Features impacting the mesopelagic layer in the ocean: a machine learning-based approach

Marit Nodland Lund

Master's thesis in Software Engineering at

Department of Computer science, Electrical engineering and Mathematical sciences,
Western Norway University of Applied Sciences

Department of Informatics,
University of Bergen

September 2022

Western Norway
University of
Applied Sciences

Abstract

Context: Recently the United Nations proclaimed a Decade of Ocean Science for Sustainable Development (2021–2030) due to threats to the productivity and health of the ocean due to human impact. The One Ocean Expedition (OOE), a circumnavigation of the world by the Norwegian tall ship Statsraad Lehmkuhl, is part of the Ocean Decade, intending to create attention and share knowledge around the crucial role of the ocean for sustainable development. There is little knowledge about the deep sea. Still, the amount and type of organisms here could be an essential factor in predicting global carbon dynamics and the effects of climate change, as well as food safety in the coming years. In marine science, the process of turning data into knowledge has long been manual or semi-manual, and automated processes are necessary for scaling up monitoring programs and making use of the extensive amount of data collected.

Research goal: In this thesis, the aim is to investigate the use of machine learning in predicting marine biomass and discovering possible correlations between biogeochemical or physical factors and the biomass of the mesopelagic zone (200-1,000 m depth) using data collected during the OOE. This will eventually lead to more knowledge about the workings of the organisms in the mesopelagic layer and the use of machine learning in marine science.

Methodology: The methods used in this thesis follow the paradigm of Design Science, where the aim is to answer questions relevant to human problems via the creation of artifacts, thereby contributing new knowledge to both the fields of marine science and data science.

Results: The findings show that it is possible to predict mesopelagic biomass with reasonable accuracy using tree-based algorithms such as random forest, which may be further enhanced using historical data. Different correlations between biomass and various biogeochemical or physical factors based on geographical area are discovered using feature importance calculated using random forests.

Acknowledgements

First, I would like to thank my supervisor Professor Rogardt Heldal for the opportunity of taking my master thesis in such an interesting field, and being a part of the One Ocean Expedition. I would also like to include Ngoc-Thanh Nguyen for his help, good ideas, and who has always been quick at answering all my questions. Thank you both for you for providing great feedback, and for spending your time on me and this project. Further, I would like to thank Professor Geir Pedersen for his expertise and help with the acoustic aspect of this study. Last but not least, I would like to thank my friends and family for their support, and without whom none of this would be possible.

Contents

List of Figures

List of Tables

Acronyms

ADCP Acoustic Doppler Current Profiler.

AI Artificial Intelligence.

API Application Programming Interface.

CDOM Chromophoric (Colored) Dissolved Organic Matter.

CHLAF Chlorophyll-A Fluorescence.

CNN Convolutional Neural Networks.

CRIMAC Center for Research Based Innovation in Marine Acoustic Abundance Estimation and Backscatter Classification.

DIC Dissolved inorganic carbon.

DOC Dissolved organic carbon.

DSL Deep Scattering Layer.

DVM Diel Vertical Migration.

FTP File Transfer Protocol.

LSSS Large Scale Survey System.

MAE Mean Absolute Error.

MAPE Mean Absolute Percentage Error.

MISQ Management Information Systems Quarterly.

MSE Mean Squared Error.

NASC Nautical Area Scattering Coefficient.

NMDC Norwegian Marine Data Center.

NORCE Norwegian Research Centre.

OBIS Ocean Biogeographic Information System.

OLS Ordinary Least Squares.

OOE One Ocean Expedition.

PCA Principal Component Analysis.

PCR Principal Component Regression.

PLSR Partial Least Squares Regression.

RMSE Root Mean Squared Error.

SL Statsraad Lehmkuhl.

SMAPE Symmetric Mean Absolute Error.

UN United Nations.

WOA World Ocean Atlas.

Chapter 1

Introduction

The ocean is vital for human beings. It is critical for food supply, climate regulation, transportation, and energy production [1]. The rapid increase in the world population and industrial development is threatening the productivity and health of the ocean due to pollution and overuse [2]. Recent pressure ultimately led to the United Nations proclaiming a Decade of Ocean Science for Sustainable Development (2021–2030), mobilizing the effort around research and technological innovation within marine science.

In this thesis, the concern is the global mesopelagic biomass, a critical component of biogeochemical research on the ocean and possibly future food supply. The mesopelagic zone extends from 200 to 1,000 meters below the ocean's surface and contains a large, mostly untouched biomass that is known very little about today. Traditional approaches are mainly visual inspections and statistical methods, which can be both time and labor-intensive. Manual methods are not suited for large amounts of data, as it might be harder to analyze all the collected samples [3]. Statistical methods also generally require a good understanding of the data, which is not always the case. Machine learning, however, can learn patterns from data with significantly less human effort and with less initial understanding of the data [4].

The marine science community is eager to apply machine learning tools to a wide range of tasks relevant to the sustainability of living ocean resources. Emerging technologies have dramatically increased data volume, exceeding manual processing capacity [3]. The application of machine learning can significantly reduce resources such as time and cost for processing data.

Previously machine learning has been applied in marine science to improve the processing and analysis of various data types collected from aerial and underwater surveys and ocean observation operations [3], [5]. The main focus has often been the use of deep learning and convolutional neural networks (CNN) for detecting and classifying imagery and acoustic data. Image processing on data from trawling and fisheries for species classification and population control is an example of a current area of interest [6].

Developing a novel automated method to predict mesopelagic biomass and the number of organisms performing diel vertical migration (DVM) in the mesopelagic could therefore be an important factor in predicting global carbon dynamics and the effects of climate change, food safety, and to bio-prospect pharmaceuticals. DVM is the synchronized movement of marine animals between the surface and deep layers in the ocean. There is a possibility to re-purpose the trained models developed as part of this thesis for new tasks, which could accelerate development

in marine technology for similar problems. In this thesis one of the goals is to move from a local to a global perspective on mesopelagic biomass.

The data used in this thesis was collected during the One Ocean Expedition (2021-2023) [7] with instruments and sensors mounted on the Norwegian tall ship Statsraad Lehmkuhl. The One Ocean Expedition (OOE) is a circumnavigation of the globe, aiming to create attention and share knowledge regarding the ocean's important role in sustainable development. The OOE is further explained in Chapter 3. Data is sent directly to the Norwegian Marine Data Center (NMDC) [8], an integrated system for the exchange of marine data between institutions in Norway, via satellite and is accessed via their open underway API.

An expedition similar to OOE was performed in 2010, called the Malaspina Expedition, named after the original scientific Malaspina Expedition that occurred between 1789 and 1794 [9]. Similar to OOE, the goal of this expedition was to assess the impact of global change on the oceans. OOE and Malaspina had different routes and durations, OOE being the longer journey, as well as practical differences in methods and equipment. Several papers have been published analyzing data collected during the Malaspina expedition. However, at the time of this thesis and to our knowledge, none of them utilized machine learning, as it was still fairly new at the time. Further information about the expedition and related research can be found in Chapter 2.

Even though the world has been aware of changes and threats related to the ocean and the need for more knowledge regarding marine environments for over a decade, utilization of machine learning in mesopelagic faunal abundance and biomass research is currently very limited. Machine learning can deal with the issue of minimal knowledge about the data by identifying patterns in large amounts of data, which traditional methods typically struggle with [10], [11]. Measurements were taken from large areas of the globe, compared to previous often very geographically limited marine research. The topic of this thesis is processing and analysing acoustic data in combination with atmospheric and oceanic properties, focusing on the ocean's mesopelagic layer.

This thesis proposes a machine learning model which purpose is to predict mesopelagic biomass. Data collected during OOE is used to train the models. The goal is that the model can also be used with similar data from other sources and possibly provide a basis for related problems within marine science. The following research questions were identified.

Research question 1: How can possible correlations between the collected physical and geochemical data, and the acoustic biomass in the ocean, be detected using machine learning?

Research question 2: How accurately can the acoustic biomass of organisms in the mesopelagic using oceanic and atmospheric data be predicted with machine learning?

Research question 3: How accurate predictions can be made about the acoustic biomass of organisms that perform the diel vertical migration (DVM) using oceanic and atmospheric data with machine learning?

Outline: The thesis starts with an overview and the motivation for the problem. Chapter 2 gives an overview of related work within both machine learning and biology-centered fields. In Chapter 3, background information is provided regarding the One Ocean Expedition, the biology of the mesopelagic zone, and key concepts within machine learning that are beneficial for understanding the implementation of the problem. The research methodology chosen for the project is presented in Chapter 4, along with the approach taken to predictive supervised machine learning. The pre-processing of the data is explained in Chapter 5, and this includes data cleaning, creation of different data sets, and other steps taken to prepare the data as input for the chosen machine learning algorithms. The results of the experiments follow in Chapter 6,

along with an interpretation of the results. In Chapter 7, threats to validity are discussed, while in Chapter 8, a discussion of the results and the project in general are presented. Finally, the thesis ends with a conclusion and suggestions for future work in Chapter 9.

Chapter 2

Related work

This chapter considers related work on machine learning, marine science, and the mesopelagic. The literature review has been split into four main categories; research related to the Malaspina Expedition 2010, biological and statistical research of the mesopelagic, the use of machine learning in mesopelagic research, and, finally, machine learning methods for estimating biomass.

2.1 Malaspina Expedition 2010

The Malaspina Expedition 2010 [9] was, similar to the One Ocean Expedition (OOE), an interdisciplinary research project to assess the impact of global change on the oceans. The expeditions had different routes and duration, but both were circumnavigations with research objectives. Hespérides, the ship performing the circumnavigation, was a mechanical research vessel in contrast to the three-masted bark Statsraad Lehmkuhl (SL) from the OOE. Of the several papers published from the Malaspina expedition, two analyze echo sounding data from the mesopelagic zone [12], [13]. The Hespérides had the EK60 echo sounding system installed, the predecessor of the EK80 system utilized by Statsraad Lehmkuhl.

Echo sounder data collected during the Malaspina expedition have resulted in several publications about mesopelagic biomass and behavior [12], [14], [15]. These suggest that the depth of the deep scattering layer (DSL), the acoustic signature of the organisms in the mesopelagic, and the behavior of the organisms performing the diel vertical migration (DVM) can be explained by horizontal patterns in physical-chemical properties of water masses, such as oxygen, temperature, and turbidity [12], [16]. Other publications based on the data from Malaspina implied relations to primary production [13], [17], and the moon phase [15]. Primary production means the creation of new organic matter from inorganic substrates via processes such as photosynthesis. Figure 2.1 shows an example of echograms at 38 kHz, indicating differences in the behavior of the mesopelagic between geographical regions.

The results of these publications are important for the thesis as they indicate possible correlated features. However, none of the studies utilized machine learning or analyzed more than a couple of variables at once, which is the approach chosen for this project.

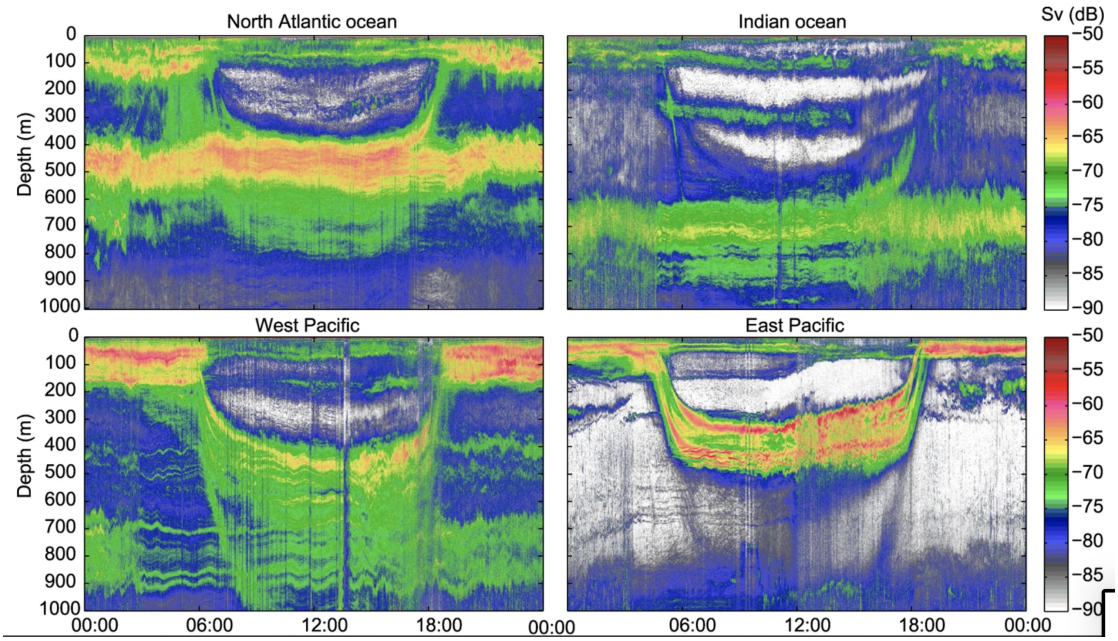


Figure 2.1: Examples of echograms at 38 kHz from the Malaspina expedition, spanning 24 hour periods, from different geographic regions. Picture from [12].

2.2 Research on the mesopelagic

This section will focus on the functions of the mesopelagic and DVM to emphasize the need for further research in this area and the limitations of existing research.

Except for publications related to Malaspina, mesopelagic biomass and DVM research has often been contained to relatively small geographic areas [18], [19], including marine science research in general. Global biomass estimates have been based on trawling, which are prone to severe underestimation [18], [20], [21], and acoustic estimates will usually exceed those from trawls. The underestimation is partly due to the fact that mesopelagic species have been shown to exhibit escape reactions to trawling[21]. Acoustic estimates also have challenges due to swim-bladder volume and target strength, as dominant species in the investigated areas may be weak scatterers [18], [22].

The fish in the mesopelagic is considered unpalatable to humans but could be an important resource for processing into fishmeal and nutritional supplements [23], [24]. Exploitation of these resources could however have global ramifications, as the DVM and sheer volume of the biomass is thought to be a critical component of the global carbon cycle and marine food webs [23], [25]–[27]. The lack of knowledge is thus important to remedy, with further research possibly being vital for events such as global warming. All of the above motivate further research of the mesopelagic. All publications cited in this section consist of manual research. In the following section, some applications of machine learning for understanding the mesopelagic will be presented. However, these publications have not focused on biomass estimation as it is in this thesis.

2.3 Machine learning and the mesopelagic

Application of machine learning in relation to the mesopelagic has often focused on target classification either from trawl mounted cameras [6] or broadband and multi-frequency echo sounder data using clustering techniques or deep learning [28]–[30].

At the time of this thesis, only one exploratory study done on the problem of mesopelagic biomass estimation using machine learning was found during literature research. In the publication by Gong and Hudson [31], they used supervised learning algorithms to model the observed mesopelagic abundance using physical and biogeochemical data. The data for mesopelagic biomass was collected from Ocean Biogeographic Information System (OBIS), and the physical and biogeochemical properties from World Ocean Atlas (WOA). The publication is a one-page poster describing few details about methods and the data used [1]. As far as can be determined from the information in the poster, the mesopelagic fauna data is based on catch data. It also seems like the biomass data and the other properties are from different years, likely 2017 and 2018 respectively. Thus the measurements have probably not been made at the same time, nor necessarily in the same place. As mentioned in the previous section, mesopelagic catch data have its disadvantages due to differences in fishing effort and trawl avoidance. The data used for the problems in this thesis was collected using the same method throughout the entire journey, and the biomass estimation was done using acoustics. Results are validated using several metrics, compared to only root mean squared error (RMSE).

2.4 Biomass estimation using machine learning

Most research within biomass estimation using machine learning is geared towards above-ground biomass such as forests and grassland using remote sensing data [32], [33]. These mainly use tree-based algorithms [34], [35].

Less progress has been made within marine biomass estimation. Research often focuses on certain species, such as mussels [36] and TUN-AI, a model developed to estimate tuna biomass [37], while others pertain to seabed coverage (corals, seagrass, algae) [38]. Tree-based algorithms are also common in these areas, along with deep learning and neural networks. Most publications use image data, with some select ones using echo sounding data [37]. In addition, several are limited to relatively small geographical areas as is common in marine science research, mentioned previously in this section. This thesis focuses on the total biomass, rather than the abundance of certain species, from a global perspective using echo sounder data.

[1] An effort was made to get in touch with the authors, but no response was received.

Chapter 3

Background

This chapter will present background information that is useful for understanding the work in this thesis. First, an overview of the expedition during which the data was collected and the sensors installed on the ship, thereafter an introduction to the biology and importance of the mesopelagic. The chapter ends with some key concepts in machine learning, with focus on evaluation metrics and algorithms relevant to the problems identified in Chapter 1.

3.1 One Ocean Expedition

In August 2021, the 98-meter-long tall ship Statsraad Lehmkuhl (SL) set sail from Arendal, Norway, for the twenty-month-long One Ocean Expedition (OOE), a circumnavigation of the globe. The goal is to create attention and share knowledge about the crucial role of the ocean for sustainable development in a global perspective. The expedition is a recognized part of the UN Decade of Ocean Science for Sustainable Development. Regional seas are different, but common challenges affect all parts of the ocean. Statsraad Lehmkuhl was equipped with modern instrumentation for collecting comparable data continuously through the journey. The quiet movement of a sailing vessel provides good conditions for collecting acoustic data from the ocean. In port, the ship is used for conferences, diplomacy, high-level meetings, and corporate hospitality.

3.2 Sensors and methods

Several systems for measurements were installed on board the Statsraad Lehmkuhl before departing for the circumnavigation. In total these systems measured 36 different variables in the ocean and atmosphere. Sensor data was sent to shore continuously over satellite through Kongsberg Maritime's Blue Insight solution and stored in the Norwegian Marine Data Center (NMDC). Raw data from the echo sounder was stored on hard drives and shipped to Bergen at regular intervals, usually when completing a leg of the expedition.

An echo sounder transmits an acoustic wave into the water. Parts of the emitted wave will be reflected by marine organisms, the sea floor, or other objects. The returning signals are called backscatter. The echo sounder can calculate the depth of the object or organism based on the time interval between emission and return of a sound pulse. The strength or "intensity" of the

backscatter indicates how hard the reflective object is. Echo sounders can be used for various tasks, such as navigation, depth measurement, and detecting fish or plankton [39].

Simrad EK80

The echo sounder installed at Statsraad Lehmkuhl is the Simrad EK80, a high-precision scientific echo sounder and Acoustic Doppler Current Profiler (ADCP) system developed by Kongsberg Maritime. ADCP systems are used to measure how fast water is moving across the entire water column using a principle of sound waves called the Doppler effect, and are primarily used in oceanography. The Doppler effect is the change in frequency of a sound wave for an observer moving relative to the source of the wave [40]. It is commonly heard when a vehicle sounding a siren or horn approaches, passes, and recedes from an observer. The received frequency is higher, compared to the emitted frequency, during the approach, identical at the instant of passing by, and lower as it moves away [40]. Continuous acoustic measurements were made with the calibrated Simrad EK80 echosounder, approximately 18° beam-width, operating at frequencies 38 kHz and 200 kHz.

FerryBox

A ferrybox is a flow-through system installed on board a ship that continuously measures physical, chemical, and biological parameters. The sensors utilized on board SL offers the opportunity to measure salinity, temperature, turbidity, chlorophyll-a fluorescence (CHLAF), chromophoric (colored) dissolved organic matter (CDOM), and oxygen saturation [41].

CDOM is an indication of light availability for primary productivity as well as a proxy for dissolved organic carbon (DOC), which can give indications about the global carbon cycle. In this case, the primary productivity might impact the biomass, which makes the CDOM sensor relevant for training the model [42].

Chlorophyll-a fluorescence (CHLAF) is a global phenomenon where chlorophyll-a molecules absorb light at one wavelength and re-emit it at different wavelengths. CHLAF of photosynthetic organisms varies as a result of changes in the biomass and is an indicator of potential primary productivity [43], [44].

pCO2 system

The pCO2-system is another flow-through system installed on board Statsraad Lehmkuhl which measures the partial CO_2 content of the ocean and atmosphere. Sea water is pumped through a closed equilibrator chamber at a rate of 2-3 liters of water per minute. The water enters the equilibrator through a spiral-shaped nozzle which makes for a conical water spray that ensures there is an equilibrium between water and air due to the increased surface area [45]. This measurement relates to the carbon cycle and the acidification of the oceans.

Other sensors

A weather station that measured air temperature, humidity, and air pressure among others was installed at the top of the main mast. An overview of the different instruments and their placements on the ship can be seen in Figure 3.1.

Figure 3.1: Overview of the scientific instruments on Statsraad Lehmkuhl.

3.3 The mesopelagic zone

The mesopelagic zone is a layer in the ocean where less than 1% of incident light reaches, usually considered to be between 200-1000 m depth [26]. This part of the ocean contains the deep-scattering layer (DSL). Initially mistaken as the ocean floor, the DSL is a horizontal zone of living organisms that ascend towards the surface in the evening and descend again at dawn. Mesopelagic fish resources are classified as one of the largest fish resources globally and are abundant in the North Atlantic. They represent a significant potential developing of new fisheries and as a source for the biomarine industry in Norway [46]. The UN's Food and Agriculture Organisation reported in 2002 that the fish-meal and fish-oil industries would need to exploit this part of the ocean in order to feed fish farms [47].

In the mesopelagic zone, also called the twilight, the light intensity is insufficient for photosynthesis. Respiration dominates as oceanic animals use this zone for feeding or avoiding predators, resulting in low oxygen concentration [26]. The mesopelagic is significant in terms of ocean volume and estimated biomass, and has inhabitants such as glacier lantern fish (Benthosema glaciale) and Mueller's pearlside (Maurolicus muelleri).

The mesopelagic biomass has usually been estimated to be approximately 1,000 million tons. However, a recent study suggests that the biomass could be at least one order of magnitude larger than previously estimated [13]. The estimate of at least 10,000 million tons, 100 times

Figure 3.2: Mueller's pearlside, Maurolicus muelleri [48].

Figure 3.3: Glacier lantern fish, Benthosema glaciale [49].

the annual global catch of all fish species [13], could still be lower than the actual biomass. The possible underestimation is because the study covered only the latitudinal range of 40°N to 40°S, whereas mesopelagic fish is still abundant in higher latitudes. The mesopelagic layer is a net consumer of oxygen due to the lack of plants. The associated fauna intercepts about 90% of organic carbon before it reaches the ocean bottom and may contribute up to 30% of the carbon dioxide in the ocean [25]. Defaecation in deep waters accelerates the carbon flux and may have important implications for the biogeochemical cycles of the ocean [13].

In *Biogeography of the Global Ocean's Mesopelagic Zone* [50], Proud *et al.* considered the effect global warming might have on fauna in the twilight, and predicted that ocean warming might benefit fisheries production in the mesopelagic. It was also predicted that both the DSL's depth and density would change in the future as a result of warming.

3.4 Machine learning

3.4.1 Overview

Machine learning is an area of computer science which concerns automatically evolving algorithms that through weighted statistical equations may learn how to perform a task [51]. In short it can be explained as the ability to teach a computer how to perform a very specific task, given very specific input. A resulting instance of machine learning is called a model, and should be able to make determinations or predictions about data it has not seen before. The main difference between machine learning and traditional software is that a person (developer) has not manually written code that instructs it how to operate - the algorithm automatically creates rules based on the data provided.

An example of machine learning is the spam filter in email software. Given examples of spam emails (e.g flagged by users) and regulars emails, it can learn to flag spam automatically. There are many different machine learning algorithms to choose from, with different strengths and

weaknesses. Which one to choose depends on the problem and what you want to solve for. Generally the algorithms are divided into four main categories [51], supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning.

Supervised learning

In supervised learning, the dataset used for training is a collection of labeled examples [51]. Labels are the solutions or answers to each sample in the dataset. For the spam filter example this means that for each email sample used to train the model there is a label saying whether the email is classified as spam or regular (not spam). After training, a model should be able to predict the label of the provided input (predictors). Supervised learning can be used for both classification and regression problems. Classification classifies data using discrete class labels, such as spam and not spam, while regression predicts a continuous value or quantity such as salary, price, amount etc [52].

Unsupervised learning

In unsupervised learning the data is a collection of unlabeled examples. The goal of unsupervised learning is to discover underlying patterns in the data, usually for clustering – dividing the data into meaningful groups – or dimensionality reduction. For this reason it is also called knowledge discovery. Unlike supervised learning, unsupervised learning methods cannot be directly applied to regression or classification problems as one does not know what the values of the output would be [53].

Semi-supervised learning

In semi-supervised learning the dataset contains both labeled and unlabeled data, a combination of supervised and unsupervised learning. Typically there will be a small amount of labeled data combined with a large amount of unlabeled data. Supervised algorithms require large amounts of data to train models with high prediction performance, in practical applications there is huge amounts of unlabeled data available which hinders the model to incorporate it as manual labeling is both time and cost consuming. Another reason for lack of labels is that the researchers might not know the specifics of the results they are looking for. Semi-supervised learning tries to deal with this issue. One of the simplest ways of semi-supervised learning is self-training. Self training works by first training the model with labeled data, letting it predict labels for unlabeled data and combining it with the labeled data, then training the model again on the now larger amount of labeled data [54], [55].

Reinforcement learning

Reinforcement learning is quite different from the previous categories. The model, called an agent, exists in an observable environment. Based on observations the agent can select and perform actions, aiming to get the highest possible "reward". There can also be "penalties" for performing the wrong actions. It must then learn by itself the optimal strategy to get the most reward over time [52]. Several impressive reinforcement learning applications have been released. In 2017 a reinforcement learning program developed by DeepMind called AlphaGo beat the reigning world champion Ke Jie at the game of Go [56], and in 2019 OpenAI Five became the first AI system to defeat the world champions of Dota 2, a very popular multiplayer online battle arena game [57].

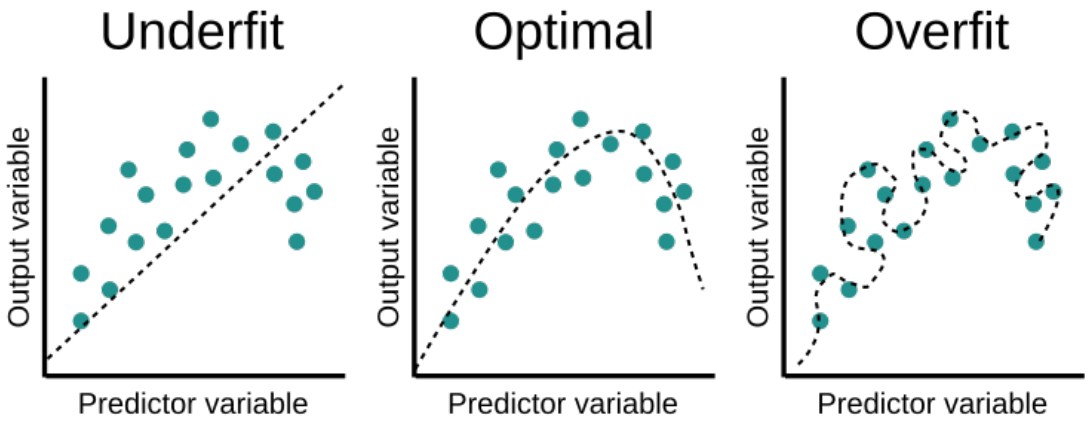


Figure 3.4: Graphs depicting underfitting, a balanced fit, and overfitting [59].

In this thesis, the focus is supervised learning and regression problems.

3.4.2 Training and evaluating machine learning models

Generalization

The goal of any machine learning algorithm is being able to generalize [58]. Generalization refers to a model's ability to adapt properly to new, unseen data similar to the one used to create the model. The importance of this is that a model could report good performance metrics on data it has previously used to train, i.e training or calibration data, while performing worse than a random guess would do for unseen data, i.e test data. This is called overfitting. Overfitting occurs when a model learns noise and details in the training data to the extent that it produces an analysis that corresponds closely or exactly to that particular dataset [58], as visualized in Figure 3.4. Most machine learning models have parameters that can be tuned to reduce the risk of overfitting, as well as techniques used in pre-processing of the training data. The opposite phenomenon is called underfitting and occurs when the machine learning model is too simple. Thus it will be unable to detect patterns in the data, and performs poorly even on the training set. Together overfitting and underfitting are some of the main reasons for poor generalization in machine learning models [51]. The optimal model has a balanced fit to the data, performing well on both training and validation data. This can be measured by different metrics depending on the problem to be solved, which will be presented later in this section.

Bias and variance

A model's generalization error can be expressed as the sum of three different types of errors. These are called bias, variance, and irreducible error [52].

- **Bias** is due to wrong assumptions, and tells us how accurate the model is on average across different possible training sets [60]. An assumption can be that the data is linear, when it is actually exponential.
- **Variance** refers to the models' sensitivity to small variations in the training set.

- **Irreducible error** is caused by noise in the data. This part of the error can only be reduced by cleaning the data in preparation for machine learning, for example detecting and removing outliers.

The bias-variance decomposition can be illustrated as trying to hit the bullseye on a dartboard, see Figure 3.5. If the hit point varies significantly, there is high variance. If they are far from the bullseye, there is high bias. The ideal is having both low bias and low variance, see Figure 3.5, however this is often difficult, introducing the bias-variance trade-off. On small data sets the error can often be decreased by for instance introducing a small bias that causes a large reduction in variance [61]. The bias-variance trade-offs is closely related to overfitting and underfitting. Models with low bias will be flexible enough to fit the training data well, but if it is too flexible it will start overfitting due to high variance. As model complexity increases, variance tends to increase as well. As such, complex models tend to have high variance and low bias and are prone to overfitting. Simple models tends to have high bias and low variance, making them prone to underfitting. A balanced model will strive for low bias and low variance to minimize generalization error [62], see Figure 3.5.

Train-test-split

In order to measure a models' ability to generalize it is common to split the dataset into several smaller sets [52]. Two of the simplest and probably most common ways to divide the data are two-part splits and three-part splits. Two-part splits randomly divides the data into one training set and one test set, while three-part splits divides into a training set, a validation set, and a test set [58]. The training set is used for training a model. The performance of the model is then measured on the validation set, and the results are used to find and optimize the best model to solve the problem. This can be a comparison between different algorithms, or several instances of the same algorithm with different hyperparameters. Hyperparameters are parameters whose values are used to control the learning process [64]. The type of parameters that are available varies with the algorithm. Finally, the test set is used to give an estimate of the performance of the selected model on unseen data.

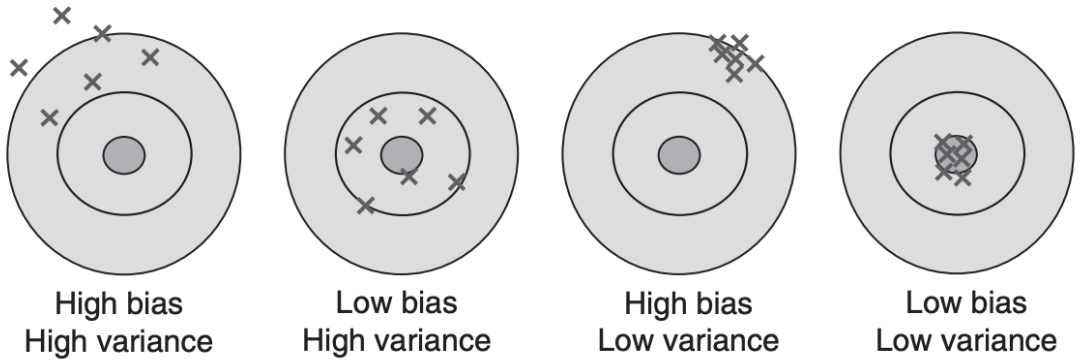


Figure 3.5: Bias Variance Decomposition illustrated as hitting the bullseye on a dartboard [63].

Evaluation

One of the core tasks in building a machine learning model is to determine if the performance of the model meets the expectations or requirements. There are many possible metrics available for evaluating the performance of a model, and choosing which to use is an important part of the process. The best choice of metric for any model depends on the problem that the model is meant to solve.

Evaluation metrics

While evaluation metrics for classification problems usually tend to focus on accuracy or precision/recall as the number of correctly predicted data points out of all the data points, the most common metric for regression is error. Accuracy and error are complements of each other. For regression problems it is unlikely that the model will perfectly predict a continuous value, as opposed to classification where the results are binary, such as spam and not spam. As the problem here is a regression task, the performance metrics represented will be geared toward regression. Even if regression analysis has been employed in a huge number of machine learning studies, no consensus has been reached on a single, unified, standard metric to assess the results of the regression itself [65]. Some of these metrics are presented here.

Root Mean Square Error (RMSE)

The most commonly used metric for regression tasks is root mean square error (RMSE). This is defined as the square root of the average squared distance between the predicted value and the observed value [64]. The biggest problem with RMSE is that it is sensitive to outliers, it is not robust. If the model performs very badly on a few data points, it may affect the average error significantly. RMSE is based on mean squared error (MSE) and is usually preferred over MSE as RMSE is measured in the same units as the target, while MSE is measured in units that are the square of the target variable [65].

Mean Absolute Error (MAE)

Compared to RMSE, mean absolute error (MAE) treats each error the same. As mentioned one big error may be enough to give a bad RMSE, while MAE gives the same importance to all errors. As such, using RMSE or MAE depends on the task at hand, but RMSE is usually preferable when the data has a normal distribution [66]. In this thesis, MAE was used in favour of RMSE, as it is more robust [67] and its interpretation is subject to less ambiguity [68]. The target in this case seemed to have a exponential distribution, see Figure 3.6, which tend to produce Laplacian-like (non-normal), which MAE also tends to yield better results for than RMSE [69].

R^2

Another popular regression metric is the coefficient of determination, namely R^2. R^2 measures the goodness of fit of a model, the proportion of the total variability in the outcome that can be explained by the model [70]. There are several ways to compute this value [71], the simplest one involves finding the standard correlation between the observed and predicted values, and squaring it. An R^2 value closer to 1 means that the model explains close to all variability in the outcome, while a value closer to zero indicates that the predictions have no linear association to the outcome.

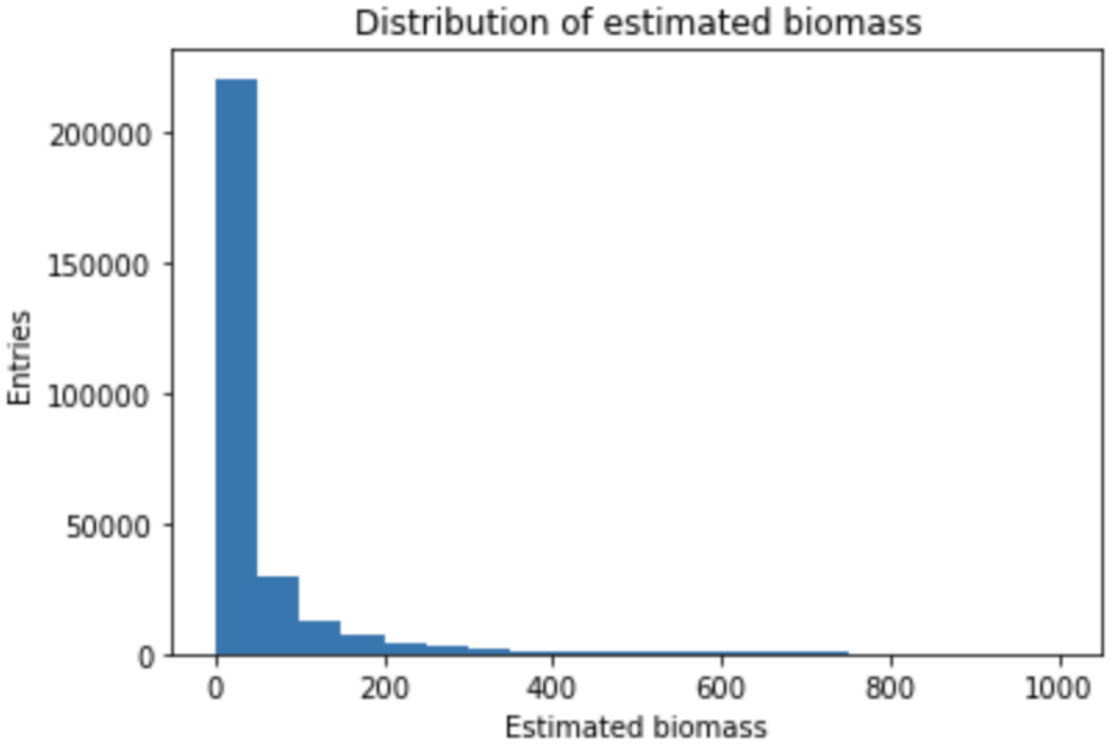


Figure 3.6: Distribution of values in the target variable, estimated biomass.

The main problem of R^2 is that it can be deceiving. It is not a measure of accuracy or error, but rather a measure of correlation. This metric alone does therefore not give a good indication of how well the predicted and observed values agree. It is possible to get a high R^2 value without the observed and predicted values conforming to a 45° line of agreement. A common example of this is in tree-based ensemble methods, where the model tends to under-predict at one end and over-predict at the other end. There are contrasting views about whether or not R^2 is a good performance metric [72], [73]. In this thesis it will for the aforementioned reasons be used in combination with other metrics to assess the performance of the resulting models.

Mean Absolute Percentage Error (MAPE)

Mean absolute percentage error (MAPE) is a widely used measure for forecast accuracy, but it is also popular for business use as it generally regarded as easy to understand. The accuracy is measured as a percentage, representing how far the prediction is from the observed value. A low score is preferred for MAPE, meaning that the prediction is close to the ground truth. It does have a significant disadvantage in that it produces infinite values for zero or near-zero actual values [74]. This led to a new metric called symmetric mean absolute percentage error (SMAPE).

Symmetric Mean Absolute Percentage Error (SMAPE)

To deal with the infinite value problem of MAPE, the symmetric mean absolute percentage error (SMAPE) was introduced. While MAPE is bounded on the low side by an error of 100%, there is no bound on the high side. SMAPE has both an upper and lower bound, sidestepping the infinite value problem of MAPE [75].

3.4.3 Regression algorithms

Regression is a statistical technique for predicting continuous variables based on the value of one or more predictors. There are two types of regression, univariate where only one output value is estimated, and the multivariate regression where more than one output value is estimated [76]. There is often confusion between multivariate and multiple or multivariable regression. Multivariate regression is as explained when there is more than one dependent variable, while multiple regression means that there is more than on independent variable [77]. The problems of this thesis thus deals with univariate multiple regression, as the target to be estimated is biomass. Five types of regression algorithms are explained, where three will be used in the process of implementing the solution. These are partial least squares regression, decision tree, and random forest. The remaining algorithms, linear regression and principal component regression, provide some background information for understanding the chosen algorithms for implementation.

Linear regression

Linear regression, also called ordinary least squares (OLS) is one of the simpler regression algorithms and assumes that the relationship between the variables can be expressed as a linear function [78]. Equation (3.1) is for multiple linear regression, where several independent variables are used to predict the target (dependent variable).

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n, \tag{3.1}$$

where Y represents the target, β_0 is the value of Y when all independent variables are equal to zero, X_1 through X_n are the independent variables, and β_1 through β_n are coefficient parameters. The model tunes the values of the coefficient parameters to optimize the accuracy, in linear regression this is typically estimated using MSE [78].

Principal Component Analysis (PCA)

Principal component analysis (PCA) is an unsupervised dimensionality reduction technique which can be used in both classification and regression problems. It is placed under regression here as it will lead us to principal component regression (PCR) and partial least squares regression (PLSR). Dimensionality reduction involves reducing the number of independent variables in a data set. PCA helps in grouping multiple variables into fewer variables without losing much information from the original set of variables [79]. This is useful when for example there are many highly correlated values in a data set, or the amount of variables is so large that the computation would be very time consuming.

PCA is based on the assumption that the higher variance our data has, the more information it provides. It projects data onto a lower-dimensional space such that it retains as much variance as possible. These linearly uncorrelated synthetic values are called principal components [80], and are linear combinations of the original variables. Figure 3.7 illustrates the first and second

principal component on a two-dimensional plot. If a dimension contains little to no variance, it is discarded. Due to these characteristics, PCA is most useful when the variance is unevenly distributed across the data set.

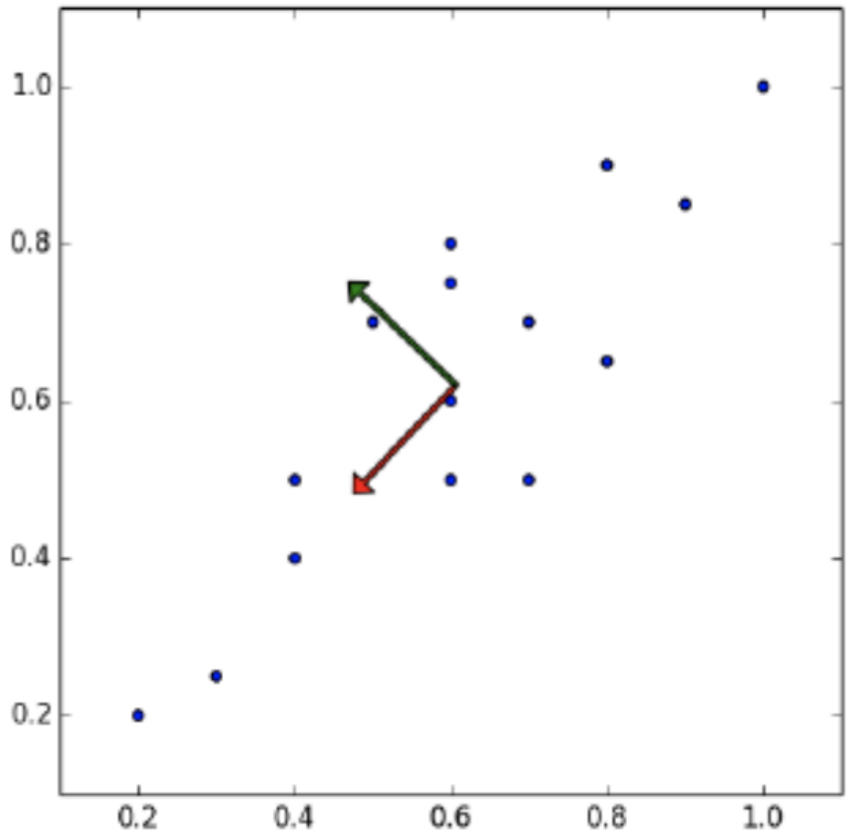


Figure 3.7: Illustration of two principal components on a data set plot. Red arrow is the first principal component, and green the second. The first component will explain more of the variation in the system than the second component. Image from [80].

Principal component regression (PCR)

PCR is a regressor composed of two steps, first PCA is applied to the training data, then a regressor is trained on the transformed samples. PCR is used for combating multicollinearity and tends to result in estimation and prediction better than ordinary least squares [81].

Partial Least Squares Regression (PLSR)

PLSR is based on the same theory as PCR, but is a supervised algorithm and thus takes the target into account when determining principal components unlike PCR. A drawback of PCR is that it only captures the variables with the most variance, which are not necessarily the ones that

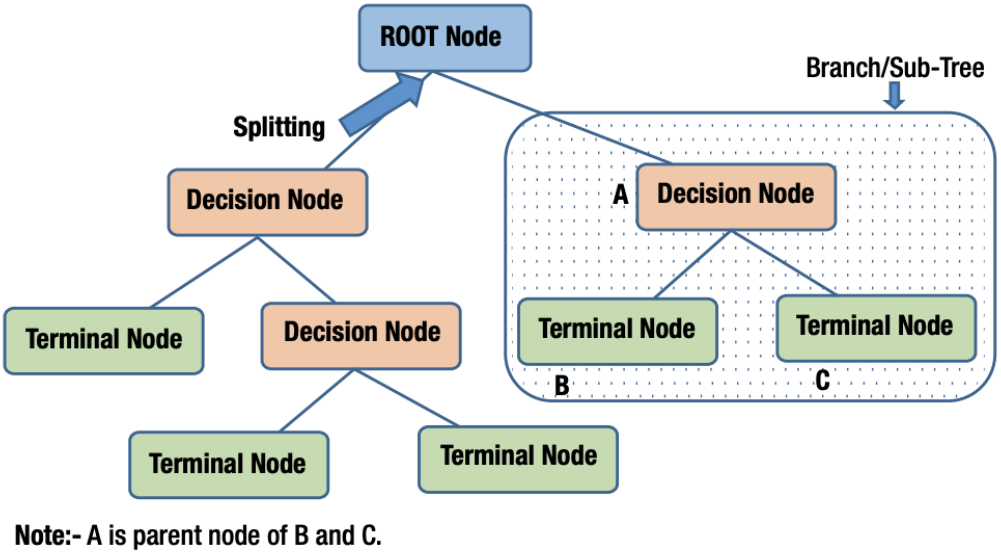


Figure 3.8: Components of a decision tree [79].

have the most predictive power. PLSR tries to identify predictive variables that capture as much information in the raw predictive variables as well as in the relation between the predictive and target variables [82]. This trait is useful for the problems in this project as some features, such as CHLAF, had a much smaller variation than other features. PLSR is preferable for dimension reduction when a target variable for a regression is specified, in this case the biomass (TOTAL), in comparison to PCR due to it being supervised [82].

Decision tree

Decision trees form the basis of all tree-based algorithms, and can perform both classification and regression tasks. A tree consists of a root node, decision nodes, and leaf nodes. A sub-section of a tree is called a branch or a sub-tree, as shown by the stippled region in Figure 3.8.

When building a decision tree regressor, the data set is split according to which variable improves some evaluation metric the most, for example mean squared error. This built-in feature importance calculation is helpful for looking at the correlation between our independent variables and the target variable. Starting at the root node (the original data set) this process is repeated usually until a max depth is reached, or there are no longer enough samples left in the data set to perform a meaningful split. Max depth of a tree and minimum samples are parameters that can be set and tuned by the developer. Depth refers to the number of "layers" in a tree, where the root node is at depth 0. The tree in Figure 3.8 has a depth of 3.

Each decision node will contain a value called a threshold. When using the tree to predict a value, it decides whether to go right or left at a decision node depending on if the value of the relevant variable is higher or lower than the threshold. When it reaches a leaf node the value of that node is returned as the prediction.

Ensemble methods

Ensemble methods are techniques that aim at improving results by combining multiple models instead of using a single model. A model is often called a learner. An ensemble contains a number of learners called base learners, these are usually instances of the same base algorithm, but some ensembles may have learners based on different base algorithms. Ensembles of a single base algorithm are called homogeneous ensembles, while ensembles with several base algorithms are called heterogeneous ensembles [83].

In ensemble learning it is often referred to weak and strong learners. Weak learners are simple models that are only slightly better than random guessing, while strong learners have relatively good accuracy [84]. Ensemble methods can generally be divided into two groups which utilize different ways of combining weak learners, boosting or bagging.

Boosting refers to algorithms that converts (boosts) weak learners to strong learners. The idea is to correct the mistakes made by a base (weak) learner, eventually resulting in a strong learner. Base learners are trained sequentially. With each new model iteration, the weights of the predicted data with the larger errors in the previous model are increased. This redistribution of weights helps the algorithm identify the parameters that it needs to focus on to improve its performance. In the end the ensemble takes a weighted decision based on the prediction of the sequence of base learners to estimate the result [85]. Popular boosting algorithms are AdaBoost and BrownBoost [83].

Bagging is short for bootstrap aggregating, where base learners are trained in parallel rather than sequentially. Each base learner is trained on a small random subset of the training set, and the final prediction is the average of all the result from all base learners. This way, the predictions are less likely to be biased due to a few outlier cases [79]. A popular bagging algorithm is random forest.

Random forest

Random forests are ensemble methods that consist of several decision trees. Single decision trees have the disadvantage of being unstable. Slight variations in the training data can cause different attribute selections at each choice point within the tree, and significantly impact the outcome [86]. The combination of several trees deals with this issue by averaging the predictions. In addition, random forests incorporate randomized feature selection [83] which provides a great basis for finding correlations using feature importance. During the building of a base learner, at each split a random subset of features is selected. Features can be ranked by importance based on the change in error affected by the presence or absence of a feature in a subset [87]. Random forests has been proven to not overfit no matter the number of trees in the forest, and are also robust to outliers and noise [88]. However, they might still overfit due to tree depth. Tree-based machine learning algorithms have gotten considerable attention in the area of above ground biomass estimation (see *Biomass estimation using machine learning* in Chapter 2) making them a good candidate for mesopelagic biomass estimation.

3.4.4 Feature engineering

A feature in machine learning is a measurable property of the thing you are trying to analyze, in data sets features appear as columns. Feature engineering is the process of selecting, manipulating, and transforming raw data into features using domain knowledge, that better represent the underlying task to be solved. As machine learning aims to address more comprehensive and

complex tasks, the problem of detecting the most relevant information in a enormous amount of data has become increasingly important [89]. Models that work well on just the unprocessed data can still benefit from feature engineering by increasing performance. Model effectiveness is influenced by many things, and different types of models each have their own sensitivities and needs. For example some models does not tolerate features that measure the same underlying property, called multicollinearity, while other models are compromised when there are any irrelevant predictors in the data [70]. Multicollinearity can for example be a data set containing both features age and date of birth, where one can be inferred from the other.

Feature selection is the process of selecting features from the available data to use for developing the model. This involves analyzing, judging, and ranking various features to determine which contribute the most to the prediction, and which are irrelevant or redundant.

Feature extraction is the automatic creation of new features by extracting them from the raw data. The goal is to get features with more predictive powers, and to lower the dimensionality of the data.

Feature selection will be an important part of this thesis, as one goal is to find which, if any, variables impact the acoustic biomass as explained in Chapters 1 and 3.

3.4.5 Feature importance

Random forest is an useful technique for feature selection. In Scikit's implementation of random forest regressor, feature importance assessed using mean decrease in impurity (MDI) computed from all the decision trees in the forest is already built-in [90]. The MDI of a feature is computed as a weighted mean of the base trees' improvement in the splitting criterion produced by each variable [91]. MDI tends to be a good measurement for feature importance, but might not reliable in situations where potential predictor variables vary in their scale of measurement or their number of categories [92].

Another means of assessing feature importance is permutation importance. Permutation feature importance measures the increase in the prediction error of the model after a feature's values was permuted, which breaks the relationship between the feature and the true outcome [93]. It is calculated using the increase in prediction error after randomly permuting (shuffling) feature values. If the prediction error increases after shuffling, the feature is deemed "important", as the model relies on it to predict the outcome. Both MDI and permutation importance were used in assessing feature importance and thereafter selection to get a more coherent picture rather than only using one of them.

3.4.6 Tuning

Hyperparameter tuning of an algorithm consists of finding a set of optimal hyperparameters. Selecting the best hyperparameter configuration for machine learning models usually has a direct impact on the model's performance [94]. Searching for the optimal hyperparameters can be done manually or automatically. Manual tuning is done by applying values using trial and error, which is very time consuming. Automated hyperparameter tuning uses algorithms to search for the optimal values. Two of the most popular algorithms are grid search and random search. In grid search a grid of possible hyperparameter values are created, then the model is fitted with every possible combination of parameters. Random search is a variation of grid search. Instead of searching over the entire grid, it only evaluates a random sample of points on the grid [33]. For

both methods, all results are recorded, and the model that performed the best is returned along with the chosen hyperparameters.

3.5 Tools and libraries

This section contains a brief overview over the most important tools used in the project. The implementation of the experiments was done in Jupyter Notebook, using the programming language Python. Jupyter Notebook was chosen due to its literate nature which combines formatted natural language text, executable code snippets, and computation results. These traits makes for a more easily understood code base when used appropriately [95].

Scikit-learn

Scikit-learn is a machine learning library providing implementations of many well-known machine learning algorithms and evaluation metrics [96]. Utilizing this library can save significant time importing methods rather than implementing them from scratch. Implementations for the different algorithms used, splitting the data, and some evaluation metrics were all imported from Scikit.

Pandas

Pandas is a data analysis library developed specifically for Python [97]. It provides tools for working with structured data sets and performing common routines in data manipulation and analysis.

Chapter 4

Methodology

4.1 Design science

Design science is a problem-solving paradigm that seeks to create innovations that define the ideas, practices, technical capabilities, and products through which the analysis, design, implementation, and use of information systems can be effectively and efficiently accomplished [98]. A designer science researcher answers questions relevant to human problems via the creation of innovative artifacts, thereby contributing new knowledge to the knowledge base. Artifacts are broadly defined as models, constructs, methods, and instantiations [99].

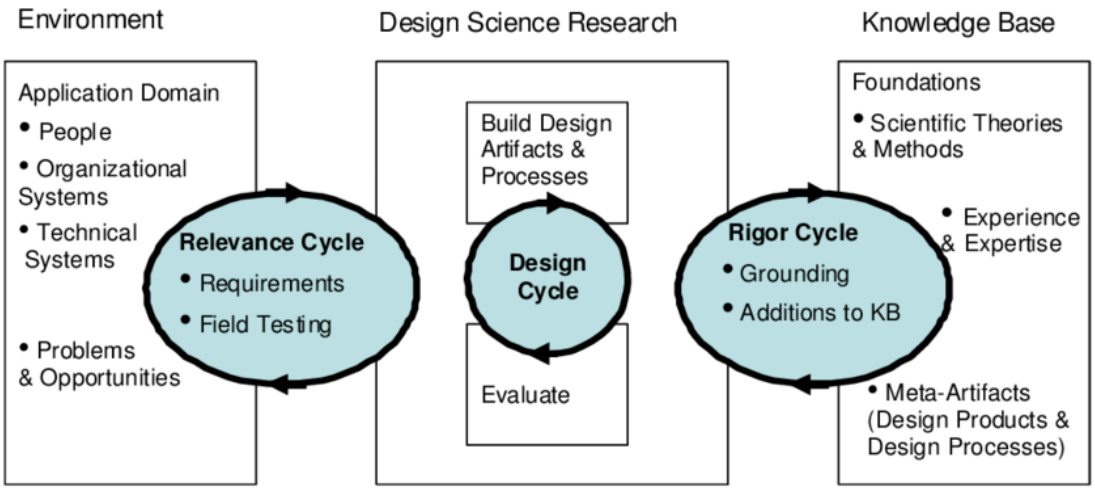


Figure 4.1: Design science research cycles [100]

In this project, a machine learning model for predicting mesopelagic biomass is trained. The solution-seeking part of this project will be to build a model that marine science researchers can use to better understand the biomass of the deep scattering layer and the diel vertical migration. Building this model is a data science activity, and the research contribution will be the knowledge gained by using this model to find relationships between physical and biogeochemical properties

of the ocean and atmosphere and estimating biomass. Therefore, on the one hand, this project is solution-seeking by building a machine learning model, but on the other hand, from the marine researcher's point of view, it is knowledge-seeking. Due to these factors it can be viewed as design science, where the artifact is the machine learning model marine science researchers can use to understand the mesopelagic biomass better. However, in this project, both the roles of data science and marine science research must be considered. In addition to creating the artifact, it is also applied to the data collected during the One Ocean Expedition. This approach is possible as both data science and marine knowledge are included in this project. It is pointed out that design science is used and not constructive research, where the focus is a lot more on the artifact's need to be a commercial product [101]. The purpose of the artifact is to gain more knowledge about the ocean and how machine learning can be applied in marine science research. In summary, in this thesis, it is shown how the artifact is produced, that is, the data science part, and the result of applying it is communicated, the empirical part, which will be new knowledge that did not exist before this project.

A conceptual framework for design science research were defined in an research essay published in Management Information Systems Quarterly (MISQ), March 2004, to provide an understanding of the research activities of design science within the information systems research discipline. This included a set of seven guidelines for conducting and evaluating good design science research. In summary, they encourage to utilise all available means to create relevant and viable artifacts that contributes to the relevant domain(s). Within the definitions of design science research, six activities were defined to guide the research process of this thesis. Each activity, save communication, will be further described in detail during the thesis.

- **Problem identification:** Great uncertainty surrounding the global abundance of mesopelagic species and ocean properties related to diel vertical migration (DVM). Currently no automated solutions for estimating the biomass or predicting the DVM of the mesopelagic layer in different areas. Further explained in Chapter 3.
- **Objective:** Automated analysis of DVM patterns in relation to physical-chemical properties of the water masses. Define a machine learning model trained for estimating biomass and DVM in an area given ocean metrics.
- **Design:** Develop (a set of) models trained to solve the identified problem.
- **Demonstration:** Demonstrate the developed model(s) on previously unseen data.
- **Evaluation:** Evaluate the models with well defined metrics.
- **Communication:** Through this thesis.

4.2 Machine learning approach

The approach taken to using predictive supervised machine learning to find correlations in this thesis consist of seven main steps. The steps rely on first identifying the problem, as it will guide the selection, aggregation, and transformation of the input data as well as the selection of machine learning algorithms and metrics. The seven steps are collecting data, preparing the data for use, choosing machine learning algorithms, training the models, evaluating the models, tuning the chosen model, and using the final model to make predictions, see Figure 4.2.

Figure 4.2: Seven steps of machine learning. Picture from [102].

Step 1: Gathering the data

Before gathering the data, one has to identify what data is needed. Generally, the more data available the better. As this thesis was based around the One Ocean Expedition, it was known that data would be available. The data collected during legs 1 to 5 of the One Ocean Expedition was aggregated to be used for training the machine learning models. In addition information about sea floor elevation from The General Bathymetric Chart of the Oceans (GEBCO) was incorporated for further analysis.

Step 2: Data preparation

The majority of the time for running machine learning end-to-end is spent on preparing the data, which includes collecting, cleaning, analyzing, visualizing, and feature engineering [103]. Data formatting, cleansing, abstraction, feature engineering, collinearity analysis, and understanding of attribute characteristics are essential steps for the selection of meaningful model input parameters [104]. Unwanted data and redundant features are removed, missing values are treated, values are transformed and so on. This process is explained in detail in Chapter 5. Before the data is used for training, it is shuffled and split into a training set, validation set, and testing set.

Step 3: Choosing algorithms

It is important to choose an algorithm that is relevant to the problem. Algorithms are developed for different tasks, such as natural language processing, image recognition, prediction and so on. In addition, some algorithms are better suited for classification tasks, while some are suited for numerical tasks.

Four algorithms were chosen for the initial implementation, a baseline algorithm, partial least squares regression, decision tree, and random forest. Except for the baseline, all of them provide a basis for feature selection which is an important part of the project. Several algorithms were chosen as this lets us compare which performed the best, and chose the most appropriate one to base feature selection on. Each algorithm was chosen based on specific traits they have, as presented in Chapter 3, which are likely be a good fit for the problems identified in this thesis. Both decision tree and random forest were selected as even though random forests usually perform better than decision trees, if they turn out to have about the same performance, the simpler option is usually preferable. Simpler models are less likely to overfit, tend to run faster, and are easier to understand.

A baseline estimator is a very simple model to to get a baseline value for the performance metrics. The baseline is used to compare more complex solutions like machine learning models with, to see if they perform better [105]. Scikit provides several dummy estimators to get a baseline, like

a model that always returns the mean of the dataset. As it is known that depth is an important factor for the estimated biomass, a slightly more advanced baseline estimator was created from scratch. The baseline estimator created for this thesis returns the mean value of the estimated biomass at a specific depth, see Listing 4.1 for the pseudocode of the baseline prediction. It takes a set of averages and a validation set as input, and returns the average biomass at the depth given for each row in the validation set.

```
Input: Set of averages and a validation set
Output: List of averages

function predict
    create an empty list
    for each row of the validation set
        get PDMEAN of the row, find the average biomass at PDMEAN using the
            averages set
        append the average to the list

    return the list of averages
```

Listing 4.1: Baseline prediction pseudocode.

Step 4: Training

The prepared data is passed to the chosen machine learning algorithms to find patterns and make predictions. In supervised learning, the algorithm learns from the input to find a model that minimizes loss. In regression tasks this is generally estimated by error, how far the predicted value is from the actual value.

Step 5: Evaluation

Three different evaluation metrics were chosen to compare the performance of the models. Using more than one metric helps give a more accurate picture of the performance, as they all evaluate the models in different ways. The metrics chosen were mean absolute error (MAE), R^2, and symmetric mean absolute percentage error (SMAPE). MAE alone does not say much about the performance of the regression concerning the distribution of the ground truth elements [65]. Two metrics that generate a high score if the majority of the elements of a ground truth group have been correctly predicted are R^2 and SMAPE. R^2 indicates how well the variation in the ground truth values are explained by the model, while SMAPE might be the easiest to understand for an audience without machine learning or information technology expertise, indicating how far from ground truth the predicted values usually are. In combination, these three metrics should give a sound basis for evaluating the model developed in the thesis. All four models trained during step 4 are evaluated, and the best performing model is selected to proceed with the next step, tuning. At this point, the model is also used to calculate feature importance. Training the model on several subsets of the original data sets lets us analyze feature importance across geographical regions. It also facilitates feature selection for the final model.

Step 6: Tuning

Once the model is created and evaluated, it is tested if the performance can be improved by tuning the hyperparameters present in the model. Hyperparameters are variables the programmer can decide, and varies by algorithm. Some of the parameters in random forest are the max depth of a tree, number of trees, max features in a tree and the minimum number of samples present to

perform a split. Tuning a model may or may not increase the accuracy of the predictions. The tuning algorithm applied in this thesis is random search.

Step 7: Prediction

After tuning, the best performing, and thus final, model is used to make predictions on unseen data. The unseen data is the test set created during step 2. The performance of the model on the test set is the expected performance of the model on other unseen data.

Chapter 5

Pre-processing

This section will explain how the data was processed before being used to train the models, this includes data cleaning in the form of removing noisy or bad data, treatment of outliers, splitting the data, and the creation of different subsets of the full data set.

5.1 Overview

The raw data was acquired from the Norwegian Marine Data Center (NMDC) [8], the sensor data is available through their Underway API and the echosounding data through FTP. Sensor data can be received as either JSON or XML. In this case JSON was used for the raw data before it was manipulated into a more usable format. Note that in the generated time series data used for training, the depth (PDMEAN) refers to a specific depth in the water column and not the depth of the sea floor. The depth of the sea floor will hereby be referred to as the sea floor elevation.

Raw echo sounding data was processed with Large Scale Survey System (LSSS), a software product developed by Norwegian Research Centre (NORCE) for processing and interpreting of marine acoustics data [106]. LSSS is a licensed software, but one can also use the preprocessing tool developed by CRIMAC which is open-source and available on github [107]. The unit used for estimated biomass when generating reports with LSSS is the Nautical Area Scattering Coefficient (NASC), denoted as sA (m^2/nmi^2). The CRIMAC tool uses volume backscattering strength, denoted Sv (dB re $1m^{-1}$) [108], thus meaning that values should be transformed to sA if using this tool in relation to the work done it this thesis.

The original set of 35 sensors was whittled down to a set of 10 to be used for further analysis and training of the models. Several sensors were removed as their purpose was to measure internal values of the different systems, such as the internal temperature of the Ferrybox. Others were removed due to measuring the same thing as another sensor, just in another unit or transformed. This included for example the conductivity measure due to its correlation to salinity, dissolved ions increase salinity as well as conductivity [109]. Finally, the sensors related to waves and wind was removed. High waves and/or wind speed negatively impacted the quality of the echo sounding data, and any correlation found here is therefore highly likely to be related to noise. Further explanation regarding this can be found in Chapter 8. Table 5.1 shows the full list of remaining sensors along with short explanations, units of measurements, and the installed system

System	Sensor	Measure	Unit
Ferrybox	CDOMFluorescence	Chromophoric dissolved organic matter fluorescence, indication of primary production and dissolved organic carbon	parts per billion (ppb)
	Temperature	Temperature of the water at approximately 5 meters depth	degrees C
	CHLAFluorescence	Chlorophyll-a fluorescence, indication of primary production	micrograms per liter (μg/L)
	Turbidity	Amount of light that is scattered by material in the water, indicates the relative clarity of the water	nephelometric turbidity unit (NTU)
	Optode Saturation	Percentage of dissolved O2 concentration relative to that when completely saturated	%
	Salinity	Concentration of salts	practical salinity unit (PSU)
Weather station	Trykk / Pressure	Atmospheric pressure	mbar
	Humidity	Concentration of water vapour in the air	%
pCO2	DIC	Dissolved inorganic carbon	μmol/kg

Table 5.1: Final list of sensors with short explanations and associated unit of measurement.

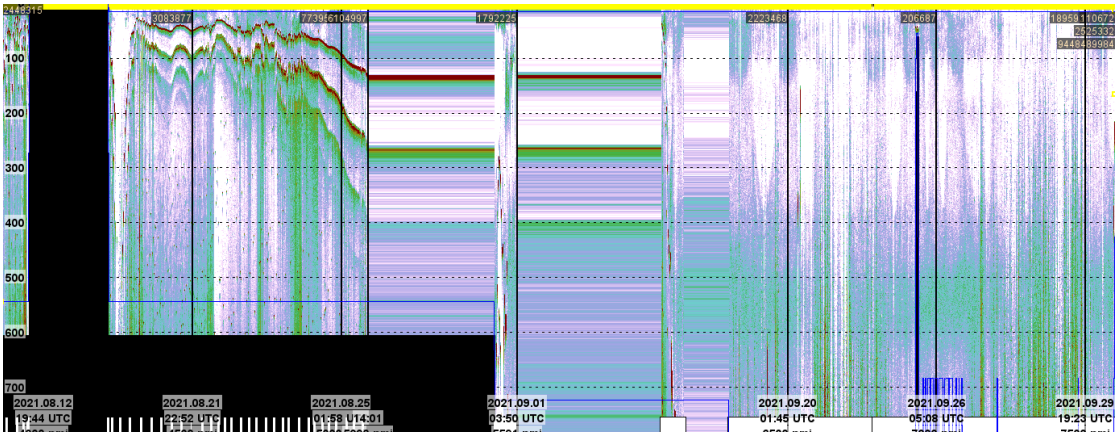
they are associated with.

5.2 Data cleaning

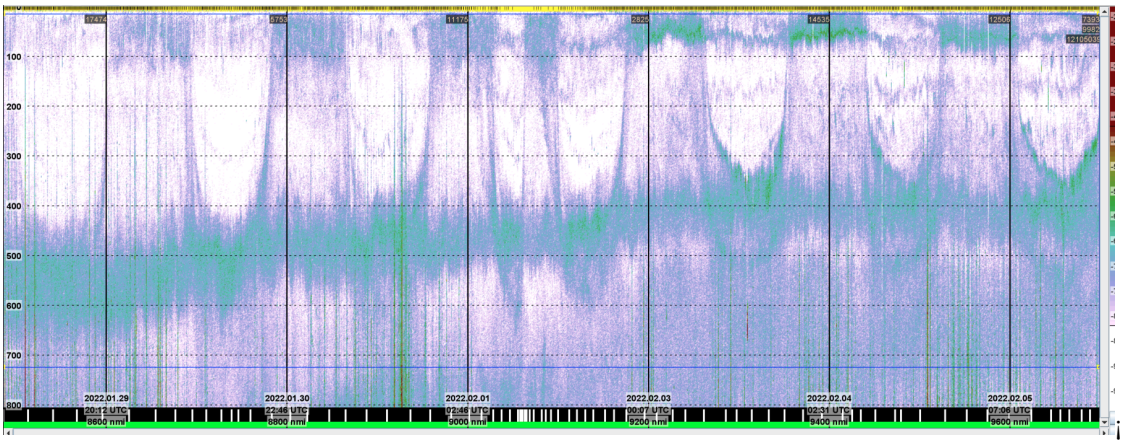
Data cleaning is the process of detecting and fixing incorrect, incomplete, duplicate or otherwise erroneous data. This can include removing or correcting outliers, noise, and other unwanted data. Outliers are data points that deviate significantly from the rest of the data points, and can be caused by for example errors in measurement. Noise is irrelevant or meaningless data which can significantly affect the data analysis tasks ahead.

5.2.1 Acoustic data

The raw acoustic data processed with LSSS output time series data with longitude, latitude, and estimated acoustic biomass at depth increments of 15 meters. The maximum depth was set to 730 meters as the deep scattering layer is generally found between 400-600 meters, and the quality of the acoustic data decreases with the depth. For some areas the data was only available for shallower depths, likely due to someone accidentally messing with the settings of the echo sounder on board the ship. The data was cleaned by loading each leg into LSS, identifying stretches of bad data visually. Bad data was then marked and excluded, before generating a time series report containing the remaining data. This process was repeated for each of legs 1 through 5, the legs of the circumnavigation available at the time of this thesis. Figures 5.1a and 5.1b show examples of bad/noisy and good data respectively that was collected during the



(a) Example of noisy or bad echo sounding data.



(b) Example of good echo sounding data.

Figure 5.1: Examples of (a) bad and (b) good echo sounding data as visualised with the LSSS software.

circumnavigation.

5.2.2 Sensordata

The first treatment of the sensor data involved removing all data points recorded while the ship was in harbour. Water intake was stopped when the ship was in harbour to counteract clogged pipes due to dirty water and littering. Therefore several sensors did not record values while in harbour, while others recorded strange measurements, such as the temperature spiking in the systems no longer taking in water. The sensor data was then visually analyzed to see which sensors might have several outliers, by comparing the min, max, and average values for each sensor. In the data from SL, the salinity measure especially were likely to have many outliers due to daily freshwater rinses of the Ferrybox system. The measurements are not stopped during this process, resulting in several data points equal or closer to zero. As the normal ocean salinity range is 33-37 grams per liter [110], all values lower than 30 were replaced by the median. The likelihood for the other Ferrybox derived values to be affected by the rinse is high, and were treated similarly

based on the indices derived from handling the salinity outliers. In addition, data points that were outside the acceptable range of the measurement were removed. Both weather station sensors seemed to be within reasonable range of their respective measurements. Table 5.2 shows an overview of acceptable ranges for measurements where known, and the approach for handling outliers for each measure. Unknown ranges contain no values, only "-", and related measurements were generally kept without any alterations. The process of replacing Ferrybox values with the median were done separately for each data subset, see Section 5.3, to ensure that the median would be representative of the recorded values in each set. Dissolved inorganic carbon (DIC) was calculated using fugacity of CO_2 (fCO_2), this is further explained in following section.

5.2.3 Dissolved inorganic carbon (DIC)

The measurements from the PCO2 system is not fit for analysis directly from the NMDC API. They have to be run through a script, and can then be accessed from the Pacific Marine Evironmental Laboratory (PMEL) [111]. This was confimed with Meike Becker, who was responsible for the PCO2 system installed at Statsraad Lehmkuhl. At the time of this thesis only data from 2021 had been processed and was accessible from PMEL [112]. The measurement of interest was fugacity of CO_2 (fCO_2), the partial pressure corrected for non-ideality of interaction between CO_2 and N_2/O_2 [113]. fCO_2 is very dependent on temperature. To circumvent the effect of the temperature, Meike advised to calculate dissolved inorganic carbon based on the fCO_2 value. First, alkalinity was calculated using sea surface salinity (SSS) and sea surface temperature (SST). The variables for the equation depends on latitude, longitude, SST, and SSS [114]. Equation 5.1 shows the base equation. After finding the alkalinity, DIC was calculated using the python toolbox PyCO2SYS [115], alkalinity values, and fCO_2 values, and was added as a new column in the sensordata.

$$A_T = a + b(SSS - 35) + c(SSS - 35)^2 + d(SST - 20) + e(SST - 20)^2 \tag{5.1}$$

5.2.4 Flipped latitude southern hemisphere

Looking at a map plot of the biomass at a specific depth, it turned out that for around -10° latitude the values had been flipped to positive again. Normally this would be expected for either the complete southern or northern hemisphere, but something must have happened with the sensor at some point during this leg of the journey. The flipped values, that is positive latitude values in the southern hemisphere, were flipped back so they would reflect the position of the ship accurately. The process was confirmed with an expert familiar with the data.

5.2.5 Combining echo sounder and sensor data

After getting both the sensor data and the echo sounding data into the same format, they had to be combined. The index for each measurement was the timestamp formatted YYYY-MM-DD HH:MM:SS, which was the best option to join the two data frames on. The timestamps did not match up to begin with, as different sensors registered measurements at different intervals. For the echo sounding this was approximately every half hour, while it varied from every minute to every ten minutes for the environmental sensors. Due to the echo sounding data containing the target value, these timestamps were chosen to be merged on. All timestamps in the echo sounding data frame were rounded to the nearest half hour. The sensors were grouped together

[1] Based on calibration values for nephelometers

Measure	Acceptable range	Approach
CDOMFluorescence	-	Values recorded during freshwater rinses of the Ferrybox system replaced with median
Temperature	-2°C to 35°C [116]	Same as above
CHLAFluorescence	-	Same as above
Turbidity	0 to 4000 NTU[1] [117]	Same as above
Optode Saturation	-	Same as above
Salinity	33 to 37 PSU [110]	All values <30 replaced with median, no values >37
Trykk / Pressure	Average 1013.25 millibar at sea level [118]	Recorded min and max were close to the average value at sea level with a min of 1006.63mb and max of 10028.92mb. Kept as is
Humidity	20% to 100% [119], [120]	Recorded min and max within the acceptable range, kept as is
DIC	-	Already processed, kept as is

Table 5.2: List of measurements with acceptable ranges where known and approach for handling outliers.

in intervals of thirty minutes and the mean for each sensor in every group was calculated. Finally the data sets were joined on the new timestamps. Part of the combined data set can be seen in Figure 5.2, not all sensors are pictured. In several consecutive rows all feature values will be the same except for PDMEAN, the depth in the water column, and TOTAL, the biomass. This is due to there being several biomass measurements for the same time stamp, one for every 15 meters. The time stamps were deconstructed into month, day, and time of day to look for possible correlations with the target.

	LATITUDE	LONGITUD	PDMEAN	TOTAL	sf_depth	NMEA.Humidity	NMEA.Trykk	FerryBox.SBE45_Salinity	FerryBox.Optode_Saturation
0	27.20153	-16.88955	7.5	0.0000	-3611.0	75.750000	1011.796173	35.8295	95.365002
1	27.20153	-16.88955	15.0	4.6863	-3611.0	75.750000	1011.796173	35.8295	95.365002
2	27.20153	-16.88955	25.0	24.4440	-3611.0	75.750000	1011.796173	35.8295	95.365002
3	27.20153	-16.88955	35.0	30.9997	-3611.0	75.750000	1011.796173	35.8295	95.365002
4	27.20153	-16.88955	45.0	33.6487	-3611.0	75.750000	1011.796173	35.8295	95.365002
...	...	...	...	...	...	...	...	...	...
279363	-23.19454	-40.66192	685.0	35.0402	-1894.0	85.858334	1008.785004	35.8820	94.059998
279364	-23.19454	-40.66192	695.0	45.1827	-1894.0	85.858334	1008.785004	35.8820	94.059998
279365	-23.19454	-40.66192	705.0	351.3631	-1894.0	85.858334	1008.785004	35.8820	94.059998
279366	-23.19454	-40.66192	715.0	42.6268	-1894.0	85.858334	1008.785004	35.8820	94.059998
279367	-23.19454	-40.66192	725.0	18.3478	-1894.0	85.858334	1008.785004	35.8820	94.059998

Figure 5.2: Example of the combined data set. Note that there are more columns than pictured.

```
Input: Echo sounding dataset: echo, sea floor elevation dataset
Output: Echo sounding dataset with a new column containing sea floor elevation in
    meters

function find sea floor elevation
    for each row in echo
        find closest measure of sea floor elevation by position
        add the sea floor elevation in a new column

    return echo sounding data set with sea floor elevation
```

Listing 5.1: Pseudocode for adding sea floor elevation to the echo sounding dataset.

Seafloor elevation

Generally, the most biologically productive regions of the oceans are coastal shallow seas [121], and seamounts and ridges are usually perceived as areas of elevated productivity and biodiversity in the ocean [122], [123]. Due to this there was also an interest to see if the seafloor elevation might impact mesopelagic biomass or the diel vertical migration. Statsraad Lehmkuhl did not record the seafloor depth, so a global grid was downloaded from General Bathymetric Chart of the Ocean (GEBCO) [124]. The grid consists of data points with longitude, latitude and elevation. The positions were compared with those in the data set from SL, and the recorded elevation closest to the position of SL was added in a new column. In total the full data set after pre-processing consisted of 279,368 rows$\times$18 columns.

5.3 Geographical and other sub data sets

As seen in the map plot, Figure 5.3, the biomass seemed to vary depending on the geographical area. This agrees with the observations made on the Malaspina expedition as mentioned in Chapter 2. In addition to analysing the full data set, several subsets of the full data set were made with regards to geographical area and maximum depth for further analysis. The three geographical area sets consisted of data from the Caribbean, Mid-Atlantic, and the coast of South America. The three groups were decided mainly by the locations of the data points on the map plot, but they also exhibit some physical differences. The Mid-Atlantic (MA) group consist of open ocean measurements, while the Caribbean and South American (SA) are closer to shore. Sutton, Clark, Dunn, *et al.* [125] proposed several mesopelagic ecoregions of the world's ocean. These regions were derived by variations in terms of biodiversity and function, and can be seen in Figure 5.4. Comparing the map plot with the proposed ecoregions, the data points of the Caribbean and MA sets are found within region 24, while the data points of the SA set are mostly in region 27, with a few in region 28. Region 24, the Central North Atlantic, is a broad area of consistent temperature-salinity-oxygen (TSO) conditions. In comparison, region 27, the Tropical and West Equatorial Atlantic, is characterized by easterly winds causing divergence and upwelling, creating a narrow band of high productivity and marked differences in the combination of TSO characteristics from adjacent areas [125]. These differences should be likely to cause some variation in feature importance for the geographical data sets.

The longitude and latitude restrictions for each zone was decided based on the data points of the map plot. The Mid-Atlantic data points where within -3°to 15°latitude, and -44°to -28°. The Caribbean data points where within 15°to 25°latitude, and -85°to -70°longitude. At the time of this thesis, all data points from the southern hemisphere were along the coast of South America.

The data points for South America were thus all simply in the southern hemisphere, meaning a negative latitude. To isolate the diel vertical migration (DVM), two subsets of the full set consisting of data with a maximum depth of 50 meters and 200 meters were created. Similarly, two subsets with max depths of 50 and 200 meters were created for each geographical data set. All of these subsets were fed separately to the algorithm that seemed to perform the best on the full data set. There were no recorded values for humidity in the Caribbean, or for the PCO2 system along South America. These columns were therefore dropped from the respective data sets. The Caribbean data set had 49,684 rows, the Mid-Atlantic set had 73,498 rows, and the South American set had 18,396 rows.

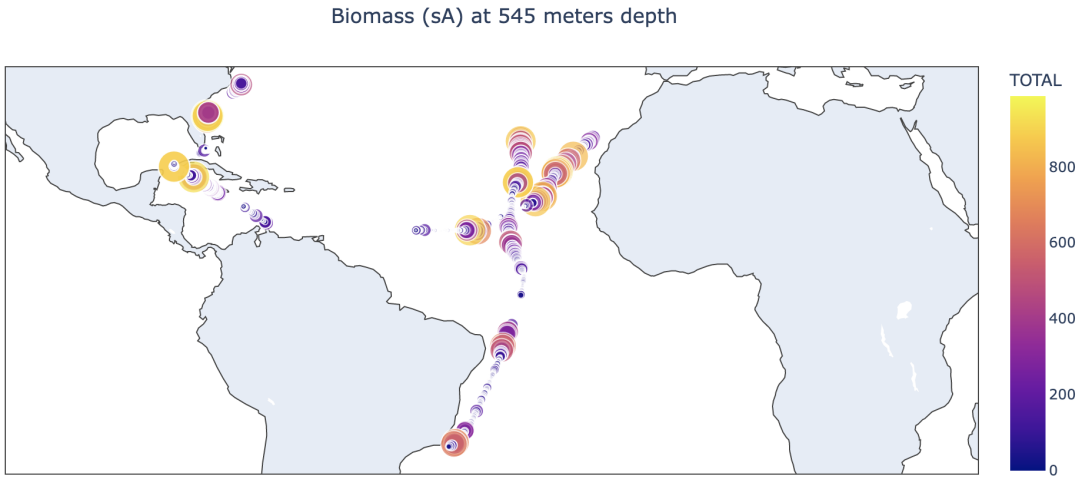


Figure 5.3: Map plot of the biomass at 545 meters depth. Larger points and lighter color indicate a larger estimate of biomass, while small and darker points indicate smaller estimates.

5.4 Train-val-test split

Time series data is usually treated in a special way compared to other data, when splitting the data set into training, validation, and test sets [126]. The reasoning behind this is that time series is usually the data form used in forecasting where a lot of historical data spanning several years is used, such as weather and stock forecasting. The data in this project is time series, but spans less than a year total. Lacking historical data, the typical time series split is not necessary, as there will not be any historical trends. For these reasons a regular random split of the data was performed to create training, validation and test sets. Empirical studies show that the best results are obtained using a 70/30 or 80/20 split [127]. The split chosen was 80/20 with 80% of the data for training and the 20% split evenly into a test and validation set, both 10% of the original data set.

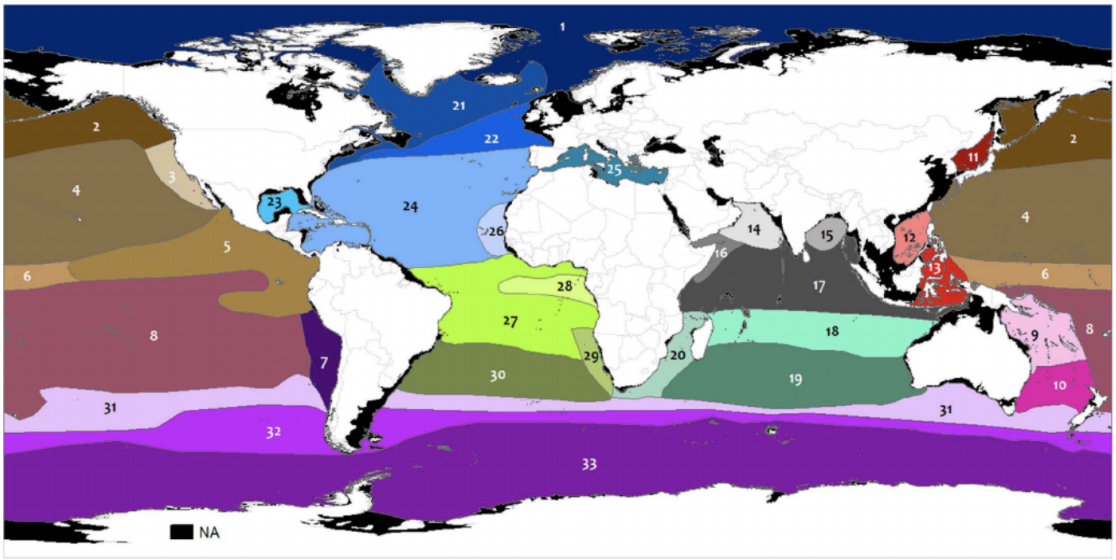


Figure 5.4: Proposed mesopelagic ecoregions by Sutton, Clark, Dunn, *et al.* Areas with depths less than 200m shaded in black. Image from [125].

Chapter 6

Results

This section will present the results obtained from the implementation. This includes choosing the best model of the original four (baseline, PLSR, decision tree, random forest), results from training the chosen model on the different data sets, feature importance and selection, and an estimate for how the model will perform on unseen data.

6.1 Choosing the best predictive algorithm

The first objective was to identify the algorithm that performed the best on the full data set. Two data scaling settings were used to train each algorithm, one with scaled data and one with unscaled. For the evaluation metrics chosen during implementation there would optimally be low values for MAE and SMAPE, and a high value for R^2.

The results were very similar with unscaled and scaled data, with the only big difference in MAE, see the "MAE" columns of Tables 6.1 and 6.2, respectively. This is due to the scaling. Scaled (smaller) estimated values give smaller absolute errors. R^2 and SMAPE values were about the same between models. PLS performed the worst out of all the algorithms, including the baseline, with the worst estimated values for all metrics. The baseline implementation was next worst, almost in line with PLS. Decision tree and random forest significantly outperformed the baseline model, with random forest performing the best in all metrics with an MAE of 15.865, R^2 of 0.697, and SMAPE of 0.315 on unscaled data. Random forest also performed the best with scaled data. Tables 6.1 and 6.2 shows the metric values for all algorithms for models trained with unscaled and scaled data, respectively, with the best in each category underlined and bold. Based on these results random forest was chosen as the algorithms used for further experiments with data sets, and for hyperparameter tuning on the full data set.

Unscaled data was used for the models in rest of this section as distance is not a factor for tree-based algorithms, and random forest performed slightly better with regards to SMAPE with unscaled data, compared to scaled.

Model	MAE	R^2	SMAPE
Baseline	47.669	0.073	1.023
PLS	50.300	0.069	1.156
Decision Tree	19.351	0.505	0.333
Random Forest	**15.865**	**0.697**	**0.315**

Table 6.1: Results of the different models on unscaled validation data.

Model	MAE	R^2	SMAPE
Baseline	0.495	0.073	1.216
PLS	0.522	0.069	1.359
Decision Tree	0.201	0.511	0.353
Random Forest	**0.165**	**0.699**	**0.324**

Table 6.2: Results of the different models on scaled validation data.

6.2 Feature importance

The second objective was to look at which features appeared to influence the target (estimated biomass) the most. As random forest performed the best of the models, it was used for assessing feature importance with both mean decrease in impurity and permutation as described in Chapter 3. Instances of random forest was trained with each of the data sets created during implementation. These data sets are the full data set, the full data set to 50 meters depth, the full data set to 200 meters depth, three data sets based on geographical regions - the Mid-Atlantic, Caribbean, and the coast of South America, and the geographical sets with maximum depths of 50 and 200 meters. The reason for creating the different data sets was to observe if different features were assessed to be important based on geographical area or depth. As such the performance of the models for these sets was not the focus, however evaluation metric values for each of the sets can bee seen in Table 6.3 and confirms the generalization ability of the random forest. The results are generally in line with the full set, with some variations in both positive and negative directions, the worst performing one being the model for the coast of South America. The South American data set was also the one with the least data points.

Table 6.4 show feature importance calculated for the original data set using MDI and permutation respectively. Each feature is listed with the corresponding permutation importance value and MDI value. For MDI the values depict how much the impurity decreases when splitting on the relevant feature, while for permutation the values explain how much the error increases (accuracy decreases) when the feature is randomly shuffled so that it no longer has a connection to the target. For example the highest rated feature for both metrics is depth, the depth in the water column. Splitting on this feature reduces the impurity by approximately 30% (MDI) or leads to a 100% or higher increase in error when randomly shuffled (permuted). The features rated highly using MDI are in accordance with features rated highly using permutation, with some slight variations, so for the sake of simplicity the feature importance values for the rest of the data sets will all be based on permutation.

The most important feature in all data sets was PDMEAN, the depth in the water column, which is in accordance with the background information presented in Chapter 3 regarding the mesopelagic. For the complete original data set the most important features were depth, sea floor elevation (sf_depth), latitude and longitude. Oxygen saturation, turbidity, and time might

Data set	MAE	R^2	SMAPE
Full	15.865	0.697	0.315
Full set 50 m	10.522	0.895	0.179
Full set 200 m	8.857	0.829	0.240
Mid-Atlantic	11.525	0.739	0.319
Caribbean	11.834	0.696	0.250
Coast of South America	19.581	0.674	0.326
Mid-Atlantic 200 m	6.899	0.884	0.231
Caribbean 200 m	14.450	0.781	0.244
Coast of South America 200 m	9.713	0.727	0.298
Mid-Atlantic 50 m	9.661	0.918	0.176
Caribbean 50 m	7.372	0.831	0.220
Coast of South America 50 m	8.108	0.924	0.283

Table 6.3: Results of random forest trained on all the different data sets.

also play a factor, but did not show a significant impact. Some differences between the feature importance for the various data sets were discovered, especially in regards to the geographical data sets. The Mid-Atlantic model relied heavily on surface temperature in addition to the depth, position being of the importance as for the full model. In the Caribbean the latitude was assessed as the second most important feature, showing a significant increase in mean accuracy when present, while none of the biogeochemical features registered. Finally the model for the coast of South America assessed the CDOMFluorescence to be of some significance, while both longitude and latitude registered much lower than for all the previous models.

Feature importance for the models created to isolate the diel vertical migration were mostly in line with the others in regards to the importance of depth, position, and sea floor elevation. Latitude, however, was far less important than longitude. In addition the time of day was estimated to be very important, again in accordance with the information provided in the background. There were not significant differences between the 50 m model and the 200 m model, however longitude and time was more important for the former.

The geographical models with a maximum depths of 50 and 200 meters did not all rely on position as most previous sets had, with the exception of the Caribbean 200 m model. Sea floor elevation registered highly for the South America 200 m model. Time and depth was the most important features for all these models, except the Caribbean 50 m model which relied little on time, and rather registered quite high for oxygen saturation. Table 6.5 presents the permutation importance values for all data sets. Values $\geq 15\%$ are underlined, and values $\geq 50\%$ are both bold and underlined. Some values are negative, this happens when the permutation importance is close to zero, but predictions improved after shuffling due to random chance.

Feature	Permutation importance	Mean decrease in impurity
Latitude	0.389	0.135
Longitude	0.302	0.120
Depth	1.149	0.323
Sea floor elevation	0.258	0.105
DIC	0.015	0.024
Humidity	0.017	0.028
Air pressure	0.042	0.036
Salinity	0.016	0.014
Oxygen saturation	0.136	0.032
Turbidity	0.142	0.040
CHLAFluorescence	0.022	0.022
Temperature	0.052	0.020
CDOMFluorescence	0	0.009
Day	0.055	0.016
Month	0.020	0.017
Time	0.164	0.059

Table 6.4: Feature importance for the full set using permutation importance and MDI.

Feature	Full set	50m	200m	Mid-Atlantic (MA)	Caribbean (C)	South America (SA)	MA 50m	C 50m	SA 50m	MA 200m	C 200m	SA 200m
Latitude	0.389	0.175	0.122	0.432	**0.861**	0.213	0.144	0.160	0.142	0.111	**0.526**	0.023
Longitude	0.302	**0.804**	**0.616**	0.269	0.359	0.160	0.025	0.092	0.091	0.076	0.223	0.027
Depth	**1.149**	**1.074**	**1.345**	**1.112**	**1.384**	**0.956**	**1.039**	**1.191**	**0.635**	**1.506**	**1.094**	**0.715**
Sea floor elevation	0.258	0.098	0.159	0.277	0.279	0.314	0.035	0.150	0.048	0.024	0.116	0.392
DIC	0.015	0.030	0.025	0.031	0.035	No values	0.012	0.050	No values	0.017	0.035	No values
Humidity	0.017	0.012	0.010	0.010	No values	0.017	0.010	No values	0.012	0.014	No values	0.010
Air pressure	0.042	0.013	0.022	0.072	-0.004	0.047	0.080	0.020	0.024	0.174	0.002	0.009
Salinity	0.016	0.007	0.009	0.002	-0.003	0.011	0.013	0.004	0.001	0.012	0	0.015
Oxygen saturation	0.136	0.005	0.057	0.113	-0.003	0.009	0.013	0.293	0.001	0.012	0.011	0.009
Turbidity	0.142	0.005	0.004	-0.004	0.005	0.002	0.024	0.001	0.007	0.006	0	0.004
CHLA Fluorescence	0.022	0.020	0.007	0.035	0	0.006	0.001	0	0.009	0.002	0	0.003
Temperature	0.052	0.008	0.012	**0.531**	0.005	0.019	0.009	0.013	0.027	0.020	0.004	0.008
CDOM Fluorescence	0	0.002	0.001	-0.003	0.001	0.094	0.003	0.003	0.002	0.001	-0.004	0.141
Day	0.055	0.017	0.055	0.001	0.003	0.003	0.006	-0.002	0.006	0.004	0.006	0.008
Month	0.020	0.006	0.004	0	0	0	0	0	0	0	0	0
Time	0.164	**0.824**	**0.685**	0.157	0.237	0.069	**0.815**	**0.170**	**1.107**	**0.765**	**0.615**	**0.587**

Table 6.5: Feature importance for all data sets calculated using permutation importance.

Model	MAE	R^2	SMAPE
Default model	**15.865**	0.697	**0.315**
Tuned model	15.876	**0.735**	0.342

Table 6.6: Comparison of evaluation metrics for default model and tuned model on validation data

Parameters	Default model	Tuned model
Number of trees	100	200
Min samples split	2	5
Min samples leaf	1	2
Max features	Auto	Auto
Max depth	None	60
Bootstrap	True	True

Table 6.7: Hyperparameters in the default random forest model and the tuned random forest model

6.3 Tuning and final result

Tuning hyperparameters of the model trained on the full data set using random search improved the R^2 score slightly, while MAE and SMAPE got slightly worse. As both improvement and deterioration of evaluation metrics were insignificant, the default model (no tuning) was kept as it had slightly better values for two of the three metrics, see table 6.6. A comparison of hyperparameters in the default model and the best model found during random search cross validation can be seen in table 6.7. Training a default random forest with only the features that scored $\geq 15\%$ for importance in the full set resulted in the following values for the evaluation metrics on the test set

$$MAE = 15.667$$

$$R^2 = 0.720$$

$$SMAPE = 0.316$$

and are the expected performance of the full model on unseen data. Removing the three features with the least importance, oxygen saturation and time, resulted in a minimal negative difference in all metrics.

$$MAE = 15.831$$

$$R^2 = 0.718$$

$$SMAPE = 0.318$$

Chapter 7

Threats to validity

The validity of a study denotes the trustworthiness of the results, to what extent the results are true and not biased by the researchers' subjective point of view [128]. To clarify potential threats to validity of this thesis we follow the scheme proposed by Runeson and Höst [128]. While this is not a case study like Runeson and Höst focused on when creating the guidelines, they operationalized the scheme for flexible design studies such that it may be used outside of case studies as well. Below we go in to the four aspects of validity based on this thesis; internal validity, external validity, construct validity, and reliability.

Internal validity

Internal validity is the extent to which the observed results represent the truth of what we are studying and are thus not due to methodological errors. Firstly due to being onboard Statsraad Lehmkuhl during leg 4 of the circumnavigation, we are familiar with a large part of possible outliers in the data such as FerryBox values being unreliable during freshwater rinses and temperatures spiking when the ship was in harbour due to the main water pump being shut off. This helps more accurately identify and replace outliers in several features. It is however not true for all features as discussed in section 5, which were kept as is with the assumption that the data was reliable due to lack of domain knowledge.

All data used in the thesis was collected using the same methods, and several different algorithms and metrics were used to evaluate the result. Data was shuffled randomly before splitting into training, validation, and test sets to better represent the overall dataset and thus help the model generalize better.

External validity

External validity is concerned with to what extent it is possible to generalize the findings. We have evaluated this directly using the test set that was set aside before training and evaluating the models with training and validation data. The test set contained data that the model had never seen before, and therefore provides an estimate as to the performance on other unseen data. Optimally this would also be tested using similar data from another source, however we could not find data that included the same variables as in this project without joining data collected from several different sources.

Construct validity

As explained in section 5 the selection of machine learning algorithms is based on scientific research, and the results were evaluated using several metrics. Additionally the thesis was conducted with regular check-ins with Geir Pedersen, a domain expert on acoustic backscatter by underwater targets, acoustic measurements, and marine data processing and analysis.

Reliability

Reliability is concerned with whether the results are dependent on the specific researcher. If another researcher later on conducts the same study, the result should be the same [128]. All sources for the data used in the thesis are provided in section 5, along with an open-source alternative for processing of echo sounding data if the researcher does not have access to the licensed software LSSS. Pre-processing is described in detail in the same section, specifying units, treatment of outliers and other relevant information pertaining to the values in the data set. Splitting of the data and training of the models all have specified seeds to ensure reproducibility. The full code for the project can be found on GitHub, see link and explanation of the code structure in appendix A.

Chapter 8

Discussion of results

In the previous sections the results of the various models were presented along with feature importance for different data sets, and the validity of the models was confirmed. In this section the results are interpreted and possible reasons behind them are discussed. The research questions are also answered.

8.1 Implementation specific

From the results in section 6 it is expected that the final full model will on average have a difference of approximately 16 between the predicted value and the actual value. This means that if the actual value is for example 100, the model will on average predict 84 or 116. In comparison the SMAPE value is approximately 0.3, meaning that on average the predicted value will be off by 30%. The difference between SMAPE and MAE may be due to the error distribution, as SMAPE is relative while MAE is absolute. Small errors on small actual values will therefore have a higher percentage error than a small error on a large actual value. Figure 8.1 show a scatter plot of the predicted values compared to the actual values from the test set of the full model. Another possibility is the way SMAPE is calculated. While it is not infinity like MAPE for actuals close or equal to zero, it can still get to unreasonable high values for near zero actual values which exist in the data set. A scatter plot of only 500 data points is provided in figure 8.2 for better readability. The model predicts actuals on the lower end of the scale more accurately than higher values. This makes sense, as can be seen from the distribution of actual values in figure 3.6 from the implementation that there is a very large dominance of smaller values.

Partial least squares regression was the worst performing model in the thesis, scoring lower than the baseline estimator. The bad performance of PLSR may be due to the assumption of linearity between the dependent variable and the independent variables, as PLSR is a linear regression method. This suggests that the relationship between the biomass and the other features is non-linear and the focus for further work should be on non-linear models.

Feature importance variation

From literature [19], [25], [50] it is known that the depth in the water column is important for biomass, and the time for diel vertical migration. Our results show that other variables such as

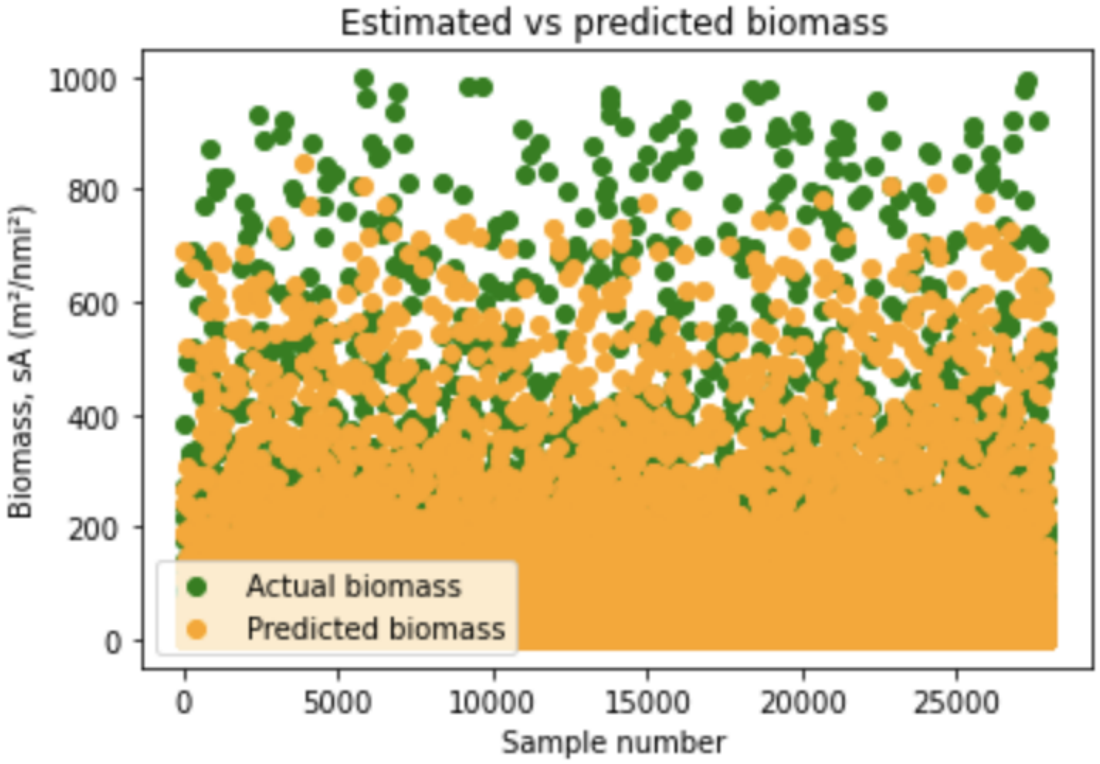


Figure 8.1: Scatter plot of predicted biomass (yellow) and actual biomass (green) for each sample in test data

position and sea floor elevation also impact the biomass of the deep scattering layer. Models based on different geographical areas show that which variables are important, and the degree of impact they have on the mesopelagic scattering layer, depends heavily on the area. There is indeed a difference between different regions as suggested by Klevjer, Irigoien, Røstad, *et al.* [12], a study based on data from the Malaspina expedition. The results show that differences are also present at smaller scales than between oceans, and supports the claim that parts of the spatial variability can be explained by horizontal patterns in physical-chemical properties of the water masses. No strong correlations were discovered between the DVM and the physical-chemical features, however the results imply that the DVM biomass is more related to longitude than latitude.

Air pressure came up as a variable that might impact the biomass for the Mid-Atlantic. This does not seem like a probable variable to impact biomass in the DSL. A study on oxygen transfer in aerobic bioreactors suggests that air pressure increases the productivity in biomass of yeast [129], however this is a very controlled study compared to biomass in the ocean, and reasons behind different features' impact should be reserved for discussion by domain experts.

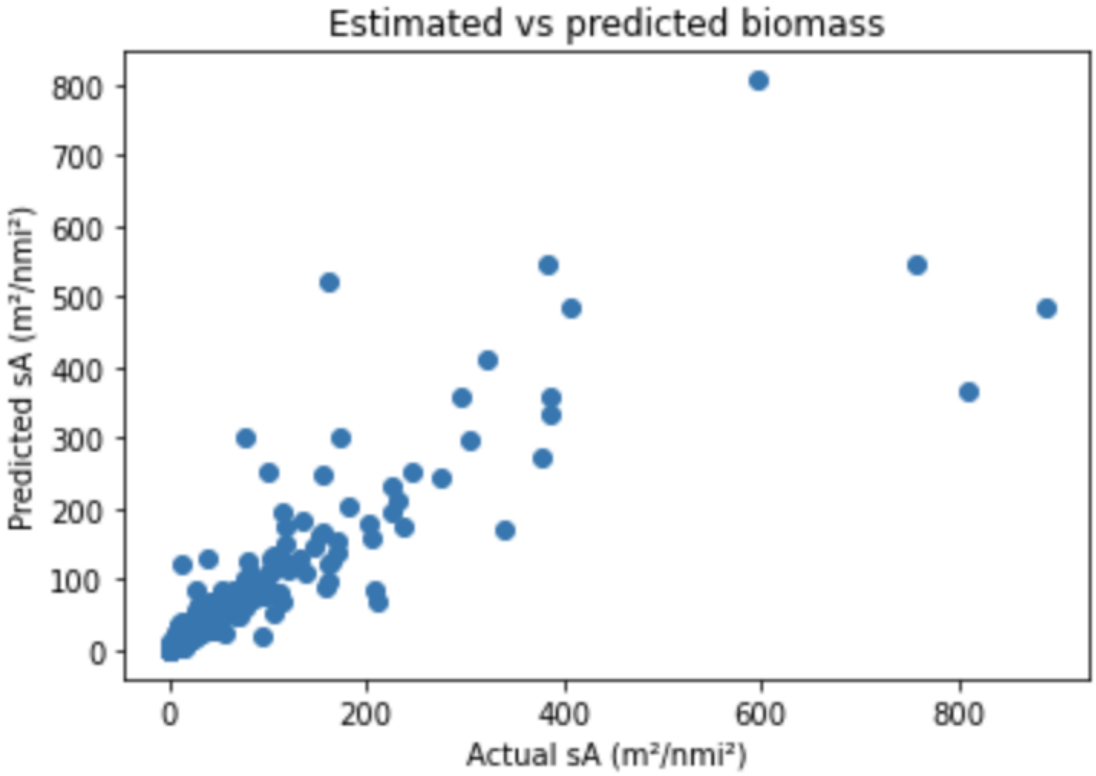


Figure 8.2: Scatter plot of predicted biomass and actual biomass for 500 data points in the test set

8.2 General

During the circumnavigation so far there have been a significant amount of days where the weather deteriorated the echo sounding measurements quite much, due to Statsraad Lehmkuhl being a sailship without modern tools like drop keel/protruding keel to avoid surface bubble layers. Due to this the usable data is not evenly distributed throughout, and some legs do not have any echo sounding data that is fit for use. At the time this thesis was written, data from leg 1-5 was available, but both leg 1 and 5 were omitted due to very noisy echo sounding data, as well as smaller parts of leg 2-4. Some data is also missing or not available for use due to issues with the research permits for the different countries that SL passed during the journey. For example, some days are missing when the ship first entered Brazilian waters because the permit was not approved until a few days later. The map plot in figure 5.3 gives an indication to where data is missing or unusable - ideally there would be an unbroken route of plots throughout.

Acoustic surveys repeatedly show a decrease in mesopelagic fish towards polar environments. Catch data from survey trawls show that the fish community switches from fish possessing gas-filled swimbladders to those lacking swimbladders as latitude increases towards the Antarctic continent, thus the surveys systemically overlook a large proportion of fish species that dominate polar seas [22]. This latitudinal community switch, from gas to non-gas dominance, has considerable implications for acoustic biomass estimation. SL did not venture close enough to polar

environments for this to be an issue with the data used in this thesis. There might have been an impact in leg 5 or 6, but leg 5 was as previously mentioned dropped due to noise, and leg 6 was not available at the time. However it is definitely an important factor for the use of acoustic surveys.

Outliers in the measurements apart from the echo sounder were removed to the best of our knowledge, however for some sensors there was a lack of domain knowledge to properly clean the data before use. Lack of domain knowledge for some features, such as the measurements from the PCO2 system, might thus be polluting the data to some extent. However these features did not register as important during the feature selection for any model, and so should not have a significant impact.

8.3 Answer to research questions

Research question 1: How can possible correlations between the collected physical and geochemical data, and the acoustic biomass in the ocean, be detected using machine learning?

Possible correlations between acoustic biomass and physical and geochemical data of the water masses can be investigated through calculating feature importance based on for example a random forest model. Permutation importance was in this case the least ambiguous and features that scored around 0.4 and higher had significant negative impacts on the performance of the model if removed, suggesting that the features are indeed connected to the biomass. Different areas showed different weights on features, and data should be treated with this in mind, like filtering data sets based on geographical area.

Research question 2: How accurately can the acoustic biomass of organisms in the mesopelagic using oceanic and atmospheric data be predicted with machine learning?

The acoustic biomass of the mesopelagic scattering layer can be predicted at a reasonable accuracy with a mean error of approximately 16. The predictions are more accurate for estimates at the lower end of the scale. As errors increase with larger estimates they should not be taken at face value, but predictions can give a general understanding of the biomass distribution in an area. Models for smaller geographical areas do not necessarily perform better than a complete/-full model.

Research question 3: How accurate predictions can be made about the acoustic biomass of organisms that perform the diel vertical migration (DVM) using oceanic and atmospheric data with machine learning?

Isolating the upper 200 meters of the ocean helps us predict the accuracy of the DVM with a mean error of approximately 11. As with the deep scattering layer this can help give a general understanding of the DVM biomass in an area.

Chapter 9

Conclusion

This chapter will provide a conclusion to the research conducted in this thesis by summarising the work done, the resulting discoveries, and its contribution to the fields of software engineering / machine learning and marine science. Further work will then be proposed.

9.1 Conclusion

This thesis proposes an approach based on random forest to estimate biomass of the diel vertical migration and deep scattering layer. Feature importance based on random forests can be used to investigate correlations between the physical and biogeochemical properties of the water and the biomass, based both on depth and geographical area. Before coming to this solution, a number of machine learning-based models have been tested. These include partial least squares regression, decision tree, and random forest, where random forests were found to give the best result. This finding saves time for future work on this topic as can apply the same method to analyze the data automatically to find out the amount of acoustic biomass in an area without needing new surveys. Our findings show that there are correlations between the physical and biogeochemical properties of the water masses and the biomass; based on depth, geographical area, and seafloor elevation. However, our results do not confirm environmental impacts such as oxygen, turbidity, or indicators of primary production on the mesopelagic layer, like the findings indicated in [12], [13], [16], [17]. Some of these features show up intermittently between the data sets, however generally with lower values. Accordingly, further investigation is needed.

9.2 Further work

9.2.1 Transforming acoustic measures into biomass

The biomass data collected during the circumnavigation were acoustic measures using echo sounding. Transforming acoustic measures into biomass involves uncertainty, dependent of the distribution of fishes according to sizes and acoustic wave lengths. Better knowledge of the composition and acoustic properties of mesopelagic species is needed to obtain further accuracy of the biomass using acoustic measures [13]. It would be interesting to compare the results of this thesis with smaller studies where a more accurate biomass estimate has been made manually, if the ship has been in the same area.

9.2.2 Using the data set from the completed expedition

This thesis was conducted before Staatsraad Lehmkuhl had completed the circumnavigation. A new model should be trained when the circumnavigation is complete, using the finished data set from the entire journey. New results should be compared to the ones obtained in this project, and feature importance for new areas should be explored. There might be differences in importance between different oceans, such as the Atlantic Ocean and the Pacific Ocean, and other smaller geographical areas might be explored based on longitude and latitude or other features.

Further analysis of features

More data was collected during the expedition than used in this thesis in the form of physical samples. Possibly the most exciting one for the problems in thesis being environmental DNA. Environmental DNA originates from cellular material shed by organisms via skin, excrement, etc. and can be used for species detection [130]. This would help gain knowledge about species distribution and relative abundance, and could be used to improve acoustic measurements as mentioned in the previous section. Other physical samples include plankton trawls and micro plastics.

All physical samples collected had to be shipped when SL was in harbour and then processed when arriving at their destination. Processing results might therefore be delayed, and no analysis results from physical samples were available at the time of this thesis.

There is also possibilities for including external data, that is data not collected on the ship, or filtering the data based on criteria other than depth and position. This could be anything that might be correlated to the biomass, such as distance to shore, information about ocean currents, meeting of water masses etc. It could also be considered to predict the depth of the deep scattering layer along with the biomass using a multi output method.

9.2.3 Inclusion of and comparison with historical data

Another interesting area is looking at the possibility of using historical data sets. This would likely mean merging data from several different sources and a lot of work pre-processing work to get everything in the same format. The most promising echo sounding data is probably what was recorded during the Malaspina expedition[1]. Several other measurements were taken during the trip such as temperature, salinity, and CDOM, but no complete data set has been found. If the full data set from the expedition can not be acquired, a new data set could be made using measurements from various sources. Including historical data could give more knowledge about how for example global warming is affecting the deep scattering layer and the diel vertical migration. If such data is acquired the procedure for splitting the data into training and test sets should be adapted for forecasting by splitting with regards to date, this method was mentioned under train-val-test split in section 5.

9.2.4 Automate data cleaning process of echo sounding data

In this thesis cleaning of the echo sounding data was done manually using the licensed software LSSS. As mentioned in section 5 there is open source software for generating time series reports from echo sounding data on GitHub, namely the CRIMAC preprocessing tool. This however gives no visuals aids for manual cleaning and simply generates a report from the raw data.

[1] https://www.nature.com/articles/s41597-021-01038-y

During the circumnaviagtion with Statsraad Lehmkuhl it was noticed that the quality of the echo sounding data deteriorated sharply when there was a lot of wind and wave action. This correlation might be useful in developing an automated cleaning process for echo sounding data, where data recorded during times of strong wind and/or waves taller than some decided threshold is removed. This would need to be further investigated, and if using acoustic data from other sources it might not be a fitting procedure due to differences in impact between for example a tall ship like Lehmkuhl or a dedicated research vessel like the Hespèrides.

Appendix A

Code structure

Code can be found here: https://github.com/maritnl/master

Some functions are repeated across notebooks, such as the function for splitting the datasets. This is due to the fact that notebooks are separate entities, and are not made for importing. If implemented as plain python files, these functions could be moved to a helper file.

algorithms.ipynb: Contains the implementation and evaluation of all the initial algorithms except the baseline. PLSR, Decision Tree, and Random Forest.
echosounderprep.ipynb: Code transforming the time series data to a usable format, and for adding sea floor elevation
fetch_data.ipynb: Contains code for getting data from the NMDC API, and the initial preparation of the sensordata
preprocessing.ipynb: Transforms sensordata into a usable format and combines it with the acoustic data
baseline.ipynb: Implementation and evaluation of the baseline model
forest.ipynb: Contains the code for creation of geographical and depth-based data sets, as well as functions for replacing values measured during freshwater rinses, and calculating dissolved organic carbon. These are placed here as they are applied to each data set to get representative values for the area. Also contains code for plots, and hyperparameter tuning of the random forest.

Bibliography

[1] Ipcc, *The Ocean and Cryosphere in a Changing Climate: Special Report of the Intergovernmental Panel on Climate Change*, 1st ed. Cambridge University Press, Apr. 30, 2022, ISBN: 978-1-00-915796-4 978-1-00-915797-1. DOI: 10.1017/9781009157964. [Online]. Available: https://www.cambridge.org/core/product/identifier/9781009157964/type/book (visited on Jul. 10, 2022).

[2] M. Visbeck, "Ocean science research is key for a sustainable future," *Nature Communications*, vol. 9, no. 1, p. 690, Feb. 15, 2018, Number: 1 Publisher: Nature Publishing Group, ISSN: 2041-1723. DOI: 10.1038/s41467-018-03158-3. [Online]. Available: https://www.nature.com/articles/s41467-018-03158-3 (visited on Jul. 10, 2022).

[3] E. W. L. Michaels, N. O. Handegard, K. Malde, and H. Hammersland-White, "Machine learning to improve marine science for the sustainability of living ocean resources," p. 108,

[4] D. Bzdok, M. Krzywinski, and N. Altman, "Machine learning: A primer," *Nature methods*, vol. 14, no. 12, pp. 1119–1120, Nov. 30, 2017, ISSN: 1548-7091. DOI: 10.1038/nmeth.4526. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5905345/ (visited on Sep. 1, 2022).

[5] M. D. Agersted, B. Khodabandeloo, Y. Liu, W. Melle, and T. A. Klevjer, "Application of an unsupervised clustering algorithm on in situ broadband acoustic data to identify different mesopelagic target types," *ICES Journal of Marine Science*, fsab167 Aug. 31, 2021, ISSN: 1054-3139. DOI: 10.1093/icesjms/fsab167. [Online]. Available: https://doi.org/10.1093/icesjms/fsab167 (visited on Sep. 22, 2021).

[6] V. Allken, S. Rosen, N. O. Handegard, and K. Malde, "A deep learning-based method to identify and count pelagic and mesopelagic fishes from trawl camera images," *ICES Journal of Marine Science*, vol. 78, no. 10, D. Demer, Ed., pp. 3780–3792, Dec. 15, 2021, ISSN: 1054-3139, 1095-9289. DOI: 10.1093/icesjms/fsab227. [Online]. Available: https://academic.oup.com/icesjms/article/78/10/3780/6429121 (visited on Dec. 16, 2021).

[7] (). One ocean expedition, One Ocean Expedition, [Online]. Available: https://oneoceanexpedition.com/ (visited on Aug. 27, 2022).

[8] P. 1. 0. 2. Oppdatert 30.04.2016. (). Norwegian Marine Data Centre (NMDC), [Online]. Available: https://nmdc.no/om-prosjektet/norwegian-marine-data-centre-nmdc- (visited on Aug. 28, 2022).

[9] C. M. Duarte, "Seafaring in the 21st century: The malaspina 2010 circumnavigation expedition," *Limnology and Oceanography Bulletin*, vol. 24, no. 1, pp. 11–14, 2015, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/lob.10008, ISSN: 1539-6088. DOI: 10.1002/lob.10008. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/lob.10008 (visited on Aug. 28, 2022).

[10] I. H. Sarker, "Machine learning: Algorithms, real-world applications and research directions," *SN Computer Science*, vol. 2, no. 3, p. 160, Mar. 22, 2021, ISSN: 2661-8907. DOI: 10.1007/s42979-021-00592-x. [Online]. Available: https://doi.org/10.1007/s42979-021-00592-x (visited on Aug. 28, 2022).

[11] A. L'Heureux, K. Grolinger, H. El Yamany, and M. Capretz, "Machine learning with big data: Challenges and approaches," *IEEE Access*, vol. PP, pp. 1–1, Apr. 24, 2017. DOI: 10.1109/ACCESS.2017.2696365.

[12] T. A. Klevjer, X. Irigoien, A. Røstad, E. Fraile-Nuez, V. M. Benítez-Barrios, and S. Kaartvedt., "Large scale patterns in vertical distribution and behaviour of mesopelagic scattering layers," *Scientific Reports*, vol. 6, no. 1, p. 19 873, Jan. 27, 2016, ISSN: 2045-2322. DOI: 10.1038/srep19873. [Online]. Available: https://www.nature.com/articles/srep19873 (visited on Nov. 20, 2021).

[13] X. Irigoien, T. A. Klevjer, A. Røstad, U. Martinez, G. Boyra, J. L. Acuña, A. Bode, F. Echevarria, J. I. Gonzalez-Gordillo, S. Hernandez-Leon, S. Agusti, D. L. Aksnes, C. M. Duarte, and S. Kaartvedt, "Large mesopelagic fishes biomass and trophic efficiency in the open ocean," *Nature Communications*, vol. 5, no. 1, p. 3271, Feb. 7, 2014, ISSN: 2041-1723. DOI: 10.1038/ncomms4271. [Online]. Available: https://www.nature.com/articles/ncomms4271 (visited on Jul. 7, 2021).

[14] X. Irigoien, T. Klevjer, U. Martinez, G. Boyra, A. Røstad, A. C. Wittmann, C. M. Duarte, S. Kaartvedt, A. S. Brierley, and R. Proud, "The simrad EK60 echosounder dataset from the malaspina circumnavigation," *Scientific Data*, vol. 8, no. 1, p. 259, Oct. 1, 2021, Number: 1 Publisher: Nature Publishing Group, ISSN: 2052-4463. DOI: 10.1038/s41597-021-01038-y. [Online]. Available: https://www.nature.com/articles/s41597-021-01038-y (visited on Jun. 30, 2022).

[15] P. Prihartato, X. Irigoien, M. Genton, and S. Kaartvedt, "Global effects of moon phase on nocturnal acoustic scattering layers," *Marine Ecology Progress Series*, vol. 544, Feb. 18, 2016. DOI: 10.3354/meps11612.

[16] D. L. Aksnes, A. Røstad, S. Kaartvedt, U. Martinez, C. M. Duarte, and X. Irigoien, "Light penetration structures the deep acoustic scattering layers in the global ocean," *Science Advances*, vol. 3, no. 5, e1602468, May 31, 2017, Publisher: American Association for the Advancement of Science. DOI: 10.1126/sciadv.1602468. [Online]. Available: https://www.science.org/doi/10.1126/sciadv.1602468 (visited on Jun. 30, 2022).

[17] S. Hernández-León, R. Koppelmann, E. Fraile-Nuez, A. Bode, C. Mompeán, X. Irigoien, M. P. Olivar, F. Echevarría, M. L. Fernández de Puelles, J. I. González-Gordillo, A. Cózar, J. L. Acuña, S. Agustí, and C. M. Duarte, "Large deep-sea zooplankton biomass mirrors primary production in the global ocean," *Nature Communications*, vol. 11, no. 1, p. 6048, Nov. 27, 2020, Number: 1 Publisher: Nature Publishing Group, ISSN: 2041-1723. DOI: 10.1038/s41467-020-19875-7. [Online]. Available: https://www.nature.com/articles/s41467-020-19875-7 (visited on Aug. 16, 2022).

[18] P. C. Davison, J. A. Koslow, and R. J. Kloser, "Acoustic biomass estimation of mesopelagic fish: Backscattering from individuals, populations, and communities," *ICES Journal of Marine Science*, vol. 72, no. 5, pp. 1413–1424, Jun. 1, 2015, ISSN: 1054-3139. DOI: 10.1093/icesjms/fsv023. [Online]. Available: https://doi.org/10.1093/icesjms/fsv023 (visited on Jul. 6, 2021).

[19] S. McClatchie and A. Dunford, "Estimated biomass of vertically migrating mesopelagic fish off new zealand," *Deep Sea Research Part I: Oceanographic Research Papers*, vol. 50, no. 10, pp. 1263–1281, Oct. 1, 2003, ISSN: 0967-0637. DOI: 10.1016/S0967-0637(03)00128-6. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0967063703001286 (visited on Jun. 30, 2022).

[20] J. Gjøsaeter, K. Kawaguchi, and F. Ã. O. o. t. U. Nations, *A Review of the World Resources of Mesopelagic Fish*. Food & Agriculture Org., 1980, 164 pp., Google-Books-ID: Zw0A_veImj8C, ISBN: 978-92-5-100924-6.

[21] S. Kaartvedt, A. Staby, and D. L. Aksnes, "Efficient trawl avoidance by mesopelagic fishes causes large underestimation of their biomass," *Marine Ecology Progress Series*, vol. 456, pp. 1–6, Jun. 7, 2012, ISSN: 0171-8630, 1616-1599. DOI: 10.3354/meps09785. [Online]. Available: https://www.int-res.com/abstracts/meps/v456/p1-6/ (visited on Jun. 30, 2022).

[22] T. Dornan, S. Fielding, R. A. Saunders, and M. J. Genner, "Swimbladder morphology masks southern ocean mesopelagic fish biomass," *Proceedings of the Royal Society B: Biological Sciences*, vol. 286, no. 1903, p. 20 190 353, May 29, 2019, Publisher: Royal Society. DOI: 10.1098/rspb.2019.0353. [Online]. Available: https://royalsocietypublishing.org/doi/10.1098/rspb.2019.0353 (visited on Sep. 22, 2021).

[23] G. Wright, "Fishing in the twilight zone: Illuminating governance challenges at the next fisheries frontier," p. 26,

[24] S. Paoletti, J. R. Nielsen, C. R. Sparrevohn, F. Bastardie, and B. M. J. Vastenhoud, "Potential for mesopelagic fishery compared to economy and fisheries dynamics in current large scale danish pelagic fishery," *Frontiers in Marine Science*, vol. 8, p. 1145, 2021, ISSN: 2296-7745. DOI: 10.3389/fmars.2021.720897. [Online]. Available: https://www.frontiersin.org/article/10.3389/fmars.2021.720897 (visited on Nov. 20, 2021).

[25] C. Robinson, D. Steinberg, T. Anderson, J. Aristegui, C. Carlson, J. Frost, J.-F. Ghiglione, S. Hernandez-Leon, G. Jackson, R. Koppelmann, B. Quéguiner, O. Ragueneau, F. Rassoulzadegan, B. Robison, C. Tamburini, T. Tanaka, K. Wishner, and J. Zhang, "Mesopelagic zone ecology and biogeochemistry - a synthesis," *Deep Sea Research Part II Topical Studies in Oceanography*, vol. 57, pp. 1504–1518, Mar. 7, 2010. DOI: 10.1016/j.dsr2.2010.02.018.

[26] M. J. Costello and S. Breyer, "Ocean depths: The mesopelagic and implications for global warming," *Current Biology*, vol. 27, no. 1, R36–R38, Jan. 9, 2017, ISSN: 0960-9822. DOI: 10.1016/j.cub.2016.11.042. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0960982216313951 (visited on Dec. 16, 2021).

[27] D. Bianchi, C. Stock, E. D. Galbraith, and J. L. Sarmiento, "Diel vertical migration: Ecological controls and impacts on the biological pump in a one-dimensional ocean model," *Global Biogeochemical Cycles*, vol. 27, no. 2, pp. 478–491, 2013, ISSN: 1944-9224. DOI: 10.1002/gbc.20031. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/gbc.20031 (visited on Dec. 16, 2021).

[28] E. D. Cotter and A. C. Lavery, "Categorization of broadband spectra of mesopelagic targets using model-generated training data," *The Journal of the Acoustical Society of America*, vol. 148, no. 4, pp. 2662–2662, Oct. 2020, Publisher: Acoustical Society of America, ISSN: 0001-4966. DOI: 10.1121/1.5147418. [Online]. Available: https://asa.scitation.org/doi/abs/10.1121/1.5147418 (visited on Jun. 30, 2022).

[29] E. Cotter, C. Bassett, and A. Lavery, "Classification of broadband target spectra in the mesopelagic using physics-informed machine learning," *The Journal of the Acoustical Society of America*, vol. 149, no. 6, pp. 3889–3901, Jun. 1, 2021, Publisher: Acoustical Society of America, ISSN: 0001-4966. DOI: 10.1121/10.0005114. [Online]. Available: https://asa.scitation.org/doi/full/10.1121/10.0005114 (visited on Dec. 16, 2021).

[30] K. J. Benoit-Bird and C. M. Waluk, "Exploring the promise of broadband fisheries echosounders for species discrimination with quantitative assessment of data processing effects," *The Journal of the Acoustical Society of America*, vol. 147, no. 1, pp. 411–427,

Jan. 2020, Publisher: Acoustical Society of America, ISSN: 0001-4966. DOI: 10.1121/10.0000594. [Online]. Available: https://asa.scitation.org/doi/10.1121/10.0000594 (visited on Jun. 30, 2022).

[31] D. Gong, "Connecting ocean physical and biogeochemical properties with the spatial distribution of mesopelagic fish abundance," p. 1,

[32] I. Ali, F. Cawkwell, E. Dwyer, and S. Green, "Modeling managed grassland biomass estimation by using multitemporal remote sensing data—a machine learning approach," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 7, pp. 3254–3264, Jul. 2017, Conference Name: IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, ISSN: 2151-1535. DOI: 10.1109/JSTARS.2016.2561618.

[33] Y. Zhang, J. Ma, S. Liang, X. Li, and M. Li, "An evaluation of eight machine learning regression algorithms for forest aboveground biomass estimation from multiple satellite data products," *Remote Sensing*, vol. 12, no. 24, p. 4015, Jan. 2020, Number: 24 Publisher: Multidisciplinary Digital Publishing Institute. DOI: 10.3390/rs12244015. [Online]. Available: https://www.mdpi.com/2072-4292/12/24/4015 (visited on Sep. 22, 2021).

[34] S. M. Ghosh and M. D. Behera, "Aboveground biomass estimation using multi-sensor data synergy and machine learning algorithms in a dense tropical forest," *Applied Geography*, vol. 96, pp. 29–40, Jul. 1, 2018, ISSN: 0143-6228. DOI: 10.1016/j.apgeog.2018.05.011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0143622818303114 (visited on Jul. 8, 2022).

[35] A. T. N. Dang, S. Nandy, R. Srinet, N. V. Luong, S. Ghosh, and A. Senthil Kumar, "Forest aboveground biomass estimation using machine learning regression algorithm in yok don national park, vietnam," *Ecological Informatics*, vol. 50, pp. 24–32, Mar. 1, 2019, ISSN: 1574-9541. DOI: 10.1016/j.ecoinf.2018.12.010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1574954118301894 (visited on Jul. 8, 2022).

[36] A. Galloway, D. Brunet, R. Valipour, M. McCusker, J. Biberhofer, M. K. Sobol, M. Moussa, and G. W. Taylor, "Predicting dreissenid mussel abundance in nearshore waters using underwater imagery and deep learning," *Limnology and Oceanography: Methods*, vol. 20, no. 4, pp. 233–248, 2022, ISSN: 1541-5856. DOI: 10.1002/lom3.10483. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/lom3.10483 (visited on Jul. 9, 2022).

[37] D. Precioso, M. Navarro-García, K. Gavira-O'Neill, A. Torres-Barrán, D. Gordo, V. Gallego, and D. Gómez-Ullate, "TUN-AI: Tuna biomass estimation with machine learning models trained on oceanography and echosounder FAD data," *Fisheries Research*, vol. 250, p. 106 263, Jun. 1, 2022, ISSN: 0165-7836. DOI: 10.1016/j.fishres.2022.106263. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165783622000406 (visited on Jul. 9, 2022).

[38] H. Mohamed, K. Nadaoka, and T. Nakamura, "Assessment of machine learning algorithms for automatic benthic cover monitoring and mapping using towed underwater video camera and high-resolution satellite images," *Remote Sensing*, vol. 10, no. 5, p. 773, May 2018, Number: 5 Publisher: Multidisciplinary Digital Publishing Institute, ISSN: 2072-4292. DOI: 10.3390/rs10050773. [Online]. Available: https://www.mdpi.com/2072-4292/10/5/773 (visited on Jul. 9, 2022).

[39] S. Kaartvedt, A. Røstad, T. Klevjer, and A. Staby, "Use of bottom-mounted echo sounders in exploring behavior of mesopelagic fishes," *Marine Ecology Progress Series*, vol. 395, pp. 109–118, Dec. 3, 2009. DOI: 10.3354/meps08174.

[40] F. I. Petrescu, *A New Doppler Effect*. Nov. 5, 2012, ISBN: 978-3-8482-2990-1. DOI: 10.13140/RG.2.1.1142.8560.

[41] Norwegian Institute of Water Research, *LEHMKUHL FerryBox manual 2021*, Aug. 20, 2021.

[42] F. E. Hoge, A. Vodacek, R. N. Swift, J. K. Yungel, and N. V. Blough, "Inherent optical properties of the ocean: Retrieval of the absorption coefficient of chromophoric dissolved organic matter from airborne laser spectral fluorescence measurements," *Applied Optics*, vol. 34, no. 30, pp. 7032–7038, Oct. 20, 1995, Publisher: Optica Publishing Group, ISSN: 2155-3165. DOI: 10.1364/AO.34.007032. [Online]. Available: https://opg.optica.org/ao/abstract.cfm?uri=ao-34-30-7032 (visited on Jun. 26, 2022).

[43] D. J. Suggett, O. Prášil, and M. A. Borowitzka, Eds., *Chlorophyll a Fluorescence in Aquatic Sciences: Methods and Applications*, Dordrecht: Springer Netherlands, 2010, ISBN: 978-90-481-9267-0 978-90-481-9268-7. DOI: 10.1007/978-90-481-9268-7. [Online]. Available: http://link.springer.com/10.1007/978-90-481-9268-7 (visited on Jun. 26, 2022).

[44] G. Öquist, Å. Hagström, P. Alm, G. Samuelsson, and K. Richardson, "Chlorophyll a fluorescence, an alternative method for estimating primary production," *Marine Biology*, vol. 68, no. 1, pp. 71–75, May 1, 1982, ISSN: 1432-1793. DOI: 10.1007/BF00393143. [Online]. Available: https://doi.org/10.1007/BF00393143 (visited on Jun. 26, 2022).

[45] Meike Becker and T. O. Kristensen, *pCO2 instrument statsraad lehmkuhl*, Jun. 17, 2021.

[46] D. Standal and E. Grimaldo, "Institutional nuts and bolts for a mesopelagic fishery in norway," *Marine Policy*, vol. 119, p. 104 043, Sep. 1, 2020, ISSN: 0308-597X. DOI: 10.1016/j.marpol.2020.104043. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0308597X19309649 (visited on Nov. 20, 2021).

[47] (). Shedding light on the ocean's "twilight zone", Food and Agriculture Organization of the United Nations, [Online]. Available: http://www.fao.org/fao-stories/article/en/c/1364889/ (visited on Dec. 16, 2021).

[48] (). Maurolicus muelleri - artsdatabanken, [Online]. Available: https://www.artsdatabanken.no/taxon/Maurolicus%20muelleri/42698 (visited on Sep. 1, 2022).

[49] (). Benthosema glaciale - artsdatabanken, [Online]. Available: https://www.artsdatabanken.no/taxon/Benthosema%20glaciale/42708 (visited on Sep. 1, 2022).

[50] R. Proud, M. J. Cox, and A. S. Brierley, "Biogeography of the global ocean's mesopelagic zone," *Current Biology*, vol. 27, no. 1, pp. 113–119, Jan. 9, 2017, ISSN: 0960-9822. DOI: 10.1016/j.cub.2016.11.003. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0960982216313288 (visited on Feb. 13, 2022).

[51] A. Burkov, "The hundred-page machine learning book," p. 152,

[52] A. Géron, "Hands-on machine learning with scikit-learn, keras, and TensorFlow," p. 851,

[53] K. El Bouchefry and R. S. de Souza, "Chapter 12 - learning in big data: Introduction to machine learning," in *Knowledge Discovery in Big Data from Astronomy and Earth Observation*, P. Škoda and F. Adam, Eds., Elsevier, Jan. 1, 2020, pp. 225–249, ISBN: 978-0-12-819154-5. DOI: 10.1016/B978-0-12-819154-5.00023-0. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128191545000230 (visited on Jun. 27, 2022).

[54] N. Al-Azzam and I. Shatnawi, "Comparing supervised and semi-supervised machine learning models on diagnosing breast cancer," *Annals of Medicine and Surgery*, vol. 62, pp. 53–64, Feb. 1, 2021, ISSN: 2049-0801. DOI: 10.1016/j.amsu.2020.12.043. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2049080120305604 (visited on Jun. 27, 2022).

[55] O. Chapelle, B. Schölkopf, and A. Zien, Eds., *Semi-supervised learning*, Adaptive computation and machine learning, OCLC: ocm64898359, Cambridge, Mass: MIT Press, 2006, 508 pp., ISBN: 978-0-262-03358-9.

[56] S. D. Holcomb, W. K. Porter, S. V. Ault, G. Mao, and J. Wang, "Overview on DeepMind and its AlphaGo zero AI," in *Proceedings of the 2018 International Conference on Big Data and Education*, ser. ICBDE '18, New York, NY, USA: Association for Computing Machinery, Mar. 9, 2018, pp. 67–71, ISBN: 978-1-4503-6358-7. DOI: 10.1145/3206157.3206174. [Online]. Available: https://doi.org/10.1145/3206157.3206174 (visited on Jun. 27, 2022).

[57] OpenAI, C. Berner, G. Brockman, B. Chan, V. Cheung, P. Debiak, C. Dennison, D. Farhi, Q. Fischer, S. Hashme, C. Hesse, R. Józefowicz, S. Gray, C. Olsson, J. Pachocki, M. Petrov, H. P. d. O. Pinto, J. Raiman, T. Salimans, J. Schlatter, J. Schneider, S. Sidor, I. Sutskever, J. Tang, F. Wolski, and S. Zhang, *Dota 2 with large scale deep reinforcement learning*, Number: arXiv:1912.06680, Dec. 13, 2019. DOI: 10.48550/arXiv.1912.06680. arXiv: 1912.06680[cs,stat]. [Online]. Available: http://arxiv.org/abs/1912.06680 (visited on Jun. 27, 2022).

[58] K. P. Murphy. (2022). Probabilistic machine learning: An introduction (adaptive computation and machine learning series): Murphy, kevin p.: 9780262046824: Amazon.com: Books, [Online]. Available: https://www.amazon.com/Probabilistic-Machine-Learning-Introduction-Computation/dp/0262046822 (visited on Aug. 28, 2022).

[59] (). Overfitting and underfitting, Educative: Interactive Courses for Software Developers, [Online]. Available: https://www.educative.io/answers/overfitting-and-underfitting (visited on Sep. 1, 2022).

[60] S. Geman, E. Bienenstock, and R. Doursat, "Neural networks and the bias/variance dilemma," *Neural Computation*, vol. 4, pp. 1–58, Jan. 1, 1992. DOI: 10.1162/neco.1992.4.1.1.

[61] D. S. Moore and G. P. McCabe, *Introduction to the practice of statistics*, ser. Introduction to the practice of statistics. New York, NY, US: W H Freeman/Times Books/ Henry Holt & Co, 1989, xix, 790, Pages: xix, 790, ISBN: 978-0-7167-1989-2.

[62] D. Rajnarayan and D. Wolpert, "Bias-variance tradeoffs: Novel applications," Nov. 6, 2008. DOI: 10.1007/978-0-387-30164-8_75.

[63] "Bias variance decomposition," in *Encyclopedia of Machine Learning and Data Mining*, C. Sammut and G. I. Webb, Eds., Boston, MA: Springer US, 2017, pp. 128–129, ISBN: 978-1-4899-7687-1. DOI: 10.1007/978-1-4899-7687-1_74. [Online]. Available: https://doi.org/10.1007/978-1-4899-7687-1_74 (visited on Aug. 28, 2022).

[64] A. Zheng. (). Evaluating machine learning models : A beginner's guide to key concepts and pitfalls [PDF] [50h6j99mgql0], [Online]. Available: https://vdoc.pub/documents/evaluating-machine-learning-models-a-beginners-guide-to-key-concepts-and-pitfalls-50h6j99mgql0 (visited on Jun. 28, 2022).

[65] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination r-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *PeerJ Computer Science*, vol. 7, e623, Jul. 5, 2021, Publisher: PeerJ Inc., ISSN: 2376-5992. DOI: 10.7717/peerj-cs.623. [Online]. Available: https://peerj.com/articles/cs-623 (visited on May 19, 2022).

[66] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)? – arguments against avoiding RMSE in the literature," *Geoscientific Model Development*, vol. 7, no. 3, pp. 1247–1250, Jun. 30, 2014, ISSN: 1991-9603. DOI: 10.5194/gmd-7-1247-2014. [Online]. Available: https://gmd.copernicus.org/articles/7/1247/2014/ (visited on Jun. 28, 2022).

[67] G. Brassington, "Mean absolute error and root mean square error: Which is the better metric for assessing model performance?," p. 3574, Apr. 1, 2017, Conference Name: EGU General Assembly Conference Abstracts ADS Bibcode: 2017EGUGA..19.3574B. [Online]. Available: https://ui.adsabs.harvard.edu/abs/2017EGUGA..19.3574B (visited on Jul. 21, 2022).

[68] C. J. Willmott and K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance," *Climate Research*, vol. 30, no. 1, pp. 79–82, Dec. 19, 2005, ISSN: 0936-577X, 1616-1572. DOI: 10.3354/cr030079. [Online]. Available: https://www.int-res.com/abstracts/cr/v30/n1/p79-82/ (visited on May 19, 2022).

[69] T. O. Hodson, "Root-mean-square error (RMSE) or mean absolute error (MAE): When to use them or not," *Geoscientific Model Development*, vol. 15, no. 14, pp. 5481–5487, Jul. 19, 2022, ISSN: 1991-9603. DOI: 10.5194/gmd-15-5481-2022. [Online]. Available: https://gmd.copernicus.org/articles/15/5481/2022/ (visited on Jul. 21, 2022).

[70] M. Kuhn and K. Johnson, *Feature Engineering and Selection: A Practical Approach for Predictive Models*. CRC Press, Jul. 25, 2019, 314 pp., Google-Books-ID: xy73DwAAQBAJ, ISBN: 978-1-351-60947-0.

[71] T. Kvalseth, "Cautionary note about r2," *The American Statistician*, vol. 39, pp. 279–285, Mar. 12, 2012. DOI: 10.1080/00031305.1985.10479448.

[72] F. Moksony, "Small is beautiful. the use and interpretation of r2 in social research," *Szociológiai Szemle*, Jan. 1, 1999. [Online]. Available: https://www.academia.edu/3880005/Small_is_beautiful_The_use_and_interpretation_of_R2_in_social_research (visited on Jun. 28, 2022).

[73] C. Onyutha, *From R-squared to coefficient of model accuracy for assessing "goodness-of-fits"*. May 4, 2020. DOI: 10.5194/gmd-2020-51.

[74] S. Kim and H. Kim, "A new metric of absolute percentage error for intermittent demand forecasts," *International Journal of Forecasting*, vol. 32, pp. 669–679, Jul. 1, 2016. DOI: 10.1016/j.ijforecast.2015.12.003.

[75] P. Goodwin and R. Lawton, "On the asymmetry of the symmetric MAPE," *International Journal of Forecasting*, vol. 15, no. 4, pp. 405–408, Oct. 1, 1999, ISSN: 0169-2070. DOI: 10.1016/S0169-2070(99)00007-2. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169207099000072 (visited on Jun. 28, 2022).

[76] T. Jo, *Machine Learning Foundations: Supervised, Unsupervised, and Advanced Learning*. Cham: Springer International Publishing, 2021, ISBN: 978-3-030-65899-1 978-3-030-65900-4. DOI: 10.1007/978-3-030-65900-4. [Online]. Available: http://link.springer.com/10.1007/978-3-030-65900-4 (visited on Jul. 11, 2022).

[77] M. Ebrahimi Kalan, R. Jebai, E. Zarafshan, and Z. Bursac, "Distinction between two statistical terms: Multivariable and multivariate logistic regression," *Nicotine & Tobacco Research*, vol. 23, no. 8, pp. 1446–1447, Aug. 1, 2021, ISSN: 1469-994X. DOI: 10.1093/ntr/ntaa055. [Online]. Available: https://doi.org/10.1093/ntr/ntaa055 (visited on Aug. 28, 2022).

[78] G. Rebala, A. Ravi, and S. Churiwala, *An Introduction to Machine Learning*. Cham: Springer International Publishing, 2019, ISBN: 978-3-030-15728-9 978-3-030-15729-6. DOI: 10.1007/978-3-030-15729-6. [Online]. Available: http://link.springer.com/10.1007/978-3-030-15729-6 (visited on Jul. 11, 2022).

[79] V. K. Ayyadevara, *Pro Machine Learning Algorithms*. Berkeley, CA: Apress, 2018, ISBN: 978-1-4842-3563-8 978-1-4842-3564-5. DOI: 10.1007/978-1-4842-3564-5. [Online]. Available: http://link.springer.com/10.1007/978-1-4842-3564-5 (visited on Jul. 12, 2022).

[80] G. Hackeling, *Mastering Machine Learning with scikit-learn - Second Edition*. Birmingham, UNITED KINGDOM: Packt Publishing, Limited, 2017, ISBN: 978-1-78829-849-0. [Online]. Available: http://ebookcentral.proquest.com/lib/bergen-ebooks/detail.action?docID=4925640 (visited on Jul. 12, 2022).

[81] T. M. V. Suryanarayana and P. B. Mistry, "Principal component analysis in transfer function," in *Principal Component Regression for Crop Yield Estimation*, ser. SpringerBriefs in Applied Sciences and Technology, T. Suryanarayana and P. B. Mistry, Eds., Singapore: Springer, 2016, pp. 17–25, ISBN: 978-981-10-0663-0. DOI: 10.1007/978-981-10-0663-0_2. [Online]. Available: https://doi.org/10.1007/978-981-10-0663-0_2 (visited on Jul. 12, 2022).

[82] S. Maitra and J. Yan, "Principle component analysis and partial least squares: Two dimension reduction techniques for regression," p. 12, 2008.

[83] Z.-H. Zhou, *Ensemble Methods: Foundations and Algorithms*. London, UNITED KINGDOM: CRC, 2012, ISBN: 978-1-4398-3005-5. [Online]. Available: http://ebookcentral.proquest.com/lib/bergen-ebooks/detail.action?docID=952003 (visited on Jul. 12, 2022).

[84] L. Rokach, *Pattern Classification Using Ensemble Methods*. Singapore, SINGAPORE: World Scientific Publishing Company, 2009, ISBN: 978-981-4271-07-3. [Online]. Available: http://ebookcentral.proquest.com/lib/bergen-ebooks/detail.action?docID=1679487 (visited on Jul. 12, 2022).

[85] V. B. Vaghela, A. Ganatra, and A. Thakkar, "Boost a weak learner to a strong learner using ensemble system approach," in *2009 IEEE International Advance Computing Conference*, Mar. 2009, pp. 1432–1436. DOI: 10.1109/IADCC.2009.4809227.

[86] Y. Zhao and Y. Zhang, "Comparison of decision tree methods for finding active objects," *Advances in Space Research*, vol. 41, no. 12, pp. 1955–1959, Jan. 1, 2008, ISSN: 0273-1177. DOI: 10.1016/j.asr.2007.07.020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S027311770700796X (visited on Jul. 12, 2022).

[87] E. V. A. Sylvester, P. Bentzen, I. R. Bradbury, M. Clément, J. Pearce, J. Horne, and R. G. Beiko, "Applications of random forest feature selection for fine-scale genetic population assignment," *Evolutionary Applications*, vol. 11, no. 2, pp. 153–165, 2018, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/eva.12524, ISSN: 1752-4571. DOI: 10.1111/eva.12524. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/eva.12524 (visited on Jul. 12, 2022).

[88] L. Breiman. (2001). Random forests — SpringerLink, [Online]. Available: https://link.springer.com/article/10.1023/A:1010933404324 (visited on Jul. 19, 2022).

[89] A. L. Blum and P. Langley, "Selection of relevant features and examples in machine learning," *Artificial Intelligence*, Relevance, vol. 97, no. 1, pp. 245–271, Dec. 1, 1997, ISSN: 0004-3702. DOI: 10.1016/S0004-3702(97)00063-5. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0004370297000635 (visited on Jun. 28, 2022).

[90] (). Python machine learning - third edition — packt, [Online]. Available: https://www.packtpub.com/product/python-machine-learning-third-edition/9781789955750 (visited on Jul. 22, 2022).

[91] M. Loecher, "Unbiased variable importance for random forests," *Communications in Statistics - Theory and Methods*, vol. 51, no. 5, pp. 1413–1425, Mar. 4, 2022, Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/03610926.2020.1764042, ISSN: 0361-0926. DOI: 10.1080/03610926.2020.1764042. [Online]. Available: https://doi.org/10.1080/03610926.2020.1764042 (visited on Jul. 22, 2022).

[92] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn, "Bias in random forest variable importance measures: Illustrations, sources and a solution," *BMC Bioinformatics*, vol. 8, no. 1, p. 25, Jan. 25, 2007, ISSN: 1471-2105. DOI: 10.1186/1471-2105-8-25. [Online]. Available: https://doi.org/10.1186/1471-2105-8-25 (visited on Jul. 23, 2022).

[93] C. Molnar, *Interpretable Machine Learning*. [Online]. Available: https://christophm.github.io/interpretable-ml-book/index.html (visited on Jul. 23, 2022).

[94] L. Yang and A. Shami, "On hyperparameter optimization of machine learning algorithms: Theory and practice," *Neurocomputing*, vol. 415, pp. 295–316, Nov. 20, 2020, ISSN: 0925-2312. DOI: 10.1016/j.neucom.2020.07.061. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231220311693 (visited on Aug. 28, 2022).

[95] J. F. Pimentel, L. Murta, V. Braganholo, and J. Freire, "A large-scale study about quality and reproducibility of jupyter notebooks," in *2019 IEEE/ACM 16th International Conference on Mining Software Repositories (MSR)*, ISSN: 2574-3864, May 2019, pp. 507–517. DOI: 10.1109/MSR.2019.00077.

[96] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, and D. Cournapeau, "Scikit-learn: Machine learning in python," *MACHINE LEARNING IN PYTHON*, p. 6,

[97] W. McKinney, "Pandas: A foundational python library for data analysis and statistics," p. 9,

[98] A. Hevner and S. Chatterjee, *Design Research in Information Systems*, ser. Integrated Series in Information Systems. Boston, MA: Springer US, 2010, vol. 22, ISBN: 978-1-4419-5652-1 978-1-4419-5653-8. DOI: 10.1007/978-1-4419-5653-8. [Online]. Available: http://link.springer.com/10.1007/978-1-4419-5653-8 (visited on Dec. 9, 2021).

[99] A. Hevner, A. R, S. March, S. T, Park, J. Park, Ram, and Sudha, "Design science in information systems research," *Management Information Systems Quarterly*, vol. 28, p. 75, Mar. 1, 2004.

[100] A. Hevner, "A three cycle view of design science research," *Scandinavian Journal of Information Systems*, vol. 19, Jan. 1, 2007.

[101] K. A. Piirainen and R. A. Gonzalez, "Seeking constructive synergy: Design science and the constructive research approach," in *Design Science at the Intersection of Physical and Virtual Design*, J. vom Brocke, R. Hekkala, S. Ram, and M. Rossi, Eds., red. by D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, and G. Weikum, vol. 7939, Series Title: Lecture Notes in Computer Science, Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 59–72, ISBN: 978-3-642-38826-2 978-3-642-38827-9. DOI: 10.1007/978-3-642-38827-9_5. [Online]. Available: http://link.springer.com/10.1007/978-3-642-38827-9_5 (visited on Aug. 16, 2022).

[102] (). 7 steps to machine learning: How to prepare for an automated future — by dr mark van rijmenam — DataSeries — medium, [Online]. Available: https://medium.com/dataseries/7-steps-to-machine-learning-how-to-prepare-for-an-automated-future-78c7918cb35d (visited on Aug. 27, 2022).

[103] Y. Roh, G. Heo, and S. E. Whang, *A survey on data collection for machine learning: A big data – AI integration perspective*, Aug. 12, 2019. arXiv: 1811.03402[cs,stat]. [Online]. Available: http://arxiv.org/abs/1811.03402 (visited on Aug. 23, 2022).

[104] S. Bhattacharya, *A Primer on Machine Learning in Subsurface Geosciences*, ser. SpringerBriefs in Petroleum Geoscience & Engineering. Cham: Springer International Publishing, 2021, ISBN: 978-3-030-71767-4 978-3-030-71768-1. DOI: 10.1007/978-3-030-71768-1.

[Online]. Available: https://link.springer.com/10.1007/978-3-030-71768-1 (visited on Aug. 23, 2022).
[105] A. Bhatia and B. Kaluza. (2018). Machine learning in java - second edition [book]. ISBN: 9781788474399, [Online]. Available: https://www.oreilly.com/library/view/machine-learning-in/9781788474399/ (visited on Jun. 29, 2022).
[106] (). MAREC, [Online]. Available: https://www.marec.no/ (visited on Aug. 28, 2022).
[107] *This repository contain the code to preprocess acoustic data for the CRIMAC project*, original-date: 2020-07-31T09:06:17Z, Jun. 29, 2022. [Online]. Available: https://github.com/CRIMAC-WP4-Machine-learning/CRIMAC-preprocessing (visited on Aug. 28, 2022).
[108] D. MacLennan and P. Fernandes. (2002). A consistent approach to definitions and symbols in fisheries acoustics, [Online]. Available: https://www.researchgate.net/publication/228787678_A_consistent_approach_to_definitions_and_symbols_in_fisheries_acoustics (visited on Aug. 27, 2022).
[109] E. L. Lewis and R. G. Perkin, "Salinity: Its definition and calculation," *Journal of Geophysical Research: Oceans*, vol. 83, pp. 466–478, C1 1978, ISSN: 2156-2202. DOI: 10.1029/JC083iC01p00466. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1029/JC083iC01p00466 (visited on Jun. 29, 2022).
[110] N. US Department of Commerce. (). NWS JetStream - sea water. Publisher: NOAA's National Weather Service, [Online]. Available: https://www.weather.gov/jetstream/seawater (visited on Jun. 29, 2022).
[111] (Mar. 16, 2014). PMEL - pacific marine environmental laboratory, NOAA Pacific Marine Environmental Laboratory (PMEL), [Online]. Available: https://www.pmel.noaa.gov/data-links (visited on Aug. 31, 2022).
[112] (). LAS UI, [Online]. Available: https://data.pmel.noaa.gov/socat/las/UI.vm#panelHeaderHidden=false;differences=false;autoContour=false;xCATID=SOCATv2022_ERDDAP;xDSID=socat_v2022_fulldata;varid=fCO2_recommended-socat_v2022_fulldata;imageSize=auto;over=xy;compute=None;constraintCount=3;constraint0=text_cr_none_cr_none_cr_WOCE_CO2_water_cr_2_cr_WOCE_CO2_water_cr_2_cr_eq_cr_;constraint1=text_cr_none_cr_none_cr_platform_name_cr_Statsraad%20Lehmkuhl_cr_platform_name_cr_Statsraad%20Lehmkuhl_cr_eq_cr_;constraint2=variable_cr_socat_v2022_fulldata_cr_fCO2_recommended-socat_v2022_fulldata_cr_fCO2_recommended_cr_NaN_cr_fCO2_recommended_cr_NaN_cr_ne_cr_;constraintPanelIndex=0token;catid=SOCATv2022_ERDDAP;dsid=socat_v2022_fulldata;varid=fCO2_recommended-socat_v2022_fulldata;avarcount=0;xlo=-180;xhi=180;ylo=-80;yhi=90;tlo=01-Jan-1957%2000:00;thi=31-Dec-2021%2000:00;operation_id=Trajectory_interactive_plot;view=xyt (visited on Aug. 30, 2022).
[113] H. Chen and R. Wanninkhof, "Measurement of fugacity of carbon dioxide in seawater: An evaluation of a method based on infrared analsysis," Aug. 1995. [Online]. Available: https://www.aoml.noaa.gov/ocd/gcc/AOML85.pdf.
[114] (). Global relationships of total alkalinity with salinity and temperature in surface waters of the world's oceans - lee - 2006 - geophysical research letters - wiley online library, [Online]. Available: https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2006GL027207 (visited on Aug. 30, 2022).
[115] M. P. Humphreys, A. J. Schiller, D. Sandborn, L. Gregor, D. Pierrot, S. M. A. C. van Heuven, E. R. Lewis, and D. W. R. Wallace, *PyCO2sys: Marine carbonate system calculations in python*, version v1.8.1, Language: en, May 18, 2022. DOI: 10.5281/ZENODO.

3744275. [Online]. Available: https://zenodo.org/record/3744275 (visited on Aug. 30, 2022).
[116] (). Sea surface temperature, [Online]. Available: https://earthobservatory.nasa.gov/global-maps/MYD28M (visited on Aug. 28, 2022).
[117] "Calibrating turbidity meters," p. 2,
[118] N. O. Ã. A. US Department of Commerce. (). NDBC - science education - what is air pressure? [Online]. Available: https://www.ndbc.noaa.gov/educate/pressure.shtml (visited on Aug. 28, 2022).
[119] (). Average annual relative humidity, Center for Sustainability and the Global Environment, [Online]. Available: https://sage.nelson.wisc.edu/data-and-models/atlas-of-the-biosphere/mapping-the-biosphere/ecosystems/average-annual-relative-humidity/ (visited on Aug. 28, 2022).
[120] (). Climate - average relative humidity — britannica, [Online]. Available: https://www.britannica.com/science/climate-meteorology/Average-relative-humidity (visited on Aug. 28, 2022).
[121] K. L. Smith, H. A. Ruhl, B. J. Bett, D. S. M. Billett, R. S. Lampitt, and R. S. Kaufmann, "Climate, carbon cycling, and deep-ocean ecosystems," *Proceedings of the National Academy of Sciences*, vol. 106, no. 46, pp. 19 211–19 218, Nov. 17, 2009, Publisher: Proceedings of the National Academy of Sciences. DOI: 10.1073/pnas.0908322106. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.0908322106 (visited on Jul. 19, 2022).
[122] (). The distribution and catch rates of deep water fish along the mid-atlantic ridge from 43 to 61n - ScienceDirect, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165783601002533 (visited on Jul. 19, 2022).
[123] T. Morato, S. D. Hoyle, V. Allain, and S. J. Nicol, "Seamounts are hotspots of pelagic biodiversity in the open ocean," *Proceedings of the National Academy of Sciences*, vol. 107, no. 21, pp. 9707–9711, May 25, 2010, Publisher: Proceedings of the National Academy of Sciences. DOI: 10.1073/pnas.0910290107. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.0910290107 (visited on Jul. 19, 2022).
[124] G. B. C. o. t. Oceans. (). GEBCO - the general bathymetric chart of the oceans, GEBCO, [Online]. Available: https://www.gebco.net/ (visited on Aug. 28, 2022).
[125] T. Sutton, M. Clark, D. Dunn, P. Halpin, A. Rogers, J. Guinotte, S. Bograd, M. Angel, J. Perez, K. Wishner, R. Haedrich, D. Lindsay, J. Drazen, A. Vereshchaka, U. Piatkowski, T. Morato, K. Blachowiak-Samolyk, B. Robison, K. Gjerde, and M. Heino, "A global biogeographic classification of the mesopelagic zone," *Deep Sea Research Part I: Oceanographic Research Papers*, vol. 126, May 1, 2017. DOI: 10.1016/j.dsr.2017.05.006.
[126] K. Banachewicz, L. Massaron, and A. Goldbloom, *The Kaggle Book: Data analysis and machine learning for competitive data science*. Packt Publishing Ltd, Apr. 22, 2022, 531 pp., Google-Books-ID: GAVsEAAAQBAJ, ISBN: 978-1-80181-221-4.
[127] A. Gholamy, V. Kreinovich, and O. Kosheleva, "Why 70/30 or 80/20 relation between training and testing sets: A pedagogical explanation," p. 7,
[128] P. Runeson and M. Höst, "Guidelines for conducting and reporting case study research in software engineering," *Empirical Software Engineering*, vol. 14, no. 2, pp. 131–164, Apr. 2009, ISSN: 1382-3256, 1573-7616. DOI: 10.1007/s10664-008-9102-8. [Online]. Available: http://link.springer.com/10.1007/s10664-008-9102-8 (visited on Jul. 30, 2022).
[129] n. Pinheiro, n. Belo, and n. Mota, "Air pressure effects on biomass yield of two different kluyveromyces strains," *Enzyme and Microbial Technology*, vol. 26, no. 9, pp. 756–762, Jun. 1, 2000, ISSN: 1879-0909. DOI: 10.1016/s0141-0229(00)00168-x.

[130] P. F. Thomsen and E. Willerslev, "Environmental DNA – an emerging tool in conservation for monitoring past and present biodiversity," *Biological Conservation*, Special Issue: Environmental DNA: A powerful new tool for biological conservation, vol. 183, pp. 4–18, Mar. 1, 2015, ISSN: 0006-3207. DOI: 10.1016/j.biocon.2014.11.019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0006320714004443 (visited on Aug. 11, 2022).