Acoust., *181*, 108141.

Mittelbach, G., McGill, B.J., 2019. Community Ecology, 2nd edn Oxford.

Mokhtar-Jamaï, K., Pascual, M., Ledoux, J.B., Coma, R., Féral, J.P., Garrabou, J., Aurelle, D., 2011. From global to local genetic structuring in the red gorgonian *Paramuricea clavata*: the interplay between oceanographic conditions and limited larval dispersal. Molecular Ecology, 20, 3291-3305.

Monika, Kumar, M., Kumar, M., 2021. XGBoost: 2D-Object Recognition Using Shape Descriptors and Extreme Gradient Boosting Classifier. In: Singh, V., Asari, V., Kumar, S., Patel, R. (eds) Computational Methods and Data Engineering. Advances in Intelligent Systems and Computing, vol 1227. Springer, Singapore.

Moraes, E.M., Manfrin, M.H., Laus, A.C., Rosada, R.S., Bomfin, S.C., Sene, F.M., 2004. Wing shape heritability and morphological divergence of the sibling species Drosophila mercatorum and Drosophila paranaensis. Heredity (Edinb), 92(5), 466–473.

Mortensen, P.B., and Buhl-Mortensen, L., 2004. Distribution of deep-water gorgonian corals in relation to benthic habitat features in the Northeast Channel (Atlantic Canada). Marine Biology 144, 1223–1238.

Mortensen, P.B., and Buhl-Mortensen, L., 2005. Deep-water corals and their habitats in The Gully, a submarine canyon off Atlantic Canada. In Freiwald, A., Roberts, J.M. (Eds.) Cold-Water Corals and Ecosystems. Erlangen Earth Conference Series. Springer, Berlin, Heidelberg.

Moulavi, D., Jaskowiak, P.A., Campello, R.J.G.B., Zimek, A., Sander, J., 2014. Density- based clustering validation. In: Proceedings of the 14th SIAM International Conference on Data Mining (SDM). Philadelphia, PA.

Mountcastle, A.M., Combes, S.A., 2014. Biomechanical strategies for mitigating collision damage in insect wings: structural design versus embedded elastic materials. J. Exp. Biol., 217(7), 1108–1115.

Movilla, J., Calvo, E., Coma, R. et al. Annual response of two Mediterranean azooxanthellate temperate corals to low-pH and high-temperature conditions. Mar. Biol., 163, 135 (2016).

Munday, P.L., Leis, J.M., Lough, J.M., Paris, C.B., Kingsford, M.J., Berumen, M.L., Lambrechts, J., 2009. Climate change and coral reef connectivity. Coral Reefs, 28, 379–395.

Murphy, K.P., 2012. Machine Learning: A Probabilistic Perspective. The MIT Press.

Murphy, S.J., Smith, A.B., 2021. What can community ecologists learn from species distribution models? Ecosphere, ESA.

Naimi, B., Hamm, Na, Groen T.A., Skidmore, A.K., Toxopeus, A.G., 2014. Where is positional uncertainty a problem for species distribution modelling. Ecography, 37, 191-203.

Nasreen, S., Azam, M.A., Shehzad, K., Naeem, U., Ghazanfar, M.A., 2014. Frequent Pattern Mining Algorithms for Finding Associated Frequent Patterns for Data Streams: A Survey. Procedia Comput. Sci., 37, 109–116.

Nguyen Hong Duc, P., Cazau, D., White, P. R., Gérard, O., Detcheverry, J., Urtizberea, F., Adam, O., 2021. Use of ecoacoustics to characterize the marine acoustic environment off the North Atlantic French Saint-Pierre-et-Miquelon Archipelago. J. Mar. Sci. Eng., 9(2), 177.

Nguyen, S.H., Webb, H.K., Mahon, P.J., Crawford, R.J., Ivanova, E.P., 2014. Natural insect and plant micro-nanostructsured surfaces: an excellent selection of valuable templates with superhydrophobic and self-cleaning properties. Molecules, 19(9), 13614–13630.

Nicolescu, G., Linton, Y.M., Vladimirescu, A., Howard, T.M., Harbach, R.E., 2004. Mosquitoes of the Anopheles maculipennis group (Diptera: culicidae) in Romania, with the discovery and formal recognition of a new species based on molecular and morphological evidence. Bull. Entomol. Res., 94(6), 525–535.

Noble, W.S., 2006. What is a support vector machine? Nature Biotechnology, 24.

Norouzzadeh, M.S., Nguyen, A., Kosmala, M., Swanson, A., Palmer, M. S., Packer, C., Clune, J., 2018. Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. Proc. Natl. Acad. Sci. U.S.A., 115(25), E5716–E5725.

Odeh, I.O.A., McBratney, A.B., Chittleborough, D.J., 1992. Soil Pattern Recognition with Fuzzy-c-means: Application to Classification and Soil-Landform Interrelationships. Soil Sci. Soc. Am. J., 56, 505–516.

Oksanen, J., Blanchet, F.G., Kindt, R., Legendre, P., Minchin, P.R., O'Hara, R., Simpson, G.L., Solymos, P., Stevens, M.H.H., Wagner, H., 2020. Vegan, Community Ecology Package. Ordination Methods, Diversity Analysis and Other Functions for Community and Vegetation Ecologists. R Package Version 2.5-7.

Olden, J.D., Lawler, J.J., Poff, N.L., 2008. Machine Learning Methods Without Tears: A Primer for Ecologists. Q. Rev. Biol., 83, 171–193.

Omran, M. G., Engelbrecht, A. P., Salman, A., 2007. An overview of clustering methods. Intell. Data Anal., 11(6), 583-605.

Opitz, D., Maclin, R., 1999. Popular Ensemble Methods: An Empirical Study. J. Artif. Intell. Res., 11, 169–198.

Oshiro, T.M., Perez, P.S., Baranauskas, J.A., 2012. How many trees in a random forest? In: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Springer, Berlin, Heidelberg, 154-168.

Ovaskainen, O., Abrego, N., 2020. Joint Species Distribution Modelling: With Applications in R (Ecology, Biodiversity and Conservation). Cambridge: Cambridge University Press.

Padovese, B., Frazao, F., Kirsebom, O.S., Matwin, S., 2021. Data augmentation for the classification of North Atlantic right whales upcalls. J. Acoust., 149(4), 2520–2530.

Pandeyz, S.K., Janghel, R.R. Gupta, V., 2021. Cardiac arrhythmia detection and classification from ecg signals using XGBoost classifier. In Machine Learning Algorithms and Applications, Eds Srinivas, M., Sucharitha, G., Matta, A., Chatterjee, P.

Park, J., Kim, D.I., Choi, B., Kang, W., Kwon, H.W., 2020. Classification and morphological analysis of vector mosquitoes using deep convolutional neural networks. Sci. Rep., 10, 1–12.

Pass, G., 2018. Beyond aerodynamics: the critical roles of the circulatory and tracheal systems in maintaining insect wing functionality. Arthropod. Struct. Dev., 47(4), 391–407.

Patterson, J.S., Klingenberg, C.P., 2007. Developmental buffering: how many genes? Evol. Dev., 9(6), 525–526.

Peterson, A.T., Vieglais, D.A., 2001. Predicting Species Invasions Using Ecological Niche Modeling: New Approaches from Bioinformatics Attack a Pressing Problem. BioScience, 51(5), 363-371.

Peterson, A.T., Soberón, J., 2012. Species Distribution Modeling and Ecological Niche Modeling: Getting the Concepts Right. Natureza & Conservacao, 10, 102-107.

Petrarca, V., Sabatinelli, G., Toure, Y.T., Di Deco, M.A., 1998. Morphometric multivariate analysis of field samples of adult *Anopheles arabiensis* and *An. gambiae* s.s. (Diptera: Culicidae). J. Med. Entomol., 35, 16–25.

Pey, A., Catanéo, J., Forcioli, D., Merle, P.L., Furla, P., 2013. Thermal threshold and sensitivity of the only symbiotic Mediterranean gorgonian Eunicella singularis by morphometric and genotypic analyses. Comptes Rendus Biologies, 336(7), 331-341.

Pfenninger, M., Schwenk, K., 2007. Cryptic animal species are homogeneously distributed among taxa and biogeographical regions. BMC Evol. Biol., 7(1), 1–6.

Piazzi, L., Balata, D., 2009. Invasion of alien macroalgae in different Mediterranean habitats. Biol. Invasions, 11, 193-204.

Pinto-Coelho, R.; Pinel-Alloul, B., Méthot, G., Havens, K.E., 2005. Crustacean zooplankton in lakes and reservoirs of temperate and tropical regions: Variation with trophic status. Can. J. Fish. Aquat. Sci., 62, 348–361.

Pivotto, I.D., Nerini, D., Masmoudi, M., Kara, H., Chaoui, L., Aurelle D., 2015. Highly contrasted responses of Mediterranean octocorals to climate change along a depth gradient. R. Soc. Open Sci., 2, 140493.

Plourde, S., Lehoux, C., Johnson, C.L., Perrin, G., Lesage, V., 2019. North Atlantic right whale (*Eubalaena glacialis*) and its food: (I) a spatial climatology of *Calanus* biomass and potential foraging habitats in Canadian waters. J. Plankton Res., 41(5), 667–685.

Pogodin, S., Hasan, J., Baulin, V.A., Webb, H.K., Truong, V.K., Phong Nguyen, T.H., Boshkovikj, V., Fluke, C.J., Watson, G.S., Watson, J.A., Crawford, R.J., Ivanova, E.P., 2013. Biophysical model of bacterial cell interactions with nanopatterned cicada wing surfaces. Biophys. J., 104(4), 835–840.

Ponti, M, Turicchia, E, Ferro, F, Cerrano, C, Abbiati, M., 2018. The understorey of gorgonian forests in mesophotic temperate reefs. Aquat. Conserv., 28, 1153– 1166.

Ponti, M., Perlini, R.A., Ventra, V., Grech, D., Abbiati, M., Cerrano, C., 2014. Ecological Shifts in Mediterranean Coralligenous Assemblages Related to Gorgonian Forest Loss. PLoS ONE, 9(7), e102782.

Popescu, A.A., Huber, K.T., Paradis, E., 2012. Ape 3.0: new tools for distance based phylogenetics and evolutionary analysis in R. Bioinformatics, 28, 1536–1537.

Purser, A., Orejas, C., Moje, A., Thomsen, L., 2014. The influence of flow velocity and suspended particulate concentration on net prey capture rates by the scleractinian coral Balanophyllia europaea (Scleractinia: Dendrophylliidae). J. MAR. BIOL. ASSOC. UK, 94(4), 687-696.

Puy, A., Lo Piano, S., Saltelli, A., Levin, S.A., 2021. sensobol: Computation of Variance-Based Sensitivity Indices. arxiv:2101.10103.

Qian, H., Ricklefs, R.E., White, P.S., 2004. Beta diversity of angiosperms in temperate floras of eastern Asia and eastern North America. Ecol. Lett.,8, 15-22.

Qiu, J., Wu, Q., Ding, G., Xu, Y., Feng, S., 2016. A survey of machine learning for big data processing. EURASIP J. Adv. Signal Process., 67.

Quenu, M. Trewick, S.A., Brescia, F., Morgan-Richards, M., 2020. Geometric morphometrics and machine learning challenge currently accepted species limits of the land snail Placostylus (Pulmonata: Bothriembryontidae) on the Isle of Pines, New Caledonia, Journal of Molluscan Studies, 86(1), 35–41.

R Core Team, 2021. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rachor, E., Reiss, H., Degraer, S., Duineveld, G.C.A., Van Hoey, G., Lavaleye, M., Willems, W., Rees, H.L., 2007. Structure, distribution, and characterizing species of North Sea macrozoobenthos communities in 2000. In Structure and dynamics of the North Seabenthos, Rees, H.L., Eggleton, J.D., Rachor, E., Vanden Berghe, E., (Eds.); ICES Cooperative Research: Copenhagen, Denmark, Volume 288, pp. 46–59.

Ragazzola, F., Foster, L.C., Form A., Anderson, P.S.L., Hansteen, T.H., Fietzke, J., 2012. Ocean acidification weakens the structural integrity of coralline algae. Glob. Change Biol., 18, 2804-2812

Rajabi, H., Shafiei, A., Darvizeh, A., Dirks, J.H., Appel, E., Gorb, S.N., 2016a. Effect of microstructure on the mechanical and damping behaviour of dragonfly wing veins. R. Soc. Open Sci. 3, 160006.

Rajabi, H., Darvizeh, A., Shafiei, A., Taylor, D., Dirks, J.H., 2015. Numerical investigation of insect wing fracture behaviour. J. Biomech., 48(1), 89–94.

Rammer, W., Seidl, R. 2019. Harnessing Deep Learning in Ecology: An Example Predicting Bark Beetle Outbreaks. Front. Plant Sci.,10, 1327.

Ray, R, Nakata, T, Henningsson, P, Bomphrey, R J, 2016. Enhanced flight performance by genetic manipulation of wing shape in Drosophila. Nat. Commun. 7 (10851).

Reaka-Kudla, M. L., 1997. The Global Biodiversity of Coral Reefs: A Comparison with Rain Forests. In M. L. Reaka-Kudla, D. E. Wilson and E. O. Wilson (Eds.), Biodiversity II: Understanding and Protecting Our Biological Resources, (pp. 83–108), Joseph Henry Press.

Recknagel, F., 2001. Applications of machine learning to ecological modelling. Ecol. Model. 146, 303–310.

Recknagel, F., Michene, W.K. 2018. Ecological Informatics Data Management and Knowledge Discovery. Springer International Publishing AG, Cham, Switzerland.

Renner, I.W., Elith, J., Baddeley, A., Fithian, W., Hastie, T., Phillips, S.J., Popovic, G., Warton, D.I., 2015. Point process models for presence-only analysis. Methods Ecol. Evol., 6(4), 366–379.

Renner, I.W., Elith, J., Baddeley, A., Fithian, W., Hastie, T., Phillips, S.J., Popovic, G., Warton, D.I., 2015. Point process models for presence-only analysis. Methods Ecol. Evol., 6(4), 366–379.

Reyes Bonilla, H., Cruz Piñón, G., 2002. Influence of temperature and nutrients on species richness of deep-water corals from the western coast of the Americas. Hydrobiologia, 471, 35–41.

Rhys, H.I., 2020. Machine Learning with R, the tidyverse, and mlr. Manning.

Ri, J.H., Kim, H., 2020. G-mean based extreme learning machine for imbalance learning. Digit. Signal Process. A Rev. J. 98, 102637.

Riessen, H.P., 2012. Costs of predator-induced morphological defences in Daphnia. Freshw. Biol., 57, 1422-1433.

Rivetti, I., Fraschetti, S., Lionello, P., Zambianchi, E., and Boero, F., 2014. Global Warming and Mass Mortalities of Benthic Invertebrates in the Mediterranean Sea. PLOS ONE, 9(12), e115655.

Rocchini, D., Marcantonio, M., Arhonditsis, G., Lo Cacciato, A., Hauffe, H.C., He, K.S., 2019. Cartogramming uncertainty in species distribution models: A Bayesian approach. Ecol. Complex., 38, 146-155.

Rodier, J., Legube, B., Merlet, N., 1999. L'Analyse de l'Eau, Dunod: Paris, France, 1987.56.

D'Auria, G., Zavagno, F. Indagine sui Bodri della Provincia di Cremona. Monogr. Pianura, 3, 5–229.

Rodolfo-Metalpa, R., Abbate, M., Peirano, A., 2001. The influence of light, temperature and feeding on the growth of the Corals *Cladocora caespitosa* and *Balanophyllia europaea.* L'influenza di luce, temperatura e alimentazione sulla crescita dei coralli mediterranei *Cladocora caespitosa* e *Balanophyllia europaea*. Risultati preliminari di una sperimentazione in ambiente controllato. Technical Report.

Rodrigues, Y.K., Beldade, P., 2020. Thermal plasticity in insects' response to climate change and to multifactorial environments. Front. Ecol. Evol., 8, 271.

Root-Gutteridge, H., Cusano, D.A., Shiu, Y., Nowacek, D.P., Van Parijs, S.M., Parks, S.E., 2018. A lifetime of changing calls: North Atlantic right whales, *Eubalaena glacialis*, refine call production as they age. Anim. Behav, 137, 21–34.

Rosenberg, E., and Ben-Haim, Y., 2002. Microbial diseases of corals and global warming. Environ. Microbiol., 4(6), 318-326

Rossi, V., Maurone, C., Benassi, G., Marková, S., Kotlík, P., Bellin, N., Ferrari, I., 2015. Phenology of *Daphnia* in a Northern Italy pond during the weather anomalous 2014. J. Limnol., 74, 74.

Rotiroti, M., Bonomi, T., Sacchi, E., McArthur, J.M., Stefania, G.A., Zanotti, C., Taviani, S., Patelli, M., Nava, V., Soler, V., Fumagalli, L., Leoni, B., 2019. The effects of irrigation on groundwater quality and quantity in a human-modified hydro-system: The Oglio River basin, Po Plain, Northern Italy. Sci. Total. Environ., 672, 342–356.

Roubens, M. 1982. Fuzzy clustering algorithms and their cluster validity. Eur. J. Oper. Res., 10, 294–301.

Rudman, S.M., Greenblum, S.I., Rajpurohit, S., Betancourt, N.J., Hanna, J., Tilk, S., Yokoyama, T., Petrov, D.A., Schmidt, P., 2022. Direct observation of adaptive tracking on ecological time scales in *Drosophila*. Science. 375(6586), 1226-1227.

Russo, A.R., 1985. Ecological observation ns on the gorgonian sea fan *Eunicella cavolinii* in the Bay of Naples. MEPS, 155-159.

Rycyk, A.M., Tyson Moore, R.B., Wells, R.S., McHugh, K.A., Berens McCabe, E.J., Mann, D.A., 2020. Passive Acoustic Listening Stations (PALS) show rapid onset of ecological effects of harmful algal blooms in real time. Sci. Rep., 10(1), 17863.

Saito, T., Rehmsmeier, M., 2015. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One, 10, e0118432.

Saito, T., Rehmsmeier, M., 2017. Precrec: fast and accurate precision-recall and ROC curve calculations in R. Bioinformatics, 33, 145–147.

Salski, A., 2007. Fuzzy clustering of fuzzy ecological data. Ecol. Inform., 2, 262–269.

Sbrocco, E. J., Barber, P.H., 2013. MARSPEC: ocean climate layers for marine spatial ecology. Ecology 94, 979.

Schindler, D.W., 2009. Lakes as sentinels and integrators for the effects of climate change on watersheds, airsheds, and landscapes. Limnol. Oceanogr., 54, 2349–2358.

Schindler, D.W. 2001. The cumulative effects of climate warming and other human stresses on Canadian freshwaters in the new millennium. Can. J. Fish. Aquat. Sci., 58, 18–29.

Schmidt, R., Morrison, A., Kunc, H.P., 2014. Sexy voices - no choices: Male song in noise fails to attract females. Anim. Behav., *94*, 55–59.

Schroeder, K., Josey, S.A., Herrmann, M., Grignon, L., Gasparini, G.P., and Bryden, H.L., 2010. Abrupt warming and salting of the Western Mediterranean Deep Water after 2005: Atmospheric forcings and lateral advection. J. Geophys., 115, C08029.

Sebens, K.P., 1991. Habitat structure and community dynamics in marine benthic systems. In Bell, S.S., McCoy, E.D., Mushinsky, H.R. (Eds.) Habitat Structure. Population and Community Biology Series, Vol. 8, Springer, Dordrecht.

Sejnowski, T. J., 2018. The Deep Learning Revolution. The MIT Press, Cambridge, Massachusetts London, England.

Senent-Aparicio, J., Soto, J., Pérez-Sánchez, J., Garrido, J., 2017. A novel fuzzy clustering approach to regionalize watersheds with an automatic determination of optimal number of clusters. J. Hydrol. Hydromech., 65, 359–365.

Sethi, S.S., Jones, N.S., Fulcher, B.D., Picinali, L., Clink, D.J., Klinck, H., Orme, C.D.L., Wrege, P.H., Ewers, R.M., 2020. Characterizing soundscapes across diverse ecosystems using a universal acoustic feature set. Proceedings of the National Academy of Sciences, 117(29), 17049–17055.

Severini, F., Toma, L., Di Luca, M., Romi, R., 2009. Le Zanzare Italiane: Generalità e Identificazione Degli Adulti (Diptera, Culicidae). Fragm. Entomol., 41, 213.

Shai, S.S., Shai, B.D., 2014 Understanding Machine Learning: From Theory to Algorithms. first editions, Cambridge University Press.

Shannon, C.E., Wiener, W., 1963. The Mathematical Theory of Communication. University of Illinois Press.

Silberschatz, A.; Tuzhilin, A., 1995 On Subjective Measures of Interestingness Discovery in Knowledge Bell Laboratories Measures. In Proceedings of the 1st International Conference on Knowledge Discovery and Data Mining, Montreal, QC, Canada, pp. 275–281.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556.

Sini, M., Garrabou, J., Trygonis, V., and Koutsoubas, D., 2019. Coralligenous formations dominated by *Eunicella cavolini* (Koch, 1887) in the NE Mediterranean: biodiversity and structure. Mediterr. Mar. Sci., 20, 174– 188.

Sini, M., Kipson, S., Linares, C., Koutsoubas, D., Garrabou, J., 2015. The yellow gorgonian *Eunicella cavolini*: demography and disturbance levels across the Mediterranean Sea. PLoS ONE, 10, e0126253.

Soberón, J., Osorio-Olvera, L., Peterson, T. 2017. Diferencias conceptuales entre modelación de nichos y modelación de áreas de distribución. Rev. Mex. Biodivers., 88, 437-441

Sofaer, H.R., Hoeting, J.A., Jarnevich, C.S., 2019. The area under the precision-recall curve as a performance metric for rare binary events. Methods Ecol. Evol. 10, 565–577.

Sommer, R.J., 2020. Phenotypic plasticity: from theory and genetics to current and future challenges. Genetics, 215(1), 1–13.

Søndergaard, M., Johansson, L.S., Lauridsen, T.L., Jørgensen, T.B., Liboriussen, L., Jeppesen, E., 2010. Submerged macrophytes as indicators of the ecological quality of lakes. Freshw. Biol., 55, 893–908.

Sospedra, J., Niencheski, L.F.H., Falco, S., Andrade, C.F.F., Attisano, K.K., Rodilla, M., 2018. Identifying the main sources of silicate in coastal waters of the Southern Gulf of Valencia (Western Mediterranean Sea), Oceanologia, 60(1), 52-64.

Sperlea, T., Kreuder, N., Beisser, D., Hattab, G., Boenigk, J., Heider, D., 2021. Quantification of the covariation of lake microbiomes andenvironmental variables using a machine learning-based framework. Mol. Ecol., 14.

Steen, V.A., Tingley, M.W., Paton, P.W.C., Elphick, C.S., 2020. Spatial thinning and class balancing: Key choices lead to variation in the performance of species distribution models with citizen science data. Methods Ecol. Evol., 12, 216– 226.

Steiner, C.F., 2004. *Daphnia* dominance and zooplankton community structure in fishless ponds. J. Plankton Res., 26, 799–810.

Sterner, R.W.; Elser, J.J. 2002. Ecological Stoichiometry: The Biology of Elements from Molecules to the Biosphere. 46–59, Princeton University Press: Princeton, NJ, USA.

Stier, A.C., Geange, S.W.; Hanson, K., Bolker, B.M., 2013. Predator density and timing of arrival affect reef fish community assembly. Ecology, 94, 1057–1068.

Stoliar, B.O., Lushchak, V.I., 2019. Environmental pollution and oxidative stress in fish. Chapter 7, from Oxidative Stress: Eustress and Distress, Helmut Sies Academic Press

Sully, S., Burkepile, D. E., Donovan, M. K., Hodgson, G., Van Woesik, R., 2019. A global analysis of coral bleaching over the past two decades. Nat. Commun., *10*(1), 1264.

Sun, J.Y., Yan, Y.W., Li, FD., Zhang, Z.J., 2021. Generative design of bioinspired wings based on deployable hindwings of *Anomala corpulenta* Motschulsky. Micron 151, 103150.

Suppa, A., Kvist, J., Li, X., Dhandapani, V., Almulla, H., Tian, A.Y., Kissane, S., Zhou, J., Perotti, A., Mangelson, H., et al., 2020. Roundupcauses embryonic development failure and alters metabolic pathways and gut microbiota functionality in non-target species. Microbiome, 8, 1–15.

Suthaharan, S., 2016. Supervised Learning Algorithms. In: Machine Learning Models and Algorithms for Big Data Classification. Integrated Series in Information Systems, vol 36. Springer, Boston, MA.

Tabak, M.A., Norouzzadeh, M.S., Wolfson, D.W., Sweeney, S.J., Vercauteren, K. C., Snow, N.P., Halseth, J.M., Di Salvo, P.A., Lewis, J.S., White, M.D., Teton, B., 2019. Machine learning to classify animal species in camera trap images: Applications in ecology. Methods Ecol. Evol., 10(4), 585–590.

Tatsuta, H., Takahashi, K.H., Sakamaki, Y., 2018. Geometric morphometrics in entomology: basics and applications. Entomol. Sci. 21 (2), 164–184.

Terrón Singler, A., López González, P.J., 2005. Cnidae variability in *Balanophyllia europaea* and *B. regia* (Scleractinia: Dendrophylliidae) in the NE Atlantic and Mediterranean Sea. Scientia Marina, 69 (1), 75-86

Thomsen, M.S., Wernberg, T., Altieri, A., Tuya, F., Gulbransen, D., McGlathery, K.J., Holmer, M., Silliman, B.R., 2010. Habitat cascades: the conceptual context and global relevance of facilitation cascades via habitat formation and modification. Integr. Comp. Biol., 50(2), 158–175.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Series B Stat. Methodol., 58(1), 267–288.

Tignat-Perrier, R., Van de Water, J.A.J.M., Guillemain, D., Aurelle, D., Allemand, D., Ferrier-Pagès, C., 2022. The Effect of Thermal Stress on the Physiology and Bacterial Communities of Two Key Mediterranean Gorgonians. Appl. Environ. Microbiol., 88(6).

Tilson, L., Excell, P., Green, R., 2005. A Generalisation of the Fuzzy C-means clustering algorithm. Remote Sens., 3, 1783–1784.

Turing, A.M.I., 1950. Computing Machinery and Intelligence, Mind, Volume LIX(236), 433–460.

Tuel, A., Eltahir, E.A.B., 2020. Why Is the Mediterranean a Climate Change Hot Spot? J. Clim., 33(14).

Ulrich, W., Gotelli, N. J., 2007. Null Model Analysis of Species Nestedness Patterns.Ecology, 88, 1824–1831.

United States Environmental Protection Agency (EPA), 2022.

Usman, A.M., Ogundile, O.O., Versfeld, D.J.J., 2020. Review of automatic detection and classification techniques for cetacean vocalization. *IEEE Access*, *8*, 105181–105206.

Vadadi-Fülöp, C., Sipkay, C., Mészáros, G., Hufnagel, L., 2012. Climate change and freshwater zooplankton: What does it boil down to? Aquat. Ecol., 46, 501–519.

Valavi, R., Guillera-Arroita, G., Lahoz-Monfort, J.J., Elith, J., 2022. Predictive performance of presence-only species distribution models: a benchmark study with reproducible code. Ecol. Monogr., 92(1), e01486.

Valderrama, J.C., 1981. Methods Used by the Hydrographical Department of the National Board of Fisheries. In Report of the Baltic Intercalibration Workshop. Annex, Grasshof, K., Ed., Interim Commission for the Protection of the Environment of the Baltic Sea: Goteborg, Sweden, 1977; pp. 14–43.54. Water Environmental Federation; American Public Health Association. Standard Methods for the Examination of Water and Wastewater; APHA: Washington, DC, USA.

Van Echelpoel, W., Boets, P., Landuyt, D., Gobeyn, S., Everaert, G., Bennetsen, E., Mouton, A., Goethals, P. L.M., 2015. Species distribution models for sustainable ecosystem management. In Y.-S.Park, S. Lek, C. Baehr, S.E., Jørgensen (Eds), Developments in Environmental Modelling, (Vol. 27, pp. 115-134), Elsevier.

Van Voorhies, W.A., 1996. Bergmann size clines: a simple explanation for their occurrence in ectotherms. Evolution, 50, 1259–1264.

Vandaele, R., Aceto, J., Muller, M., Peronnet, F., Debat, V., Wang, C.W., Huang, C.T., Jodogne, S., Martinive, P., Geurts, P., Maree, R., 2018. Landmark detection in 2D bioimages for geometric morphometrics: a multi-resolution tree-based approach. Sci. Rep. 8, 538.

VanDerWal, J., Shoo, L.P., Graham, C., Williams, S.E., 2009. Selecting pseudo-absence data for presence-only distribution modeling: How far should you stray from what you know? Ecol. Model., 220(4), 589-594.

Velliangiri, S., Alagumuthukrishnan, S., Thankumar S.I.J., 2019. A Review of Dimensionality Reduction Techniques for Efficient Computation. Procedia Comput., 165, 104-111.

Venables, W.N., Ripley, B. D., 2002. Modern Applied Statistics with S (fourth edition). Springer.

Verdonschot, R.C.M., Keizer-Vlek, H.E., Verdonschot, P.F.M., 2011. Biodiversity value of agricultural drainage ditches: A comparativeanalysis of the aquatic invertebrate fauna of ditches and small lakes. Aquat. Conserv. Mar. Freshw. Ecosyst., 21, 715–727.

Verdura, J., Linares, C., Ballesteros, E., Coma, R., Uriz, M.J., Bensoussan, M., Cebrian, E., 2019. Biodiversity loss in a Mediterranean ecosystem due to an extreme warming event unveils the role of an engineering gorgonian species. Sci. Rep., 9, 5911.

Vester, H., Hallerberg, S., Timme, M., Hammerschmidt, K., 2017. Vocal repertoire of long-finned pilot whales (*Globicephala melas*) in northern Norway. J. Acoust, *141*(6), 4289–4299.

Vezzulli, L., Pezzati, E., Huete-Stauffer, C., Pruzzo, C., Cerrano, C., 2013. 16S rDNA Pyrosequencing of the Mediterranean Gorgonian *Paramuricea clavata* Reveals a Link among Alterations in Bacterial Holobiont Members, Anthropogenic Influence and Disease Outbreaks. PLOS ONE, 8(6), e67745.

Viana, D.S., Keil, P., Jeliazkov, A., 2022. Disentangling Spatial and Environmental Effects: Flexible Methods for Community Ecology and Macroecology." Ecosphere, 13(4), e4028.

Vidal, P.O., Suesdek, L., 2012. Comparison of wing geometry data and genetic data for assessing the population structure of *Aedes aegypti*. Infect. Genet. Evol., 12, 591–596.

Viscosi, V., Cardini, A., 2011. Leaf morphology, taxonomy and geometric morphometrics: a simplified protocol for beginners. PLoS One 6, 25630.

Vrijenhoek, R.C., Johnson, S.B., Rouse, G.W., 2009. A remarkable diversity of bone- eating worms (Osedax; Siboglinidae; Annelida). BMC Biol., 7(1), 1–13.

Wang, X., Cheng, J., Wang, L.A., 2022. Reinforcement learning-based predator-prey model. Ecol. Complexity, 42, 100815.

Warren, D., Dinnage, R., 2022. ENMTools: Analysis of Niche Evolution using Niche and Distribution Models. R package version 1.0.6.

Wei, W., Chen, R., Wang, L., Fu, L., 2017. Spatial distribution of crustacean zooplankton in a large river-connected lake related to trophic status and fish. J. Limnol., 76, 546–554.

Weidman, P.R., Schindler, D.W., Thompson, P., Vinebrooke, R.D, 2014. Interactive effects of higher temperature and dissolved organic carbon on planktonic communities in fishless mountain lakes. Freshw. Biol., 59, 889–904.

Wellborn, G.A., Skelly, D.K., Werner, E.E., 1996. Mechanisms Creating Community Structure across a Freshwater Habitat Gradient. Annu. Rev. Ecol. Syst., 27, 337–363.

West-Eberhard, M.J., 2005. Developmental plasticity and the origin of species differences. Proc. Natl. Acad. Sci. U.S.A., 102(1), 6543–6549.

Wilke, A.B.B., De Oliveira, C.R., Multini, L.C., Vidal, P.O., Wilk-Da-Silva, R., de Carvalho, G.C., Marrelli, M.T., 2016. Morphometric wing characters as a tool for mosquito identification. PLoS One, 11(8), 1–12.

Williams, J.N., Seo, C., Thorne, J., Nelson, J. K., Erwin, S., O'Brien, J.M., Schwartz, M.W., 2009. Using species distribution models to predict new occurrences for rare plants. Divers. Distrib., 15, 565-576.

Wright, D.H., Reeves, J.H., 1992. On the meaning and measurement of nestedness of species assemblages. Oecologia, 92, 416–428.

Wright, M. N., Ziegler, A., 2017. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. J. Stat. Softw., 77(1), 1-17.

Yang, Y., Sun, H., Zhang, Y., Zhang, T., Gong, J., Wei, Y., Duan, Y.G., Shu, M., Yang, Y., Wu, D., Yu, D., 2021. Dimensionality reduction by UMAP reinforces sample heterogeneity analysis in bulk transcriptomic data. Cell Rep., 36 (4), 109442.

Yesson, C., Taylor, M.L., Tittensor, D.P., Davies, A.J., Guinotte, J., Baco, A., Black, J., Hall-Spencer, J.M. Rogers, A. D., 2012. Global habitat suitability of cold-water octocorals. J. Biogeogr., 39, 1278-1292.

Yuval, B., Wekesa, J.W., Lemanager, D., Kauffman, E.E., Washino, R.K., 1993. Seasonal variation in body size of mosquitoes (Diptera: Culicidae) in a rice culture agroecosystem. Environ. Entomol. 22, 459–463.

Zadeh, L.A., 1965. Fuzzy Sets. Information and control, 8, 338-353.

Zelditch, M.L., Mezey, J., Sheets, H.D., Lundrigan, B.L., Garland, T., 2006. Developmental regulation of skull morphology II: ontogenetic dynamics of covariance. Evol. Dev., 8 (1), 46–60.

Zhang, S., Wu, X. 2011. Fundamentals of association rules in data mining and knowledge discovery. Wiley Interdiscip. Rev. Data Min. Knowl. Discov., 1, 97-116.

Zhang, Z., Xu, S., Capinha, C., Weterings, R., Gao, T., 2019. Using species distribution model to predict the impact of climate change on the potential distribution of Japanese whiting *Sillago japonica*. Ecol. Indic., 104, 333–340.

Zheng, J.H., Nie, H.T., Yang, F., Xi-Wu, J., 2019. Genetic variation and population structure of different geographical populations of *Meretrix petechialis* based on mitochondrial gene COI. J. Genet. 98, 68.

Zhou, L., Pan, S., Wang, J., Vasilakos, A.V., 2017. Machine learning on big data: Opportunities and challenges. Neurocomputing, 237, 350-361.

Zhou, Z.H., 2021. Machine Learning. Springer Nature.

Zibrowius, H., 1980. Les Scléractiniaires de la Méditerranée et de l'Atlantique nord-oriental. Mémoires de l'Institut océanographique, Monaco, 87-89.

Zibrowius, H., 1983. Nouvelles données sur la distribution de quelques scléractiniaires “méditerranéens” à l’est et à l’ouest du détroit de Gibraltar. Rapp. Comm. Int. Mer. Medit., 28, 307-309.

Zimmermann, H.J., 1999. Practical Applications of Fuzzy Technologies. Springer: Boston, MA, USA.