A data-driven approach for prediction and optimization of ship fuel consumption

Rafael Mesquita Lacerda Madureira

Thesis to obtain the Master of Science Degree in

Naval Architecture and Ocean Engineering

Supervisors: Professor Ângelo Manuel Palos Teixeira,
Professor João Ricardo Centeno da Costa

Examination Committee

Chairperson: Professor Yordan Garbatov

Supervisors: Professor Ângelo Palos Teixeira
Professor João Ricardo Centeno da Costa

Members of the Committee:

Professor Roberto Vettor
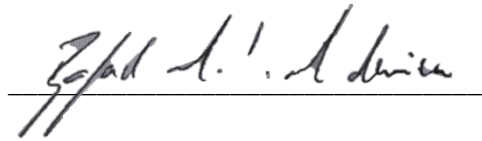
October 2021

TÉCNICO LISBOA

A data-driven approach for prediction and optimization of ship fuel consumption

Rafael Mesquita Lacerda Madureira

DECLARATION

I declare that this document is an original work of my own authorship and that it fulfils all the requirements of the Code of Conduct and Good Practices of the Universidade de Lisboa.

Signature

ACKNOWLEDGEMENTS

First, I want to thank my wife, Paula, who always supported and helped me in all phases of our life.

To my mother, who even from a distance, always gave us the strength to continue to fight for what we think is right.

To my friends and colleagues, who make life lighter, even more so in these last strange times we live in.

Finally, to professors Ângelo Teixeira and Ricardo Centeno for giving me the opportunity to develop a thesis on this topic. I learned new and interesting subjects and techniques that will certainly be useful to me in the future.

RESUMO

O aumento contínuo das emissões anuais de gases com efeito estufa pressiona todos os setores da indústria a reduzir as emissões, mudar para tecnologias verdes e se tornarem mais eficientes.

Na indústria naval existem diversos projetos e tecnologias para a redução das emissões pelos navios, tanto ao nível do projeto como da operação. Uma solução possível para navios existentes é a instalação de sistemas de otimização de combustível que atuam automaticamente nas rotações do motor e no passo da hélice de forma a aumentar a eficiência propulsiva do navio. Estes sistemas permitem a otimização do uso de combustível em rota em função de um conjunto de variáveis como a velocidade, os parâmetros do sistema propulsivo e as condições ambientais, que são monitorizadas de forma contínua ao longo das viagens.

O objetivo desta dissertação é desenvolver modelos de aprendizagem automática que representem o funcionamento de um sistema de otimização de combustível e desenvolver um protótipo de um sistema de apoio à decisão que forneça previsões do consumo ótimo de combustível da máquina principal do navio.

Para tal, é utilizada uma amostra de um ano de dados recolhidos de um sistema de otimização automática de combustivel de um navio, que incluem os parâmetros do sistema propulsivo, as condições ambientais e consumo de combustível do navio em operação.

Esta amostra de dados é primeiro analisada e pré processada e, depois, usada em tarefas de aprendizagem com algoritmos de Rede Neural Artificial e Máquinas de Vetores de Suporte. O desempenho dos algoritmos é avaliado e, posteriormente, é desenvolvido um modelo com dois estágios para prever a velocidade e consumo de combustível do navio em condição de operação.

Por fim, os modelos desenvolvidos são usados num sistema de apoio a decisão que é desenvolvido e demonstrado em diferentes cenários operacionais.

Palavras-chaves: consumo de combustível, otimização de combustível, aprendizagem automática, rede neural artificial, máquina de vetores de suporte, sistema de apoio a decisão

ABSTRACT

The continuous increase in annual greenhouse gas emissions pressures all industrial sectors to reduce emissions, move to green technologies, and become more efficient.

In the shipping industry, there are several projects and technologies for reducing emissions by ships, both at the design and operation levels. One possible solution for existing ships is the installation of fuel optimization systems that automatically adjust engine rotations and propeller pitch to increase the propulsive efficiency of the ship. These systems optimize the fuel consumption in the route as a function of a set of variables such as speed, propulsive system parameters, and environmental conditions, which are monitored continuously throughout the voyages.

The objective of this dissertation is to develop machine learning models that represent the operation of a fuel optimization system and to develop a prototype of a decision support system that provides predictions of the optimal fuel consumption of the ship's main engine.

For this purpose, a one-year sample of data collected from a ship's automated fuel optimization system is used, which includes the propulsion system parameters, environmental conditions, and fuel consumption of the ship in operation.

This dataset is first analysed and pre-processed and then used in learning tasks with Artificial Neural Network and Support Vector Machines algorithms. The performance of the algorithms is assessed and then a two-stage model is developed to predict the speed and fuel consumption of the ship under operating conditions.

Finally, the developed models are used in a decision support system that is developed and demonstrated in different operational scenarios.

Keywords: fuel consumption, fuel optimization, machine learning, artificial neural network, support vector machine, decision support system

Table of contents

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

ANN – Artificial Neural Network

AWA – Apparent Wind Angle

AWS – Apparent Wind Speed

CFD – Computational Fluid Dynamics

CII – Carbon Intensity Indicator

DCS – Data Collection System

DFN – Deep Feed-Forward Neural Network

DSS – Decision Support System

EEDI – Energy Efficiency Design Index

EEOI – Energy Efficiency Operational Indicator

EEXI – Energy Efficiency Design Index for Existing Ships

FOC – Fuel Oil Consumption

GCNN – Geodesic Convolutional Neural Network

GHG – Greenhouse Gas

IEA – International Energy Agency

JIT – Just in Time

LASSO – Least Absolute Shrinkage and Selection Operator

MCR – Maximum Continuous Rating

MEPC - Marine Environment Protection Committee

ReLU – Rectified Linear Unit

RFR – Random Forest Regressors

SEEMP – Ship Energy Efficiency Management Plan

SOG – Speed Over Ground

STW – Speed Through Water

SVM – Support Vector Machines

SVR – Support Vector Regression

TWA – True Wind Angle

TWS – True Wind Speed

WASP – Wind Assist Propulsion

1. INTRODUCTION

1.1 Motivation

Despite the shipping industry being one of the most efficient modes of transport, it produces around 3% of the global greenhouse gas (GHG) emissions [1] and according to International Energy Agency (IEA) is a hard-to-abate industry [2], due to the difficulties in electrification, long lifespan of the fleet and high dependency of the fossil fuels.

The first main step of the shipping industry on GHG emissions was in 2000 when the International Maritime Organization (IMO) published the first GHG study [3] that describes the GHG impact of the world merchant fleet in the environment, in that time IMO predicted an impact of about 1.8% of the total greenhouses gases emitted. After that, it took another nine years for the IMO to publish a second study, in 2009, where a deeper study was conducted, concluding that the maritime industry contributes to 2.7% of the total CO_2 emissions [4]. Moreover, in the same year, the IMO's Marine Environment Protection Committee 59 (MEPC) [5] published a not mandatory index, named Energy Efficiency Operational Indicator (EEOI), which would become a reference as a carbon intensity indicator. EEOI is defined as the ratio of the mass of CO_2 emitted per unit of transport work - $[gCO_2/tnm]$, using in its equation the mass of all consumed fuel in a specified route, converted in CO_2 mass, divided by the total cargo carried in tonnes and the travelled distance in nautical miles.

Since the beginning of the last decade, IMO has been doing a major effort to discuss and approve new rules and indices to reduce not only GHG but also the pollutants from burning fossil fuels on ships. The first rule focused on saving energy was approved in 2011 [6], with the inclusion of a new index, the Energy Efficiency Design Index (EEDI) measured in $[gCO_2/tnm]$, and a new mandatory report, the Ship Energy Efficiency Management Plan (SEEMP), which have both entered into force in 2013. The EEDI is an index focused on the design phase of the ship, calculated based on the ship's mechanical installations that measure the theoretical ship's efficiency. Also, this index is cut every five years, wherein this period ship designs need to be increasingly efficient compared to the 2008 base.

The SEEMP was created to be a ship energy management and monitoring report. Although it does not have in it any mandatory plan or obligation of GHG emissions reduction, it provides the basis for any initiative-taking improvement of the ship energy management. The SEEMP can contain the efficiency of each engine of the ship, procedures to improve each system, hull monitoring data, energy management analysis, weather routing analysis and other parameters that could help reduce the fuel consumption on the ship. Still, a specific guideline was published in 2016 [4] with some good practices and operating tactics, to help the shipowner to improve the operational efficiency of the ship.

The MEPC in 2018 [8], was approved the IMO Decarbonisation Strategy, where the focus is to reduce the carbon intensity by 40% in 2030, compared with the 2008 base, reducing this index to 70% in 2050 with a total GHG reduction of all world fleet in 50% (Figure 1-1). Although there are several criticisms by politicians and countries about the numbers involved, this is the first time an entire industry has agreed to be more efficient and reduce the impact on the global environment.

Recently, in 2021 [9], IMO approved a new index called Energy Efficiency Design Index for Existing Ships (EEXI) applicable to all existing ships with more than 400 GT. This index reflects the energy efficiency of the vessel based on the 2008 baseline and introduced an operational index, the Carbon Intensity Indicator (CII), to be discussed and will be entered into force in 2023. CII has the objective to link GHG emissions to the amount of cargo transported over the distance travelled by each ship and need to be included in the SEEMP. It will be mandatory for each ship above 5000 GT and it will rate the energy efficiency of the ship as A, B, C, D and E, where A is the best rating. Moreover, for a ship classified as D in three consecutive years or classified as E, the SEEMP will be mandatorily revised, and a correction plan must be prepared.

Since 2011, IMO has published two studies [1], [10], about the impact of the shipping industry in the atmosphere regarding GHG emissions. The last one, published in 2020, showed that despite the carbon intensity indicator having decreased by 31% (Figure 1-1), the total CO_2 emissions have reduced only about 8%. This occurred due to the increase in the global economy and with that, the maritime transportation, and ships construction [1], also, the actual tendency is for the world fleet to increase by 6.3% until 2026 according to a recent report from BIMCO [11].

Figure 1-1 presents a general overview of the rules and plans to reduce GHG emissions, based on the IMO GHG study [1], the scenario since the first rules applied with their impacts, and what is the objective discussed by operators today on the horizon of 2050.
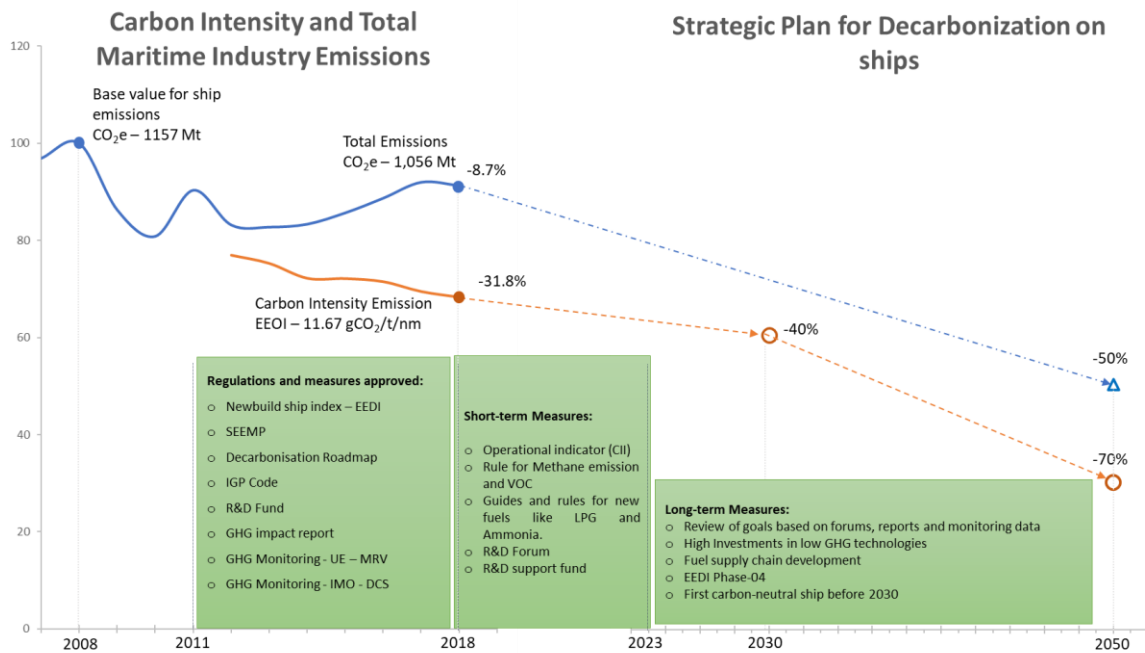
Figure 1-1 - Regulations, measures, and indices of shipping industry along the time

Despite IMO has been accelerating the discussion and the approvement of new rules focused on decarbonisation in the shipping industry, there is pressure to accelerate further this process. The European Union has been leading this process for some time now. In 2015, they were pioneers to approve a regulation named Monitoring, Reporting and Verification (MRV), which requires ships to

monitor and report their GHG emissions [12], fuel consumption, transport work and average energy efficiency. This regulation entered into force in January of 2018 and has the objective to monitor the GHG emissions and carbon intensity in the maritime transportation of ships leaving and entering European ports. IMO approved a similar regulation in the MEPC 70 [13] in 2016, called Data Collection System (DCS), where the shipowner needs to send fuel consumption, distance travelled, type of fuel and other data for each ship, based on annual consumption. With this, the IMO can study the impacts of ships over the years and verify the impact of measures taken to reduce GHG. This regulation entered into force in 2018 and IMO started receiving the data in 2019.

Moreover, in 2020 [14], the European Union (EU) approved to include the ships passing or coming to the European ports to enter in the European Union Emissions Trading Emission (ETS). It means that each ship whose voyage starts or ends in the EU or use for some reason a berth in EU ports may be taxed for carbon emissions [15]. The objective is to encourage investment in improvements to reduce GHG emissions through the reduction of the cost difference between new fuels and green technologies and traditional maritime fuels, e.g., HFO and MGO. Still, this measure according to Mundaca et al. [16], using a tax of US$ 40 per ton of CO_2 emitted will reduce the ship CO_2 emissions by 7.65% using as a reference the current level of international trade.

To achieve these challenging goals, there are six major group measures with high potential of GHG mitigation, according to Bouman et al. [17]. The first one concerns hull design; as examples of this group, for a new ship, one can study its shape through numerical tools such as CFD and the use of test tanks to check the ship's resistance due to speed and thus try to optimize its shape. Also, lighter materials can be evaluated, better ballast management can be studied, the use of paints that avoid the fouling formation and the use of air lubrication system can be used to reduce the ship's resistance, in this last one by creating an air layer around the bottom of the ship.

The economy of scale is the second group, it can be done to companies that work in deep-sea routes, since larger ships tend to be more efficient per freight unit, as it is well knowing that when doubled the cargo-carrying capacity, the required power increases by about two-thirds, thus reducing fuel consumption per freight unit.

The third group of measures are focused on power and propulsion and are related to the design of power systems and machinery in a general way, using waste heat recovery to maximize the energy use or to install new wind propulsions systems, such as the use of kite and sails, to minimize the use of the propulsion system. Hybrid power systems could also be used, such as the use of batteries, which can be utilized in low power operations, such as manoeuvring or berthing operations.

The fourth group is related to the fuel and alternative energy sources, i.e., the use of new fuels to substitute or at least in part the bunker fuels, such as HFO or MGO. As an example, the use of LNG and methanol that is already up and running on 619 and 39 ships [18], respectively, are intermediary fuels, since they reduce the production of CO_2 per kWh [1]. In the future, it is expected to use green hydrogen or green ammonia to reach a total free carbon fuel, not only in the use of the fuel in the engines but in the entire supply chain.

Weather routing and logistic scheduling, the fifth group, consists of studying the operation of the ship or fleet to find the best logistics, where one can find an optimum speed to a pre-determined ship configuration, combined with the weather prediction along the route. This allows the shipowner to meet port demand on time and minimize fuel consumption per cargo per mile during the voyage.

Last but not the least, the sixth group is about the operational speed of the vessel and its relationship with its design speed. Conventionally, the ships are designed to operate at their hydrodynamic boundary speed. However, the required power is proportional to the product of speed and resistance, and the hull resistance curve starts to rise exponentially as the speed increases. Therefore, a reduction in operating speed causes a reduction in fuel consumption per cargo per mile.

Figure 1-2 shows some of the mentioned technologies along with their potential for GHG emission reductions. There are several studies about each mitigation measured, the presented values are based on [1], [10], [17] reports and studies.

Figure 1-2 – Reduced GHG for each technology[1]

As one can see, the carbon-free stage is only achieved by changing the fuel to non-fossil fuels such as hydrogen and ammonia, thus reducing greenhouse gas emissions completely. Still, in the middle ground, ships can be carbon neutral, i.e., using blue fuels such as biofuels and synthetic LNG and synthetic methanol.

Also, the major players in the shipping industry started to invest in technology, studies, and development to produce neutral carbon ships and soon, free carbon ships. As an example, Maersk has implemented a research centre focused only on decarbonization in ships, CMA-CGM already has 26 ships in construction and 15 operating using LNG, Waterfront operates with 11 ships using methanol and they

---

[1] Figure developed by the author with the base figure source: https://conceptbunny.com/container-ship-johannesburg/

have more than eight ordered, to be constructed until 2023, the Eastern Pacific Shipping is testing biofuel is their fleet, and NYK is expecting a ship that could be run with ammonia in 2022.

The class societies are developing rules and good practices to help shipowners to develop new projects and operate their ships with green technologies. As an example, the American Bureau Society (ABS) developed guides to ship fuelled with ammonia [19] and using wind propulsion [20]. To finance all those projects and new vessels, a global framework was established by big financial companies to draft the Poseidon Principles, with the objective to enable financial institutions to align their ship financing portfolios with environmentally responsible behaviour and to encourage the decarbonization of international shipping.

There is no single solution to the problem, a container ship operating on a route between Asia and Europe will certainly not have the same technical or operational solution compared to a smaller bulk carrier that sails only cabotage routes. A catchphrase commonly used by the shipping industry is that "there is no silver bullet", i.e., it will not be simple or easy to reduce emissions and for each ship and operation, a unique solution must be developed, at least in the short-term horizon.

1.2   Problem

The average lifespan of a merchant ship is around 25 years [21]. It is a challenge for shipowners to have energy-efficient ships during their entire lifespan. Even for a new building ship, it is a challenge to meet the new standards and rules to come and still be competitive and efficient in the shipping market.

Improvements of existing ships can be divided into operational and technical, according to Psaraftis [22]. The former concerns operational improvements such as trim, weather routing, and ship monitoring and propulsion control systems, among others, and the latter involves equipment changes, like wind assist propulsion (WASP), new engines and others, as already mentioned.

These improvements can reduce the ship's fuel consumption, however, most of them are better applied to new ships, where the ship can be better designed to adapt appropriately to the green technologies. As an example, the use of WASP to an existing ship in its best scenario barely achieve the 20% of the GHG reduction [23], mainly because the hull and the propeller are not prepared to a high leeway angle, which causes an increase in the resistance of the ship in addition of vibration in the propeller and shaft [24]. New ship projects such as the Oceanbird can achieve a GHG reduction of 90% with that technology [25]. Even if it does not reach this mark in the future, it is far above what is achieved by the current fleet.

That difference in the impact of the technologies between existing and new ships occurs not only with WASP, but also with the use of new fuels, where the engines need to be adapted and new tanks must be installed for the storage of fuels such as methanol, LPG and ammonia, and this together with the loss of cargo space, can cause an increase in drag due to poor distribution of tanks on deck, which ends up not compensating the installed equipment. Not to mention the high investment required and the off-hire period for the installation of the equipment, which requires good planning to achieve a financial return within the ship's lifetime.

The challenge is how to continuously optimize the fuel consumption of ships to achieve the targets and rules, and still be competitive, without installing new equipment, systems, or even new storage tanks from time to time, since this process is expensive and sometimes it is not possible to install or use new technology in an existing ship. Also, traditional fuel consumption represents an average of 66% of a voyage cost [26] and tends to increase due to the high cost associated with new fuels [21], as shown in Table 1-1. So, the problem does not only consist in reducing the GHG emissions but also, be more efficient throughout the operational life of the ship.

Table 1-1 – Consolidated Fuel prices in the DNV GL study [21]

	Fuel	**Price (USD/GJ)**	**Price (USD/toe)**
Fossil	MGO	13.8	578
	VLSFO	12.0	502
	LNG	7.8	327
	LPG	10.2	427
Carbon-neutral	Ammonia	22.9	959
	Methanol	29.8	1248
	MGO	40.0	1675
	LNG	30.7	1285

One of the improvements that can be used by shipowners in existing ships, is to install a fuel optimization system that automatically adjusts engine rotations and propeller pitch to increase the propulsive efficiency of the ship. These systems optimize the fuel consumption in the route as a function of a set of variables such as speed, propulsive system parameters, and environmental conditions, which are monitored continuously throughout the voyages. The data can be used not only to calculate fuel consumption but also, to identify points of improvement, such as operational bottlenecks, preventive maintenance or situations that hamper the ship efficiency improvement.

Such a system has been installed on a container ship with 126 m length of a Portuguese shipowner. The system is intended to reduce the impacts caused by environmental conditions during navigation, controlling the maximum fuel consumption. The automated optimization system can adapt the shaft rotation, the pitch angle of the propeller and the fuel rack position of the main engine to ensure that the fuel consumption does not exceed the set value, according to the inputs from the environment and fuel consumption and shaft thrust meter. Data collected from this ship's fuel optimization system is used in this dissertation. The system is further explained in Chapter 4.1.

The ship sails off the continental coast of Portugal and to the Portuguese islands of Madeira and Azores. Figure 1-3 shows all voyages of the ship in one year. The voyages are normally made in a counterclockwise direction, that is, from the port of Lisbon to the Azores, to Madeira Island and then returning to the mainland.

The data available for this study were collected during one year of the ship operation. The dataset includes data related to the propulsive system such as main engine power, shaft rotation and propeller pitch, also, from the auxiliary power generation system such as shaft generators and auxiliary engines,

as well as voyage related data like vessel position, speed over ground and wind condition. This generated a dataset of over 20000 records with more than 20 monitored variables. To find patterns, correlations or to use the data to improve the operation it is necessary to use numerical and statistical tools, otherwise, it will be difficult and time-consuming to analyse each group of data and try to find patterns to be used in an operational improvement.

This is a challenge with the growth of ship automation, as larger datasets are generated that cannot be fully analysed by experts without proper numerical tools. So, automated processes with Artificial Intelligence (AI) and Machine Learning (ML) models have been used to help analyse these large datasets and to develop decision support tools [27]. The correlations between the data collected should be studied to use in the ML method. Also, pre-processing should be made in the data if necessary, and outliers and redundant variables should be removed so that the models can be more reliable. In addition, there are several ML methods, which should be studied and tested to assess which one can meet and predict the expected results. Each ship has its size, equipment, route, and own operation, so a model that fits each ship configuration must be found since there is no general model that can be used for all ships.
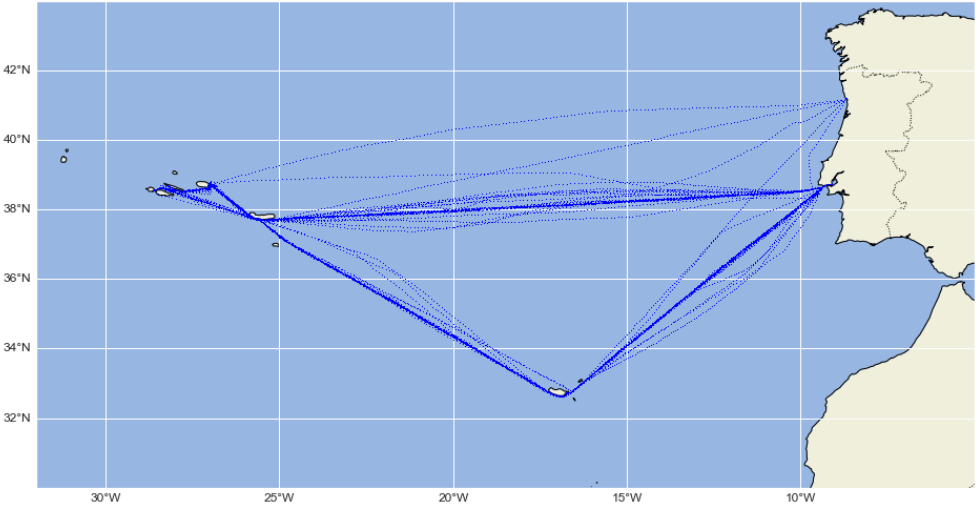
Figure 1-3 – Ship route during a year

1.3   Objectives

The objective of this dissertation is to develop machine learning models that represent the operation of a fuel optimization system and to develop a prototype of a decision support system that provides predictions of the optimal fuel consumption of the ship's main engine.

For this purpose, a one-year sample of data collected from a ship's automated fuel optimization system is used, which includes the propulsion system parameters, environmental conditions, and fuel consumption of the ship in operation

The objective is to analyse different operational scenarios, such as the speed and the use or not of the shaft generator, and to check the total fuel consumption on the voyage.

The first step of this study consists of analysing the data collected during one year from a container ship operating off the coast of Portugal and the Atlantic islands. Statistical methods are used to analyse the correlations of each variable, remove outliers and identify patterns on system operation variables.

With the dataset generated by the analysis process, machine learning methods are adopted and derived. The performance of the methods is assessed and then a model is developed to predict the speed and the fuel consumption of the ship. To build a model with good accuracy, a two-stage model is proposed, using robust ML methods, such as Artificial Neural Network (ANN) and Support Vector Machine (SVM). The first stage has the objective of finding a model that can predict the ship's speed, and the second model, the main engine fuel consumption.

Then, the two-stage model is used to build a decision support system. This system simulates the ship on a voyage under specific environmental conditions, changing the propulsive configurations to achieve a minimum desirable speed and minimizing fuel consumption.

A code in Python computer language is developed to perform the data pre-processing, the statistical analysis and the machine learning tasks, as well as to predict the velocity and fuel consumption using the two-stage model and the decision support system.

1.4 Work Structure

Chapter 2 has the literature review. It describes the studies related to the use of machine learning in the maritime industry, not only to develop models for speed and fuel consumption prediction but also models used for weather routing, to predict delays and other predictions using machine learning.

The theoretical background on machine learning is presented in Chapter 3, with a brief description of the machine learning models used to develop the prediction model, namely the artificial neural network (ANN) and the support vector machine (SVM).

Chapter 4 analyses the data extracted from the ship's automatic fuel optimization system. It describes the data pre-processing and the new data that has been collected to enrich the dataset, such as draught, trim and all wave characteristics.

The data correlation analysis is performed in Chapter 5, where an analysis of the voyages recorded by the ship is conducted. Also, a study on the impact of the use of the ship's fuel optimization system, and an analysis of the impact of the use of the shaft generator are conducted.

Chapter 6 details the construction of the machine learning models, with the hyperparameters used to find the best models that could predict with good reliability the main engine fuel consumption and the speed over ground, with a focus on the construction of the decision tool.

The decision support system is presented and demonstrated in Chapter 7 and conclusions and future works are provided in Chapter 8 that also discusses other uses of the machine learning models developed in this study.

2. LITERATURE REVIEW

2.1 Machine Learning in the maritime industry

Machine learning (ML) algorithms are programs that can learn from data and improve from experience, without human intervention. Even in the shipping industry machine learning algorithms have been used for a wide variety of studies and improvements. Jain and Deo [28] reviewed the state of the art on use of the Neural Network method in ocean engineering. The authors found in that time that this method was mainly used to predict a result of some random variables such as wave height (both temporal and spatial cases), tidal levels, wind speeds and other metocean parameters, moreover, there were a few studies to predict structural forces due to wind and wave loads and structural damage indicators.

More recently with the growth of the knowledge about machine learning, computers processors, and digitalization of all the areas, there is a growing use of ML techniques to analyse other behaviours of the shipping industry that not only about the environmental condition, resistance force and fuel consumption prediction. As an example, Rawason et al. [29] studied several different machine learning methods to predict the risk of an incident due to severe environmental conditions. The studies resulted in a good prediction of incident cases, but with many false-positive cases, given that the complexity of an accident and incident causes. The authors discussed similar cases where if there is an accident, there are also several non-accidents before that, and that is quite complex to analyse with just a binary ML model classification.

In the same way, Viellechner and Spinler. [30] studied different regression prediction models to foretell the time that a ship will be delayed considering weather disaster, piracy risk and chokepoint congestion. Like other studies, they found that Neural Network (NN) and Support Vector Machine (SVM) have the best accuracy, which considers the precision of both delay and on-time predictions, compared to all delay predictions. Also, the models have a good sensitivity, which considers only the prediction of delay values to this problem.

Tsaganos et al. [31] used the open-source software Wake, a cost-free data mining tool, to study different ML algorithms to predict the marine engines faults, due to its importance to guarantee high reliability in ships operations. The authors found a good predictive model using the regression trees algorithm with the monitored variables from the engine, such as total power, the rotational speed of the engine, exhaust gas temperature, ignition, and others.

Outside the areas of technical and dynamic analysis, ML has also been used in studies involving different subjects. Wang et al. [32] developed a Bayesian model to predict the probability of bankruptcy of oil tanker shipping companies. Fabregat et al. [33] used machine learning tools to estimate and compare the impact of cruisers on the pollution around the ports. Sanfilippo and Chikkagoudar [34] applied unsupervised machine learning methods to find illicit trade using merchant ships. Zhou and Thai [35] studied models to predict personal injury accidents in tanker shipping companies. Bramer et al. [36] analysed a Naïve Bayes model to identify anomalous behaviour of the shipping companies in the business.

There are also studies using machine learning to speed up the process of analysis of models already consolidated. These models use results of numerical models as a dataset for the learning model, thus having a prediction model that can give much faster answers than the numerical model studied. As an example, Boogaard et al. [37] studied a Geodesic Convolutional Neural Network (GCNN) model to predict the thrust and torque coefficient and the open water efficiency of a ship propeller based on Computational Fluid Dynamics (CFD) simulations. Moreira et al. [38] studied an artificial neural network model based on ship travelling simulations of a weather routing system, to predict ship speed and fuel consumption. Also, Krata et al. [39] used the same weather routing system to develop a Bayesian Network to predict the ship's speed based on propulsion system variables and wave configuration. The authors focused to develop a trustable model to be used with real observed ship data in the near future.

2.2 Predict ship resistance in the design phase

Rudzki and Tarelko [40] studied many methods to calculate ship resistance and argued that there are two stages where it can be performed some influence in the ship propulsion performance. The first would be the design or project stage, and the second is the operational stage. The first stage is the study of the ship's behaviour and design, where its hull shape will be developed, the propeller designed, the necessary torque in the shaft defined, and all the other systems and features that influence fuel consumption will be studied and projected. The operational stage is influenced not only by the results of the first stage but also by several other operational factors that may not have been fully mapped in the first stage, such as fouling in the hull and propeller, ship vibrations and equipment malfunction.

To have a fuel-efficient hull design and an efficient propulsion system, various methods can be applied in the first stage to estimate the ship resistance for each ship configuration studied. These theoretical models are based on physical principles, i.e., they formulate the physical behaviour of the ship to calculate each part of the total resistance equation, e.g., the hydrodynamic relationship between the hull shape and sea condition, or hydrostatic calculations to determine the trim and stability of the ship.

To calculate the total resistance, it could be used in the design phase empirical or semi-empirical methods. These methods are regression formulas based on a statistical analysis of empirical data. From the former method, it can be used the recommended procedure by the International Towing Tank Conference (ITTC) [41], where an analytical methodology is presented to predict the delivered power and the propeller rate of revolution. For the semi-empirical methods, as an example, it can be used the Hollenbach method [42] or the Holtrop & Mennen method [43], [44], where the authors present methods to calculate the still water resistance based on regression analysis of test models in towing tanks. Being the latest, one of the most used models for ship performance studies and predictions in the project phase [45].

But those methods, according to Bertram [46], are outdated and overestimate the total resistance of modern ships as they do not consider the mechanical and hull efficiencies presented in more recent designs. Also, Aldous [47], [48] argued that the extrapolation is based on a small number of experiments and extrapolations to different types of ships could have high uncertainty.

The added resistance in waves can be calculated by the method developed by Faltinsen et al. [49], the method uses the transverse drift force and the yaw moment on a ship in regular waves to calculate the added mass based on the strip theory of the Salvesen et al. [50]. Also, numerical methods can be used to calculate the total resistance in waves since the well-known Mitchel theory [51], [52] and Rankine source method, as in [53].

Some of those cited methods can be used to calculate ship resistance in still water conditions. To try to compare these methods with existing ships, Dinham-Peren and Dand [54] studied and analysed measurements of an operational ship intending to determine the actual performance of the vessel in calm waters. To verify and compare the monitored data with the calm water predictions the authors proposed to limit the data points used, using them only if the following conditions are fulfilled:

- Wind speed less than 10 knots.
- Significant wave height less than 2 metres.
- Drift angle less than 5 degrees.
- Current less than 1 knot.
- Shaft rotation between 0 and 120 rpm.
- Successive speed samples with a difference of less than 0.5 knot/minute.

With these configurations, the authors aimed to get a calm water configuration inside a normal operational measurement, and if comparing with different operational periods, it can be verified if the ship is performing better or worse, or even verify the fouling in the hull or the wind effect. Moreover, Lakshmynarayanana et al. [55] used the same method to separate and calculate the wave resistance in an operational data set, using that to develop a parametric model to wave resistance force for that specific ship. Both studies show how distant the theoretical forecast is when compared to the operational data, as they differ from the expected results.

This difference occurs because the total ship resistance has many factors that can influence it, as shown in Larsson and Baba [56], and many factors need to be monitored to calculate it to decrease the uncertainties, as seen in Pedersen and Larsen [57], also the operational routine has many changes that are not mapped or not properly controlled as hull cleaning, fouling in the propeller, bad machine operations, trim change, leeway angle, vibrations in machine operation, among others, as shown in Figure 2-1 and Figure 2-2.

As an example, Park et al. [58] investigated the uncertainties of the KVLCC2 using the ITTC method [41] and concluded that there is an increased resistance caused by short and moderate wavelength cases, these been 16% and 9% respectively. So, to correct predict or reduce the uncertainties of the prediction of the resistance it must be calculated the uncertainties to each variable that can influence the resistance as in Figure 2-1, this may be unrealistic to do even for a small class of ships. Even the use of CFD as in [59], [60] cannot absorb all the variables and well predict a ship operational resistance with great accuracy due to the lack of operational information and difficulty to simulate all events that may occur on the route, such as breaking waves conditions, as in [61].
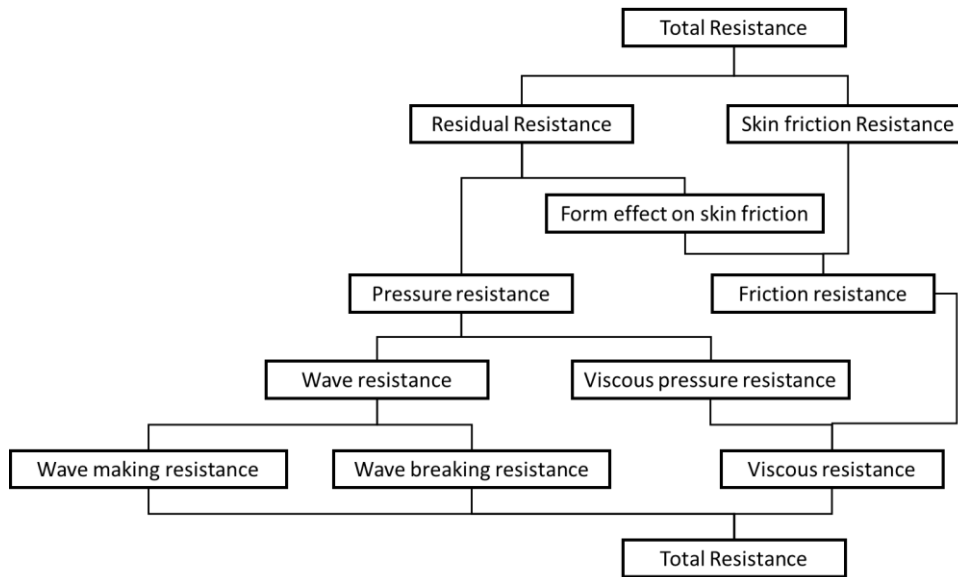
Figure 2-1 - Resistance Decomposition based on Larson and Baba division [56]
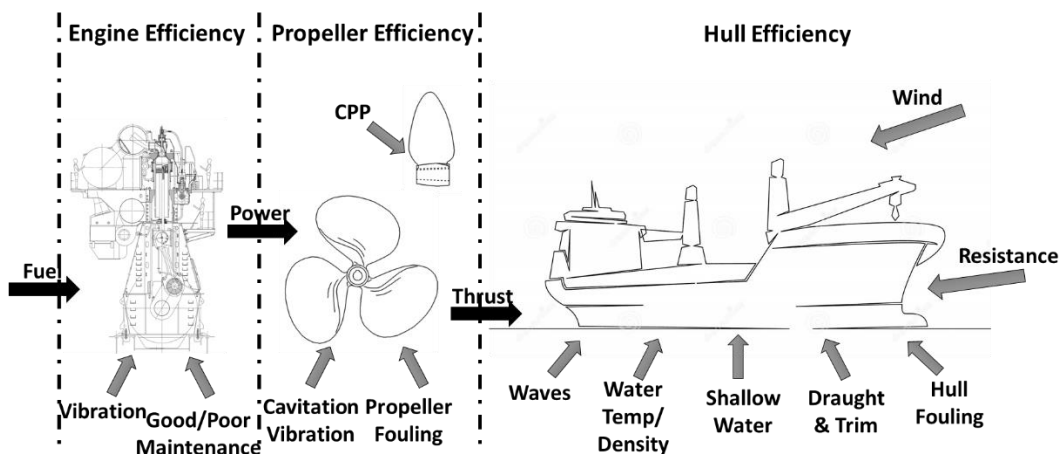
Figure 2-2 – Ship efficiency based on Pedersen and Larsen [57]

2.3   Machine Learning in Fuel Oil Consumption prediction

To try to correctly predict the fuel consumption and avoid errors due to the uncertainties present in the operational routine, some studies have been developed in the so-called black-box model, when the model is driven only by data and statistical methods, like the use of ship dataset and machine learning.

Parkes et al. [62] studied the data from three sister ships with more than 120.000 records to predict the shaft power of each one. They used as input variables the wave height, true wind speed, apparent wind direction, draught, trim and the speed over ground to predict the power required by the vessel and developed a sensibility analysis around the number of hidden layers and parameters. Also, a comparison with regression methods presented in [63] was analysed, where the model fits well until the maximum measured speed. It is important to notice that the author did randomised analyses, and the results are more profitable where they had more trained data, between 11 and 17 knots, after and before

this range the results have a higher variance. When a machine learning solution is developed, it is important to have a good balance in the data weight and know the boundary limits of the prediction. The authors investigated the number of data required for model convergence, and the model limits were investigated, comparing, and analysing which variable is more necessary to reduce model uncertainties. The final error was obtained around 8%.

Wang et al. [64] developed a model that uses a wavelet neural network (WNN) combined with the parametric ship models to optimize a ship energy efficiency for a ship cruise working in short distances. The model was developed to predict the water depths and the wind speed using the neural network and uses it as an input to the parametric ship resistance calculation, thereby calculating the curves correlating the engine speed with the fuel consumption. The model presented a good result and was implemented in the ship operational routine, reducing by 19% of the total fuel consumption.

Gkerekos and Lazakis [65] developed a weather route model based on the documented ship performance and climate conditions, i.e., wave and current characteristics, where the input to calculate the optimal route is based on ANN model training combined with a weather prediction consideration and a modified Dijkstra's algorithm. This model calculates the best route to minimise the Fuel Oil Consumption (FOC), on a fixed route between France and Angola.

Peterson et al. [66] investigated and compared the use of ANN and Gaussian process models to predict the propulsion power of a ferry in the Faroe Islands. It was used two-month data from the speed through water (STW), fuel consumption, relative wind speed and angle, propeller pitch, rudder angle, trim angle, and the distance from the sea surface of two points along the ship's route. Despite both models having a reliable performance, the Gaussian model had a poor accuracy nearby the limit of the scale, which means that one must be careful to choose the limits of the variables to predict consumption after the model has been built, otherwise the results may be unrealistic. The study concludes that the ANN model is more robust and tolerates projections outside the range studied compared to the Gaussian model.

Bal Besikci et al. [67] studied an ANN model to develop a decision support system (DSS) to assist the crew to reduce fuel consumption. The resulted determination coefficients (R^2) were low compared to other studies presented in this chapter, i.e., 0.834 for training data and 0.759 for validation data, this occurred due to the use only of the noon report, which in this case was only one per day and with low accuracy due to human errors. With that, the authors only have about 200 samples to train, test and validate the model, which is also low compared to other studies. The DSS developed can verify different configurations of trim, RPM, velocity, and pitch of the propeller to find the best sailing configuration to reduce the fuel consumption, estimating the best solution to reach the destination on time.

Pederson et al. [57] were pioneers to investigate the use of ANN to predict ship resistance based on the full-scale measurements of ship speed, wind speed and direction, sea and air temperature, in different load conditions. Also, they compared the results with the empirical and data-driven methods based on hydrodynamics relationships (e.g., Holtrop and Mennen) and concluded that the use of ANN is a better model to use in predicting operating resistance with differences ranging from 5 to 20 percentage points compared to theoretical models.

Farag et al. [68] studied an ANN prediction model for an oil tanker with a route between Sultan Qaboos Port, in Oman, and Rotterdam, in the Netherlands. The authors studied the correlation of 11 variables, related to wave, wind, current and ship speed, with the fuel consumption and modelling an ANN using a feed-forward neural network with the polynomial regression model in the hypotheses function. The results presented show a good prediction model with R^2 around 0.98. The authors also used the same route studied as an example of how to predict the fuel consumption before starting to navigate and how to study a just-in-time (JIT) scenario using the model.

A deep feed-forward neural network (DFN) was developed by Lazakis et al. [69], where the authors developed a process to use the ocean environmental data, like wave period, wave height, wind speed and angle, and water temperature, and the ship configuration as draught, speed over ground and heading, to develop and analyse the loss of the cost function of a fuel prediction model. They developed 540 models using DFN to find which has better results, changing the hyperparameters as the number of hidden layers, the number of neurons in hidden layers, the learning rate, the gradient optimizer, and the dropout parameter, this last one was used to prevent overfitting. The final model has an error of 3.5% compared to the test data.

Tuan Hoang et al. [70] developed an ANN model to predict diesel engine performance and exhaust emissions using biofuel instead of diesel. The authors were successful in estimating the characteristics of the emissions, e.g., NO_x, CO and CO_2, and its performance, e.g., torque, main engine power and exhaust gas temperature. Although the study was not focused on the maritime industry, the developed model is important because it can have high precision in the engine performance with different fuel mixture proportions, which is extremely useful for the maritime area due to the current moment of migration towards less polluting fuels and with less GHG emissions, with all the uncertainties about the new dual-fuel engines.

A combination of ANN and Monte Carlo simulation was used in Tien-Anh [71] to develop a prediction method to fuel consumption in the main engine of a bulk carrier. The authors studied a two-year dataset from noon reports to develop this model, also dividing it into three operational states of the main engine based on the Maximum Continuous Rating (MCR), been them 65%, 85% and 100% of the MCR. The ANN was used to develop the prediction model and using the Monte Carlo simulation to analyse the correlation between each variable used, the author achieved a good model with a high coefficient of determination, also, it can analyse at each stage of main engine operation, which variables have the greatest impact on fuel consumption.

Despite ANN being one of the most used models in FOC prediction, it is not always the best prediction model to be used, even if it can be scaled into a deep learning method, and with that find patterns, those other methods cannot. In Gkerekos et al. [72], a large comparison study was conducted to either compare two databases, noon reports and automated data, and different learning methods, e.g., Support Vector Machines (SVM), Artificial Neural Network (ANN), Random Forest Regressors (RFRs), among others. Using the features as main engine RPM, ship speed, wind speed and direction, sea state and direction, and draught, and changing the hyperparameters for each training model, a large analysis was done, analysing the convergence of each model, and comparing the best results. All models

reached good prediction results of the testing dataset, with a coefficient of determination greater than 0.85, with the SVM being the best result with 0.91.

Other studies analysing fuel prediction within machine learning, in Hu et al. [73] compares the Back Propagation Neural Network (BPNN) with the Gaussian Progress Regression (GPR), where both methods resulted in a high value of the coefficient of determination about 0.98, to predict a fuel consumption using wind, wave and ship routing configuration as input data.

Kim et al. [74] studied the difference between the models of Artificial Neural Network (ANN) and Multiple Linear Regression (MLR) to a deep-sea route of a 13000 TEU container ship, where in this case ANN had a better model with a score of 0.97. The authors also studied the use of the least absolute shrinkage and selection operator (LASSO) regularization to investigate the curve fitting of the input variables, getting a better result when this regularization was applied, as it discards the variables that least influence the fuel consumption.

A support vector regression (SVR) model was developed by Kim et al. [75] for a 200,000-ton cargo bulk carrier, the objective, in addition to developing the main engine power prediction model, was to compare the machine learning method to the method presented in ISO 15016. The result showed that the model predicted from the data collected directly from the ship are more reliable than those derived from ISO 15016 for that specific vessel. The model presented a good prediction with a coefficient of determination about 0.89, better than the model presented by the ISO that resulted in 0.3. This difference occurred, according to the authors, because the ISO model assumes static sea conditions, and this is more evident when compared with the ML model for cases with the severe sea.

3. THEORETICAL BACKGROUND

Machine learning is often defined as a "*Field of study that gives computers the ability to learn without being explicitly programmed*", a definition attributed by Arthur Samuel based on his studies in the field [76], where the author explains that by programming computers to learn, it is possible to eliminate the need for much of the effort in detailed programming, the author has studied computer games and artificial intelligence. A more technical definition is used based on Thomas Mitchell definition, "*A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E*", in the Introduction Chapter of his book [77].

The methods can be divided basically into supervised and unsupervised learning. The first is used to learn from well-defined input and output sources, comparing each iteration with the predicted results and the correct results given, and minimizing the error for the next analysis and learning iteration. It can be a regression model when it predicts continuous-valued output, e.g., price of commodities, fuel consumption or estimating life expectance, or classification model, when it is an output of discretized variables, in other words, there is a classification in the predicted result, some examples are: verify the colours of an image, diagnosis of diseases such as cancer, facial recognition and identify financial fraud.

Unsupervised learning is a method that can discover hidden patterns in the data without the need to direct what is expected. It is a powerful tool to analyse large datasets because it can identify and segment patterns when it is impractical for a human to propose a trend with a large amount of data provided. This method is largely used to segment the customers in shops and websites, or biology studies for genetics and species, and also, in day-by-day when the websites suggest a new book, film or news, it is an unsupervised machine learning method acting behind the curtains.

Currently, there are many types of machine learning algorithms, part of them is exemplified and categorised as supervised and unsupervised and shown in Table 3-1. There are also the categories of Semi-supervised and Reinforcement, as well as other ML models. In [78] one can find the use of each one and some descriptions on how they work.

Table 3-1 – Example of Machine Learning Model Types

Supervised	**Classification**	**Naïve-Bayes**
		K-Nearest Neighbors Classification (KNN)
		Suport Vector Machines
		Artificial Neural Network
	Regression	Linear Regression
		Suporte Vector Regression (SVR)
		K-Nearest Neighbors Regression (KNN)
		Artificial Neural Network
Unsupervised	Dimensionality Reduction	Multidimensional Scaling
	Clustering	K-Means
		Gaussian Mixture Models

3.1 Artificial Neural Network Method

One of the most used methods to predict results and translate and recognise pictures and sounds is the Artificial Neural Network method (ANN). The method has been used for many functionalities such as face-recognizing [79], automated driving cars [80], sound recognition [81], among others. In the shipping sector, some studies are using it to predict fuel consumption and ship speed, as well as for other prediction models as shown in Chapter 2.1.

Artificial Neural Network (ANN) is a type of machine learning method inspired by the decision-making process of the biological human central nervous system that uses nerve cell networks (neurons), where the dendrites (Figure 3-1) receive all the sensory, auditory, visual and taste information, processes this information in the nucleus and forwards the information to the axons. This output is what it uses to send signals to other neurons and restart the process, with new assimilation, analysis and new signal. This occurs until a response to the stimulus or thought is formed [82]. The ANN mimics the brain, making the same composition and using algorithms to solve problems and find patterns as shown in Figure 3-1.

Conventional computer programs are used to replace, enhance or speed up a human calculation or other activity, where the computer follows a set of instructions (algorithms) to solve a problem. Machine learning methods, as the ANN, oppositely, learn by example, using a large number of highly interconnected processing elements (nodes) that work in parallel to solve a specific problem.
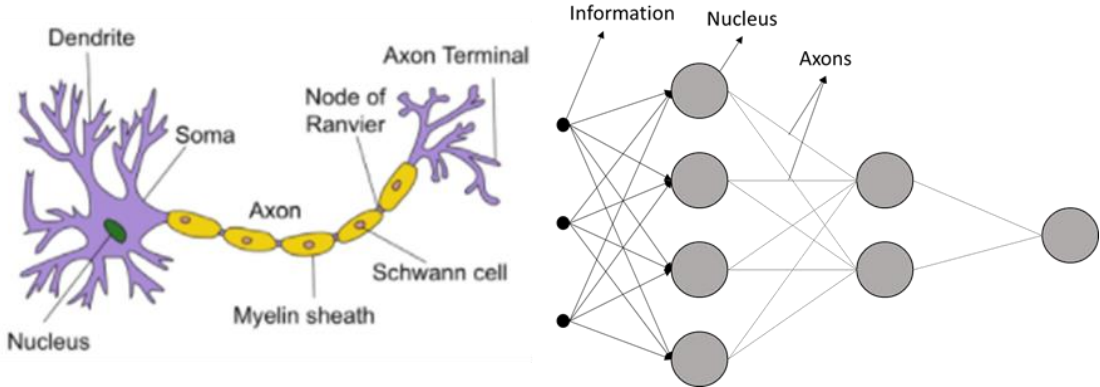
Figure 3-1 – Structure of a typical neuron and a typical ANN[2]

This method is widely used in data mining from large datasets due to the increased use of the internet, automation, and internet of things (IoT) [83] such as medical records, web clicking, biology, quantum chemistry, face identification, engineering development and other systems with many nonlinear relationships involved between its elements. It is used for applications that cannot be programmed by hand, e.g., autonomous vehicles, images recognition, stock market prediction and medical diagnosis.

Neural networks are specific, in other words, ANN is built to solve a specific type of problem. In a general way, it can be used as a tool for prediction, forecasting, estimation, classification, and pattern recognition

---

[2] First figure source: Southampton Machine Learning Course.

in different areas as explained above, but there is no ANN built to a universal purpose, solution or study, to each case one needs to build a particular network.

3.1.1 Layers

ANN has three basic sections each one composed of nodes, i.e., Input Layer, Hidden Layer and Output Layer, as shown in Figure 3-2. The Input Layer, as the name suggests, is the first layer of a neural network where the input nodes are allocated, each of these nodes represents an input variable to be analysed, called as features and represented as $x_j^{(i)}$, which is the value of the feature j in the (i^{th}) training example for the training set. All features are stored in a vector form represented by $a^{(1)}$, as shown in Figure 3-2. It is important to mention that in cases with just a few features the matrix and vector solution is faster compared to the other solvers.

The hidden layer is a layer of nodes between the input and output layers (Figure 3-2) and can be single or multiple. Neural Networks with more hidden layers can go deeper into the learning process and find patterns and associations that are not possible with a few layers, this is an example of the so-called deep learning. The hidden nodes belonging to this layer can also have different quantities for each case, and even for each hidden layer in the network. There are some studies about how many hidden nodes one can use to have a good performance of the model [84], [85], also, an analysis of the number of hidden nodes and layers was performed in this analysis to find the better ANN configuration that predicts a good result without overfitting. The nodes are represented by $a_j^{(l)}$, where j is the position of the value in the hidden layer l.

Finally, the output layer within the output nodes can be with single or multiple nodes, where it can be a binary or probabilistic value (classification problem), like cancer analysis or face identification, or a continuous value (regression problem) like the fuel consumption, gas emission or ship speed. This layer contains the results from the prediction solution, and it is represented by the vector $h_\Theta(x)$, called hypothesis function or activation function [82].
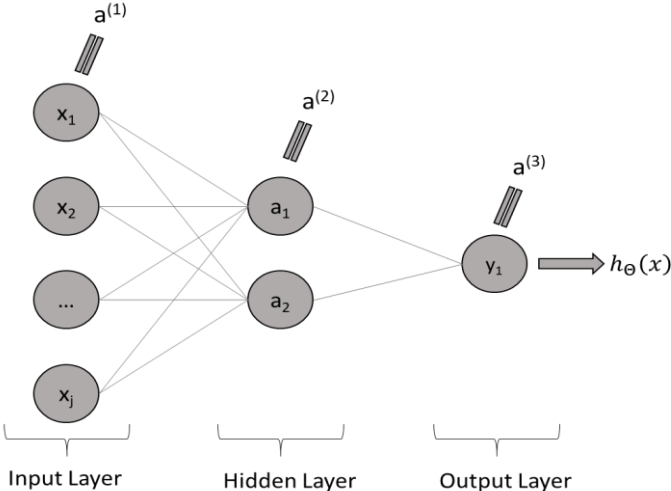

Figure 3-2 – Neural Network Example

3.1.2 Hypothesis Function

The hypothesis function, or activation function, is the function that calculates the output of a given node. This function is the result of the summation of the multiplication of each node in the previous layer l to a parameter $\Theta^{(l)}$, also called weight as in equation (3.1), or it can be represented in a vector form, as a multiplication of two vectors as in equation (3.2). The results of this process represent the total influence of each previous node in the actual active node.

Still, the parameters or weights, represented by $\Theta^{(l)}$, are the variables that represent the influence between two nodes to obtain the net input of a subsequent node in the subsequent layer. One can see a graphical example in Figure 3-4.

$$h_\theta\big(x^{(i)}\big) = \sum_{j=0}^{n} \theta_j x_j \tag{3.1}$$

$$h_\theta\big(x^{(i)}\big) = \theta^T x^{(i)} = \begin{bmatrix} \theta_1 & \theta_2 & \dots & \theta_n \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{bmatrix} \tag{3.2}$$

Some mathematical functions can be used as an activation function, as an example, one of the most used functions to calculate probabilistic prediction is the sigmoid function (3.3), as it results in a value from 0 to 1, and the probability of something happening is also in the range of 0 and 1, it fits perfectly as the activation function in classification problems.

$$h_\theta\big(x^{(i)}\big) = g(\theta^T x^{(i)}) = \frac{1}{1 + e^{-\theta^T x^{(i)}}} \tag{3.3}$$

However, nor the probabilistic analysis or the classified analysis fit in this study since almost all the variables are continuous, so some of the functions that fit for this regression method are presented.

The hyperbolic tangent activation function (3.4) is similar to the sigmoid function, a called s-shape curve, but the results are between the values -1 and 1, with that the input that causes negative impacts can be mapped as strongly negative results and also, the near-zero input will be mapped as a near-zero result.

$$h_\theta\big(x^{(i)}\big) = tanh(\theta^T x^{(i)}) = \frac{e^{\theta^T x^{(i)}} - e^{-\theta^T x^{(i)}}}{e^{\theta^T x^{(i)}} + e^{-\theta^T x^{(i)}}} \tag{3.4}$$

Another very useful and used function is the Rectified Linear Unit (ReLU) (3.5), as one can see, the ReLU is half rectified (from the bottom). The function $R(\theta^T x^{(i)})$ is zero when the result is less than zero and is equal to the result value when it is greater than or equal to zero. This function is one of the most used in neural network models.

$$h_\theta\big(x^{(i)}\big) = R(\theta^T x^{(i)}) = \max(0, \theta^T x^{(i)}) \tag{3.5}$$

In Figure 3-3 one can see the expected result for each function presented.
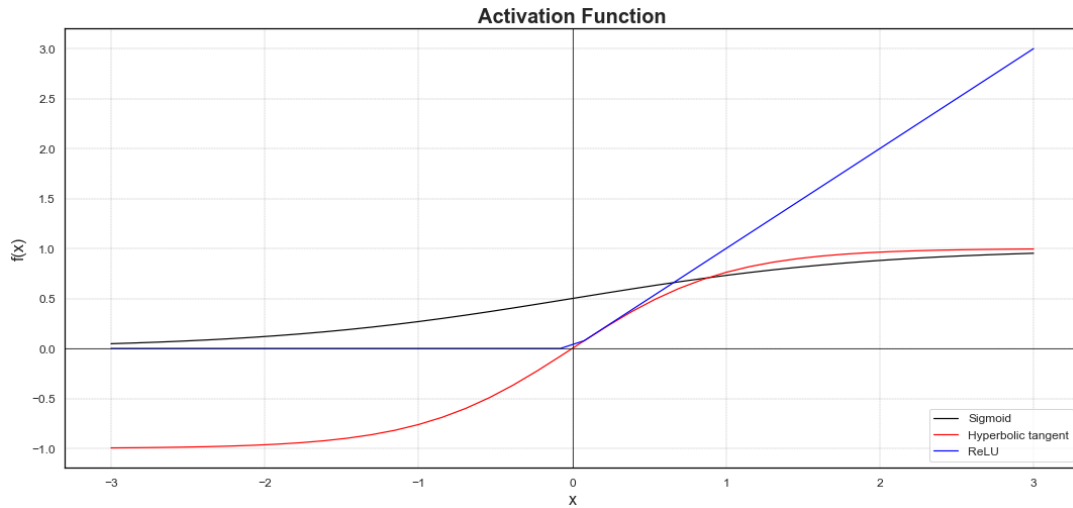
Figure 3-3 - Activation function plot examples

3.1.3 Bias Node

Another important component of the ANN is the bias node. This node is added in each input and hidden layer and normally has the value as 1 or -1, as exemplified with the unitary neuron in Figure 3-4. This node is also multiplied by a weight, and it is intended to provide to each equation a unitary constant value that could be used to shift the solution, as shown in Figure 3-5, and help to converge the weight calculation and efficiently minimize the cost function, as it complements the equation with a constant value.
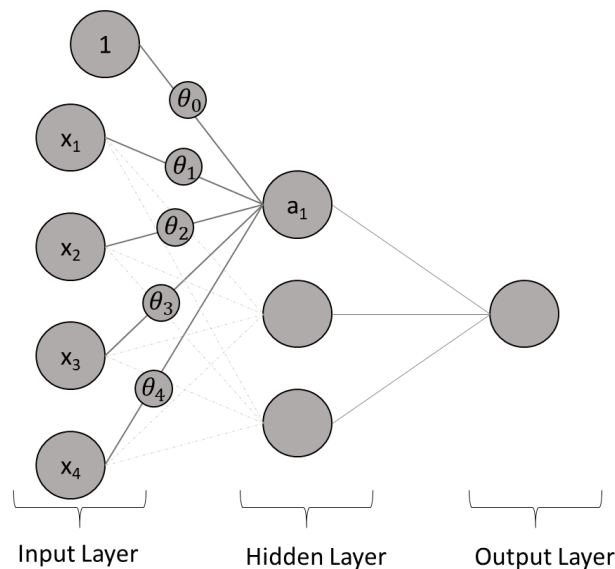
Figure 3-4 - Example of use of hypothesis function in the first layer
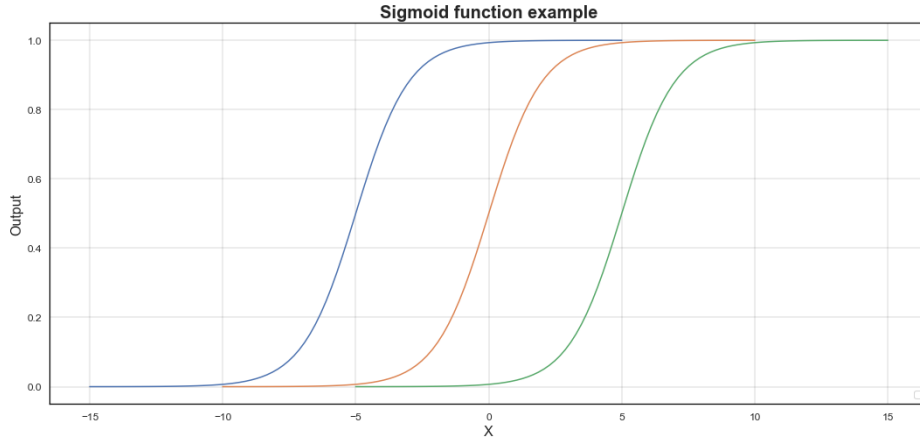
Figure 3-5 - Example of the impact of bias node in the output

3.1.4 Cost Function

The cost function is the function that calculates the total error of the network comparing the predicted results with the expected results. This error value represents the loss or cost associated with the network, which is sometimes referred to as objective function due to the objective of the ANN is to minimize this function.

There are a few types of cost functions that can be used in ANN, e.g., Mean Squared Error (MSE), Squared Error (SE), Sum of Squared Errors (SSE) are the popular ones, there are other functions that can also be used as objective functions, for instance, Exponential, Hellinger Distance and Cross-entropy.

In this study, the MSE equation (3.6) is used to evaluate the cost of each iteration until the convergence and to evaluate the cost of the testing and validations sets.

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2 \tag{3.6}$$

Where m is the number of sample data, $h_\theta(x^{(i)})$ is the predicted value calculated using the hypothesis function, and $y^{(i)}$ is the real result for the ith example.

Still, to avoid the Overfitting problem it is common to use a so-called regularization value, represented by λ, that penalizes the parameters' values θ. The regularization value is used to avoid the overfitting problem since it helps minimize the influence between the nodes. The new cost function is represented by the equation, and one can see that the parameter penalizes the square of the sum of all the parameters.

$$J(\theta) = \frac{1}{2m} \left[\sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2 + \lambda \sum_{j=1}^{n} \theta_j^2 \right] \tag{3.7}$$

where n is the total number of parameters. It is important to notice that the parameter θ_0, corresponding to the bias value influence, is not penalized.

3.1.5 Gradient Descent Algorithm

Gradient Descent is the optimization method used in an ANN to find the combination of weights that will minimize the cost function. With this method, at each iteration, the weights are updated descend down the slope of the gradients to find the minimum error point as in equation (3.8).

$$\theta_j := \theta_j - \alpha \frac{\delta}{\delta \theta_j} J(\theta) \qquad (3.8)$$

Where α is the learning rate of the network, it is a value that determines the size of the step that is used to move towards the global solution, i.e., global minimum. It can be static, or it can change in function of the network's error rates, if the error falls, the learning rate can decrease. In Figure 3-6, one can see how the algorithm works, the size of each step is defined by the size of the learning rate, and it is expected with a good selection of the learning rate, the cost function will decrease and stay constant after a certain number of iterations.
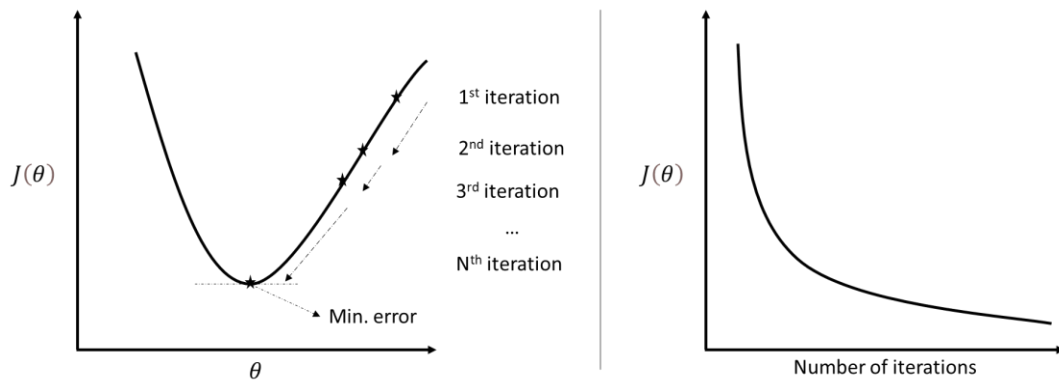
Figure 3-6 - Graphic example of how the gradient descent works

It is important to mention that the choice of learning rate influences the velocity to find the result and sometimes the non-convergence of the solution. For very low values the solution will take many iterations to find the minimum point as the descend step size is too small, for high values in the opposite way, it may have a non-convergence since the solution is oscillating around the minimum solution and never reach the minimum error.
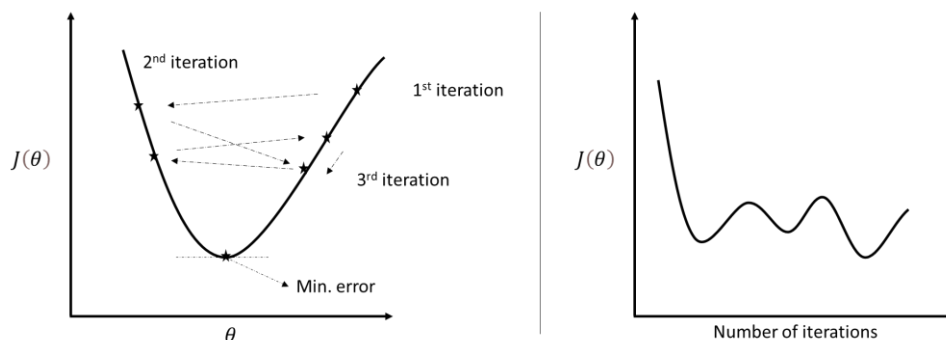
Figure 3-7 – Example of a large learning rate with a non-convergence

3.1.6 Overfitting and Underfitting

Three types of datasets are used to develop an ANN. The first one is the training set that corresponds usually to 60% of the samples and is used to calculate the parameters of the network. The second, the cross-validation set, is used to calculate the cost of each iteration and to minimize the overfitting problem and stop the iteration when it is already converged and stopped improving, this set is about 20% of the samples. The testing set, with the final 20% of the samples, is used to check how accurately the network has been trained.

To verify how a model is performing one need to compute the train set error, using the Cost Function $J(\theta)$, but this is an optimistic estimation of the generalization error since it is calculated just in the training set. Therefore, it is important to also calculate the error of the test and cross-validation sets and compare each, to find if the model fits in all sets.

When a learning algorithm did not perform as expected with high error it will be due to either a high bias problem or a high variance problem, also called underfitting and overfitting problems. An underfit machine learning model will have high values on the cost function performance on the training and test data and will not predict well, it occurs when the model can't perform in the training data nor generalize to test data, due to a lack of features, of a low number of samples.

Overfitting occurs when a network performs well with a specific training set and minimizes the error, but when used with a testing set, the error rate is much larger, which means it has a high generalization error. This often occurs when a network pays too much attention to unnecessary details called noise instead of paying attention to the signal.

An example of underfitting and overfitting is shown in Figure 3-8. One can see in that case a linear solution is underfitting the model and a high degree polynomial is overfitting, where all the training data match with the model results but will not work properly with other amounts of data. Figure 3-9 shows what the cost function might look like concerning the order of the polynomial, where J_{cv} is the error from the cross validation set and J_{train} is the training set error.

High bias - Underfitting
$h(x) = \theta_0 + \theta_1 x$

"Good model"
$h(x) = \theta_0 + \theta_1 x + \theta_2 x^2$

High variance - Overfitting
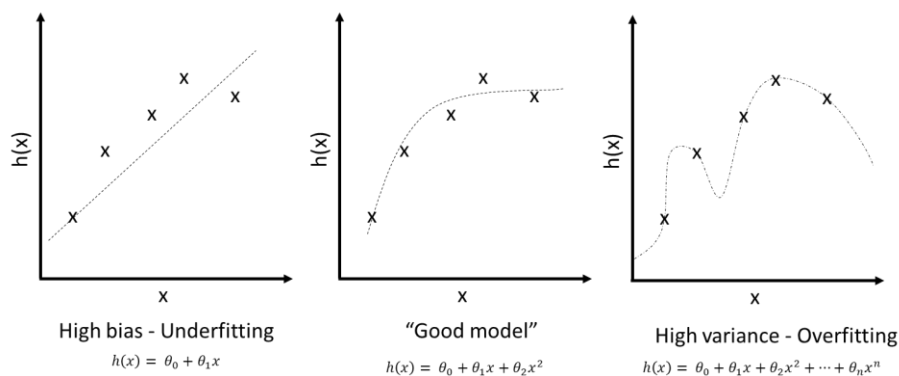$h(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \cdots + \theta_n x^n$

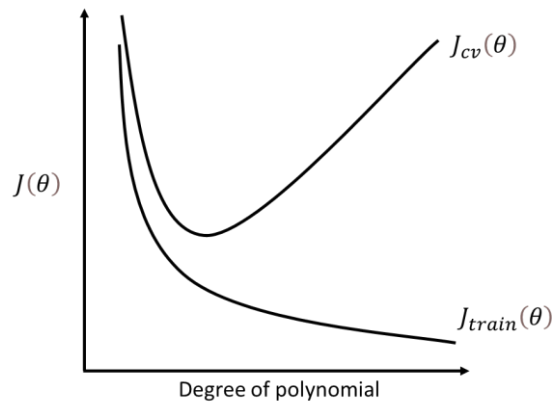*Figure 3-8 - Example of Underfitting and Overfitting*

Figure 3-9 - Cost function result in the function of the polynomial degree

To avoid the underfitting and overfitting problem, one can change some variables and methods to calculate the neural network:

- Add polynomial features– Verify not only if the linear solution fits well but try different inputs with the interaction between the variables. This could also increase the effectiveness of the model, where one can find a model that needs less training set to achieve the result.
- Change regularization value – Change from a small to a large value, e.g., 0.001 to 100. The decrease of λ tends to fix the high bias and oppositely, increase λ, tends to fix the high variance.
- Try a different set of features – Study which feature has an impact on the result.
- Get more training data – Get more samples and datasets to improve the model.

3.1.7 Stages

The stages in a feedforward artificial neural network are Forward Propagation, Error Calculating, Gradient Calculating, and Weights Updating. They are briefly explained below.

Forward Propagation

In this first step of the ANN, each value allocated in each node is calculated using the hypothesis function with the values from the previous layer, as explained in chapter 3.1.2, in that way calculating each node until the final result. Normally, the parameters are used with a unitary value, but there are some scripts to randomize these first values to try to improve the efficiency of the solution.

In Figure 3-10, one can see an example of how the forward propagation works, where each input is multiplied by a parameter and summed to others, including the bias node, to have as an output the next node in the next layer.
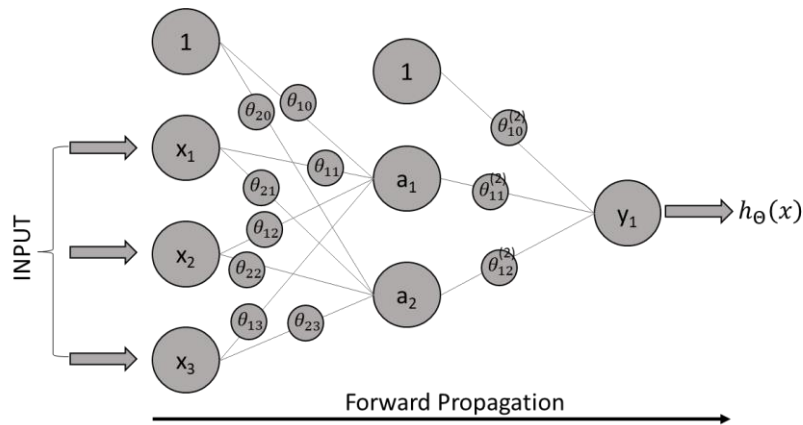
Figure 3-10 – Forward Propagation Example

Calculate the total error

The cost function is calculated as explained in chapter 3.1.4, to verify the error of the actual parameters in the ANN and to conclude whether it has an acceptable error and converged or has failed to converge and need new iterations to find a better solution. Also, it compares with the previous cost function result and verifies whether if the minimum value has already been reached. In this stage, one can previously choose the value of the regularization parameter to prevent outfitting, if necessary.

Calculate the gradients and update the weights

If the difference of the calculated cost function with the previous solution is large, a new iteration is needed. So, with the total error of the neural network calculated, the gradient descent algorithm is applied, as explained in chapter 3.1.5, in each node to find each node error and adjust each node parameter, with this, a new iterative process is restarted until the minimum value of function cost is reached.

3.2 Support Vector Machine (SVM)

Another type of supervised machine learning method is the Support Vector Machines (SVM) also used to analyse data for classification and regression analysis. It was developed by Boser et al. in 1992 [86] and it works differently as in ANN and other machine learning methods. Instead of minimizing the cost function, as explained in Chapter 3.1.4, this method is a training algorithm with the objective to maximize the margin between the training patterns and the decision boundary Figure 3-11, i.e., it finds a pattern in the data using vectors to define the decision boundary. SVM is largely used in classification problems, but one can modify the so-called kernel, as explained in Chapter 3.2.2, to be used as a clustering model, i.e., unsupervised learning, or a regression model, called Support Vector Regression (SVR) to be used with continuous variables as in the present study.
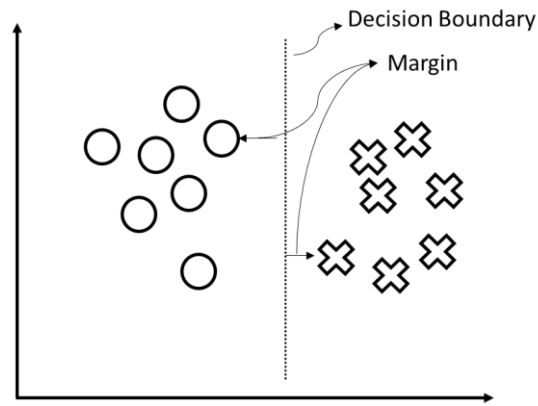
Figure 3-11 - Example of SVM

3.2.1 Optimization objective

This method uses a modified logistic regression method, previously presented, to do the supervised classification of the results. Using the sigmoid function, Eq. (3.9), for an expected result equal to 1, it is desired that $h_\theta(x)$ is approximately 1, that is, that the model has a good prediction and for this $\theta^T x^{(i)}$ must be much larger than zero. On the opposite side for a final value of zero, one expects $h_\theta(x)$ to be close to zero and thus $\theta^T x^{(i)}$ to be much smaller than 0. In Figure 3-12 one can see a graphical explanation.

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}} \tag{3.9}$$
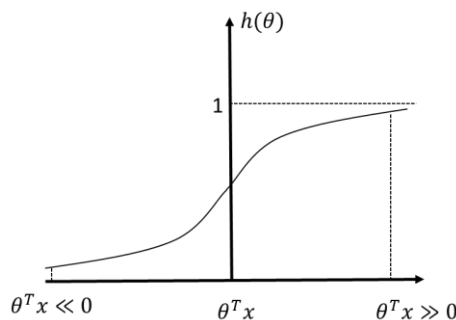
Figure 3-12 – Sigmoid function sketch

The form of the logistic regression cost function is modified to facilitate computational calculations and simplify the optimization method. For the best visualization of the solution, the cost function (Eq. (3.10)) and its altered form (Eq. (3.11)) are shown in Figure 3-13 for the conditions where the result is zero or one. The proposed modification makes that for cases where the result is equal to 1, for $\theta^T x^{(i)}$ values greater than 1, the cost function has its results at zero, being a straight line as in Figure 3-14 and for values less than 1, it is modified to a line segment parallel to the original cost function, similarly, the curve for the results (y) equal to zero is also modified, as shown in Figure 3-14. These new curves are

called as $cost_1$ and $cost_0$, with the first function to get the expected results as y=1 and the second as y=0.

$$J(\theta) = -(y \log h_\theta(x) + (1-y)\log(1 - h_\theta(x)))$$
(3.10)

$$J(\theta) = -(y \log \frac{1}{1 + e^{-\theta^T x}} + (1-y)\log(1 - \frac{1}{1 + e^{-\theta^T x}}))$$
(3.11)
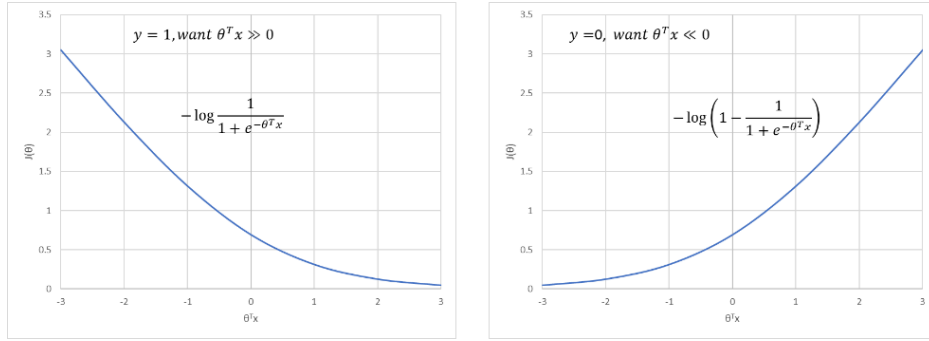
Figure 3-13 – Cost function of the logistic regression function
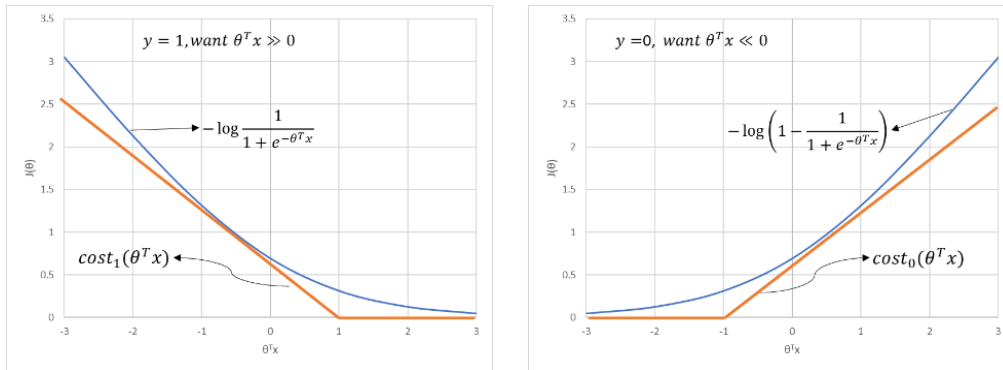
Figure 3-14– Cost function and the altered cost function of the logistic regression function

The total cost function is presented in Equation (3.12), but this is now adapted with the new functions $cost_1$ and $cost_0$, and in addition the regularization parameter λ is changed so that instead of giving high weight to the second part of the equation, it will be focused on the first part, i.e., unlike the ANN method, this method prioritizes to minimize the first part of the cost function and not only the sum of θ_j parameters (3.13). The new parameter $C = 1/\lambda$ is also called as regularization parameter.

$$min_\theta \frac{1}{m}\left[\sum_{i=1}^{m} y^{(i)}\left(-\log h_\theta(x^{(i)})\right) + (1 - y^{(i)})\left(-\log(1 - h_\theta(x^{(i)}))\right)\right] + \frac{\lambda}{2m}\sum_{j=1}^{n}\theta_j^2$$
(3.12)

$$min_\theta C\left[\sum_{i=1}^{m} y^{(i)} cost_1\left(\theta^T x^{(i)}\right) + (1 - y^{(i)}) cost_0\left(\theta^T x^{(i)}\right)\right] + \frac{1}{2}\sum_{j=1}^{n}\theta_j^2$$
(3.13)

With this new cost function $\theta^T x^{(i)}$ results are now calculated with a margin since it is no longer only greater or equal to zero (i.e., $\theta^T x^{(i)} \geq 0$) or less than zero (i.e., $\theta^T x^{(i)} < 0$) as in the ANN solution, but the solution has to be greater than 1 or less than -1 as already presented. This creates a distance between the results and the decision boundary, as shown in Figure 3-15. The figure shows the division between two classified results. Besides this, parameter C will be determinant to define the position of the decision boundary and its created margin, depending on the problem to be studied, an analysis of parameter C must be done to prevent overfitting and underfitting, since larger values of C tends to cause high variance and with small values, high bias.
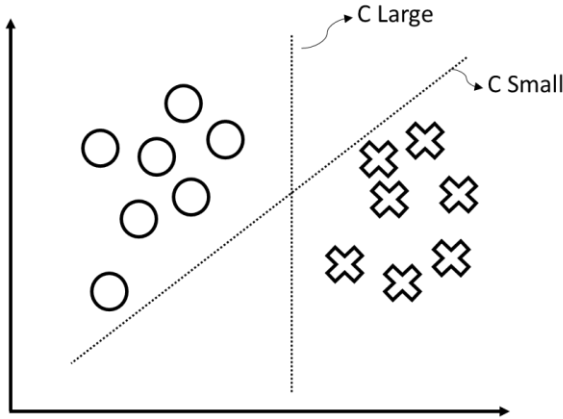
Figure 3-15 - Example of a Decision boundary definition

3.2.2 Kernel

The linear method cannot always find a decision boundary for the solution, even more, when there are non-linearly correlated data, as is the case in this study, so to use the SVM one must use a mathematical method that linearizes the relationship between the variables, referred as to the *kernel*.

The SVM can be used linearly as presented, where the system searches a set of θ to find the best solution to the linear problem (Eq. (3.12)), or it can use a kernel, which is a method of creating a new variable from a relationship between two features. So, the values of θ will be calculated using the new relationship (Eq. (3.15)), called also, of similarity, so the new equation to be solved by the SVR method remains as shown in Eq. (3.16).

$$y = 1 \; if \; \theta^T x \geq 0$$

$$\theta_0 + \theta_1 x_1 + \cdots + \theta_n x_n \geq 0$$

(3.14)

$$f_i = similarity(x, x^{(i)})$$

(3.15)

$$y = 1 \ if \ \theta^T f \geq 0$$

$$\theta_0 + \theta_1 f_1 + \cdots + \theta_n f_n \geq 0$$

(3.16)

To better exemplify, Figure 3-16 shows a classification relationship between two variables, one can see that this relationship is not linear, to solve this particular problem, a kernel function can be used such as in Eq. (3.17) to find a relationship between these variables and with that, be able to find the decision boundary. Still, Figure 3-16 shows how this new feature f modified the solution, making easier to find a decision boundary.

$$f = x^2 + y^2$$

(3.17)
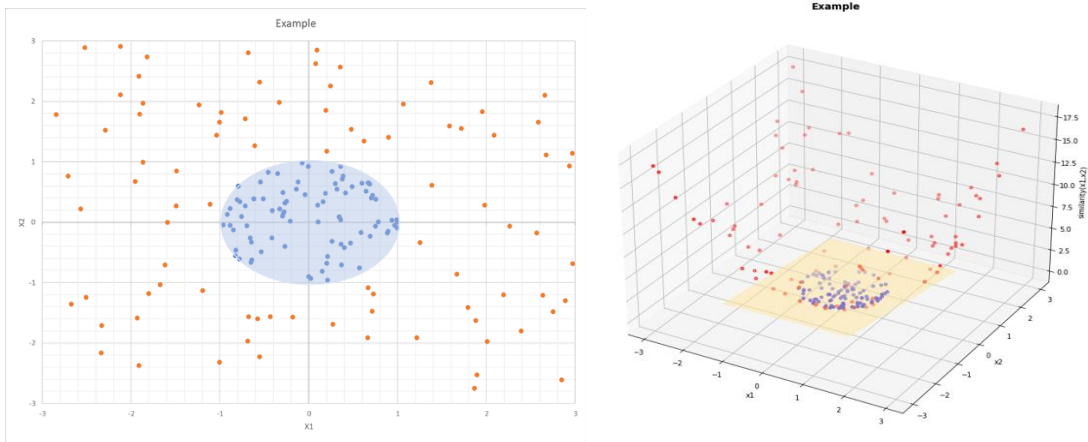
Figure 3-16 – Kernel new feature example

Several kernels can be used to solve non-linear problems. One of the most used classification problems is the Gaussian kernel, as in Eq. (3.18), where σ can be changed to solve higher bias or higher variance since these variable changes how smooth is the result of the kernel function.

$$f(x, x') = e^{-\frac{\|x - x'\|^2}{2\sigma^2}}$$

(3.18)

To the regression analysis, the default kernel used is the Radial Basis Function (RBF), as in Eq. (3.19). It is similar to the Gaussian equation but there is a γ value, where the defined default value is as in Eq. (3.20). It defines how much a single training example influences the entire model, also used to control the bias and variance of the problem.

$$f(x, x') = e^{-\gamma \|x - x'\|^2}$$

(3.19)

$$\gamma = \frac{1}{n_{features} * \sigma^2}$$

(3.20)

4. CASE STUDY

4.1 Vessel Automated Optimization System

An automated optimization system has been installed on the ship to optimize the use of propulsion system power and reduce fuel oil consumption [87]. The system proposes to reduce the impacts caused by environmental conditions during navigation, controlling the maximum fuel consumption set by the captain or chief engineer, with a real-time monitored system, receiving inputs from the environment and fuel consumption and shaft thrust meter. The automated system can adapt the shaft rotation, the pitch angle of the propeller and the fuel rack position of the main engine to ensure that the fuel consumption does not exceed the set value. In Figure 4-1, one can see a schematic of how the system works integrated into the Engine Control Room (ECR) and in Figure 4-2 is the panel example installed in the ship bridge. The ship has the main dimensions and machinery characteristics as presented in Table 4-1.

Figure 4-1 - Overview of the installed system [87]

Figure 4-2 - Bridge panel of the system [87]

Table 4-1 - Ship Characteristics

Ship Feature	Value	Ship Feature	Value
Ship Type	Container Ship	Propeller's Diameter	4400
Length - L	126 [m]	Propeller's Blade	4
Breadth - B	19 [m]	TEU	620
Draught - D	7.5[m]	DWT	8450 [t]
Main engine type	MAK8M552 6000 kW @500 RPM	Design Speed	16 [knots]
		Build-in	1994
Propulsion	Single CPP	Reduction Gear	3.36:1

As an example of how the system works, Figure 4-3 shows the graphs of the speed over ground (SOG), the fuel consumption of the system propulsion, the set maximum fuel consumption and the main configuration of the propulsion system, i.e., shaft rotation speed, fuel rack position and propeller pitch. This data shown refers to the route from Lisbon to the Azores, as in Figure 4-3, travelled in the period studied. One can verify as the set maximum fuel consumption changes, the system changes the configuration of the propulsion system to counterbalance the subsystems and find the best combination at which the propulsion system can deliver the maximum power, also it can be verified how accurate the system is since the target FOC and the actual FOC are quite identical.
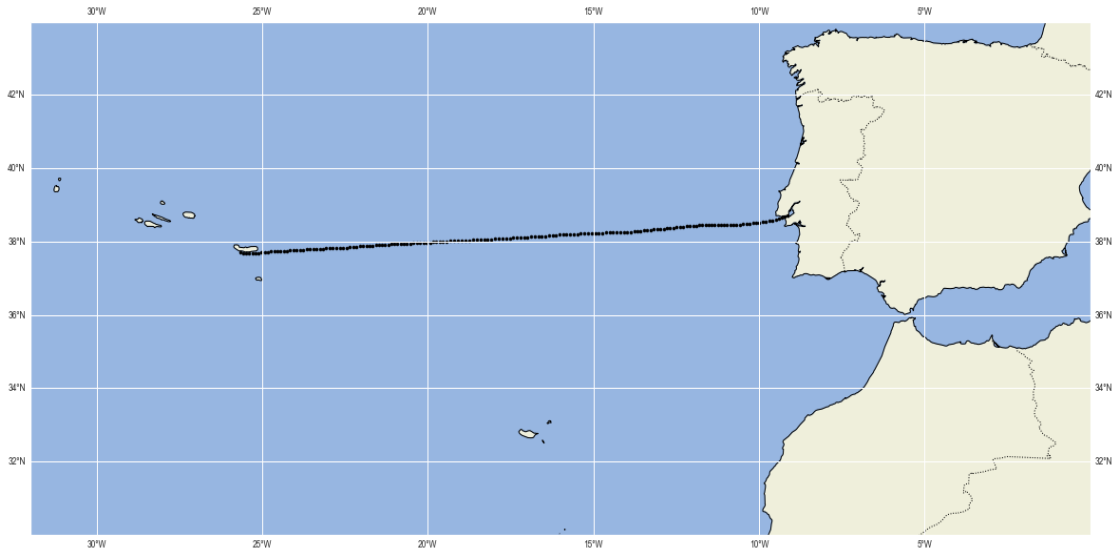

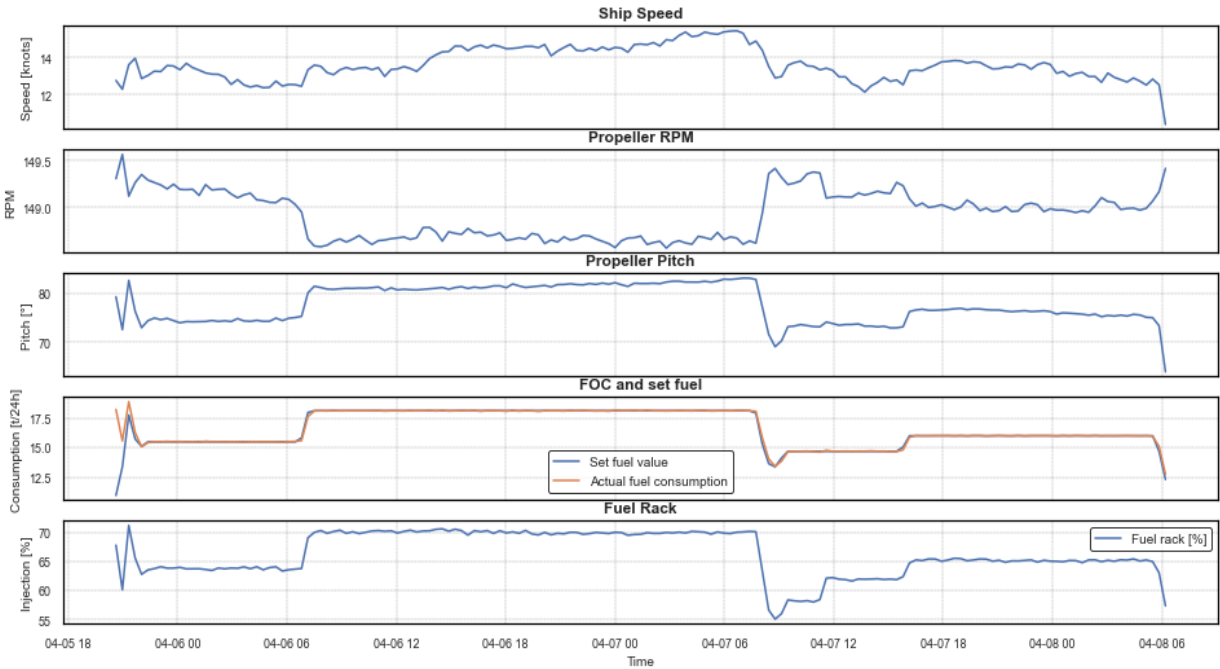*Figure 4-3 – Example of Route - Lisbon to the Azores*

Figure 4-4 – Sample data for the monitored routes

4.2 Dataset

Most of the accumulated data come from the automated recording system installed onboard. The system monitors various subsystems, all correlated with fuel consumption and energy spending. The list of variables recorded by the automated system is explained below.

- **Time:** The exact time of record, with seconds, minutes, hours, day, month, and year.
- **Latitude and Longitude:** The exact GPS position in degrees.
- **Speed Over Ground:** The measured momentary ship speed relative to the ground in knots.
- **Speed Through Water:** The measured momentary ship speed relative to the water in knots.
- **Apparent Wind Angle**: Angle at which the vessel faces the wind in degrees.
- **Apparent Wind Speed**: Speed of the feeling wind in [m/s].
- **Total Fuel Consumption**: Total fuel consumption of the ship in [t/24h] considering propulsion consumption, shaft generator and auxiliary engines consumption.
- **Total Propulsion Consumption**: Fuel consumption only for the propulsion system, without shaft generator, in [t/24h].
- **Total Propulsion Power**: Power generated only for the propulsive system in [kW].
- **Total Shaft Generator Power**: Power generated by the shaft generator in [kW].
- **Total Main Engine Power**: Total power in [kW] provided by the main engine for the propulsive system and shaft generator.
- **Main engine rotation per minute**: Rotation of the engine registered in [RPM]
- **Total Auxiliary Engine Power**: Total power in [kW] supplied by the auxiliary engines.
- **Auxiliary Engine Power 01, 02 and 03**: Power in [kW] supplied by each auxiliary engine.
- **Total Auxiliary Engine Consumption**: Auxiliary engine fuel consumption in [t/24h].
- **Propeller rotation per minute**: The shaft rotation speed in [RPM].
- **Propeller pitch angle**: This Ship has a CPP, so, this feature measures the angle between the rotor disc horizontal plane and the chord line of the propeller in degree.
- **Fuel temperature**: HFO temperature to be burned in the main engine in [°C].
- **System fuel optimization (ON, OFF)**: Feature that provides the information if the automated system is on or off using the Booleans numbers, as 0 when it is OFF and 1 if it is on.
- **Fuel consumption settled**: Indicator showing the maximum fuel consumption to be optimised by the system in [t/24h].

The total dataset available for this study has about 25019 samples, recorded every 20 minutes for exactly one year. Although having a large dataset with different variables, important variables were not in the dataset and were obtained separately. The draught and trim were obtained from the route reports, as explained in Chapter 4.5.3, and the local wave data by the Copernicus Climate Data Store [88], as explained in Chapter 4.6.

4.3 Redundant data

To have a reduced database, to facilitate the verification of correlations, reduce the size of computational memory usage and the computational calculation time, a redundancy analysis of the variables is carried out to verify which of them could be discarded.

This system presents some unnecessary data for this analysis, some of the variables that represent the energy generation are not independent and can be represented using linear relationships between them, this can avoid overfitting in the model. As an example, the Total Main Engine Power is the sum of the Total Propulsion Power and the Total Shaft Generator Power being these three variables redundant between them, also, each generator has its indicator, but, for this study, only the sum of the power generated by all auxiliary engines is sufficient. Table 4-2 summarises the variables that have been replaced or not used because they are redundant in the ship's power system.

Table 4-2 – Redundant Power Data

Not used Variables	Used Variables
Total Main Engine Power	Total Shaft Generator Power
Auxiliar engine 01 Power	Total Propulsion Power
Auxiliar engine 02 Power	Total Auxiliar Engine Power
Auxiliar engine 03 Power	

4.4 Original Data and Data Split

As showed in Figure 1-3, all the data points are registered with all the variables already listed, but, since the objective is to study the fuel consumption during ship navigation and its influences by the environment and operational decisions, it is necessary not only to check the reliability and redundancy of the dataset but also to identify the cut-off points to be applied on the data, to try to group them with the same patterns.

Figure 4-5 shows all the distribution of the velocity over ground datapoint and also, verify that there are two very clear distinct groups, a "zero-speed to a low-speed" situation group, where the ship is whether in manoeuvring to proceed towards or departs to the port or she is at the pier loading and unloading operation, and an "ongoing" situation group with a characteristic of a VOG greater than 10 [knots].

A data division is applied based on the distribution shown in Figure 4-5, having the speed of 10 [knots] as a cut-off value. One can verify how accurate is this division in Figure 4-6, Figure 4-7 and Figure 4-8 where they show the two new groups, the blue dots are the situation where the ship sails with more than 10 [knots] and the red dots when she has less than that speed, with this, it is clear that the proposed division can fit the objective of studying the fuel consumption used in the routes.
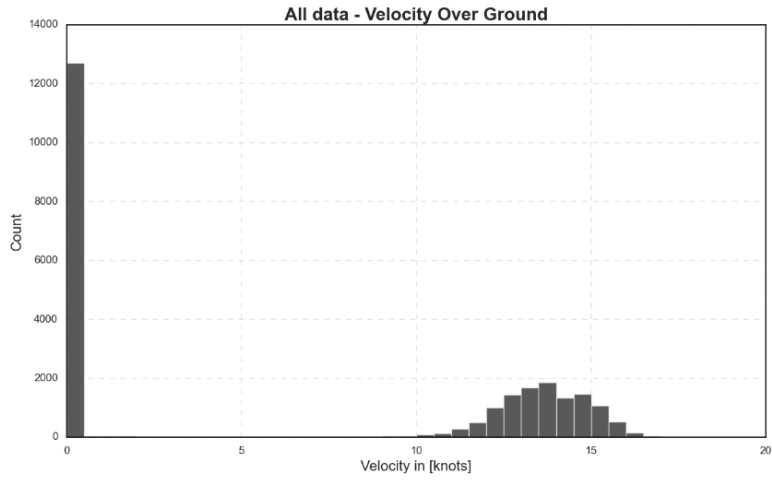
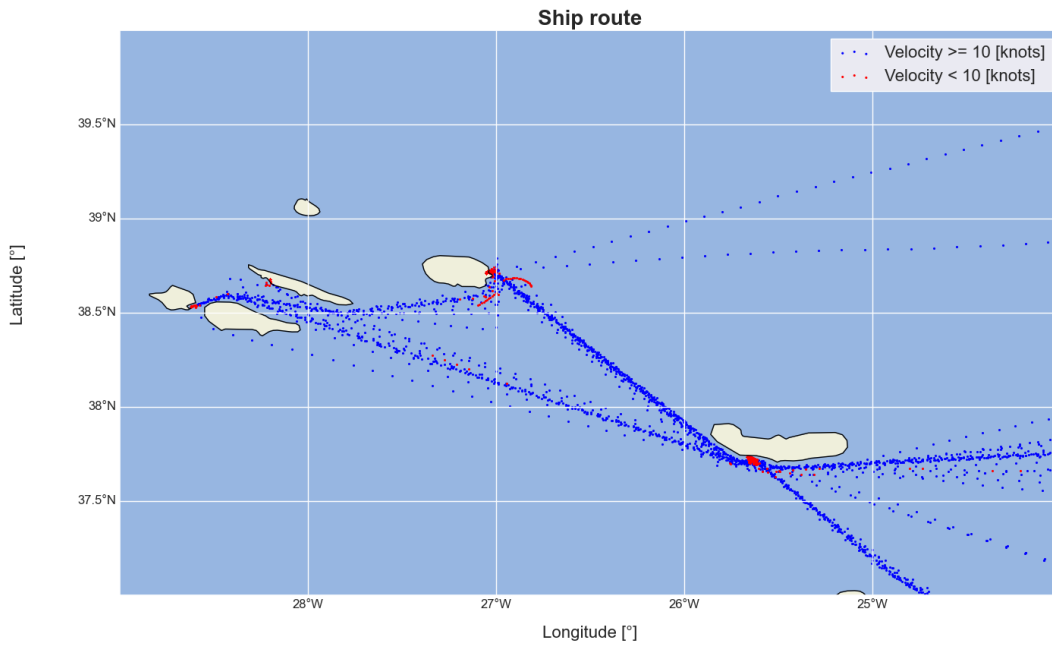*Figure 4-5 – Frequency Distribution of all Velocity Dataset*
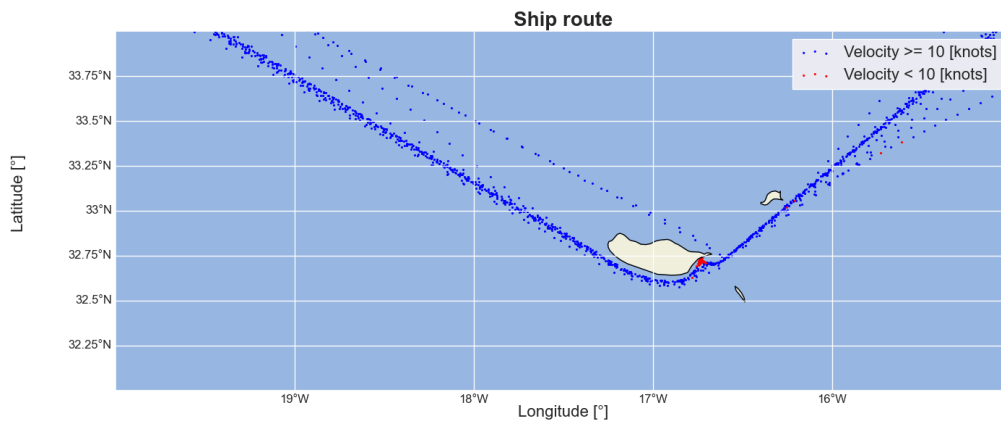

Figure 4-6 – Azores region
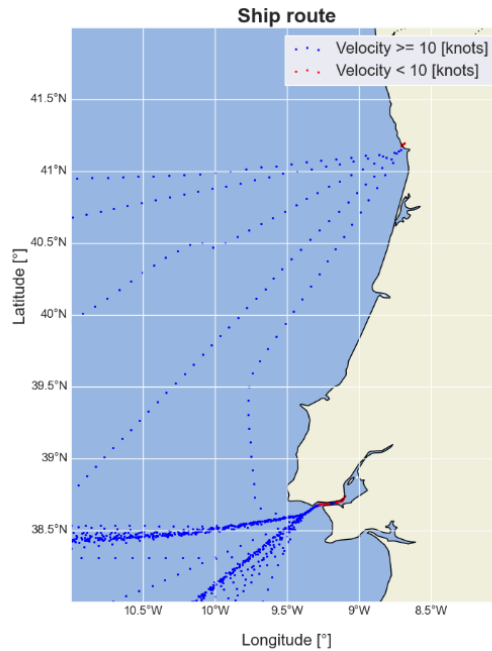

Figure 4-7 – Madeira Island

Figure 4-8 – Portugal Coast

4.5 Features treatments and previous analysis

After splitting the database to use only navigation with a speed above 10 knots, some treatments and adjustments are applied to the features to be able to verify the correlations between them, their consistencies and to analyse their behaviour.

4.5.1 System Fuel Optimization Setup and Set Fuel Consumption

The operability of the automated optimization system has two possibilities: on or off, registered as 1, if on, or 0, if off, in the dataset. But there are some situations the system shows numbers between 0 and 1, out of a Boolean pattern. Also, the fuel consumption set has a recorded number higher than zero whether the system is on or off. So, it is necessary to filter the data and get only two situations, if the system is on and operating with a set maximum value or if the system is off with the value set to zero, this is important to understand and find patterns between the set fuel consumption and the real fuel consumption, without the intermediate situation that can only cause noise in the system.

A filter is created so that if the system configuration values are different from zero and one, the data would be removed, also, if the system is off with the value defined as zero, the variable "set fuel consumption" is considered as zero. Figure 4-9 shows the initial system functional data without any filter nor cut-off, where one can see the outliers' points. Figure 4-10 shows the results when either the filter described and the 10 [knots] cut-off are used. The new condition has variability in the set FOC only when the system is switched on, avoiding correlations of falsely set values that do not represent the operational reality of the system.
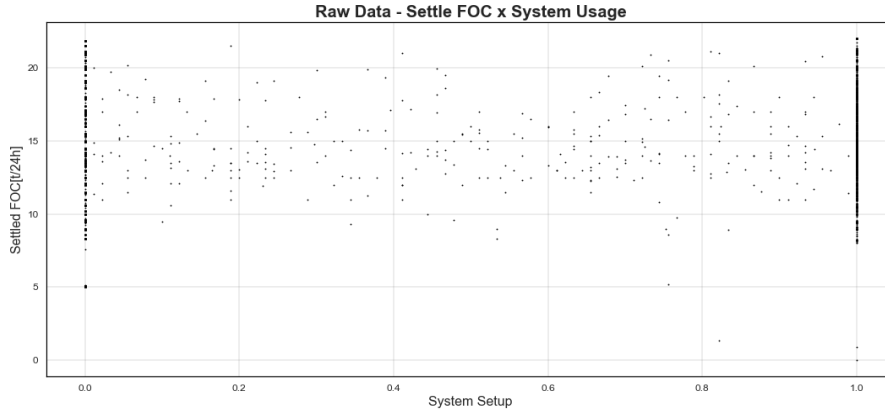
Figure 4-9 – Initial Data of the System setup
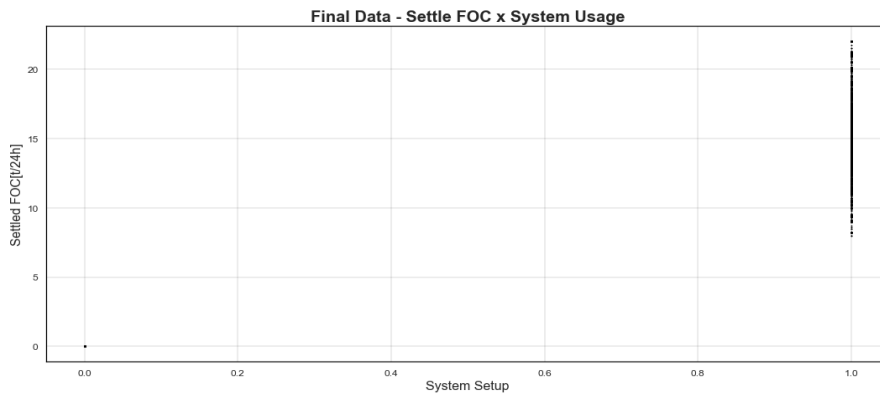
Figure 4-10 – Final Data of the System setup

4.5.2 The bearing of the route

The heading of the ship is needed to calculate the true wind angle also, it is an input of the learning model as it is correlated with the influence of the wave and wind in the ship resistance. As the instantaneous values of the ship heading are not available in the dataset, it is approximated by the bearing angle between two consecutive coordinates (i.e., latitude and longitude) recorded during the routes.

To calculate the bearing of the ship ξ_6, it is used the angle between the latitude and longitude coordinates, using the particular case of the law of haversines [89], that relates the points and angles in spherical trigonometry. For two given coordinates $(LatA, LongA)$ and $(LatB, LongB)$, the 2-argument arctangent relationship as in Eq. (4.1) is used for this calculation.

$$\xi_6 = atan2(X, Y) \tag{4.1}$$

where

$$X = \cos(LatB) * sin(\Delta L) \tag{4.2}$$

$$Y = \cos(LatA) * \sin(LatB) - \sin(LatA) * \cos(LatB) * \cos(\Delta L) \tag{4.3}$$

with $\Delta L = LongB - LongA$.

Figure 4-11 shows the distribution of the calculated bearing of the ship ξ_6 using the already cleaned up dataset, where 0° represents North and 270°, the West position. That means, if the ship has a heading angle of 270⁰, she is navigating to the west. One can note that the distribution of the course angle has a high concentration around a few values, as the ship is a liner only making a fixed round trip with regular schedule legs, this behaviour is expected.
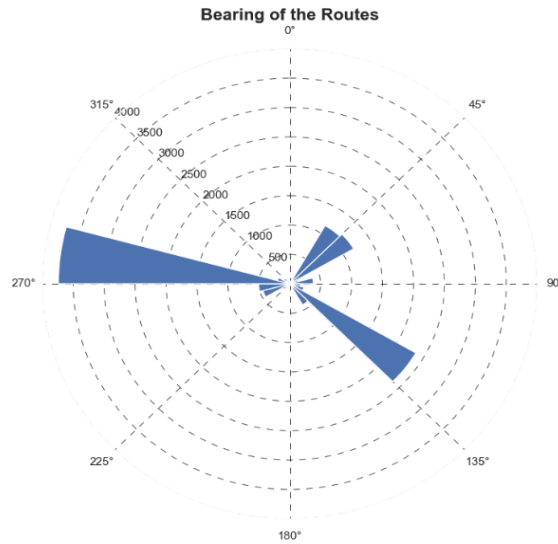
Figure 4-11 – Calculated Ship Route Bearing

4.5.3 Trim and Mean Draught Calculation

The forward and aft draught are recorded not in an automated way. They are given from the initial route reports when the ship is departing from a port. To calculate the mean draught and the trim, the well-known Equations (4.4) and (4.5) [90] are used.

Within the recorded draughts, the mean draught and trim are calculated, but it is important to notice that it is not necessary to calculate those new data to use them in ML methods, as they are pretty robust models and can learn even using brute data, understanding their impacts only using the forward and aft draught. But, to have more visibility of the impact and correlation between the trim and draught with other variables, also be conservative, as most of the studies use these features, those variables are calculated and implemented.

$$T_M = \frac{T_A + T_F}{2} \tag{4.4}$$

$$\Phi = \tan^{-1}\left(\frac{T_A - T_F}{L}\right) \tag{4.5}$$

where, T_A is the draft at the after perpendicular and T_f is the draft at the forward perpendicular.

Figure 4-12 and Figure 4-13 show the results where one can verify that in most cases the trim did not pass the 0.5°, but in the draught analysis, one can verify that she has a good dispersion between 5 [m]

and 7.5 [m], this is an expected draught variation for this ship type, as it operates in ballast and laden condition.

Figure 4-12 – Density Draught Distribution
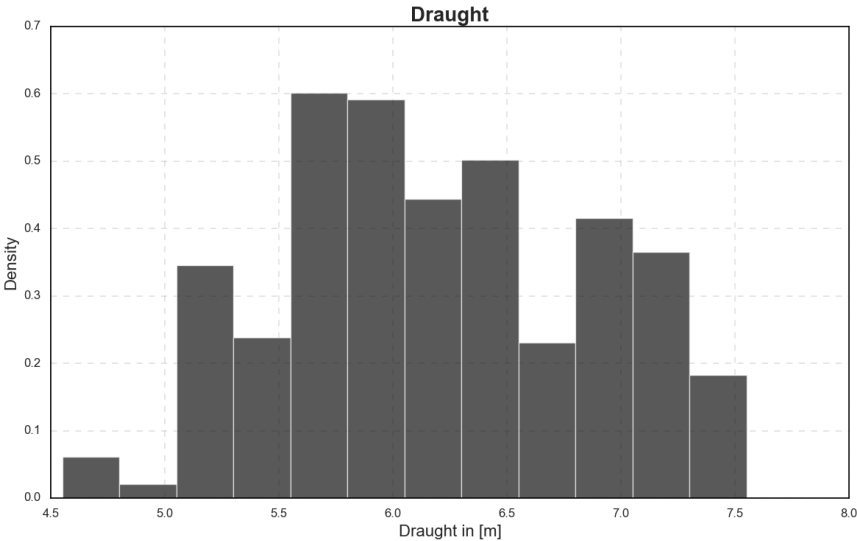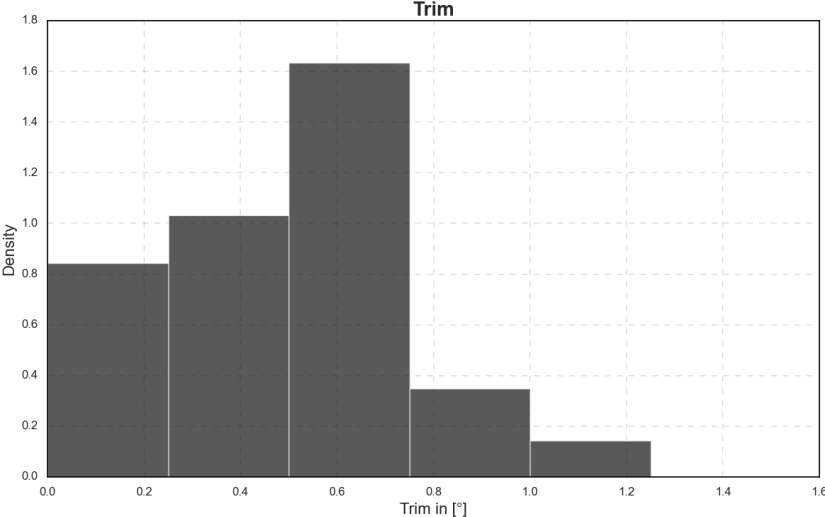

Figure 4-13 – Density Trim Distribution

4.5.4 True Wind Angle and Velocity

The apparent wind is the wind that the ship feels when in motion, it is a combination of the true wind and the observer relative velocity as delineated in Figure 4-14 and is described by a speed and an angle.
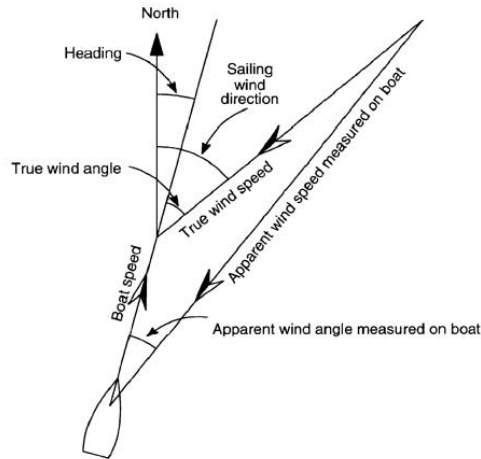
Figure 4-14 – Wind velocity triangle [91]

The wind measured in the ship is the apparent wind speed (AWS) and apparent wind angle (AWA) with the angle references as shown in Figure 4-15, but the true wind needs to be calculated to correct develop the prediction model because, as the operator does not know the future speed of the ship, the wind used as input data for the prediction model must be the true wind, with speed and angle, in the route region.

To calculate the true wind speed (TWS) and true wind angle (TWA) it is used the relationship between the vectors of the speeds as the sum of the apparent wind speed vector with ship speed vector is equal to the true wind speed, which results in the following equations (4.6) and (4.7).

$$TWS = \sqrt{S^2 + AWS^2 - [2.S.AWS.\cos(AWA)]} \qquad (4.6)$$

$$TWA = \cos^{-1}\frac{[AWS^2 - TWS^2 - S^2]}{2.TWS.S} \pm \xi_6 \qquad (4.7)$$

where, ξ_6 is the true heading angle of the ship in radians.

Figure 4-16 shows the apparent wind angle and apparent wind speed, and Figure 4-17 show the calculated results of the angles and velocities. As expected, the apparent wind angle is more recurrent at the bow due to the vectorial summation of speeds, as shown in Figure 4-14. Still, the perceived speed is higher than the real speed, as also expected by the formulation.
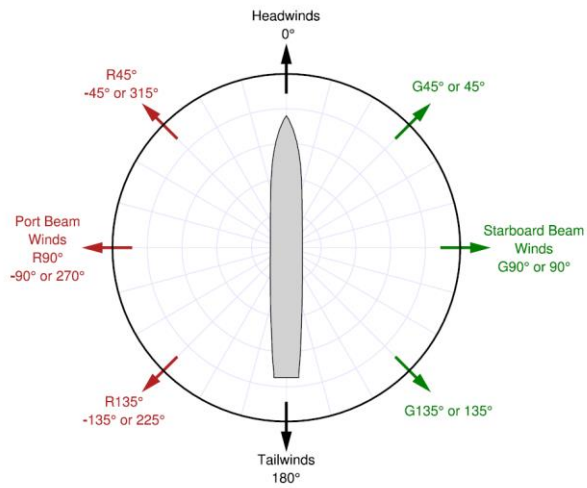
Figure 4-15 – Wind and wave definition reference angle [92]
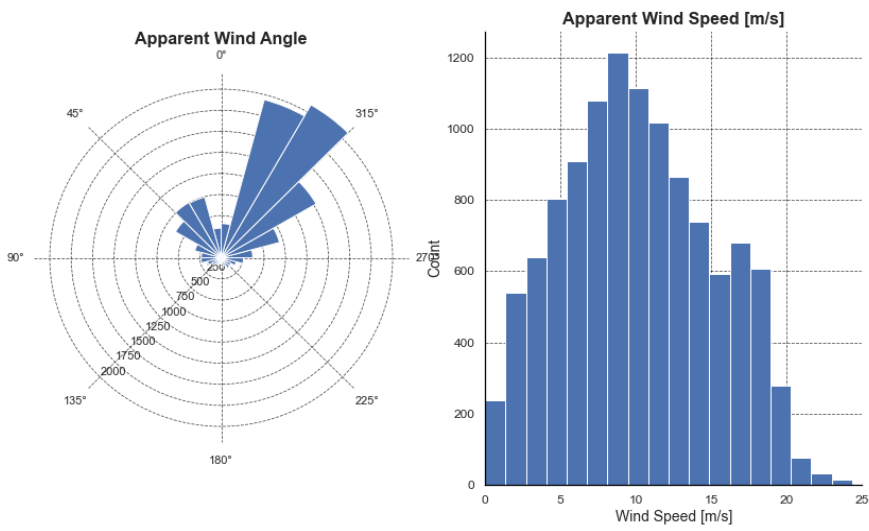
Figure 4-16 – Apparent Wind Angle and Apparent Wind Speed
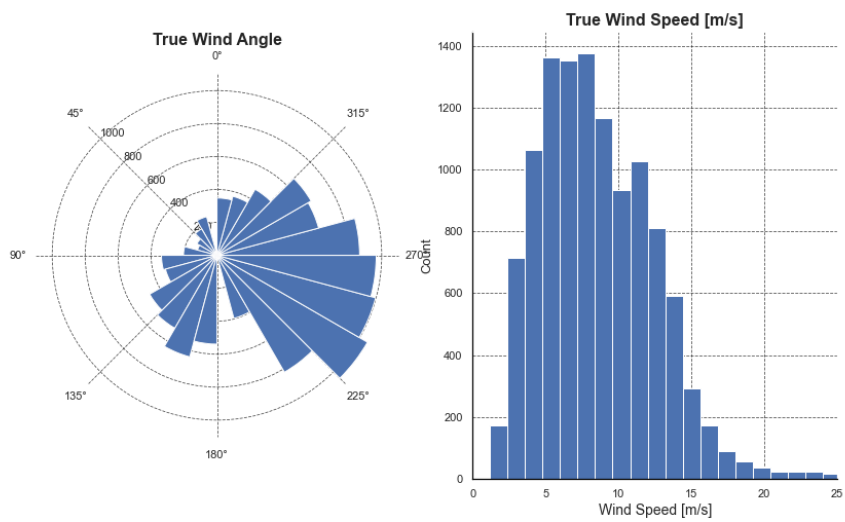
Figure 4-17 – True Wind Angle and True Wind Speed

4.6 Wave Dataset

The wave is one of the main factors contributing to the increase in the ship's resistance [58] and in consequence of that, the fuel consumption. To add these features the Copernicus Climate Data Store [88] is used, from this data store, it is obtained the "Mean Wave Period", the "Significant height of combined wind waves and swells" and the "Mean wave direction", to all the area and time of the ship´s route recorded.

The recovered data is between the 1st of January 2020 and the 31st of December 2020, recorded every hour, collecting data within the area between Latitude 42°N and 32°N and Longitude 5°W and 30°W. Using the opensource XyGrib [93] to visualize the data in Figure 4-18, one can see the area studied and compare it with the previous route in Figure 1-3. In this example, the software shows the significant wave height distribution at a random time in the chosen area.
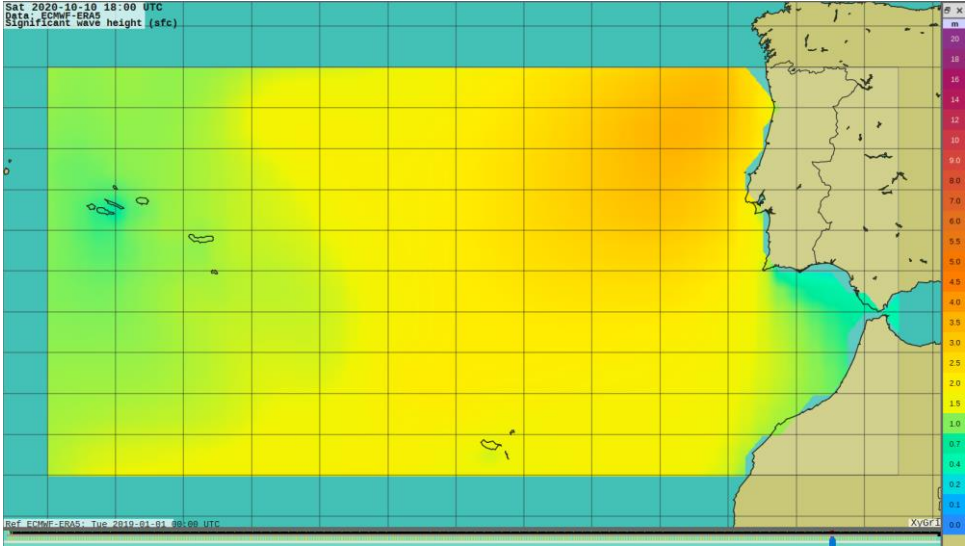
Figure 4-18 – XyGrib example of a Significant Wave Height plot data

A Python script is implemented to mount and analyse the wave data set. Considering each latitude and longitude position with its specific day and time, the sea state is investigated and added to the dataset for each ship position, during the studied period. Figure 4-19 shows the distribution and characteristics of the waves that the ship encountered in the voyages.
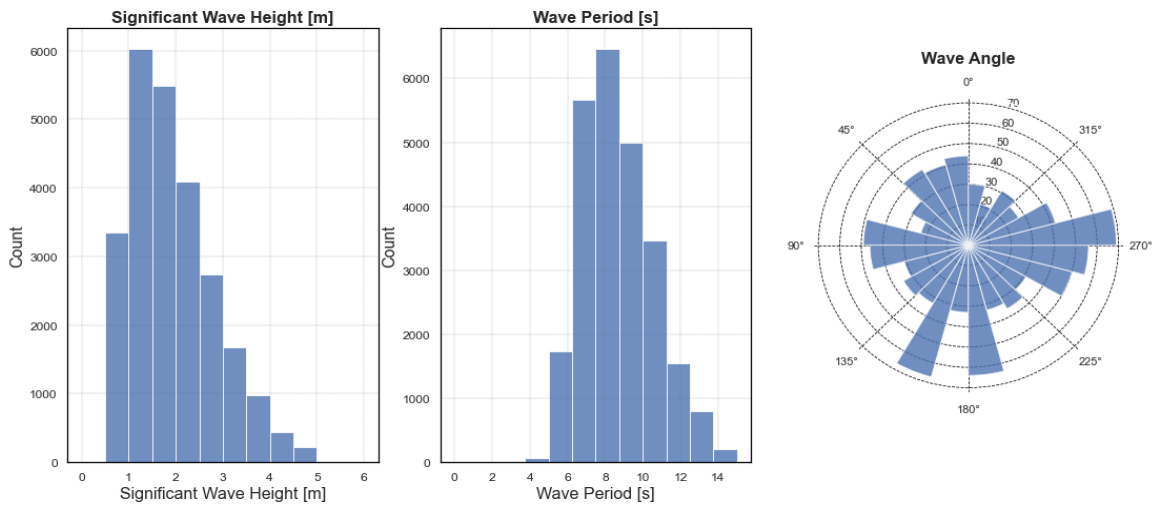
Figure 4-19 – Wave characteristics

4.7   Final Dataset

For the original dataset with 25019 records, only 11597 have the velocity equal to or above 10 [knots], as represented in Figure 4-20. Also, it is verified that in one route the draught configuration and trim values were missing and this dataset, along with all variables related to that specific time, have been removed from the dataset to guarantee the reliability of the results, since it is not guaranteed that all the learning models and scripts can handle with missing values. Table 4-3 summarises the dataset final characteristics.

Table 4-3 – Dataset cleaning summary

Dataset	Number of cases
Full database	25019
Equal or above 10 [knots]	11597
Less than 10 [knots]	13422
Missing datasets	258
Outliers	302
Final database	11116
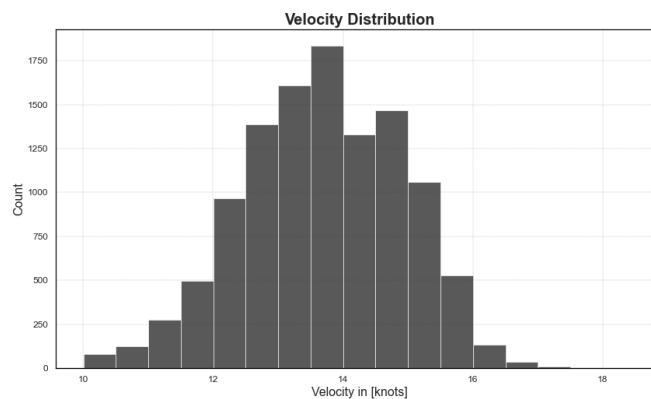
Figure 4-20 – Final Velocity Over Ground Distribution

5. DATA ANALYSIS

5.1 Spearman's Rank-Order Correlation

After splitting the databases, applying filters, and transforming and including other variables a Spearman rank-order correlation is applied to verify the correlation between each variable. That method was applied in [62] and [94] and helped the authors to find the characteristics that have the highest correlation, as this is not easy to visualise due to the non-linearity between the variables.

The Spearman rank-order correlation or Spearman's rho is a nonparametric correlation analysis that is used to rank the relationship between rank-ordered variables, similar to the database studied. This method assesses monotonic relationships and is appropriate for both continuous and discrete variables, linear or not. When a perfect monotone function between two variables occurs the correlation coefficient will be +1 or −1, as shown in Table 5-1, where it presents the coefficient range and its meaning between two studied variables. The general equation of Spearman's rho is represented in Eq. (5.1).

Table 5-1 – Spearman's rho coefficient [95]

| Correlation Coefficient for a Direct Relationship	Correlation Coefficient for an Indirect Relationship	Relationship Strength of the Variables
0.0	0.0	None/trivial
0.1	-0.1	Weal/small
0.3	-0.3	Moderate/medium
0.5	-0.5	Strong/large
1.0	-1.0	Perfect

$$\rho_{r_x, r_y} = \frac{cov(r_x, r_y)}{\sigma_{r_x} \sigma_{r_y}} \tag{5.1}$$

where $cov(r_x, r_y)$ is the covariance of ranked data r_x and r_y, and σ_{r_x} and σ_{r_y} are the standard deviation of r_x and r_y.

Due to the size of the complete result of the Spearman's rho analysis, the result of the comparison of all variables is in Appendix I, and partial analysis is discussed in the following chapters so that the relationships between subsystems and environmental variables can be better understood and studied.

5.2 Route analysis

The ship has fixed routes, as already shown in Figure 1-3, moreover, in Figure 5-1, one can see that the duration of the longer route does not exceed 65 [h] or nearly 3 days. The shorter routes last only a few hours, between 5 and 11 hours, which means that this ship makes a short sea shipping operation and differs from other container ship studies, as in [96]–[98], where they have deep-sea routes. This is a factor that hinders the study for route optimisation and the development of a weather route, as shown in [45], [75] and [76] because route optimisation using environmental variables would not have much effect due to the small variation in weather on the route under consideration. A model to improve the

ship operational performance, such as in [64], where the author studied the best engine speed for the optimal energy efficiency under different working conditions, fits better in this ship operation, and in fact, it is already underway with the past initiative of the shipowner to install an automated optimization system to maximize the use of the fuel consumption in this ship as explained in Chapter 4.1.
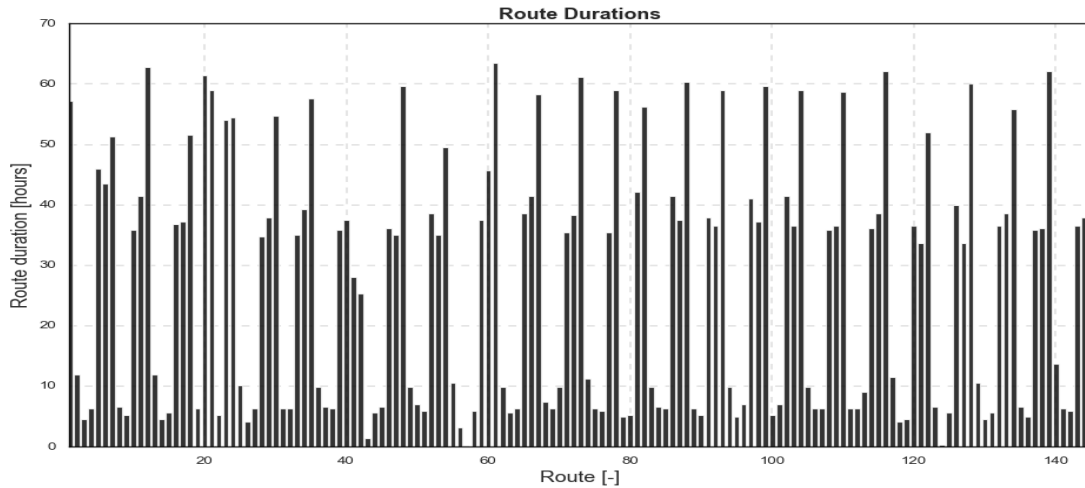
Figure 5-1 – Route duration in hours

With the correlation analysis done in Chapter 5.1, it is possible to verify the influence of environmental conditions on speed and fuel consumption using the database. One can verify in Figure 5-2 that there is a small influence of the weather on the speed and main engine consumption There are some hypotheses that can be formulated to understand why this unexpected lack of influence occurred. It could be because it is a short route, the variation between the weather conditions during the navigation time is small, so they end up not influencing the speed and fuel consumption in a way to be perceived by the numerical model. Or, still, the automated system combined with the constant changes of the set maximum consumption, works as a filter where the influence of the weather condition is felt on the fuel consumption and not on the speed. This could explain why weather has a higher Spearman's rho value with FOC compared to SOG.
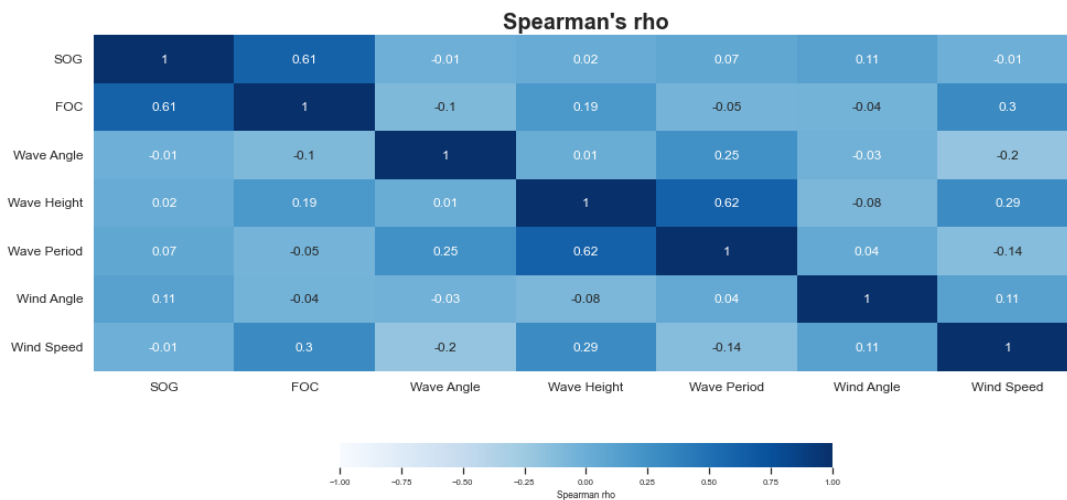
Figure 5-2 – Spearman's rho analysis between SOG and FOC with weather conditions

Furthermore, analysing the two variables that most influence fuel consumption, wave height and wind speed, the latter varying slightly during the year, as shown in Figure 5-3. A long-term moving average is used only to visualize the wind speed trend that the ship encountered in its route, it can be seen that it is practically the same throughout the year. The same is done for the significant wave height, however, this one can be verified that it is lower during summer compared to winter.
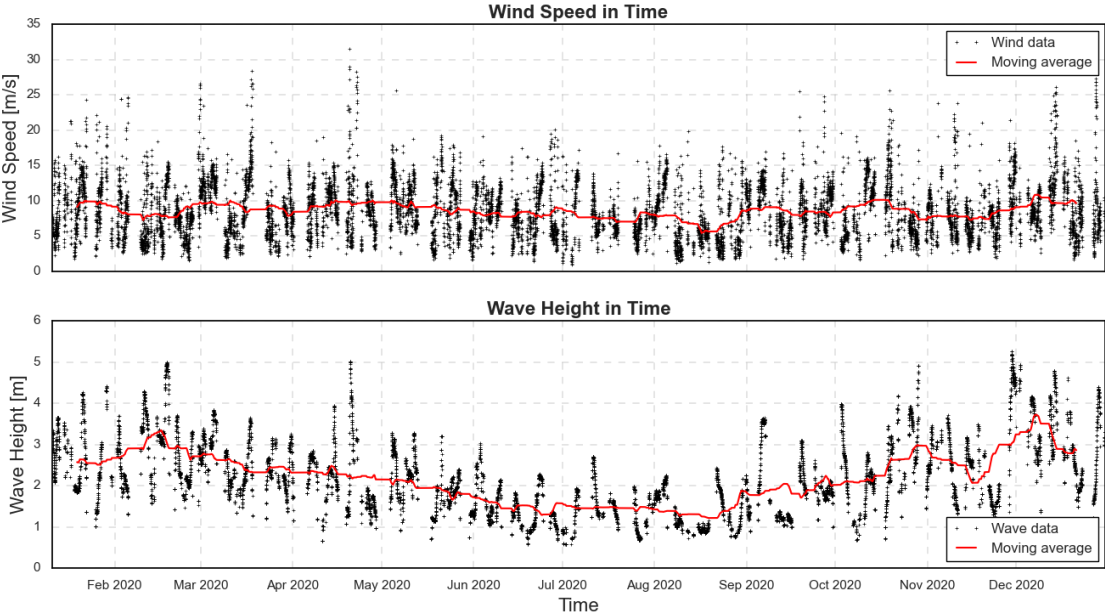
Figure 5-3 – Wave Height and Wind Speed distribution in time

Figure 5-4 shows the relationship between the ship's operating conditions, with its draught, trim and fuel consumption set. The trim and draught have little influence on speed and fuel consumption, something also not expected, as seen in some studies [72], [101]. This occurs because draft and trim are not varying with time, i.e., the variables are static throughout the routeing period, so this correlation fails to capture the influence of the two variables on speed and fuel consumption. And one can verify how high is the relationship of set consumption and speed and fuel consumption, the latter being almost perfect.
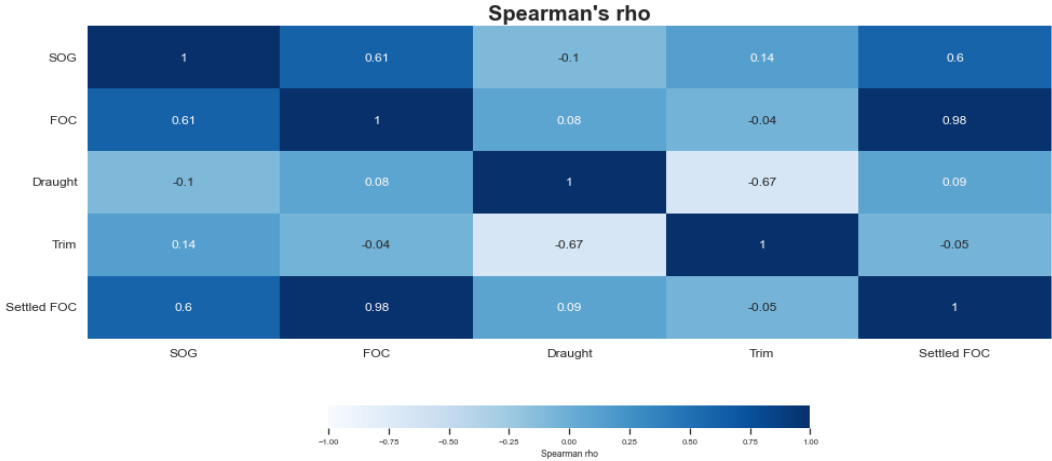
Figure 5-4 – Spearman's rho analysis between SOG and FOC with operational conditions

5.3 Automated System Analysis

Within the previous Spearman correlation results, in Table 5-2, it is analysed the relationship between the set fuel consumption variable with the actual propulsion consumption and its subsystems. The system receives some influence from wind speed and a slight, but not ignored, influence from wave height. This influence may occur due to variations in sea conditions during the route, where the captain/chief engineer is required to increase the maximum fuel consumption allowed to increase speed or at least reach a speed at which the ship will not delay its arrival at the port of destination. Further information needs to be collected on how the operator uses this system and mapping the decision-making process for changing the maximum consumption in the system.

Still, one can verify that the system has a high influence on the fuel rack position and propeller pitch, and these are the two subsystems most used to control the propulsion consumption by the automated system. An example is presented in Figure 4-4, where one can verify the system changing the fuel rack position, propeller pitch angle and shaft rotation speed to achieve the desirable fuel consumption. Moreover, the shaft rotation has a low correlation with the automated system due to the use of the shaft generator, when this is in use, the shaft needs to have a constant rotation, so the automated system is not able to change this rotation speed. This case is discussed in Chapter 5.4 and has an example of this particular situation in Figure 5-13.

Table 5-2 – Spearman Rho results for the automated system

	SOG	Propulsion Consumption	Propeller pitch	Fuel rack position	Shaft RPM	Wave Height	Wind speed
System Settled FOC	0.59	0.98	0.47	0.97	0.02	0.19	0.3

Linear regression is applied relating the set fuel consumption with the actual fuel consumption, as shown in Figure 5-5, where one can verify the high correlation between these two features, with the detailed results presented in Table 5-3. This regression has a coefficient of determination (R^2) of about 0.98, confirming the high correlation of Spearman's rho results. One can see that the propulsion consumption is almost the same as it is set in the system, in other words, the propulsion system always operates within the limit imposed by the system operator.

Table 5-3 – Linear regression analysis – Settled FOC x Actual FOC

Coefficient of Determination – R^2	**0.9833**
Intercept (Independent term)	0.5611
Slope	0.9606
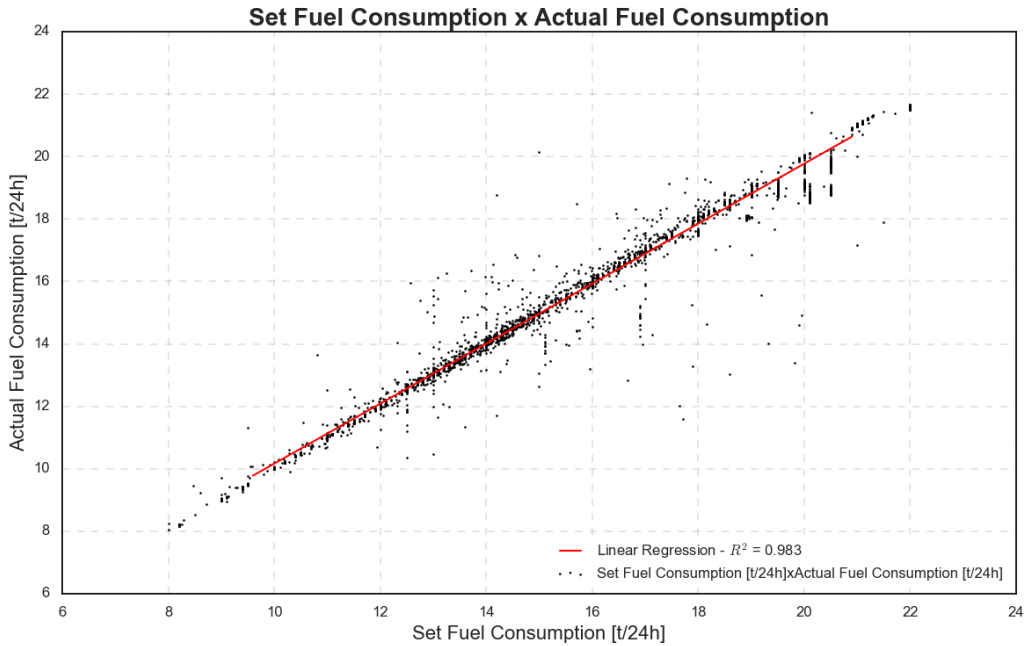
Figure 5-5 – FOC – settled versus actual fuel consumption

It is difficult to develop a model to predict the fuel consumption with this high correlation between the actual fuel consumption and the set fuel consumption, in addition, there is a high usage rate of the system, which is about 98% of the time travelled. This means that the ship will always operate at the system limit so there is no way to predict the fuel consumption as this will be whatever the system operator sets. The model then needs to be made to study whether it is possible to achieve a speed with given fuel consumption. This explains the low correlation between weather conditions and ship speed.

According to the result presented in Table 5-2, the system actuates more in the fuel rack position than in other propulsion subsystems. Moreover, in the scatterplot (Figure 5-6) it can be seen the high correlation between the set fuel oil consumption and the fuel rack position. Furthermore, a linear regression analysis is made to verify this relationship, shown in Table 5-4, where the resultant coefficient of determination is about 0.96, demonstrating the high inference of the automated system on the main engine fuel injection control system, or fuel rack.

Table 5-4 – Linear regression analysis – Settled FOC x Fuel Rack

Coefficient of Determination – R²	**0.9575**
Intercept (Independent term)	28.9750
Slope	2.2425
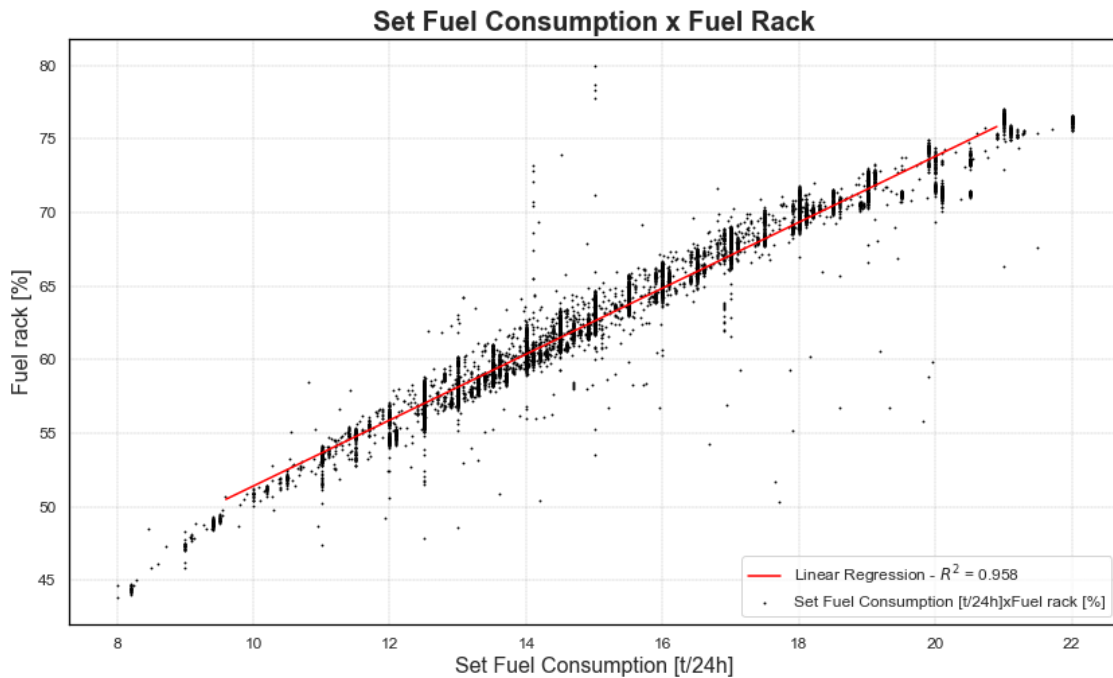
Figure 5-6 – FOC – set versus Fuel Rack

Figure 5-7 shows the dispersion graph of the ship's speed and the propulsion power of the main engine. One can see how this differs from the results expected by the already well known parametric relationships, such as Holtrop's [44] or ITTC [41], thus reinforcing, as already mentioned in Chapter 2, that there are many variations and uncertainties in ship operations that end up influencing the speed and fuel consumption relationship so that a prediction model can present high uncertainties.

The other two propulsion subsystems influenced by the automated system, are the propeller pitch and the shaft rotation speed. One can verify the scatter plot of them with the main engine power in Figure 5-8 and Figure 5-9, where on the propeller pitch graph is clear to see that there are some distinct operating patterns, some of them are mapped out and discussed in Chapter 5.4.

Also, on the shaft rotation speed graph, a clear straight line can be seen as the upper limit between the relationship between the shaft rotation speed and propulsion power, the same occurs at the lower limit of shaft speed, around 119 RPM. This represents the engine's minimum rotation point.

There is a clear distinct operating condition where either the propulsion system works at the operational limits of the ship's shaft with the engine rotation, or in an operational situation where the rotation is around 150 RPM. But in the latter case without any apparent pattern of operational limit condition nor high correlation behaviour. This behaviour is discussed in Chapter 5.4.

Figure 5-7 – Ship speed x Total Propulsion Power

Figure 5-8 – Propeller pitch angle x Total Propulsion Power

Figure 5-9 – Shaft rotation speed x Total Propulsion Power

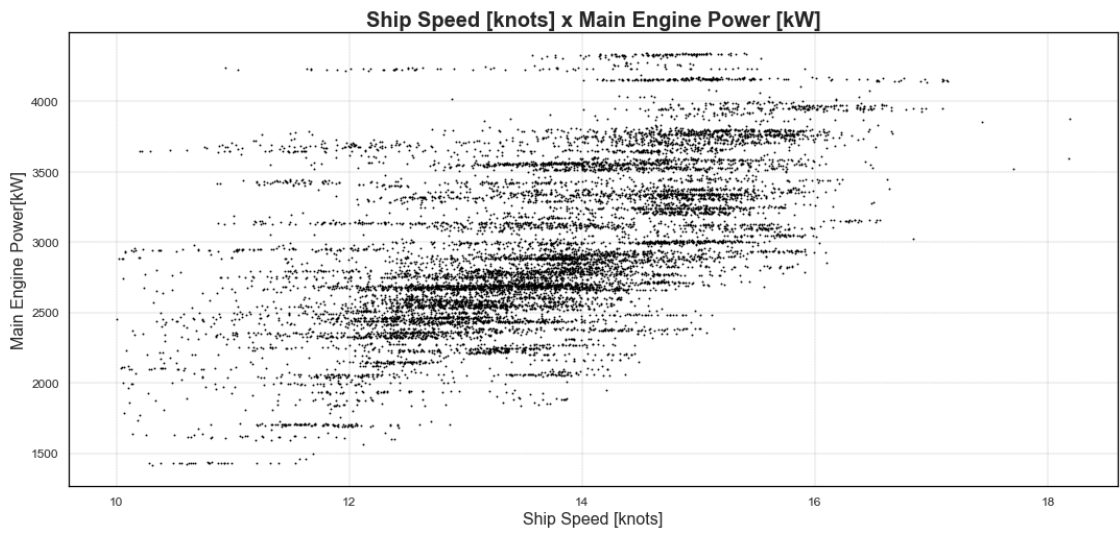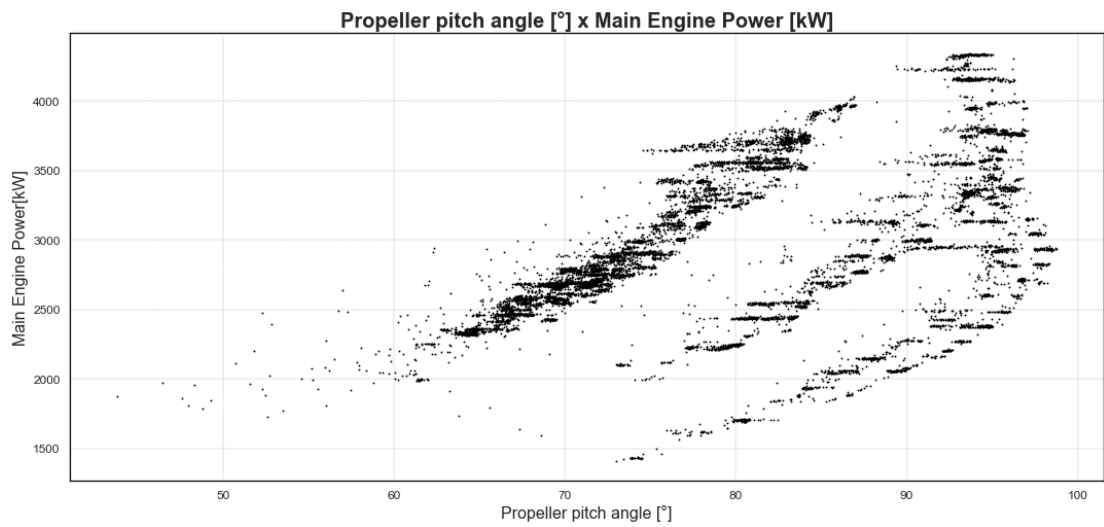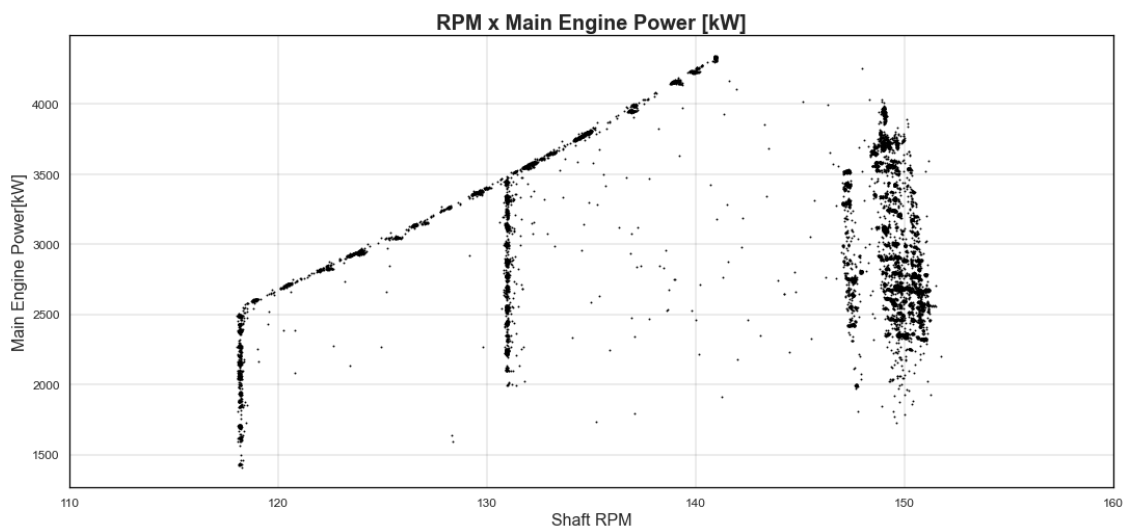## 5.4 Power Shaft Generator Analysis

With the analyses presented in Chapter 5.3, with different operational patterns in the operating conditions of the propulsive system, another system that influences the main engine is studied, the shaft generator. Where this system is directly accoupled in the shaft and use its speed rotation to generate energy. But this system needs some specific conditions to properly works, one of them is the fixed rotation of the shaft, so the automated system is not allowed to change the rotation speed to limit the fuel consumption, it needs to remain constant.

As one can see in Figure 5-10, the two shaft rotation patterns are divided based on the use of the shaft generator, where when the system is in use, the shaft needs to operate at a constant rotation speed and needs to compensate for the torque loss, operating at a high rotation speed.
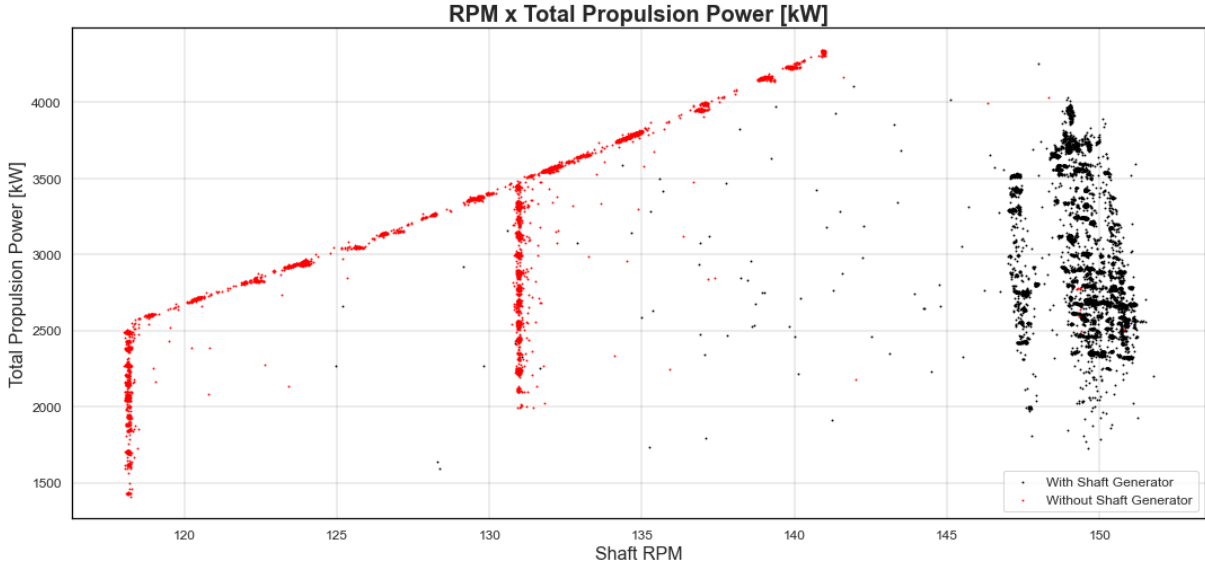
Figure 5-10 – RPM x Propulsion Power – Shaft Generator Analysis

Figure 5-11 shows the relationship of the propeller pitch angle and the total propulsion power, splitting the data between using or not using the shaft generator. It can be verified that when the shaft generator is running, the propeller angle is smaller to counterbalance the increase in shaft rotation speed and maintain the maximum fuel consumption established in the automation system. Furthermore, in Figure 5-12, one can see that the fuel rack does not show different patterns with the shaft generator on or off.

As an example, the related data to a whole route is shown in Figure 5-13, where one can see the speed, shaft rotation, propeller pitch, fuel consumption, fuel rack position and power generation by the auxiliary system over time. It can be seen how the system adjusts when switching from the use of the auxiliary generator to the shaft generator. Thus, one can see the different behaviour before and after the change, where, after the change, the shaft rotation speed increased by around 150 [RPM] and remains constant until the end of the route. The propeller pitch angle is reduced to compensate for the increased shaft speed and still achieve the maximum fuel consumption set. So, it is important to consider the use of the shaft generator in the decision model, as it influences the propulsion configuration.
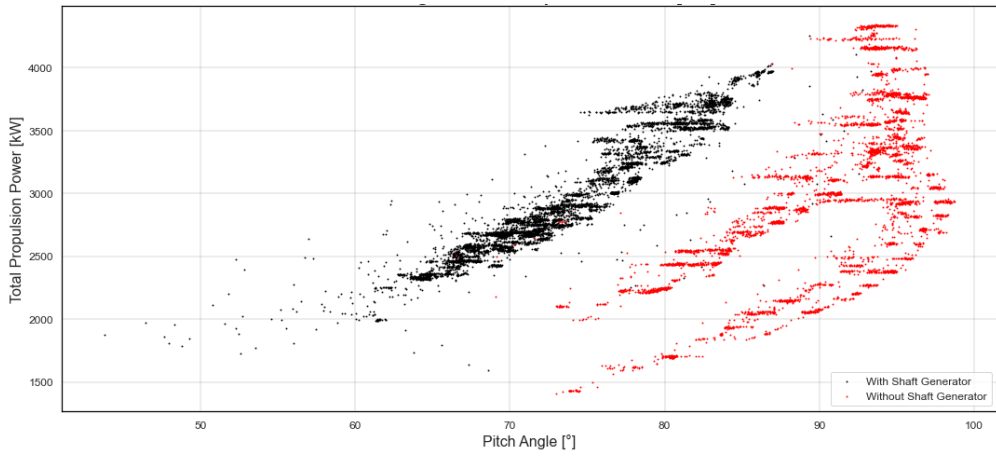
Figure 5-11– Pitch Angle x Propulsion Power – Shaft Generator Analysis
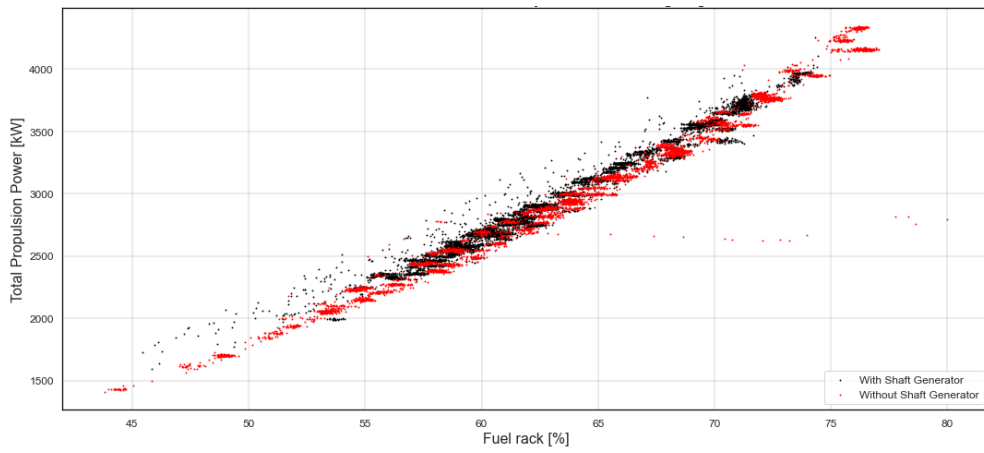
Figure 5-12 – Fuel Rack x Propulsion Power – Shaft Generator Analysis
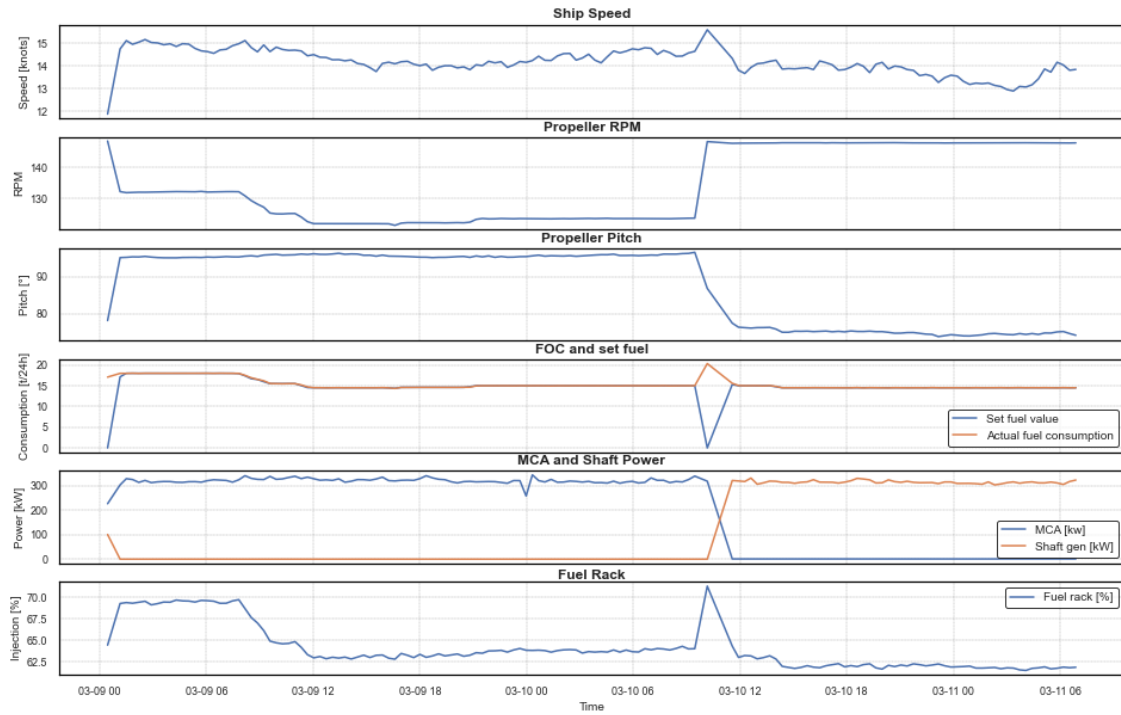
Figure 5-13 - Route data analysis for shaft generation

5.5 Auxiliary engine analysis

An analysis is performed to verify the use of the ship's auxiliary power generation system. The dataset of power generation by the shaft generator and of the total power production of the auxiliary engines is analysed. The results are presented in Table 5-5, where one can see that the two systems generate approximately the same amount of power for the ship. Also, they worked almost half of the time in the year each. However, it can be seen in Figure 5-14 that the shaft recovery system was not used at the beginning of the studied period.

Table 5-5 – Power generated by Shaft generator and Auxiliary Engines

	Shaft Generator	Auxiliary Engine
Using time [%]	56.8%	44.2%
Mean Power Generated [kW]	275.5 kW	270.2 kW
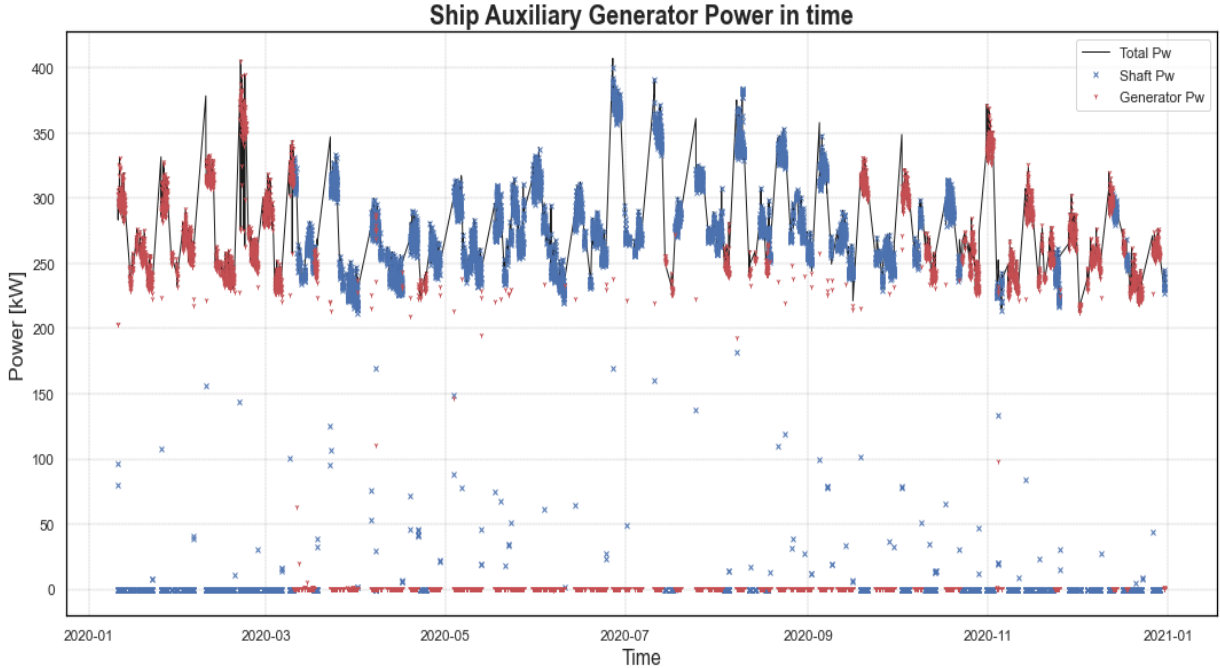
Figure 5-14 – Total Auxiliary Power [kW]

In Spearman's rho analysis (Appendix I), one can see that the systems are related to each other and the significant wave height and wave period. However, this could be incorrectly related due to some malfunction or other condition that caused the shaft generator to not work at the beginning of the period studied. New variables and data should be collected and studied to be able to relate which of the energy generation groups should be used in each situation and why.

6. FUEL CONSUMPTION PREDICTION MODEL

The fuel consumption prediction model aims to evaluate, for a given speed, what would be the propulsion fuel consumption of the main engine, and thus be a decision support tool for the shipowner or to the ship's crew, to analyse the possibility of reaching the port in a certain time, by spending a certain amount of fuel.

As already mentioned in Chapter 2.3, the prediction models using ML have achieved good results and are in line with the operational reality of ships, without the excessive cost of a complete hydrodynamic analysis, such as CFD, or even using parametric analyses that in some cases may have low performance due to several variables of the ship operation.

The first attempt to develop the model is performed according to the literature, e.g., [71], [72], where the variables of the ship configuration and weather conditions are used to find a model to predict the fuel consumption for a given voyage. However, this approach has not succeeded well, having an unsatisfactory performance with a score of about 0.88. This could occur because, in this situation, the automated optimization system installed in the ship interferes with the expected relationship between the environment and operation conditions and the speed of the vessel. This model and discussion are presented in Chapter 6.1.

To develop a machine learning model for fuel consumption analysis, a predictive model with the variables representing the propulsion system is proposed. These variables have as a characteristic the direct influence of the automated optimization system. This is a way to use as an input the automated system, without using the "set maximum fuel consumption" variable since it has a high bias, as already explained in Chapter 5.3. If this variable is used as input, whatever the value of this variable, a remarkably close value will be the result of the prediction model, in the same way as the data presented in Figure 5-5. This occurs because the learning method will give a large weight to this characteristic in its final solution, ignoring the other existing variables.

A two-stage model is proposed to solve this high influence problem. The first stage is developed to study and get a reliable prediction method for the speed over ground from a ship's configuration and weather conditions. The general model is presented in Chapter 6.3. The second-stage model is aimed to find an ML model to predict the fuel consumption for the propulsion system, as explained in Chapter 6.4.

The objective of the two-stage model is to verify in the first stage if a given SOG, for a given weather condition and ship configuration, is feasible and under what operating conditions, i.e., shaft rotation speed, propeller pitch angle and fuel rack position. The corresponding operating conditions are then used in the second stage to predict the propulsion fuel consumption, as shown in Figure 6-1.
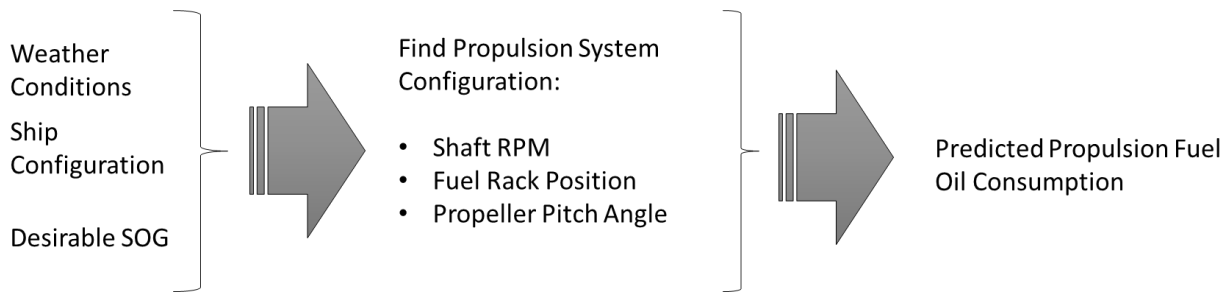
Figure 6-1 - Schematic of the two-stage model

6.1   Preliminary prediction model

The first attempt to develop a fuel consumption prediction model is based on all variables that could interfere with the ship's speed and thus cause an increase or decrease in fuel consumption, as shown in studies and literature already discussed in Chapter 2. These variables are the ship's operating conditions and the weather conditions of the voyage, as shown in Table 6-1, where one can see that the model aims to predict the fuel consumption as a function of wave and wind condition, the draught and trim configuration and the ship speed.

Table 6-1 – Training dataset to the first attempt model

Training dataset

Wave Angle	Mean Draught
Wave Period	Trim
Wave Height	Heading
Wind Angle	Wind Speed
	Ship Speed

The machine learning model used initially for developing the prediction model is the ANN, for which the range of hyperparameters is shown in Table 6-2. One can see that the number of hidden layers is tested from one to three so that it can be seen how the model improves with more layers. Also, different activation functions, such as hyperbolic tangent, ReLU, and the influence of the regularization value[3] are tested, similar to [68], [72], [102].

Table 6-2 – Hyperparameters of the ANN model – First model

Hyperparameter	Values/Type of function
Regularization Term - λ	$\in [0,1.28]$
Number of hidden layers	[1,2,3]
Number of nodes	$\in [2,100]$
Solver	Adam
Activation function	"tanh", "identity", "relu",

---

[3] This term is called α in the sktlearn used function neural_network.MLPRegressor

Although this model has performed well in other studies, with score values around 0.98 [68], in this case, its performance is not satisfactory. The resulting values for each number of layers are shown in Table 6-3, and, even when adding more layers, the results are not satisfactory.

Table 6-3 – ANN results to fuel prediction

Size of ANN	Number of Nodes	Regularization Term – λ	Activation	Score - R^2
1 hidden layer	(100)	0.0	tanh	0.7929
2 hidden layers	(100,80)	0.08	tanh	0.8538
3 hidden layers	(160, 100, 40)	0.02	tanh	0.8817

Still, Figure 6-2 shows the dispersion between the predicted values and the observed values, and one can see a great dispersion in the relationship between them, which cannot be acceptable in an operational prediction tool.
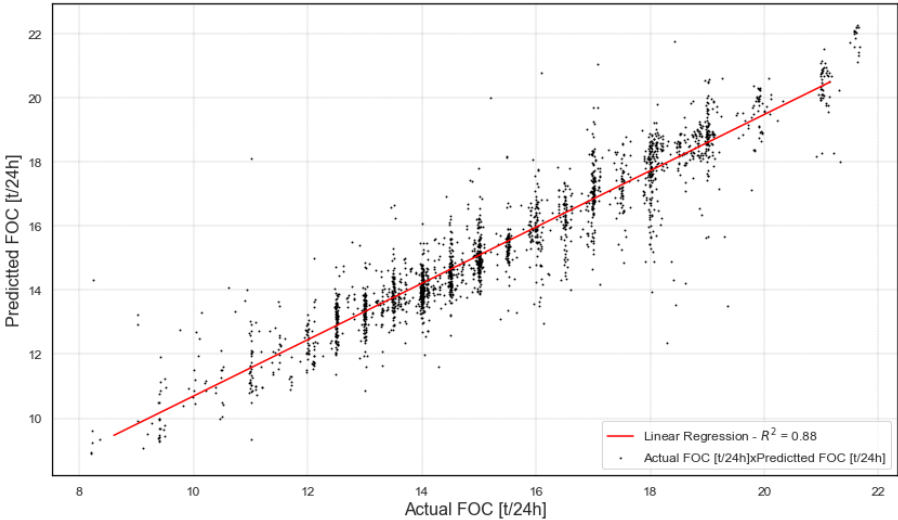
Figure 6-2 – Scatter distribution Predicted and Observed FOC – (ANN)

The Support Vector Machine (SVM) model is also used as it has also provided good results in other studies [72], [75], [103] and is a powerful machine learning method that can learn from data with non-linear relationships like the ones presented. The analysis is made by changing the hyperparameters, as shown in Table 6-4.

Table 6-4 – Hyperparameters of SVM model in the First model

Hyperparameter	Value
Gamma	$\in [2^{-15}, 2^0]$
C – Regularization parameter	$\in [2^0, 2^8]$
Epsilon	$\in [10^{-4}, 1]$
Kernel	Radial Basis Function

As expected, the SVM model has also not performed well, and its best result is shown in Table 6-10, being the best score of 0.88. In Figure 6-3 one can see that although the result has an expected trend, it presents many outliers, these being more dispersed at the limits of the calculated conditions and a little more concentrated between the 12 to 16 [ton/24h] conditions, but with many points outside the expected region.

Hyperparameter	Value
Gamma	1.45e-03
C – Regularization parameter	64
Epsilon	1.43e-01
Score (R²)	0.88
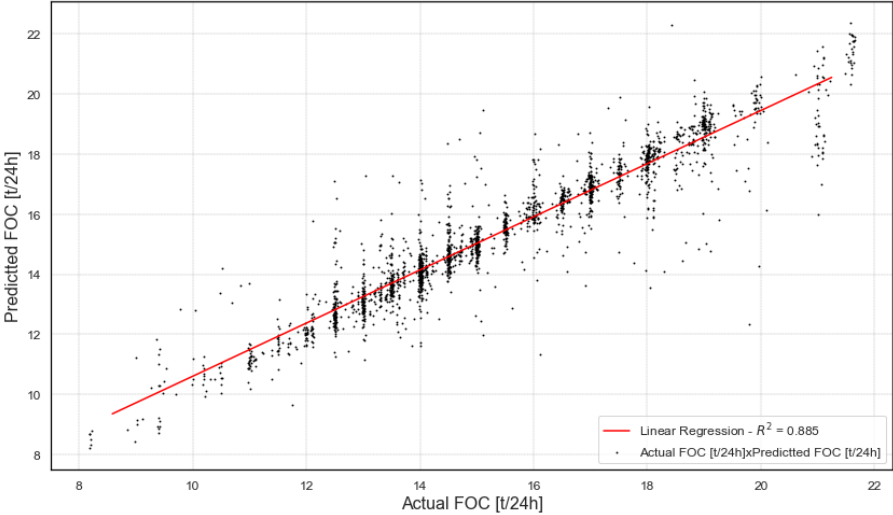
Figure 6-3 – Scatter distribution Predicted and Observed FOC – (SVR)

This dispersion becomes more evident when comparing the predicted results with the observations of the voyages studied. Figure 6-4 shows the consumption of four of the 145 voyages collected. It can be seen that several points are quite different from the original points. This can cause problems in predicting fuel consumption, where the best solution may not be chosen due to the distorting results as presented.

Moreover, the model can fit very well in some cases, as in Route 77 (Figure 6-4), but in other cases, varied greatly, as in the Route 138 case, where the real fuel consumption appears almost constant all the time, but the predicted FOC appears sometimes with 2 [ton/24h] less than the real case.

Moreover, the low prediction performance of the model may be also due to the variables not having great correlations among them. In Figure 6-5 one can see that only the wave height, wind speed and ship speed variables have some influence on fuel consumption. The lack of correlated parameters may be the cause of the underfitting of the model. This is not acceptable for a fuel consumption prediction model, where the main objective is to develop a reliable prediction tool. So, another model is proposed to achieve better prediction performances.
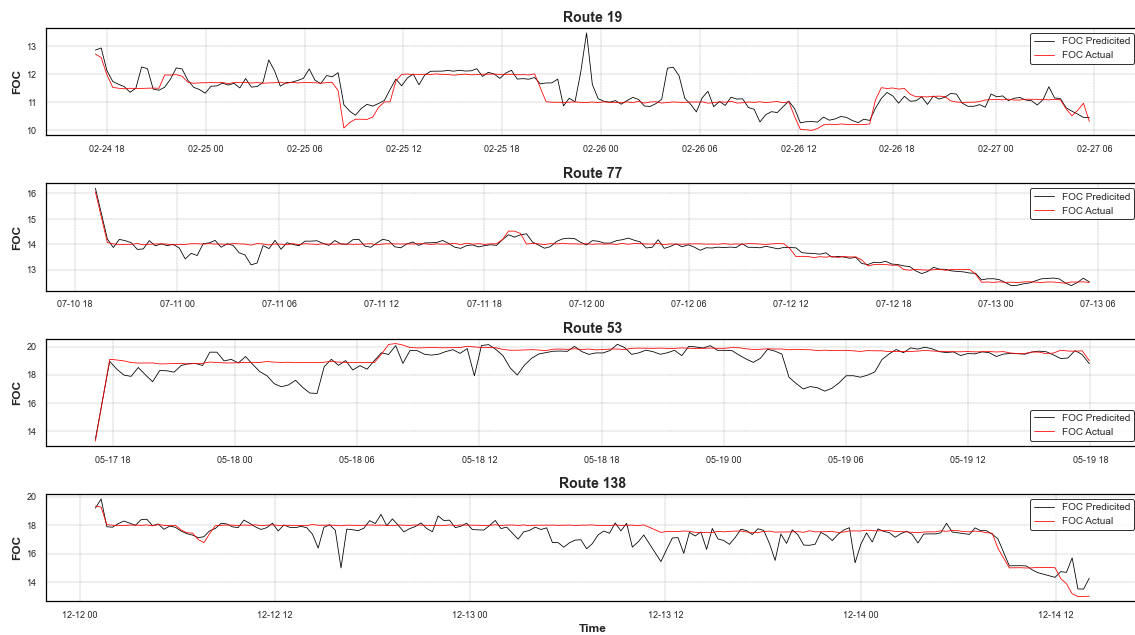
Figure 6-4 - Model comparison in different routes
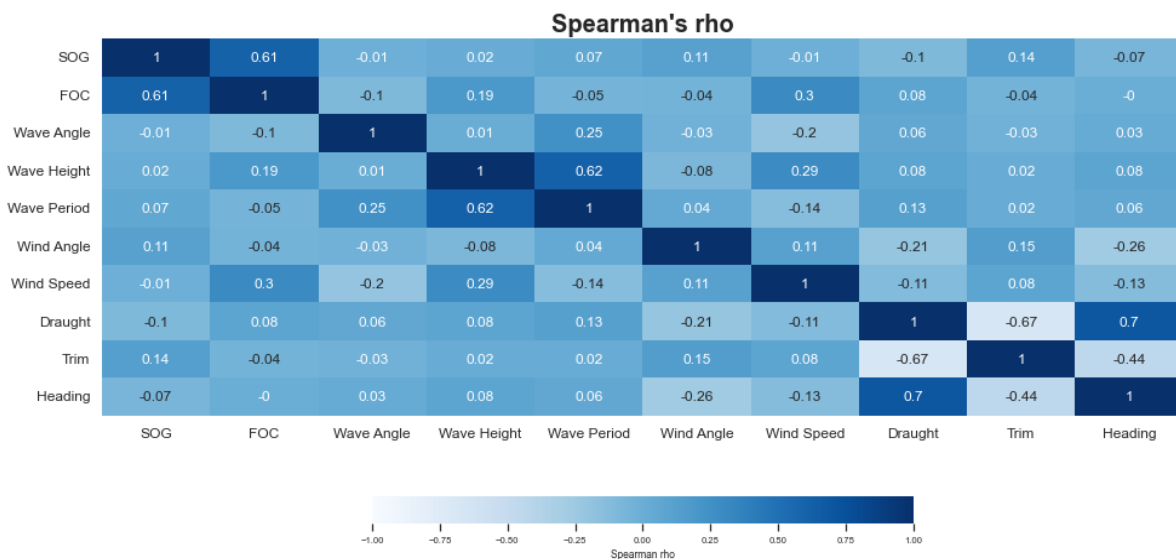
Figure 6-5 - Spearman´s Rho values

6.2 Adding new features

As described in Chapter 3.1.6, when the model is underfitting, new variables can be considered and checked to achieve a model with better accuracy. So, to build a new model, Spearman's relationship is used to analyse which variables have the highest correlation with the main engine fuel consumption. Figure 6-6 shows the results for the subsystems of the propulsive system, namely pitch and fuel rack position. As expected, they have a high correlation with main engine fuel consumption and ship speed. Also, since these two parameters have a direct influence from the automated system, it is included the shaft speed rotation, not only because it directly interferes with the ship´s speed, but because it can also

be used to analyse and calculate when the shaft generator system is used, as shown in Chapter 5.4, and how it will impact in the speed of the ship.

However, despite the system having very good predictability as will be seen in Chapter 6.4, to be used as a decision support tool it needs a previous step to analyse if the desired speed can be achieved and with which configuration of the propulsion system this can occur. This way, the model ensures that the ship's speed is consistent with the propulsion system parameters that are used for the fuel consumption prediction.
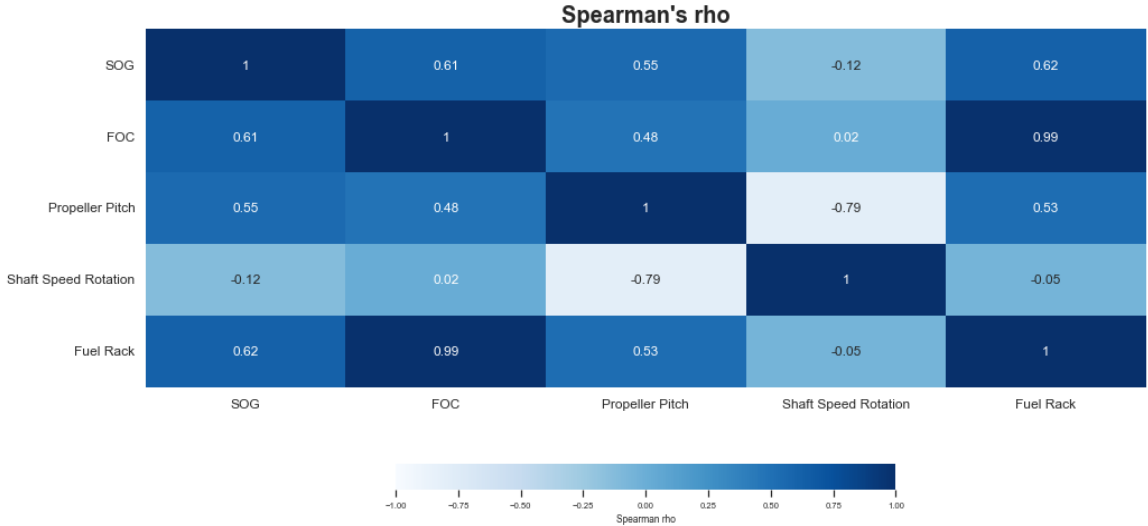
Figure 6-6 – Spearman´s coefficient of the propulsion system with speed and FOC

6.3   Two-stage model - the first stage

A machine learning first stage prediction model is developed using the features of the weather conditions, i.e., wave and wind, for each position of the studied routes; the ship configuration, i.e., draught, and trim and the propulsion configuration as fuel rack position, shaft rotation speed and propeller pitch angle. A schematic of the model is in Figure 6-7, where one can see the list of variables and methods used with their hyperparameters.
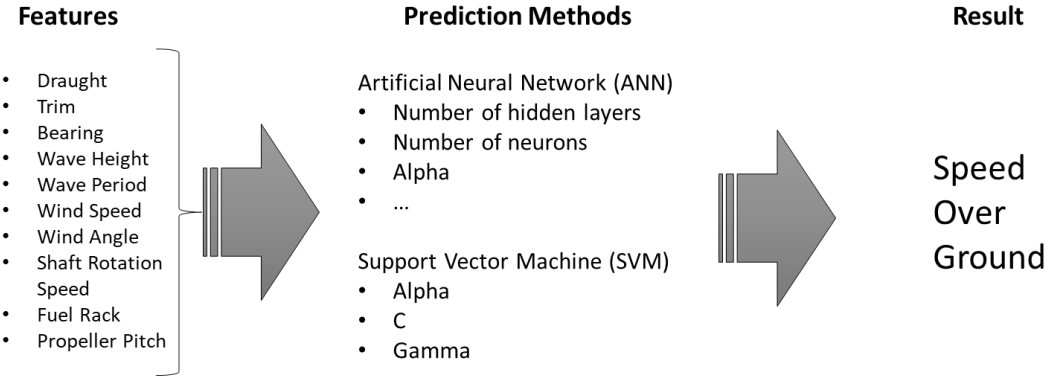
Figure 6-7 – Schematic of first-stage Prediction Model

Again, the first method used is the Artificial Neural Network (ANN). The combination of hyperparameters presented in Table 6-6 was tested to find a better convergence for this specific solution. The λ value is the regularization term that penalizes the cost weights in the cost function, which is used to avoid overfitting. The number of hidden layers and their nodes is modified to analyse how deep need to be this neural network to achieve a good prediction. Also, the solver Adam [104] is used to optimize the weight calculations and it is assessed which activation function provides the best result, evaluating the hyperbolic tangent, the ReLU and the identity functions.

Table 6-6 – Hyperparameters of ANN model

Hyperparameter	Values/Type of function
Regularization Term - λ	$\in [0,1.28]$
Number of hidden layers	[1,2,3]
Number of nodes	$\in [2,100]$
Solver	Adam
Activation function	Tanh, ReLU, identity

The data used to train the model to contain variables related to the local environmental conditions, route conditions, ship characteristics on the route and propulsion system data, as in Table 6-7. This initial dataset is based on the studies already cited, with the weather and ship configuration features, and the correlation analysis in Chapter 5.1, where it is included the configuration of the propulsion system.

Table 6-7 – Training dataset to the first stage

Training dataset – First stage	
Wave Angle	Mean Draught
Wave Period	Trim
Wave Height	Heading
Wind Angle	Shaft Rotation Speed
Wind Speed	Fuel Rack
	Propeller Pitch

To verify which ANN prediction set fits better within this dataset, a Python script is developed to calculate and analyse each configuration. The script calculates for each group of variables and provides a final score, and all the input and final results are included in a file for posterior analysis. There are a few software and scripts that could be used for this purpose. In this study, the Scikit-learn [105] is used, which is free opensource software in Python, largely used in academic research and studies to develop machine learning and artificial tools. Moreover, the function MLPRegressor[4] from this library is used.

The analysis used the dataset already pre-processed, i.e., with outliers and inconsistencies removed, as explained in Chapter 4. The dataset is split into two parts, the first is the learning dataset with 80% of the total data, and the 20% is the test dataset to calculate the coefficient of determination or score

---

[4] https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPRegressor.html

(R^2). Also, the MLPRegressor divided the 20% to use as a cross-validation test, to improve the efficiency of the learning method. Still, the dataset used for learning is random, that is, the dataset is randomised, with the records not being time consecutive, to prevent the model from having a high bias or causing overfitting.

Table 6-8 presents the results for each size of the neural network. One can see that with the growth of the network the performance of the model increases. Also, the regularization term is low, which means that the weights applied to the variables are not causing overfitting. Still, the activation function as a hyperbolic tangent is expected since in similar studies [72], [106] it had a better performance compared to the others.

Table 6-8 – ANN results to speed prediction

Size of ANN	Number of Nodes	Regularization Term - λ	Activation	Score - R^2
1 hidden layer	(100)	1.28	tanh	0.8128
2 hidden layers	(100,80)	5e-3	tanh	0.8567
3 hidden layers	(200, 100, 80)	1e-5	tanh	0.8887

To test and compare with other machine learning methods and to try to improve the score, a Support Vector Machine (SVM) model is developed, as it has been successfully implemented in some fuel prediction studies as already mentioned. In Table 6-9 one can see the hyperparameters that are used to analyse and find a good configuration for the prediction model.

Table 6-9 – Hyperparameters of SVM model

Hyperparameter	Value
Gamma	$\in [2^{-15}, 2^0]$
C – Regularization parameter	$\in [2^0, 2^8]$
Epsilon	$\in [10^{-4}, 1]$
Kernel	Radial Basis Function

The same dataset used in the ANN model is used by the SVM model, with the training and test data split, also. A script in Python is developed to calculate the prediction model of each set of hyperparameters configuration and each score using the Scikit-learn library already mentioned, using the Epsilon-Support Vector Regression[5] function to analyse the data. The best solution found is shown in Table 6-10 and the scatter plot of the results compared with the data is in Figure 6-8.

The model obtained has better accuracy than the ANN. Besides that, the regularization parameter did not extrapolate to the maximum that it could, that is, the solution obtained avoided overfitting with high regularization parameters. The gamma value found is small and shows that the model found a solution where the kernel calculation does not vary smoothly, having minor variation, which could cause overfitting in some models.

---

[5] https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html

Table 6-10 – SVM results to speed prediction

Hyperparameter	Value
Gamma	4.00e-04
C – Regularization parameter	32
Epsilon	1.42e-01
Score (R²)	0.9256

Predict Ship Speed x Actual Ship Speed
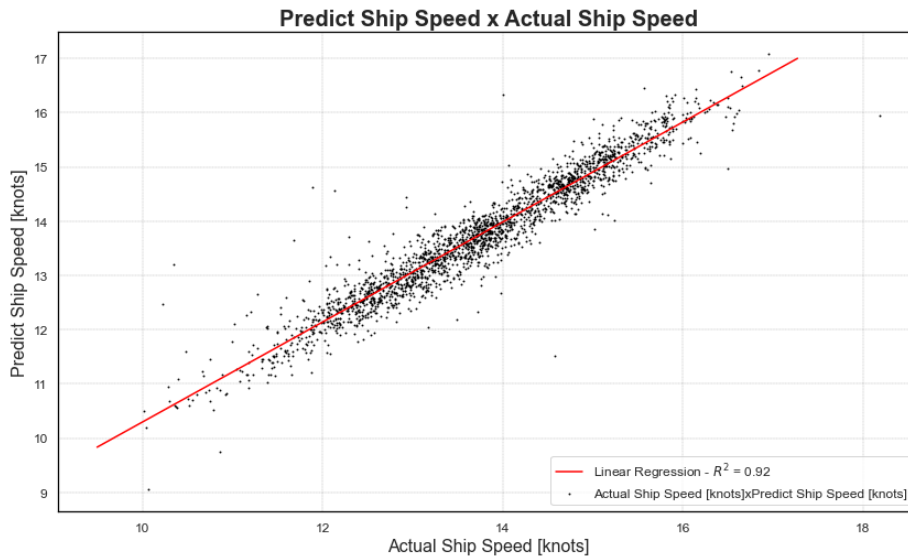
Figure 6-8 – Scatter plot result of the predicted and actual value - SOG

A sensitivity analysis was performed to check what is the minimum dataset size for the result to convert to the best result. In Figure 6-9 one can see the result for cost function for each test set size, where one can see that the model converges with a dataset starting at 5000 samples. Similar occurred in Figure 6-10, which shows the score of each test set size.

SVM Speed Model Cost Function
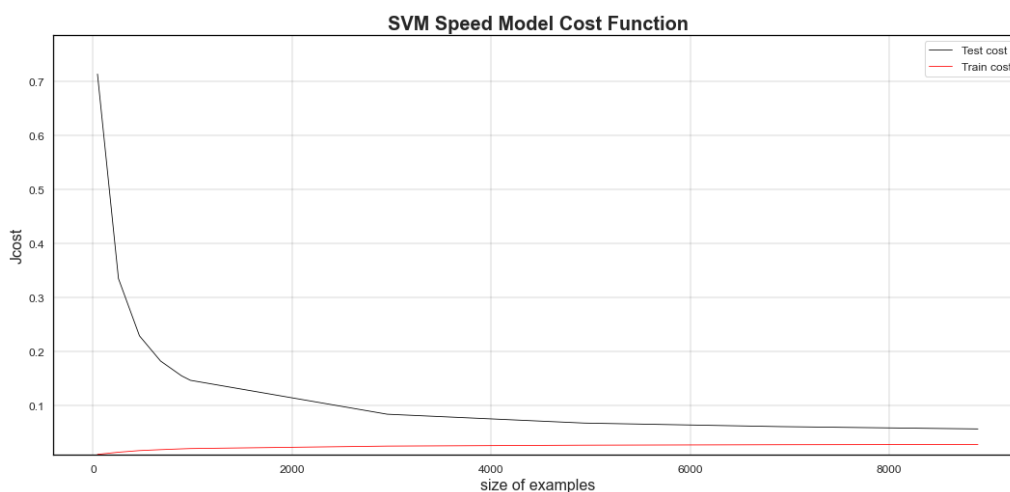
Figure 6-9 – Cost functions analysis in the function of training set size
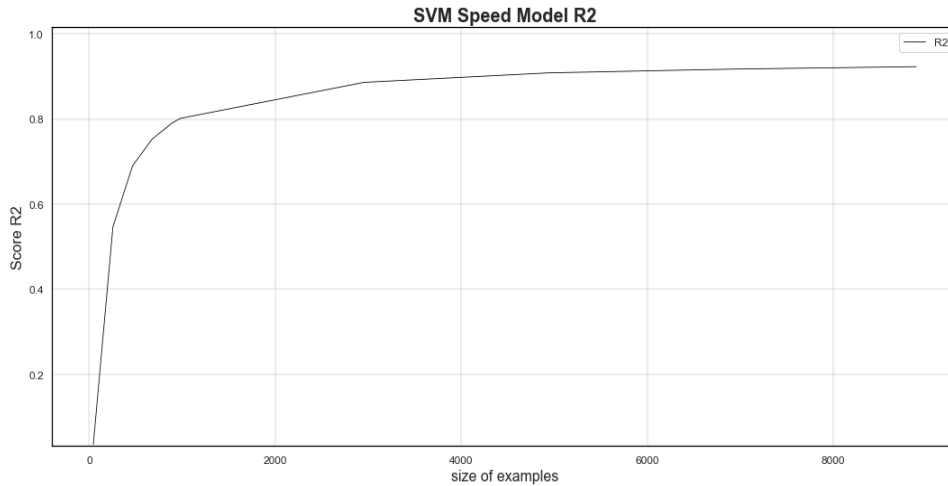
Figure 6-10 – Score analysis in the function of training set size

6.4 Two-Stage Model - Second Stage

6.4.1 First model

A second stage model is developed to generate fuel consumption predictions for the propulsive system. It is proposed that this model is trained to receive the data resulting from the first stage model and thus indicate what the consumption would be for a given route and expected speed. Figure 6-11 shows the proposed scheme of the second stage model. Like the first stage, the second stage model is developed using the same features and, additionally, the SOG, as shown in Table 6-11. Also, the same hyperparameters are used in both machine learning models, ANN as in Table 6-6 and SVM as in Table 6-9.
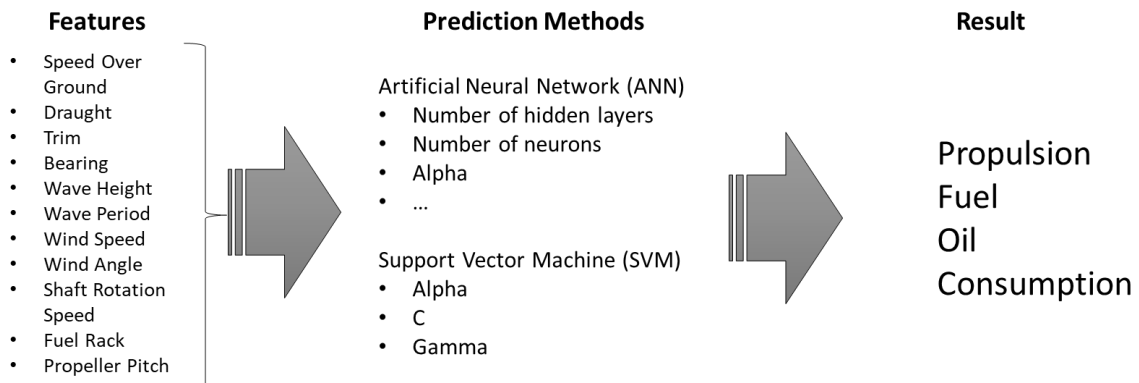
Figure 6-11 – Schematic – Second stage model prediction

Table 6-11 – Training dataset for the second stage model

Training dataset – the second stage	
SOG	Mean Draught
Wave Angle	Trim
Wave Period	Heading
Wave Height	Shaft Rotation Speed
Wind Angle	Fuel Rack
Wind Speed	Propeller Pitch

The results of the ANN model are shown in Table 6-12. As expected, the results have higher accuracy with the addition of the variables related to the ship propulsion compared to the first model described in Chapter 6.1. This model contains the input data with more correlation between them so that the simpler neural network system already showed satisfactory results, with $R^2 \approx 0.98$, very similar to those presented with more hidden layers.

Table 6-12 - ANN results to fuel consumption prediction

Size of ANN	Number of Nodes	Regularization Term - λ	Activation	Score - R^2
1 hidden layer	(100)	1.0e-4	tanh	0.9795
2 hidden layers	(100,100)	0.002	tanh	0.9878
3 hidden layers	(200, 100, 40)	0.01	tanh	0.9888

For the SVM model, the same hyperparameters shown in Table 6-9 are used, and the same inputs are used in the ANN model, with the same training set and test set. The best result found can be seen in Table 6-13. It can be seen that the model results fit the observed data very closely, as shown in Figure 6-12.

Table 6-13 - SVM results to fuel consumption prediction

Hyperparameter	Value
Gamma	1.10e-04
C – Regularization parameter	256
Epsilon	1.00e-04
Score (R²)	0. 9971
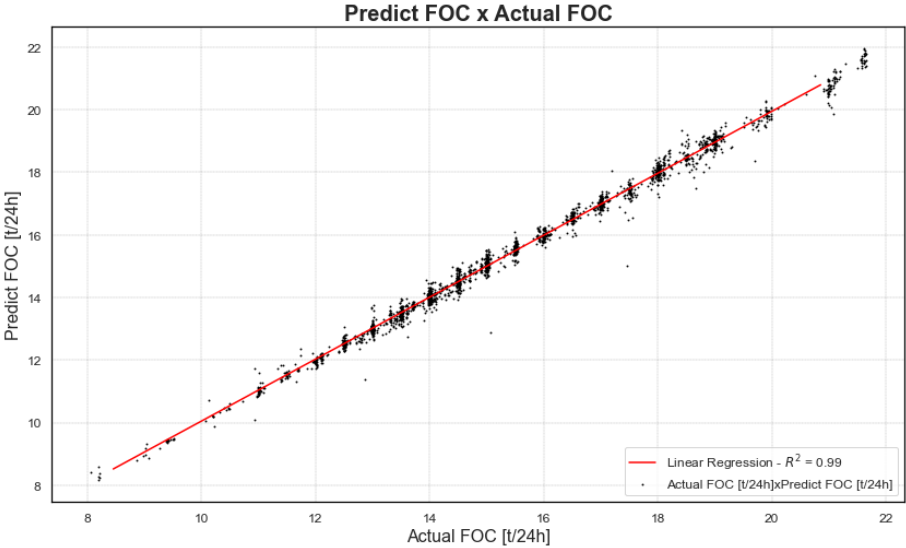
Figure 6-12 - Scatter plot result of the predicted and actual value – FOC

Furthermore, it can be seen from Figure 6-13 and Figure 6-14 that the convergence of this model requires fewer test data. With less than 2000 samples the model has already converged to an optimal

solution since the cost of the train and test data converged in a small number and the score (R^2) is close to 1.

Figure 6-13 – Cost result from the Test data and Training data for different samples sizes
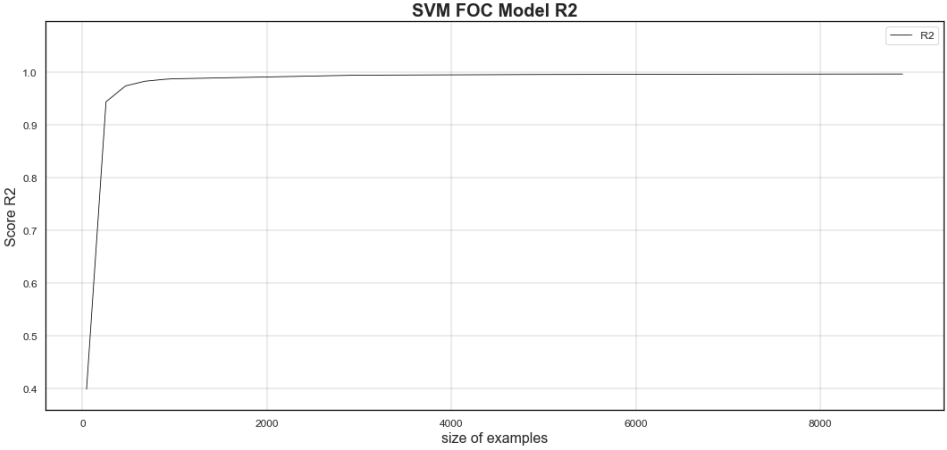
Figure 6-14 – Score result (R^2) for different sample sizes

6.4.2   Second Model

Since the model has shown to have excellent accuracy, it is assessed whether one of the variables could be removed so that, besides the computational gain, it could better balance the relationship between the other variables and avoid having a high bias. Thus, the fuel rack position that is the variable with the highest correlation with the fuel consumption is removed from the dataset and a new analysis is performed using only the SVM method with the same variations of hyperparameters shown in Table 6-9.

As can be seen in Figure 6-15 that even removing the variable with a correlation value of 0.95 with the main engine fuel consumption (Figure 5-6), the model provides good predictions, maintaining a high value of the coefficient of determination.

Table 6-14 - SVM results to fuel consumption prediction

Hyperparameter	Value
Gamma	1.10e-04
C – Regularization parameter	256
Epsilon	1.00e-04
Score (R²)	0. 9928
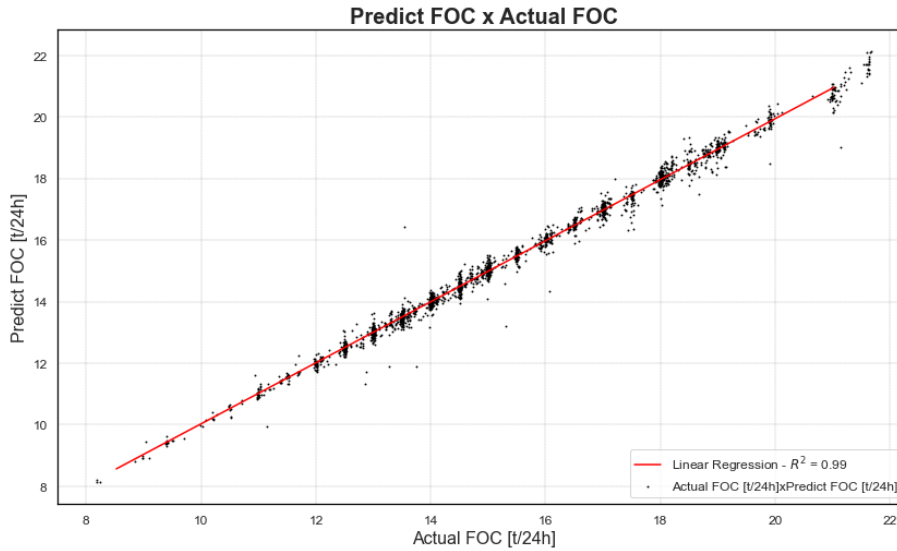
Figure 6-15 – Second Model FOC prediction without fuel rack as an input variable

Also, a similar analysis is conducted as before, to calculate the cost of the solution for each sample size. The results are shown in Figure 6-16 and Figure 6-17. One can see that the model converges with just a small part of the dataset size, which can be useful reduce the computational time in future analyses.
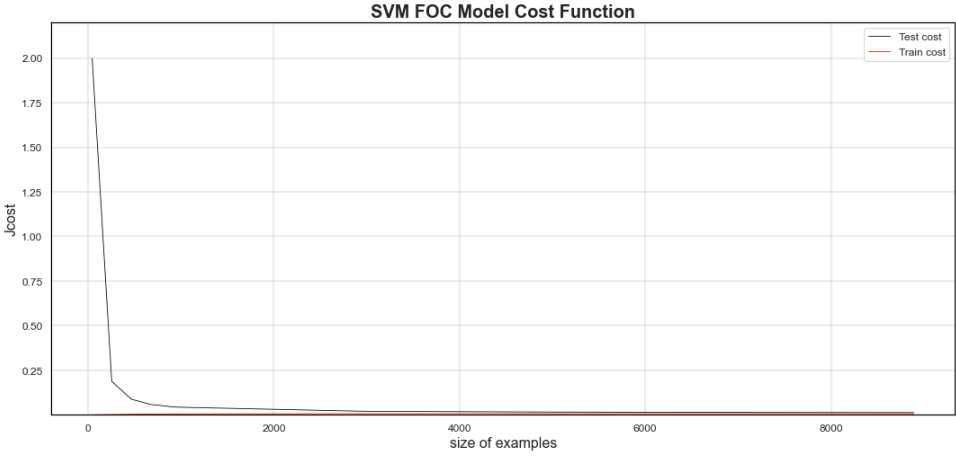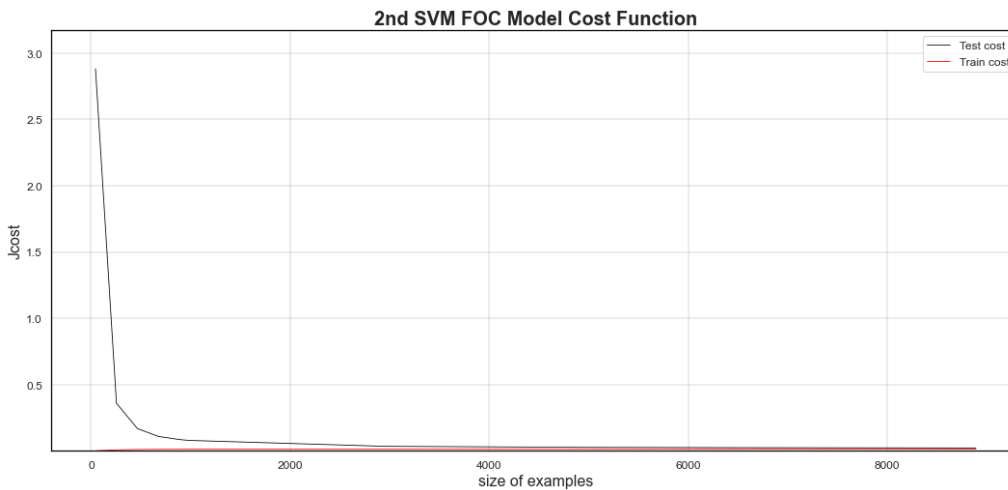

Figure 6-16 - Cost result from the Test and Training data for different sample sizes
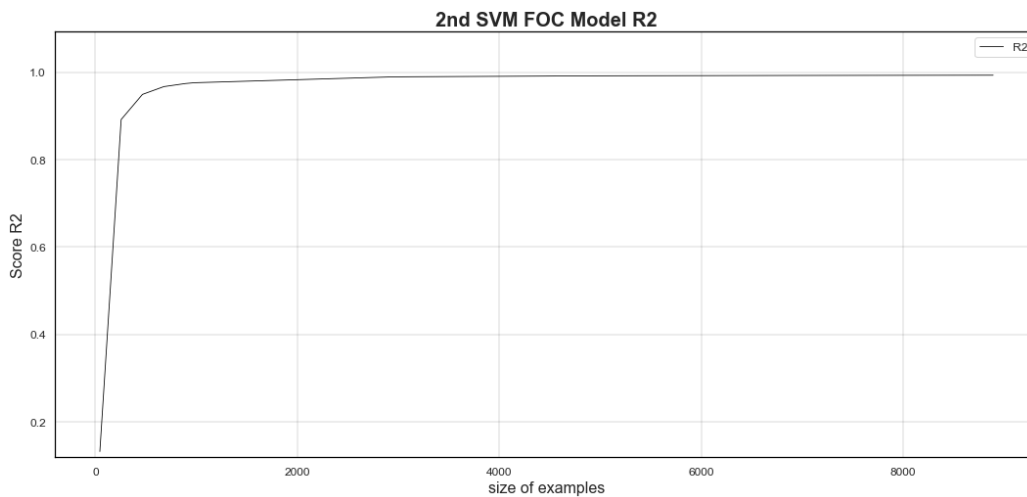
Figure 6-17 - – Score result (R^2) for the samples size

6.5 Two-stage model results and predictions

With the two-stage models already defined, a script is developed so that the first stage model is an input for the second stage model, as outlined in Figure 6-18. Thus, the two-stage model proposed can analyse each set of propulsive systems, predicting the speed and consequently analysing the consumption of the chosen system.
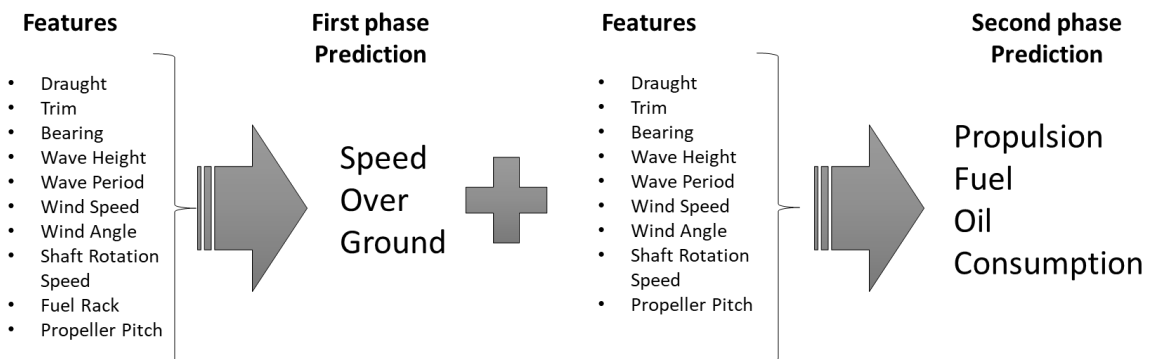
Figure 6-18 – Two-stage fuel consumption prediction model

The models chosen for the solution are for the first stage, the SVM with the hyperparameters of Table 6-10, and for the second stage, the SVM model with the hyperparameters of Table 6-14 and Table 6-13 since they both have a good prediction score. Also, it is used the model without the fuel rack position, since the FOC prediction model has a good performance without this variable, as demonstrated in Chapter 6.4.2.

A tool is implemented in Python to combine the two stages and to provide the speed and fuel consumption predictions. Figure 6-19 shows the results predicted by the model compared to the real observed consumption. It can be seen that the model has a good adherence, with a coefficient of determination of around 0.99.
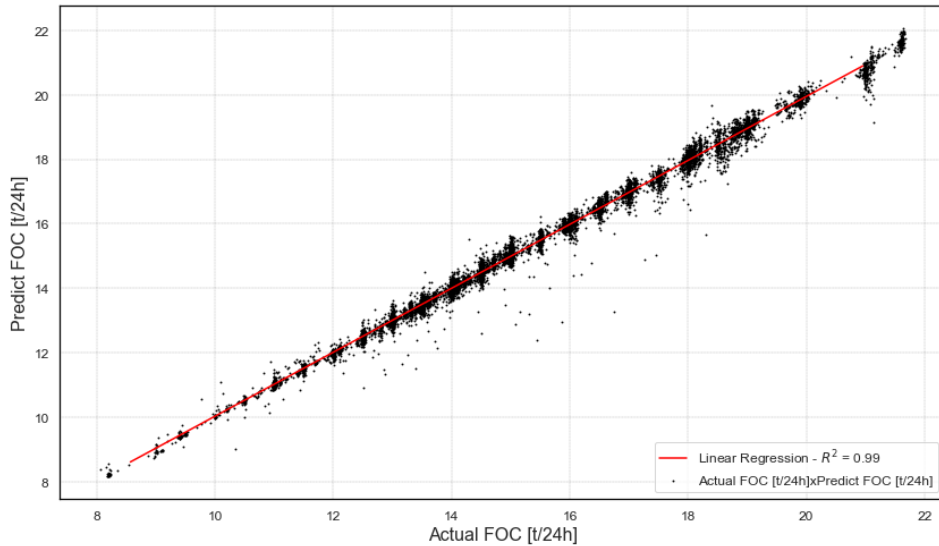
Figure 6-19 - Scatter plot of predicted and observed FOC values – Two-stage model

A brief analysis of four of the routes is shown in Figure 6-20 and Figure 6-21, where the first presents the speed-related results and the second presents the fuel-related results. As one can see in the speed prediction, although the predictions follow the general trends, at some points of the route there is a difference of 0.5 knots between the predicted and the actual data. This does not affect the result of the fuel consumption prediction because this is a combination of factors, and the learning model is basing its results on other parameters.
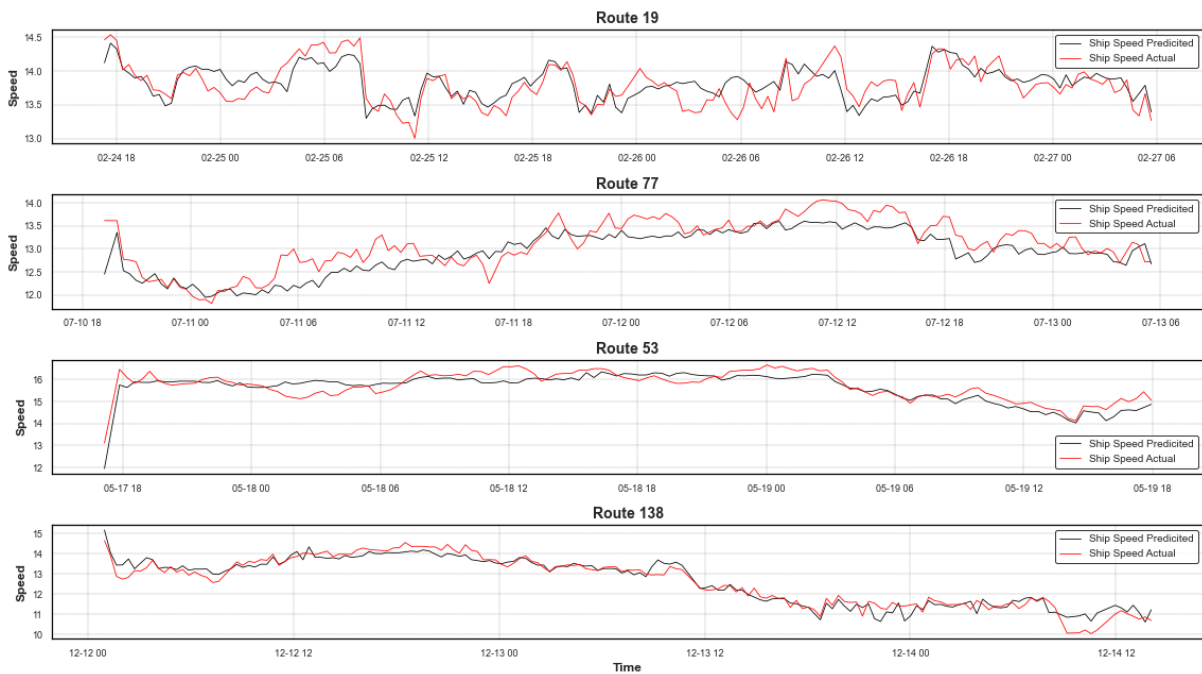
Figure 6-20 – Prediction Speed x Actual Speed
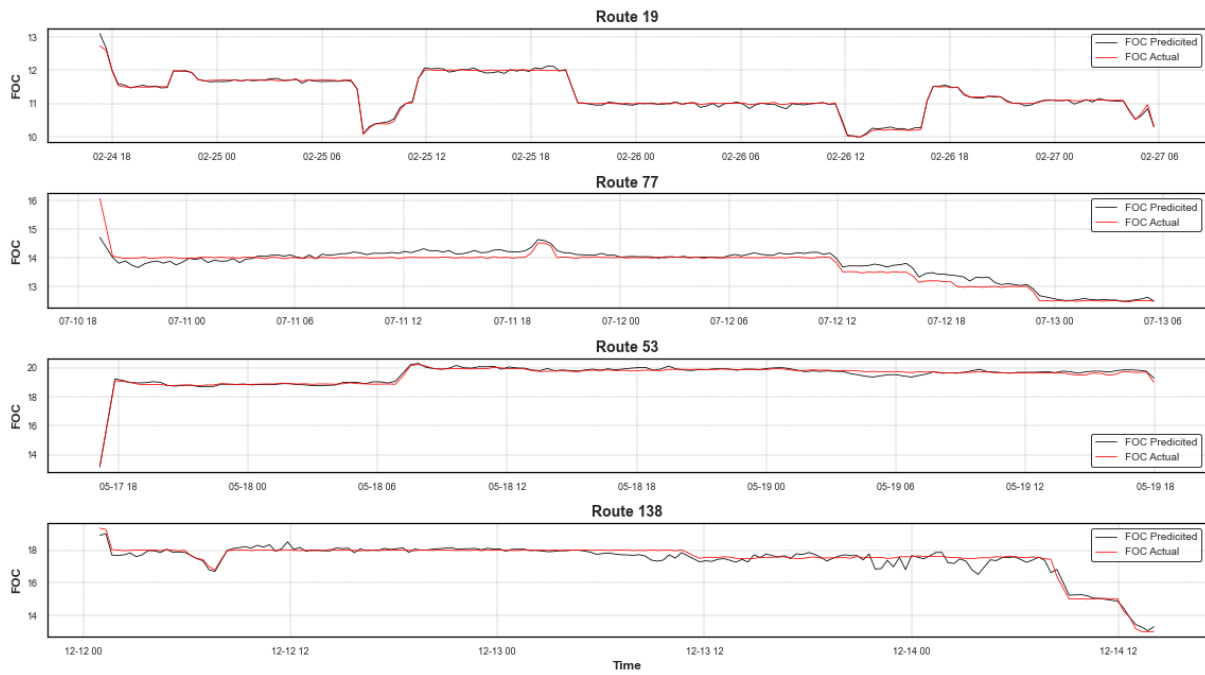
Figure 6-21 – Prediction FOC x Actual FOC

Table 6-15 shows the total fuel consumption by the main engine calculated for each of the above routes. The errors are acceptable and less than 1%. Also, an analysis is conducted for all 145 routes, comparing the predicted and observed total fuel consumptions. In this case, the maximum error is 2.27% and the mean error is about 0.44%.

Table 6-15 – Total fuel consumption comparison

Route	Predicted total FOC [tons]	Actual total FOC [tons]	Error
19	28.346	28.324	0.08%
77	33.490	33.308	0.55%
53	39.646	39.631	0.04%
138	44.214	44.574	-0.81%

7. DECISION SUPPORT SYSTEM

A decision support system (DSS) is developed to demonstrate how the developed ML models can be used in practice. Figure 7-1 shows the schematic design of the code developed in Python for this analysis.

The user must define the target speed and indicate whether the shaft generator is used or not. In this analysis, data from the studied routes such as latitude, longitude, wind and wave conditions are used as input to the model. However, the system can cope with specific conditions of the operation included by the user. The same is done for the ship conditions, i.e., draught and trim data.

Furthermore, the system takes all the propulsive system configuration variables (i.e., fuel rack, shaft rotation speed and propeller pitch), to ensure that the condition of the propulsive configuration is feasible. As already mentioned in the previous chapters, the learning methods make predictions based on the range of the training dataset. Therefore, if the input values used are outside the range used for training, the results may deviate from reality. So, the values adopted by the analysis are distributed as shown in Chapter 4, Chapter 5.3, Appendix II and Appendix III.

The first step of the model consists of building the dataset to be analysed. The parameters to be introduced are the route, the speed, the use or not of the shaft generator and the draught and trim conditions of the ship. With these data, the system actively or passively adds the environmental data and calculates the ship's heading based on the route. It thus assembles the data related to the route to be investigated.

In the second step, the model verifies for each position of the route with each configuration of the propulsion system, if it reaches a speed greater or equal to the requested speed. Then, in the third step, the predicted fuel consumption is calculated for each configuration that attains the desired speed, thus knowing which is the lowest fuel consumption for that speed.
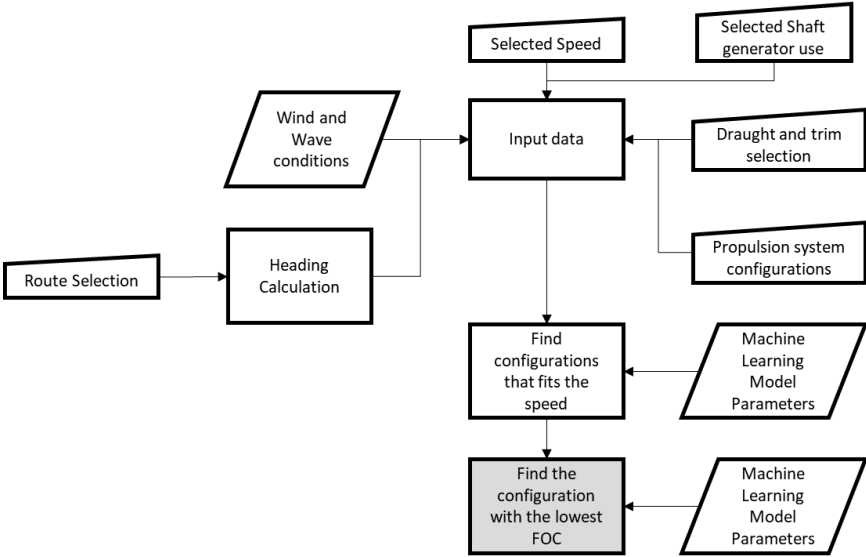
Figure 7-1 – Decision Support System

This calculation is made for each point along the route, verifying which operational conditions could be used and their fuel consumption. An analysis is performed based on one of the recorded routes (Figure 7-2), with the original weather and operating conditions. This route has about 970 nautical miles.
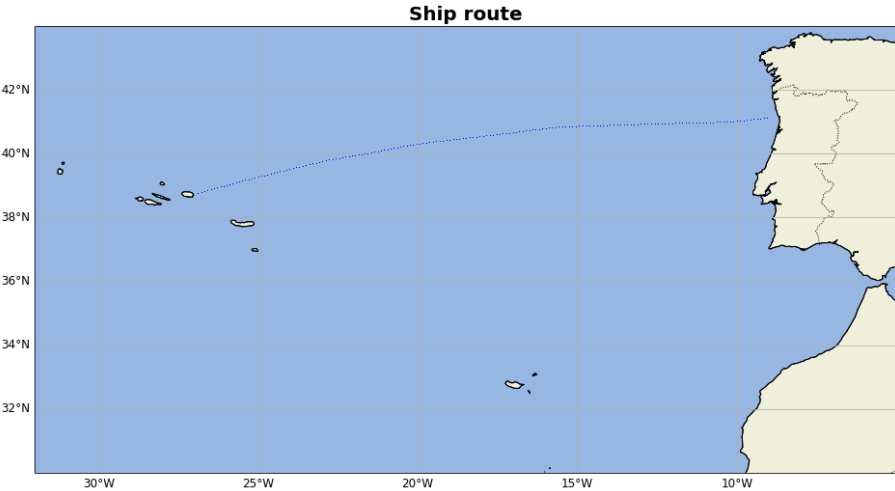
Figure 7-2 - Route 19

The DSS is programmed to switch on the shaft generator in one-third of the route travelled, and then to switch off in the final section. The model is set to calculate the FOC for a minimum speed of 13.5 [knots]. The results are shown in Figure 7-3. One can see that when the shaft generator is switched on, the propeller pitch reduces to compensate for the increased rotational speed of the shaft. In addition, the consumption increases slightly to ensure the requested minimum speed. A study on the impact of the use of the shaft generator on the speed and FOC should be conducted to ensure that this modelling approach is in line with reality. The graph is discretised by space, i.e., the results presented represent the variation of the systems along the route.
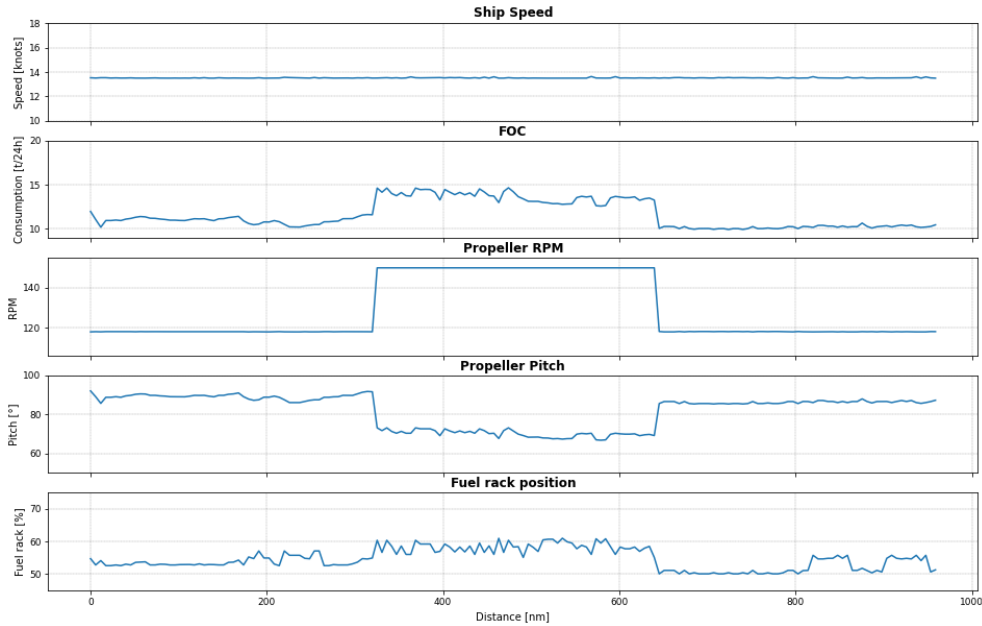
Figure 7-3 – DSS Simulation results along the route

Figure 7-4 shows the predictions for different speeds, without the use of the shaft generator. As expected in this case for higher speeds there is higher fuel consumption, and one can see that the fuel rack follow the same behaviour of the FOC. The shaft rotation at the speeds of 12 and 14 knots is low and increases for the speed of 16 knots to compensate for the limitations of the propeller pitch angle, which was already close to the limit at the speed of 14 knots. Thus, for the 16 knots case, the model increased the shaft rotation to obtain more thrust in the propulsion system. Further analysis is needed to understand why the model prefers to use 119 rpm rather than 130 rpm, and whether it represents the ship's current propulsion system.
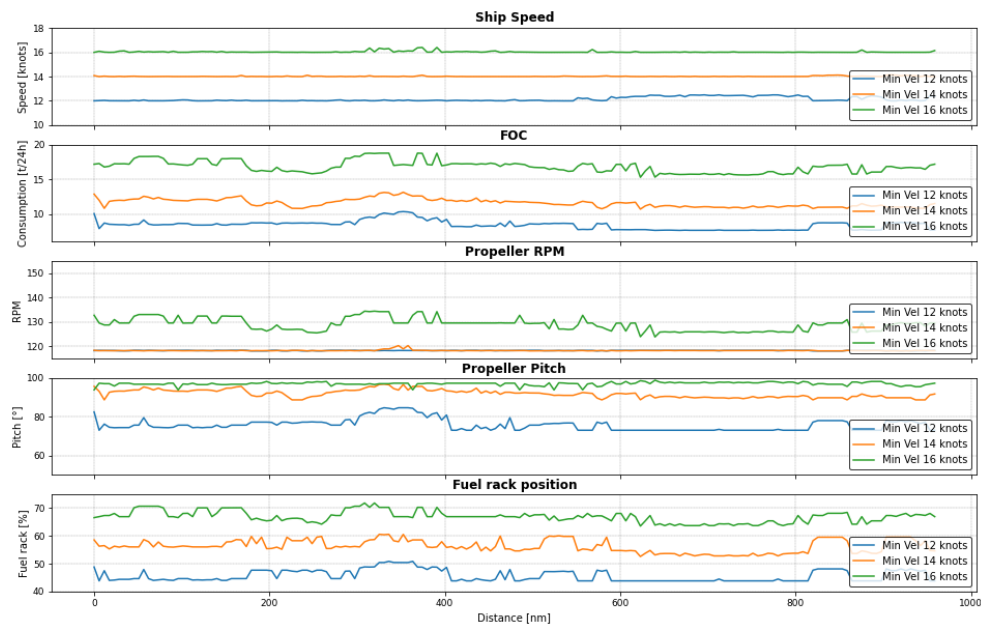
Figure 7-4 – DSS result for different speeds without the use of shaft generator

A sensitivity analysis is also performed to asses how the model behaves with some variables. An analysis is conducted simulating a calm water condition, with winds of 4 [m/s] and a significant wave height of 0.7 [m]. The environmental values are not set to zero because as mentioned before, the model learns up to the limits of the studied ranges, so the lower limits of the wave and wind conditions are used, as shown in Figure 4-19 and Figure 4-17.

With these environmental conditions, the cases with a ship's speed from 12 to 16 knots are simulated. Figure 7-4 shows the results for the total fuel consumption for the entire route for each ship's speed. One can see that the model understood the relationship between speed and consumption as a polynomial function relationship, as expected.

Similar to the previous simulation, a new analysis is conducted, but in this case, varying the significant wave height. The wind condition remained the same, with the wind at 4 [m/s], the ship's speed is chosen as 13.5 knots, and the analysed wave conditions vary from 1.0 [m] to 3.5 [m] of significant wave height. Figure 7-6 shows an almost linear relationship between the significant wave height and predicted.
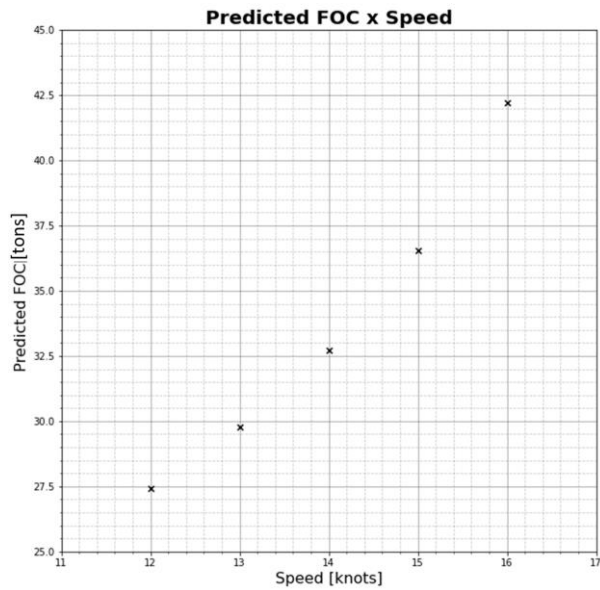
Figure 7-5 – Predicted FOC based on different Ship speed
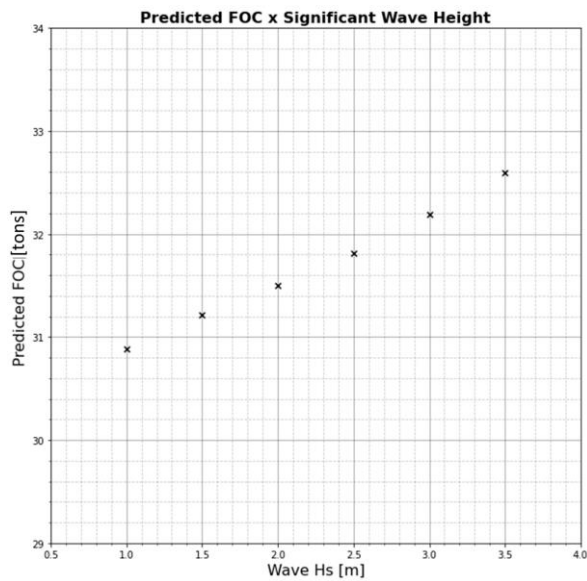
Figure 7-6 - Predicted FOC for different significant wave height

Also, an analysis changing the wind speed was performed, where the ship's speed was kept at 13.5 knots, the significant wave height at 0.7 [m] and varying the wind speed between 2.5 [m/s] and 15 [m/s]. In Figure 7-7 one can see the result for the total consumption on the route. Where at lower speeds, the predictions are nearly the same with small variations within the error range of the model. The consumption only increases from a wind speed of 10 [m/s]. This behaviour needs to be investigated if the DSS model shows a low correlation with reality. If the model has low accuracy, one of the reasons may be that it is underestimating the resistance force due to the wind.

A final analysis is conducted to asses whether the model can capture the impact of changing the ship draught on the fuel consumption. In Figure 7-8, one can see the variation of draught from 5 [m] to 7.5

[m] and its impact on total fuel consumption. The analysis is performed using the calm water condition already explained and also with a speed of 13.5 knots.

These analyses serve the purpose of understanding how the machine learning models perceived the influence of each variable related to this particular ship under specific operational conditions. It can be seen that the model identified a large influence of the ship's speed and draught, as already expected. However, the significant wave height and wind speed have a low impact on the FOC predictions. This may be due to the low correlation observed between the environmental variables and ship speed and ship consumption (Figure 5-2), as investigated in Chapter 5.
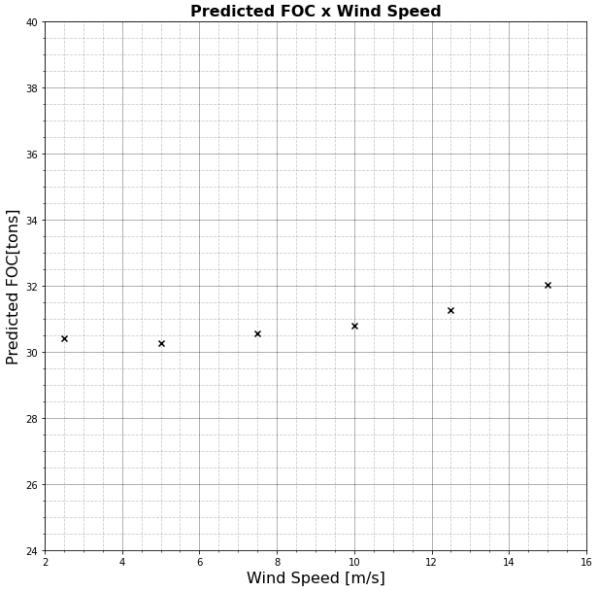
Figure 7-7 - Predicted FOC based on different Wind Speed
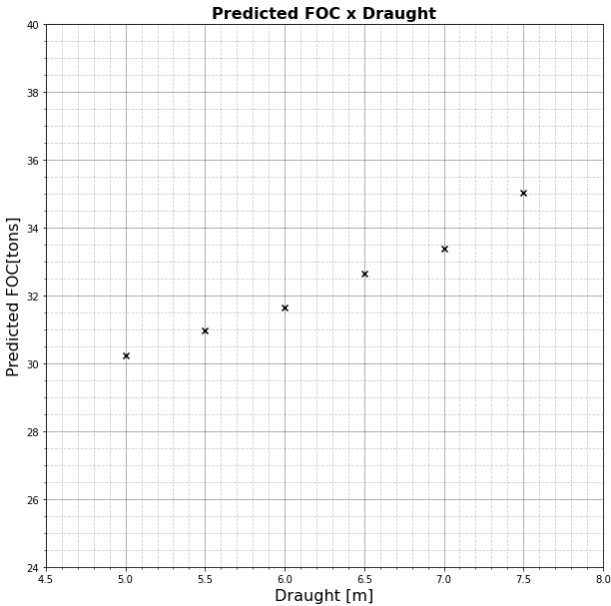
Figure 7-8- Predicted FOC for different ship's draughts

8. CONCLUSION AND FUTURE WORK

8.1 Conclusion

This dissertation has developed machine learning models that represent the operation of a fuel optimization system and has developed a prototype of a decision support system that provides predictions of the optimal fuel consumption of the ship's main engine.

The study was developed based on a one-year sample of data collected from a ship's automated fuel optimization system, which includes the propulsion system parameters and fuel consumption of the ship in operation as well as data on the environmental conditions and other variables that were added to enrich the dataset.

The analyses performed on the dataset demonstrated that the automated system for fuel consumption optimization is reliable since the actual fuel consumption along the ship voyages is always very close to the set value. Moreover, by optimizing the propulsive system all the time to guarantee the set consumption, this system makes the environmental variables appear uncorrelated with the speed and with low impact with the fuel oil consumption, as shown in Chapter 5.

This has affected the preliminary Machine Learning models proposed to predict the fuel consumption, because the variables used by the model present low correlation, making the final model not perform well. As shown in Chapter 6.1, both preliminary Artificial Neural Network and Support Vector Machines models using the ship and weather conditions as a training dataset did not perform as expected. The models showed a coefficient of determination of about 0.88, which is not acceptable for a fuel consumption forecast, since the shipowner needs to have a good prediction of how much fuel the ship will spend on the voyage and if the ship will reach the required carbon index limit.

Therefore, a 2-stage Machine Learning prediction model was proposed using new features such as the shaft rotation speed, fuel rack and propeller pitch, as explained in Chapter 6.2. In the first stage, Artificial Neural Network and Support Vector Machines models were developed to predict the ship's speed based on the ship configuration, weather conditions and propulsion system configuration, with the Support Vector Machines the method showing best prediction accuracy on the dataset, with a score of 0.92.

The second stage model was developed to predict the fuel consumption of the main engine also using Artificial Neural Network and Support Vector Machines models. In this case, both methods had a good performance, with scores of 0.99 and 0.98, respectively.

The proposed two-stage model proved to have good accuracy, as shown in Chapter 6.5, reaching 0.99 in the score value when using the Support Vector Machines model in both stages, as it presented a better result in each stage.

The Decision Support System was proposed and formulated based on the 2-stage prediction model, to evaluate what would be the fuel consumption for a chosen minimum speed, using or not the shaft generator. The first stage assesses if the target speed is reached for a given operating condition, and in the second stage, the fuel consumption for each condition of the first stage is calculated. Chapter 7

shows some examples of how to use the system and a sensibility analysis that provides indications of the influence of selected variables on the total ship's fuel consumption. It was shown the large influence of the ship's speed and draught, as already expected and the low impact of the environmental conditions. This reinforces the findings of Spearman's correlation analysis, presented in Chapter 5, where the correlation between the environmental conditions and the ship's speed and FOC are minimal. Despite

The developed models need to be tested and adjusted for ship operation. New variables and larger datasets properly pre-processed may be required. It is suggested to collect more data or, if possible, to try to use the model to find out how adherent this model is with reality. New tests can be made including new variables such as engine speed and fuel temperature, as long as it is known how much the optimisation system influences these variables.

8.2 Future work

The 2-stage model has been shown to provide good predictions, as presented in Chapter 6.5. Even removing one of the variables, the system shows a good score of 0.99 in the FOC prediction. However, further studies should be carried out and compared to verify the validity of the use of these models in a DSS, since it is not clear that the machine learning method captures all the influences necessary to predict the FOC using this decision support system. Also, this model was developed with a focus on fuel consumption and not to find out the impact of each component that interferes in consumption. For that, a specific model must be created, with specific dataset treatments for each case under study, similar to what was done by Dinham-Peren et al. [54], in the study of ship resistance in calm water conditions using operational data.

The input in the DSS model can be complemented with theoretical models, creating a so-called grey model, combining the parametric models with the ML methods. There are a few studies in that area, as by Leifsson et al. [107] and Haranen et al. [108]. These models can prevent the system from choosing a set of solutions that would not be feasible for the ship's propulsion system to achieve a speed with the desired fuel consumption. Thus, avoiding a possible bias or overfitting of the prediction model.

The SVM model seemed to provide better results in this dataset, but in future work, after comparing with new real observations, it is necessary to verify if there is some kind of unbalance between the variables that may require a reanalysis and new treatments of the dataset, which could be subjected to standardization to achieve better results. Also, in future work, deeper ANNs could be tested, including non-linear relationships between the variables.

Analyses of new scenarios can be made, but these must be in accordance with the range of the model, because any prediction model has its limits in the range of the input data, and extrapolating values outside of this range may not correspond to reality.

In this study, it was not possible to access the types of sensors and their uncertainties. It is known that there are biases and uncertainties in the measurement instruments. The ISO 19030 - Ship Performance and Condition Monitoring [109] provides guidance and introduce good practices on that. Future work would be to analyse these uncertainties and implement them in the model to know the total uncertainties

of the model. Some studies have analysed these uncertainties as by Hagestuen et al. [110], Thornhill et al. [92], Aldous et al. [111] and Aldous [47], but they have not coupled them to a machine learning prediction model.

Better models can be built with other Python libraries, like Keras [112]. This library contains more models and machine learning functionalities and possibilities to work with parallel computing, which would speed up the analysis process.

Several applications can be developed using this model as a starting point, in addition to the decision model for shipowners focusing on the speed analysis and fuel consumption by the main engine. It can be used to plan a route based on environmental and ship conditions, as for this it would be only necessary to have the environmental forecasts and the ship's operational condition. For this, a model based on the Dijkstra algorithm should be used, changing it to allow the software to consider not only the distance but also the consumption for each section analysed [102].

A just in time (JIT) tool can be constructed, as in Farag and Ölçer [68], to try to optimize the sailing time and the waiting time for port entry, thus being able to optimize the speed during the route and consuming less fuel, also making the ship not wait in a queue, reducing the consumption of more expensive fuels such as MDO. This tool could work integrated within the ship system.

The prediction model can also be used to check if the ship complies with regulatory standards, verifying if the carbon index is within the limits imposed for that class of ships. For example, using the EEOI, one can calculate the carbon index based on the total FOC of the routes, as previously shown in Table 6-15.

A digital twin can be modelled, as in Coraddu et al. [113], to study how the ship would behave in new scenarios, equipment changes or marine fouling growing. This type of development involves larger datasets enriched with additional variables and a complete understanding of how they relate to each other.

9. REFERENCES

[1] International Maritime Organization - IMO, "Fourth IMO Greenhouse Gas Study," 2021.

[2] IEA *et al.*, "World Energy Outlook 2020," vol. 2050, no. October, pp. 1–461, 2020, [Online]. Available: https://www.oecd-ilibrary.org/energy/world-energy-outlook-2020_557a761b-en

[3] International Maritime Organization - IMO, "Study of Greenhouse Gas Emissions from Ships Final Report to the International Maritime Organization," 2000.

[4] International Maritime Organization - IMO, "Second IMO Greenhouse Gas Study," 2009.

[5] International Maritime Organization - IMO, "MEPC.1/Circ.684 - GUIDELINES FOR VOLUNTARY USE OF THE SHIP ENERGY EFFICIENCY OPERATIONAL INDICATOR (EEOI)." 2009.

[6] International Maritime Organization - IMO, "Resolution MEPC.203(62) - Amendments to the Annex of the Protocol of 1997 - 2011 Revised MARPOL Annex VI," vol. 1, no. 22. p. 19, 2011.

[7] International Maritime Organization - IMO, "MEPC.278(70) - GUIDELINES FOR THE DEVELOPMENT OF A SHIP ENERGY EFFICIENCY MANAGEMENT PLAN," vol. 278, no. October 2016. pp. 1–7, 2016.

[8] International Maritime Organization - IMO, "MEPC.304(72) - Resolution- Initial IMO Strategy on reduction of GHG emission from ships.," vol. 304, no. April. 2018.

[9] International Maritime Organization - IMO, "MEPC.328(76) - Resolution - Amendments to the Annex of the Protocol of 1997 - 2021 Revised MARPOL Annex VI," vol. 148. pp. 148–162, 2021.

[10] International Maritime Organization - IMO, "Third IMO Greenhouse Gas Study 2014," 2014. doi: 10.1007/s10584-013-0912-3.

[11] BIMCO, "Seafarer Workforce Report - The global supply and demand for seafarers in 2021," 2021.

[12] Council of the European Union, "Regulation (Eu) 2015/757 of the European Parliament," *Official Journal of the European Union*, vol. L, no. April. pp. 55–76, 2015. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32015R0757&from=EL

[13] International Maritime Organisation - IMO, "Resolution MEPC.278(70) - Amendments to MARPOL Annex VI - Data collection system for fuel oil consumption of ships - 2016," vol. 278, no. October 2016. pp. 1–7, 2016.

[14] European Parliament, "The European Green Deal. European Parliament resolution of 15 January 2020 on the European Green Deal," *https://www.europarl.europa.eu/doceo/document/TA-9-2020-0005_EN.html*. 2020.

[15] European Parliament, "More efficient and cleaner maritime transport," *The Routledge Handbook of Gender and EU Politics*. pp. 107–119, 2021. doi: 10.4324/9781351049955-11.

[16] G. Mundaca, J. Strand, and I. R. Young, "Carbon pricing of international transport fuels: Impacts on carbon emissions and trade activity," *Journal of Environmental Economics and Management*, p. 102517, 2021, doi: 10.1016/j.jeem.2021.102517.

[17] E. A. Bouman, E. Lindstad, A. I. Rialland, and A. H. Strømman, "State-of-the-art technologies, measures, and potential for reducing GHG emissions from shipping – A review," *Transportation Research Part D: Transport and Environment*, vol. 52, pp. 408–421, May 2017, doi: 10.1016/j.trd.2017.03.022.

[18] DNV, "DNV's Alternative Fuels Insight platform," 2021. https://afi.dnvgl.com/Statistics

[19] A. Bureau, "Guide for Ammonia Fueled Vessels," no. September, 2021.

[20] A. Bureau, "Guide for Wind Assisted Propulsion System Installation 2021," no. August, 2021.

[21] DNVGL, "Maritime Forecast To 2050," *Energy Transition Outlook 2019*, p. 118, 2021.

[22] H. N. Psaraftis, "Green Maritime Transportation: Market Based Measures," 2016. doi: 10.1007/978-3-319-17175-3_8.

[23] T. Chou, V. Kosmas, M. Acciaro, and K. Renken, "A comeback of wind power in shipping: An economic and operational review on thewind-assisted ship propulsion technology," *Sustainability (Switzerland)*, vol. 13, no. 4, pp. 1–16, 2021, doi: 10.3390/su13041880.

[24] N. J. van der Kolk, J. A. Keuning, and R. H. M. Huijsmans, "Part 1: Experimental validation of a RANS-CFD methodology for the hydrodynamics of wind-assisted ships operating at leeway angles," *Ocean Engineering*, vol. 178, no. December 2018, pp. 375–387, 2019, doi: 10.1016/j.oceaneng.2018.12.041.

[25] "The Oceanbird." https://www.theoceanbird.com/ (accessed Oct. 09, 2021).

[26] M. Stopford, *Maritime Economics*. 2013. doi: 10.4324/9780203442661.

[27] T. P. V. Zis, H. N. Psaraftis, and L. Ding, "Ship weather routing: A taxonomy and survey," *Ocean Engineering*, vol. 213, no. March, p. 107697, 2020, doi: 10.1016/j.oceaneng.2020.107697.

[28] P. Jain and M. C. Deo, "Neural networks in ocean engineering," *Ships and Offshore Structures*, vol. 1, no. 1, pp. 25–35, 2006, doi: 10.1533/saos.2004.0005.

[29] A. Rawson, M. Brito, Z. Sabeur, and L. Tran-Thanh, "A machine learning approach for monitoring ship safety in extreme weather events," *Safety Science*, vol. 141, no. March, p. 105336, 2021, doi: 10.1016/j.ssci.2021.105336.

[30] A. Viellechner and S. Spinler, "Novel Data Analytics Meets Conventional Container Shipping: Predicting Delays by Comparing Various Machine Learning Algorithms," *Proceedings of the 53rd Hawaii International Conference on System Sciences*, vol. 3, pp. 1278–1287, 2020, doi: 10.24251/hicss.2020.158.

[31] G. Tsaganos, N. Nikitakos, D. Dalaklis, A. I. Ölcer, and D. Papachristos, "Machine learning algorithms in shipping: improving engine fault detection and diagnosis via ensemble methods," *WMU Journal of Maritime Affairs*, vol. 19, no. 1, pp. 51–72, 2020, doi: 10.1007/s13437-019-00192-w.

[32] G. W. Y. Wang, Z. Yang, D. Zhang, A. Huang, and Z. Yang, "Application of Bayesian networks in analysing tanker shipping bankruptcy risks," *Maritime Business Review*, vol. 2, no. 3, pp. 177–198, 2017, doi: 10.1108/MABR-12-2016-0032.

[33] A. Fabregat, L. Vázquez, and A. Vernet, "Using Machine Learning to estimate the impact of ports and cruise ship traffic on urban air quality: The case of Barcelona," *Environmental Modelling and Software*, vol. 139, no. February, 2021, doi: 10.1016/j.envsoft.2021.104995.

[34] A. Sanfilippo and S. Chikkagoudar, "Automated detection of anomalous shipping manifests to identify illicit trade," in *2013 IEEE International Conference on Technologies for Homeland Security, HST 2013*, 2013, pp. 529–534. doi: 10.1109/THS.2013.6699059.

[35] Q. Zhou and V. v Thai, "Application of Data-Mining Techniques for Personal Injury Evaluation in Tanker Shipping Industry," *International Journal of Computing, Communication and Instrumentation Engineering*, vol. 2, no. 2, 2015, doi: 10.15242/ijccie.er1215103.

[36] L. Bramer, S. Chatterjee, A. Holmes, S. Robinson, and S. Bradley, "A Machine Learning Approach for Business Intelligence Analysis using Commercial Shipping Transaction Data," *Int'l Conf. Data Mining | DMIN'15*, pp. 162–167, 2015, [Online]. Available: https://search.proquest.com/openview/c79b355e6e441a514bc22127f6ea7a66/1?pq-origsite=gscholar&cbl=1976357

[37] M. van den Boogaard, G. Alessi, B. Mallol, D. Wunsch, and N. Clero, "Accelerating marine propeller development in early design stages using machine learning".

[38] L. Moreira, R. Vettor, and C. G. Soares, "Neural network approach for predicting ship speed and fuel consumption," *Journal of Marine Science and Engineering*, vol. 9, no. 2, pp. 1–14, 2021, doi: 10.3390/jmse9020119.

[39] P. Krata, R. Vettor, and C. Guedes Soares, "Bayesian approach to ship speed prediction based on operational data," in *Developments in the Collision and Grounding of Ships and Offshore Structures - Proceedings of the 8th International Conference on Collision and Grounding of Ships and Offshore Structures, ICCGS 2019*, 2020, pp. 384–390. doi: 10.1201/9781003002420-47.

[40] K. Rudzki and W. Tarelko, "A decision-making system supporting selection of commanded outputs for a ship's propulsion system with a controllable pitch propeller," *Ocean Engineering*, vol. 126, no. June 2015, pp. 254–264, 2016, doi: 10.1016/j.oceaneng.2016.09.018.

[41] ITTC, "ITTC Performance Prediction Method." pp. 1–10, 2011.

[42]   K. U. Hollenbach, "Estimating resistance and propulsion for single-screw and twin-screw ships," *Ship Technology Research*, vol. 45, no. 2, pp. 72–76, 1998.

[43]   J. Holtrop, "STATISTICAL RE-ANALYSIS OF RESISTANCE AND PROPULSION DATA.," *International Shipbuilding Progress*, vol. 31, no. 363, pp. 272–276, 1984.

[44]   J. Holtrop and G. G. J. Mennen, "An Approxima Power Prediction Method," *International Shipbuilding Progress*, vol. 29, no. 335, 1982, [Online]. Available: http://mararchief.tudelft.nl/file/33815/

[45]   J. Prpić-Oršić and O. M. Faltinsen, "Estimation of ship speed loss and associated CO_2 emissions in a seaway," *Ocean Engineering*, vol. 44, pp. 1–10, 2012, doi: 10.1016/j.oceaneng.2012.01.028.

[46]   V. Bertram, *Practical Ship Hydrodynamics*. 2012. doi: 10.1016/C2010-0-68326-X.

[47]   L. G. Aldous, "Ship operational Efficiency: Performance Models and Uncertainty Analysis," 2015.

[48]   L. Aldous, T. Smith, R. Bucknall, and P. Thompson, "Uncertainty analysis in ship performance monitoring," *Ocean Engineering*, vol. 110, pp. 29–38, Dec. 2015, doi: 10.1016/j.oceaneng.2015.05.043.

[49]   O. M. Faltinsen, K. J. Minsaas, N. Liapis, and S. O. Skjordal, "PREDICTION OF RESISTANCE AND PROPULSION OF A SHIP IN A SEAWAY.," 1981, no. C, pp. 505–529.

[50]   N. Salvesen, E. O. Tuck, and O. M. Faltinsen, "Ship Motions and Sea Loads - Salvesen, Tuck, Faltinsen 1970.PDF."

[51]   J. H. Michell, "The Wave-Resistance of a Ship," *Philosophical Magazine*, vol. 45, pp. 106–123, 1898.

[52]   E. O. Tuck, "The wave resistance formula of J.H. Michell (1898) and Its significance to recent research in ship hydrodynamics," *The Journal of the Australian Mathematical Society. Series B. Applied Mathematics*, vol. 30, no. 4, pp. 365–377, 1989, doi: 10.1017/s0334270000006329.

[53]   H. Söding, A. von Graefe, O. el Moctar, and V. Shigunov, "Rankine source method for seakeeping predictions," *Proceedings of the International Conference on Offshore Mechanics and Arctic Engineering - OMAE*, vol. 4, no. August 2014, pp. 449–460, 2012, doi: 10.1115/OMAE2012-83450.

[54]   T. Dinham-Peren and I. Dand, "The Need for Full Scale Measurements," *RINA Conference, William Froude …*, no. November, pp. 1–13, 2010, [Online]. Available: http://media.bmt.org/bmt_media/resources/33/RINA Froude Conf - The Need for Full Scale Measurements.pdf.pdf

[55]   P. A. Lakshmynarayanana and D. A. Hudson, "Estimating Added Power in Waves for Ships Through Analysis of Operational Data," no. September 2018, 2017.

[56]   L. Larsson and E. Baba, *Ship resistance and flow computations*, vol. 5. 1996.

[57] B. P. Pedersen and J. Larsen, "Modeling of Ship Propulsion Performance," *World Maritime Technology Conference*, no. 2008, pp. 1–10, 2009.

[58] D. M. Park, J. Lee, and Y. Kim, "Uncertainty analysis for added resistance experiment of KVLCC2 ship," *Ocean Engineering*, vol. 95, pp. 143–156, 2015, doi: 10.1016/j.oceaneng.2014.12.007.

[59] F. S. Pereira, L. Eça, and G. Vaz, "Verification and Validation exercises for the flow around the KVLCC2 tanker at model and full-scale Reynolds numbers," *Ocean Engineering*, vol. 129, no. November 2016, pp. 133–148, 2017, doi: 10.1016/j.oceaneng.2016.11.005.

[60] A. Farkas, N. Degiuli, and I. Martić, "Assessment of hydrodynamic characteristics of a full-scale ship at different draughts," *Ocean Engineering*, vol. 156, no. January, pp. 135–152, 2018, doi: 10.1016/j.oceaneng.2018.03.002.

[61] H. J. Shaw and C. K. Lin, "Marine big data analysis of ships for the energy efficiency changes of the hull and maintenance evaluation based on the ISO 19030 standard," *Ocean Engineering*, vol. 232, no. 3, p. 108953, 2021, doi: 10.1016/j.oceaneng.2021.108953.

[62] A. I. Parkes, A. J. Sobey, and D. A. Hudson, "Physics-based shaft power prediction for large merchant ships using neural networks," *Ocean Engineering*, vol. 166, pp. 92–104, Oct. 2018, doi: 10.1016/j.oceaneng.2018.07.060.

[63] J. S. Carlton, *Marine propellers and propulsion*. 2018. doi: 10.1016/C2014-0-01177-X.

[64] K. Wang, X. Yan, Y. Yuan, and F. Li, "Real-time optimization of ship energy efficiency based on the prediction technology of working condition," *Transportation Research Part D: Transport and Environment*, vol. 46, pp. 81–93, 2016, doi: 10.1016/j.trd.2016.03.014.

[65] C. Gkerekos and I. Lazakis, "A novel, data-driven heuristic framework for vessel weather routing," *Ocean Engineering*, vol. 197, no. November 2019, p. 106887, 2020, doi: 10.1016/j.oceaneng.2019.106887.

[66] J. P. Petersen, D. J. Jacobsen, and O. Winther, "Statistical modelling for ship propulsion efficiency," *Journal of Marine Science and Technology*, vol. 17, no. 1, pp. 30–39, 2012, doi: 10.1007/s00773-011-0151-0.

[67] E. Bal Beşikçi, O. Arslan, O. Turan, and A. I. Ölçer, "An artificial neural network based decision support system for energy efficient ship operations," *Computers and Operations Research*, vol. 66, no. January 2013, pp. 393–401, 2016, doi: 10.1016/j.cor.2015.04.004.

[68] Y. B. A. Farag and A. I. Ölçer, "The development of a ship performance model in varying operating conditions based on ANN and regression techniques," *Ocean Engineering*, vol. 198, no. January, 2020, doi: 10.1016/j.oceaneng.2020.106972.

[69] I. Lazakis, Y. Raptodimos, and T. Varelas, "Predicting ship machinery system condition through analytical reliability tools and artificial neural networks," *Ocean Engineering*, vol. 152, no. November 2017, pp. 404–415, 2018, doi: 10.1016/j.oceaneng.2017.11.017.

[70]   A. Tuan Hoang *et al.*, "A review on application of artificial neural network (ANN) for performance and emission characteristics of diesel engine fueled with biodiesel-based fuels," *Sustainable Energy Technologies and Assessments*, vol. 47, no. January, p. 101416, 2021, doi: 10.1016/j.seta.2021.101416.

[71]   T. Tien-Anh, "Comparative analysis on the fuel consumption prediction model for bulk carriers from ship launching to current states based on sea trial data and machine learning technique," *Journal of Ocean Engineering and Science*, 2021, doi: 10.1016/j.joes.2021.02.005.

[72]   C. Gkerekos, I. Lazakis, and G. Theotokatos, "Machine learning models for predicting ship main engine Fuel Oil Consumption: A comparative study," *Ocean Engineering*, vol. 188, Sep. 2019, doi: 10.1016/j.oceaneng.2019.106282.

[73]   Z. Hu, Y. Jin, Q. Hu, S. Sen, T. Zhou, and M. T. Osman, "Prediction of fuel consumption for enroute ship based on machine learning," *IEEE Access*, vol. 7, pp. 119497–119505, 2019, doi: 10.1109/ACCESS.2019.2933630.

[74]   Y. R. Kim, M. Jung, and J. B. Park, "Development of a fuel consumption prediction model based on machine learning using ship in-service data," *Journal of Marine Science and Engineering*, vol. 9, no. 2, pp. 1–25, 2021, doi: 10.3390/jmse9020137.

[75]   D. Kim, S. Lee, and J. Lee, "Data-driven prediction of vessel propulsion power using support vector regression with onboard measurement and ocean data," *Sensors (Switzerland)*, vol. 20, no. 6, Mar. 2020, doi: 10.3390/s20061588.

[76]   A.~L.~Samuel, "Some Studies in Machine Learning Using the Game of Checkers," *IBM Journal of Research and Development*, vol. 3, no. 3, pp. 210–229, 1959, [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5392560

[77]   T. M. Mitchell, *Machine Learning*. McGraw Hill Higher Education, 1997.

[78]   A. Burkov, *The Hundred-Page Machine Learning Book*. 2019.

[79]   K. G. Shanthi, S. Sesha Vidhya, K. Vishakha, S. Subiksha, K. K. Srija, and R. Srinee Mamtha, "Algorithms for face recognition drones," *Materials Today: Proceedings*, no. xxxx, 2021, doi: 10.1016/j.matpr.2021.06.186.

[80]   A. Gupta, A. Anpalagan, L. Guan, and A. S. Khwaja, "Deep learning for object detection and scene perception in self-driving cars: Survey, challenges, and open issues," *Array*, vol. 10, no. February, p. 100057, 2021, doi: 10.1016/j.array.2021.100057.

[81]   S. D. H. Permana, G. Saputra, B. Arifitama, Yaddarabullah, W. Caesarendra, and R. Rahim, "Classification of bird sounds as an early warning method of forest fires using Convolutional Neural Network (CNN) algorithm," *Journal of King Saud University - Computer and Information Sciences*, no. xxxx, 2021, doi: 10.1016/j.jksuci.2021.04.013.

[82] D. Graupe, *Principals of Artificial Neural Networks*. World Scientific Publishing Co. Pte. Ltd., 1999.

[83] M. Taylor, *Neural Networks: A Visual Introduction for Beginners*, no. 2. Blue Windmill Media, 2017.   [Online].   Available: https://eur-lex.europa.eu/legal-content/PT/TXT/PDF/?uri=CELEX:32016R0679&from=PT%0Ahttp://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:52012PC0011:pt:NOT

[84] S. Karsoliya, "Approximating Number of Hidden layer neurons in Multiple Hidden Layer BPNN Architecture," *International Journal of Engineering Trends and Technology*, vol. 3, no. 6, pp. 714–717, 2012.

[85] T. Vujičić and T. Matijevi, "Comparative Analysis of Methods for Determining Number of Hidden Neurons in Artificial Neural Network," *Central European Conference on Information and Intelligent Systems*, pp. 219–223, 2016.

[86] B. E. Boser, V. N. Vapnik, and I. M. Guyon, "Training Algorithm Margin for Optimal Classifiers," *Perception*, pp. 144–152, 1992.

[87] Lean Marine, "GENERAL INTRODUCTION - OPTIMIZATION OF PROPULSIVE POWER." pp. 1–6, 2017.

[88] "Copernicus Climate Data Store." https://cds.climate.copernicus.eu/

[89] C. C. Robusto, "THE COSINE-HAVERSINE FORMULA," *The American Mathematical Monthly*, vol.   64,   no.   1,   pp.   38–40,   1957,   [Online].   Available: https://www.jstor.org/stable/2309088%0AJSTOR

[90] A. Biran and R. López-Pulido, *Ship Hydrostatics and Stability: Second Edition*, Second. 2013. doi: 10.1016/C2011-0-07795-5.

[91] Mark Chisnell, *Tips from the Top, Chisnell on Instrument Techniques*, First. 1992.

[92] E. Thornhill, A. Wall, S. McTavish, and R. Lee, "Ship anemometer bias management," *Ocean Engineering*, vol. 216, Nov. 2020, doi: 10.1016/j.oceaneng.2020.107843.

[93] "XyGrib." https://opengribs.org/en/

[94] Q. Meng, Y. Du, and Y. Wang, "Shipping log data based container ship fuel efficiency modeling," *Transportation Research Part B: Methodological*, vol. 83, pp. 207–229, 2016, doi: 10.1016/j.trb.2015.11.007.

[95] G. W. Corder and D. I. Foreman, *Nonparametric Statistics for Non-Statisticians*, vol. 47, no. 07. 2009. doi: 10.1002/9781118165881.

[96] W. Mao, I. Rychlik, J. Wallin, and G. Storhaug, "Statistical models for the speed prediction of a container ship," *Ocean Engineering*, vol. 126, pp. 152–162, Nov. 2016, doi: 10.1016/j.oceaneng.2016.08.033.

[97] P. KARAGIANNIDIS, "Data-driven Ship Propulsion modeling with applications in the Performance Analysis and Fuel Consumption prediction," 2019.

[98] T. Uyanık, Ç. Karatuğ, and Y. Arslanoğlu, "Machine learning approach to ship fuel consumption: A case of container vessel," *Transportation Research Part D: Transport and Environment*, vol. 84, Jul. 2020, doi: 10.1016/j.trd.2020.102389.

[99] L. Walther, A. Rizvanolli, M. Wendebourg, and C. Jahn, "Modeling and Optimization Algorithms in Ship Weather Routing," *International Journal of e-Navigation and Maritime Economy*, vol. 4, pp. 31–45, 2016, doi: 10.1016/j.enavi.2016.06.004.

[100] W. Du, Y. Li, G. Zhang, C. Wang, P. Chen, and J. Qiao, "Estimation of ship routes considering weather and constraints," *Ocean Engineering*, vol. 228, no. July 2020, p. 108695, 2021, doi: 10.1016/j.oceaneng.2021.108695.

[101] S. Wang, B. Ji, J. Zhao, W. Liu, and T. Xu, "Predicting ship fuel consumption based on LASSO regression," *Transportation Research Part D: Transport and Environment*, vol. 65, no. October 2017, pp. 817–824, 2018, doi: 10.1016/j.trd.2017.09.014.

[102] C. Gkerekos and I. Lazakis, "A novel, data-driven heuristic framework for vessel weather routing," *Ocean Engineering*, vol. 197, no. November 2019, p. 106887, 2020, doi: 10.1016/j.oceaneng.2019.106887.

[103] A. Pagoropoulos, A. H. Møller, and T. C. McAloone, "Applying Multi-Class Support Vector Machines for performance assessment of shipping operations: The case of tanker vessels," *Ocean Engineering*, vol. 140, no. February, pp. 1–6, 2017, doi: 10.1016/j.oceaneng.2017.05.001.

[104] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, pp. 1–15, 2015.

[105] "Scikit-Learn." https://scikit-learn.org/stable/

[106] P. KARAGIANNIDIS, "Data-driven Ship Propulsion modeling with applications in the Performance Analysis and Fuel Consumption prediction," 2019.

[107] L. T. Leifsson, H. Sævarsdóttir, S. T. Sigurdsson, and A. Vésteinsson, "Grey-box modeling of an ocean vessel for operational optimization," *Simulation Modelling Practice and Theory*, vol. 16, no. 8, pp. 923–932, 2008, doi: 10.1016/j.simpat.2008.03.006.

[108] M. Haranen, P. Pakkanen, R. Kariranta, and J. Salo, "White, Grey and Black-Box Modelling in Ship Performance Evaluation," *1st Hull Performance & Insight Conference*, vol. 1, no. April, pp. 115–127, 2016.

[109] "ISO 19030 Ships and marine technology." [Online]. Available: https://www.iso.org/standard/63774.html

[110] E. Hagestuen, B. Lund, and C. Gonzalez, "Continuous Performance Monitoring – A Practical Approach to the ISO 19030 Standard," in *1st Hull Performance & Insight Conference*, 2016, vol. 2014, no. March 2016, pp. 49–61.

[111] L. Aldous, T. Smith, R. Bucknall, and P. Thompson, "Uncertainty analysis in ship performance monitoring," *Ocean Engineering*, vol. 110, pp. 29–38, 2015, doi: 10.1016/j.oceaneng.2015.05.043.

[112] "Keras." https://keras.io/

[113] A. Coraddu, L. Oneto, F. Baldi, F. Cipollini, M. Atlar, and S. Savio, "Data-driven ship digital twin for estimating the speed loss caused by the marine fouling," *Ocean Engineering*, vol. 186, no. March, p. 106063, 2019, doi: 10.1016/j.oceaneng.2019.05.045.
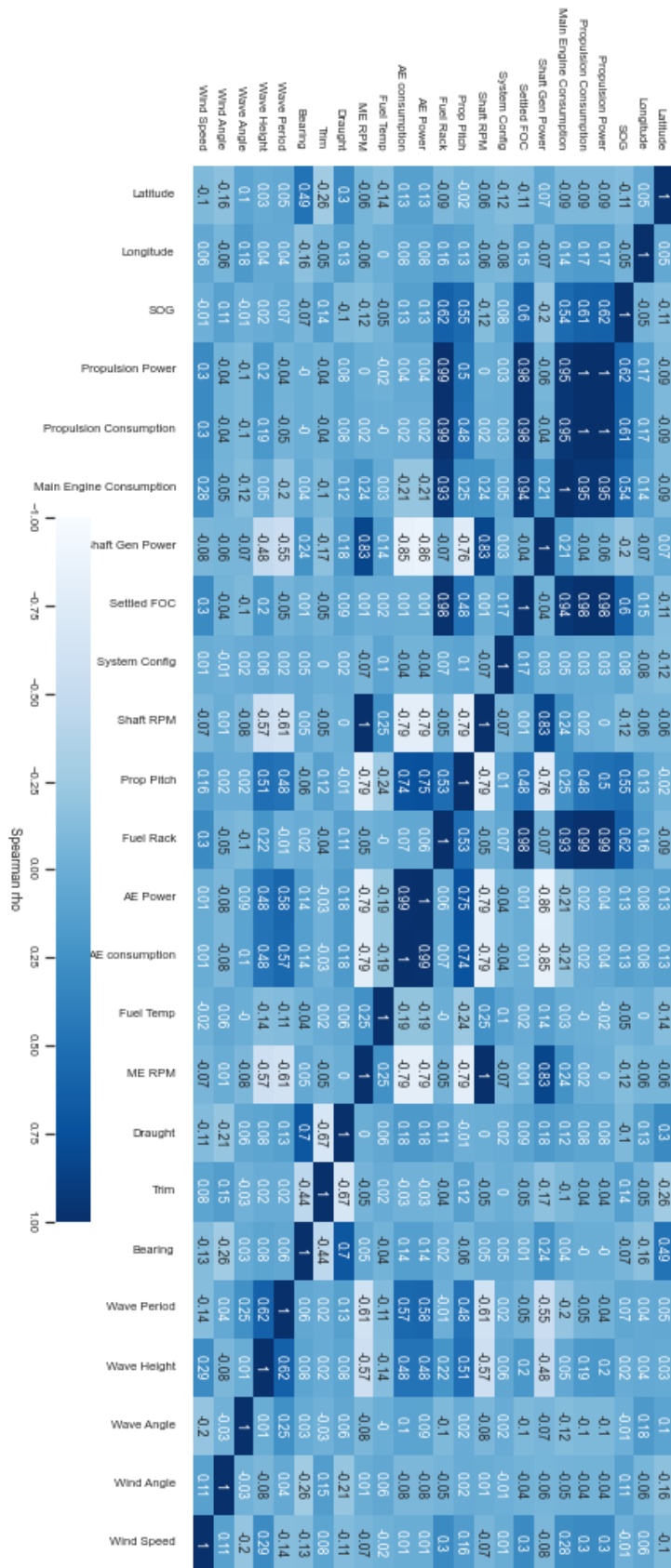
APPENDIX I – SPEARMAN'S RHO CORRELATION

Figure – Spearman's Correlation of all variables

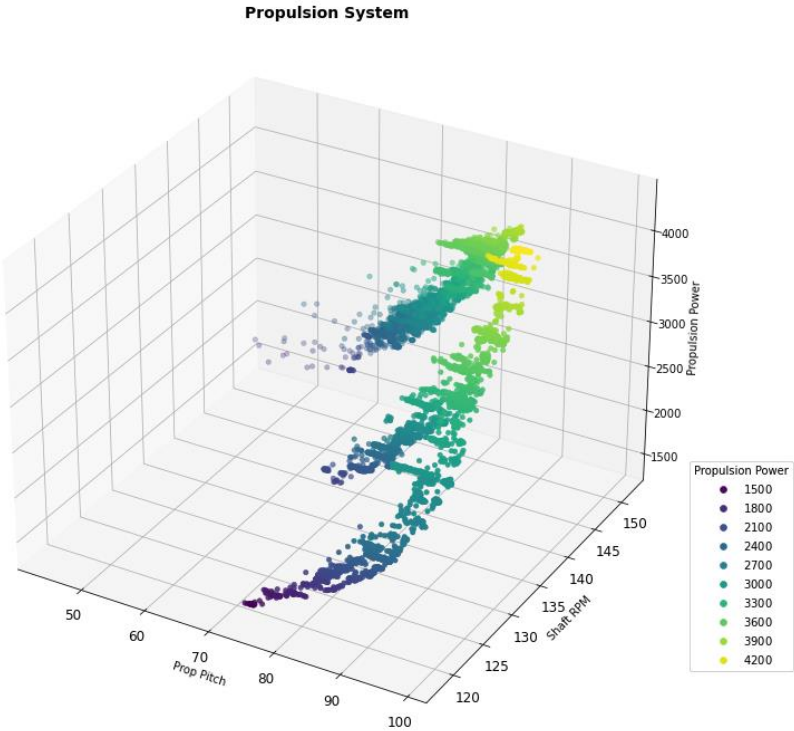# APPENDIX II – PROPULSION 3D GRAPHS

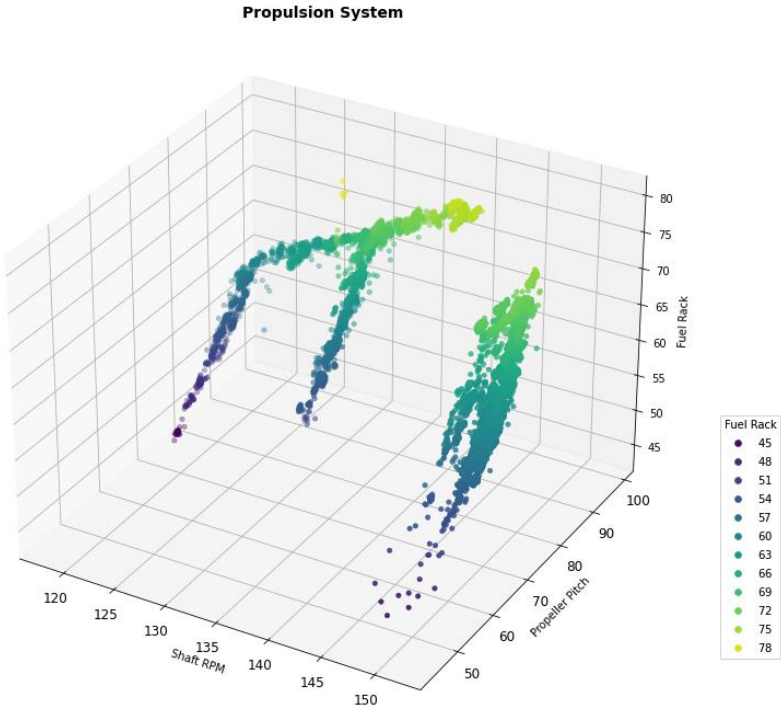*Figure – 3D Graph – Propeller Pitch, Shaft Rotation and Propulsion Power*

Figure - 3D Graph – Propeller Pitch, Shaft Rotation and Fuel Rack

APPENDIX III – PROPULSION SYSTEM VARIABLES DISTRIBUTION

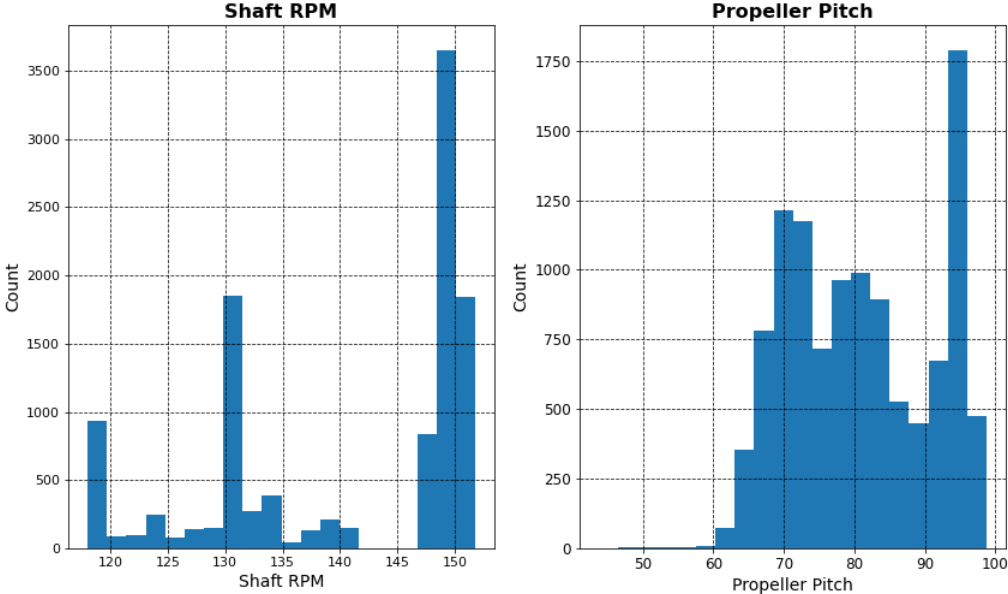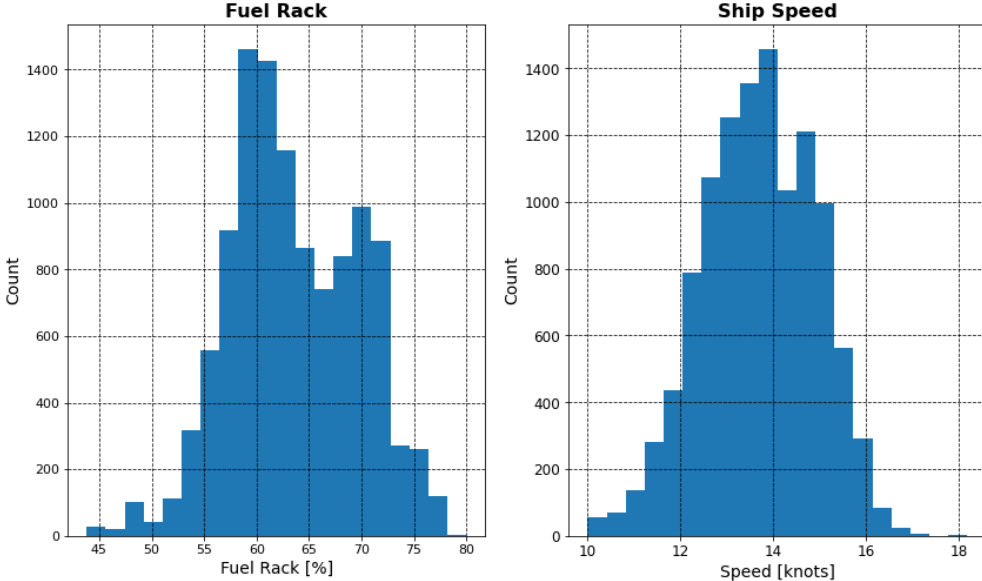*Figure – Shaft rotation and Propeller pitch distribution*

Figure – Fuel rack and ship speed distribution