Norwegian School of Economics
Bergen, Fall 2020

Estimating vessel environmental performance

A machine learning approach for predicting vessel fuel consumption and transparently quantifying the environmental sustainability impact of vessel exhaust gases

Fridtjof Gustav Jebsen & Sander Skogsrud Mathiesen

Supervisor: Roar Os Ådland

Master thesis, Economics and Business Administration

Major: Business Analytics

NORWEGIAN SCHOOL OF ECONOMICS

This thesis was written as a part of the Master of Science in Economics and Business Administration at NHH. Please note that neither the institution nor the examiners are responsible – through the approval of this thesis – for the theories and methods used, or results and conclusions drawn in this work.

Preface

Initially, we had several economical subjects that we wanted to investigate. After discussing topics throughout last fall, it became clear to us that we wanted to combine our common passion for business analytics and strong frustration over corporate 'greenwashing' in our thesis. From this, we concluded that the shipping industry would be an interesting subject due to its importance to world trade and impact to global greenhouse gas emissions. Writing this thesis has been highly educational, but also challenging at times. First and foremost, we would like to thank our supervisor, Roar Os Ådland, for constructive feedback and support. Your guidance and extensive knowledge within the field of maritime economics has truly been helpful. We would also like to thank Marianne Tronstad Lund for valuable input regarding climate research. In addition, we would like to express our gratitude to Jonas Andersson for our discussions concerning statistical modeling. Lastly, we would also like to thank the Terravera Foundation for valuable support.

Norwegian School of Economics

Bergen, December 2020

F.G. Jebsen

Fridtjof Gustav Jebsen

Sander S. Mathiesen

Sander Skogsrud Mathiesen

Abstract

The shipping industry faces a number of challenges regarding its share of total anthropogenic emissions worldwide. A range of measures have been initiated, both by official and private parties. Nevertheless, there are indications that the variety of approaches and the lack of recognized industry standards are creating confusion, resulting in ineffective action against the significant problem of climate change. The growing sense of urgency in relation to global warming, as well as disappointment in the International Maritime Organizations as the main regulatory body for global shipping, have created uncertainty and made it more difficult for decision makers to predict the future of the industry. At the same time, recent advances in data analysis mean that decision makers are able to produce better empiric modeling of emissions estimations, potentially improving their operations, regulations, or policies. In this thesis, we propose a machine learning approach for estimating the environmental performance of vessels. The theoretical framework established by Gibson et al. (2019) serves as a foundation for our research. Accordingly, we establish a transparent and quantitative approach for estimating environmental sustainability impact of vessel exhaust gases. Our approach uses predicted operational data derived from the Gradient boosting method, together with the best available measurements of emissions impacts, to portray the complexity of environmental sustainability. Our findings show the value that empirical modeling, in the form of machine learning, can provide to internal and external decision makers who compute and apply emissions estimates, both in the short and long term.

Keywords – Machine learning, Vessel environmental performance, Shipping, Sustainability

Contents

List of Figures

List of Tables

1 Introduction

International shipping has been a fast-growing sector of the global economy and its 2.89% share of total anthropogenic emissions is significant, causing effects on climate, air quality, and human health (Fuglestvedt et al., 2009; Smith et al., 2014; IMO, 2020b). The principal exhaust gas emissions from vessels are released into the atmosphere from ship stacks and diluted with the ambient air.

By 2050, emissions from shipping are projected to increase by 90-130% of 2008 levels for a range of plausible long-term economic and energy scenarios (IMO, 2020b). This is inconsistent with the internationally agreed goal of keeping global temperature increases to less than 2°C compared with pre-industrial levels, which requires worldwide emissions to be at least halved over the same period (European Commission, 2007). The issue is therefore attracting increasing international scrutiny (Jia et al., 2017). Yet more than 90% of all cargo in the world is transported by sea (Wu and Xia, 2018) and significant and economically feasible changes cannot be made overnight.

Environmental sustainability is defined as the effects of human activity on the environment (Moldan et al., 2012). The International Maritime Organization (IMO), as the main regulatory body for international shipping, has committed to regulate energy efficiency in the industry, controlling marine greenhouse gas (GHG) emissions via technological development, operational measures, and the use of alternative fuels (Jia et al., 2017; IMO, 2020b). Examples of marine sector regulations and optional instruments include the Ship Energy Efficiency Management Plan (SEEMP), the Energy Efficiency Design Index (EEDI), and the Energy Efficiency Operational Indicator (EEOI) (Smith et al., 2014; IMO, 2020b).

Additionally, there are a number of independent initiatives which embody the European Commission's notion that the benefits of acting to limit global warming far outweigh the costs of reducing GHGs (Gibson et al., 2019). There are a range of research groups, national maritime bodies, port authorities, and classification societies who promote their own energy efficiency indices, such as the Clean Shipping Index (CSI), the Environmental Ship Index (ESI), Rightship's Existing Vessel Design Index (EVDI), the Sea Cargo Charter, and the Poseidon Principles. Moreover, certain shipping companies also promote an increased focus on sustainability in their business models through green investments and the publication of sustainability reports.

Despite these initiatives, the growing sense of urgency and increasing disappointment at the slow pace of change has forced the IMO to propose additional amendments to cut ship emissions, such as energy efficiency measures, carbon intensity targets and rating systems (IMO, 2020a,b). The EU, meanwhile, is pushing for additional decarbonizing of shipping with their growth strategy, the European Green Deal, with the ambition of having a climate-neutral Europe by 2050 (European Commission, 2019). This strategy proposes to extend emissions trading to encompass maritime transport (European Commission, 2020a). To further ensure that the shipping industry reduces its emissions, the EU may use other regulatory and non-regulatory tools such as climate law, carbon pricing and new alternative fuel rules, as well as funding and market-proving innovations through different forms of research and development (European Commission, 2020b,c). Pressure on the IMO as a regulatory body has arguably intensified, in accordance with the increased demand for better and more effective regulations. This makes it more difficult for decision makers to predict future trends within the shipping industry.

Our research and feedback from different participants in the shipping industry indicate that there is confusion and room for improvement related to environmental sustainability. First and foremost, there is no standardized way of measuring energy efficiency, as several indices and calculation methods exist. This is understandable, given the complexity of both environmental calculations and the intricacy of transport and trade links between continents. Working in any of these fields requires a fundamental understanding of complex systems in addition to years of industry experience. Thus, one can understand the difficulty of implementing additional environmental sustainability measures without further standardization.

At the same time, the quantity and quality of data available in recent years have created new opportunities to model complex relationships without deep industry experience. Proper data analysis could improve decision makers' ability to make calculations and estimations and, hence, their capability of making informed choices. This raises the important question of whether those in authority can make the transition to more empirical modeling of emissions estimations, and thereby potentially improve their operations, regulations, and policies.

In this thesis, we propose a machine-learning approach to estimate vessels' environmental performance. Machine learning cannot directly replace specific field experience or human intuition and is dependent on the quality of data input. Nonetheless, through the use of machine learning methods, one can model relationships too complex and time-consuming for humans to calculate. These in turn can provide valuable insights that decision makers can use, distinguishing machine learning from other methods used and assumptions made in the literature. In particular, it can be usefully applied to the interaction between a vessel and its environment (in terms of speed, direction, sea conditions, and wind), which can lead to highly complex and non-linear relationships between variables.

Data from 16 medium-sized oil tanker vessels from a three/four-year time period have been used in our thesis. There are various reasons why oil tankers are of relevance when modeling vessels' environmental performance. The vessel type represents 25.1% of the total world fleet, as measured in deadweight tonnage (DWT). Together with container ships and bulk carriers, oil tankers are responsible for the majority of shipping's GHG emissions (IMO, 2020b) and their homogeneous cargo is produced and consumed in every corner of the world. This provides us with a standard type of vessel experiencing greatly varying conditions, which is beneficial for machine-learning predictions. Lastly, vessel owners benefit from near-continuous usage which, in combination with regulatory demands for reporting, means that oil tankers typically generate a solid foundation of data for analysis.

Previous studies have applied machine learning to solve different problems within shipping. For example, the method has been used to predict vessel fuel consumption, either with the objective of reducing operational costs or emissions - as fuel is the main driver for both (Stopford, 2008; Greene and Lewis, 2019). Other researchers have looked into why and how the calculation of emissions are applied by the industry. For example, Gibson et al. (2019), argue that there is room for improvement, and present a theoretical framework for calculating and applying vessel environmental performance to this end.

This thesis fills a gap in the literature by proposing a machine learning approach for the transparent estimation of a vessel's environmental performance. To the best of our knowledge, this is the first study that uses machine learning-predicted fuel consumption results in order to estimate vessel environmental sustainability. Our research is focused on the effects and choices regarding vessel exhaust gases, and builds further upon the theoretical framework of Gibson et al. (2019).

The contribution of our work is threefold. Firstly, we enhance the literature on machine learning predictions of vessel fuel consumption and emissions, by clarifying existing methods through the use of a larger sample than comparable studies. Secondly, we contribute to the literature on environmental sustainability within shipping by analyzing the environmental sustainability impact of vessel exhaust gases. Thirdly, we assist decision makers with estimation and implementation of sustainability, by combining insights from both machine learning and environmental sustainability through a practical tool.

The remainder of this thesis is structured as follows. In Section 2, we will review relevant literature regarding environmental sustainability, performance measurement, and machine learning predictions of vessel fuel consumption and emissions. Section 3 includes a discussion of sample data and related decisions of importance. In Section 4, machine learning theory, regression methods and evaluation metrics are presented. Section 5 explains the methodology by which this theory is implemented, as well as our approach to processing the results through the use of the theoretical framework devised by Gibson et al. (2019). Section 6 involves our results and a discussion and evaluation of our approach, based on two specific examples. In Section 7 we discuss limitations to the thesis and considerations for further research. Lastly, in Section 8 we present a conclusion to the thesis.

2 Literature Review

The literature review will first cover relevant past work on environmental sustainability and the use of performance measurement within the shipping industry. Afterwards, we will focus on literature relevant to modeling and the prediction of vessel fuel consumption and emissions.

2.1 Environmental sustainability

Sustainable maritime transportation involves complex decisions and multiple actors (Smith et al., 2014). Subsequently, sustainability considerations involve several and usually conflicting objectives. As an example, minimizing fuel emissions, an indicator of environmental sustainability, and maximizing service levels, a performance metric for economic prosperity, cannot be achieved at the same time (Moldan et al., 2012). As a result, the majority of the existing decision tools for maritime transportation focus on cost and/or operational performance indicators (Mansouri et al., 2015).

Growing attention to shipping's substantial effect on the global climate has had consequences, however. For instance, Mansouri et al. (2015) find that there has been a sizable increase in the number of scientific publications focusing on environmental sustainability within shipping. Furthermore, the majority of the papers apply some form of technical approach as a solution to their perceived problem. Through our research of the available literature, we find that such proposals include: alternative fuels (Bengtsson et al., 2011; Balcombe et al., 2019); fleet deployment (Fagerholt et al., 2009); fuel cells (Shih et al., 2014); fuel lifecycle calculations (Greene et al., 2020); hull cleaning (Adland et al., 2018); route optimization (Bui-Duy and Vu-Thi-Minh, 2020; Jia et al., 2017); scheduling optimization (Brouer et al., 2013; Johnson and Styhre, 2015; Lam, 2010); slow steaming (Cariou, 2011; Yin et al., 2014; Woo and Moon, 2014); speed optimization (Adland et al., 2020; Adland and Jia, 2016, 2018; Ballou et al., 2008; Kim et al., 2014; Psaraftis and Kontovas, 2013; Sheng et al., 2014; Wang et al., 2018a); the use of solar energy (Yu et al., 2018; Wang et al., 2019a), trim (Islam and Soares, 2019); vessel design (Motley et al., 2012; Doulgeris et al., 2012); wave energy (Alujevic et al., 2019); weather routing (Balmat et al., 2011; Windeck and Stadtler, 2011); and wind energy (Ionescu et al., 2015; Rehmatulla et al., 2017).

Mansouri et al. (2015) claim that the research on environmental sustainability has proven to be helpful theoretically, but overall lacks practical application. Despite the extensive research, the authors argue that many of the proposed solutions are not implemented in practice, mainly due to a knowledge gap between academia and the industry. Additionally, Lister et al. (2015) argue that the number and diversity of environmental initiatives available to shipping companies may cause confusion, create an administrative burden, or even hinder the progress of sustainability, due to the widely different audiences they are designed to target.

An exploratory study by Gibson et al. (2019) sought to explore the effectiveness of existing environmental initiatives, in order to promote improvements beyond regulatory requirements. Generally speaking, with the exception of regulations developed by the IMO, the majority of initiatives analyzed were found to be flawed for several reasons. Firstly, 80% of initiatives located in the literature were inaccessible via the public domain, effectively making them impossible to use. Secondly, those available for analysis were found to be lacking in transparency, as few demonstrated the calculation of vessel environmental scores, and those that did had registration and membership requirements. Thirdly, there were notable differences in environmental scope between the initiatives, with some focusing on single pollutants, and others examining multiple emissions and discharges to the environment. Furthermore, in cases where several emissions were evaluated, there were considerable differences in weighting factors chosen to measure their damaging effects. The rationale behind the allocation of pollutant weighting factors was therefore perceived as unclear, with a risk of subjective differences between initiatives, as they could be biased towards certain vessels or targets. Lastly, the authors point out that vessel pollutants are mostly assessed based on design parameters rather than actual performance criteria. Murphy et al. (2013) argue through the use of SO_x and CO_2 assessments that this method can be misleading and not necessarily representative of actual fuel consumption.

As a consequence of their findings, Gibson et al. (2019) propose a more suitable framework for transparently determining and quantifying the environmental performance of vessels, through the use of objective weighting factors based on the environmental impact of pollutants and operational data.

2.2 Prediction of vessel fuel consumption and emissions

In choosing a method for processing operational data and devising an estimation tool, it is necessary to review the literature regarding the modeling of vessel fuel consumption and emissions.

Traditional "resistance modelling" is the theoretical foundation of ship fuel consumption (Telfer, 1926; Todd, 1967), and aims to estimate the ship's total resistance (power requirement) as a function of speed and external factors such as wind and waves. Formally, fuel consumption can be calculated through the following Equation 2.1:

$$FC = \frac{(P_s + P_w + P_a)}{\eta} \tag{2.1}$$

where P_s is the effective horse power for propulsion at the propeller for a given speed (Lewis, 1988; Lloyd, 1989; Lindstad et al., 2013; Meng et al., 2016), P_w and P_a account for the additional power requirement due the impact of waves and wind respectively, and η is defined as propulsion efficiency. The latter is a product of the efficiency of the main engine, propeller efficiency, and their interaction at different engine loads and under various external operating environments. Hence, it is a function of engine load, the specific fuel consumption function (SFC), and refers to the fuel required per kWh (g/kWh) produced.

$$SFC_{ME,i} = SFC_{base} * (0.455 * Load_i^2 - 0.710 * Load_i + 1.280) \tag{2.2}$$

Source: IMO (2020b)

The function derived from Equation 2.2 has a U-shape due to sources of non-linearity: engine efficiency and wave-making resistance. Studies reveal a multitude of alternative determinants of fuel consumption with complex, stochastic, and potentially nonlinear interactions. For example, in addition to the main contributor, speed, the fuel consumption of a specific ship can be influenced by hull fouling, draught, trim, wave height and direction, wind force and direction, and water salinity, depth and temperature. Thus, traditional resistance modeling has been found to be too simple for the complex task of calculating vessel fuel consumption. Subsequently, several other functions, or white box methods (WBMs), have been proposed in the literature (Wang and Meng, 2012; Yao et al., 2012;

Psaraftis and Kontovas, 2013; Xia et al., 2015; He et al., 2017; Wang et al., 2019a; Meng et al., 2016). In general, the studies find the vessels' speed to be the principal determinant of fuel consumption. Yet residual resistance from weather conditions can affect its relative importance (Lo and McCord, 1995; Meng et al., 2016; Adland et al., 2018; Du et al., 2019).

Even though these studies achieve similar results, they differ in terms of their complexity and approach to modeling, from applying principles of naval architecture to their use of raw data. In one comprehensive example of the former, Kristensen (2019) estimates the ship-fuel elasticities under different weather conditions for three vessel types (seven tankers, six bulk carriers and five container ships) through traditional resistance modelling methods. Pedersen and Larsen (2009) revolutionized the latter, arguing for the application of artificial neural networks (ANN) to predict propulsive power from the variables that might influence ship resistance, such as ship speed, relative wind speed and direction, and air and sea water temperatures. Other recent empirical studies have also applied ship log data from noon reports (Wang and Meng, 2012; Du et al., 2019; Adland et al., 2018, 2020), or used vessel positions indicated by data from the Automatic Identification System (AIS) (Yang et al., 2019a). Some studies have leveraged the increased availability and quality of empirical data in recent years by further including "black-box" machine learning methods (BBM), such as ANN, in their studies. This might involve combining these methods together with elements from naval architecture principles, WBMs, in order to create hybrid "gray-box" models (GBM) for the optimization of trim (Coraddu et al., 2018), analysis of fuel efficiency (Meng et al., 2016), or estimation of the fuel-consumption function (Leifsson et al., 2008; Yang et al., 2019b). In our thesis, we use BBMs and the following passage will recognize relevant scientific studies.

Bocchetti et al. (2013) used multiple linear regression (MLR) from a set of navigation parameters to estimate fuel consumption intervals for specific voyages. The authors used data from two twin cruise ships and collectively 846 voyages. They argued that the prediction interval allowed management to better analyze the effectiveness of operations and prepare future remedial actions. Two years later, the authors conducted a similar study with a single cruise ship and 361 voyages, this time including additional parameters such as weather features (Bocchetti et al., 2015).

Bal Besikci et al. (2016) conclude that ANN-based fuel prediction is suitable when using noon reports to analyze the bunker fuel efficiency of a single oil tanker from 233 noon reports. Jeon et al. (2018) further show that ANN outperforms more traditional methods, such as polynomial regression and support vector machine (SVM) learning models, in terms of both accuracy and efficiency, when using data from the main engine of a vessel to predict fuel consumption. However, Wang et al. (2018b) suggest that, due to the complexity of vessel fuel consumption calculations, predictions should be performed through the use of least absolute shrinkage and selection operator (LASSO) regression. The study considered a dataset of approximately 800 voyages for 97 different vessels, and found that LASSO outperforms ANN, SVM, and the gaussian process in terms of both accuracy and interpretability.

As the number of studies and use of different machine learning methods has grown, Lepore et al. (2017) decided to conduct a comparison of 12 different approaches, with the objective of predicting CO_2 emissions per voyage for a cruise ship. Dependent on the voyage length, the study found that LASSO and Gradient boosting (GB) were the most accurate methods. More recently, Uyanık et al. (2020) conducted a similar comparative study, but for the prediction of fuel consumption using a broader spectrum of variables, including 724 noon reports and engine and sensor data from a container ship. In the study, they compare 14 different machine learning methods and evaluate them through the use of the performance metrics RMSE, MAE and R^2. The study finds the best performing methods to be Bayesian ridge regression (BRR), Kernel ridge regression, MLR and ridge regression (RR). Abebe et al. (2020) employ a variety of different machine learning regression techniques to improve the energy efficiency of shipping. The study uses machine learning with AIS and weather data and finds that linear regression and polynomial models achieve inaccurate predictions due to the highly non-linear tendency of a vessel's speed over ground with time, supporting the opinion that non-linear models may be more suitable for voyage and vessel predictions.

With this context in mind, a central part of our thesis is the notion that most models within the shipping literature apply assumptions for the calculation of fuel use and emissions. This also includes the IMO's methodology for calculating aggregated emission estimates and future shipping emissions scenarios. For an overall understanding of how this is done, see Figure 27 and Chapter 2 (especially 2.2.5) in the IMO GHG study of 2020 (IMO, 2020b). In short, they apply AIS-data, in combination with other data sources, and standard models with several assumptions to estimate and aggregate emission figures. One example is estimating main engine operational power demand, where ship performance under design conditions needs to be altered by several correction factors, such as for weather, fouling, speed, and vessel design specifics. Conversely, Adland et al. (2020) argue through empirical findings that the constant speed-consumption elasticity is only valid in practice near the design speed. Taking this examples into account, it can be argued that some assumptions in the IMO's estimations may be imperfect. Hence, a proposed alternative is to employ greater use of data-driven models for the estimation of aggregated fuel consumption and emissions.

Overall, we find that substantial work has been done on environmental sustainability within shipping, including machine learning predictions of vessel fuel consumption. However, we also observe that confusion persists in different areas. Despite academia's substantial research and technical propositions, the industry is apparently struggling to apply academic findings in their operations. Furthermore, machine learning seems to be a proven method that adds value, but existing studies are limited to rather small samples and differ in their opinion of which method performs best. Our contribution to the literature is therefore the following. Clarifying the performance of different machine learning methods to best predict vessel fuel consumption with a large dataset, structured analysis of the environmental sustainability impact of vessel exhaust gases, and proposal of a approach on how to practically combine the disciplines.

3 Data

This section of the thesis aims to describe our relatively large sample of data and related decisions of importance. For this study we have made use of operational data from 16 oil tankers from a single anonymous international shipping company. The oil tankers are divided into two categories, with 10 Product tankers and six Suezmax class tankers with 120,000 DWT and 159,000 DWT capacity respectively. Data for the former is taken from January 2012 to March 2016, and for the latter from January 2013 to December 2016.

A “one size fits all” approach is not appropriate within the shipping industry (Gibson et al., 2019). Even though both types of vessels in our example are oil tankers, we have decided to keep the data in two separate datasets: Product tankers and Suezmax. Our reasoning behind this is twofold. Firstly, we aim to exhibit how different-sized vessels react and respond to the same circumstances. If both datasets were analyzed together, we might lose aspects that are significant for one specific type of tanker but not the other; with this separate approach, we can evaluate behavior better for different conditions. Secondly, we want to highlight the possibility of comparing vessels with respect to environmental performance. With the use of standard units, we enable analysis across vessel types. It is also worth mentioning the common practice within scientific literature of splitting data in order to create greater insight. For example, it becomes possible to remove the influence of route or vessel-specific variations (Lepore et al., 2017; Adland et al., 2020). We chose not to split the datasets any further, however, as we wanted to maintain the possibility for type comparison, and to accommodate the machine-learning algorithm’s capability to recognize how variations can affect output.

A noon report is a manually prepared report which is conducted by the chief engineer of the vessel. The report includes operational vessel data, geographic position, and perceived sea and weather conditions. Our data consist of a total of 14,098 noon reports with variables such as: report date; vessel name; departure and destination ports; longitude and latitude; draft in metres; two measures of average daily speed in knots (speed over water (GPS-speed) and speed through water (LOG-speed)); fuel consumption in metric tons per day; daily distance in nautical miles; whether the vessel is ballast or laden; relative wind and swell direction and wind type; sea state and swell state. Broadly speaking, the noon report functions as a snapshot of the vessel's condition on a daily basis, which is why it is favored by the shipping industry for vessel assessment and analysis.

The following Table 3.1 exhibits the descriptive statistics of fuel consumption for the two vessel types. From the table we can observe that the bigger vessels, Suezmax, on average burn more fuel than Product tankers (36.08 metric tons a day versus 31.21 t/d).

	Product tankers	**Suezmax**
Mean	31.21	36.08
Median	32.60	36.70
St. dev.	11.76	15.21
Min	0.10	0.20
Max	76.00	82.40
Number of observations	9257	4841

Table 3.1: Descriptive statistics of daily fuel consumption for both vessel types

We recognize that there is some uncertainty related to the use of noon reports as a data source (Aldous et al., 2013). For example, given that shipping companies evaluate ship performance in terms of fuel consumption, and that the noon reports are manually computed by the vessel crew, one could argue that the vessel crew may have incentives to report worse conditions than actually experienced – this is known as "the principal agent problem". Subsequently, we find it necessary to conduct an analysis of the uncertainty related to weather, prior to the use of this data as our sample.

From the Copernicus ERA5 dataset we obtained components of northward and eastward wind at the height of 10 meters above the surface of the earth, and the significant height of combined wind and wave swells, for all available latitudes and longitudes, at the same time (12:00 hours), for the relevant years (Hersbach et al., 2018). The former components, in combination, give the horizontal wind speed. Both variables were hence transformed to the same unit of measurement as in the noon reports: knots and meters, respectively. As the crew onboard the vessel register wind and swell type in intervals, these variables were transformed to the average within each interval. The results of this analysis are exhibited in Figure 3.1 and Figure 3.2.

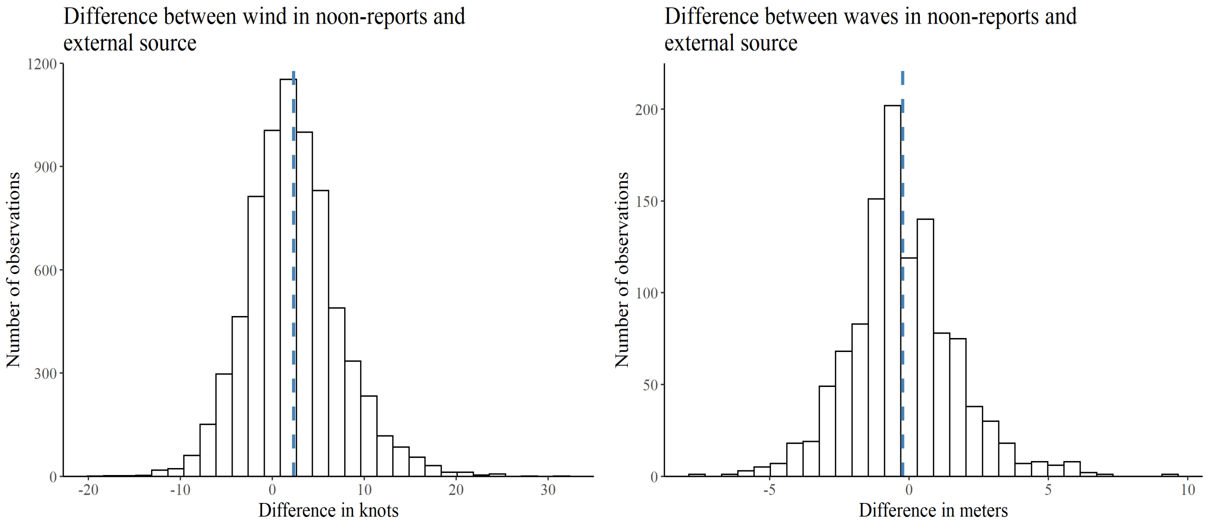


Figure 3.1: Analysis of reported wind and waves in noon reports versus external source for Product tankers. The blue dashed line illustrates the mean difference

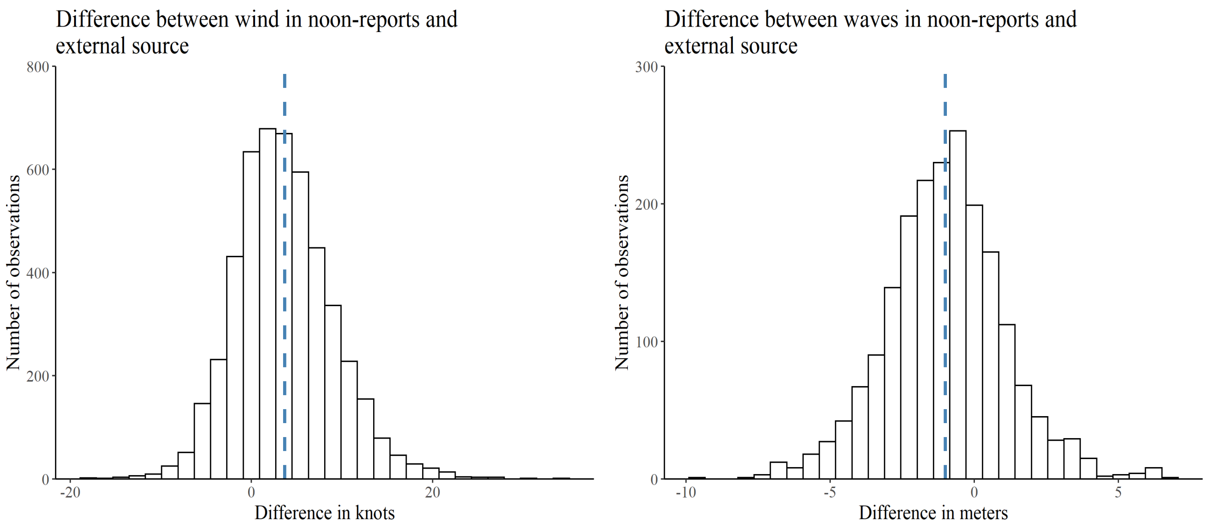


Figure 3.2: Analysis of reported wind and waves in noon reports versus external source for Suezmax. The blue dashed line illustrates the mean difference

Figure 3.1 and Figure 3.2 illustrate the difference in knots or meters between data recorded in noon reports and external information from the Copernicus dataset. The blue dashed line illustrates the average difference in either meters or tonnes, dependent on the figure. Positive values indicate that the noon-reports records greater figures. From Figure 3.1, we observe that the mean for Product tankers with regards to wind speed is slightly above zero, while the mean for wave height is below zero. Figure 3.2 reports a similar trend for Suezmax vessels. However, these deviations may be due to one or more of the following reasons. Firstly, the information obtained from the noon reports are recorded in intervals and by averaging them, we may have altered the data to show trends that do not exist. Secondly, it should be taken into consideration that even though both sources is a snapshot of the conditions at noon, the noon reports could be registered with some deviation in time (Aldous et al., 2013). Thus, potentially result in another observation than the external source. Consequently, the weather reported in the noon reports can be evaluated as plausible and hence are included in further analysis. For further visualization of this analysis on different parts of the dataset, see Appendix A.

Next, with the intention of improving predictability, we expand and modify our datasets. First, we include summer deadweight tons (SDWT) and vessel ages, collected from the company's own website. Furthermore, as the Copernicus ERA5 dataset contains hundreds of variables describing ocean surface and climate conditions globally, we have included weather features evaluated as adding value to the existing data from the noon reports. These are matched against vessel locations at a given noon report. The following 11 weather features were added: sea surface temperature; wave directional width; mean wave period; coefficient of drag with waves; total precipitation; and six dummy variables for different types of precipitation (rain, freezing rain, snow, wet snow, mix, and ice). The sea surface temperature is the temperature of the sea water near the surface. Wave directional width is a measure which indicates whether the waves are coming from similar directions or from a wide range of directions. The mean wave period is the average time, in seconds, between two consecutive wave crests. Finally, the coefficient of drag with waves is often referred to as the "friction coefficient" and is a parameter measuring the resistance that the ocean waves exert on the atmosphere.

Secondly, with base in the included variables we calculate trim and sea current, and create control variables. Trim is the difference between draft forward and draft aft. Sea current is the difference between LOG-speed and the GPS-speed: when the first measure is higher than the second, the resulting sea current will be negative – effectively the vessel will have a countercurrent – and vice versa. Furthermore, we apply 11 dummy variables for different types of wind (calm, light air, light breeze, gentle breeze, moderate breeze, fresh breeze, strong breeze, near gale, strong gale and storm), nine dummy variables for heights of waves (no wave, calm, smooth, slight, moderate, rough, very rough, high, and very high) and nine dummy variables for the type of swell (no swell, very low, low, light, moderate, moderate rough, high, rough, and very high). Additionally, we create dummy variables for the relative directions of the wind (headwind, sidewind or aftwind) and swell (head swell, side swell or aftswell).

In order to obtain the datasets in our preferred manner, some data processing was necessary. We can divide this process into the transformation of and exclusion of data. Longitude and latitude were transformed from a traditional format of degree, minutes, and seconds, into a decimal format, as used by Copernicus. Other cases where a transformation was needed include when data points were either located on land or were clearly an outlier in relation to other observations. These transformations were essential to match our observations with the Copernicus database. Additionally, they resulted in a larger proportion of observations available for analysis, and enhanced our visual plotting of voyages.

Furthermore, several observations were excluded, mainly due to two reasons. Firstly, all noon reports are manually plotted, making human errors possible. Secondly, outliers can affect our analysis greatly. Hence, we excluded all observations as follows: speed under 5 knots and over 18 knots; sea current below -10 knots and over 10 knots; travelled distance over 500 nautical miles a day; and registered fuel consumption below 0 or above 100 metric tons per day. We assume that all observations with an average speed of below 5 knots are associated with entry and exit of harbors and are therefore not of interest here. Even though the design speed of Product tankers and Suezmax are about 15 and 15.5 knots respectively, we set the limit of average speed at 18, as such observations are at the limit between possible outcomes and outliers or human errors. Furthermore, we assume that the current cannot have an absolute value larger than 10, and that both travel distance and fuel consumption below 0 and above 500 miles and 100 tons respectively are seen as impossible observations. Overall, we wanted to include as many plausible observations in our analysis as possible. Subsequently, the exclusionary constraints are arguably looser than strictly necessary.

As a result of this process of selecting and handling data, two final datasets are created and used as our sample. Table B1.1 in Appendix B exhibits descriptive statistics for each dataset, with 52 and 49 variables, respectively.

Lastly, to further visualize the observations in each of our datasets, Figure 3.3 exhibits every observation on a world map. The red dots illustrate vessels from the dataset of Product tankers while the blue dots indicate vessels from the Suezmax dataset. From this map, we can observe that the vessels travel to harbors across the globe, but Suezmax vessels mainly transport cargo in the Atlantic Ocean while Product tankers vessels operate with a greater variety of destinations.

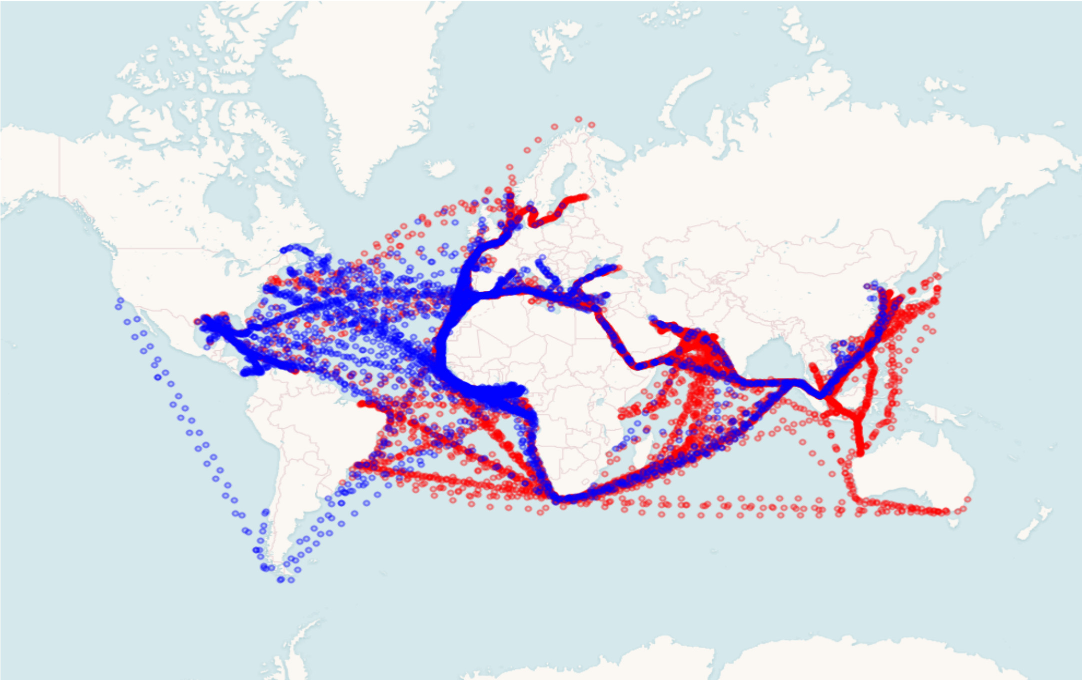


Figure 3.3: Illustration of vessel positions in our datasets. Red dots exhibit Product tankers. Blue dots exhibit Suezmax

4 Machine learning theory

4.1 Prediction with machine learning

In essence, statistical learning refers to a set of approaches for estimating a function f. There are two main reasons that we may wish to estimate f, either because we want to predict the output of the function (prediction), or because we aim to understand how changes in the functions input affects the output (inference) (James et al., 2013). For prediction purposes there is less focus on the relation between the dependent and the independent variables, as the goal is to create a function that best predicts the dependent variable (outcome). On an aggregated level most statistical learning methods can be characterized as either parametric or non-parametric (Breiman, 2001; James et al., 2013). The first one, parametric methods, assumes a specific stochastic data model (functional form or shape) with a randomized selection of independent variables. The second one, non-parametric methods, attempts to estimate a function that best explains the dependent variables based on given independent variables. Machine learning is part of the latter, within a category called algorithmic modeling, and contrasts from statistics by "concentrating on prediction by using general-purpose learning algorithms to find patterns in often rich and unwieldy data" (Bzdok et al., 2018).

Machine learning consists of mainly two elements: 1. a learning process that decides the best fit of the independent variables, and 2. an algorithm that based on the learning, attempts to model the context between the dependent and independent variables (Jung et al., 2018). The learning process can be split into two categories: supervised and unsupervised learning (Hastie et al., 2009; James et al., 2013). Supervised learning is about issues where there is little doubt about the connection between the independent variables and what kind of output is expected from the model. The challenge is to produce the correct quantification of the dependent variable, and the effect from the various independent variables. Unsupervised learning is more about an undefined problem, where you collect a large set of data to uncover connections. In this way, the independent variables can be categorized based on the relationships the algorithm finds. Based on machine learning theory and literature in the context of predicting fuel consumption within shipping, we only apply methods from the first category as we aim to predict a given dependent variable. More on our chosen machine learning methods and how they function is further explained in Section 4.2.

Prediction modeling demands some crucial elements in order to ensure credible results to learn from (James et al., 2013). Firstly, in order to apply a machine learning method, the sample data has to be clearly split into data that are used to train/fit the model and data which is used to evaluate the model's predictive ability. The train part of the dataset describes the part of the original data sample which the model fits its parameters on. Generally, the training set constitutes a considerable part of the total sample. After one has created the train set, the remaining part of the original dataset is used as the test set. However, the ratio is dependent on the model specific abilities, for example the need for computational power. Our chosen approach is further specified in Section 5. By testing the method with data which has already been used to fit the model, one will struggle to understand the model's true predictive capability on unseen data. Furthermore, in the literature, training and test sets are also referred to as "in-sample" and "out-of-sample". Thus, will we from now on apply these definitions.

Machine learning models perform extensive "tuning" through resampling methods to adapt the model so that the precision of the out-of-sample predictions are optimized. There exist several resampling methods, but we apply k-fold Cross-Validation (CV) in our thesis. This resampling method is preferred in the literature because of its strong ability and computational advantage of evaluating the success of applied estimation methods and to avoid overfitting (James et al., 2013; Molinaro et al., 2005). When fitting our machine learning methods, k-fold CV first divides the in-sample into a chosen number of k subsets of similar size. Next, one of the subsets is discarded and described as a validation set. The rest of the folds, k-1, makes up the CV-training set and is used to fit the model. The performance of this fit is evaluated by the predicted accuracy of the dependent variable within the validation set. Then, this is repeated k times, leaving each fold as a validation set one time. Next, the total machine learning method fit of the whole in-sample is the average of the k subsets' fit and is ultimately used to predict on the out-of-sample. For each fold performed one obtains more information about the fitted model, which means that various model parameters can be adjusted optimally. In this way, overfitting is counteracted, a problem where over-emphasis on in-sample observations reduces a machine learning method's predictive ability on unseen data. Our datasets consists of a great number of observations and multiple variables. Therefore, in order to make precise predictions, cross-validation is crucial in order to capture relevant characteristics (James et al., 2013). Moreover, this process of k-fold CV is also performed for hyperparameter tuning within each machine learning method.

An alternative resampling method to k-fold CV is to split the data into a train-, validation- and test-set. The intuition by doing so is similar to CV, by resampling the training data while still tuning the model and providing an unbiased evaluation of its performance on unseen data, before evaluating its final performance on the hold-out test-set (Russell et al., 2010). Yet, by applying k-fold CV in both training of the model and tuning the hyperparameters, the validation set is created within the training-set. Hence, removing the need for a separate validation set (Kuhn, 2008).

4.2 Regression methods

From our literature review we found similar studies that have predicted fuel consumption with use of machine learning on both design- and operational data. However, there are dissimilarities between what machine learning methods they find to give the most accurate predictions. Therefore, we have chosen to examine which of the suggested methods best predicts on our samples. The suggested methods are; Multiple linear regression (MLR), Ridge regression (RR), LASSO regression (LASSO), Bayesian ridge regression (BRR) and Gradient boosting (GB). Moreover, as a natural extension of RR and LASSO and GB, Elastic net (EN) and eXtreme gradient boosting (XGB) are also included. Furthermore, the Ensemble learning method stacking (EM) has shown interesting results in other fields (Wang et al., 2011; Divina et al., 2018; Wang et al., 2019b) and thus also added.

4.2.1 Multiple linear regression

Multiple linear regression is a mathematical model that can estimate a dependent output with multiple variable inputs. The MLR method is given in Equation 4.1.

$$Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n + \epsilon \tag{4.1}$$

Where Y is the independent variable, β_0 is the intercept and β_i is the coefficient to each independent variable x_i. Epsilon is the error term that the model cannot predict. (James et al., 2013).

4.2.2 Ridge regression

Ridge regression is very similar to ordinary least squares (OLS) regression, where the objective of the function is estimating a coefficient $\hat{X}$ that best represents the true parameter X, by finding the coefficients that minimize the sum of the squared residuals. However, RR solves the optimization problem by reducing the coefficients of the variables that correlate the most through use of a shrinkage penalty. Thus, reducing variance while keeping all variables. The function chooses the coefficient estimates $\hat{\beta}^R$ that minimizes the following:

$$\sum_{i=1}^{n}\left(y_i - \ \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 \tag{4.2}$$

Where the first part of the equation is the residuals sum of squares (RSS). The last part of the equation is the ridge shrinkage penalty which is determined by λ. (James et al., 2013).

4.2.3 LASSO regression

Least Absolute Shrinkage and Selection Operator (LASSO) was first introduced by Tibshariani in 1996. The shrinkage method is similar to RR by being based on OLS and reducing the coefficients that are the most correlated through a penalty. Nonetheless, the penalty works differently in LASSO as it provides the possibility of forcing some of the coefficient estimates to be exactly zero. This happens when the tuning parameter λ is sufficiently large, in practicality excluding variables through variable selection rather than reducing them as in RR. The method chooses the coefficient estimates $\hat{\beta}^L$ that minimizes:

$$\sum_{i=1}^{n}\left(y_i - \ \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j^2| \tag{4.3}$$

where the first part of the equation is the RSS. The last part of the equation is the LASSO shrinkage penalty which is determined by λ. (Tibshirani, 1996; James et al., 2013).

4.2.4 Elastic net

Elastic net is a combination of the OLS regression and the penalty introduced from both RR and LASSO. It is a model that can adapt to a wide range of applications and scenarios. Through selection of λ and α it is able to include both small contributions from a group of predictors while eliminating others. As λ becomes smaller the total shrinkage penalty is reduced. An $\alpha = 1$ results in only the LASSO penalty being used and opposite, if $alpha = 0$, then only the RR penalty is applied. The method chooses the coefficient estimates $\hat{\beta}^{EN}$ that minimizes the following equation

$$\sum_{i=1}^{n}\left(y_i - \ \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\left[(1-\alpha)\,\beta_j^2 + \alpha\left|\beta_j^2\right|\right] \tag{4.4}$$

(Hastie and Zou, 2005).

4.2.5 Bayesian ridge regression

Bayesian ridge regression differentiates from ordinary RR as both parameters alpha and lambda can come from a prior distribution. Instead of setting values for the parameters, the algorithm treats them as variables to be estimated from the sample. Through an estimation of a probabilistic model of the regression the algorithm can obtain more fits, which can result in a better fit than regular RR. The advantages of BRR is a better adaption to the data in hand, but can at the same time be more computationally demanding and give less inference. (Neal, 1996; Assaf et al., 2019).

4.2.6 Gradient boosting

Gradient boosting was introduced by Friedman (1999) and is a regression method sequentially fitting a decision tree, with a predetermined maximum depth, to minimize a loss function based on the "psuedo"-residuals from the previous tree. By giving each tree a small learning rate, nu, the algorithm takes a small step closer to the optimal value by each iteration. The model will do so either until it has reached a satisfied value or the maximum number of predetermined trees - $mstop$. The loss function fitted to each tree is given in Equation 4.5, which is the mean square error.

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \tag{4.5}$$

$\hat{y}_i$ represents every estimated output value, while y_i is the actual value. (Friedman, 2002).

4.2.7 eXtreme gradient boosting

eXtreme gradient boosting is a machine learning algorithm derived from traditional GB. It works similarly by making several trees learn sequentially from the "psuedo"-residuals from the previous tree, with the objective of finding the optimal values. What makes the eXtreme gradient boosting different is how each tree is made and fitted. Each tree is fitted by minimizing Equation 4.6.

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 + \gamma\, T + \frac{1}{2}\,\lambda\sum_{j=1}^{p}\beta_j^2 \tag{4.6}$$

The equation consists of a loss function, a tree pruning term and a penalty term. The loss function is equal to the one used in GB, as the sum of the residuals between the predicted and actual value. The tree pruning term consists of γ and T. γ is a relative penalty term determining how sensitive the split of trees should be, and T is the number of leaves allowed in each tree. Finally, the penalty term is approximately equal to the regularization term in RR. Both the regularization parameter λ and γ intend to reduce the prediction's sensitivity to individual observations and prevent overfitting. Next, similar to GB, each fitted decision tree is given a learning rate η, and the method generates trees until it either reaches a sufficient fit or the predetermined maximum number of trees - *nrounds*. Other hyperparameters used by the method is *min_ child_ weight*, *colsample_ bytree* and *subsample*. *min_ child_ weight* determines the minimum number of leaves in a tree for it to be included. *colsample_ bytree* and *subsample* is two parameters randomizing correlating trees with the objective of reducing variance. (James et al., 2013; Cho et al., 2014).

4.2.8 Ensemble method

Ensemble modeling is a technique that combines individual machine learning methods into one prediction model, with the objective of improving stability and a higher performance. By combining different methods the algorithm aims to use the strengths from all added methods and thus offset individual method variance and bias. The applied base methods should preferably be as diverse as possible. All methods are implemented through an ensemble learning technique called stacking. The technique runs each method separately before combining them. The meta-regressor, with all the methods, is then fitted through a multiple linear regression with a suitable weighting based on each method's individual performance. In our thesis, we include the former methods listed, as they are, as mentioned in Section 4.2, argued to be the best methods for prediction of vessel fuel consumption. (Wolpert, 1992; Breiman, 1996; Zhou, 2012).

4.3 Performance metrics

To evaluate the methods performance we need to quantify to which extent the predicted out-of-sample value is close to the true value. In this thesis, three different primary metrics (Botchkarev, 2018), are chosen for error measure; RMSE, MAE and MAPE. Additionally, the coefficient of determination R^2 has been included for model accuracy.

4.3.1 RMSE

RMSE stands for root mean square error and measures how well the final fitted model performs in relation to the true value in the sample. RMSE is calculated through the following:

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{4.7}$$

The formula takes the sum of the squared residuals and divides them by the number of observations. The residuals are squared to assess both negative and positive deviations. Because the residuals are squared, greater deviations are punished more than smaller ones. Interpretability-wise, RMSE performs better than mean square error (MSE), due to the counteraction of squaring. (Holmes, 2000).

4.3.2 MAE

MAE stands for mean absolute error and it calculated through Equation 4.8.

$$\frac{1}{n}\sum_{i=1}^{n} | \hat{y}_i - y_i| \tag{4.8}$$

Similar to RMSE, the formula divides the sum of the residuals by the number of observations. However, instead of squaring the residuals, MAE uses the absolute value to assess both negative and positive deviations on the same basis. Because it is on the same level as the data, it becomes easier to interpret, compared to other forms of error metrics. (Willmott et al., 2009)

4.3.3 MAPE

MAPE stands for Mean Absolute Percentage Error, and is calculated through the following formula:

$$\frac{1}{n}\sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \tag{4.9}$$

where error is defined as actual or observed value minus the forecasted value. Commonly, the results are multiplied by 100 to show the ratio as a percentage, and therefore often referred to as a percentage metric, where a smaller percentage indicates a good forecast. (Botchkarev, 2018; Swamidass, 2000)

4.3.4 R^2

R^2 is a statistical measure that represents the proportion of the variance in Y, explainable by the independent variable X. It is derived from the following equation:

$$1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{4.10}$$

(James et al., 2013)

4.3.5 Assessment of performance metrics

Botchkarev (2018) states that debate on which performance metric to use is common in the literature. Usually, discussions are based on the premise that there could be a single "ideal" metric that beats all others in all situations. However, other studies point out that the selection of appropriate error measures always will be a problem, because no single metric gives an unambiguous indication of performance, while the use of multiple measures makes comparisons between methods difficult and unwieldy (Mathews and Diamantopoulos, 1994; Goodwin and Lawton, 1999; Botchkarev, 2018).

Although, this thesis uses the current most popular metrics, they have also received criticism and rejection (Botchkarev, 2018). For instance, Willmott et al. (2009) argue that RMSE has disturbing characteristics and is inappropriate to be used as error metric. In addition to, strongly advising the literature to apply the metric MAE instead. However, Chai and Draxler (2014) partially dispute these findings, and argue that one distinct advantage of RMSEs over MAEs is that it avoids the use of absolute value, which is highly undesirable in many mathematical calculations. Foss et al. (2003) argue that MAPE is an unreliable selection criterion, and Kim and Kim (2016) argue that the scale-independent and interpretability advantages of the metric, should be viewed in context of its disadvantages when actual values are zero or close to zero, which results in large or infinite MAPEs.

Moreover, there is an ongoing debate of the importance of R^2 in the literature. On the one hand, Li (2017) argues that R^2 should not be used as measures to assess the accuracy of predictive models for numerical data as it can be biased, insufficient or misleading. Moreover, it is worth mentioning that R^2 always will increase when more variables are added, even though it does not improve the fit of the model (James et al., 2013). It has been proven that even with almost a perfect fit of a model, R^2 has been significantly low and vice versa (Ford, 2015). On the other hand, Lewis-Beck and Skalaban (1990) argued that although R^2 is not a unique measure of predictive capability, "it informs us, as no other statistic can, of the relative predictive capability of the model. Intuitively, it suggests how much we are reducing prediction error, relative to how much potential error there is".

In summary, no consensus on the "ideal" metric has yet been achieved and researchers now express a more practical view that there is no need to strive for a single best metric (Botchkarev, 2018). A combination of metrics, and an understanding of the different advantages and disadvantages they represent in relation to different types of problems and data, will arguably be the ideal option.

5 Methodology

This section of the thesis describes our application and modification of the theoretical framework drawn up by Gibson et al. (2019) for transparent quantification and determination of vessel environmental performance. The process consists of five parts. First, we establish the thesis's operational data through implementation of our samples and machine learning theory, in order to predict vessel fuel consumption. Then, we reconfigure the result of our machine learning method to vessels' greenhouse gas and pollutant emissions. Thirdly, we calculate pollutant scores for each emission based on vessel, cargo, and voyage-specific data. Next, pollutant weighting factors are established for the relevant emissions. Finally, we summarize and exhibit how to estimate vessel environmental performance through our methodology.

5.1 Implementation of machine learning theory and operational data

This subsection aims to describe our thesis's use of data and machine learning theory through the programming language R.

Firstly, the samples are split into in-sample and out-of-sample. As there is no definite standard in the literature and several types of splits are applied in comparable studies – 66%-33% and 80%-20% (Uyanık et al., 2020; Lepore et al., 2017) – we apply the average and general machine learning split of 75%-25%. Instead of splitting the samples randomly, we divide it with respect to equal distribution of the dependent variable $fuel$, with the *createDataPartition* function. In this way, we ensure balanced splits of the variable in our samples (Kuhn, 2008). Secondly, because some algorithms demand data in a certain form and are sensitive to the measurement units used (Lepore et al., 2017), the samples are standardized. Samples are thus both centered and scaled based on the mean and standard deviations of the in-sample. Next, the different machine learning methods are carried out through use of the Caret-package (Kuhn, 2008).

With regards to tuning, the method fit and hyperparameters are resampled through k-fold CV. K equal to 10 has been used, as this is most recommended in the literature (Witten and Frank, 2005; Molinaro et al., 2005). When tuning the hyperparameters, the most used strategy is a combination of grid and manual search (LeCun et al., 1998; Larochelle et al., 2007; Hinton, 2010; Bergstra and Bengio, 2012). In a grid search, one assembles every possible combination of the parameters and sequentially fits the method with each combination. Furthermore, one uses a grid search in combination with a manual search, to specify a region per parameter that the search can be applied to. The advantages of this strategy are its high degree of intuitiveness and insight. However, any strategy is only as good as one's access to a high dimensionality of computational power. Also, if the number of hyperparameters becomes substantial, this strategy can suffer from the curse of dimensionality (Bellman, 1961). Moreover, an overly precise and specific tuning of hyperparameters might result in overfitting. As an alternative strategy, Bergstra and Bengio (2012) argue that a random search for hyperparameters can provide similar or better results when the number of hyperparameters becomes considerable. Additionally, in response to the notion that a random search can perform worse in situations where there are few hyperparameters with a small area of possible values, they argue that the trade-off between a slightly worse method fit and vastly less computational demand is a favorable one. Simultaneously, the strategy possesses the same practical advantage of implementation and conceptual simplicity. Therefore, in our thesis we apply a random search for hyperparameters with a tune length of 5.

Then, with the use of different performance metrics, we determine the most suitable and accurate machine learning method. Lastly, the method is applied using its $finalmodel$ to predict fuel consumption and lay the foundation for the next step in our process.

5.2 Pollutant emissions

In order to estimate vessel environmental performance, predicted vessel fuel consumption has to be converted to equivalent emissions. Jia et al. (2017) state that the production of greenhouse gases and other pollutants are proportional to the amount of fuel burned, with the proportionality constant being known as the "emission coefficient". For the calculations of emissions in this study, we apply aggregated vessel/fuel-based non-machinery-type-specific coefficients from the fourth IMO GHG study of 2020 (IMO, 2020b), which are exhibited in Table 5.1. We chose to use fuel-based estimates rather than energy-based ones, as our machine learning method predicts vessel fuel consumption, not the amount of energy used. Moreover, we recognize that the values of the emission coefficients are not absolute and alter according to different factors. In this thesis, however, we apply the IMO's approach to fuel-based emission factors, which "leverages world energy statistics provided by IEA to estimate global shipping emissions for the period 2012-2017 and applies emissions factors based on the total mass of pollutants divided by the total mass of fuel consumption" (IMO, 2020b). These factors are estimated using a bottom-up approach, where engine types and loads for each vessel class is specified. We apply these figures, as the IMO argues that these are of the highest possible quality. Thus, emissions of relevant GHGs and pollutants are computed through the multiplication of predicted fuel consumption and each exhibited emission factor.

Pollutant	**Marine HFO** (g/g fuel)	**Marine LSHFO** (g/g fuel)	**Marine MGO/MDO** (g/g fuel)
CO_2	3.114	3.114	3.206
CH_4	0.00005	0.00005	0.000045
N_20	0.000175	0.000175	0.00018
NO_x	0.077255	0.077255	0.05488
CO	0.00285	0.00285	0.002465
NMVOC	0.00316	0.00316	0.002285
SO_x	0.04773	0.0196	0.00215
PM	0.007245	0.007245	0.000945

Table 5.1: Emission factors (IMO, 2020b)

5.3 Pollutant score

For assessment and comparison of environmental performance, we find it essential to consider emissions with information specific to the vessel, cargo, and voyage in mind.

As mentioned in the literature review, considerable research has been conducted to determine energy efficiency standards, benchmarks, and indicators for the shipping industry. One of the indicators, the energy efficiency operational indicator (EEOI), was introduced by the IMO (IMO, 2009) as a monitoring tool that companies can use for implementation of their ship energy efficiency management plan (SEEMP). It later became a requirement under the EU MRV Regulation of 2015 (Council of the European Union, 2015; Fridell et al., 2018). The EEOI is simply defined as the ratio of mass of CO_2 emitted per unit of transport work, where transport work "represents the actual maritime transport service determined by multiplying the distance travelled with the amount of cargo carried" (European Commission, 2020a). As the indicator varies according to the actual cargo carried over a given distance, it arguably reflects the emission intensity of the transport service rendered by each individual vessel. Hence, for proper assessment and comparison of vessel emissions, we find the EEOI indicator structure highly suitable for calculation of pollutant scores. Therefore, in order to objectively examine different vessels and voyages on the same scale, based on well-known standards, we create pollutant scores by dividing each emission by the given vessel's transport work, as exhibited in Figure 5.1.

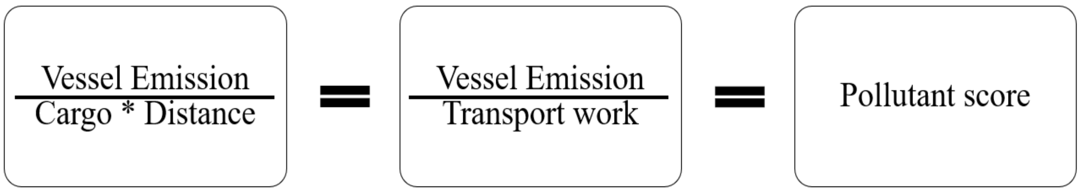


Figure 5.1: Illustration of the calculation of Pollutant score

5.4 Pollutant weighting factor

A number of important assumptions underlie the process of determining a suitable weighting factor for estimation of a vessel's environmental performance. For instance, it must again be emphasized that emissions resulting from shipping affect a number of aspects of life, having considerable effects on the earth's climate and ecosystem, on sea and air quality, and on human health (Fuglestvedt et al., 2009). Despite the broadly varying effects of emitted substances, one often aims through weighting factors or metrics to place their impact on a common scale, with the intention of enhancing comparison across sources. However, metrics are subject to a range of different considerations such as timelines, lifecycles, objectives, the nature of the problems addressed, and degrees of transparency and sophistication. Thus, it has become clear to us through this project that no single perfect objective metric exists that can fully explain all the harmful effects of emissions. Indeed, the question of which metrics are an ideal fit for measurement of environmental performance in the shipping industry fall outside the scope of this thesis, as we simply aim to create an estimation tool. Thus, as suggested by Gibson et al. (2019), our thesis uses the widely industry established indicator of Global Warming Potential (GWP) as a weighting factor. It is therefore necessary to explain the basics of the GWP indicator.

The GWP is a widely used indicator in the context of climate change for the assessment of greenhouse gases, ozone pre-cursors, and aerosols (Fuglestvedt et al., 2010). Any method that aims to aggregate such substances has to account for the different nature of the emissions, for instance their timescale of decay or temperature effects. The GWP indicator solves this by placing the emission of the substances on a common CO_2-equivalent scale. It does so by measuring the radiative forcing of a pulse emission of an pollutant, relative to the radiative forcing of a pulse emission of carbon dioxide, both integrated over a common time horizon for some background atmospheric state. The length of time horizon reflects the consideration of short- and long-term priorities of mitigation and warming targets. Typical horizon-values are 20, 50, 100 or 500 years. The Kyoto Protocol and the Paris Agreement have adopted GWP for a time horizon of 100 years (GWP100) (Fuglestvedt et al., 2010; UNFCCC, 2018) . Subsequently, the same horizon is relevant for the shipping industry as the IMO targets are consistent with the Paris Agreement's temperature and global decarbonization goals. Thus, our thesis will mainly apply GWP100, but also include figures for a time horizon of 20 years (GWP20) and 500 years (GWP500), to illustrate how the choice of time horizon might affect results. In those cases where current scientific literature disagrees on the effect and magnitude of the emission measured in GWP, the average has been applied, as suggested by Lund (2020). Thus, the following Table 5.2 exhibits the calculated GWP20, GWP100 and GWP500 CO_2-equivalents of the vessel emissions listed in Table 5.1.

Pollutant	**Name**	**GWP20**	**GWP100**	**GWP500**
CO_2	Carbon dioxide	1	1	1
CH_4	Methane	72	25	7.6
N_20	Nitrous oxide	289	298	153
NO_x	Nitrogen oxide	-53.5	-30.5	-9.3
CO	Carbon monoxide	7.5	2.53	0.76
NMVOC	Non-methane volatile organic compound	14	4.5	1.4
SO_x	Sulphur oxide	-140	-40	-12
PM	Black carbon*	1600	460	140
	Organic carbon*	-240	-69	-21

* Weighting of Black carbon and Organic carbon within PM is 15% and 39%, respectively.
(Lack et al., 2009)

Table 5.2: Pollutant weighting factor GWP

5.5 Vessel environmental performance

Finally, with the purpose of estimating vessel environmental performance, we apply and combine the elements from above, as exhibited in Figure 5.2. This final outcome of our methodology is our attempt in this thesis to explain a relatively complex picture of the environmental sustainability impact of vessel exhaust gases through a single score.

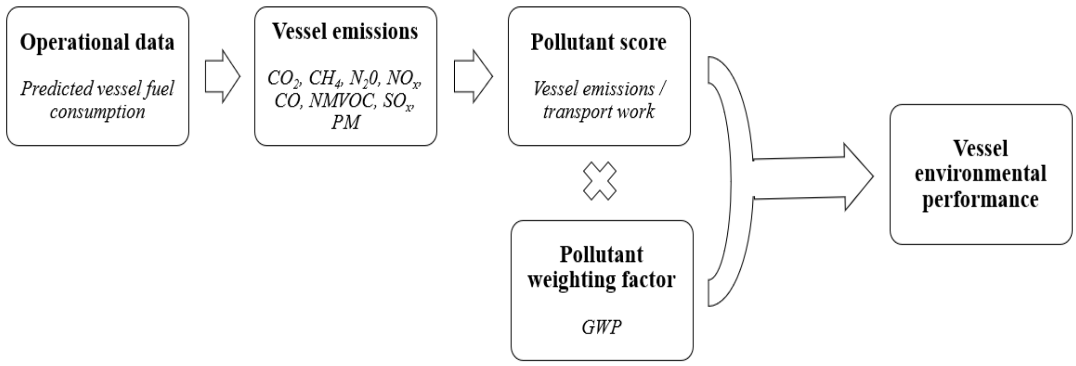


Figure 5.2: Illustration of our methodology to obtain vessel environmental performance

6 Results and discussion

This section describes the research findings of our thesis. Firstly, our process of using different machine learning methods to predict vessel fuel consumption for operational data is presented, discussed, and assessed. Secondly, vessel environmental performance is estimated through the use of computed vessel emissions, pollutant scores, and weighting factors. Following this subsection, we present, discuss and evaluate our approach with the use of two examples.

6.1 Evaluation of predictive methods and prediction of vessel fuel consumption

For the purpose of estimating vessel environmental performance, it is necessary to find which machine learning method predicts vessel fuel consumption on our out-of-sample most accurately. Table 6.1 describes each of the methods' hyperparameters and Table 6.2 exhibits each of the methods' performance results on the out-of-sample set.

	Performance metrics				**Computational time**
	RMSE	R2	MAE	MAPE	
Product tankers					
Multiple linear regression	0.46099	0.78622	0.35473	35.56 %	2,6 seconds
Ridge regression	0.46222	0.78616	0.35679	32.24 %	10,7 seconds
LASSO	0.46047	0.78664	0.35469	35.20 %	6,6 seconds
Elastic net	0.46048	0.78663	0.35470	35.16 %	3,6 seconds
Bayesian ridge	0.46033	0.78679	0.35443	34.77 %	37 minutes 39,8 seconds
Gradient boosting	0.35116	0.87602	0.26626	14.03 %	11 minutes 55,9 seconds
eXtreme gradient boosting	0.36428	0.86643	0.27377	16.93 %	22 minutes 58 seconds
Ensemble method	0.35279	0.87487	0.26716	15.50 %	2 hours 27 minutes 55,9 seconds
Suezmax					
Multiple linear regression	0.41701	0.83708	0.30832	25.73 %	1,6 seconds
Ridge regression	0.42013	0.83515	0.30990	23.64 %	8,9 seconds
LASSO	0.41785	0.83640	0.30852	25.56 %	5,2 seconds
Elastic net	0.41823	0.83601	0.30867	25.24 %	3 seconds
Bayesian ridge	0.41573	0.83797	0.30820	24.73 %	12 minutes 35,8 seconds
Gradient boosting	0.33167	0.89702	0.23411	13.23 %	7 minutes 34,9 seconds
eXtreme gradient boosting	0.33836	0.89266	0.23768	13.92 %	15 minutes 2,7 seconds
Ensemble method	0.33797	0.89288	0.24252	14.63 %	58 minutes 38 seconds

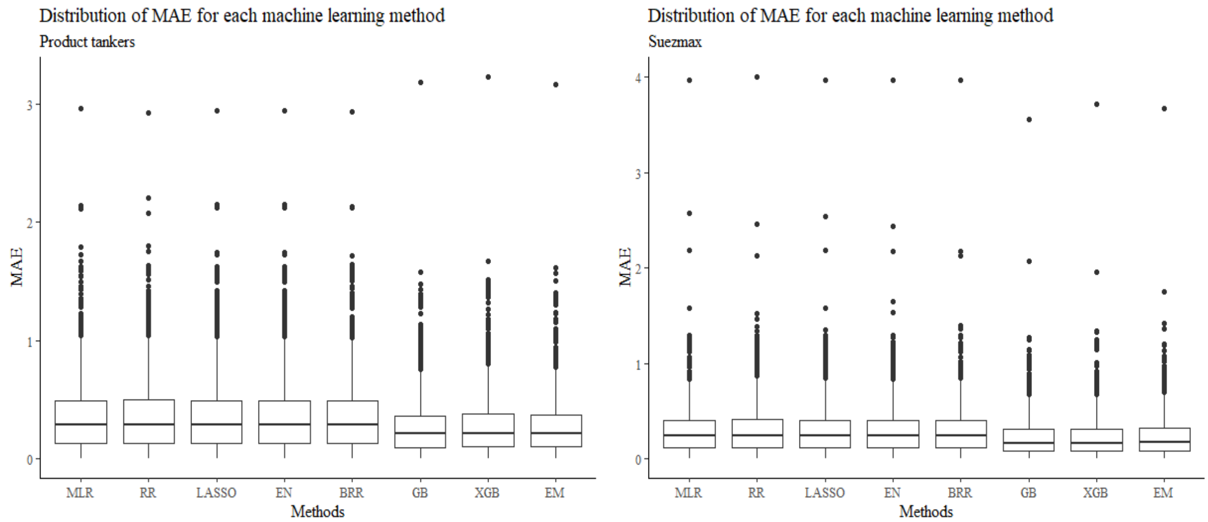
Table 6.2: Performance for each machine learning method

As expected, based on the difference in nature between the methods, there are dissimilarities in performance metrics and computational time.

Machine learning method	**Hyperparameters**	**Product tanker**	**Suezmax**
Multiple linear regression		None	None
Ridge regression	Lambda	0.076	0.077
LASSO	Lambda	0.002	0
Elastic net	Alpha	0.55	0.55
	Lambda	0.00332	0.00339
Bayesian ridge		None	None
Gradient boosting	mstop	250	150
	max_depth	5	5
	nu	0.1	0.1
eXtreme gradient boosting	Gamma	0	0
	min_child_weight	1	1
	nrounds	200	200
	max_depth	3	3
	eta	0.3	0.3
	colsample_bytree	0.6	0.8
	subsample	1	1
Ensemble method	Intercept	0.00118	0.00028
	MLR	1.36500	0.01471
	Elastic net	-1.32800	0.12130
	Bayesian ridge	-4,119e-14	1,688e-14
	Gradient boosting	0.53900	0.39870
	XGB	0.43200	0.47570

Table 6.1: Hyperparameters for each machine learning method. Within the Ensemble method, EN includes RR and LASSO.

Adland et al. (2020) have used the same samples as this thesis. Naturally, will it be interesting to compare their results with ours. Their study uses the performance metrics: R^2, RMSE and MAE. The latter two are metrics at the same aggregation level as data. As we standardize our samples differently, it becomes unnatural to compare these values amongst our studies. With regard to R^2, it must be noted that we do not emphasize R^2, due to the number of variables our methods use, ref Section 4.3.5. Furthermore, it can be argued that there are significant methodological differences between our chosen methods. Thus, we only compare MLR-results as this method does not require tuning of hyperparameters. We observe from Adland et al. (2020) (see Table B, Appendix A in (Adland et al., 2020)) that they achieve a MLR with R^2 equal to 0.824 for Product tankers and 0.873 for Suezmax. In our thesis, the MLR-method obtained a R^2 of 0.786 for Product tankers and 0.837 for Suezmax. We argue that the results are relatively similar, indicating consistent results. The minor difference can be due to dissimilar variables included and number of total observations, as our thesis' exclusionary constraints are looser for the latter.



(a) Product tankers

(b) Suezmax

Figure 6.1: Boxplot illustrating the distribution of MAE for each machine learning method

From Table 6.2, we observe that the GB method performs slightly better in terms of RMSE, MAE, R^2 and MAPE. Thus, outperforming the other methods. However, it must be pointed out that there is a relatively small difference between the methods with good performance. Figure 6.1a and Figure 6.1b further visualizes this, as we can observe how GB, XGB and EM, in terms of MAE, outperforms the rest of the methods. Thus, it can be argued that one also should consider other factors than solely error metrics. In terms of computational time, the EM is vastly inferior compared to XGB and GB due to its complexity. Furthermore, the parsimony principle states that when several methods have the same performance for a regression problem, one should use the simpler method (Seaholtz and Kowalski, 1993; Lepore et al., 2017). Subsequently, the GB method appears to be the better method for prediction of vessel fuel consumption. Thus, we find it necessary to evaluate the method's tuning further.

The final GB hyperparameters imply relationships and patterns in the samples. As exhibited in Table 6.1, the random search configures the GB hyperparameters of *mstop*, *max_depth* and *nu* equal to 250/150, 5 and 0.1, respectively. This corresponds well to empirically preferred boosting iterations between 100 and 500, tree depth in the range of 4 to 8 and a low shrinkage value to ensure slow learning (Friedman, 1999; Ridgeway, 2007; Hastie et al., 2009). These hyperparameters were also found to be optimal for other tunelengths and tunegrids, which favors the notion that random search can find better solutions than one might find blindly through use of a gridsearch. Alternatively, the hyperparameters and area searched could have been increased. However, the GB method's results are arguably good enough to be used for this thesis objective of estimating environmental performance, and the increase of searched area becomes a question of computational power versus the potential of either overfitting or a higher out-of-sample accuracy.

In summary, after our evaluation of machine learning methods, we find GB to be the most favorable alternative for prediction of vessel fuel consumption in our thesis. Next, in order to create a greater understanding of the methods ability to predict, we have exhibited different aspects of how the method performs for the out-of-sample on Product tankers in Figure 6.2 and Suezmax in Figure 6.3. Positive values indicates a under-prediction and negative values indicates a over-prediction from our GB method.

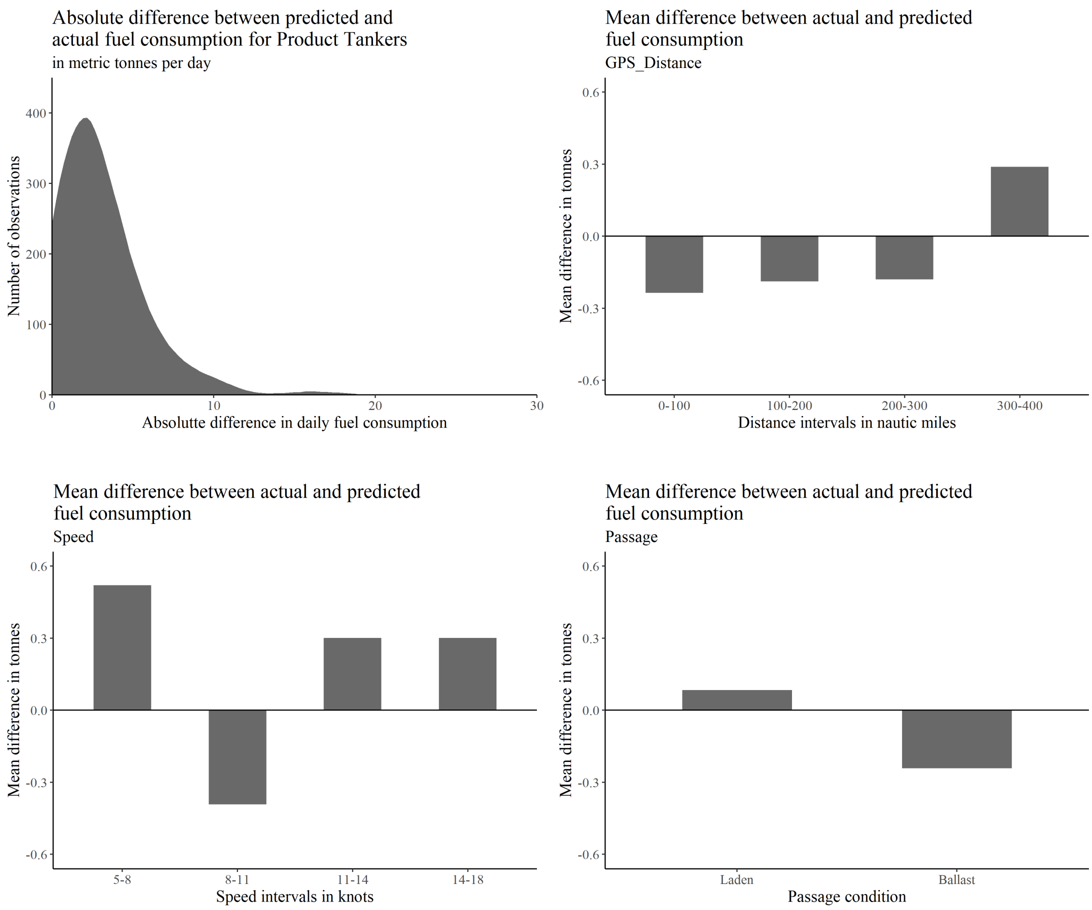


Figure 6.2: Illustration of the GB method's performance of predicting fuel consumption on unseen data in comparison to actual for Product tankers vessel type

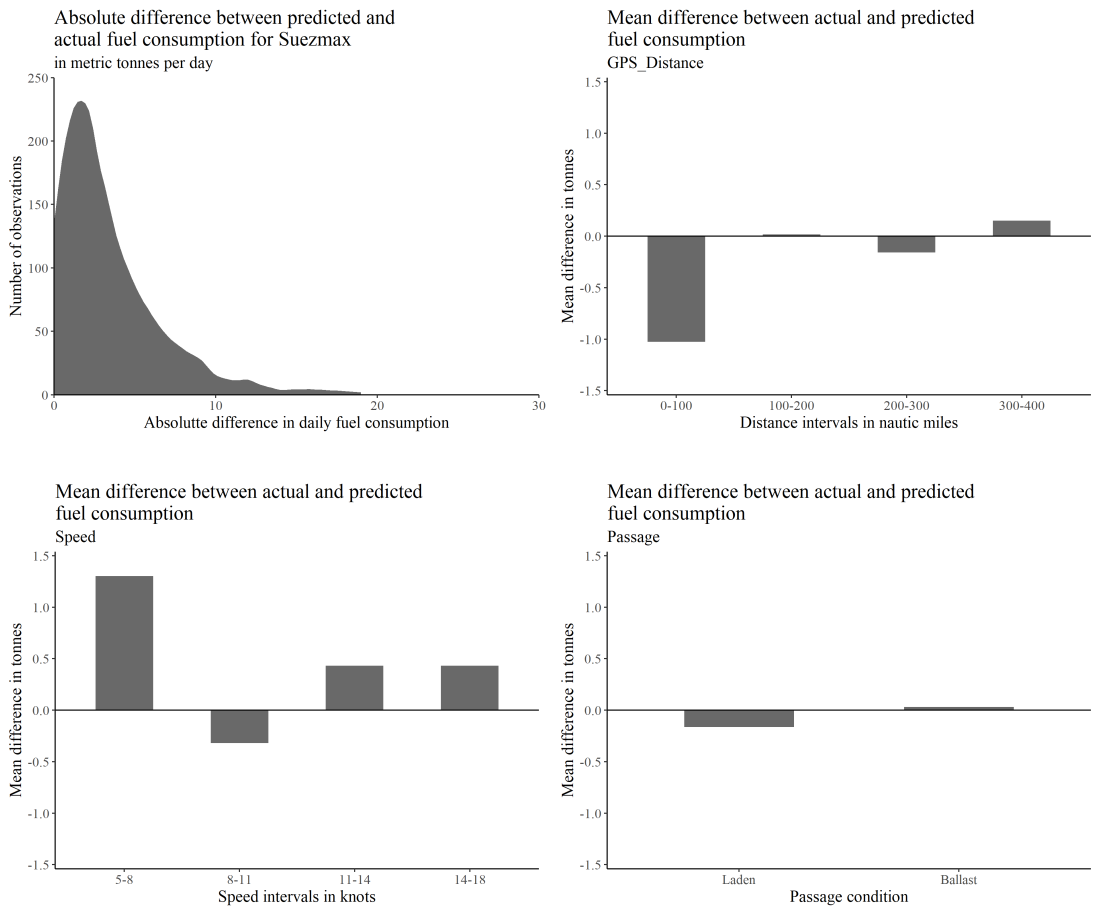


Figure 6.3: Illustration of the GB method's performance of predicting fuel consumption on unseen data in comparison to actual for Suezmax vessel type

The plots exhibit how the methods perform under different conditions in regard to GPS distance, Speed, and passage (whether vessel is laden); we can observe from plots on the GB method's variable importance in Section 6.3, that these are the three most important features. From Figure 6.2, we can observe that the majority of observations have an absolute difference close to zero. Taking the MAPE of 14.03% into consideration as well, this further supports our choice of prediction model. It is notable that the model on average is less accurate when travelling over longer distances, at lower speeds, or with ballast, all else being equal. However, the mean difference between the actual and predicted average fuel consumption is evaluated to be low with no means lesser or greater than -0.6/0.6 metric tonnes over the time period registered in our out-of-sample observations.

From Figure 6.3, we observe that the GB method presents similar results to those in Figure 6.2. The absolute difference for all observations in the out-of-sample is arguably close to zero. Furthermore, we see that under different conditions for GPS distance, Speed, and Passage, the mean difference is generally larger, being under/over -1/1 metric tonnes. This could be due to the fact that the total sample for Suezmax is smaller than that of Product tankers. Notably, we observe a substantial difference between the GB method's average ability to predict daily fuel consumption with regard to speed and distance, as the method under-predicts at low speeds, a substantial positive difference, and over-predicts at lower distances, a negative difference.

Until this point, we have illustrated our results for the unseen out-of-sample. Preferably, going forward with our approach, we would have used our prediction method on new forecasted operational data. This could be achieved by predicting the data obtained from our noon reports, using information regarding contractual aspects, vessel specifications, and weather forecasts. Thus, we could generate the same datasets of operational data applied in this approach.

Despite this possibility, we consider the forecasting of noon reports, and independent variables such as weather and sea data, as falling outside the scope of this thesis. Thus, the complete original samples have instead been used as a proxy. We argue that action is valid, because, based on a Welch t-test, we find that the means of the difference between prediction and true value for the out-of-sample and full sample sets are significantly similar. In other words, this enables us to predict on the basis of the complete original samples, without obtaining implausible results. The results of our statistical tests are shown in Table B2.1 in Appendix B. Hence, the GB's $finalmodel$ is used to predict new daily fuel consumption and thus our operational data.

6.2 Estimating vessel emissions, pollutant scores and pollutant weighting factor

Vessel emissions are estimated, as outlined in our methodology, through multiplication of predicted fuel and each emissions factor listed in Table 5.1. From our sample we note that the Product tankers only use high fuel oil (HFO) as fuel source, while the Suezmaxes use HFO, low sulphur high fuel oil (LSHFO) and marine gas oil (MGO). Hence, we take the assumption, based on our samples, that the proportion of HFO, LSHFO, and MGO used by the Suezmaxes' is 90.3%, 2.5%, and 7.2%, respectively. We further assume this proportion to be consistent throughout our calculations, even though we recognize that this ratio could change both due to the “IMO2020”-regulation and different emission control areas.

The following Table 6.3 presents the descriptive statistics for the estimated operational data of Product tankers and Suezmax.

Pollutant	CO_2	CH_4	N_2O	NO_x	CO	NMVOC	SO_x	PM
Unit	(tonne)	(kg)	(kg)	(kg)	(kg)	(kg)	(kg)	(kg)
Product tankers								
Mean	97.21	2.00	5.00	2412.00	89.00	99.00	1490.00	226.00
Median	104.00	1.67	5.85	2580.20	95.19	105.54	1159.00	241.97
Min	-8.42	-0.10	-0.50	-209.00	-8.00	-9.00	-129.00	-20.00
Max	202.14	3.00	11.00	5015.00	185.00	205.00	3098.00	470.00
Std. Dev.	34.49	1.00	2.00	856.00	32.00	35.00	529.00	80.00
Obs	9257	9257	9257	9257	9257	9257	9257	9257
Suezmax								
Mean	112.65	2.00	6.00	2731.00	102.00	112.00	1579.00	245.00
Median	116.10	1.85	6.52	2814.27	105.00	115.22	1627.50	252.67
Min	-1.01	-0.02	-0.10	-25.00	-1.00	-1.00	-14.00	-2.00
Max	230.85	4.00	13.00	5596.00	209.00	229.00	3236.00	502.00
Std. Dev.	45.16	1.00	3.00	1095.00	41.00	45.00	633.00	98.00
Obs	4841	4841	4841	4841	4841	4841	4841	4841

Table 6.3: Descriptive statistics of estimated emissions per vessel type

For the computation of a pollutant scores, we apply Figure 5.1 from our methodology section. An ideal data sample would contain information on the current mass of cargo onboard. This is not the case for our data, as only one of our datasets register this. Hence, we have made assumptions regarding the cargo transported. We find from our Suezmax dataset that the vessels’ average percentage of total SDWT when laden to be 79.5% of maximum. Thus, we have assumed this percentage for both vessel types when laden and ballast, due to lack of ballast-specific data. This assumption is supported by the GLEC Framework, as it argues that weight can be used as a cargo measure even though it has its limitations (Greene and Lewis, 2019). Therefore, in this case, the pollutant score is calculated by dividing vessel emissions by distance per noon report, both when ballast or laden, multiplied by cargo, exhibited in Table 6.4. Finally, vessel environmental performance is calculated by multiplying pollutant scores with GWP20, GWP100 and GWP500 as weighting factors.

Vessel type	Total SDWT	Average % of total SDWT	Cargo (in metric tonnes)
Product Tanker	120 000	79.50 %	95 400
Suezmax	159 000	79.50 %	126 405

Table 6.4: Calculation of *cargo* per vessel type

6.3 Discussion and evaluation of the method used

This subsection highlights, discusses and evaluates the results of our approach through the use of two specific examples. As exhibited in the following examples, we argue that our thesis enables an increased degree of decision making ability, as our tool enables the estimation and comparison of aspects of environmental sustainability.

6.3.1 Interpretation of results

Figure 6.4 illustrates the estimated environmental performance for average voyages by our two vessel types from José Terminal (Venezuela) to Houston (USA), with GWP100 as the pollutant weighting factor. As we do not have forecast data, as mentioned in Section 6.1, for this example we have applied the average of historical voyages for each vessel, but with identical weather conditions. This is not optimal, but arguably sufficient for illustrative purposes. Table B3.1 in Appendix B depicts the day-specific information. This information serves as the basis for our illustrations, and could hence be useful for upward aggregation over time.

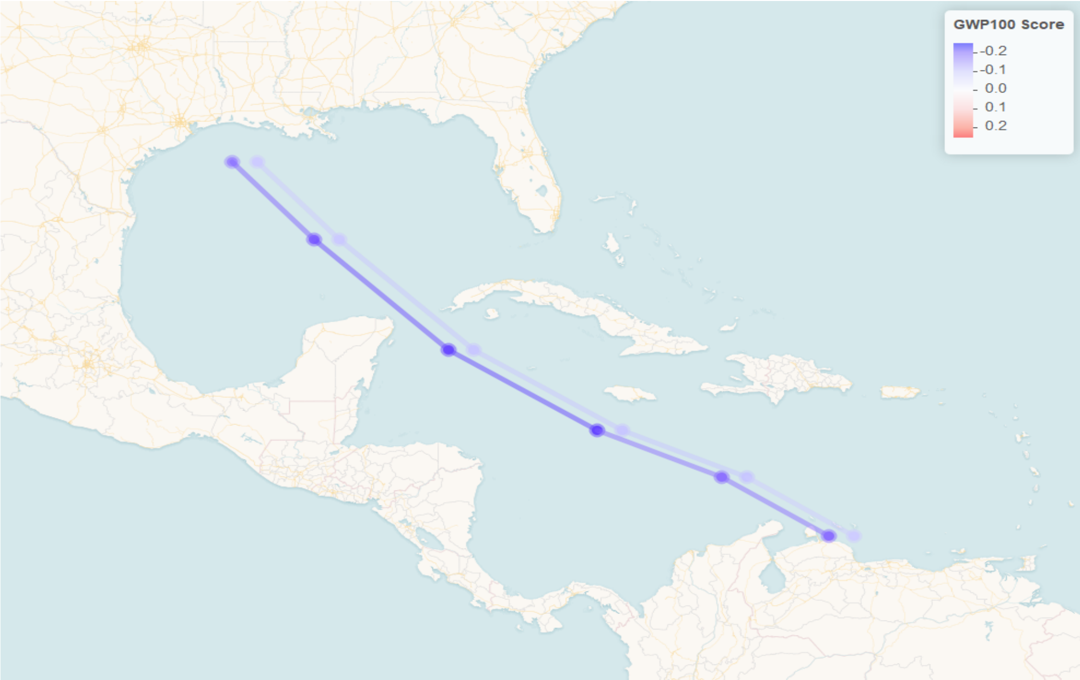


Figure 6.4: Example 1 - illustration of a voyage between José Terminal (Venezuela) and Houston (USA) for Suezmax (lower line) and Product tanker (upper line). Each dot is a estimated GWP100 score at given point of the voyage. The GWP score is multiplied by 100 000.

From Figure 6.4, we observe that an average voyage takes six days to complete, as each dot represents a noon report. Furthermore, we register that Product tankers (upper line) and Suezmax vessels (lower line) both obtain negative scores. Yet, from the ledger we observe that it is also possible to get positive values. When faced with these results and with comparison purposes in mind, decision makers must remember that our approach estimates performance through reconfiguring fuel to CO_2 equivalent figures. As a result, we can obtain both negative and positive numbers due to the different emissions temperature effect. In other words, the notion that the vessel exhaust gases emission has a cooling or warming effect for the given time horizon (Fuglestvedt et al., 2009). In the light of this observation, our thesis has visualized one of the more discussed parts of environmental sustainability within shipping: the overall climate effect of principal exhaust gas emissions from vessels.

The nature of this contribution to climate change is complex. In addition to warming the climate by emitting CO_2 emissions, emissions of sulphur (SO_x) cause short-term cooling through effects of atmospheric particles and clouds. At the same time, nitrogen oxides (NO_x) increases the level of the greenhouse gas ozone (O_3) and reduces the GHG methane

(CH_4), causing both warming and cooling, respectively. In our example, for both Product tankers and Suezmax vessels, we find that the short-term cooling effects from NO_x, SO_x and Organic carbon overpower the long-term warming effects of CO_2, CH_4, N_2O, CO, NMVOC and Black carbon. The vessel type Suezmax uses, in addition to HFO, "cleaner" fuels such as LSHFO and MGO with lower SO_x and NO_x content, which should indicate a warming effect compared with Product tankers. Nevertheless, as previously exhibited in Table 3.1, Suezmax vessels use on average more fuel than Product tankers and exhibit higher emissions per transport work in this example. This results in a higher SO_x and NO_x CO_2-equivalent emission per transport work and thus has a strong cooling effect.

Historically, the overall effect of international shipping has been found to be similar to our Suezmax vessels, strongly negative (Fuglestvedt et al., 2009). However, Fuglestvedt et al. (2009) argue that this will change in the following years due to new constricting regulations for SO_x and NO_x emissions. For example, the "IMO 2020"-regulation has the objective of limiting sulphur content in fuel from 3.5% to 0.5%. From the previous Table 5.2, we can observe that SO_x is one of the emissions that has a negative CO_2 equivalent impact. Hence, by legally limiting the amount of sulphur in fuel, total SO_x emissions are minimized, resulting in a reduction of the industry's cooling effect. Consequently, shipping as a transportation industry will, relative to its historical impact, impart a net "double warming" effect in the decades to come: one from CO_2, and one from the reduction of SO_x. On the one hand, one could argue that such regulations contradict international temperature goals and the IMO's strategy of being consistent with the Paris Agreement. On the other hand, SO_x and NO_x clearly have harmful effects on the environment beyond the climate. Thus, using "dirty" fuels in order to limit short-term climate change will never be a sensible measure (Fuglestvedt et al., 2009; Lund, 2020).

Figure 6.5 strengthens our understanding of the effect of such considerations, by visualizing how a choice of time horizon can affect decision makers' perception of operations. Table B4.1 in Appendix B exhibits the vessel day specifics. The plot depicts vessel environmental performance for a single Product tanker, transporting cargo over a longer voyage from Rotterdam in the Netherlands to Singapore. For the lower line, middle line and upper line, a time horizon of 20 years, 100 years and 500 years has been applied, respectively. From the example, we can observe how the vessel's environmental performance changes from negative to positive by a change in time horizon. From Table 5.2 we can observe this to happen because the cooling effect of NO_x, SO_x and organic carbon diminishes substantially with a change in time horizon from 20 to 500 years: by approximately 82.6%, 91.4%, and 91.2%, respectively.

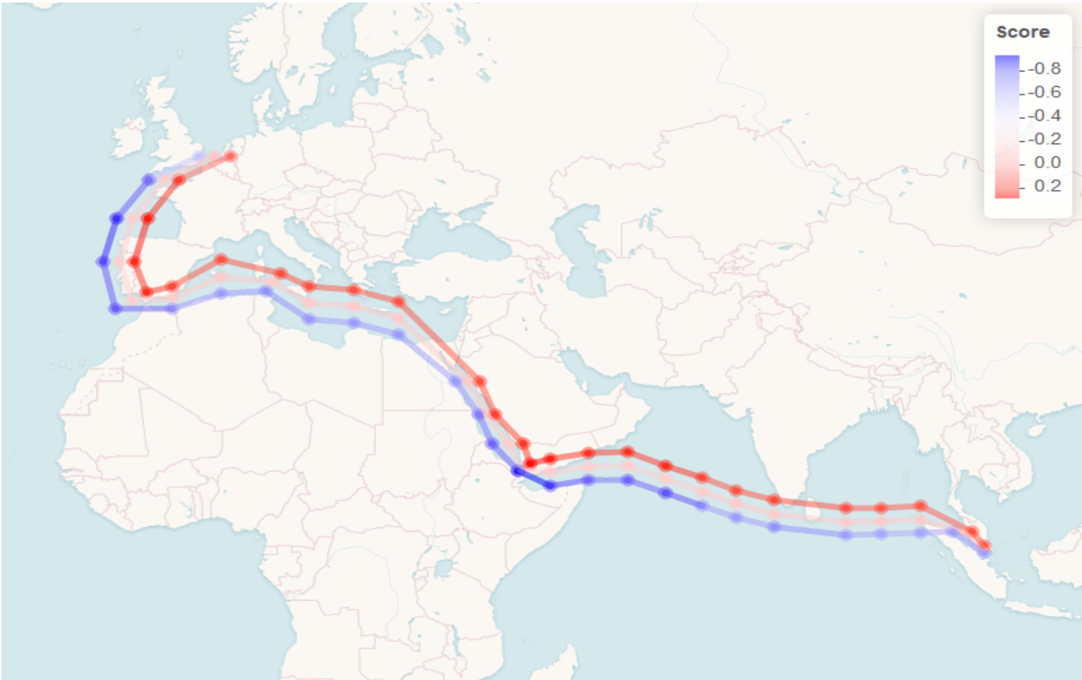
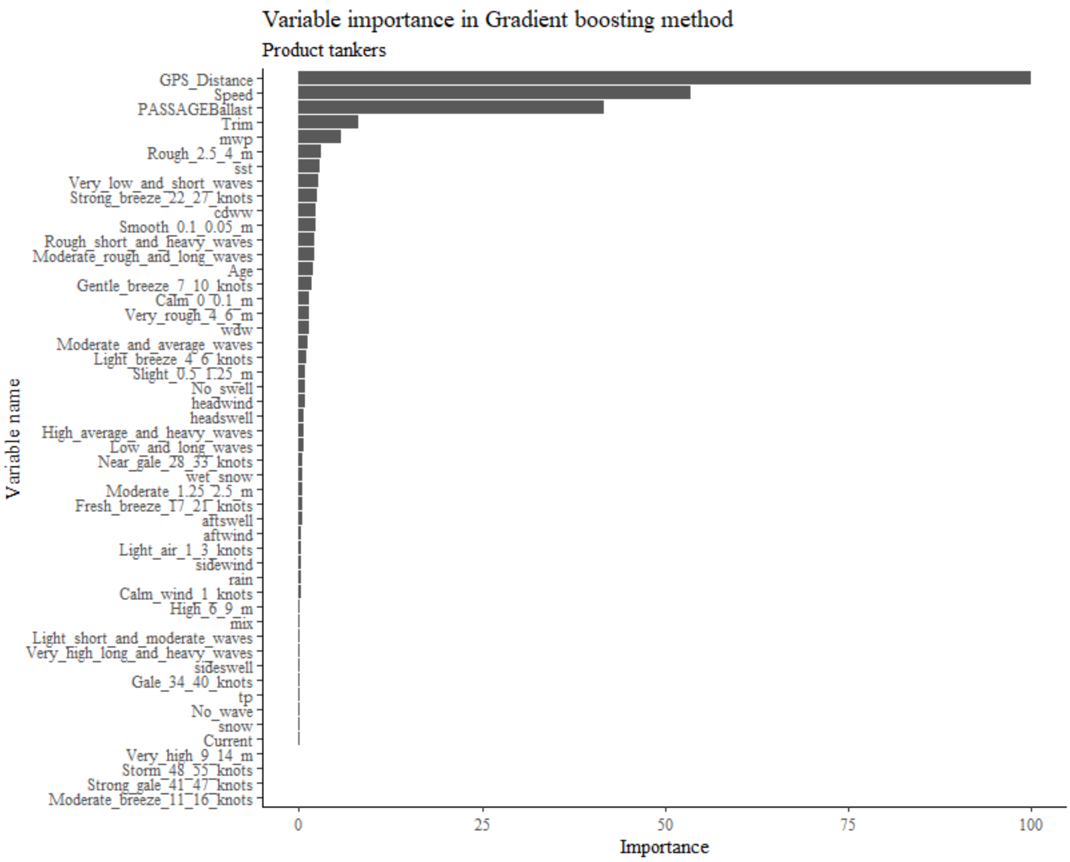


Figure 6.5: Example 2 - illustration of a voyage between Rotterdam and Singapore for a Product tanker with different GWP time horizons. The lower line (blue) is GWP20, the middle line (pink) is GWP100 and the upper line (red) is GWP500. Each dot is a estimated score at given point of the voyage. The GWP score is multiplied by 100 000.

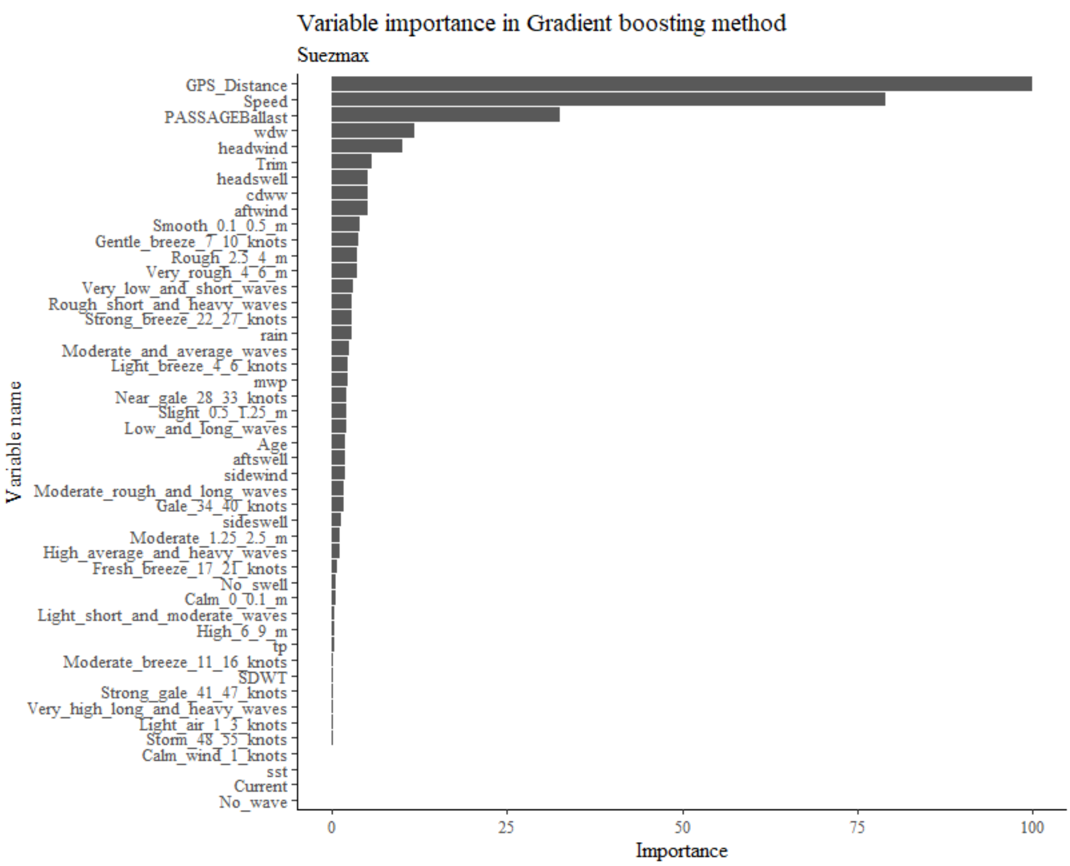
Even though GWP, as mentioned in section 5.4, strives to be an objective indicator, subjective decisions within the use of the metric can greatly affect decision makers' perception of operations when implementing it as a weighting factor (Gibson et al., 2019). Hence, the use of GWP as a metric has been strongly debated in the scientific literature (Rotmans and Den Elzen, 1992; Eckaus, 1992; Skodvin and Fuglestvedt, 1997; O'Neill, 2000; Smith and Wigley, 2000; Manne and Richels, 2001; Bradford, 2001; Fuglestvedt et al., 2003) and there are substantial differences within academia for calculation of the atmospheric effect of certain emissions (Fuglestvedt et al., 2003, 2010). Furthermore, there is no clear definition of what GWP is an indicator of, as regards which aspects of climate change it is a proxy for. Some studies have evaluated the equivalence of GWP-weighted emissions and underlined that it does not imply any equivalent temperature response except in certain idealized situations. As a valid alternative to GWP, one could have used the “global temperature change potential” (GTP), as this may be more appropriate in some circumstances. For example, in terms of temperature goals, the GTP-indicator measures the apparently resulting temperature change after the given time horizon of emissions (Fuglestvedt et al., 2010). However, the GTP metric is considered more technically demanding to calculate, as well as no more representative as a proxy than any other metric. Therefore, no alternative indicator has so far become as widely accepted as GWP.

By further inspection of each line in Figure 6.4 and Figure 6.5, as well as Table B3.1 and Table B4.1 in Appendix B, we can observe that the score changes during the different stages of the voyage. These changes are further visualised on an aggregated GWP100 level for all observations of Product tankers and Suezmax with Figure B5.1 in Appendix B. To what degree can decision makers interpret and explain the reason for these changes in estimated environmental performance? To answer this, we need to evaluate our method's design in relation to its explanatory ability.

In all simplicity, the results acquired by using our approach originate from our machine learning method of predicting fuel consumption. Thus, in terms of interpretive ability, this becomes a question of how much explanatory power we can extract from the assessments our GB method makes in relation to its input. As previously mentioned in Section 4.1, when trying to most accurately predict an output, we are not trying to establish the most accurate relation between our independent variables and the output variable. Hence, our ability to explain changes in score will not be as strong compared with other forms of regression which focus on inference. Yet, as mentioned in the literature review, there already exist numerous studies with the objective of explaining fuel consumption. In light of these, it can still be interesting to recognize which variables the GB algorithm places greatest emphasis on. The following Figure 6.6 exhibit our GB method's variable importance for each out-of-sample when the method was fitted.



(a) Product tankers



(b) Suezmax

Figure 6.6: Variable importance for our GB method

In accordance with our literature review, we observe from the tables that GPS distance traveled, speed, and whether the vessel is ballasted or laden (PassageBallast), to be the most important variables for the prediction of fuel consumption. Additionally, vessel trim and weather/sea conditions appear to have some influence, as suggested in other scientific studies. Subsequently, by closer inspection of our GB method's weighting of variables, decision makers may gain some insight from changes within scores.

At the same time, it is interesting to split these variables into categories and compare them against each other. For instance, we observe that for both vessel types, the most important variables are features the operator can manage in some way - such as distance travelled, speed, trim and the loading condition. The remaining features concerning weather are arguably out of operator's control. Furthermore, are the variables included from the Copernicus ERA5 dataset interesting to analyze more in depth. We observe that mean wave period, wave directional width, coefficient of drag with waves and sea surface temperature in general are the most influencing. However, there are great differences between the two vessel types and the importance of the respective features. For example, note the different importance of *sst* in Figure 6.6a and Figure 6.6b. Reasons for this could be due to matching between the data from Copernicus and our noon-report observations, or how GB as a BBM reacts differently to similar variables.

The latter part of our thesis' process reconfigures the results of our GB method through the use of other sources, assumptions, and studies. Changes in these could alter the vessel's environmental performance substantially. Regarding changes in this part of the approach, it is important to recognize how certain changes only alter the scale of scores, as all observations would be affected equally. For instance, by adjusting our method's use of emission coefficients and/or calculation of pollutant score, variations within each vessel's voyage as exhibited in Figure 6.4 and Figure 6.5 would be identical. Therefore, unless there are modifications in the approach's reconfiguration of fuel and transport work, decision makers will struggle to understand variations within performance, solely based on the environmental sustainability aspect of our approach. Hence, due to the nature of machine learning, our thesis's focus on prediction, and our approach to methodology design, decision makers cannot strongly interpret and explain changes in estimated vessel environmental performance.

6.3.2 Application of approach

Overall, we find our approach to be a valid tool for the estimation of vessel environmental performance, as illustrated in Figure 6.4 and Figure 6.5. Subsequently, three questions arise:

1. Why should one estimate figures through our suggested approach rather than using historic absolute numbers? 2. Why estimate the final figure of vessel environmental performance? 3. How can potential decision makers make use of this approach?

Firstly, as mentioned in the introduction of this thesis, the shipping sector is arguably more difficult to navigate and predict for decision makers now than before. Reasons include the increase of focus on sustainability and regulatory uncertainty from the EU. One could argue that these factors affect both internal and external parts of the shipping sector. Internally, there are mounting pressures on the IMO to establish more accurate estimates and regulations, and/or directly alter the operations of firms within shipping. Examples include changes in allowed emissions, fuel types or allocation of socio-economic costs from vessel exhaust gases. Externally, this pressure influences how investors and consumers relate to shipping companies, as well as the choices they make in connection with them. Sources to this pressure can be experienced through institutional grants and financial incentives, the sustainability requirements and mandates of private funds, recommendations from investment analysts, or the demands from consumers and companies. By applying estimated values rather than using historical figures, decision makers can potentially establish a greater understanding of the future in connection with their alternatives.

Secondly, with our approach, decision makers will estimate vessel environmental performance, a complex score dependent on several variables and assumptions. How one is to apply this score further, and consider if a result is good or bad, depends on several factors discussed in the paragraph below. However, this does not suggest that users must only utilize the final estimated scores. One could argue that in the current market it would be sufficient to stop earlier in the methodology and use the estimation of absolute values through vessel emissions, pollutant scores, or the multiplication of vessel emissions and pollutant weighting factors. Nonetheless, vessel environmental performance is a figure that portrays a relatively complex picture through an understandable and well-known format, CO_2-equivalence. Arguably, it is preferable for further use.

Lastly, in terms of how decision makers can potentially make use of this approach, it is important to note that the degree of usefulness will depend on whether the user is internal or external. Other considerations include their objective for using the method, and how much information they have available already. One the one hand, internal or external parties might use the tool as a supplement to current systems or to obtain completely new insights. On the other hand, organizations could, for example, use this data-driven approach as an alternative to estimate and extrapolate future emissions scenarios. Furthermore, we would argue that the estimation tool can be applied in both short- and long-term decision making.

From Figure 6.4 and Figure 6.5, we can observe that it is possible to estimate at a daily level how vessels will perform in terms of emissions and environmental performance. This is relevant for operators within the industry, given the notion that there are socio-economic costs associated with various emissions. A market for pricing of CO_2 emissions already exists in the form of the European Union's Emissions Trading System (ETS), and in terms of shipping, the discussion of inclusion has now changed from "if" to "when" (European Commission, 2020b,d). Thus, there will be a need for estimations of how short-term costs will depend on various emissions. For example, shipping operators accordingly need to plan for the demand and supply of quotas, for the taxation of emissions, or for contractual agreements with pricing of both estimated freight and environmental impact (Manne and Richels, 2001). They will arguably want to prepare for change, and thus strive for an increased understanding of their operations.

Furthermore, by aggregation of the values our method estimates, one can create a more accurate long-term "bottom-up" micromodel than one would do by solely relying on historical figures. For example, our GB method predicts the total amount of fuel used, versus that actually used, with an error margin of 0.061% and 0.21% over the whole period for Product tankers and Suezmax, respectively. Hence, decision makers can gain an increased understanding of how a company will perform in the future. We argue this would be relevant for both internal and external parties. For operators, given the notion that there are costs associated with emissions, they could better plan for the future with improved sensitivity, better cash flow and cost analyses, and more accurate estimations of emissions from transportation and purchased goods (GHG Scope 3). Moreover, a greater degree of accuracy to their estimations, organizations such as the IMO and EU, who do not have access to the same operational high-frequency data that operators have, could improve their scenarios of future shipping emissions, and thus be enabled to set more accurate regulations.

By use of this tool and the comparison of results, external parties could better compare and understand the future trends of shipping companies. This would inform their choices as customers, port authorities, investment analysts, insurance experts, investment funds or other stakeholders, based on the shipping company's environmental sustainability position in the market. Potentially, decision makers would prefer and incentivize vessels and companies that are predicted to perform well in the future, both economically and environmentally. This could operate similarly to how the proposed Just Transition Mechanism, an essential part of the EUs Green Deal, aims to (among other things) differentiate companies on their ability to change for the better in terms of sustainability (European Commission, 2020e). Regardless of their current business profile, if a company demonstrates through its actions how they can adapt their operations, the mechanism might be more likely to support it strategically or financially.

In summary, we find our approach to be a valid transparent and quantifiable tool for the estimation of vessel environmental performance, as illustrated in Figure 6.4 and Figure 6.5. However, due to the nature of machine learning and the thesis's focus on prediction, we cannot strongly explain the significance of each independent variable for the estimation of an environmental score in terms of design. Nevertheless, we believe we have identified several useful aspects of environmental sustainability, in addition to proposing why and how potential decision makers could make use of our estimation method.

7 Limitations and further research

7.1 Limitations

The limitations of the approach taken in this paper stem from the assumptions made throughout the process. Examples include our interpretation, processing, and use of noon reports. Newer and more accurate sensor data could provide more precise results and enable identification of other trends and relations. Moreover, we only considered one type of vessel, i.e. medium-size crude oil tankers. Another example originates from our machine learning choices: the chosen split of samples and seeds, the performance measures and methods used, as well as the number of methods and degree of resampling and hyperparameter tuning applied. It is also worth mentioning again the interpretability limitations with the use of BBMs.

7.2 Further research

We see three potential directions for further work on this topic: quantitative, qualitative, and a combination of the two. Quantitatively, the estimation approach should be extended through a forecast method with the objective of forecasting the input applied in our approach. In this way, one can effectively forecast and estimate future vessel environmental scores. Furthermore, keeping our limitations in mind, it would be interesting to know how the inclusion of more and other machine learning methods, independent of what other studies have found useful in the past, might affect the results. Likewise, how could the inclusion of more extensive and different vessel data, as well as an increase in hyperparameter tuning, influence the accuracy of predicted outputs? Although this would be more computationally demanding, it should in theory provide more precise results. Moreover, one could also study the addition of other explanatory variables, as there might exist examples, besides those included in this study, that increase our GB method's predictive power. For instance, one could apply AIS data and their objective variables to produce a more precise prediction compared with only noon report variables, as they do not depend on manual and subjective entry. Another approach could be to add variables related to hull cleaning, fouling, contractual specifics, types of fuel, brackish water, ocean depth, and ocean salinity (the Copernicus ERA5 database does not register the latter three).

Qualitatively, further work could involve mapping of whether this approach covers the market need our thesis has pursued to cover. This would require analysis of feedback from the implementation of this method amongst relevant decision makers over time. Furthermore, it would have been just as interesting to research to what extent environmental scores affect decisions, as stakeholders may often emphasize earning potential over other considerations.

At the intersection of quantitative and qualitative research, might it be interesting to discover whether this machine learning approach of transparently estimating environmental sustainability scores could be applied elsewhere? Examples include other industries, value chains, markets or forms of transportation. It could also be interesting to evaluate the use of environmental areas and the effect of the "IMO2020"-regulation in terms of shipping environmental sustainability.

8 Conclusion

This thesis aims to propose a machine learning approach for the estimation of vessel environmental performance. The theoretical framework established by Gibson et al. (2019) served as a foundation for our approach. Accordingly, we established a transparent and quantitative approach for estimating environmental sustainability in terms of vessel exhaust gas emissions. The results from the thesis give an indication of whether machine learning and predicted fuel consumption can be reconfigured to environmentally sustainable performance, and whether this provides decision makers with an increased ability to estimate and implement consequences of emissions in the decision-making process.

Our approach used a large sample of data from 16 medium-sized oil tankers over a comparatively long time period, with several variables describing both operational and weather features derived from noon reports and the Copernicus ERA5 database. With this data, we formulated a supervised regression problem to predict vessel fuel consumption, through regression methods that were found to provide the best results in the existing literature. Furthermore, we reconfigured the results from the best-performing method to give an estimation of vessel environmental performance. This measure is obtained through the use of the IMO's emission factors, pollutant scores inspired by the EEOI indicator and the GWP indicator as weighting factor.

Our results show that the Gradient boosting method is the best performing means of predicting vessel fuel consumption. We find that the method accurately predicts fuel consumption on unseen data, and is therefore favorable for our approach as it emphasizes the validity of estimated vessel environmental performance. However, while we find eXtreme gradient boosting and our Ensemble method to provide accurate results, these come at a higher cost in terms of complexity and computational power.

Given our estimated vessel performance, we note that the choice of weighting factor and associated time horizon is decisive for the outcome, while at the same time being a subjective assessment based on preferences and objectives. From our results, we observe how these choices yield different environmental sustainability impacts of vessel exhaust gases, in terms of warming or cooling global warming potential in comparison with CO_2.

In conclusion, we find our approach of estimating vessel environmental performance to be valid. Our approach uses predicted operational data and the best available measurement of emissions impact to portray the complexity of environmental sustainability in a relatively understandable manner. To the best of our knowledge, this is the first study that successfully uses machine learning and estimates shipping environmental sustainability in light of vessel exhaust gases. We believe that our results and assessments demonstrate the value that empirical modeling, in the form of machine learning, can provide to internal and external decision makers who compute and apply emission estimates, both in the short and long term.

We hope that this thesis motivates further research that can promote the applicability of machine learning for estimating sustainability within shipping.

References

Abebe, M., Shin, Y., Noh, Y., Lee, S., and Lee, I. (2020). Machine learning approaches for ship speed prediction towards energy efficient shipping. *Applied Sciences (Switzerland)*, 10(7).

Adland, R., Cariou, P., Jia, H., and Wolff, F. C. (2018). The energy efficiency effects of periodic ship hull cleaning. *Journal of Cleaner Production*, 178:1–13.

Adland, R., Cariou, P., and Wolff, F. C. (2020). Optimal ship speed and the cubic law revisited: Empirical evidence from an oil tanker fleet. *Transportation Research Part E: Logistics and Transportation Review*, 140(October 2019):101972.

Adland, R. and Jia, H. (2018). Dynamic speed choice in bulk shipping. *Maritime Economics and Logistics*, 20(2):253–266.

Adland, R. O. and Jia, H. (2016). Vessel speed analytics using satellite-based ship position data. *IEEE International Conference on Industrial Engineering and Engineering Management*, 2016-Decem:1299–1303.

Aldous, L., Smith, T., and Bucknall, R. (2013). Noon report Data Uncertainty. *Low Carbon Shipping Conference Londen 2013*, page 13.

Alujevic, N., Catipovic, I., Malenica, S., Senjanovic, I., and Vladimir, N. (2019). Ship roll control and energy harvesting using a U-tube anti-roll tank. *Ocean Engineering*, 172:857–870.

Assaf, A., Tsionas, M., and Tasiopolous, A. (2019). Diagnosing and correcting the effects of multicollinearity: Bayesian implications of ridge regression. *Tourism Management*, 71:1–8.

Bal Besikci, E., Arslan, O., Turan, O., and Olcer, A. (2016). An artificial neural network based decision support system for energy efficient ship operations. *Computers & Operations Research*, 66:393–401.

Balcombe, P., Brierley, J., Lewis, C., Skatvedt, L., Speirs, J., Hawkes, A., and Stafell, I. (2019). How to decarbonise international shipping: Options for fuels, technologies and policies. *Energy Conversion and Management*, 182:72–88.

Ballou, P., Chen, H., and Horner, J. (2008). Advanced Methods of Optimizing Ship Operations to Reduce Emissions Detrimental to Climate Change. In *Oceans 08 Conference*.

Balmat, J., Lafont, F., Maifret, R., and Pessel, N. (2011). A decision-making system to maritime risk assessment. *Ocean Engineering*, 38(1):171–176.

Bellman, R. (1961). Adaptive Control Processes: A Guided Tour. *Princeton Legacy Library*.

Bengtsson, S., Andersson, K., and Fridell, E. (2011). A comparative life cycle assessment of marine fuels liquefied natural gas and three other fossil fuels. *Proceedings of the Institution of Mechanical Engineers, Part M: Journal of Engineering for the Maritime Environment*, 225(2):97–110.

Bergstra, J. and Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13:281–305.

Bocchetti, D., Lepore, A., Palumbo, B., and Vitiello, L. (2013). A statistical control of the ship fuel consumption. *RINA, Royal Institution of Naval Architects - Design and Operation of Passenger Ships 2013*, pages 87–92.

Bocchetti, D., Lepore, A., Palumbo, B., and Vitiello, L. (2015). A statistical approach to ship fuel consumption monitoring. *Journal of Ship Research*, 59(3):162–171.

Botchkarev, A. (2018). Performance Metrics (Error Measures) in Machine Learning Regression, Forecasting and Prognostics: Properties and Typology. *arXiv:1809.03006*.

Bradford, D. F. (2001). Time, money and tradeoffs. *Nature*, 410(6829):649–650.

Breiman, L. (1996). Stacked regressions. *Machine Learning*, 24:49–64.

Breiman, L. (2001). Statistical Modeling: The Two Cultures. *Statistical Science*, 16(3):199–231.

Brouer, B., Dirksen, J., Psinger, D., Plum, C., and Vaaben, B. (2013). The Vessel Schedule Recovery Problem (VSRP) – a MIP model for handling disruptions in liner shipping. *European Journal of Operational Research*, 224(2):362–274.

Bui-Duy, L. and Vu-Thi-Minh, N. (2020). Utilization of a deep learning-based fuel consumption model in choosing a liner shipping route for container ships in Asia. *Asian Journal of Shipping and Logistics*.

Bzdok, D., Altman, N., and Krzywinski, M. (2018). Points of Significance: Statistics versus machine learning. *Nature Methods*, 15(4):233–234.

Cariou, P. (2011). Is slow steaming a sustainable means of reducing CO2 emissions from container shipping? *Maritime Policy & Management*, 16(3):260–264.

Chai, T. and Draxler, R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, 7(3):1247–1250.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv:1406.1078*, 3.

Coraddu, A., Oneto, L., Baldi, F., and Anguita, D. (2018). Vessels fuel consumption: a data analytics perspective to sustainability. *Soft Computing for Sustainability Science*, 11(48).

Council of the European Union (2015). Regulation (Eu) 2015/757 of the European Parliament. *Official Journal of the European Union*, L(April):55–76. Retrieved from: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32015R0757&from=EL.

Divina, F., Gilson, A., Goméz-Vela, F., Torres, M. G., and Torres, J. F. (2018). Stacking ensemble learning for short-term electricity consumption forecasting. *Energies*, 11(4):1–31.

Doulgeris, G., Korakianitis, T., Pilidis, P., and Tsoudis, E. (2012). Techno-economic and environmental risk analysis for advanced marine propulsion systems. *Applied Energy*, 99:1–12.

Du, Y., Meng, Q., Wang, S., and Kuang, H. (2019). Two-phase optimal solutions for ship speed and trim optimization over a voyage using voyage report data. *Transportation Research Part B: Methodological*, 122:88–114.

Eckaus, R. (1992). Comparing the Effects of Greenhaus Gas Emissions on Global Warming. *The Energy Journal*, 13(1):25–36.

European Commission (2007). Limiting Global Climate Change to 2 degrees Celsius. (January). Retrieved from: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX%3A52007DC0002&from=EN.

European Commission (2019). The European Green Deal. *European Commission*, 53(9):24. Retrieved from: https://eur-lex.europa.eu/resource.html?uri=cellar%3Ab828d165-1c22-11ea-8c1f-01aa75ed71a1.0002.02%2FDOC_1&format=PDF.

European Commission (2020a). 2019 annual report on co2 emissions from maritime transport, brussel 2020. page 77. Retrieved from: https://ec.europa.eu/clima/sites/clima/files/transport/shipping/docs/swd_2020_82_en.pdf.

European Commission (2020b). Addressing shipping Emission in the International Maritime Organization. Retrieved from: https://ec.europa.eu/commission/commissioners/2019-2024/valean/announcements/addressing-shipping-emission-international-maritime-organization_en.

European Commission (2020c). CO2 emissions from shipping – encouraging the use of low-carbon fuels. Retrieved from: https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12312-FuelEU-Maritime-.

European Commission (2020d). EU Emissions Trading System (EU ETS). Retrieved from: https://ec.europa.eu/clima/policies/ets_en#tab-0-0.

European Commission (2020e). The Just Transition Mechanism: making sure no one is left behind. Retrieved from: https://ec.europa.eu/info/strategy/priorities-2019-2024/european-green-deal/actions-being-taken-eu/just-transition-mechanism_en.

Fagerholt, K., Johnsen, T., and Lindstad, H. (2009). Fleet deployment in liner shipping: a case study. *Maritime Policy & Management*, 36(5):397–409.

Ford, C. (2015). Is R-squared Useless? Retrieved from: https://data.library.virginia.edu/is-r-squared-useless/.

Foss, T., Stensrud, E., Kitchenham, E., and Myrtveit, I. (2003). A simulation study of the model evaluation criterion MMRE. *IEEE Transactions on Software Engineering*, 23(11).

Fridell, E., Sköld, S., and Bäckström, S. (2018). Transport work and emissions in MRV; methods and potential use of data. (April):32. Retrieved from: https://www.lighthouse.nu/sites/www.lighthouse.nu/files/mrv_webb.pdf.

Friedman, J. (1999). Greedy Function Approximation: A Gradident Boosting Machine.

Friedman, J. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4):367–378.

Fuglestvedt, J., Berntsen, T., Eyring, V., Isaksen, I., Lee, D. S., and Sausen, R. (2009). Shipping emissions: From cooling to warming of climates - and reducing impacts on health. *Environmental Science and Technology*, 43(24):9057–9062.

Fuglestvedt, J. S., Berntsen, T. K., Godal, O., Sausen, R., Shine, K. P., and Skodvin, T. (2003). Metrics of climate change: Assessing radiative forcing and emission indices. *Climatic Change*, 58(3):267–331.

Fuglestvedt, J. S., Shine, K. P., Berntsen, T., Cook, J., Lee, D. S., Stenke, A., Skeie, R. B., Velders, G. J., and Waitz, I. A. (2010). Transport impacts on atmosphere and climate: Metrics. *Atmospheric Environment*, 44(37):4648–4677.

Gibson, M., Murphy, A. J., and Pazouki, K. (2019). Evaluation of environmental performance indices for ships. *Transportation Research Part D: Transport and Environment*, 73(July):152–161.

Goodwin, P. and Lawton, R. (1999). On the asymmetry of the symmetric MAPE. *International Journal of Forecasting*, pages 405–408.

Greene, S., Jia, H., and Rubio-Domingo, G. (2020). Well-to-tank carbon emissions from crude oil maritime transportation. *Transportation Research Part D: Transport and Environment*, 88(October):102587.

Greene, S. and Lewis, A. (2019). Global Logistics Emissions Council Framework for Logistics Emissions Accounting and Reporting v2.0. Technical report, Smart Freight Center.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction (Second edition)*. Springer.

Hastie, T. and Zou, H. (2005). Regularization and variable selection via the elastic net. *Journal of the American Statistical Association*, 67(2):301–320.

He, Q., Zhang, X., and Nip, K. (2017). Speed optimization over a path with heterogeneous arc costs. *Transportation Research Part B: Methodological*, 104:198–214.

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Dee, D., and Thépaut, J.-N. (2018). ERA5 hourly data on single level from 1979 to present.

Hinton, G. (2010). A practical guide to training restricted Boltzmann machines. *Neural Networks: Tricks of the Trade*, pages 599–619.

Holmes, S. (2000). RMS Error.

IMO (2009). Guidelines for voluntary use of the Ship Energy Efficiency Operational Indicator (EEOI). Retrieved from: https://wwwcdn.imo.org/localresources/en/OurWork/Environment/Documents/Circ-684.pdf.

IMO (2020a). IMO Environment Committee approves amendments to cut ship emissions. Retrieved from: https://www.imo.org/en/MediaCentre/PressBriefings/pages/42-MEPC-short-term-measure.aspx.

IMO (2020b). Reduction of GHG emissions from ships. Fourth IMO GHG Study 2020. *International Maritime Organization.*

Ionescu, R., Szava, I., Vlase, S., Ivanoiu, M., and Munteanu, R. (2015). Innovative Solutions for Portable Wind Turbines, Used on Ships. *Procedia Technology*, 19:722–729.

Islam, H. and Soares, G. (2019). Effect of trim on container ship resistance at different ship speeds and drafts. *Ocean Engineering*, 183:106–115.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R.* Springer.

Jeon, M., Noh, Y., Shin, Y., Lim, O., Lee, I., and Cho, D. (2018). Prediction of ship fuel consumption by using an artificial neural network. *Journal of Mechanical Science and Technology*, 32:5785–5796.

Jia, H., Adland, R., Prakash, V., and Smith, T. (2017). Energy efficiency with the application of Virtual Arrival policy. *Transportation Research Part D: Transport and Environment*, 54(July 2017):50–60.

Johnson, H. and Styhre, L. (2015). Increased energy efficiency in short sea shipping through decreased time in port. *Transportation Research Part A: Policy and Practice*, 71:167–178.

Jung, J., Patnam, M., and Ter-Martirosyan, A. (2018). An Algorithmic Crystal Ball: Forecasts-based on Machine Learning. *IMF Working Paper*, 18(230).

Kim, J., Kim, H., and Lee, P. (2014). Optimizing ship speed to minimize fuel consumption. *Transportation Letters The International Journal of Transportation Research*, 6(3):109–117.

Kim, S. and Kim, H. (2016). A new metric of absolute percentage error for intermittent demand forecasts. *International Journal of Forecasting*, 32(3):669–679.

Kristensen, H. (2019). Evaluation of Different Measures for Reduction of Green-house gas (GHG) Emissions for Different Ship Types.

Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software*, 28(5):1–26.

Lack, D. A., Corbett, J. J., Onasch, T., Lerner, B., Massoli, P., Quinn, P. K., Bates, T. S., Covert, D. S., Coffman, D., Sierau, B., Herndon, S., Allan, J., Baynard, T., Lovejoy, E., Ravishankara, A. R., and Williams, E. (2009). Particulate emissions from commercial shipping: Chemical, physical, and optical properties. *Journal of Geophysical Research Atmospheres*, 114(4):1–16.

Lam, J. (2010). An integrated approach for port selection, ship scheduling and financial analysis. *NETNOMICS: Economic Research and Electronic Networking*, 11(1):33–46.

Larochelle, H., Erhan, D., Courville, A., Bergstra, J., and Bengio, Y. (2007). An empirical evaluation of deep architectures on problems with many factors of variation. In *Proceedings of the Twenty-fourth International Conference on Machine Learning (ICML'07)*, pages 473–480.

LeCun, Y., Bottou, L., Orr, G., and Muller, K. (1998). Efficient backprop. *Neural Networks: Tricks of the Trade*, pages 9–48.

Leifsson, L. T., Sævarsdóttir, H., Sigurdsson, S. T., and Vésteinsson, A. (2008). Grey-box modeling of an ocean vessel for operational optimization. *Simulation Modelling Practice and Theory*, 16(8):923–932.

Lepore, A., dos Reis, M. S., Palumbo, B., Rendall, R., and Capezza, C. (2017). A comparison of advanced regression techniques for predicting ship CO2 emissions. *Quality and Reliability Engineering International*, 33(6):1281–1292.

Lewis, E. (1988). *Principles of Naval Architecture, vol 2*. The Society of Naval Achitects and Marine Engineers.

Lewis-Beck, M. and Skalaban, A. (1990). The R-Squared: Some Straight Talk. *Political Analysis*, 2:153–171.

Li, J. (2017). Assessing the accuracy of predictive models for numerical data: Not r nor r2, why not? Then what? *PloS one*, 12(8).

Lindstad, H., Asbjørnslett, B., and Jullumstrø, E. (2013). Assesment of profit, cost and emissions by varying speed as a funtion of sea conditions and freight market. *Transport Reviews*, Part D 19:5–12.

Lister, J., Poulsen, R., and Ponte, S. (2015). Orchestrating transnational environmental governance in maritime shipping. *Global Environ. Change*, 34:185–195.

Lloyd, A. (1989). *Seakeeping, Ship Behaviour in Rouch Weather*. Ellis Horwood.

Lo, H. and McCord, M. (1995). Routing through dynamic ocean currents: General heuristics and empirical results in the gulf stream region. *Transportation Research Part B: Methodological*, 29(2):109–124.

Lund, M. T. (2020). personal communication. https://cicero.oslo.no/en/employee/45/marianne-tronstad-lund.

Manne, A. and Richels, R. (2001). An alternative approach to establishing trade-offs among greenhouse gases. *Nature*, 410:675–677.

Mansouri, S., Lee, A., and Aluko, O. (2015). Multi-objective decision support to enhance environmental sustainability in maritime shipping: A review and future directions. *Transportation Research Part E: Logistics and Transportation Review*, (78):3–18.

Mathews, B. and Diamantopoulos, A. (1994). Towards a taxonomy of forecast error measures: a factor-comparative investigation of forecast error dimensions. *Journal of Forecasting*, 13:409–416.

Meng, Q., Du, Y., and Wang, Y. (2016). Shipping log data based contained ship fuel efficiency modeling. *Transportation Research Part B: Methodological*, 83:207–229.

Moldan, B., Janouskova, S., and Hak, T. (2012). How to understand and measure environmental sustainability: Indicators and targets. *Ecological Indicators*, 17:4–13.

Molinaro, A., Simon, R., and Pfeiffer, R. (2005). Prediction error estimation: a comparison of resampling methods. *Bioinformatics*, 21(15):3301–3307.

Motley, M., Nelson, M., and Young, Y. (2012). Integrated probabilistic design of marine propulsors to minimize lifetime fuel consumption. *Ocean Engineering*, 45:1–8.

Murphy, A., Landamore, M., Pazouki, K., and Gibson, M. (2013). Modelling ship emisson factors and emission indices. In *Low carbon shipping conference*.

Neal, R. (1996). *Bayesian Learning for Neural Networks*. Springer.

O'Neill, B. (2000). The Jury is Still Out on Global Warming Potentials. *Climate Change*, 44:427–443.

Pedersen, B. and Larsen, J. (2009). Prediction of Full-Scale Propulsion Power using Artifical Neural Netwoorks. In *COMPIT '09 : 8th International Conference on Computer and IT Applications in the Maritime Industries*, pages 537–550.

Psaraftis, H. N. and Kontovas, C. A. (2013). Speed models for energy-efficient maritime transportation: A taxonomy and survey. *Transportation Research Part C: Emerging Technologies*, 26:331–351.

Rehmatulla, N., Parker, S., Smith, T., and Stulgis, V. (2017). Wind technologies: Opportunities and barriers to a low carbon shipping industry. *Marine Policy*, 75:217–226.

Ridgeway, G. (2007). Generalized Boosted Models: A guide to the gbm package. *Compute*, 1(4):1–12.

Rotmans, J. and Den Elzen, M. (1992). A model-based approach to the calculation of global warming potentials (GWPs). *International Journal of Climatology*, 12:865–874.

Russell, S., Norvig, P., and Davis, E. (2010). *Artificial intelligence: a modern approach (third edition)*. Prentice Hall.

Seaholtz, M. and Kowalski, B. (1993). The parsimony principle applied to multivariate calibration. *Anal Chim Acta*, 277(2):165–177.

Sheng, X., Lee, L., and Chew, E. (2014). Dynamic determination of vessel speed and selection of bunkering ports for liner shipping under stochastic environment. *OR Spectrum*, 36:455–480.

Shih, N., Weng, B., Lee, J., and Hsiao, Y. (2014). Development of a 20 kW generic hybrid fuel cell power system for small ships and underwater vehicles. *International Journal of Hydrogen Energy*, 39(25):13894–13901.

Skodvin, T. and Fuglestvedt, J. S. (1997). A Comprehensive Approach to Climate Change: Political and Scientific Considerations. *Ambio*, 26(4):351–358.

Smith, S. J. and Wigley, M. L. (2000). Global Warming Potentials: 1. Climatic implications of emissions reductions. *Climatic Change*, 44(4):445–457.

Smith, T. W. P., Jalkanen, J. P., Anderson, B. A., Corbett, J. J., Faber, J., Hanayama, S., O'Keeffe, E., Parker, S., Johansson, L., Aldous, L., Raucci, C., Traut, M., Ettinger, S., Nelissen, D., Lee, D. S., Ng, S., Agrawal, A., Winebrake, J. J., and Hoen, M., A. (2014). Third IMO Greenhouse Gas Study 2014. *International Maritime Organization (IMO)*, page 327.

Stopford, M. (2008). *Maritime economics (Third edition*. Routledge.

Swamidass, P. (2000). MEAN ABSOLUTE PERCENTAGE ERROR (MAPE). Retrieved from: https://link.springer.com/referenceworkentry/10.1007%2F1-4020-0612-8_580.

Telfer, E. (1926). *The Practical Analysis of Merchant Ship Trials and Service Performance*. North East Coast Institution of Engineers and Shipbuilders.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. *Journal of the Royal Statistical Society. Series B*, 58(1):267–288.

Todd, F. (1967). *Principles of naval architecture. Chapter VIII, Resistance and Propulsio*. Society of Naval Architecture and Marine Engineers.

UNFCCC (2018). Report of the Conference of the Parties serving as the meeting of the Parties to the Paris Agreement on the third part of its first session, held in Katowice from 2 to 15 December 2018. Retrieved from: https://unfccc.int/sites/default/files/resource/cma2018_3_add2_new_advance.pdf.

Uyanık, T., Karatuğ, Ç., and Arslanoğlu, Y. (2020). Machine learning approach to ship fuel consumption: A case of container vessel. *Transportation Research Part D: Transport and Environment*, 84(May).

Wang, G., Hao, J., Ma, J., and Jiang, H. (2011). A comparative assessment of ensemble learning for credit scoring. *Expert Systems with Applications*, 38(1):223–230.

Wang, H., Oguz, E., Jeong, B., and Zhou, P. (2019a). Life cycle and economic assessment of a solar panel array applied to a short route ferry. *Journal of Cleaner Production*, 219:471–484.

Wang, K., Yan, X., Yuan, Y., Jiang, X., Lin, X., and Negenborn, R. R. (2018a). Dynamic optimization of ship energy efficiency considering time-varying environmental factors. *Transportation Research Part D: Transport and Environment*, 62(May):685–698.

Wang, S., Ji, B., Zhao, J., Liu, W., and Xu, T. (2018b). Predicting ship fuel consumption based on LASSO regression. *Transportation Research Part D: Transport and Environment*, 65(October 2017):817–824.

Wang, S. and Meng, Q. (2012). Sailing speed optimization for container ships in a liner shipping network. *Transportation Research Part E: Logistics and Transportation Review*, 48(3):701–714.

Wang, Y., Wang, D., Geng, N., Wang, Y., Yin, Y., and Jin, Y. (2019b). Stacking-based ensemble learning of decision trees for interpretable prostate cancer detection. *Applied Soft Computing*, 77:188–204.

Willmott, C., Matsuura, K., and Robeson, S. (2009). Ambiguities inherent in sums-of-squares-based error statistics. *Atsompheric Environment*, 43(3):749–752.

Windeck, V. and Stadtler, H. (2011). A liner shipping network design–routing and scheduling impacted by environmental influences. In *INOC 2011: Network Optimization*, pages 574–576.

Witten, H. and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques. (2nd edition)*. Morgan Kaufmann.

Wolpert, D. (1992). Stacked generalization. *Neural Networks*, 5(2):241–259.

Woo, J. and Moon, D. (2014). The effects of slow steaming on the environmental performance in liner shipping. *Maritime Policy & Management2*, 41(3):176–191.

Wu, Z. and Xia, X. (2018). Tariff-driven demand side management of green shipping. *Solar Energy*, 170:991–1000.

Xia, J., Li, K., Ma, H., and Xu, Z. (2015). Joint planning of fleet deployment, speed optimization, and cargo allocation for liner shipping. *Transportation Science*, 49(4):922–938.

Yang, D., Wu, L., Wang, S., Jia, H., and Li, K. X. (2019a). How big data enriches maritime research–a critical review of Automatic Identification System (AIS) data applications. *Transport Reviews*, 39(6):755–773.

Yang, L., Chen, G., Rytter, N. G. M., Zhao, J., and Yang, D. (2019b). A genetic algorithm-based grey-box model for ship fuel consumption prediction towards sustainable shipping. *Annals of Operations Research*.

Yao, Z., Ng, S., and Lee, L. (2012). A study on bunker fuel management for the shipping liner services. *Computers & Operations Research*, 39(5):1160–1172.

Yin, J., Fan, L., Yang, Z., and Li, K. (2014). Slow steaming of liner trade: its economic and environmental impacts. *Maritime Policy & Management,*, 41(2):149–158.

Yu, W., Zhou, P., and Wang, H. (2018). Evaluation on the energy efficiency and emissions reduction of a short-route hybrid sightseeing ship. *Ocean Engineering*, 162:34–42.

Zhou, Z. (2012). Ensemble Methods: Foundations and Algorithms. *CRC Press*.

Appendix A

A1 Analysis of internal versus external weather

A1.1 Product tankers

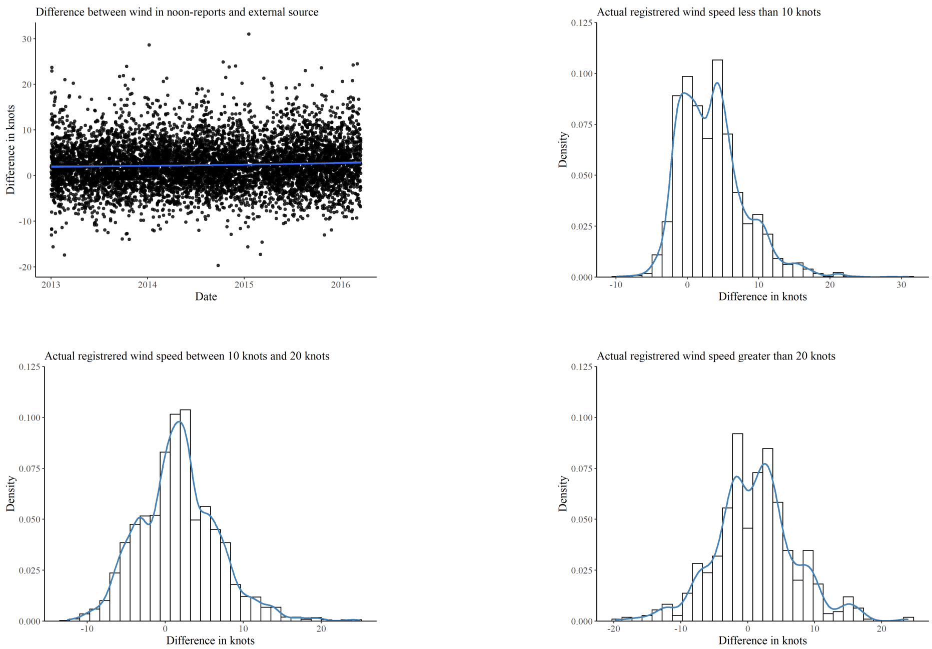


Figure A1.1: Wind analysis for Product tankers

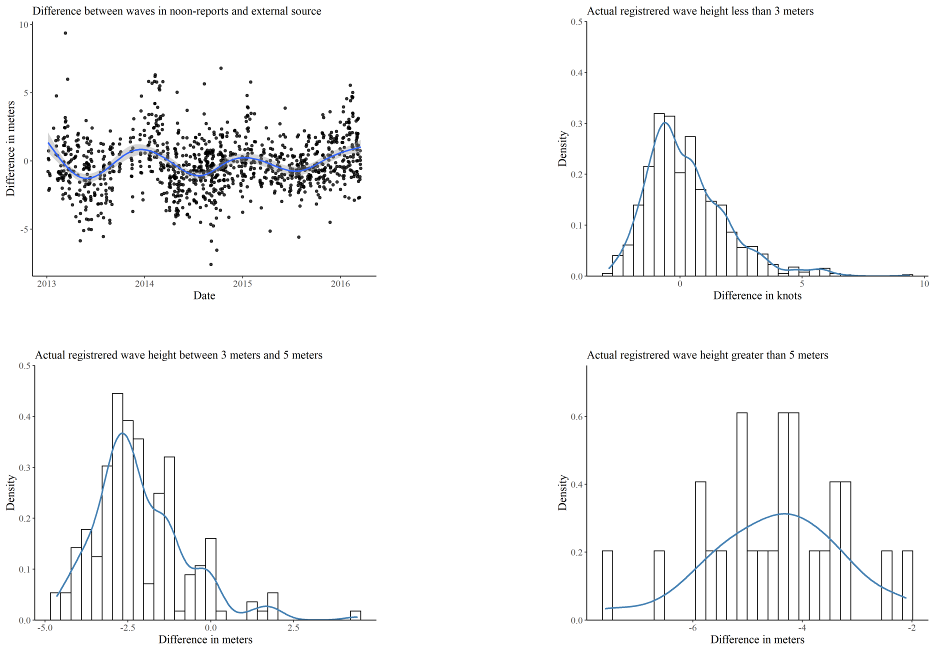


Figure A1.2: Wave analysis for Product tankers

A1.2 Suezmax

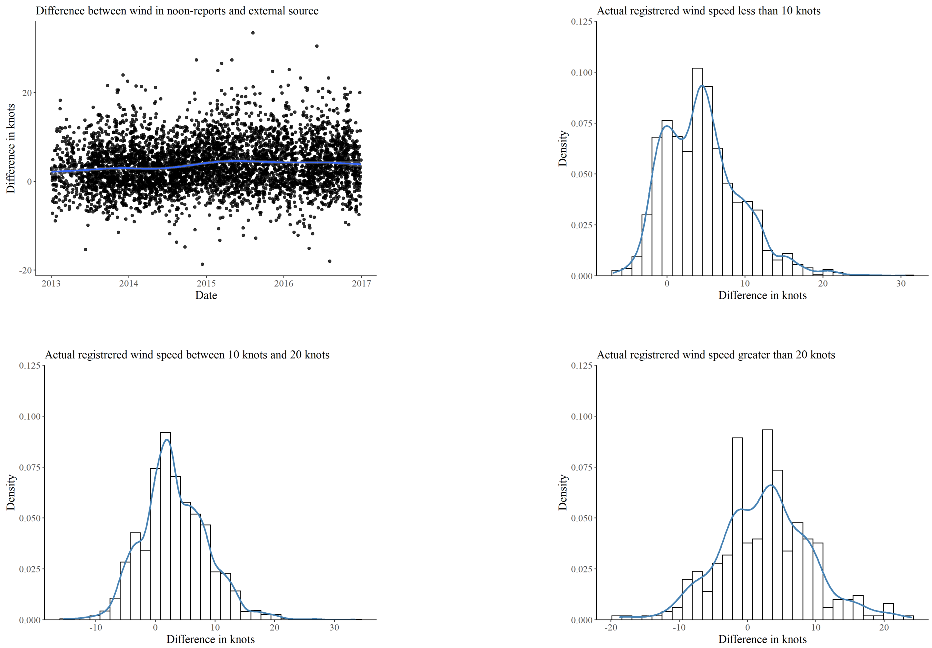


Figure A1.3: Wind analysis for Suezmax

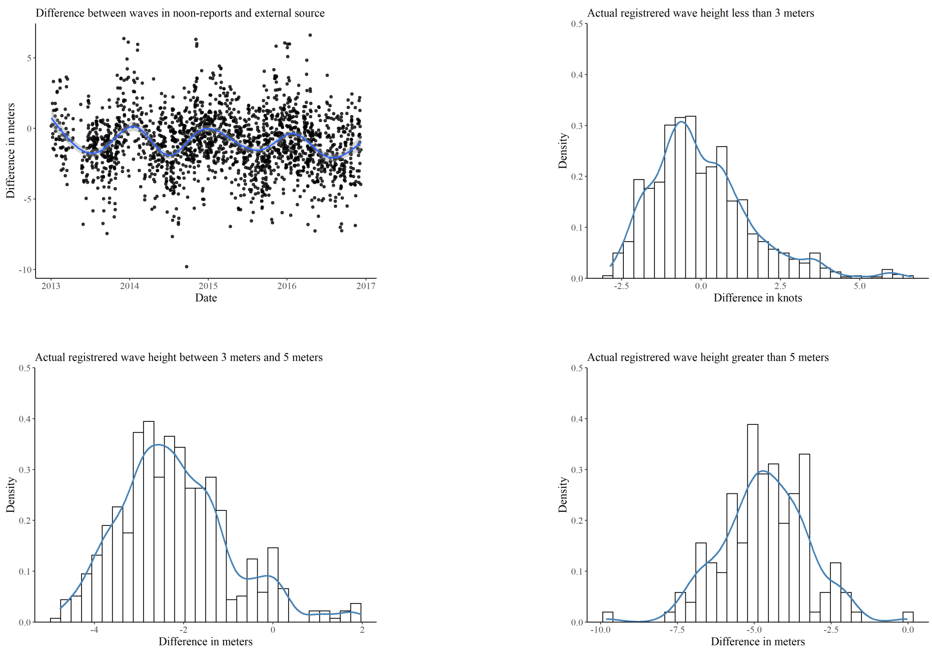


Figure A1.4: Wave analysis for Suezmax

Appendix B

B1 Descriptive statistics

Variabel	Product Tankers	Suezmax
Fuel	31.21461	36.08800
Speed	11.97393	11.93611
GPS_Distance	272.78640	273.38040
Trim	-0.26517	-0.80317
Current	0.04232	0.00964
Age	3.08028	9.05621
PASSAGEBallast	0.32166	0.50734
Calm_wind_1_knots	0.00065	0.00083
Light_air_1_3_knots	0.00627	0.00434
Light_breeze_4_6_knots	0.05650	0.04661
Gentle_breeze_7_10_knots	0.21696	0.21781
Moderate_breeze_11_16_knots	0.26116	0.33085
Fresh_breeze_17_21_knots	0.15916	0.21926
Strong_breeze_22_27_knots	0.08990	0.12275
Near_gale_28_33_knots	0.02680	0.04216
Gale_34_40_knots	0.00875	0.01157
Strong_gale_41_47_knots	0.00085	0.00186
Storm_48_55_knots	0.05683	0.04732
No_wave	0.00130	0.00103
Calm_0_0.1_m	0.01218	0.01819
Smooth_0.1_0.5_m	0.13798	0.15003
Slight_0.5_1.25_m	0.31162	0.35193
Moderate_1.25_2.5_m	0.22010	0.32031
Rough_2.5_4_m	0.10827	0.11635
Very_rough_4_6_m	0.02971	0.03431
High_6_9_m	0.00583	0.00641
Very_high_9_14_m	0.00022	0.03430
No_swell	0.05122	0.04980
Very_low_and_short_waves	0.12577	0.09320
Low_and_long_waves	0.13798	0.19539
Light_short_and_moderate_waves	0.22042	0.21802
Moderate_and_average_waves	0.17709	0.32155
Moderate_rough_and_long_waves	0.66451	0.07378
High_average_and_heavy_waves	0.00918	0.10539
Rough_short_and_heavy_waves	0.03793	0.03451
Very_high_long_and_heavy_waves	0.00130	0.00083
headwind	0.53517	0.49845
sidewind	0.14587	0.19446
alterwind	0.31896	0.30626
headswell	0.48460	0.44658
sideswell	0.16823	0.23063
alterswell	0.34706	0.32114
sst	24.30472	22.77931
tp	0.06669	0.02371
wdw	0.09777	0.37282
mwp	7.12018	6.90686
cdww	0.00110	0.00123
rain	0.48579	0.44389
snow	0.00065	N/A
wet_snow	0.00173	N/A
mix	0.00054	N/A
SDWT	119456	159000
Number of observations	9257	4841

Table B1.1: Descriptive statistics per vessel type

B2 Welch Two Sample t-test

Type	T-value	P-value	Deg. of Free.	95 % conf. int.		Sample estimates	
Product tanker	-0.16776	0.8668	3094.2	-0.01970783	0.1660124	-0.020706875	-0.005173923
Suezmax	-0.31301	0.7543	1599.6	-0.3517295	0.2549186	-0.06451840	-0.01611294

Table B2.1: Welch two sampled t-test for determining if the mean difference of predicted out-of-sample fuel and complete sample fuel is significantly different

B3 Example 1

Type	Product Tanker	Suezmax
Day	**GWP100**	**GWP100**
Day 1	-0.093	-0.184
Day 2	-0.105	-0.182
Day 3	-0.104	-0.216
Day 4	-0.104	-0.211
Day 5	-0.097	-0.193
Day 6	-0.094	-0.162
Average	-0.100	-0.191

* GWP values are multiplied with 100 000

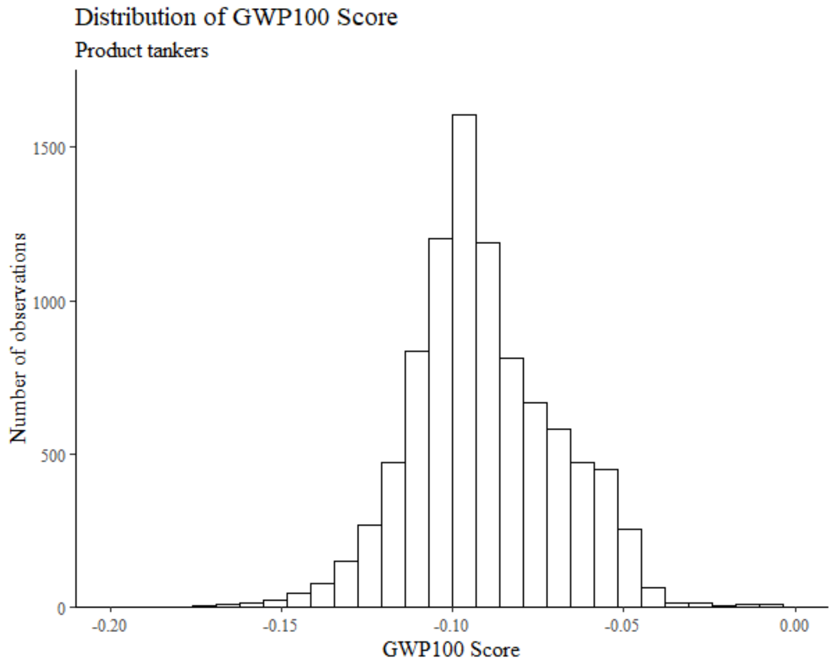
Table B3.1: GWP-score day-to-day for Example 1 in Figure 6.4

B4 Example 2

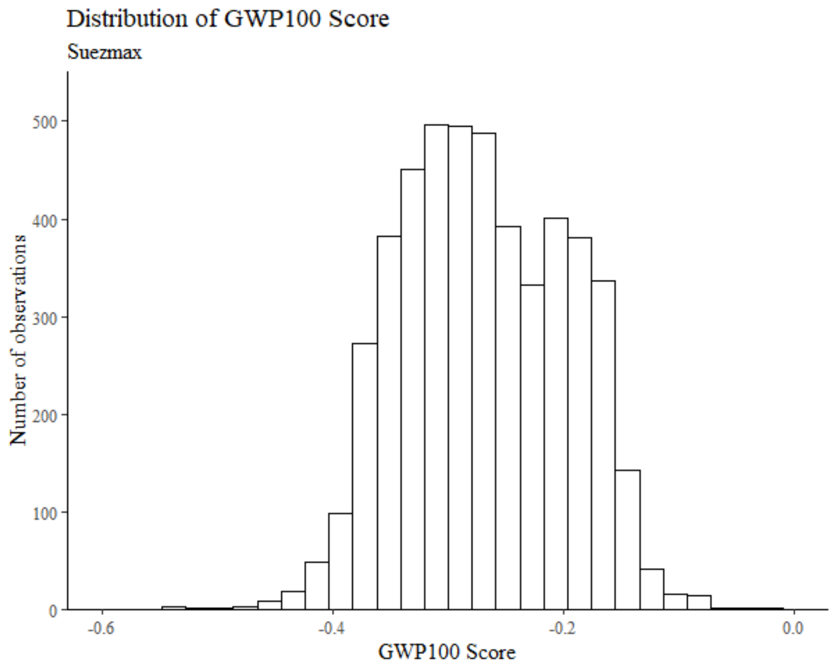
Day	**GWP20**	**GWP100**	**GWP500**
Day 1	-0.577	-0.068	0.172
Day 2	-0.853	-0.101	0.255
Day 3	-0.910	-0.108	0.272
Day 4	-0.853	-0.101	0.255
Day 5	-0.843	-0.100	0.252
Day 6	-0.721	-0.085	0.216
Day 7	-0.724	-0.086	0.216
Day 8	-0.758	-0.090	0.226
Day 9	-0.725	-0.086	0.217
Day 10	-0.725	-0.086	0.217
Day 11	-0.727	-0.086	0.217
Day 12	-0.731	-0.086	0.218
Day 13	-0.815	-0.096	0.244
Day 14	-0.840	-0.099	0.251
Day 15	-0.930	-0.110	0.278
Day 16	-0.864	-0.102	0.258
Day 17	-0.819	-0.097	0.245
Day 18	-0.839	-0.099	0.251
Day 19	-0.864	-0.102	0.258
Day 20	-0.720	-0.085	0.215
Day 21	-0.721	-0.085	0.216
Day 22	-0.761	-0.090	0.227
Day 23	-0.705	-0.083	0.211
Day 24	-0.686	-0.081	0.205
Day 25	-0.700	-0.083	0.209
Day 26	-0.734	-0.087	0.219
Day 27	-0.700	-0.083	0.209
Average	-0.772	-0.091	0.231

Table B4.1: GWP-score day-to-day for Example 2 in Figure 6.5

B5 Aggregated GWP100 score



(a) Product tankers



(b) Suezmax

Figure B5.1: Aggregated GWP100 for all observations for both vessel types