Ocean
and Coastal
Research

METHODS
Special Article Collection
Sousa, S.H.M.; Carreira, R.C.; Fonseca, G. (eds),
Subsidies for ecosystem-based management in an
offshore oil and gas exploration area, Santos Basin, Brazil
http://doi.org/10.1590/2675-2824071.22044gf
ISSN 2675-2824

Overcoming the challenges of data integration in ecosystem studies with machine learning workflows: an example from the Santos project

Gustavo Fonseca[1,*], Danilo Candido Vieira[1]

[1] Instituto do Mar. Universidade Federal de São Paulo – Campus Baixada Santista. R. Silva Jardim, 136 – Vila Matias, Santos, SP – 11015-020, Brazil.

* Corresponding author: gfonseca.unifesp@gmail.com

ABSTRACT

Integrating intricate environmental data within a unified analytical framework for extensive conservation and monitoring initiatives encounters several challenges. These challenges encompass defining a conceptual model outlining cause-and-effect relationships, addressing dissimilarities in data source quantity and information content, grappling with missing or noisy data, fine-tuning model optimization, achieving accurate predictions, and tackling the issue of imbalanced observations across factors. In the context of the Santos project, dedicated to comprehending the spatio-temporal dynamics of benthic, pelagic, and physical systems for the facilitation of conservation and monitoring programs, the application of machine learning's random forest (RF) technique for modeling univariate data offers notable advantages. This approach adeptly handles non-linearity, covariation, and interactive effects among predictors. For modeling multivariate data sets, a hybrid strategy combining a self-organizing map (SOM) and RF is harnessed to effectively tackle the challenges. Addressing missing values, the bagging imputation technique demonstrated superior performance compared to other methods. Both machine learning techniques discussed herein exhibit resilience against the impact of noisy data, yet the identification of noisy data remains feasible based on model outputs. In scenarios of imbalanced data sets, we investigate the correlation between the RF model's overall statistics and those of individual classes. The joint interpretation of these statistics aids in comprehending model limitations and facilitates discussions on the environmental mechanisms shaping observed patterns. We propose two analytical workflows that not only enable the exploration and enhancement of model accuracy but also facilitate the investigation of potential cause-and-effect relationships inherent in the data. Furthermore, these workflows lay the foundation for implementing long-term learning algorithms, a pivotal increment for monitoring initiatives. Notably, these workflows, alongside the discussed analytical challenges, can be seamlessly implemented within iMESc, an open-source application.

Keywords: Self-organizing map, Random forest, Oceanography, Modeling, Santos basin

INTRODUCTION

The conservation of natural ecosystems is one of the biggest challenges faced by humanity today. Pollution, climate change, habitat loss and resource exploitation are pushing natural ecosystems to the critical point of no return (Walker, 2006; Newman, 2019). Critical issues for the conservation of natural systems are the understanding of the spatio-temporal variation and identification of the key structuring processes. To have a first glimpse at the data variability, such baseline knowledge generally starts by summing published material and sampling multiple stations at different spatial scales and time periods (Rhodes and Jonzén, 2011; Borja et al., 2016). In addition, to guarantee the characterization of the main structuring processes,

Submitted: 05-May-2022
Approved: 10-Apr-2023

Editor: Rubens M. Lopes
Associate Editor: Silvia Helena Sousa

multiple parameters from multiple areas of science are collected in a coordinated manner (Grehan et al., 2017; Dailianis et al., 2018). Nevertheless, the integration of environmental data into a single framework to be used for conservation and monitoring purposes usually comes with several challenges (Rollinson et al., 2021) that have to be considered: (1) definition of the conceptual model (cause-effect relationships), (2) differences in the quantity and/or information content of data sources, (3) detection of noisy data, (4) missing data, (5) model optimization, and (6) imbalanced number of observations across factors. Although these challenges can occur in analyses that integrate data at any scale, they tend to be exacerbated in large-scale environmental studies (Levy et al., 2014).

Depending on the complexity of the system and of the database, going through all these issues to transform raw data into scientific knowledge and further into management decisions involves multiple analytical steps. This sequence of steps, also termed analytical pipeline or analytical workflow, is a method, much like any laboratory protocol, that must be precisely documented to guarantee reproducibility (Perkel, 2019). Although both terms are usually used as synonyms, they can be distinguished by its nature: a pipeline is machine oriented based on a fixed routine, eventually allowing for an initial parametrization without human intervention during the process; a workflow is human oriented, and the sequence of analytical steps are set allowing some decisions and reanalysis along the processes (Stoudt et al., 2021). In both cases, the analytical process should be kept as simple as possible, ensuring some important steps such as data download, verification and quality control, pre-processing, analysis, storage, and visualization (Figure 1). But depending on the goals, a pipeline or a workflow can reach over one hundred steps.

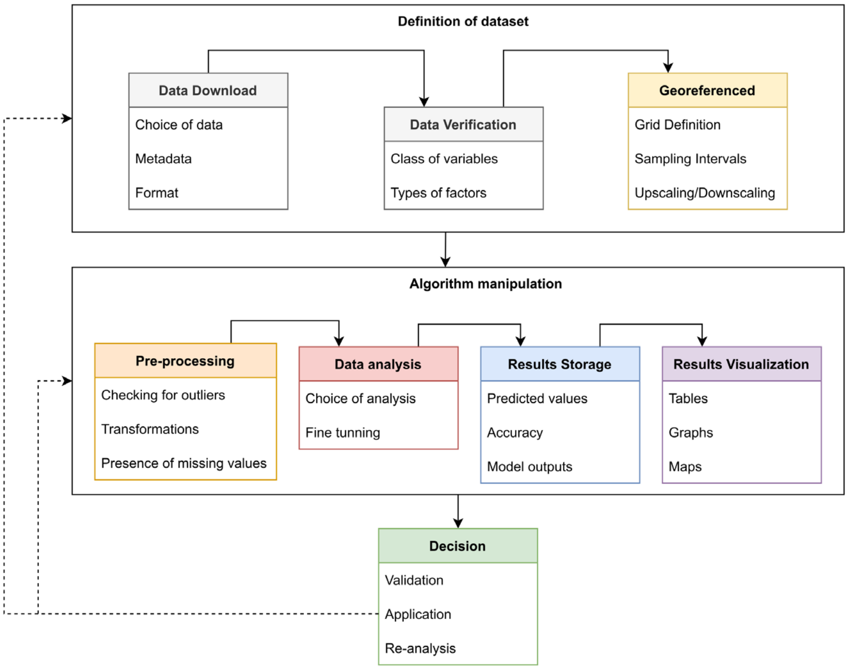


Figure 1. Typical steps (boxes) involved in a reproducible analytical workflow. Full arrows represent the main steps, while dashed arrows indicate alternative ways for re-analysis.

Particularly for the understanding and monitoring of complex and interacting environmental processes, an important step to achieve an effective pipeline is the definition of the research question and of a statistical modelling workflow (Schaub and Abadi, 2011; Michener and Jones, 2012). If the analytical problem is very complex and deals with multiple interacting parts of an ecosystem (e.g. Butenschön et al., 2016), it can be broken into smaller and simpler models in such a way that the results of one subset can be used as the predictor of another subset. Although the use of ML techniques is well stablished in some scientific fields (Sarker, 2021; Zhong et al., 2021), in oceanography and environmental monitoring it is still incipient, especially when considering biological data (Ditria et al., 2022; Jiang and Zhu, 2022). Particularly for baseline and monitoring research programs, which aim to anticipate potential hazardous effects on unaffected areas, the construction of analytical workflows based on machine learning (ML) techniques have the advantage of continuous learning and evaluation of model performance, (Hino et al., 2018; Stupariu et al., 2021). In addition, ML algorithms have the principle of predictive analytics and of life-long learning, meaning that each newly collected data in a monitoring program based on a ML workflow is evaluated whether it is within the predicted range or out of it. In the second case the model should be recalculated to generate better predictions (L'Heureux et al., 2017). Eventually, the prediction of this new data will demand the incorporation of new explanatory variables not previously considered in the model. This feedback between model performance and monitoring program is the principle to conduct adaptative monitoring programs (Nichols and Williams, 2006; Lynam et al., 2016). Moreover, in comparison with traditional statistics, when a large amount of information is available, ML algorithms can better handle some of the challenges mentioned above (L'Heureux et al., 2017; Kaur et al., 2020). Nonetheless, classical statistics and mechanists' approaches are not direct competitors, but complementary approaches. On the one hand, statistical and mechanistic models aim to test if a causality is significant, but their oversimplified assumptions and extremely specific nature prohibit the universal predictions achievable by machine learning algorithms (Baker et al., 2018). On the other hand, depending on the method, machine learning techniques miss the specificity of cause-effect relationships. The objective of this study is to give a brief introduction on machine learning techniques, present solutions to overcome the challenges mentioned above and provide examples of ML workflows that could be used on large-scale baseline ecosystem. For the challenges, we provide examples from the benthic system of the Santos Project – Santos Basin Environmental Characterization– coordinated by the oil company Petróleo Brasileiro S.A. (PETROBRAS) that were explored by other authors in this issue. The Santos Basin is highly valued commercially by its oil and gas production potential (Moreira et al., 2023). The aims of the workflows presented here are to integrate different types of data, facilitate data processing, data analysis, provide a graphical interface for results visualization and make accurate predictions. The study is organized into four further sections: 1) machine learning algorithms for ecosystem baseline studies; 2) overcoming the analytical challenges of the Santos project, 3) machine learning workflows. The challenges explored hereafter are those related to the Santos Project, other challenges may be found in other types of datasets (Gligorijević and Pržulj, 2015; Zipkin et al., 2021).

Machine learning Algorithms for ecosystem baseline studies

Machine learning is about designing algorithms that allow the computer to find statistical regularities or other patterns in the data (Ayodele, 2010). According to its goals, the algorithms are usually separated into supervised, semi-supervised, unsupervised, reinforcement, transduction and learning-to-learn. Reviewing all of them is beyond the scope of this paper and can be found elsewhere (Ayodele, 2010; Bonaccorso, 2017; Mahesh, 2020). An important aspect to point out is that ML uses the benefits of computation capacity, which includes processing large amounts of information, resampling techniques and performing analytical loops (i.e., feed model with new data and monitoring the results to make sure that the models continue to improve in value) (Pope and McNeill, 2013; Kaur et al., 2020). Unlike classical statistics, where the learning process comes from hypothesis testing against a given probability distribution (frequentist) or a prior knowledge of a distribution (Bayesian and Maximum likelihood), ML techniques make the assumption that when dealing with large quantities of data, a part of the data can be used to learn, while a second part can be used to test (Mahesh, 2020). Thus, ML techniques create their own hypothesis with the training-part of the data (internal validation) and evaluate it with the test-part (external validation).

Internal validation is an examination of model performance in the same dataset that was used to develop it. The cross-validation procedure is the most commonly used form of internal validation. It uses different portions of the data to test and train a model on different iterations, with model selection performed independently in each fold to avoid selection bias (Refaeilzadeh et al., 2009). The number of folds to be used depends on the amount and heterogeneity of the data. In theory, a higher number of folds means that more data is used to train the model at each run, leading to a lower prediction error. In contrast, a lower number of folds means that the training set is small, and the test set is large, enhancing the average chance of error. Nevertheless, using a large number of folds increases processing time. Like for the number of trees, the recommendation here is to run multiple folds increasing slowly to evaluate whether the gain in accuracy compensates the processing time (e.g., 3, 5, 10, 15...). External validation (EV) is the process of examining a prediction model's performance in data independent to that used for model development. Since training data may not be truly representative, any trained model should be regarded as potentially non-generalizable (Ho et al., 2020). Therefore, before generalizing predictions, it is recommended to evaluate a learned model via EV. A well-trained model that captures informative features is robust and will continue to perform well even if repeatedly imputed with new data (Ho et al., 2020).

Here we explore two algorithms that can be used to handle the issues encountered in the Santos Project. Above all, the research goals of the project are to generate an accurate prediction that supports a monitoring program and at the same time understand the oceanographic processes that are shaping the benthos. Tree-based algorithms are probably the most appropriate ones to disentangle potential covariate and interaction effects among the predictors, a characteristic of complex systems. The model structure between the response variable and the predictors are organized in branches, without requiring the researcher to have any preconception on this matter (Rahmati et al. 2019). An additional advantage of tree-based models is that they deal from small to large data sets (Markham et al., 2000; Razi and Athappilly, 2005). They also deal with non-linear data and a large number of predictors by removing the irrelevant ones from the analysis for improved accuracy (Gardner and Dorling, 2000). Tree-based models are also considerably powerful, simple, and flexible for regression, classification, and survival analysis (compared to other algorithms such as naïve Bayes and logistic regression), making them attractive and widely used in different fields (Rahmati et al., 2019). Nonetheless, there are different types of tree-based models (Rahmati et al., 2019), which plays an important role in model performance. Among them, the ensemble methods, such as random forests and boosting trees, markedly outperform the other methods (Bertolino et al., 2020; Wang et al., 2020). Random Forest has an additional advantage of retrieving the variable importance, partial dependence, and variable interactions plots, which are useful to better comprehend the complexity of the oceanographic processes. As such, for the Santos project, we decided to use this technique to model univariate response data.

Yet, if the goal is to model multiple response variables simultaneously (e.g., species composition), among the most used methods in environmental science are the multi-target regression (MTR), neural networks (Virts et al., 2020), and hybrid modelling. These methods have the advantages of modeling nonlinear associations with a variety of data types and require no specific assumptions concerning the distributional characteristics of the independent variables (Webb et al., 2017). While in the MTR the relationships between inputs and outputs can be highly structured reducing the accuracy in the test data, the neural network and the hybrid modelling has the benefit of learning a continuous function and thus making better predictions on new collected data outside the range of training data. Among the neural networks, self-organizing maps (SOM) have been used successfully as an unsupervised approach to disentangle complex patterns of biological and environmental data (Kohonen, 2001; Tison, 2004; Kangur et al., 2007). The generated map of neurons can be further clustered into major groups. In the case of a species composition data set, such cluster would represent potential association of species. Thus, to find the set of environmental variables that best predicts the species association, a hybrid approach can be performed, combining neural network with a tree-based approach, for instance. In this case the clusters obtained from the SOM analysis are considered as the response variable (Y) and the environmental variables are the predictors (X).

In the next sections, we provide a brief description of each machine learning technique to be explored in the Santos Project.

Random Forests (RF)

RF is among the most used machine learning techniques in environmental studies (Effrosynidis and Arampatzis, 2021). The RF is a tree-based method that consists of non-parametric statistical approaches for conducting regression and classification analyses by recursive partitioning of the data in function of the predictors (Hastie et al., 2009). As no implicit statistical distribution assumptions are needed, recursive partitioning of the data is a useful approach when the response variable is heterogeneous, and the predictors may be associated in some non-linear fashion. The RF methodology (Breiman, 2001) uses a collection of decision trees to increase prediction accuracy (Breiman, 1996; Freund and Schapire, 1996; Bartlett et al., 1998). Decision trees build the rule by recursive binary partitioning into regions (also called nodes) that are increasingly homogeneous with respect to the class variable (Cutler et al., 2007). At each step in fitting a classification tree, an optimization is carried out to select a node, a predictor variable, and a cut-off or group of codes (for numeric and categorical variables respectively) that result in the most homogeneous subgroups for the data, as measured by the Gini index (Breiman, 2001). This process continues until further subdivision no longer reduces the Gini index (known as fully grown trees). The terminal nodes encompass the sampling error and thus pruning the lower branches is an important modelling step of decision trees (Breiman, 2001).

However, while decision trees handle multiple predictors, it has been shown that single tree models are inaccurate to make predictions out of the range observed in training set (Rahmati, 2019). The RF is formed by a combination of unrelated trees. The forest of trees is built on bootstrap samples of the observations and on a random selection of the predictors to be used to determine the best split at each node. A bootstrap procedure is done with replacement and for each tree a specific portion of the data is used for constructing the tree, while the remaining (out-of-bag) are left out for estimating the predictions and errors of the model. Because a different bootstrap sample is used to grow each tree, there is a different set of out-of-bag observations for each tree. RF produces an importance score for each variable across all the trees that can be used to recognize the most important variables (Figure 2). Using unrelated ensembles of trees increases model accuracies and guarantees that, when multicollinearity is present among predictors, the inclusion/exclusion of single variables will have small individual effects on the overall accuracy of the model (Cutler et al., 2007)

One of the most interesting aspects of tree-based methods is the possibility to understand the interaction among predictors. The interaction term does not need to be explicitly specified to be utilized. This is particularly important for large-scale heterogeneous data sets, where structuring processes are not homogenous across the study area. For single classification trees approach, the interaction effects are easily represented and interpreted across the nodes of the tree. Yet when ensembling multiple trees, the interpretation and representation of potential interaction effects is less straightforward, and the interaction effects may be confounded with their marginal effect (Lunetta et al., 2004). A way to explore the interaction effects among variables in RF is to retrieve all the trees that share the variable that occurred most of the time in the root node (e.g., Depth, Figure 2). Having fixed the trees, the mean minimum node position of the other variables in trees is calculated (conditional mean_min_depth) and compared with the mean minimum node position of this interaction considering all the trees in the model (unconditional mean_min_depth; Figure 3). Note that the unconditional mean minimum depth is also represented from nodes 1 to 22 in Figure 2.

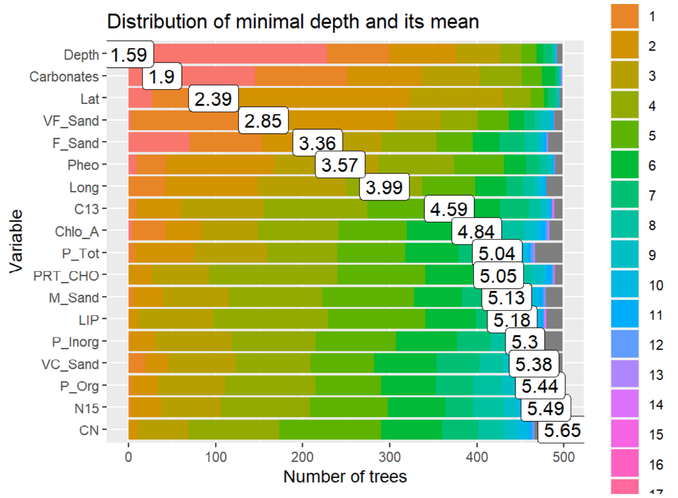


Figure 2. Example of a Random Forest feature importance histogram. List of variables returned after calculating 500 trees and keeping the significant ones. The ranking is done according to their mean depth distribution among all trees. The smaller the value the more important is the variable. Colors legend indicates the node position in the tree. Abbreviations used in this figure are listed in Gallucci et al. 2023.

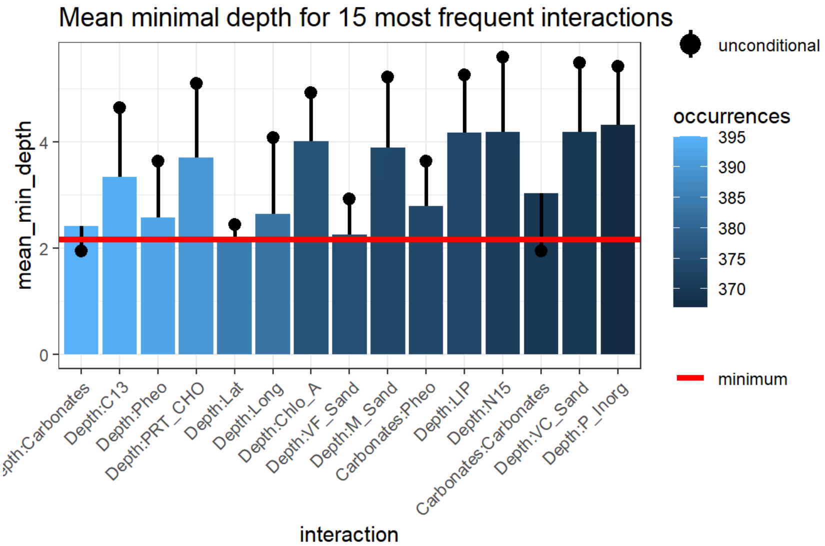


Figure 3. Example of an interaction effects plot with mean minimum depth of the 15 most relevant variables among the analyzed trees. Bars indicate the constrained results (top trees) while black dots the unconstrained results (all trees). Red line represents the minimum depth of an interaction effect among all trees. Blue gradient represents the number of occurrences among all trees.

Self-organizing maps (SOMs)

Like other multivariate techniques, such as principal component analysis (PCA), principal coordinate analysis (PCoA) and non-metric multidimensional scale (nMDS), K-means, t–Stochastic Neighbourhood Embedding (t-SNE) or Uniform Manifold Approximation (UMAP), SOMs allow to visualize objects onto a bi-dimensional plane in such a way that similar objects are close together and dissimilar objects are far away from each other (Wehrens and Kruisselbrink, 2018). The main distinctions of the SOM from these techniques are the possibility to project new objects into the known bidimensional space and the discretization of the mapping into units (also termed neurons), meaning that the bidimensional space is not continuous. As such, while the continuous space of a PCA or nMDS plot, for instance, highlights data points dissimilarities, the discretization of the space from the SOMs emphasizes the similarities among them (Wehrens and Buydens, 2007). Such discretization of the multidimensional space can be much better to handle, for instance, non-linear variables. Each unit, represented as a hexagon or square (Figure 4), is associated with a codebook vector. The codebook corresponds to the average of all objects mapped to that unit, representing a "typical" object for that area of the map. Mapping data to a SOM is the process of calculating the distance of new data points to the codebook vectors and assigning each object to the unit with the most similar codebook vector (the best matching, or "winning", unit). During training, objects are repeatedly and randomly presented to the map. The unit most similar to the current training object ("winning unit") will be updated to become even more similar; the weighted average is used in the subsequent step. The learning rate (α) is the amount of distance that will be considered to update the codebook; it is a small value in the order of 0.05 that decreases constantly so that the map converges. As such, in the beginning of the learning cycle, each new data point has a large effect on the network, but as soon as more data is presented the network tends to stabilize (Figure 4). At the end, neighboring units in a SOM tend to have similar codebook vectors. In addition to the codebook, analysis returns the profile of the learning process of the network, the number of observations per unit and the distance between adjacent units (U-matrix, Figure 4). SOMs have been successfully used in several scientific fields and a detailed explanation on the mathematical principles can be found elsewhere (Chon, 2011; Van Hulle, 2012; Clark et al., 2020). In environmental science in particular, the use of SOM has been applied in various fields, including engineering, ecology, agriculture, health, etc. (Chon, 2011). For example, SOMs were efficiently used to pattern macrofauna and fish communities (Penczak et al., 2005; Park et al., 2007) and to predict the performance of a wetland agroecosystem and assessment of nutrient removal performance (Zhang, J.-T. et al., 2008; Zhang, L. et al., 2008).

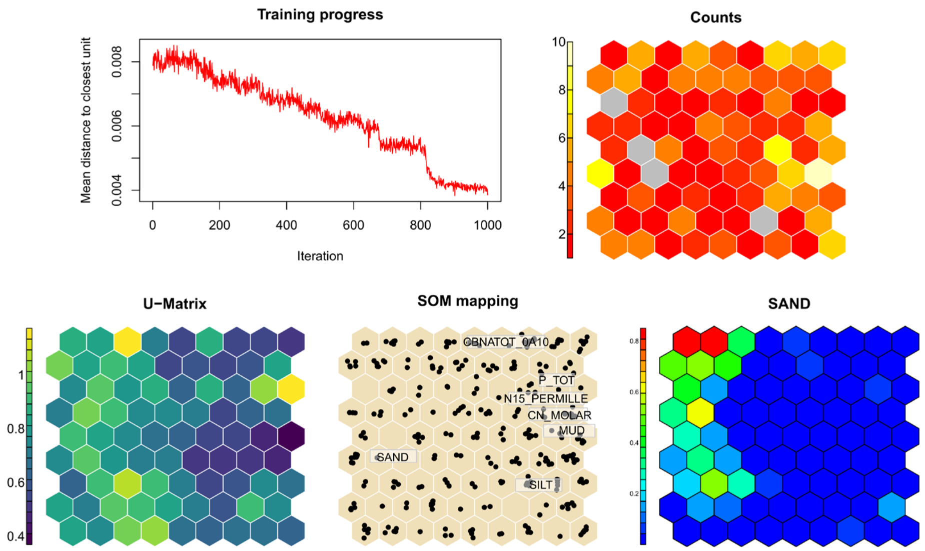


Figure 4. Examples of five graphical outputs of an unsupervised SOM analysis performed on a network based on hexagon units. Training progress indicates the learning rates which is represented by the mean distance of a sample to its closest unit along the interaction. Counts indicates the number of samples allocated in each unit of the network, with shading colors representing its quantities, grey units represent empty cells. Umatrix represents the relative distance between neurons, with shading representing the distance. SOM mapping with dots representing the samples with variables superimposed according to its relevance on the codebook. Single variable map (e.g., Sand) showing the values of the codebook in the network; shading representing the quantities.

Hybrid modelling: RF + SOM

In machine learning, the hybridization approach has been an active research area to improve the classification/prediction performance over single learning approaches (Jain and Kumar, 2007; Tsai and Chen, 2010; Chou et al., 2013; Park et al., 2013). In general, hybrid models are based on combining two or more machine learning techniques. For the Santos project, the hybrid model consisted of using the clustering results obtained from the SOM or from the hierarchical clustering (unsupervised learning) as a response variable in the RF (supervised learning). Such hybrid approach had been already explored before (e.g. Bilski, 2017; Ma et al., 2021). The assumption of the framework used for the multivariate data is that the association of neurons obtained with SOM becomes a response variable (Y) which can be predicted with RF by an independent data set (X). In this hybrid approach, it is also possible to perform long term learning procedures. When new data is collected, it is tested against the distribution of the predictions generated by the RF and depending on the heterogeneity of the new data, it could be either classified in an already known group (i.e., it is within the predicted interval) or attributed to a new group (out of the predicted interval). In both cases, the new data will become part of the model and used to make new predictions. In this way, the model is constantly learning with the acquisition of new data (L'Heureux et al., 2017).

Overcoming the analytical challenges of the Santos project

Challenge 1: definition of the conceptual model (cause-effect relationships)

Understanding complex systems requires approaches from different scientific fields. This knowledge is built independently, like pieces of a fragmented puzzle. The visualization of the complete puzzle requires thus an additional effort to connect the parts. The logical exercise of establishing the links between the compartments of the system (e.g., granulometry, macrofauna, meiofauna, organic matter) is the first and certainly the most crucial step to achieve a unified understanding of the system's complexity. Such conceptualization, commonly referred to as conceptual model, focuses on synthesizing the current scientific knowledge, guiding data collection and subsequent data analysis, simulating, and predicting unknown scenarios and, most importantly, supporting management decisions (Franks, 2018). A conceptual model does not require any implementation of paradigm or software solution. It involves the abstraction of a model from the real system, identifying what must be modelled and how (Furian et al., 2015). The formulation of a conceptual model involves 5 steps (modified from Robinson, 2008): problem situation (research question), objectives, content (entities, relationships), data input and result outputs. All these steps assumptions and simplifications must be made to achieve the simplest model that still meets the proposed objectives.

As an example, for the Santos project, the problem is the sustainable use of the Basin which encompasses one of the largest oil and gas reserves in the world. The objective of the project is to understand the spatio-temporal dynamics of the benthic, pelagic, and physical systems to give support for conservation and monitoring programs (Moreira et al., 2023 (Figure 5). Each system is further subdivided into abiotic and biotic components, each of them is composed by several research areas, such as chemistry, geology, biology, oceanography, and so on. The goal proposed by the Santos Project is characteristic of complex systems. It encompasses a large amount of information collected from different disciplines, covers a large and heterogeneous geographical area, and it is characterized by the natural interdependencies of the components of the system. For the benthic system, the proposed conceptual model has been constructed to answer seven research questions (Figure 5): (1) Are the sediment properties of the Basin spatially structured, characterizing distinct sedimentary processes? And, if yes, can the basin geomorphology, bathymetry, and bottom-currents explain them? (2) Are the indicators of quantity and quality of the organic matter associated with the sedimentary environments? (3) Are the concentrations of alkanes covarying with the organic matter as an indication of diffuse pollution? (4) Are the metals in the sediment indicating any source of pollution? (5) Are the species from each benthic compartment (meiofauna, macrofauna, microbiota and foraminifers) organized into taxonomic associations and are they explained by the sedimentary environment, quantity and quality of food sources and potential pollutants? (6) Are the responses of the benthic groups congruent? (7) Which are the most appropriate indicators (abiotic and biotic) to monitor the benthos of the Santos Basin? The motivations, hypotheses and detailed methodologies supporting each of these questions will be addressed elsewhere (e.g. Carreira et al., 2023; Moura et al., 2023; Gallucci et al., 2023).

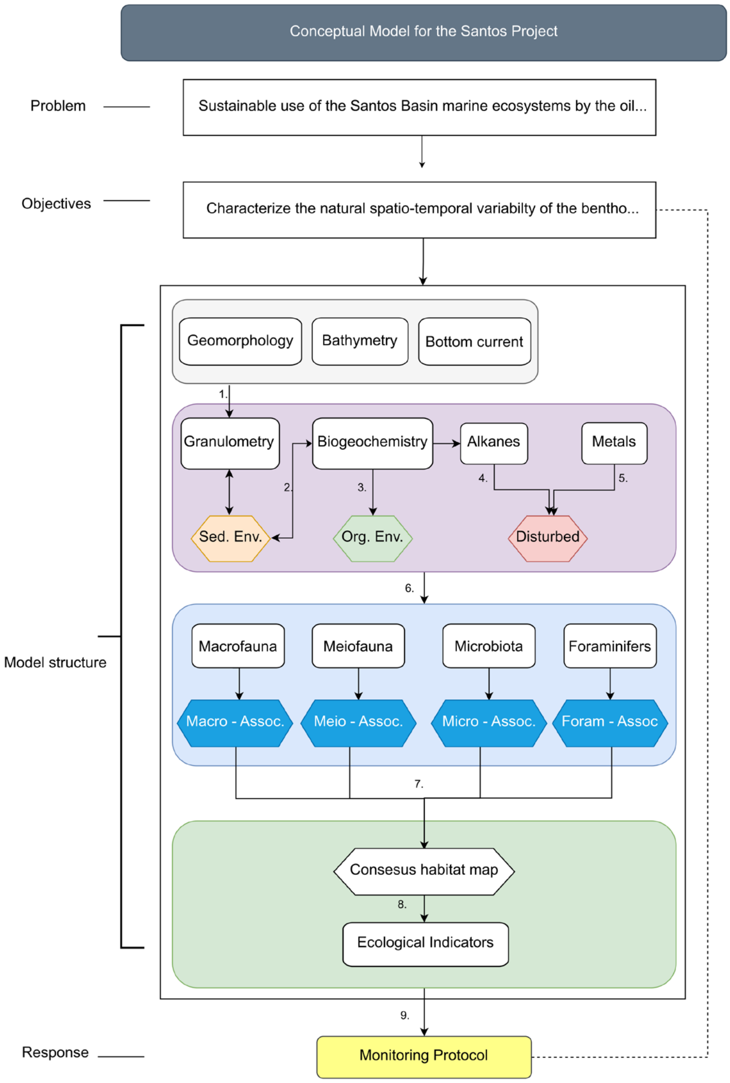


Figure 5. An example of a conceptual framework that has been applied in the Santos project to achieve an integrated monitoring program of the benthic system.

It should be noted that the proposed framework is composed of collected data (rounded rectangles, Figure 5) and analyzed data (diamonds). For instance, the diamond variable “Sediment Environment” is a categorical variable with the types of sedimentary environments that have been determined after analyzing the granulometric properties of the sediment in response to the bathymetry, geomorphology, and bottom currents (Figueiredo Jr. et al., 2023).

Variables generated from an analysis carries the collective property of the respective data set and the interpretation of the authors. Thus, the use of the analyzed variables as predictors in a subsequent research question (e.g., research questions 3 and 6; Figure 5) improves data interpretation. As mentioned before, the framework was built to permit a continuous insertion of new data allowing establishing a long-time learning approach (dashed line Figure 5).

CHALLENGE 2: DIFFERENCES IN THE QUANTITY OF DATA SOURCES IN THE STUDY AREA

Integrating data across disciplines becomes critically challenging when dealing with different data resolutions. In the Santos project, for example, there were two main issues: one dealing with the differences in the number of replicates per sampling station and another dealing with the spatial resolution of a given variable. Most of the biological variables (e.g., meiofauna and macrofauna) were collected in triplicates from 100 geo-located stations at two distinct periods (2019 and 2020). Other variables, such as the microbiota (Paula et al., 2023) and radioisotopes (Moreira et al., 2023), were sampled without replicates per sample station. In this example, integrating meiofauna and radioisotopes data requires either upscaling the radioisotope data by copying it across the replicates or downscaling the meiofauna data with the mean value per station. Particularly, for the Santos project the downscaling is recommended to reduce the small-scale variability that appears to be highly heterogeneous and unpredictable across the Basin (Gallucci et al., 2023).

Regarding spatial resolution, there are two distinct sets: one is the sampling design on the 100 stations with dozens of variables simultaneously collected and the second is the high-resolution bathymetrical data, arranged in a 2 km by 2 km) resolution grid totalizing more than 100,000 points (Figure 6). In this case, to have the distribution of a biological variable over the high-resolution bathymetrical map, the modelling approach was performed in two-steps: the first steps are the base-model and the meta-model. The base model consists of modelling the response variable (e.g., species richness) in relation to the environmental variables across the 100 stations. In this model bathymetry and coordinates are included to guarantee the interpretation across both models and potential autocorrelation of the response variable. The predicted data obtained from the base-model is further used on the meta-model, which is based solely on bathymetry and geographical coordinates. The predictions of the meta-model can be now expanded for the whole high-resolution bathymetrical grid. Note that the same procedure can be repeated for the error estimates of the base-model permitting an integrated interpretation of the predictions and errors across the Basin. An example on how to perform this modelling approach is available at Gallucci et al. (2023). All the analytical steps involved are presented below as analytical workflows.

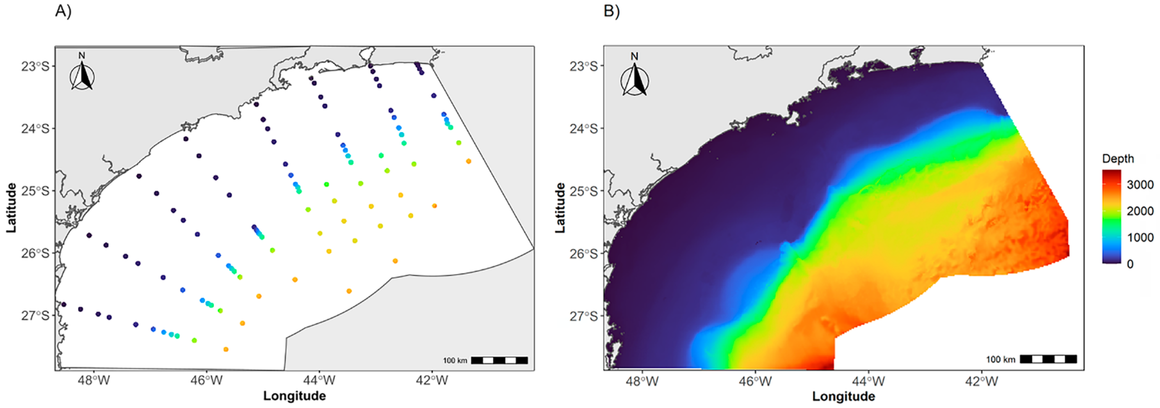


Figure 6. Two bathymetric spatial resolutions of the Santos Project: (A) sampling design composed of 100 stations; (B) bathymetric data containing more than 100,000 points.

Challenge 3: detection of noisy data

Large datasets formed by multiple variables are frequently characterized by distorted or corrupted data because of methodological inaccuracy, typing errors, or any other uncontrolled situation. Also known as noisy data or outliers, they usually disrupt the expected distribution of the data. The presence of noisy data may dramatically affect model performance and interpretation (Gupta and Gupta, 2019). Scanning, recognizing and eventually removing or replacing them are thus a prerequisite of any analytical framework. Nonetheless, sometimes it is hard to distinguish between noisy examples and true exceptions, and henceforth many techniques have been proposed to deal with noisy data sets with different degrees of success. The detection of noisy data is in practice a classification problem that tests whether a given observation belongs or not to the distribution of the statistical population under analysis (Loureiro et al., 2004). For a systematic review on the topic, we refer to García et al., (2015b) and Gupta and Gupta (2019). The detection of noisy data goes along with the analytical approach. As mentioned before, for the Santos project, Random Forests and Self-Organizing Maps are being recommended to achieve a better understanding of the oceanographic processes structuring the benthic communities. These two techniques classify the observations into classes and therefore are among the algorithms termed as robust leaners (sensu García et al., 2015b), meaning that the presence of few extreme values have little effect on the analysis outcome. When using the RF, the detection of noisy data can be done while analyzing the different types of error estimations between predicted and observed data. Yet a way to evaluate the presence of noisy data in a multivariate dataset is to explore the outcomes of the network (Muñoz and Muruzábal, 1998). As mentioned earlier, SOM is a classification analysis. When performing the SOM, it is possible to retrieve the number of observations per unit of the network (also termed neuron), the position of the neuron in the network, as well as the relative distance of each neuron to its neighbors (the U-Matrix) (Ultsch, 2003). When comparing these three outcomes together it is possible to check whether an observation is an outlier within the data. A common characteristic of noisy data is to be isolated in a neuron, which is close to the borders of the network and with a high relative distance from its neighbor (Gupta and Gupta, 2019) (Figure 7). In both cases, when noisy data is present, it is worth checking the observation and, eventually, re-analyzing the data without it. If removed, this empty cell must be treated as a missing value as explored anteriorly. Nevertheless, before removing a data point, we suggest a careful investigation of its nature. Not all outliers are meaningless data. For instance, in a monitoring context an outlier could be an early sign of a disturbed sample (Yang et al., 2019; Yotova et al., 2021).

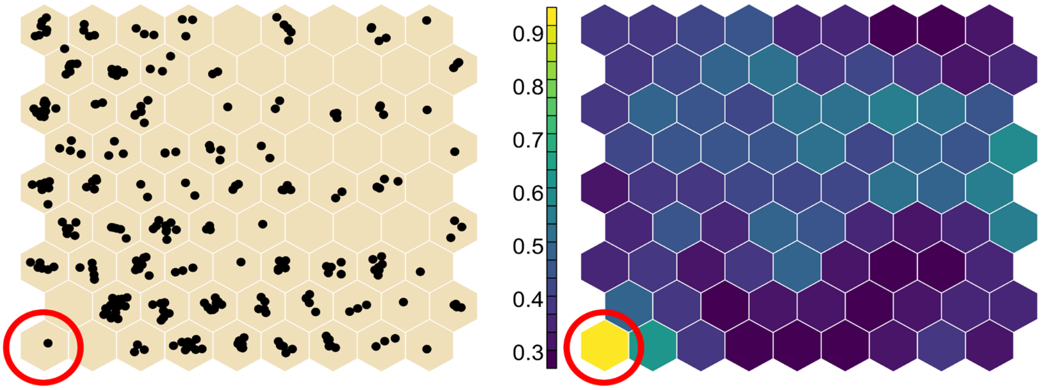


Figure 7. Graphical representation of the SOM analysis. Left: Best matching unit with dots representing the observations per hexagon. Right: U-matrix with colors range representing the relative distance between adjacent units. Red circle represents a potential outlier of the data set.

Challenge 4: missing data

Missing data is present in almost all large-scale databases. Nevertheless, to run a statistical model a complete dataset is usually desired. Removing information is one way to solve it but depending on the dataset this may limit the analysis and compromise the study. Imputing the empty cells with values is an alternative process to complete the dataset and the practice is already well established in the literature (García et al., 2015a). There are different techniques for dealing with missing data (Poulos and Valle, 2018; García et al., 2015a) and they can be roughly classified as supervised or unsupervised. The supervised approaches explicitly model the response variable and retrieve the predicted value (or category) correspondent to the missing observation. The unsupervised approaches infer the missing data from a probability distribution (Little and Rubin, 2002). They can be done by means of clustering methods, density estimations or basic statistical measures, such as the median or mean value.

As an example on how to handle missing data, in the Santos project, density of meiofauna has 13 missing observations, which limit an analytical comparison across the benthic groups. As explored elsewhere (Gallucci et al., 2023, this issue), meiofauna density depends on a variety of environmental conditions. When considering the mean density per stations, the modelling accuracy was 74% in the training data and 79% in the test portion. Nevertheless, as discussed by the authors, the model was incapable of predicting density of meiofauna in each replicate. So, if the objective is to use the mean value per station, the same parametrization used by Gallucci et al. (this issue) will return accurate predictions for the missing values. But if one intends to fulfill the missing replicates, the supervised modelling will return inaccurate values and, in this case, alternative unsupervised methods should be explored.

For instance, in addition to the 13 missing values of the meiofauna (Gallucci et al., 2023, this issue), we have simulated different percentages of missing values up to 10% and performed distinct imputation techniques, namely: k-nearest neighbors (with number of k varying from 2 to 30), median and bag impute. The analysis showed that for 1% of missing values, all of them have little effect on predicting the real data (Figure 8). But, as soon as the amount of missing data increases, accuracy decreases. The best performance came from the bag imputation, which even after removing 10% of the data, reduced 1% of the R². (Figure 8). As stated before, model accuracy decreases with increasing numbers of empty cells (Jordanov et al., 2018) and the performance of the method used depends on the characteristics of the dataset (Platias and Petasis, 2020).

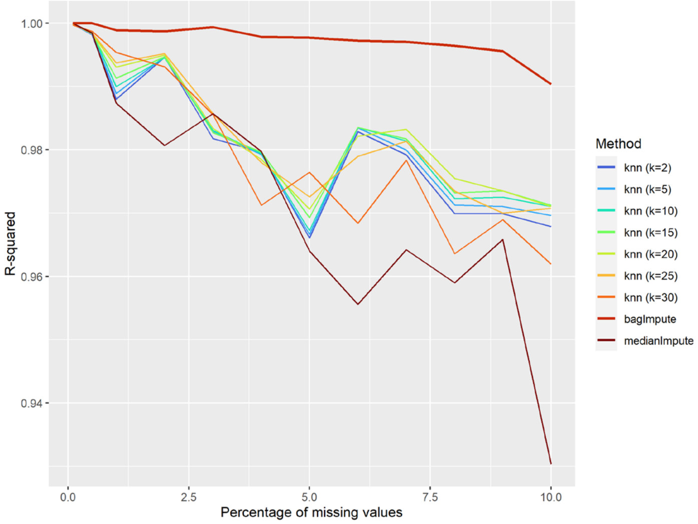


Figure 8. R-squared between observed and imputed data: k-nearest neighbors (with number of k varying from 2 to 30), median and bag impute.

Challenge 5: model optimization

Machine learning model optimization is the process of fine-tuning a model to achieve the best performance. The optimization process can involve a variety of techniques, including selecting the appropriate model architecture and adjusting hyperparameters.

Hyperparameters are the parameters of a model that are set before training begins, and they can have a significant impact on the performance of the model. Two popular methods for tuning hyperparameters are grid search and random search. Grid search involves specifying a set of potential values for each hyperparameter and training the model with every combination of these values, while random search randomly samples from the specified range of values for each hyperparameter.

The most important hyperparameters for the RF and SOM algorithms are described below:

RF hyperparameters

The main hyperparameters of the RF models are the number of trees and the number of variables randomly sampled as candidates at each split (commonly termed as '*mtry*') (Biau and Scornet, 2016; Mahesh, 2020).

In general, the greater the number of trees, the better, as the out-of-bag errors tend to reduce and stabilize in the long run. However, the calculation and storage of large forests increases the time of computer processing. The threshold is then between information gain versus processing time, making it unnecessary to process thousands of trees at each run for a little increase in model accuracy (Oshiro et al., 2012). The recommendation is therefore to use a reasonable number of trees that gives the chances that a large proportion of the dataset (observations and predictors) can be used during the learning process. A possible strategy is to run the model with a different number of trees and stop the training when the last model does not improve performance by more than one choice level (Probst and Boulesteix, 2017).

The number of variables randomly selected at each split (hereinafter referred as *mtry*) is probably one of the most influential RF hyperparameters (Probst, Boulesteix, 2019). The selection of few variables at each split leads to more different and less correlated trees in the forest, but given a large number of trees, it increases the stability of the model (Probst, Wright, 2019). The number of relevant predictor variables strongly influences the optimal *mtry*. On the one hand, if there are many relevant predictor variables, *mtry* should be set small since it will allow to explore multiple models that have considered solely less influential variables, therefore providing small but relevant performance gains (Bernard et al., 2009). On the other hand, if there are only a few relevant variables out of many, *mtry* should be set high, so that the algorithm can find the relevant variables (Goldstein et al., 2011). A strategy to find the optimal *mtry* is to make a grid or a random search across multiple *mtry* (e.g., ten variables interval steps) and select the one with the best accuracy. Nevertheless, the processing time increases proportionally by increasing the search. At the end, we have to deal with a trade-off between stability of the model, accuracy of the single trees and overall processing time. Eventually, an increase *mtry* intervals may promote a gain in accuracy in the third decimal place of the AUC or R^2, without causing major changes in the ranking of variable importance (Fox et al., 2017).

SOM hyperparameters

The SOM optimization involves hyperparameters related to grid topology and learning; exploring the interactions of all possible combinations and how they may affect the results is a matter of further investigation (Liu et al., 2006). Briefly, the number nodes and their arrangement is defined in four steps (Vesanto and Alhoniemi, 2000):

1. Determine the number of map nodes using the heuristic recommendation:

$$M=5\sqrt{N}$$

where N is the number of observations in the input dataset,
2. Determine the eigenvectors and eigenvalues in the data from the autocorrelation matrix,
3. Set the ratio between the two sides of the grid equivalent to the ratio between the two largest eigenvalues,
4. Scale the side lengths so that their product (xdim * ydim) is as close as possible to the number of map units determined above.

Other two parameters related to the network topology are the neighborhood factor and the map dimension. Considering the neighborhood, two functions are the most used: “bubble” and “Gaussian.” Bubble is a constant function in the defined neighborhood of the winning neuron, that is, every neuron in the neighborhood is updated with the same proportion of the difference between the neuron and the presented sample vector (Stefanovič and Kurasova, 2011). This function can be either set as a constant or in a linear decreasing fashion by adjusting the starting and ending radius of the learning process. In comparison, the Gaussian function is a function that already smoothly decreases the defined neighborhood distance of the winning neuron throughout the learning process. Overall, the Gaussian function returns smaller errors (Natita et al., 2016; Ng and Chan, 2019); however, depending on the size of the dataset, it is computationally demanding as the exponential function has to be calculated during the learning processes. In this latter case, the bubble function, together with the choice of a starting radius (r1) larger than the final radius (r2), is a good compromise between the computational cost and the approximation of the Gaussian.

The choice between performing a SOM in a two-dimensional space (planar shape) or a three-dimensional form, such as a torus, depends on the type of data and objectives of the analysis. During training, best fitting units at the margins of the array influence fewer neighbors than those at the center (Lo and Bavarian, 1991). If the objective is to find clusters of similar units, it can happen that similar units will be discontinuously arranged along the margins of the SOM space (Mount and Weaver, 2011). Such discontinuity may promote incorrect interpretation of the topological relations between units. The solution for the discontinuity problem can be solved by using a continuous 3-D mapping, such as the Torus-shape. Performing a toroidal map means that the units on one edge of the planar map are connected to adjacent units on the opposite side (Figure 9). In comparison to the planar shape, the use of a toroidal SOM reduces the number of splitted clusters and the estimation of these observations that were relatively close to the expected value (small errors), but it did not differ in performance for the larger errors (Mount and Weaver, 2011). Additionally, while for the planar map larger errors will be placed closer to the borders, at the torus shape they are more diffusely distributed across the map. The torus map also adds some complexity in data visualization and in the end also ends up in a planar projection.

An additional setting possibility is the choice of learning algorithm. In Kohonen 3.0 implemented in R and in the iMESc (Vieira and Fonseca, 2023) three possibilities exist: “online” (also termed sequential), “batch” and “pbatch.” In all cases, training objects are compared to the current set of codebook vectors. The difference between the online and batch SOM algorithms is that the update of the winning unit(s) in the online algorithm is done after each individual object is presented to the network, whereas in the batch algorithm the objects are partitioned into groups to speed up the learning processes (Kohonen, 1990; Vesanto et al., 1999, 2000; Wehrens and Kruisselbrink, 2018). In the batch mode each group is used to update the corresponding weight vector and the final codebook is updated after all groups have been presented. The “pbatch” is similar to batch but the task of finding the best matching units for all records in the dataset is split up over different cores (Lawrence et al., 1999). In the batch and pbatch algorithms, the learning rate function (alpha) of the sequential algorithm is no longer needed, but, like the sequential algorithm, the radius of the neighborhood may decrease during the learning process (Liu et al., 2006). It is worth noting that even with identical settings, repeated training of a SOM will lead to different mappings, because of the random initialization. Setting a seed solves this issue, but it is always wise to train several maps before making conclusions.

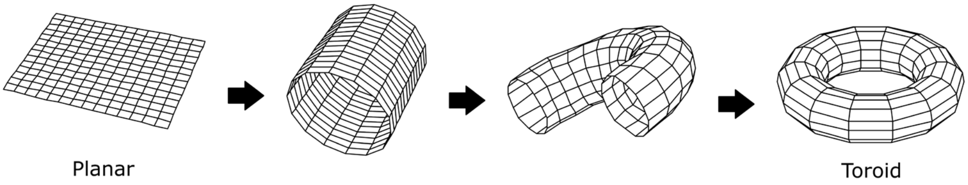


Figura 9. Exemplification of the relationship between the planar and toroid arrangements of a SOM.

Challenge 6: imbalanced balanced number of observations across factors

In large-scale studies, such as the Santos project, it is common to have the distribution of observation unequally across the categories. This may result in models that may have a high overall accuracy, but poor predictive performance for the minority category. One possible solution is partitioning the training and test parts in a balanced fashion (Chawla et al., 2002), also known as data-level methods (Krawczyk, 2016). Yet, for imbalanced designs this simple solution means losing substantial amounts of data during the training phase hampering the construction of an accurate model across the categories. As such, understanding the effects of imbalanced data on model accuracy is a major issue that has gained considerable attention recently (Krawczyk, 2016; Chicco et al., 2021). An important tool to explore the accuracy of the model across categories is to calculate the confusion matrix and the additional indices that can be derived from it (e.g., specificity, precision, recall rates; Table 1) (Chicco and Jurman, 2020; Chicco et al., 2021). Among the indices we call attention to those that normalize the observed accuracy by the expected accuracy, such as the Cohen´s Kappa statistics (Landis and Koch, 1977; Chicco and Jurman, 2020; Chicco et al., 2021). The expected accuracy in the Kappa statistics is based on the frequency of observations among the categories and therefore will cope with the skewed distributions typical of imbalanced studies (Jeni et al., 2013).

For instance, for the Santos project it has been suggested that the meiofauna in the Basin is arranged in six benthic zones with distinct number of stations (LPP: La Plata Plume - 9; CCS: Central Continental Shelf- 8; CFU: Cabo Frio Upwelling - 15; CB: Carbonate zone -8; CS: Continental Slope -21; DS: Deep-Sea -38). Assuming that these zones will be used in a monitoring program, we can assess how accurate our environmental model is performing in each zone. To facilitate the interpretability, we will explore this example following the same rationale and parametrization proposed by Gallucci et al. (2023); i.e., perform the random forest algorithm using 38 predictors, and in this case our supervisor is the benthic zones. The overall statistic returned a model with 92% of accuracy for the training portion and 95% for the test (Table 2). This accuracy is significantly higher than the non-information rate of 39%. The kappa statistic was lower than overall accuracy pointing that the model performance varied among the benthic zones as a result of the imbalanced number of stations.

When analyzing the accuracy individually by zone, indeed the performance of the model varied from 98% and 78% in the training phase and between 100% and 50% in the test phase (Figure.. 10). Note that the lowest performance in the training phase was for the CSS, while for the test phase it was for the CB. Actually, in the test phase, except for the CB, all others have 100% accuracy. When compared to the other groups, the lower performance of the model for the CSS in the training phase is explained by the lower proportion of the positive class correctly predicted (sensitivity) and consequently a lower average of the true positive and true negative rates (balanced accuracy) (Table 3). The misclassified stations from the CCS were placed in the adjacent zones (LPP and CB), an indication of similarity between them. An interesting aspect to be observed is that this variation in the performance during the training phase was not mirrored in the test phase (Table 3). For the test phase, the lower accuracy observed for the CB is associated with the lower sensitivity and F1-score. The F1 score is the harmonic mean between the proportion of positive identifications that were correct classified (Precision) and sensitivity, that is, it considers both false positives and false negatives. In the particular case of the CB zone, there is a prevalence of false positives (i.e., the model classified as CB, but they were actually CCS). The interpretation of the confusion matrix with these additional indices help to understand the limitation of the model and discuss the environmental processes shaping the observed pattern.

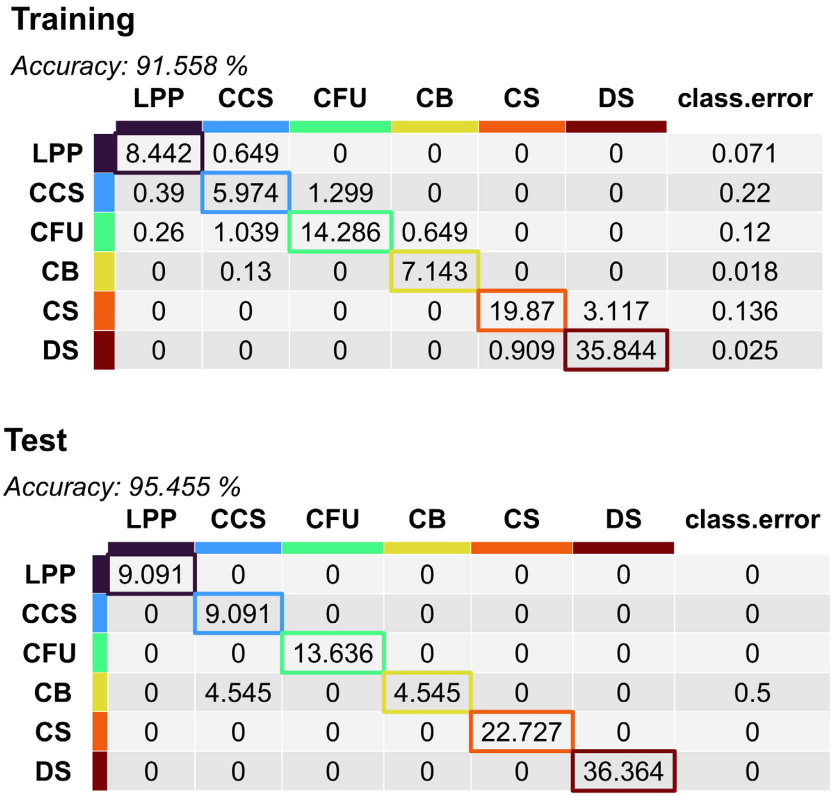
Table 1. Statistical measures that can be calculated from a confusion matrix

1	Accuracy	Acc	the proportion of predictions that the model classified correctly
2	Misclassification rate	Mis	The proportion of predictions that the model misclassified
3	Confidence Interval	CI	a likelihood that the true accuracy for this model lies within this range
4	No-information rate	NIR	the largest proportion of the observed classes.
5	Kappa	k	the accuracy of the classifier normalized by the expected accuracy simply by chance
6	p-value	p-value	the significance of the accuracy performing better the no-information rate
7	Sensitivity or Recall	Sens	the proportion of the positive class correctly predicted
8	Specificity	Spec	the proportion of the negative class correctly predicted
9	Precision	Prec	The proportion of positive identifications that were correct
10	Prevalence	Prev	the frequency of the positive class in the model
11	F1 Score	F1	the harmonic means between precision and sensitivity
12	Positive Predictive Value	PPV	the number of the positive class correctly predicted as a proportion of the total positive class predictions
13	Negative Predictive Value	NPV	the number of the negative class correctly predicted as a proportion of the total negative class predictions
14	Detection Rate	DR	the number of correct positive class predictions made as a proportion of all of the predictions
15	Detection Prevalence	Dprev	the number of positive class predictions as a proportion of all predictions
16	Balanced Accuracy	BA	The average between the true positive and true negative rates

Table 2. Overall statistical measures of the meiofauna benthic zones model

	Training	Test
Acc	0.92	0.95
k	0.89	0.94
95%CI	(0.89 - 0.93)	(0.77-0.99)
NIR	0.39	0.36
p-value	<0.01	<0.01

The following workflows were developed to handle the research problems from the Santos project (see the conceptual model). Basically, two distinct approaches have been created: 1) to predict a continuous variable (e.g., density of a taxon) and (2) to detect and predict species associations (multivariate data set). Both approaches are composed of a base- and a meta-model (see challenge 2) which aim to spatialize the predictions of the base-mode into a higher resolution bathymetrical map. They are easily applied to abiotic and biotic research questions (see the conceptual model).



Training

Accuracy: 91.558 %

	LPP	CCS	CFU	CB	CS	DS	class.error
LPP	8.442	0.649	0	0	0	0	0.071
CCS	0.39	5.974	1.299	0	0	0	0.22
CFU	0.26	1.039	14.286	0.649	0	0	0.12
CB	0	0.13	0	7.143	0	0	0.018
CS	0	0	0	0	19.87	3.117	0.136
DS	0	0	0	0	0.909	35.844	0.025

Test

Accuracy: 95.455 %

	LPP	CCS	CFU	CB	CS	DS	class.error
LPP	9.091	0	0	0	0	0	0
CCS	0	9.091	0	0	0	0	0
CFU	0	0	13.636	0	0	0	0
CB	0	4.545	0	4.545	0	0	0.5
CS	0	0	0	0	22.727	0	0
DS	0	0	0	0	0	36.364	0

Figure 10. Confusion matrices of the training and test predictions of the six benthic zones (Gallucci et al. 2022): LPP: La Plata Plume; CCS: Central Continental Shelf; CFU: Cabo Frio Upwelling; CB: Carbonate zone; CS: Continental Slope; DS: Deep-Sea.

Table 3. Statistical measures from the confusion matrix by meiobenthic zones (Gallucci et al. 2022): LPP: La Plata Plume; CCS: Central Continental Shelf; CFU: Cabo Frio Upwelling; CB: Carbonate zone; CS: Continental Slope; DS: Deep-Sea.

		Sens	Spec	PPV	NPV	Prec	F1	Prev	DR	DP	BA
Training	CB	0.92	1.00	0.98	0.99	0.98	0.95	0.08	0.07	0.07	0.96
	CFU	0.92	0.98	0.88	0.98	0.88	0.90	0.16	0.14	0.16	0.95
	CCS	0.77	0.98	0.78	0.98	0.78	0.77	0.08	0.06	0.08	0.87
	DS	0.92	0.99	0.98	0.95	0.98	0.95	0.39	0.36	0.37	0.95
	CS	0.96	0.96	0.86	0.99	0.86	0.91	0.21	0.20	0.23	0.96
	LPP	0.93	0.99	0.93	0.99	0.93	0.93	0.09	0.08	0.09	0.96
Test	CB	0.50	1.00	1.00	0.95	1.00	0.67	0.09	0.05	0.05	0.75
	CFU	1.00	1.00	1.00	1.00	1.00	1.00	0.14	0.14	0.14	1.00
	CCS	1.00	0.95	0.67	1.00	0.67	0.80	0.09	0.09	0.14	0.98
	DS	1.00	1.00	1.00	1.00	1.00	1.00	0.36	0.36	0.36	1.00
	CS	1.00	1.00	1.00	1.00	1.00	1.00	0.23	0.23	0.23	1.00
	LPP	1.00	1.00	1.00	1.00	1.00	1.00	0.09	0.09	0.09	1.00

Machine learning workflows

A workflow for predicting a continuous variable

As an example of predicting a continuous variable, we tackle the density of Nematodes (Gallucci et al., 2023, this issue). This problem is based on a Random-Forest regression model, where the aim is to predict the density of Nematodes (Y) using a multivariate environmental dataset (X) and then generate a full-coverage prediction map based on 100,000 data-points. The base model includes 34 explanatory variables, whereas the meta-model includes only the explanatory variables available for the 100,000 data-points (longitude, latitude, and bathymetry). For both models, there are a total of 288 observations and one response variable (i.e., the observed density of nematodes). The workflow can be implemented using the iMESc application (Vieira and Fonseca, 2023), which provides a quite easy interface for training RF models, allowing automatic parameter tuning and reducing the requirements on the researcher's programming knowledge. Concrete examples on how these workflows can be used to interpret ecological data are given by Gallucci et al. (this issue), Carrera et al. (this issue) and Moura et al. (this issue).

An important aspect in any machine learning workflow is the model selection that consists of the search of a set of parameters in order to find the model with the best performance in predicting a particular set of data (Anguita et al., 2010; Probst, Philipp et al., 2019). As such, the model selection is strictly linked with the estimation of the generalization ability of a classifier. The generalization is assessed by the error rate attainable on unobserved data. The chosen model is characterized by the smallest estimated generalization error. Unfortunately, the tuning of the hyperparameters is not a trivial task and represents an open research problem (Aken et al., 2017; Probst, Philipp et al., 2019). Ideally, multiple parameters should be explored simultaneously allowing for a more complete search.

For the Santos project we suggest the following workflow.

The base model

1. Load both X and Y datasets (this has to be previously defined, see challenge 1).
2. Explore the quality of the data, such as the presence of noisy data (challenges 2 and 3).
3. Consider using data imputation techniques (e.g., k-nearest neighbor method) to "fill in" any missing values (challenge 4).
4. To allow for reproducible results, set a seed value and randomly split the datasets into two subsets: for example, 80% of the data are used for training, and 20% of the data are used for testing (validation).
5. Search for hyperparameters of an RF regressor model (challenge 5):
   a. Set multiple number of trees (e.g., 250, 500).
   b. Set the search type and length for the *mtry* hyperparameter (e.g., random; 10, 20).
   c. Set the resampling method (e.g., 5-fold cross validation, repeated 10 times) for internal validation.
   d. Train the RF models using the training dataset.
   e. Evaluate the internal model performances using metrics like Mean squared error (MSE), Root mean squared error (RMSE), R-squared (R2), Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE) to select the optimal model. In iMESc the tuning parameters with the lowest RMSE are automatically selected as the optimal model.
6. Assess the reproducibility of the optimal model reproducibility of the selected model through internal and external validation (section 1).
7. Explore the importance of variables from the optimal model.
8. Evaluate the interaction effects of the most important variables.
9. Explore the biplots of the interaction effects on the response variable.

The meta model

1. Restrict X dataset used in the base-model to the variables available for the 100,000 data-points (i.e., latitude, longitude, and bathymetry).
2. Use the base model to make predictions on the whole dataset (training and testing data). Use its predictions as the response variable (Y) (challenge 5).
3. Train the meta-model by repeating steps 2 to 7 from the base-model.
4. Use the meta-model to make predictions on the 100,000 data-points.
5. Spatialize the obtained predictions in a high-resolution map.

A workflow for predicting species association

The solution for this problem is based on the combination of the Self-Organizing maps, Hierarchical clustering, and Random Forest classification analysis. The aim here is to cluster the multivariate species abundance dataset (X_{sp}) into classes (associations), predict these associations using a multivariate environmental dataset (X_{envi}) and finally, generate a full-coverage prediction map (100,000 data-points). The workflow architecture comprises 29 steps across 4 major analysis: (1) Training an unsupervised SOM model using a species abundance dataset (X_{sp}); (2) performing a hierarchical clustering (HC) to identify patterns of species association; (3) creating an RF base-model using X_{envi} and the HC results (Y_{hc}) as explanatory and response variables, respectively; (4) creating an RF meta-model using the predictions from the RF base-model as response variable and the variables available for the 100,000 data-points as explanatory variables. The current workflow can be implemented using the iMESc application (Vieira and Fonseca, 2023).

The SOM model

1. Load the species (X_{sp}) dataset (this has to be previously defined, see challenge 1).
2. If desirable, remove the rare species (e.g., density < 1% and/or frequency < 1%).
3. If desirable, apply transformations to increase the weights (i.e., importance) of the species with low abundance (e.g., log (X+1), scale between 0 and 1 over the range of minimum and maximum abundance for each species).
4. Set the SOM parameters (challenge 5):
   a. Define the size and shape of the network (number of row and column nodes in the two-dimensional neuron map).
   b. Set the distance metric to calculate distance between node and data-points (e.g., Bray-Curtis).
   c. Choose the topology of the map (e.g., hexagonal).
   d. Choose between a planar or toroidal network.
   e. Set the maximal number of iterations (e.g., 1000).
   f. Set the learning mode (e.g., online).
   g. To allow for reproducible results, set a seed value.
5. Train the SOM model.
6. Evaluate the model metrics (e.g., topographic and quantization errors).
7. Explore the results (e.g., training progress, counting plot, U-matrix, BMU plot).
8. Evaluate the presence of outliers. If present, remove them and return to step 2 (challenge 2 and 3).

Hierarchical clustering

1. Retrieve the codebook from the SOM model (i.e., the final matrix of neuron weights) and use it as the input for the HC analysis (challenge 1).
2. Start the HC analysis by choosing a distance measure (e.g., Bray-Curtis). iMESc automatically uses the SOM training distance metric.
3. Choose a clustering method (e.g., Ward).
4. Explore the dendrogram of the codebook.
5. Define the optimal number of clusters using a clustering validation technique. For example, the Elbow method looks at the

total within-cluster sum of squares (WSS) as a function of the number of clusters. The location of a “knee” in the plot is usually considered as an indicator of the appropriate number of clusters.

6. Explore the obtained clusters in the SOM map.
7. The clustered SOM map is then used for classifying the observations.
8. Create a vector of the obtained clusters (Y_{hc}) for the observations.

Random forest base model

1. Load the X_{envi} dataset (this has to be previously defined, see challenge 1).
2. Consider using data imputation techniques (e.g., bag input, k-nearest neighbor method) to “fill in” any missing values.
3. To allow for reproducible results, set a seed value and randomly split the datasets into two subsets: for example, 80% of the data are used for training, and 20% of the data are used for testing (validation). Make sure that random sampling occurs within each class and preserves the overall class distribution of the data (balanced data partition).
4. Define the Y_{hc} vector as the response variable.
5. Search for hyperparameters of an RF classifier model (challenge 5):
   a. Set multiple number of trees (e.g., 250, 500).
   b. Set the search type and length for the *mtry* hyperparameter (e.g., random; 10, 20).
   c. Set the resampling method (e.g., 5-fold cross validation, repeated 10 times) for the internal validation.
   d. Train the RF model using the training dataset.
   e. Evaluate the internal model performances using metrics like Accuracy (MSE), and Kappa. Tuning parameters with the highest accuracy are automatically selected as optimal model.
6. Assess the reproducibility of the optimal model through internal and external validation (section 1).
7. Evaluate the Confusion Matrix of the optimal model (training and testing data).
8. Explore the importance of variables from the optimal model.
9. Evaluate the interaction effects of the most important variables.
10. Explore the biplots of the interaction effects on the response variable.

Random forest meta model

1. Restrict X-matrix used in the base-model to the variables available for the 100,000 data-points (i.e., latitude, longitude, and bathymetry).
2. Use the base model to make predictions on the whole dataset (training and testing data). Use its predictions as the response variable (Y).
3. Train the meta-model by repeating the steps 2 to 8 from the base-model above.
4. Use the meta-model to make predictions on the 100,000 data-points.
5. Spatialize the obtained predictions in a high-resolution geographical map.

Conclusion

This study explored the use of the random forest technique, self-organizing maps, and a hybrid approach between them to model and understand complex oceanographic processes. We explored the main challenges that permeated the Santos Project in order to implement such models and provided recommendations on how they should be handled. Additionally, two analytical workflows are given to guide future ecosystem baseline studies on modelling univariate and multivariate response variables. These workflows allow to explore and optimize model accuracy and, at the same time, to explore potential cause-effect relationships within the data. In addition, they will serve as base to implement long-term learning algorithms, which is an important increment for monitoring programs. These workflows, as well as all the analytical challenges discussed, can be easily implemented on iMESc, an open-source application.

Acknowledgments

We thank PETROBRAS for the access to data, information, and funding. Special acknowledgements are paid to the support of the Brazilian National Petroleum Agency - ANP (grant and research funds under resolution ANP PD&I) and to the Federal University of São Paulo. We thank the anonymous reviewers for their important contributions.

Author contributions

G.F.: Conceptualization and Writing – original draft; Writing – review & editing; Supervision; Resources; Project Administration; Funding Acquisition;

D.C.V.: Methodology; Software; Formal Analysis; Investigation; Writing – review & editing.

References

Aken, D. V., Pavlo, A., Gordon, G. J. & Zhang, B. 2017. Automatic Database Management System Tuning Through Large-scale Machine Learning. *In*: *Proceedings of the 2017 ACM International Conference on Management of Data* (pp. 1009–1024). ACM. DOI: https://doi.org/10.1145/3035918.3064029

Anguita, D., Ghio, A., Greco, N., Oneto, L. & Ridella, S. 2010. Model selection for support vector machines: Advantages and disadvantages of the Machine Learning Theory. *In*: *The 2010 International Joint Conference on Neural Networks (IJCNN)*. IEEE. DOI: https://doi.org/10.1109/ijcnn.2010.5596450

Ayodele, T. 2010. New Advances in Machine Learning. *In*: Zhang, Y. (ed.), *New Advances in Machine Learning* (Vol. 3, pp. 19–48). InTech. DOI: https://doi.org/10.5772/9385

Baker, R., Peña, J.-M., Jayamohan, J. & Jérusalem, A. 2018. Mechanistic models versus machine learning, a fight worth fighting for the biological community? *Biology Letters*, *14*(5), 20170660. DOI: https://doi.org/10.1098/rsbl.2017.0660

Bartlett, P., Freund, Y., Lee, W. & Schapire, R. 1998. Boosting the margin: a new explanation for the effectiveness of voting methods. *The Annals of Statistics*, *26*(5), 1651–1686. DOI: https://doi.org/10.1214/aos/1024691352

Bernard, S., Heutte, L. & Adam, S. 2009. Multiple Cassifier Systems. *In*: Benediktsson, J. A., Kittler, J., & Roli, F. (eds.), *Multiple Classifier Systems* (pp. 171–180). Berlin: Springer Berlin Heidelberg. DOI: https://doi.org/10.1007/978-3-642-02326-2_18

Bertolino, A., Guerriero, A., Miranda, B., Pietrantuono, R. & Russo, S. 2020. Learning-to-rank vs ranking-to-learn: Strategies for regression testing in continuous integration. *In*: *Proceedings - International Conference on Software Engineering* (pp. 1261–1272). Wahington, DC: ICSE. DOI: https://doi.org/10.1145/3377811.3380369

Biau, G. & Scornet, E. 2016. A random forest guided tour. *TEST*, *25*(2), 197–227. DOI: https://doi.org/10.1007/s11749-016-0481-7

Bilski, P. 2017. Unsupervised learning-based hierarchical diagnostics of analog circuits. *In*: *Workshop on Technical Diagnostics* (Vol. 119, pp. 99–104). Budapest: Elsevier BV. DOI: https://doi.org/10.1016/j.measurement.2018.01.029

Bonaccorso, G. 2017. *Machine learning algorithms*. Birmingham: Packt.

Borja, A., Elliott, M., Andersen, J., Berg, T., Carstensen, J., Halpern, B., Heiskanen, A.-S., Korpinen, S., Lowndes, J., Martin, G. & Rodriguez-Ezpeleta, N. 2016. Overview of Integrative Assessment of Marine Systems: The Ecosystem Approach in Practice. *Frontiers in Marine Science*, *3*(20). DOI: https://doi.org/10.3389/fmars.2016.00020

Breiman, L. 1996. Bagging predictors. *Machine Learning*, *24*(2), 123–140. DOI: https://doi.org/10.1007/bf00058655

Breiman, L. 2001. Random Forests. *Machine Learning*, *45*, 5–32. DOI: https://doi.org/10.1023/A:1010933404324

Butenschön, M., Clark, J., Aldridge, J., Allen, J., Artioli, Y., Blackford, J., Bruggeman, J., Cazenave, P., Ciavatta, S., Kay, S., Lessin, G., Van leeuwen, S., Van der molen, J., De Mora, L., Polimene, L., Sailley, S., Stephens, N. & Torres, R. 2016. ERSEM 15.06: a generic model for marine biogeochemistry and the ecosystem dynamics of the lower trophic levels. *Geoscientific Model Development*, *9*(4), 1293–1339. DOI: https://doi.org/10.5194/gmd-9-1293-2016

Carreira, R. S., Lazzari, L., Rozo, L. & Ceccopieri, M. 2023. Bulk and isotopic characterization of organic matter in cross-margin transect sediments in the Santos Basin, south-western Atlantic Ocean.

Chawla, N., Bowyer, K., Hall, L. & Kegelmeyer, W. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. DOI: https://doi.org/10.1613/jair.953

Chicco, D. & Jurman, G. 2020. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*(1), 6. DOI: https://doi.org/10.1186/s12864-019-6413-7

Chicco, D., Tötsch, N. & Jurman, G. 2021. The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData Mining*, *14*(1), 13. DOI: https://doi.org/10.1186/s13040-021-00244-z

Chon, T.-S. 2011. Self-Organizing Maps applied to ecological sciences. *Ecological Informatics*, *6*(1), 50–61. DOI: https://doi.org/10.1016/j.ecoinf.2010.11.002

Chou, J.-S., Tsai, C.-F. & Lu, Y.-H. 2013. Project dispute prediction by hybrid machine learning techniques. *Journal of Civil Engineering and Management*, *19*(4), 505–517. DOI: https://doi.org/10.3846/13923730.2013.768544

Clark, S., Sisson, Scott. & Sharma, A. 2020. Tools for enhancing the application of self-organizing maps in water resources research and engineering. *Advances in Water Resources*, *143*, 103676. DOI: https://doi.org/10.1016/j.advwatres.2020.103676

Cutler, D., Edwards, T., Beard, K., Cutler, A., Hess, K., Gibson, J. & Lawler, J. 2007. Random Forests for classification in ecology. *Ecology*, *88*(11), 2783–2792. DOI: https://doi.org/10.1890/07-0539.1

Dailianis, T., Smith, C., Papadopoulou, N., Gerovasileiou, V., Sevastou, K., Bekkby, T., Bilan, M., Billett, D., Boström, C., Carreiro-Silva, M., Danovaro, R., Fraschetti, S., Gagnon, K., Gambi, C., Grehan, A., Kipson, S., Kotta, J., Mcowen, C., Morato, T., Ojaveer, H., Pham, C. & Scrimgeour, R. 2018. Human activities and resultant pressures on key European marine habitats: An analysis of mapped resources. *Marine Policy*, *98*, 1–10. DOI: https://doi.org/10.1016/j.marpol.2018.08.038

Dalto, A. G., Moura, R. B., Sallorenzo, I. & Lavrado, H. P. 2023. Habitat quality assessment using the benthic macrofauna in Santos Basin continental shelf, SW Atlantic.

Ditria, E., Buelow, C., Gonzalez-Rivero, M. & Connolly, R. 2022. Artificial intelligence and automated monitoring for assisting conservation of marine ecosystems: A perspective. *Frontiers in Marine Science*, *9*, 918104. DOI: https://doi.org/10.3389/fmars.2022.918104

Effrosynidis, D. & Arampatzis, A. 2021. An evaluation of feature selection methods for environmental data. *Ecological Informatics*, *61*, 101224. DOI: https://doi.org/10.1016/j.ecoinf.2021.101224

Figueiredo Jr., A. G., Carneiro, J. C., Santos Filho, J. R., Cecilio, A. B., Rocha, G. J., Santos, S. T. V., Oliveira, A. S., Ferreira, F. & Luz, M. R. 2023. Sedimentary processes as a set-up conditions for living benthic communities in Santos Basin, Brazil.

Fox, E., Hill, R., Leibowitz, S., Olsen, A., Thornbrugh, D. & Weber, M. 2017. Assessing the accuracy and stability of variable selection methods for random forest modeling in ecology. *Environmental Monitoring and Assessment*, *189*(7), 316. DOI: https://doi.org/10.1007/s10661-017-6025-0

Franks, P. 2018. Global Ecology and Oceanography of Harmful Algal Blooms. *In*: Glibert, P. M., Berdalet, E., Burford, M. A., Pitcher, G. C., & Zhou, M. (eds.), *Ecological Studies* (Vol. 232, pp. 359–377). Springer International Publishing. DOI: https://doi.org/10.1007/978-3-319-70069-4_19

Freund, Y. & Schapire, R. E. 1996. Experiments with a new boosting algorithm. *In*: *Proceedings of the Thirteenth International Conference on International Conference on Machine Learning* (Vol. 13, pp. 148–156). San Francisco: Scientific Research Publishing, Inc. DOI: https://doi.org/10.4236/iim.2010.26047

Furian, N., O'sullivan, M., Walker, C., Vössner, S. & Neubacher, D. 2015. A conceptual modeling framework for discrete event simulation using hierarchical control structures. *Simulation Modelling Practice and Theory*, *56*, 82–96. DOI: https://doi.org/10.1016/j.simpat.2015.04.004

Gallucci, F., Corbisier, T. N., Gheller, P., Brito, S., Vieira, D. C. & Fonseca, G. 2023. Spatial distribution of meiofauna communities at the Santos Basin.

García, S., Luengo, J. & Herrera, F. 2015a. Dealing with missing values. *In*: *Data Preprocessing in Data Mining* (Vol. 72, pp. 59–105). Berlin: Springer International Publishing. DOI: https://doi.org/10.1007/978-3-319-10247-4_4

García, S., Luengo, J. & Herrera, F. 2015b. Dealing with noisy data. *In*: *Data Preprocessing in Data Mining* (Vol. 72, pp. 107–145). Berlin: Springer International Publishing. DOI: https://doi.org/10.1007/978-3-319-10247-4_5

Gardner, M. & Dorling, S. 2000. Statistical surface ozone models: an improved methodology to account for non-linear behaviour. *Atmospheric Environment*, *34*(1), 21–34. DOI: https://doi.org/org/10.1016/S1352-2310(99)00359-3

Gligorijević, V. & Pržulj, N. 2015. Methods for biological data integration: perspectives and challenges. *Journal of The Royal Society Interface*, *12*(112), 20150571. DOI: https://doi.org/10.1098/rsif.2015.0571

Goldstein, B., Polley, E. & Briggs, F. 2011. Random Forests for Genetic Association Studies. *Statistical Applications in Genetics and Molecular Biology*, *10*(1). DOI: https://doi.org/10.2202/1544-6115.1691

Grehan, A., Arnaud-Haond, S., D'onghia, G., Savini, A. & Yesson, C. 2017. Towards ecosystem based management and monitoring of the deep Mediterranean, North-East Atlantic and Beyond. *Deep Sea Research Part II: Topical Studies in Oceanography*, *145*, 1–7. DOI: https://doi.org/10.1016/j.dsr2.2017.09.014

Gupta, S. & Gupta, A. 2019. Dealing with Noise Problem in Machine Learning Data-sets: A Systematic Review. *Procedia Computer Science*, *161*, 466–474. DOI: https://doi.org/10.1016/j.procs.2019.11.146

Hastie, T., Tibshirani, R. & Friedman, J. 2009. *The Elements of Statistical Learning*. New York: Springer New York. DOI: https://doi.org/10.1007/978-0-387-84858-7

Hino, M., Benami, E. & Brooks, N. 2018. Machine learning for environmental monitoring. *Nature Sustainability*, *1*(10), 583–588. DOI: https://doi.org/10.1038/s41893-018-0142-9

Ho, S., Phua, K., Wong, L., Bin & Goh, W. 2020. Extensions of the External Validation for Checking Learned Model Interpretability and Generalizability. *Patterns*, *1*(8), 100129. DOI: https://doi.org/10.1016/j.patter.2020.100129

Jain, A. & Kumar, A. 2007. Hybrid neural network models for hydrologic time series forecasting. *Applied Soft Computing*, *7*(2), 585–592. DOI: https://doi.org/10.1016/j.asoc.2006.03.002

Jeni, L., Cohn, J. & De La Torre, F. 2013. Facing Imbalanced Data–Recommendations for the Use of Performance Metrics. *In*: *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction* (Vol. 61, pp. 245–251). Geneva: IEEE. DOI: https://doi.org/10.1109/acii.2013.47

Jiang, M. & Zhu, Z. 2022. The Role of Artificial Intelligence Algorithms in Marine Scientific Research. *Frontiers in Marine Science*, *9*, 1–4. DOI: https://doi.org/10.3389/fmars.2022.920994

Jordanov, I., Petrov, N. & Petrozziello, A. 2018. Classifiers Accuracy Improvement Based on Missing Data Imputation. *Journal of Artificial Intelligence and Soft Computing Research*, *8*(1), 31–48. DOI: https://doi.org/10.1515/jaiscr-2018-0002

Kangur, K., Park, Y.-S., Kangur, A., Kangur, P. & Lek, S. 2007. Patterning long-term changes of fish community in large shallow Lake Peipsi. *Ecological Modelling*, *203*(1–2), 34–44. DOI: https://doi.org/10.1016/j.ecolmodel.2006.03.039

Kaur, H., Pannu, H. & Malhi, A. 2020. A Systematic Review on Imbalanced Data Challenges in Machine Learning. *ACM Computing Surveys*, *52*(4), 1–36. DOI: https://doi.org/10.1145/3343440

Kohonen, T. 1990. The self-organizing map. *Proceedings of the IEEE*, *78*(9), 1464–1480. DOI: https://doi.org/10.1109/5.58325

Kohonen, T. 2001. *Self-Organizing Maps*. Springer: Berlin.

Krawczyk, B. 2016. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, *5*(4), 221–232. DOI: https://doi.org/10.1007/s13748-016-0094-0

Landis, J. & Koch, G. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, *33*(1), 159–174. DOI: https://doi.org/10.2307/2529310

Lawrence, R., Almasi, G. & Rushmeier, H. 1999. A scalable parallel algorithm for self-organizing maps with applications to sparse data mining problems. *Data Mining and Knowledge Discovery*, *3*(2), 171–195. DOI: https://doi.org/10.1023/A:1009817804059

Levy, O., Ball, B., Bond-Lamberty, B., Cheruvelil, K., Finley, A., Lottig, N., Punyasena, S., Xiao, J., Zhou, J., Buckley, L., Filstrup, C., Keitt, T., Kellner, J., Knapp, A., Richardson, A., Tcheng, D., Toomey, M., Vargas, R., Voordeckers, J., Wagner, T. & Williams, J. 2014. Approaches to advance scientific understanding of macrosystems ecology. *Frontiers in Ecology and the Environment*, *12*(1), 15–23. DOI: https://doi.org/10.1890/130019

L'Heureux, A., Grolinger, K., Elyamany, H. & Capretz, M. 2017. Machine Learning With Big Data: Challenges and Approaches. *IEEE Access*, *5*, 7776–7797. DOI: https://doi.org/10.1109/access.2017.2696365

Little, R. & Rubin, D. 2002. *Statistical Analysis with Missing Data*. Hoboken: John Wiley & Sons, Inc. DOI: https://doi.org/10.1002/9781119013563

Liu, Y., Weisberg, R. & Mooers, C. 2006. Performance evaluation of the self-organizing map for feature extraction. *Journal of Geophysical Research*, *111*(C5), C05018. DOI: https://doi.org/10.1029/2005jc003117

Lo, Z.-P. & Bavarian, B. 1991. On the rate of convergence in topology preserving neural networks. *Biological Cybernetics*, *65*(1), 55–63. DOI: https://doi.org/10.1007/bf00197290

Loureiro, A., Torgo, L. & Soares, C. 2004. Outlier Detection using Clustering Methods: a data cleaning application. *In*: *Proceedings of KDNet Symposium on Knowledge-based Systems for the Public Sector*. Sankt Augustin: KDnet.

Lunetta, K., Hayward, L., Segal, J. & Van Eerdewegh, P. 2004. Screening large-scale association study data: exploiting interactions using random forests. *BMC Genetics*, *5*(1), 32. DOI: https://doi.org/10.1186/1471-2156-5-32

Lynam, C., Uusitalo, L., Patrício, J., Piroddi, C., Queirós, A., Teixeira, H., Rossberg, A., Sagarminaga, Y., Hyder, K., Niquil, N., Möllmann, C., Wilson, C., Chust, G., Galparsoro, I., Forster, R., Veríssimo, H., Tedesco, L., Revilla, M. & Neville, S. 2016. Uses of Innovative Modeling Tools within the Implementation of the Marine Strategy Framework Directive. *Frontiers in Marine Science*, *3*, 1–18. DOI: https://doi.org/10.3389/fmars.2016.00182

Ma, E.-Y., Kim, J.-W., Lee, Y., Cho, S.-W., Kim, H. & Kim, J. 2021. Combined unsupervised-supervised machine learning for phenotyping complex diseases with its application to obstructive sleep apnea. *Scientific Reports*, *11*(1), 4457. DOI: https://doi.org/10.1038/s41598-021-84003-4

Mahesh, B. 2020. Machine Learning Algorithms - A Review. *International Journal of Science and Research*, *9*(1), 381–386. DOI: https://doi.org/10.21275/ART20203995

Markham, I., Mathieu, R. & Wray, B. 2000. *Kanban* setting through artificial intelligence: a comparative study of artificial neural networks and decision trees. *Integrated Manufacturing Systems*, *11*(4), 239–246. DOI: https://doi.org/10.1108/09576060010326230

Michener, W. & Jones, M. 2012. Ecoinformatics: supporting ecology as a data-intensive science. *Trends in Ecology & Evolution*, *27*(2), 85–93. DOI: https://doi.org/10.1016/j.tree.2011.11.016

Moreira, D. L., Marcon, E. H., Toledo, R. G. A. & Bonecker, A. C. T. 2023. Multidisciplinary Scientific Cruises for Environmental Characterization in the Santos Basin – Methods and Sampling Design. DOI: https://doi.org/10.5281/ZENODO.7702291

Mount, N. J. & Weaver, D. 2011. Self-organizing maps and boundary effects: quantifying the benefits of torus wrapping for mapping SOM trajectories. *Pattern Analysis and Applications*, *14*(2), 139–148. DOI: https://doi.org/10.1007/s10044-011-0210-5

Muñoz, A. & Muruzábal, J. 1998. Self-organizing maps for outlier detection. *Neurocomputing*, *18*(1–3), 33–60. DOI: https://doi.org/10.1016/s0925-2312(97)00068-4

Natita, W., Wiboonsak, and W. & Dusadee, S. 2016. Appropriate Learning Rate and Neighborhood Function of Self-organizing Map (SOM) for Specific Humidity Pattern Classification over Southern Thailand. *International Journal of Modeling and Optimization*, *6*(1), 61–65. DOI: https://doi.org/10.7763/ijmo.2016.v6.504

Newman, E. A. 2019. Disturbance Ecology in the Anthropocene. *Frontiers in Ecology and Evolution*, *7*. DOI: https://doi.org/10.3389/fevo.2019.00147

Ng, S. & Chan, and M. 2019. Effect of Neighbourhood Size Selection in SOM-Based Image Feature Extraction. *International Journal of Machine Learning and Computing*, *9*(2), 195–200. DOI: https://doi.org/10.18178/ijmlc.2019.9.2.786

Nichols, J. D. & Williams, B. K. 2006. Monitoring for conservation. *Trends in Ecology & Evolution*, *21*(12), 668–673. DOI: https://doi.org/10.1016/j.tree.2006.08.007

Oshiro, T. M., Perez, P. S. & Baranauskas, J. A. 2012. How Many Trees in a Random Forest? *In*: Perner, P. (ed.), *Machine Learning and Data Mining in Pattern Recognition* (Vol. 7376, pp. 154–168). New York: Springer. DOI: https://doi.org/10.1007/978-3-642-31537-4_13

Park, Y.-S., Chung, Y.-J. & Moon, Y.-S. 2013. Hazard ratings of pine forests to a pine wilt disease at two spatial scales (individual trees and stands) using self-organizing map and random forest. *Ecological Informatics*, *13*, 40–46. DOI: https://doi.org/10.1016/j.ecoinf.2012.10.008

Park, Y.-S., Song, M.-Y., Park, Y.-C., Oh, K.-H., Cho, E. & Chon, T.-S. 2007. Community patterns of benthic macroinvertebrates collected on the national scale in Korea. *Ecological Modelling*, *203*(1–2), 26–33. DOI: https://doi.org/10.1016/j.ecolmodel.2006.04.032

Penczak, T., Kruk, A., Park, Y. S. & Lek, S. 2005. Modelling Community Structure in Freshwater Ecosystems. *In*: Lek, Sovan, Scardi, M., Verdonschot, P. F. M., Descy, J.-P., & Park, Y.-S. (eds.), *Modelling Community Structure in Freshwater Ecosystems* (pp. 100–113). Springer-Verlag. DOI: https://doi.org/10.1007/3-540-26894-4_10

Perkel, J. 2019. Workflow systems turn raw data into scientific knowledge. *Nature*, *573*(7772), 149–150. DOI: https://doi.org/10.1038/d41586-019-02619-z

Platias, C. & Petasis, G. 2020. A Comparison of Machine Learning Methods for Data Imputation. *In*: *11th Hellenic*

Conference on Artificial Intelligence (pp. 150–159). New York: ACM. DOI: https://doi.org/10.1145/3411408.3411465

Pope, D. & McNeill, F. 2013. *From Big Data to Meaningful Information*. Cary: SAS.

Poulos, J. & Valle, R. 2018. Missing Data Imputation for Supervised Learning. *Applied Artificial Intelligence*, *32*(2), 186–196. DOI: https://doi.org/10.1080/08839514.2018.1448143

Probst, P., Bischl, B. & Boulesteix, A.-L. 2019. Tunability: Importance of Hyperparameters of Machine Learning Algorithms. *The Journal of Machine Learning Research*, *20*(1), 1934–1965. DOI: https://doi.org/10.5555/3322706.3361994

Probst, P. & Boulesteix, A.-L. 2017. To tune or not to tune the number of trees in random forest. *The Journal of Machine Learning Research*, *18*(1), 1934–1965. DOI: https://doi.org/10.48550/ARXIV.1705.05654

Probst, P., Wright, M. & Boulesteix, A. 2019. Hyperparameters and tuning strategies for random forest. *WIREs Data Mining and Knowledge Discovery*, *9*(3), e1301. DOI: https://doi.org/10.1002/widm.1301

Rahmati, O., Falah, F., Naghibi, S., Biggs, T., Soltani, M., Deo, R., Cerdà, A., Mohammadi, F. & Tien Bui, D. 2019. Land subsidence modelling using tree-based machine learning algorithms. *Science of The Total Environment*, *672*, 239–252. DOI: https://doi.org/10.1016/j.scitotenv.2019.03.496

Razi, M. & Athappilly, K. 2005. A comparative predictive analysis of neural networks (NNs), nonlinear regression and classification and regression tree (CART) models. *Expert Systems with Applications*, *29*(1), 65–74. DOI: https://doi.org/10.1016/j.eswa.2005.01.006

Refaeilzadeh, P., Tang, L. & Liu, H. 2009. Encyclopedia of Database Systems. *In*: Liu, L. & Özsu, M. T. (eds.), *Encyclopedia of Database Systems* (pp. 532–538). Boston: Springer US. DOI: https://doi.org/10.1007/978-0-387-39940-9_565

Rhodes, J. & Jonzén, N. 2011. Monitoring temporal trends in spatially structured populations: how should sampling effort be allocated between space and time? *Ecography*, *34*(6), 1040–1048. DOI: https://doi.org/10.1111/j.1600-0587.2011.06370.x

Robinson, S. 2008. Conceptual modelling for simulation Part II: a framework for conceptual modelling. *Journal of the Operational Research Society*, *59*(3), 291–304. DOI: https://doi.org/10.1057/palgrave.jors.2602369

Rollinson, C., Finley, A., Alexander, M., Banerjee, S., Dixon Hamil, K.-A., Koenig, L., Locke, D., Demarche, M., Tingley, M., Wheeler, K., Youngflesh, C. & Zipkin, E. 2021. Working across space and time: nonstationarity in ecological research and application. *Frontiers in Ecology and the Environment*, *19*(1), 66–72. DOI: https://doi.org/10.1002/fee.2298

Sarker, I. H. 2021. Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, *2*(3), 160. DOI: https://doi.org/10.1007/s42979-021-00592-x

Schaub, M. & Abadi, F. 2011. Integrated population models: a novel analysis framework for deeper insights into population dynamics. *Journal of Ornithology*, *152*(S1), 227–237. DOI: https://doi.org/10.1007/s10336-010-0632-7

Stefanovič, P. & Kurasova, O. 2011. Influence of Learning Rates and Neighboring Functions on Self-Organizing Maps. *In*: *WSOM 2011: Advances in Self-Organizing Maps* (Vol. 6731, pp. 141–150). Berlin: Springer Berlin Heidelberg. DOI: https://doi.org/10.1007/978-3-642-21566-7_14

Stoudt, S., Vásquez, V. & Martinez, C. 2021. Principles for data analysis workflows. *PLOS Computational Biology*, *17*(3), e1008770. DOI: https://doi.org/10.1371/journal.pcbi.1008770

Stupariu, M.-S., Cushman, S., Pleşoianu, A.-I., Pătru-Stupariu, I. & Fürst, C. 2021. Machine learning in landscape ecological analysis: a review of recent approaches. *Landscape Ecology*, *37*(5), 1227–1250. DOI: https://doi.org/10.1007/s10980-021-01366-9

Tison, J. 2004. Use of unsupervised neural networks for ecoregional zoning of hydrosystems through diatom communities: case study of Adour-Garonne watershed (France). *Archiv Für Hydrobiologie*, *159*(3), 409–422. DOI: https://doi.org/10.1127/0003-9136/2004/0159-0409

Tsai, C.-F. & Chen, M.-L. 2010. Credit rating by hybrid machine learning techniques. *Applied Soft Computing*, *10*(2), 374–380. DOI: https://doi.org/10.1016/j.asoc.2009.08.003

Ultsch, A. 2003. U*-matrix: a tool to visualize clusters in high dimensional data.

Van Hulle, M. 2012. Handbook of Natural Computing. *In*: Rozenberg, G., Bäck, T., & Kok, J. N. (eds.), *Handbook of Natural Computing* (pp. 585–622). Berlin: Springer Berlin Heidelberg. DOI: https://doi.org/10.1007/978-3-540-92910-9_19

Vesanto, J. & Alhoniemi, E. 2000. Clustering of the self-organizing map. *IEEE Transactions on Neural Networks*, *11*(3), 586–600. DOI: https://doi.org/10.1109/72.846731

Vesanto, J., Himberg, J., Alhoniemi, E. & Parhankangas, J. 1999. Self-organizing map in Matlab: the SOM Toolbox. *In*: *Proceedings of the Matlab DSP Conference* (pp. 35–40). Espoo.

Vesanto, J., Himberg, J., Alhoniemi, E. & Parhankangas, J. 2000. SOM Toolbox for Matlab 5.

Vieira, D. & Fonseca, G. 2023. iMESc: An Interactive Machine Learning App for Environmental Science. DOI: https://doi.org/10.5281/zenodo.6484391

Virts, K., Shirey, A., Priftis, G., Ankur, K., Ramasubramanian, M., Muhammad, H., Acharya, A. & Ramachandran, R. 2020. A Quantitative Analysis on the Use of Supervised Machine Learning in Earth Science. *In*: *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium* (pp. 2252–2255). Waikoloa: IEEE. DOI: https://doi.org/10.1109/igarss39084.2020.9323770

Walker, G. 2006. The tipping point of the iceberg. *Nature*, *441*(7095), 802–805. DOI: https://doi.org/10.1038/441802a

Wang, F., Shi, Z., Biswas, A., Yang, S. & Ding, J. 2020. Multi-algorithm comparison for predicting soil salinity. *Geoderma*, *365*, 114211. DOI: https://doi.org/10.1016/j.geoderma.2020.114211

Webb, J., Arthington, A. & Olden, J. 2017. Models of Ecological Responses to Flow Regime Change to Inform Environmental Flows Assessments. Water for the Environment: From Policy and Science to Implementation and Management. *Water for the Environment*, 287–316. DOI: https://doi.org/10.1016/B978-0-12-803907-6.00014-0

Wehrens, R. & Buydens, L. 2007. Self- and Super-organizing Maps in *R*: The Kohonen Package. *Journal of Statistical Software*, *21*(5), 1–19. DOI: https://doi.org/10.18637/jss.v021.i05

Wehrens, R. & Kruisselbrink, J. 2018. Flexible Self-Organizing Maps in kohonen 3.0. *Journal of Statistical Software*, *87*(7), 1–18. DOI: https://doi.org/10.18637/jss.v087.i07

Yang, P., Wang, D., Wei, Z., Du, X. & Li, T. 2019. An Outlier Detection Approach Based on Improved Self-Organizing Feature Map Clustering Algorithm. *IEEE Access*, *7*, 115914–115925. DOI: https://doi.org/10.1109/access.2019.2922004

Yotova, G., Varbanov, M., Tcherkezova, E. & Tsakovski, S. 2021. Water quality assessment of a river catchment by the composite water quality index and self-organizing maps. *Ecological Indicators*, *120*, 106872. DOI: https://doi.org/10.1016/j.ecolind.2020.106872

Zhang, J.-T., Dong, Y. & Xi, Y. 2008. A comparison of SOFM ordination with DCA and PCA in gradient analysis of plant communities in the midst of Taihang Mountains, China. *Ecological Informatics*, *3*(6), 367–374. DOI: https://doi.org/10.1016/j.ecoinf.2008.09.004

Zhang, L., Scholz, M., Mustafa, A. & Harrington, R. 2008. Assessment of the nutrient removal performance in integrated constructed wetlands with the self-organizing map. *Water Research*, *42*(13), 3519–3527. DOI: https://doi.org/10.1016/j.watres.2008.04.027

Zhong, S., Zhang, K., Bagheri, M., Burken, J., Gu, A., Li, B., Ma, X., Marrone, B., Ren, Z., Schrier, J., Shi, W., Tan, H., Wang, T., Wang, X., Wong, B., Xiao, X., Yu, X., Zhu, J.-J. & Zhang, H. 2021. Machine Learning: New Ideas and Tools in Environmental Science and Engineering. *Environmental Science & Technology*, *55*(19), 12741–12754. DOI: https://doi.org/10.1021/acs.est.1c01339

Zipkin, E., Zylstra, E., Wright, A., Saunders, S., Finley, A., Dietze, M., Itter, M. & Tingley, M. 2021. Addressing data integration challenges to link ecological processes across scales. *Frontiers in Ecology and the Environment*, *19*(1), 30–38. DOI: https://doi.org/10.1002/fee.2290