Norwegian School of Economics

Bergen, Spring 2021

Estimating weather margin seasonality in shipping using machine learning

Joakim Nilsson & Marcus Nilsson

Supervisor: Roar Os Ådland

Master thesis, Economics and Business Administration, Business Analytics

NORWEGIAN SCHOOL OF ECONOMICS

Abstract

Accurate predictions of fuel consumption are an essential tool in the pricing of forward cargo contracts. This thesis develops a predictive model for fuel consumption using noon report data from Handysize and Supramax vessels. In the process, we employ a wide selection of machine learning algorithms, including decision trees, shrinkage models, and an artificial neural network. Furthermore, we replace all weather and oceanographic variables with third-party data. The replacement ensures the model is independent of noon report weather data and allows us to generate predictions using historical weather conditions from the last decades. The trained models are used to study the seasonal patterns of weather margins for two case routes. Estimated weather margins and fuel consumption may be used by chartering managers to improve cost predictions and facilitate more profitable contract selection.

Acknowledgments

We want to express our gratitude to the parties that have assisted us during this process. Firstly, we would like to thank our supervisor Roar Os Ådland for his excellent guidance, connecting us with his valuable network within the shipping sector, and his rapid and thorough responses to any questions. Secondly, we would like to thank Western Bulk for providing us with the data necessary to complete our studies and a special thanks to Tobias W. Balchen for guiding us through the dataset's contents.

Keywords:
Ship fuel prediction
Weather margins
Machine learning
Decision trees
Shrinkage models
Artificial Neural Network

Contents

1. Introduction

Maritime shipping is the backbone of world trade. According to UNCTAD's review of maritime transport (2020), around 80% of the volume of international trade in goods is carried by sea, and the percentage is even higher for most developing countries. UNCTAD (2020) further estimates the total volume of maritime trade in 2019 at 11.08 billion tonnes. For a given freight, almost two-thirds of the expenses are attributed to fuel consumption (Stopford, 2009). This makes accurate fuel consumption predictions highly valuable for dry bulk operators who must price cargo that will be lifted several weeks ahead. Fuel predictions are, however, a complex task, and many factors must be taken into consideration.

Some of these factors can be grouped under the common term weather margin, sometimes also referred to as sea margin. We will use weather margin to refer to the increase in consumption due to weather compared to consumption in ideal conditions. A rule of thumb is to use a weather margin of 10-15%. For example, Nabergoj and Prpi (2007) found a weather margin of 10% when studying a passenger ship, but also mentions that 15-30% are typical values used by ship designers. However, weather conditions can be highly volatile, and traditional weather forecasts cannot provide reasonable long-term forecast accuracy. According to Hu and Skaggs (2009), the National Oceanic and Atmospheric Administration's 6- to 10-day forecasts are only correct 40% of the time. With forward cargoes being signed weeks in advance, chartering managers have to make contract pricing decisions based on unreliable weather forecasts or find alternative tools to reduce the weather margin uncertainty.

Previous authors have used a wide array of methods to predict fuel consumption. Some focus on modeling ships' physical features and relationships, often referred to as white-box models. In recent times, data-driven methods referred to as black-box models have increased in popularity. These methods are purely data-driven and use data to determine the historical relationships of ship features. Machine learning (ML) is an important tool in this process. Authors have applied a wide range of ML algorithms, from decision trees to artificial neural networks, with varying results.

A complicating factor in research on fuel consumption in shipping is the availability of high-quality data. As exemplified by this thesis, many prediction models are based on data from noon reports, which are often incomplete and not perfectly accurate. Moreover, noon reports are generally not publicly available, and their contents and formats may differ. As we will see in our literature review, the difference in data quality and format makes it difficult to determine the comparative performance of predictive models. These differences also mean that developed models cannot directly be applied to new data sources.

This thesis contributes to the literature by showing how free and publicly available third-party weather data can be used to reduce the problem of data availability and model generalizability. We show that all weather-related data from the noon reports can be discarded and replaced with standardized third-party weather data while achieving comparable predictive accuracy. We further find indications that higher-resolution weather data available from 2019 onwards may boost predictive accuracy beyond what is achievable with noon report data only. If these results are replicated by others, then the use of third-party weather data may serve as a step toward developing generalized predictive models for fuel consumption in the shipping industry.

The second contribution of our thesis is to demonstrate how the third-party weather data can further be used to estimate expected consumption on any voyage, thus mitigating some of the uncertainty stemming from the lack of long-term weather forecasts. To achieve this, we use our trained models together with 25 years of weather data to predict consumption given the historical weather conditions along the routes. The resulting predictions give us insight into how seasonal weather patterns translate into changes in expected consumption and variance at different times of the year and on different voyages.

We believe our approach may prove to be a viable method for improving cost estimates and subsequently enabling more accurate pricing of forward cargo and a better understanding of the risk associated with a given voyage. Margins in the shipping industry are relatively modest. For example, Fidan (2019) estimates the industry average to be around 6-10%. This means the economic margins are comparable in size to the variation in weather margins, underlining the economic importance of accurate weather margin predictions.

The remaining sections of this thesis are structured as follows. Section 2 will study the theoretical framework for fuel consumption and prediction. Section 3 covers the existing literature within the shipping analytics field. Section 4 describes our methodological approach. Section 5 will present our modeling results. Section 6 will analyze weather margins for two real-world cases. Section 7 will outline the limitations of our study and propose future areas of research. Finally, section 8 will present our overall conclusions.

2. Theory

2.1 Vessel fuel consumption

There are many features influencing fuel consumption, such as vessel speed, draft, trim, waves, wind, sea current and propeller slip (Gkerekos et al., 2019; B. J. S. Wang et al., 2018). Vessel speed is one of the most important predictors of fuel consumption, as has been shown by numerous authors in the past (e.g., Adland et al., 2020; Gkerekos et al., 2019; B. J. S. Wang et al., 2018). As described by Meng et al. (2016), vessel speed V primarily impacts consumption by increasing the total resistance R_T according to the formula:

$$P_E = R_T \times V \tag{1}$$

where P_E denotes the effective power necessary to move the ship forward at the given speed and is closely related to fuel consumption. According to the authors, R_T consists of three components:

$$R_T = R_F + R_R + R_A \tag{2}$$

where R_F represents the frictional force of the hull and the propeller, R_R is the residual resistance mainly caused by waves, and R_A is the air resistance. Although the exact proportions can vary, the authors suggest that the three resistance components are in proportion to V^2. Vessel speed also influences the relationship between P_E and fuel consumption by affecting the efficiency of the engines, the propellers and more (MAN Diesel & Turbo, 2015). For example, the highest efficiency for electronically and mechanically controlled MAN engines is obtained at 70% and 80% of maximum power, respectively (MAN Diesel & Turbo, 2015).

The magnitude of the impact from waves (through residual resistance R_R) is dependent on factors such as wave height and modal period (Arribas, 2007). Similarly, the magnitude of air resistance R_A depends on factors such as wind direction, wind speed and the size of the superstructures determining the total resistance (Magnussen, 2017). Even though the frictional resistance R_F is a large part of the total resistance R_T (Meng et al., 2016), the impact of sea current speed and current direction on consumption are relatively low (Abebe et al., 2020; Adland et al., 2020). Hull condition is another factor that impacts R_F, and as will be discussed in section 4.1.3, this variable can have a large influence on fuel consumption.

The complexity of vessel fuel consumption makes it difficult to model accurately. Variables can be correlated with each other and themselves in forms that are not always easily reproduced in linear regression (LR) models. For example, the cubic law of ship speed claims that fuel consumption can be well approximated by a cubic function of speed (S. Wang & Meng, 2012). While the approximation fit has later been challenged Adland et al. (2020), it is clear there may exist non-linear relationships between predictors and fuel consumption. Another example is that a vessel's draft is determined by its weight, which again is determined by its cargo. The shape of a hull also means the marginal increases in water displacement rise with increasing draft size. There are also many more influential predictors for fuel consumption, which we will introduce and study later in the thesis.

2.2 Machine learning

In an attempt to more accurately model these complex, and at times nonlinear relationships with fuel consumption, authors have implemented various data-driven machine learning approaches. Machine learning can be broadly defined as computational methods using experience to improve performance or make accurate predictions (Mohri et al., 2018). Machine learning algorithms take many different forms but can be grouped based on similarities. We will now study some of the relevant groups for our thesis.

Regression methods estimate the relationship between a dependent variable and one or more independent variables based on historical data and iteratively minimizes the estimation errors. The most common method is Ordinary Least Squares (OLS) regression, where the estimation errors are measured as the sum of squared differences. Shrinkage models, or regularization algorithms, are an extension of other algorithms which penalize increased model complexity, such as added predictors (Brownlee, 2020a). The regularization methods are often combined with regression methods, as described above. Ridge and Lasso are two examples of shrinkage algorithms that are based on regression.

Another group of machine learning algorithms is instance-based models. These algorithms generate instances or examples of training data deemed important or required for the model instead of using the training data itself (Brownlee, 2020a). These algorithms are useful when the target function is very complex but can be broken down into less complex generalizations. Examples of instance-based algorithms include support vector machines (SVMs) and k-nearest neighbor algorithms.

Decision trees are amongst the most popular groups of algorithms in machine learning. They work by forking decisions in a tree-like structure until a final classification or prediction is reached (Brownlee, 2020a), hence their name. Decision trees are often highly accurate and come in many different variants, including Random Forests (RF), which uses bootstrap replicas and optimal splits, and Extra Trees (ET), which uses the whole learning sample and randomly selected splits (Geurts et al., 2006). Cubist is a third variant that combines decision trees with regression.

Bayesian machine learning methods are based on the principle of Bayes' rule (Tipping, 2004). They can use a non-parametric approach; instead of learning exact values for every parameter in a function, the Bayesian approach infers a probability distribution over all possible values. Examples of Bayesian methods include the Naïve Bayes, the Gaussian Naïve Bayes and the Multinomial Naïve Bayes.

The last group we examine has grown immensely in the last few years. Artificial Neural Networks (ANNs) are a class of pattern recognition algorithms that uses interconnected nodes with associated weights and activation functions to make predictions. ANNs contain nodes structured in an input layer, one or more hidden layers, and an output layer. ANN algorithms can differ in workflow. In Feed-Forward Networks, the data flows in one direction from start to finish. Recurrent Neural Networks are a more advanced form, where data can be fed back into the input layer or previous traversed hidden layers.

2.3 Performance metrics

For evaluation and comparisons of the performance of the different models, we need to select performance metrics. There is no ideal "one-fits-all" performance metric. Each has drawbacks and advantages, so it is important to study the measure's purpose (Swalin, 2018) and possibly include several measures. For this thesis, we need a metric that is well suited to compare a large number of models. Secondly, we need a scale-free measure to compare our achieved accuracy with the accuracy of other authors in the literature. Lastly, we want a measure that is easily interpretable for our readers. To fulfill all these purposes, we decided to include several measures. The selected measures are RMSE, nRMSE, sMAPE and R squared. In the following, we will explain the measures and why we found them appropriate for our purposes.

The Root Mean Squared Error is defined as

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{3}$$

The RMSE takes the difference between the predicted value $\hat{y}_i$ and the true value y_i i.e., the prediction error, and squares it so that negative and positive errors are weighted equally. Squaring the errors has the collateral effect of penalizing larger prediction errors more harshly than smaller errors, which is often desirable in an economic setting where risk has a cost. This makes the measure suitable for our thesis. All prediction errors are then summed and divided by the number of observations to find the average error, before the square root is applied. Lower RMSE corresponds to better model performance. RMSE is scale-dependent but a good measure to select between models on the same dataset (Chugh, 2020; Swalin, 2018), and is widely used in literature and comparable studies from our literature review. As such, we will use RMSE to determine our best-performing model.

The Normalized Root Mean Squared Error is defined as

$$nRMSE = \frac{RMSE}{\sigma_y} \tag{4}$$

The RMSE also has several normalized variants that are scale-free, including the standard deviation normalized variant presented above. The normalization makes the measure suitable for comparing models with different units or scales, such as if a dependent variable is to be compared with a log-transformed dependent variable (Otto, 2019). The standard deviation variant of nRMSE is a good choice as it represents the ratio between the variation not explained by the regression versus the overall variation in the dependent variable (Otto, 2019). A nRMSE score of 0 means all variation is captured by the model, while a score of 1 means the model captures no variation. Consequently, lower values represent better performance.

The Mean Absolute Percentage Error is defined as

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{5}$$

In contrast to RMSE, which squared the prediction errors, MAPE divides them with the true value before averaging them. This ensures that the measurement is scale-free and comparable to other author's models and datasets, which is one of the purposes we wanted to fulfill. One key difference compared to RMSE is that absolute errors are used rather than squared errors, making it less sensitive to outliers and less attractive for model selection. On the other hand, the measure is more interpretable than RMSE. Lower MAPE values correspond to better model performance.

The Symmetric Mean Absolute Percentage Error is defined as

$$sMAPE = \frac{1}{n}\sum_{i=1}^{n} \frac{|y_i - \hat{y}_i|}{(|y_i| - |\hat{y}_i|)/2} \tag{6}$$

While the normal MAPE is a good measure, it has some drawbacks that are corrected for in the symmetric MAPE presented above. Firstly, MAPE can go over 100% for positive values but not for negative values, so it tends to weigh positive errors higher than negative errors (Lewinson, 2020). Additionally, MAPE is undefined when actuals are zero. The sMAPE mitigates these drawbacks by setting 200% as the upper bound and setting actual observation values of zero equal to the upper bound. The described corrections convert sMAPE into a similar but improved version of the normal MAPE, and as such, we will only be using the symmetric version of MAPE in our results.

The R squared is defined as

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_{\,i})^2} = 1 - \frac{Unexplained\ variance}{Total\ variance} \tag{7}$$

R squared subtracts from one the sum of the squared prediction errors divided by the squared differences of the true values and the average actual value. In simpler terms, it calculates the proportion of the variation of a dependent variable explained by the independent variable(s) (Fernando, 2020). A model preferably explains as much of the variance as possible, thus a higher R squared means a better performing model. This measure is well known and easily interpretable (Swalin, 2018), so we will include it for the purpose of explaining the models' performance in an interpretable and familiar manner for our readers.

3. Literature review

3.1 White-box models

As mentioned, white-box models are based on the physical characteristics of the ships and the environment. Jalkanen et al. (2009) modeled fuel consumption using their Ship Traffic Emission Assessment Model (STEAM) based on data from Automatic Identification System (AIS) data, with prediction errors within 5%. They further improved the accuracy in their 2012 publication (Jalkanen et al., 2012). Tillig and Ringsberg (2019) used Monte Carlo simulations to estimate fuel consumption during the ship design phase. The simulations were built on numerous empirical methods applied to data about the various physical ship features. They achieved prediction errors below 4% in the later design phases. Goldsworthy and Goldsworthy (2015) made a generic model for predicting fuel consumption and emissions, using a combination of AIS data and ship mechanical data from Lloyd's database. Their prediction errors reached below 3%. Magnussen (2017) modeled ship resistance using ISO standards and further estimated sea margins for a case ship sailing three different routes, and found the sea margins to be 18-20%. Eide (2015) modeled the sea margins based on data from noon reports, and found that the proposed margin by the ship designer of 15% was accurate at design speed, but inaccurate at lower speeds.

3.2 Black-box models

Linear regression is one of the simplest types of black-box models. It provides excellent interpretability in combination with accurate predictions. Adland et al. (2020) used linear regression to calculate fuel consumption, with R squared scores of 82.4% and 87.3% for Aframax and Suezmax vessels, respectively. Similarly, Erto et al. (2015) used linear regression to predict the fuel consumption of a cruise ship in the Mediterranean Sea, and achieved an R squared of 94% on their training data.

Continuing with machine learning models, Pedersen and Larsen (2009) used linear and nonlinear regression and ANNs on data from noon reports, onboard sensory data and hindcasts of weather and sea information to predict full-scale propulsion power. The highest accuracy was achieved using ANNs on onboard sensory data. They further found that introducing hindcast data reduced the prediction errors and gave the best solutions in general. Petersen et al. (2012) used publicly available data in ANNs and Gaussian Process (GP) models. While the GP models have the advantage of quantifying the uncertainty, they fall short of the accuracy of the ANNs. Their paper highlights the difficulty of comparing models across different datasets and encourages the release of more data to the public. Jeon et al. (2018) developed ANN models to predict ship fuel consumption with accurate results. The ANNs outperformed both Polynomial Regression and SVMs on the dataset. Uyanik et al. (2019) also used ANN to predict ship fuel consumption from 23 days' worth of data from a voyage, though with lower performance metrics scores than Jeon et al. (2018).

One potential pitfall in machine learning modeling is overfitting, where the model performs well on the training data but generalizes poorly. Shrinkage models attempt to counteract this phenomenon by restricting the total weights that can be allocated to the variables. The method often leads to reduced variance at the cost of slightly more bias in the fitting process. Soner et al. (2019) applied the shrinkage-based Ridge and LASSO models on the same data as Petersen et al. (2012). They achieved a comparable prediction accuracy that was lower than the ANNs but higher than the GP models. Wang et al. (2018) proposed a LASSO regression to predict consumption, resulting in highly accurate results combined with high interpretability and low running time. The model outperformed the ANN model on the same data, despite ANNs having proved to be amongst the more accurate model types.

Gkerekos et al. (2019) compares a large selection of data-driven regression algorithms on both noon reports and onboard sensory data, and focuses on giving the models equal grounds for comparison. They find that the RF models provide the most accurate predictions of fuel oil consumption, closely followed by ANNs and SVMs. The much simpler LR model also provides comparable results. Similarly, Abebe et al. (2020) proposes a maritime data analysis framework based on AIS and marine weather data to predict ship speed over ground (SOG). They used a combination of AIS satellite data and noon-report weather data of 14 tankers and 62 cargo ships, and applied various machine learning algorithms. Like Gkerekos et al. (2019), they found that ET and RF achieved the most accurate results.

Based on the literature alone, it is difficult to tell which models perform best. Algorithms that perform better in some studies perform worse in others. For example, Petersen et al. (2012) and Jeon et al. (2018) achieved the highest accuracy with ANNs, while both Gkerekos et al. (2019) and Abebe et al. (2020) found that ET outperformed ANNs. However, as neither

Petersen et al. (2012) or Jeon et al. (2018) applied ET algorithms, we cannot rule out that they would have achieved better results than the ANNs.

The comparisons become even more complicated when we take into consideration the data the models are based on. If one author achieved better results with ANNs than another achieved with ET, we cannot rule out that it was not just higher quality or quantity of data (or both) that led to the disparity in results. Furthermore, models may require or perform better on data with certain characteristics. Algorithms that perform best on some datasets are not necessarily the best on other datasets. For example, it is unclear why Gkerekos et al. (2019) achieved high accuracy with SVM, while Jeon et al. (2018) achieved sub-par performance from the same algorithm.

The modeling insight from the literature therefore brings us back to Petersen et al. (2012). They highlighted the difficulty of comparing models across dissimilar data, and emphasized the need for more publicly available data for easier comparisons of models. Gkerekos et al. (2019) also emphasize that there may be a larger gain from testing different algorithms compared to meticulously tuning a single algorithm. Until more data becomes publicly available, it will therefore be necessary to apply several of the competing algorithms to ensure the best algorithm is not excluded. In our thesis, we will follow this recommendation and test a wide selection of algorithms that have proven to give reasonable results in the literature or in our own testing. These are ANN, ET, RF, LASSO, Cubist, SVM, and GP. For the SVM and GP models, we will apply two variants of kernels; polynomial and radial. We also include LR to serve as a baseline model.

4. Data and methods

The data and methodology section is structured chronologically, i.e., in the same order as the modeling procedures were carried out. We have followed the workflow used by Abebe et al. (2020), illustrated in Chart 1 below, and the steps are described in greater detail in their corresponding paragraphs. Using this workflow ensures important data modeling principles are followed, such as setting aside an unseen test set for model evaluation while training and tuning the models on separate training and validation sets. In our case, cross-validation is used to create the train and validation sets. We will start by describing our data acquisition, and the other steps will follow after that. All implementation of methodology was performed in R, using a Windows 10 operating system.

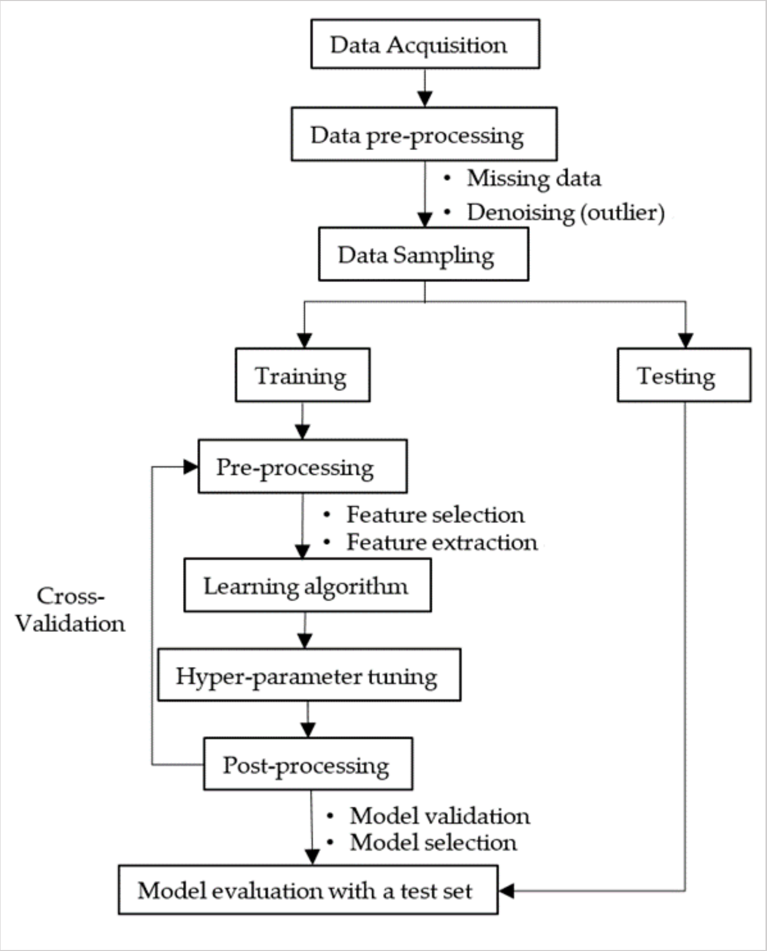

Chart 1. Methodology workflow (Abebe et al., 2020).

4.1 Data acquisition

4.1.1 Noon report data

Many authors, including us, rely on noon reports for their data collection. Noon reports are prepared once per day at noon, usually by the ship's captain or chief engineer, with standardized data to assess the ship's performance based on its speed and environmental forces, including weather conditions (Anish, 2019). Other authors rely on more accurate data from automated onboard sensors. These can provide additional data parameters and have an update frequency as low as seconds. However, they are used to a lower extent than noon reports, and acquiring data from the same number of ships might be challenging.

Noon reports are not necessarily perfectly accurate due to their many possible sources of error, which may negatively impact model accuracy. When noon reports are prepared manually, it exposes them to the risk of human errors such as misinterpreted readings and input errors, and chief engineers might use different units, rounding, or even leave parts of reports empty or fail to deliver reports at all. Sensors may also fail, be inaccurate or uncalibrated, or give erroneous readings for other reasons. Aldous et al. (2013) studied the uncertainty of noon reports as a data source. They suggest additional sources of uncertainty, including failure to adjust for time zones, using different sensors to populate the same fields, and the low resolution of reporting units, such as the Beaufort scale or binary values for load status. Their study fitted a regression model that captured as much as possible of the information affecting fuel consumption and ensured the remaining residuals were normally distributed, leading to a model that closely approximated the true underlying model. Their regression results showed relative standard errors in the range of 1-8% for various types of oil tankers, and 15.8% for LNG carriers, which they argue is due to the aleatory and measurement uncertainty present in noon report data. To address the high uncertainty present in noon reports, it is clear that a rigorous pre-processing routine is required before the data can be used in our models.

For this thesis, we received access to an unprocessed dataset from the international shipping company Western Bulk, consisting of 8,995 noon reports from November 2015 to April 2021. Of these, 6,580 are sourced from a fleet of approximately 100 Supramax bulk carriers of the same design. The remaining are from approximately 25 Handysize bulk carriers, all of which are also of the same design. For an overview of the scope of the dataset, noon reports from the respective vessel type can be seen plotted on world maps in Figure 1 below. We see a high degree of route overlap for the two designs, with the Supramax carriers possibly being a bit more present in the East and the Indian Ocean than the Handysize carriers. Certain shipping lanes account for significant parts of the noon reports, such as the coastlines surrounding Africa.

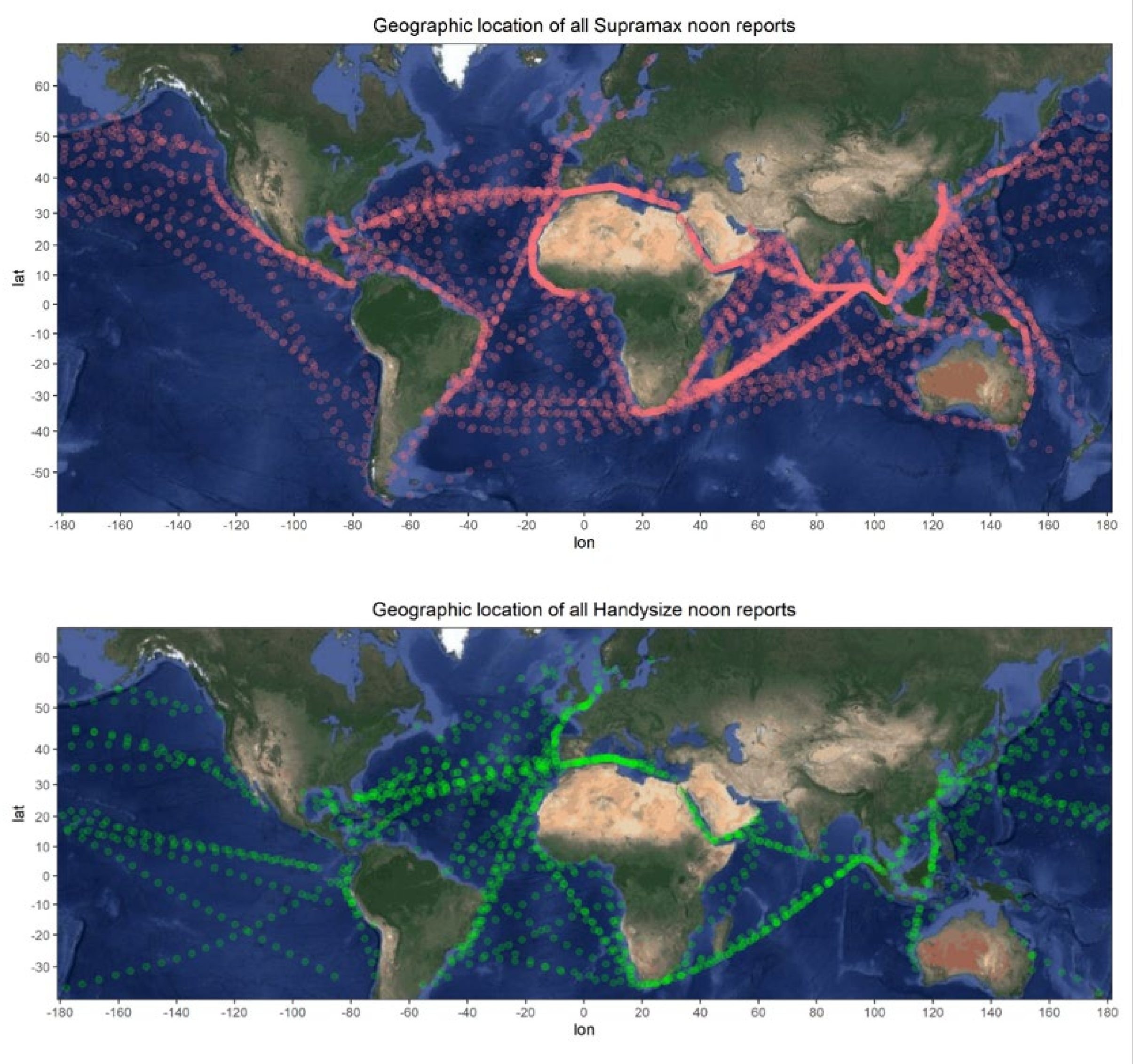

Figure 1. Map of the geographic locations of all Handysize and Supramax noon reports.

4.1.2 Third-party weather data

In addition to the noon reports, we have retrieved third-party weather and oceanographic data. Table 1 summarizes the datasets and variables we have used.

Dataset (Source)	Storage Size	Temporal Resolution	Spatial Resolution	Temporal Coverage (Period Used)	Variables Used (Variable Identifier in Dataset)
ERA5 hourly data on single levels from 1979 to present[1] (CDS)	186 GB	1 hour	0.5°	1950-01-01 - Current (1995-01-01 - Current)	Mean wave period (mwp)
					Mean wave direction (mwd)
					Combined height of wind waves and swell (swh)
	476 GB	1 hour	0.25°	1950-01-01 - Current (1995-01-01 - Current)	Eastward component of wind (u10)
					Northward component of wind (v10)
Global Sea Physical Analysis and Forecasting Product[2] (CMEMS)	1,060 GB	1 hour	1/12°	2019-01-01 - Current (All)	Sea water temperature (thetao)
					Eastward component of current (uo)
					Northward component of current (vo)
	54 GB	6 hours	1/12°	2019-01-01 - Current (All)	Sea water salinity (so)
GLORYS12V1 - Global Ocean Physical Reanalysis Product[3] (CMEMS)	538 GB	24 hours	1/12°	1993-01-01 - 2019-12-31 (1995-01-01 - 2018-12-31)	Sea water temperature (thetao)
					Eastward component of current (uo)
					Northward component of current (vo)
					Sea water salinity (so)

Table 1. All used third-party weather datasets and variables. For variables available at different depths, we have used values at a depth of 0.5 meters. The wind variables are "surface level", which is defined as an altitude of 10 meters. CDS: Coperernicus Climate Data Store; CMEMS: E.U. Copernicus Marine Service Information.

We have used the original resolution of all variables. It is worth noting that the highest resolution dataset for temperature, current, and salinity is available only from 2019 and later. For noon reports earlier than this date, we have instead used a different dataset with a daily (rather than hourly) resolution. The lower resolution is likely to have a more significant impact on sea current accuracy than salinity, since salinity is essentially time-invariant. Figure 2 illustrates this, with significant variations in sea current and close to constant salinity levels.

Ocean temperature and salinity are less commonly used as input variables for prediction fuel consumption but may still have some predictive power. According to Abebe et al. (2020), these variables are directly proportional to the viscosity and density of the seawater. A higher viscosity or a higher density of water will increase the frictional resistance of the vessel. Since water starts decreasing in density when warming past 4 degrees Celsius, fuel consumption

[1] (Copernicus Climate Change Service (C3S) Climate Data Store (CDS), 2018)
[2] (E.U. Copernicus Marine Service Information, 2019)
[3] (E.U. Copernicus Marine Service Information, 2018)

may be lower in warmer waters. On the other hand, higher temperatures are related to higher biofouling rates, which we will describe in more detail in the next section.

In the later case studies, we will be using historical weather data going back to 1995 to estimate seasonal patterns in consumption. This analysis relies on the assumption that weather patterns in this entire period are representative of current weather patterns, meaning that there must not be any influential long-term trends. Figure 2 shows long-term changes in the weather variables. The figure does not reveal any notable long-term trends or cycles, indicating that our assumption holds.

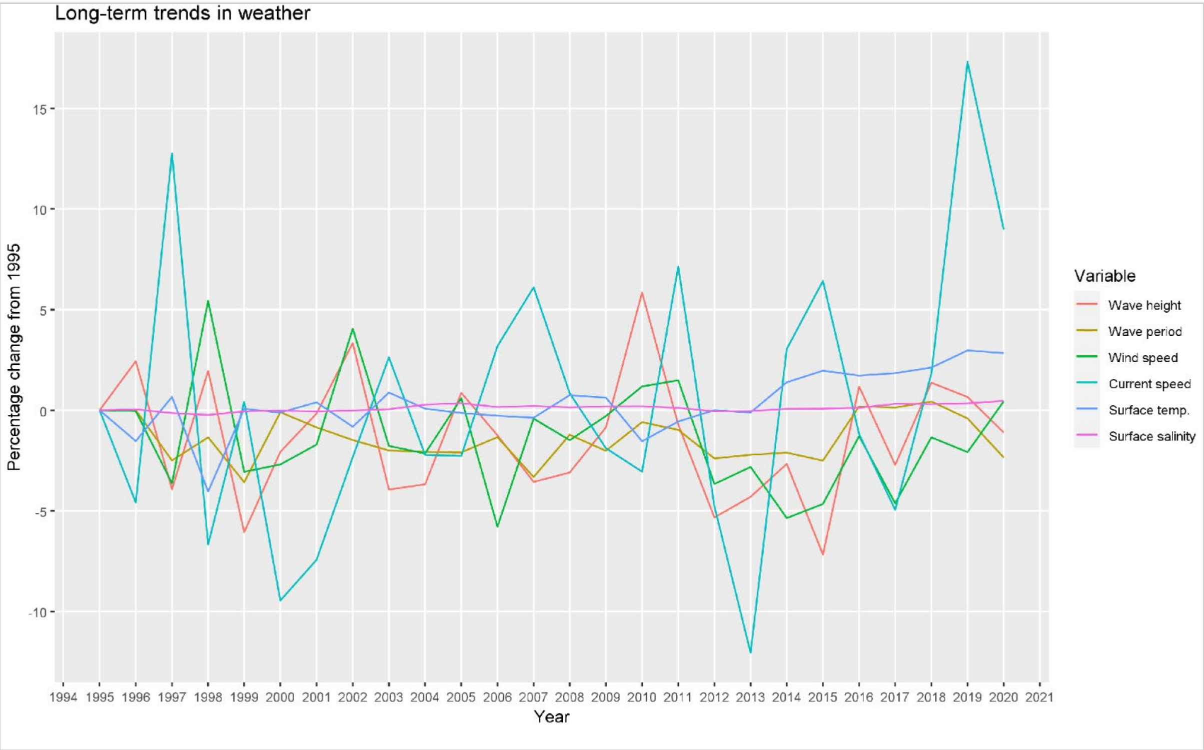

Figure 2. Long term trends in weather. Yearly means are calculated from monthly samples from 19 locations evenly distributed along the transatlantic route that we will be presenting in Section 6.3. This gives a total of approximately 500 samples per variable per year. The graph shows "true" weather forces (as experienced by a stationary observer).

Based on our testing, the long-term trends shown in Figure 2 were not notably different when sampled from other routes or locations. The same cannot be said for the seasonality in weather patterns, which can differ significantly depending on the location from which it is sampled. For this reason, we have included separate figures (Figure 14 and Figure 17) showing seasonality for the two routes we will be examining in the case studies. These can be found in Sections 6.2 and 6.3.

4.1.3 Clarksons' World Fleet Register

Hull fouling describes the deterioration of the hull condition over time, mainly due to marine growth (biofouling). By increasing the roughness of the hull, full fouling increases the frictional resistance from moving through water which in turn increase fuel consumption. This problem can be attenuated by periodic hull cleaning. Adland et al. (2018) show that the daily fuel consumption of oil tankers is reduced by 17% after hull cleaning in dry docks and by 9% after underwater hull cleaning. This finding shows the large impact hull fouling can have on fuel consumption.

Biofouling generally occurs when vessels are stationary or at speeds below 2 knots but is also highly dependent on water temperature, salinity, and sea currents (Dürr & Thomason, 2010). Salinity affects self-polishing rates and biocide release rates and, consequently, the ability to prevent or limit biofouling (Lindholdt et al., 2015). Increased salinity corresponds to more leaching of the protective coating (Lindholdt et al., 2015) and, as a consequence, may increase biofouling. As illustrated in Figure 2, the salinity levels in oceans are nearly constant, which means the observed effect of salinity changes largely depends on location. Higher temperatures typically lead to higher biofouling intensity (Lindholdt et al., 2015), and are associated with higher rates of polishing and dissolution of protective paints on the hull (Kiil et al., 2002). These effects lead to increases in biofouling when sailing in warmer and more saline waters and mean that if included, the temperature and salinity variables may pick up some of the added effects of biofouling from sailing in these conditions.

We retrieved data on dry docking events from the Clarksons' World Fleet Register (Clarksons Research Services Limited, 2021). Although a complete record could not be obtained, we managed to determine the most recent dry docking date for about half of our noon reports. In addition, we will use data about vessel age in combination with the typical inter-docking interval to impute the remaining values which, according to Bohlander (2009), is five years. The data was then added to the dataset as a "time since last dry docking" feature, with the goal of capturing the negative effect hull fouling has on fuel consumption.

We also retrieved data on vessel age. The physical degradation of ships is a gradual process, and older ships generally have a higher operating cost (Stopford, 2009). Rakke (2016) found that engine age could affect fuel consumption by as much as 10%. Unless exceptionally well maintained, hull fouling will also gradually reduce the maximum operating speed (Stopford, 2009). These factors indicate that vessel age may be an important feature for fuel consumption, where higher ages are related to lower fuel efficiency. In addition to vessel age, we also included a factor variable for the main engine model to capture potential efficiency differences. Finally, we included a variable describing whether the propeller has been fitted with either a propeller duct or a boss cap fin. Research has shown that some types of ducts can increase efficiency by up to 12% (Yilmaz et al., 2013), and similarly, boss cap fins have been shown to increase the open water efficiency by 1-5% (Xiong et al., 2013). Based on this, we expect ships with these efficiency augmentations to have slightly higher fuel efficiency.

Table 2 shows the final selection of variables used as input to the machine learning models, grouped by their data sources. A more thorough description of some of the transformed variables follows in the next chapter.

Final selection of input parameters

	Parameter	**Unit**
Noon report data		
1	Speed over ground	kts
2	Draft	m
3	Trim	m
4	Latitude	deg.
5	Longitude	deg.
6	Imo number	-
Clarksons' World Fleet Register data		
7	Time since dry docking	years
8	Vessel age	years
9	Eco Propellar	T/F
10	Main engine model	-
Copernicus data		
11	Sea surface temperature	°C
12	Sea surface salinity	psu
13	Mean wave period	sec
14	Wave height	m
15	Wave height/direction interaction	m · deg. forwards
16	Wind speed	kts
17	Wind speed/direction interaction	kts · deg. forwards
18	Current speed	kts
19	Current speed/direction interaction	kts · deg. forwards

Table 2. The final selection of input parameters for the machine learning models, grouped by data sources.

4.2 Pre-processing and transformation

4.2.1 Scope of the model

In this thesis, we have limited our models to predict only the fuel consumption of the main engines. There may also be additional fuel consumption from auxiliary engines that generate electricity for the vessel, but their consumption is dependent on a different set of factors than the main engines. Combining them in the same models could therefore have reduced the accuracy of the predictions.

The low temporal resolution of noon reports puts some limits on what can be modeled. Activities such as maneuvering near ports and shores or acceleration cannot be accurately represented in the noon reports but still impact consumption. Using these observations in our model could have led to an overestimation of the fuel consumption under normal sailing conditions on the open sea. Therefore, we limit our analysis to fuel consumption during open-sea sailing.

Some of the criteria we set to achieve this include filtering noon reports where the current status was marked as anything other than open-sea sailing. For the same reason, reports with SOG below 7 or above 15 knots were removed. Noon reports filed sooner than 20 hours or later than 27 hours after the previous report were also discarded, as many of the possible reasons the reports are not registered on time may also imply sailing activities not representative of open-sea sailing.

An interesting observation about the scope of the model is that a narrower scope often leads to increased performance metrics. The reason for this can be illustrated with an exaggerated example. By narrowing the allowed values of fuel consumption down to just values between 19-20 tonnes and training the model on these observations, a good model would rarely miss by more than 1 tonne, which would lead to seemingly good scores of metrics like sMAPE. This further complicates the already tricky performance comparisons with models of other authors, who may have used different scopes.

4.2.2 Transformations

We previously described how some of the predictors in our model, such as draft and speed, may have a nonlinear relationship to the dependent variable. Log-transforming fuel consumption to adjust for this in the model was thus tested during the modeling phase. The results of the comparison are seen in Table 3 below. The performances are relatively similar between the two, with five models performing better for the log-transformed dependent variable and three models performing better with the untransformed dependent variable. As the log-transformed version scored slightly better, we decided to apply the log-transformation on our dependent variable. We performed a similar test for log-transforming the speed over ground (SOG) feature but decided not to apply it as the performance was slightly worse with the transformation applied. The results of this comparison are available in Appendix C.

Log/level comparison for Supramax		
Transformation	Log-transformed	Level
Model	sMAPE (%)	sMAPE (%)
Linear Regression	7.77	7.77
Neural Network	6.42	6.28
Extra Trees Regression	3.62	3.62
Random Forest	6.47	6.57
LASSO	7.81	7.83
Ridge	7.83	7.85
SVM Poly	6.01	6.19
SVM Radial	5.20	4.71
GP Poly	5.81	7.80
GP Radial	15.31	15.25

Table 3. Comparison of model performance between a model with log-transformed fuel consumption and the same model using level fuel consumption.

The vessel's bearing is a key variable as it is a prerequisite for direction-dependent effects like sea current and wind. The noon reports did not include this information, so it was preferentially added from external AIS data by matching based on IMO number and the midpoint time between the current and the previous noon report. In total, AIS data for bearing and position was available for only 3,743 noon reports, with data missing for the remaining 5,252 noon reports.

To fill the remaining information, we imputed the values based on the position and time of consecutive noon reports. Each noon report provided the current time and the elapsed time since the last noon report, which could be used to calculate the expected time of the previous report. Thus, we grouped the reports by vessels using their IMO numbers, calculated the expected time of the previous noon report, and searched for any matching reports. An error margin of two hours was applied for matching the expected time with the actual time. Given the importance of knowing a vessel's bearing, we discarded the approximately 400 observations where this information could not be determined.

For every observation matched with its preceding observation, we now had its starting and ending position for the past approximately 24 hours. We used this to estimate the midpoint position and average bearing. Since AIS data was unavailable for these noon reports, the estimation had to be made with the assumption that the vessels held a constant speed and sailed in a straight line. Over 24 hours, the bearing of a vessel sailing in a straight line may change up to several degrees due to Earth's curvature. We used the vessel's straight-line bearing calculated from the midway point to approximate the average bearing over this period.

A problem with the described approach is the mentioned assumptions that vessels were sailing in a straight line at a constant speed. In particular, vessels routinely navigate around land to reach their destination, sometimes resulting in a significant difference between the straight-line distance and the reported sailing distance. To mitigate this, we compared the implied straight-line distance traveled with the distance traveled as stated on the noon reports. A straight-line distance lower than the reported distance implies maneuvering, invalidating our straight-line assumption and, in turn, our bearing and midpoint calculation. A greater straight-line distance implies a data or rounding error. We discarded a total of 197 rows with a straight-line distance lower than 90% or greater than 105% of the reported distance.

A more precise approach would have been calculating the shortest *possible* distance, accounting for known land masses. However, since noon reports provide only the cumulative fuel consumption since the previous report, it is unclear whether it would have resulted in a more accurate model. For instance, it would not be possible to disentangle the fuel impact of head wind before a 90-degree turn from the effects of port side wind after the turn.

4.2.3 Matching and processing third-party weather

Our third-party weather data was organized along three dimensions: latitude, longitude, and time. Depending on the year, there may only be one data point every 24 hours, similar to the noon reports. Since fuel consumption from the noon reports details the cumulative consumption since the previous report, each report was matched with the weather at the time and location of the midpoint between itself and the preceding report. These values were matched with weather data using nearest-neighbor interpolation.

A different interpolation technique could have possibly achieved more accurate results. In particular, we believe that linear interpolation in the time dimension might improve the result. We were, however, unable to attempt this due to the excessive amount of computational time required (with disk reading speed being the bottleneck).

Initially, noon report data for wind, sea current, and waves include separate variables for direction and force. However, the direction of these forces does not matter with respect to fuel consumption if there is no wind, current, or waves. To facilitate our model to interpret the effects of these forces on the vessels accurately, we replaced the directional variables with interaction terms of the direction and speed of the forces. This meant that we first had to transform the forces' directions to a scale from 90° to -90°, where -90° are forces moving the same direction as sailing direction and 90° are forces moving the opposite of sailing direction. This transformation is illustrated in Figure 3. We then multiplied the directional variables with the forces of the effects. This ensured wind from astern and wind from ahead of equal speed affect the fuel consumption in opposite directions. The relevant interaction variables are between the direction and speed of the wind, the direction and speed of the sea current, and the wave direction and wave height.

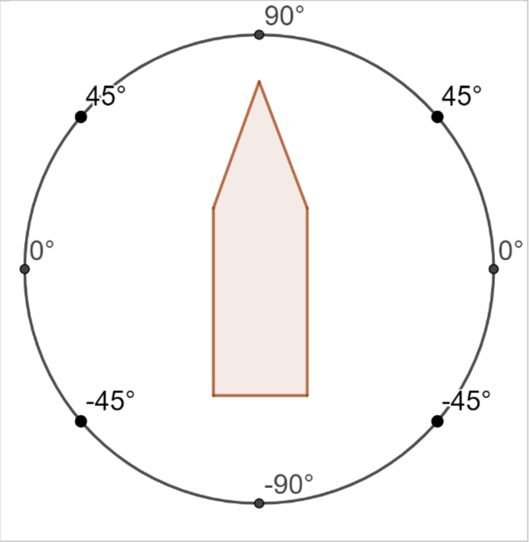

Figure 3. Illustration of how directional variables are transformed.

While the interaction variables capture much information of the directional effects of the forces, they do not necessarily capture all the effects. The interaction terms give wind and sea currents directly from the sides a neutral weighting as they are multiplied by zero, but the vessels still need to steer slightly into the sea current to avoid going off course. Wind may also lead to a slight tilt which can increase hull drag or propeller slip. These effects are captured by retaining the variables representing the weather forces (wave height, wind speed and current speed).

4.2.4 Cargo and draft

The noon reports contain variables for both cargo weight, draft, and load status. Load status and cargo weight provide essentially the same information, and draft and cargo weight are also highly correlated.

Accordingly, observations with conflicting values in these columns were discarded as likely incorrect, and a relatively high amount of missing draft values were imputed using a simple linear regression model with cargo weight as the predictor. Jia et al. (2019) performed a similar regression, where they predicted cargo using draft as a predictor and achieved accuracy of 91%, which indicates our regression should be sufficiently accurate. We opted to keep only the draft variable to keep our model comparable to most previous work. Nevertheless, due to our overarching goal of providing actionable results, our estimated regression equations are included in Appendix A. The equations allow for simple conversion between cargo and draft for each vessel class.

With all filtering procedures carried out, and as many missing values salvaged as possible, there were still 2,247 missing values left. These could not remain missing as some of our models could handle missing input values. The distribution of these across the predictors are shown in Table 4 below. These missing values were consequently imputed by replacing them with the mean value of their respective columns to reduce their influence on the predictions.

Remaining missing values	
Sea surface salinity	284
Trim	247
Draft/trim interaction	247
Sea surface temperature	198
Current speed	198
Current speed/direction interaction	198
Mean wave period	121
Wave height	121
Wave direction/height interaction	121
Wind speed	43
Wind speed/direction interaction	43
SUM	**1,821**

Table 4. Distribution of remaining NA values that were imputed before model training.

The pre-processed datasets now contain 1,672 rows of Handysize noon reports and 4,720 rows of Supramax noon reports. Summary statistics for the final selection of input variables for the two vessel designs are shown in Table 5 below. We find that most variables have similar values. Most notable is the difference of 1.9m in mean draft between the smaller Handysize and larger Supramax designs.

Handysize - Descriptive statistics of input variables (N = 1672)					
Feature	Unit	Mean	St. Dev.	Min	Max
Speed over ground	kts	11.5	1.3	7.1	15.0
Draft	m	9.1	1.5	5.1	10.7
Trim	m	0.8	0.8	-2.1	3.1
Latitude	deg.	13.9	23.7	-36.8	62.1
Longitude	deg.	7.2	85.3	-179.6	179.8
Time since dry docking	years	2.7	1.2	0.04	4.8
Vessel age	years	2.8	1.1	0.1	5.4
Sea surface temperature	°C	23.2	5.9	-0.4	33.3
Sea surface salinity	psu	35.2	1.9	7.1	40.0
Mean wave period	sec.	8.0	2.1	2.5	15.1
Wave height	m	1.9	0.9	0.1	7.0
Wave height/direction interaction	m · deg. forwards	-3.9	109.6	-631.5	373.6
Wind speed	kts	12.8	5.7	0.9	37.5
Wind speed/direction interaction	kts · deg. forwards	-47.9	824.1	-3,033.0	2,402.7
Current speed	kts	0.4	0.4	0.002	3.6
Current speed/direction interaction	kts · deg. forwards	0.4	34.4	-201.2	280.6
Supramax - Descriptive statistics of input variables (N = 4720)					
Feature	Unit	Mean	St. Dev.	Min	Max
Speed over ground	kts	11.6	1.5	7.0	15.0
Draft	m	11.0	2.5	5.5	13.6
Trim	m	0.8	0.9	-0.1	4.2
Latitude	deg.	7.4	24.5	-55.6	62.6
Longitude	deg.	32.4	82.6	-180.0	179.2
Time since dry docking	years	2.5	1.6	0.000	5.0
Vessel age	years	3.6	1.7	0.03	7.7
Sea surface temperature	°C	24.0	6.1	1.4	32.8
Sea surface salinity	psu	34.9	1.9	5.6	40.6
Mean wave period	sec.	7.9	2.1	2.1	15.9
Wave height	m	1.9	0.9	0.1	7.0
Wave height/direction interaction	m · deg. forwards	-7.8	106.5	-516.1	424.1
Wind speed	kts	12.4	5.9	0.2	40.6
Wind speed/direction interaction	kts · deg. forwards	28.2	781.6	-2,830.8	2,655.1
Current speed	kts	0.5	0.4	0.005	3.7
Current speed/direction interaction	kts · deg. forwards	0.3	37.4	-222.1	319.7

Table 5. Descriptive statistics of input variables for Handysize and Supramax vessels.

4.2.5 Standardization and train-test split

The variables in our dataset have varying ranges and signs. For some machine learning models, this can lead to variables with higher values being weighted unproportionally and for the models to become unstable and converge slower (Jaitley, 2018). To prevent this, we standardize all variables to zero mean and unit variance, which is achieved by subtracting the mean and then dividing by the standard deviation (Gkerekos et al., 2019). The transformation is shown in equation (8).

$$x' = \frac{x_i - \mu}{\sigma} \tag{8}$$

An alternative approach would be min-max scaling which maps all values between 0 and 1. However, scaling based on the maximum values makes the approach sensitive to outliers and other high values in the predictors (Aggarwal, 2015, p. 37), and could become problematic due to the diverse nature of our predictors.

Training and selecting models based on their performance on the test data directly, or simply knowing their performance on the test data, can lead to overfitting (Aggarwal, 2015, p. 335), as the models to some degree become tailored to the specific data. To compute an unbiased measure of the models' performance, we instead split the data into a train and a test set, where the test set remains unseen until the final performance measurement. There is no single optimal ratio between the train and test data. Literature usually defines ratios between 50-50 and 80-20 as common (Brownlee, 2020b), and comparable studies to ours have used 70-30 (Jeon et al., 2018) and 80-20 (Du et al., 2019). Based on this, we find 70-30 to be an appropriate ratio. When splitting into a train and a test set, random sampling is often used to avoid autocorrelation between adjacent rows. However, random sampling may still lead to a disproportionate distribution of values of the dependent variable, and consequently, a sub-par trained model. To ensure our train and test sets both are randomly sampled and contain a proportionate distribution of the dependent variable's values, we use a splitting function that randomly samples from different quantiles of the target variable.

4.3 Parameter tuning and model training

When tuning a machine learning model, it is the model parameters we are trying to optimize. These parameters decide the features of the algorithms, such as how many hidden layers an

ANN should have, or the penalty value of a shrinkage model. There are several methods to decide which combinations of parameters to test. The two most common are grid searches that test all possible combinations of a pre-set list of values, and random search which tests randomly selected configuration within a predefined range. With the high dimensionality of our dataset, the computational load of the grid searches can become substantial. Random search tends to perform just as well or better than grid search at a fraction of the computational cost (Bergstra & Bengio, 2012), so we will apply the random search method in our tuning approach.

The underlying principle of machine learning algorithms is that they improve by learning from their achieved predictive performance for their given parameters. Since the test data set must remain untouched, we require an additional set for measuring accuracy during model training, commonly referred to as a validation set. There are several methods available for this, including holdout, cross-validation, and bootstrap. Holdout implies a single validation set, and while the method is computationally fast, it is subject to biased results if the train-test split is not representative of the complete data. The bootstrap method selects several test sets randomly, but may still lead to biased results if the same data is selected as test data several times. It does, however, generally lead to lower variance than cross-validation (Abraham, 2017).

Cross-validation (CV) is sometimes split into leave-one-out CV and k-fold CV. Leave-one-out CV loops over all observations and uses each observation once as a test set with the remaining data as the train set. With the number of observations and variables in our data, this method becomes computationally infeasible. K-fold CV uses the same principle as leave-one-out, but instead of using single observations, it groups them into subsets ("folds") of equal size without replacement. More folds can increase the predictive accuracy but is more computationally costly. A typical number of folds is 10 (Aggarwal, 2015, p. 336). The reduced exposure to bias by avoiding replacement makes K-fold an attractive method for our purposes, and as such, we will apply this method with the typical 10 folds. As our cross-validation implementation creates train and validation folds only from within the training set, the test set remains separate and unseen.

5. Results and discussion

5.1 Results of models with third-party weather data

5.1.1 Model selection

To determine which of the applied machine learning algorithms achieve highest predictive performance, each algorithm was fitted to the training data separately for the two vessel designs, and their performance and computation times were measured on the previously unseen test set. All metrics are thus based on out-of-sample predictions. The models are trained and tested on the same dataset with the same dependent variable, so the RMSE scores can and will be used to determine the best-performing model. The results are shown in Table 6.

Model comparison for Handysize					
Model	sMAPE (%)	R^2 (%)	RMSE	nRMSE (%)	Train duration (sec)
Linear Regression	10.33	56.29	2.047	66.05	2.2
Neural Network	7.08	75.26	1.540	49.69	9.2
Extra Trees Regression	4.46	86.83	1.123	36.25	22.1
Random Forest	6.29	80.17	1.378	44.49	44.9
LASSO	10.30	56.54	2.041	65.85	0.8
Ridge	10.29	56.39	2.044	65.97	0.7
SVM Poly	9.86	55.73	2.060	66.47	3.3
SVM Radial	6.98	70.88	1.670	53.91	4.6
GP Poly	9.35	43.23	2.332	75.27	8.4
GP Radial	7.75	73.61	1.590	51.32	8.3
Cubist	5.21	82.37	1.300	41.94	61.9
Model comparison for Supramax					
Model	sMAPE (%)	R^2 (%)	RMSE	nRMSE (%)	Train duration (sec)
Linear Regression	7.77	65.50	2.241	58.72	3.0
Neural Network	6.42	74.00	1.946	50.97	9.0
Extra Trees Regression	3.62	87.60	1.331	35.20	16005.0
Random Forest	6.47	75.85	1.875	49.13	188.4
LASSO	7.81	65.18	2.251	58.99	1.0
Ridge	7.83	64.88	2.261	59.24	1.0
SVM Poly	6.01	65.12	2.253	59.04	869.1
SVM Radial	5.20	78.58	1.766	46.27	60.7
GP Poly	5.81	77.46	1.811	47.46	274.2
GP Radial	15.31	-0.78	3.830	100.36	278.7
Cubist	4.96	83.11	1.568	41.09	344.4

Table 6. Model performance comparison for Handysize and Supramax vessels, using third-party weather data.

For the smaller Handysize dataset with N = 1,672, the Extra Trees (ET) model achieved the clearly best performance with an RMSE score of 1.123. ET was also best scoring model for predicting ship speed in the comparison performed by Abebe et al. (2020). The ET model we trained explained 86.8% of the observed variance and had a sMAPE of 4.5%. The Cubist was ranked second with an RMSE score of 1.300, followed by the Random Forest and Artificial Neural Network models. On the lowest ranks, we find the linear models LR, Ridge and LASSO, as well as SVM polynomial and lastly GP Polynomial, all scoring within an RMSE interval of 2.041 to 2.332.

Continuing with the results from the Supramax dataset with N = 4,720, we also find that the ET model has the highest performance, achieving an RMSE of 1.331. The model's R squared was 87.6%, and its sMAPE was 3.62%. Cubist again ranked second, but this time the SVM Radial model ranked third. The worst performing models for Supramax were GP Radial and the linear models. We found the results from the SVM and GP algorithms to be highly unpredictable. Their performance appears tightly knit with the data structure compared to the kernel type used and the selection of hyperparameters (Pedregosa et al., 2011). The one negative R squared score means that the model, in this case, was less accurate than a hypothetical model predicting the mean test set value for all observations.

In the data acquisition section, we discussed the data uncertainty of noon reports as a data source and referred to literature claiming that it could be between 1% and 16%, depending on ship types. This means it is uncertain whether the models can be significantly improved given the data available.

The ET models have a high training time compared to the other algorithms, with 22.1 seconds for the Handysize dataset and closer to four and a half hours for the Supramax dataset. The significant difference in training duration between the two datasets is mainly due to IMO numbers being used as a factor variable to capture vessel-specific effects and that Supramax has considerably more unique IMO numbers. With dummy variables for IMO numbers, this translates to high dimensional data and an exponentially increasing number of branches. All performance metrics presented were from a model with input variables described earlier in Table 2 and with a 10-fold cross-validation training procedure with a random hyperparameter search. These training times were achieved using a personal computer with a 6-core, 4.1 GHz CPU and 32GB RAM.

5.1.2 Feature importance

Determining feature importance in machine learning is not always as straightforward as with models such as OLS Regression. While we in regression can use the model's coefficients to determine importance, our most accurate machine learning model is in the form of a decision tree. One can use several feature importance functions, but we found that they generally gave inconsistent results that could change drastically between runs with similar model configurations. An alternative to the feature importance functions is to use the correlation coefficients between the predictors and the dependent variables. These will not pick up model-specific variations in feature importance or possible non-linear relationships the models might find. However, based on their robustness, we considered them to give the most accurate representation of the true variable importance. The plot of correlation coefficients can be seen in Figure 4.

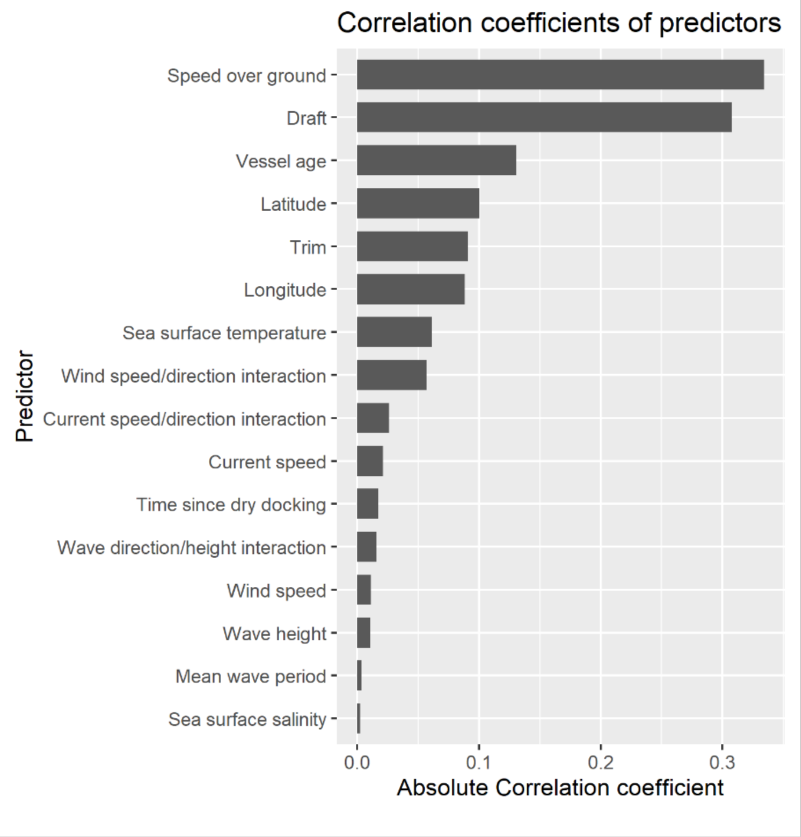

Figure 4. Absolute correlation coefficients of all predictors.

Studying the correlation coefficients, we see that SOG appears to be the most important predictor. As described earlier, the relationship between speed and fuel can be roughly approximated as a cubic function of speed, at least near the common operating speeds. Therefore, we would expect SOG to be the most important predictor. The draft appears to be the next most important factor. Deeper drafts correspond to more of the hull being exposed to water friction and resistance, so it is as expected that a higher value for the draft is correlated with more drag and higher fuel consumption. Since cargo weight and draft are almost perfectly correlated, the former's effect on consumption is captured by the draft variable. The third most important variable is vessel age. As described previously, there are several factors dependent on age that may correlate with fuel consumption. Among other factors, we described how newer engines could be around 10% more efficient, and as such, it is expected that vessel age has a significant effect.

Latitude and longitude also seem highly correlated with fuel consumption. However, since comparable models in the literature achieve high accuracy despite rarely including these variables, we consider it unlikely also in our case that the variables truly are highly influential. Therefore, their relatively high correlation with fuel consumption underlines an important point when interpreting our correlation matrix, which is that correlation does not necessarily imply causation. We suspect that most of the observed correlation will disappear when accounting for other variables, but the high correlation nevertheless justified a closer examination of their impact on prediction accuracy. To test their impact, we trained our models both with and without the latitude and longitude variables while holding all else equal. The results, available in Appendix C, showed an improvement in RMSE from 1.383 to 1.331 when adding the variables. They were therefore kept. We hypothesize that their remaining predictive power results from local effects not captured by other variables, such as ocean depth, local regulations, areas necessitating maneuvering, or other unknown influences.

One interesting finding from the figure is that the variable describing the time since the last dry docking has a non-negligible correlation with fuel consumption. This variable aims to pick up how hull fouling affects fuel consumption. As described in greater detail in section 4.1.3, hull fouling can significantly increase fuel consumption. The correlation we find here indicates that our variable may at least partly capture this effect. However, the variable is potentially correlated with vessel age, which could mean it captures some effect actually caused by the ship's aging. We trained our models with and without the dry docking variable to further test the effects and compared the RMSE values. With dry docking included, the best achieved RMSE was 1.331, and with dry docking excluded, the best RMSE was 1.382. The comparison results are available in Appendix C. This indicates that the variable leads to minor improvements in prediction accuracy.

We only had data on the last dry docking time for around half of the noon reports, and the rest were imputed with an uncertain method based on typical dry docking intervals. Other key data we were missing included underwater hull cleaning and accurate data on time spent stationary by vessels in our dataset. Based on this, larger improvements in accuracy could not be expected. As we know, the effect of fouling can be significant on fuel consumption, and further work might use more complete data sources to better capture the effect of hull fouling.

Salinity and temperature show some correlation to fuel consumption. We previously hypothesized that the variables might capture some of the effects from additional biofouling in warmer and more saline waters, as well as changes in the frictional resistance due to water viscosity and density. As before, we tested the hypothesis by running the model with and without the variables. However, we found worse performance with the variables added this time, with the RMSE increasing from 1.331 to 1.343. Comparison results are available in Appendix C. There are a few reasons why accuracy does not always increase when adding variables correlated to the dependent variable. One reason could be that the model exhibits overfitting, where it finds relationships in the training data that are not present in the unseen data. Overfitting may happen if the added correlation is low, as in this case. Another reason could be that by increasing the complexity of the model, the current hyper-parameter tuning length may no longer be sufficient to model the most important predictors as accurately as before. Based on the comparison results, the models could not make use of the low correlation present, and we decided to exclude the variables from our models.

The interaction terms of both waves and wind seem much more important than corresponding variables measuring their magnitude. The direction of the weather forces alone may not necessarily explain resistance adequately as they cannot differentiate between low and high-speed winds and sea currents, or low and high waves. However, the factor directions combined with their force can determine whether and how much resistance the vessel is subject to. This is the effect we hoped to capture by creating the interaction terms, and the figure indicates that we may have been successful in doing so. As described earlier, weather forces perpendicular to the vessel heading can still increase the resistance by requiring the vessel to steer into the forces to stay on course and a possible tilting that might increase hull drag and propeller slip. As such, it was as expected that the weather forces still retain some correlation with fuel consumption.

5.1.3 Prediction accuracy analysis

Figure 5 and Figure 6 illustrate the predictive accuracy of the models. These show that the models can explain most of the variance in the actual values, and that more significant prediction errors often occur on more extreme observations.

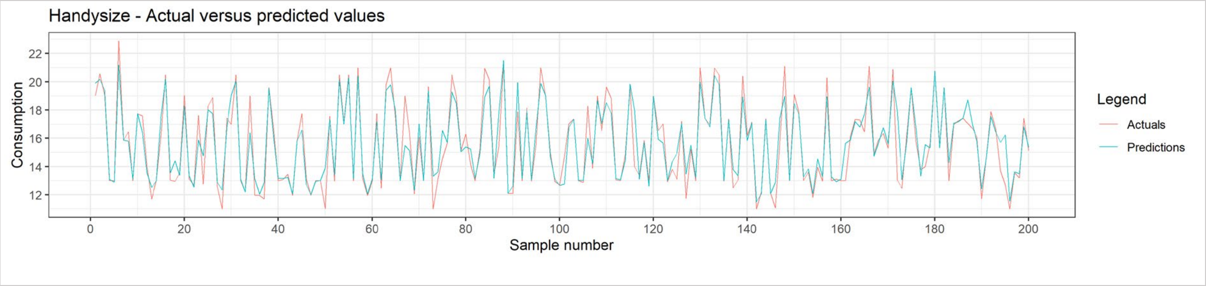

Figure 5. Handysize: Actual observations of daily fuel consumption plotted against the predictions from our best performing ML models. The plot includes a random sample of 200 observations from the test set.

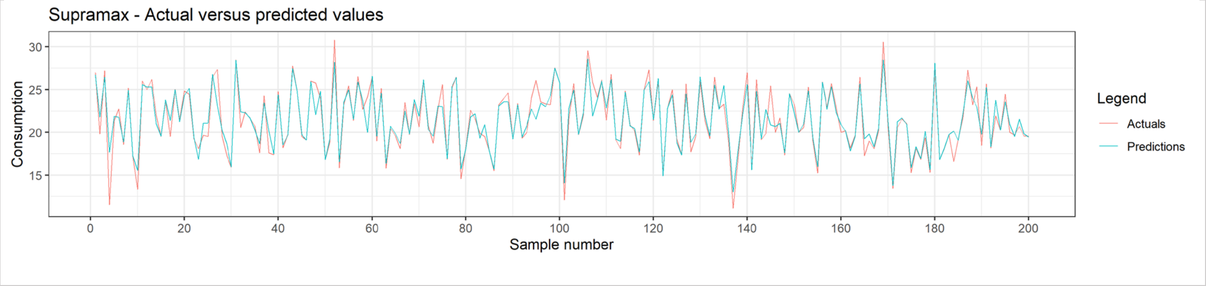

Figure 6. Supramax: Actual observations of daily fuel consumption plotted against the predictions from our best performing ML models. The plot includes a random sample of 200 observations from the test set.

For the next part of the analysis, we study how the prediction errors are distributed, both in sum and in relation to varying predictor values. We start with the Supramax dataset. In Figure 7 below, we see that the prediction error density resembles the normal distribution. The mean prediction error is close to the mode prediction error. Figure 8 shows how absolute prediction errors are distributed along with varying values for SOG. Each dot in these types of plots represents one prediction of the observations in the test set, and the red line is the rolling mean of the prediction errors. We find that the mean errors are relatively uniform for different values of SOG, with the lowest values near 13 knots. There are two likely reasons for this. The first is that speeds around 13 knots are the most common in the dataset, as can be seen by the dots, meaning that the model has trained on more observations in this range. Another reason could be that 13 knots often corresponds to standard operation in the open seas, which is easier to predict than, e.g., the more unpredictable patterns found while navigating a canal.

For draft, shown in Figure 9, we find the model displays sporadic differences in accuracy that are not necessarily correlated with the density of the observations. Some of the most accurate intervals lay between 10 to 13 m, but the differences are, for the most part, minor. For prediction errors by wave height, shown in Figure 10, we find that the prediction errors display a slight increase in correlation with increasing wave heights. With stronger forces in motion, it would be expected with some reduction in accuracy. The errors peak at around 3.5 m, but we see that there are much fewer observations in this area. This means that the apparent lower accuracy could be due to a few outliers and that it might have changed considerably given more test data.

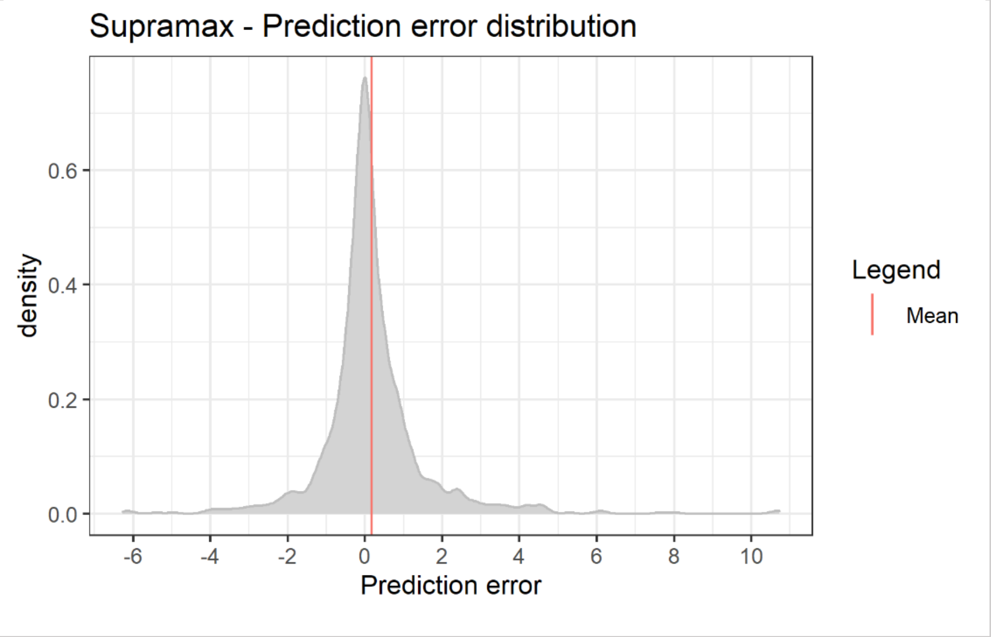

Figure 7. Prediction error distribution on fuel consumption reported in Supramax noon reports.

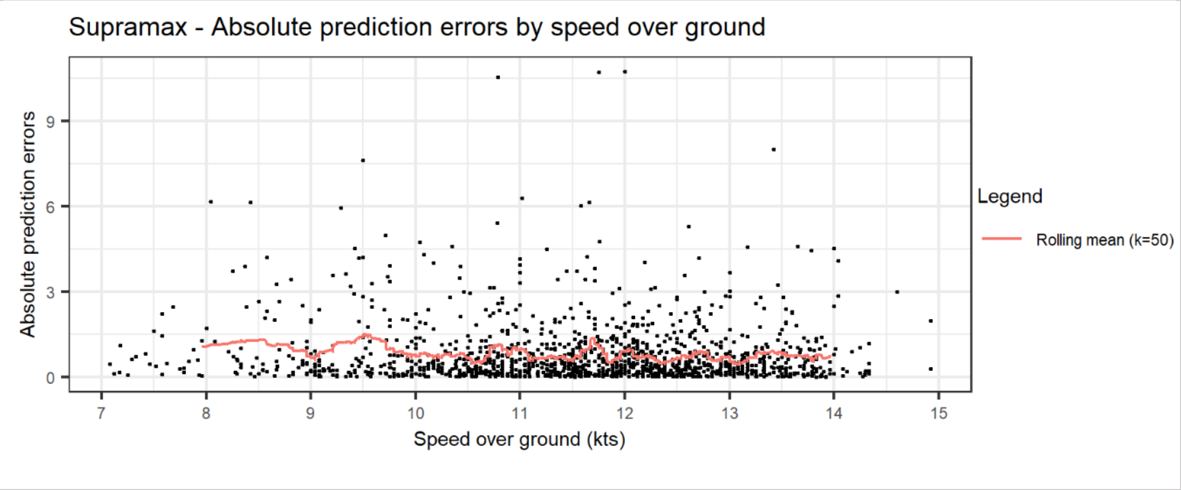

Figure 8. Absolute prediction errors on fuel consumption reported in Supramax noon reports, sorted by speed over ground, with a rolling mean of $k = 50$.

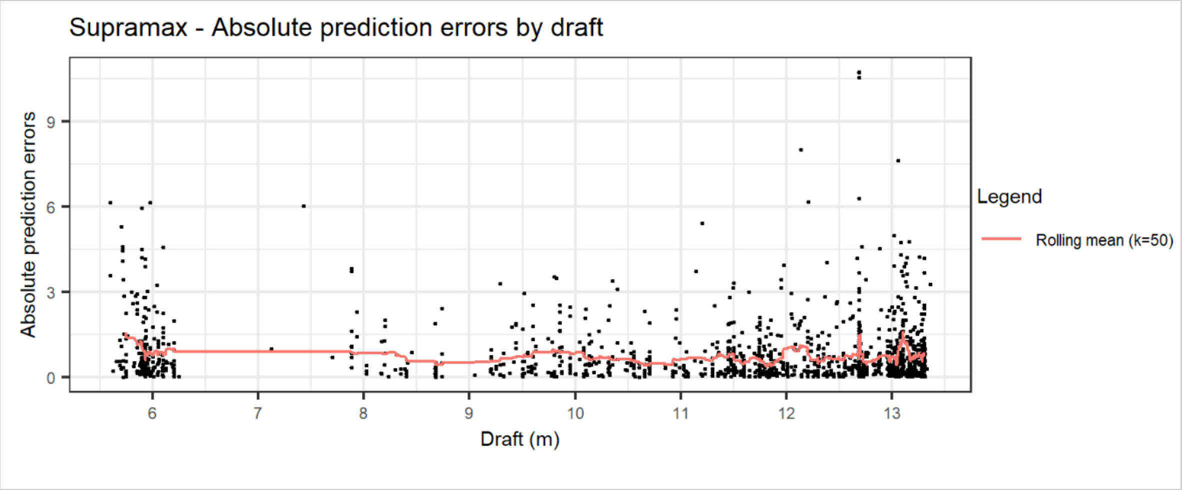

Figure 9. Absolute prediction errors on fuel consumption reported in Supramax noon reports, sorted by draft, with a rolling mean of $k = 50$.

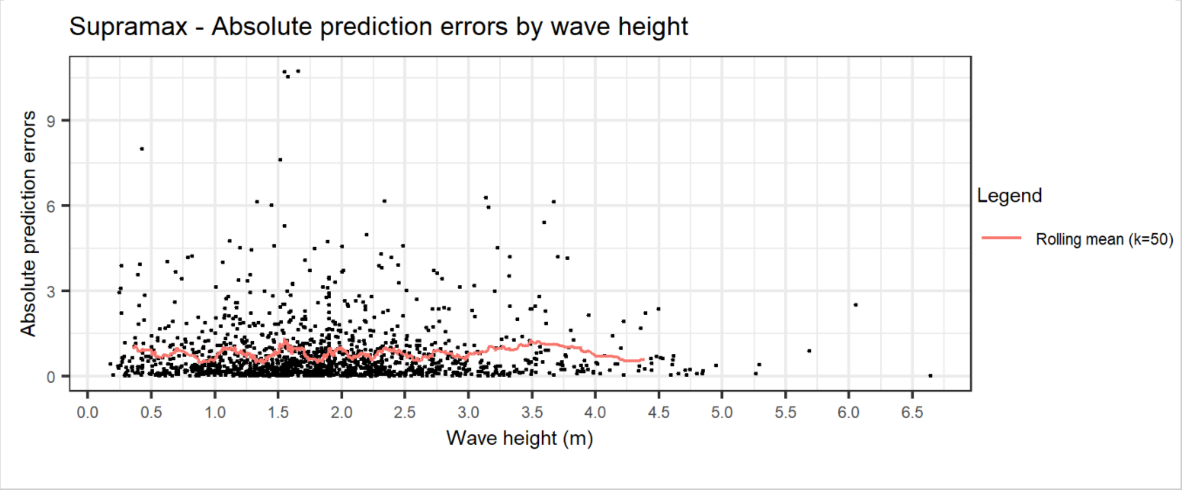

Figure 10. Absolute prediction errors on fuel consumption reported in Supramax noon reports, sorted by wave height, with a rolling mean of $k = 50$.

The same set of figures for the Handysize dataset are available in Appendix D. For Handysize vessels, we found some of the same trends as we did with the Supramax data above. However, there are noticeably fewer observations in the Handysize dataset and a slight increase in variance in the prediction errors. For the mean prediction errors by SOG in Figure 26, there appears to be a slight increase in correlation with increased SOG. For the draft in Figure 27, we find that the model seems most accurate between 9.5 and 10.5m, which corresponds to sailing in laden for this design. Lastly, in Figure 28, wave height appears similar to Supramax but with fewer signs of an increasing trend in errors.

5.1.4 Cumulative voyage prediction errors

So far, we have analyzed our models' prediction errors in detail on the level of individual predictions. In practice, the model may be more useful for generating predictions for an entire voyage or route, a use case which we will demonstrate in the case studies in chapter 6. For this purpose, we would also like to quantify prediction errors in the voyage level to see whether daily prediction errors tend to cancel out or accumulate throughout the voyage.

In Figure 11, noon reports have been grouped by individual voyages[4]. Since we want to compute out-of-sample prediction errors, observations from the training data have been excluded before calculating prediction errors and voyage duration. This means the voyage durations presented may be artificially lower than the actual durations. For each voyage, the error term is calculated as the difference between the sum of actual fuel consumption on a given voyage and the sum of the predicted consumptions. Voyages are then grouped by their duration to show the distribution of voyage level errors for voyages of different durations.

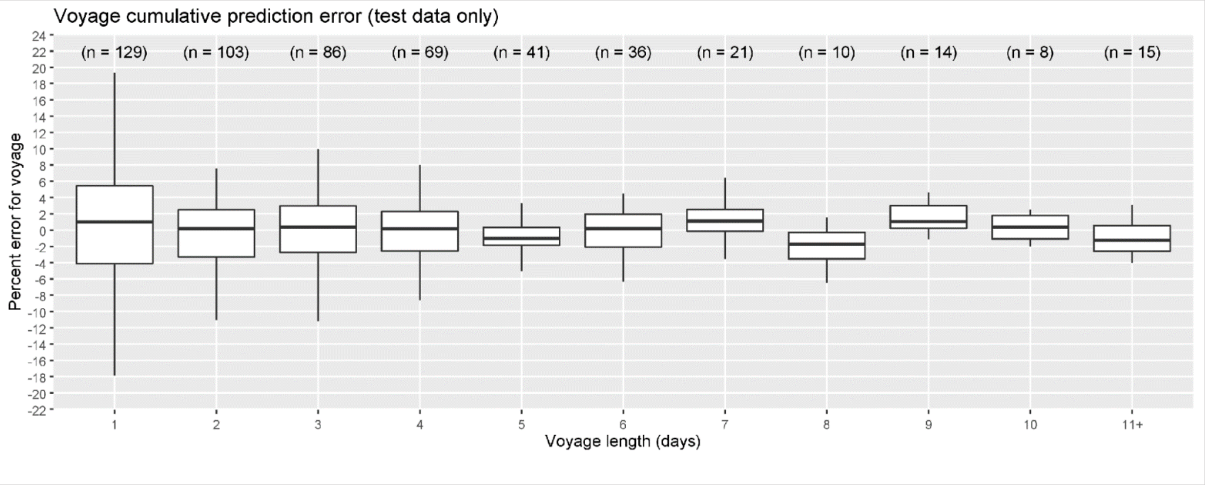

Figure 11. Distribution of prediction errors on the voyage level for voyages of different durations. Errors from both the Supramax and Handysize models showed similar trends, so they have been pooled in this plot. The boxes show median error, 25-75% quartiles, and whiskers to the closest of min/max/1.5 × Inter-quartile range.

The figure shows that although predictions are largely unbiased regardless of voyage duration, the errors seem to quickly cancel out, resulting in a marked improvement in accuracy. For voyages four days or longer, there are almost no voyage-level errors above 4%. Table 7 presents the performance metrics on total fuel consumption from voyages that are a minimum of 7 days long. We achieve an R squared of 99.5% and 98.6% for Handysize and Supramax designs, respectively. For comparison, Abebe et al. (2020) computed their performance

[4] Noon reports within four days of each other and with matching IMO numbers were considered as part of the same voyage or route.

measures for a single route of different vessels and achieved an R squared of 98.5% with their best model.

Predictive accuracy for fuel consumption on voyages minimum seven days long

Vessel design	Voyages	sMAPE (%)	R^2 (%)	RMSE	nRMSE (%)
Handysize	18	1.80	99.49	3.147	6.92
Supramax	50	2.17	98.64	5.425	11.53

Table 7. Performance metrics for total fuel consumption of all voyages at least seven days long. All metrics are from out-of-sample predictions. The total number of individual predictions is 162 for Handysize and 463 for Supramax.

Based on these results, we can conclude that the model can generate highly accurate fuel consumption predictions if the weather along a route is known. Knowing the weather in advance is, of course, often not realistic. In the case studies below, we will therefore demonstrate how our models may still be used to estimate consumption for a given voyage without prior knowledge of weather.

5.2 Impact of third-party weather data on prediction accuracy

For our later weather margin analysis, we needed the model to be trained on only third-party weather variables that could be sampled many years back in time. In addition, we had to remove all weather data from noon reports that could otherwise have captured some of the effects of the third-party weather variables. In this section, we want to study how these requirements affected the prediction accuracy of the model. We measure this effect by comparing our models with baseline models that were not subject to these requirements. The optimal feature selection for the baseline models is seen in Table 8 below.

Final selection of input parameters		
	Parameter	**Unit**
Noon report data		
1	Speed over ground	kts
2	Draft	m
3	Trim	m
4	Latitude	deg.
5	Longitude	deg.
6	Imo number	-
7	Wind speed	kts
8	Wind speed/direction interaction	kts · deg. forwards
9	Swell height	m
10	Swell height/direction interaction	m · deg. forwards
Clarksons' World Fleet Register data		
11	Time since dry docking	years
12	Vesssel age	years
13	Eco Propellar	T/F
14	Main engine model	-

Table 8. Final selection of features for baseline models without third-party weather data.

The modeling procedure for the baseline models was kept similar to the models with third-party weather data to ensure a fair basis of comparison. We previously presented the performance of the models with third-party weather data in Table 6. The performance for the baseline models is presented in Table 9 below. Based on the RMSE metric, Extra Trees (ET) was the best performing algorithm on both vessel designs for the baseline models. Since the models with and without third-party weather data have the same dependent variable, the RMSE metric can also compare the models' relative performance. The results show that predictions for both vessel designs are similar across the models. For the Handysize dataset, the model with third-party weather data performs slightly worse with an RMSE of 1.123 versus 1.089 for the baseline model. The model with third-party weather data also performs slightly worse on the Supramax dataset with an RMSE of 1.331 versus 1.284 for the baseline model.

The changes in accuracy are mainly caused by the replacement of weather variables from noon reports with third-party data. Noon report variables should reflect average experienced values since the last noon report. This could make them more accurate than our third-party weather variables, which retrieve weather data from a virtual midpoint location between two given noon reports. On the other hand, the third-party weather data we add consists of more variables such as sea currents and the directions of the weather forces relative to the vessel heading.

With more information, the prediction accuracy may increase. Nevertheless, the total effect is a minor reduction in accuracy compared to the baseline models.

Model comparison for Handysize					
Model	sMAPE (%)	R^2 (%)	RMSE	nRMSE (%)	Train duration (sec)
Linear Regression	10.73	52.24	2.139	69.04	3.3
Neural Network	6.82	76.06	1.515	48.88	8.0
Extra Trees Regression	4.13	87.62	1.089	35.15	19.1
Random Forest	5.64	82.16	1.308	42.20	39.6
LASSO	10.75	52.38	2.136	68.94	0.8
Ridge	10.77	52.02	2.144	69.20	1.1
SVM Poly	9.92	49.34	2.203	71.11	174.0
SVM Radial	7.39	69.25	1.716	55.40	3.5
GP Poly	8.40	64.78	1.837	59.29	8.4
GP Radial	7.48	73.32	1.599	51.60	8.9
Cubist	5.07	82.97	1.277	41.23	46.5
Model comparison for Supramax					
Model	sMAPE (%)	R^2 (%)	RMSE	nRMSE (%)	Train duration (sec)
Linear Regression	8.44	60.04	2.412	63.19	3.2
Neural Network	6.82	71.15	2.049	53.69	9.7
Extra Trees Regression	3.33	88.68	1.284	33.63	14598.3
Random Forest	5.82	79.28	1.737	45.50	161.4
LASSO	8.46	60.01	2.413	63.22	1.0
Ridge	8.48	59.93	2.415	63.28	1.0
SVM Poly	5.40	74.44	1.929	50.54	599.6
SVM Radial	5.37	76.13	1.864	48.83	70.7
GP Poly	8.49	59.88	2.417	63.31	274.1
GP Radial	15.31	-0.78	3.830	100.36	277.1
Cubist	4.56	84.90	1.483	38.85	324.0

Table 9. Model performance comparison for Supramax and Handysize vessels, using only noon report weather variables.

Table 9 showed that the baseline models' results were comparable to the models using third-party weather data. While the Extra Trees model achieved the best single performance based on noon report data, the majority of the models performed better when trained on third-party weather data. In light of the above discussion, we hypothesize that the higher-resolution third-part weather data only available in 2019 and later could have changed the outcome had it been available for the entire period covered by the noon reports. This dataset had a temporal resolution of 1 hour, as compared to the 24-hour resolution of the alternative dataset. We compared the predictive accuracy of models trained only on data from 2019 and later to explore this possibility. Although this almost halves the available training data, we found improved prediction accuracy compared to models trained on the entire dataset, as shown in Table 10.

Comparison of models trained on different datasets

Data source	Noon reports only	With third-party weather data	With third-party weather data (2019 and later)
Model	RMSE	RMSE	RMSE
Linear Regression	2.412	2.241	2.156
Neural Network	2.049	1.946	1.666
Extra Trees Regression	1.284	1.331	1.299
Random Forest	1.737	1.875	1.755
LASSO	2.413	2.251	2.144
Ridge	2.415	2.261	2.177
SVM Poly	1.929	2.253	2.302
SVM Radial	1.864	1.766	1.535
GP Poly	2.417	1.811	2.163
GP Radial	3.830	3.830	1.769

Table 10. Comparative performance of models trained on different training data. Results in the rightmost column are from models trained on slightly more than half the number of observations compared to the other results.

The baseline ET model proved the most accurate overall and is therefore the preferred choice for retrospective predictions on the test set data. However, when using the baseline models, the only observations we have of the noon report weather variables are from when the trip took place. If we want to study how the weather would have been if the trip took place a week before, we would not know which values to use for the noon report weather variables. Since we aim to investigate weather margins more generally, we depend on being able to study fuel consumption under various historical weather conditions. This means that the baseline models do not fulfill our requirements for weather margin analysis. Moreover, as we will show in the case studies, we do not have enough training data to accurately predict consumption in unusually rough weather conditions. Therefore, we also consider the models based only on 2019 and later observations to be unsuitable for use in our case studies.

By training the models on the third-party weather data from the Copernicus database, we can simulate routes where we sample historical weather many years further back in time. This allows us to make more accurate estimations of the weather margins we want to study. Nevertheless, the comparison with the baseline models gives us a useful impression of how much accuracy had to forego to facilitate the sampling of decades of historical weather data, and overall, we found only a slight reduction in accuracy.

6. Case studies

Western Bulk has expressed a desire to understand better how seasonal weather patterns impact weather margins, thereby improving cost predictions related to forward pricing of cargos. Predictions would be particularly useful beyond the 10-day forecast horizon where weather forecasts can no longer provide reasonable accuracy. This section aims to use our trained models to explore in greater detail how weather impacts fuel consumption for a given route. To this end, we picked one voyage for each vessel type from our data, as shown on the map in Figure 12. The dotted lines are the actual paths taken by two ships traveling these routes. These are voyages commonly sailed by vessels in our dataset and serve as a representative example of applying the model to a real-world scenario. Western Bulk mentioned the North Pacific voyage specifically as a route with highly variable fuel consumption driven by strong weather effects and seasonality.

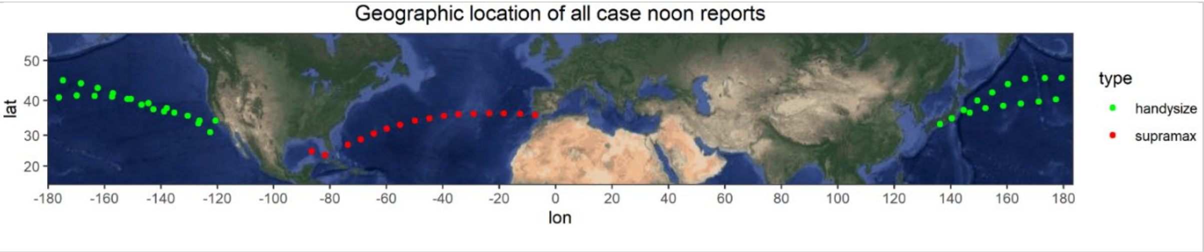

Figure 12. Map showing the actual paths taken by ships sailing the case routes. The Handysize case from Japan to North America is plotted with green dots, and the Supramax case from Gibraltar to Houston is plotted with red dots.

To study the routes in a broader weather context, we include a map in Figure 13 showing the annual wind power density. The northern hemisphere winter season is shown in the top panel, and the northern hemisphere summer season is shown in the lower panel. Red and white colors represent high power winds, while green and blue represent lower power winds. Comparing this map with Figure 12 shows that two case routes pass through the North Pacific and the North Atlantic oceans, respectively. The wind map shows these areas experience heavy winds during the winter seasons but much milder winds during the summer seasons.

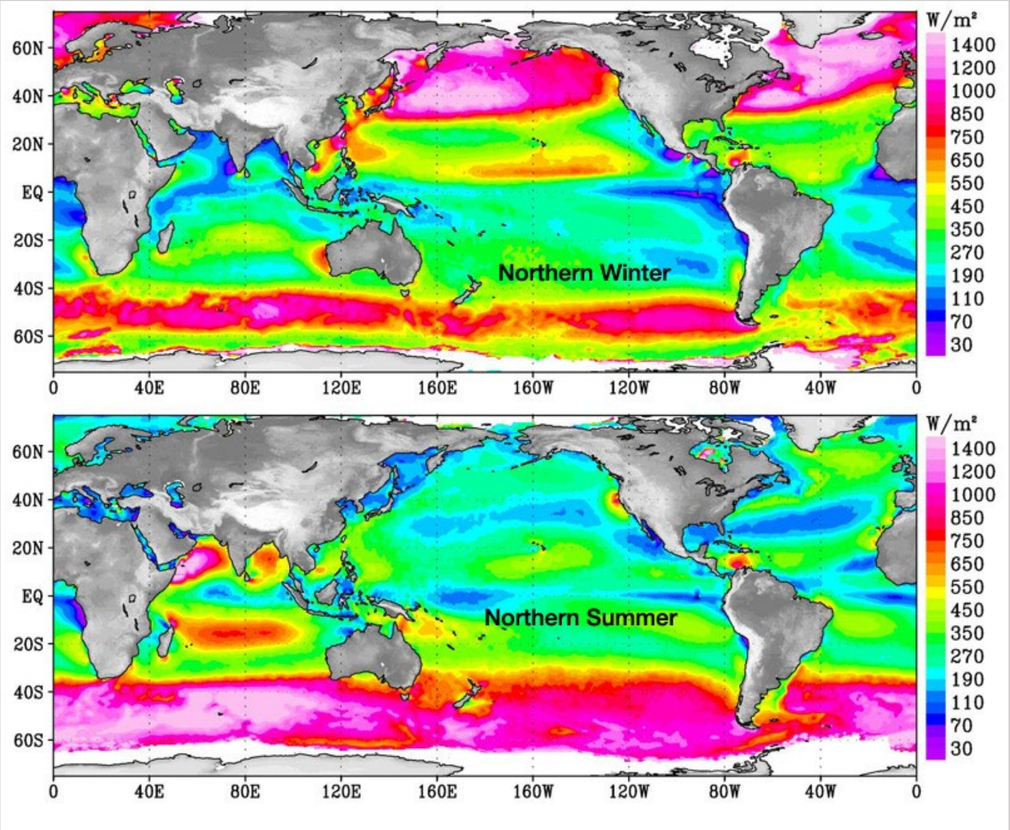

Figure 13. Annual wind power density. Top panel: Northern hemisphere winter. Bottom panel: Northern hemisphere summer season. Red and white colors represent high power winds, while green and blue represent low power winds.
Source: (ScienceX, 2008).

6.1 Procedure

Chart 2 shows the procedure we have followed for each of the two case routes. In this section, we will describe key steps and the reasoning behind them.

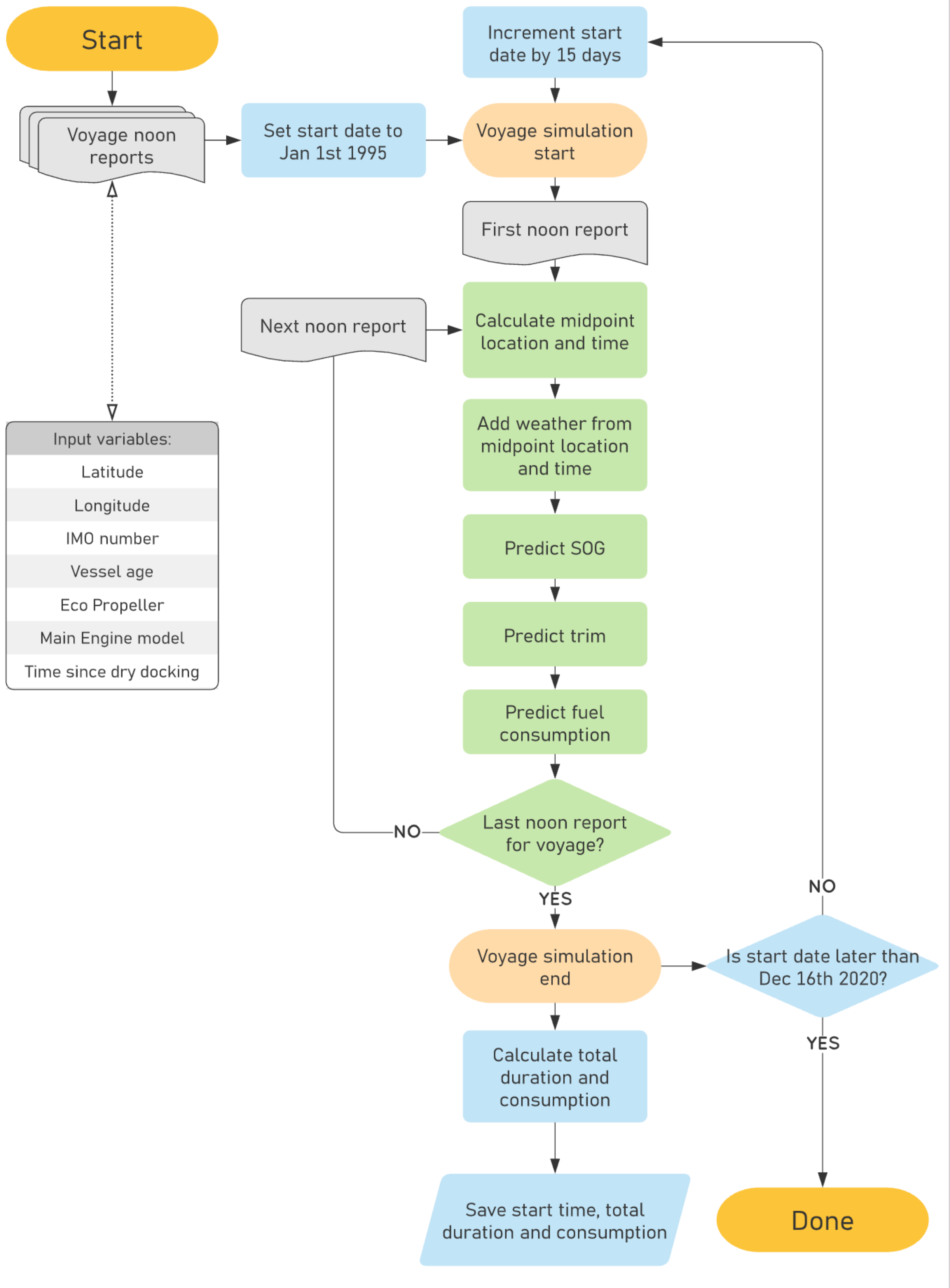
Start
Voyage noon reports
Set start date to Jan 1st 1995
Voyage simulation start
Increment start date by 15 days
First noon report
Next noon report
Calculate midpoint location and time
Add weather from midpoint location and time
Predict SOG
Predict trim
Predict fuel consumption
Last noon report for voyage?
NO
YES
Voyage simulation end
Is start date later than Dec 16th 2020?
NO
YES
Calculate total duration and consumption
Save start time, total duration and consumption
Done
Input variables:
Latitude
Longitude
IMO number
Vessel age
Eco Propeller
Main Engine model
Time since dry docking

Chart 2. Procedure for calculating weather margins for case routes.

For each case study, we have used a real voyage from our dataset as a starting point. All input variables shown in the figure were left unchanged. For example, this means that a vessel's age increases during the voyage but is identical for the first noon report of all simulated voyages (regardless of the simulated start date). The January 1st, 1995 start date for the first simulated voyage was chosen due to it being the earliest date for which we have historical weather data.

As previously described (see section 4.2.3), the weather for any given noon report should be retrieved from the temporal and spatial midpoint between itself and the preceding noon report. The spatial midpoint could easily be calculated as we knew the location of both the current and previous noon reports. However, determining the temporal midpoint posed more of a challenge because SOG and elapsed time since the previous report was unknown at this step due to our decision to make SOG dependent on weather conditions. The reasoning behind this decision will be discussed in the subsequent paragraphs. To solve this problem, we used the SOG from the preceding noon report as a best guess of the current SOG. This SOG was then used to calculate the time it would have taken to reach the spatial midpoint, which enabled calculating the temporal midpoint, and finally, to retrieve weather data for this midpoint.[5]

In the next step, we predict SOG based on the original input data and the weather conditions we just retrieved. Changing SOG based on weather conditions does have some drawbacks for our analysis. First, as described above, not knowing SOG when retrieving weather adds uncertainty to the exact time that should be used when retrieving weather data. Second, it means that our case study results now show the combined effect of changes to weather and SOG on consumption, rather than the isolated effect of weather.

However, there are also drawbacks associated with keeping SOG invariant to the weather due to the clear correlation between SOG and adverse weather. In particular, there is a correlation of -0.4 between wave height and SOG. One possible explanation for this is that strong adverse weather increases frictional resistance, sometimes increasing the power and rpm requirement to maintain SOG to beyond the engine's limits (Tillig & Ringsberg, 2019). In other words, if SOG is not reduced in adverse weather, our simulation results may be based on impossible combinations of SOG and weather.

In addition to involuntary speed loss, it is also likely that voluntary adjustments are made to the planned voyage depending on weather conditions. For instance, if there is severe weather

[5] Higher accuracy could have been achieved by retrieving weather along a smaller segment of the path, predicting SOG for this segment, then repeating this for the entire path between the location of the two noon reports. Unfortunately this approach proved computationally infeasible.

on a given route, the operator may postpone the voyage, or the captain may change course to avoid the worst parts of the weather or reduce speed to maintain safe operation. Some adjustments may, in part, be motivated by increased fuel efficiency. For instance, Du et al. (2019) has conducted a study on optimal SOG and trim for a given route configuration and weather state. They found that using optimal SOG or trim could lead to 7.5% or 5-6% fuel savings, respectively. As these features' effect on fuel consumption is interwoven, the simultaneous optimization of the two features enabled fuel savings of 8.25%.

Based on the above, adjusting speed depending on weather enables more realistic and representative results, and we believe these benefits outweigh the mentioned drawbacks. We aim to set a value for SOG that is realistic and representative for the given voyage and weather conditions. As we previously determined that the Extra Trees model performs well on our dataset, we trained such a model using the same input variables as our previous models, except for fuel consumption and trim. Fuel consumption was left out as a predictor because it is unknown when predicting SOG for these case studies. Trim will be discussed in the following paragraph.

As previously mentioned, trim and SOG can be jointly optimized to increase fuel efficiency. However, it is traditionally set based on trim tables that indicate trim based on speed and displacement (equivalent to draft; Du et al., 2019). Such a method of setting trim seems to approximate our data well, as the correlation coefficient between trim and draft is -0.78, and 0.22 between trim and SOG. Therefore, we will assume a causal relationship between SOG and trim, specifically in the direction of SOG → trim. This causal relationship explains why using trim as a predictor when predicting SOG would be inappropriate and why adjustments to SOG necessitates adjustments to trim to achieve realistic results in these case studies. As with SOG, we trained an Extra Trees model to predict trim, this time leaving out only fuel consumption as a predictor.

The SOG and trim prediction models were trained on the same training data as was used to train the consumption prediction models. A brief summary of these models’ performance on the test data is shown in Table 11.

	SOG		Trim	
Design	sMAPE	RMSE	sMAPE	RMSE
Supramax	1.28 %	0.2009	14.70 %	0.02713
Handysize	1.25 %	0.2008	9.60 %	0.03258

Table 11. Performance of the models predicting SOG and trim.

So far, we have estimated midpoint, retrieved weather, and predicted SOG and trim. The only remaining step in processing a single simulated noon report is predicting fuel consumption. The procedure for this was identical to predictions on our test set data, although we used our second-best Cubist models for this purpose due to its superior extrapolation capacity, which we will discuss in detail in section 6.4.

The above describes the processing of a single noon report, which was done sequentially for each noon report in a simulated voyage. The starting date of the voyage was then set to 15 days later before the entire process was repeated, for a total of 631 simulated voyages per case route.

6.2 Handysize case voyage

The selected route for the Handysize model starts in late April from Japan. From there, it sails in ballast across the North Pacific Ocean to the North American west coast, where it loads around 37,000 tonnes of cargo before returning to Japan. Originally, the average SOG during open-sea sailing was around 11 knots, for a total duration of 36 days.

From the 631 generated datasets, 47 were discarded due to missing weather data. The following is based on the remaining 584 datasets. Figure 14 shows the weather conditions encountered at different times, revealing a clear seasonal pattern with stronger wind and wave forces in the winter season. There is a slight increase in the wave and wind variance in the winter, but this change in variance appears minor compared to the change in means. A similar plot showing the less influential or less seasonal sea temperature, current, and salinity variables is shown in Figure 24 (Appendix B).

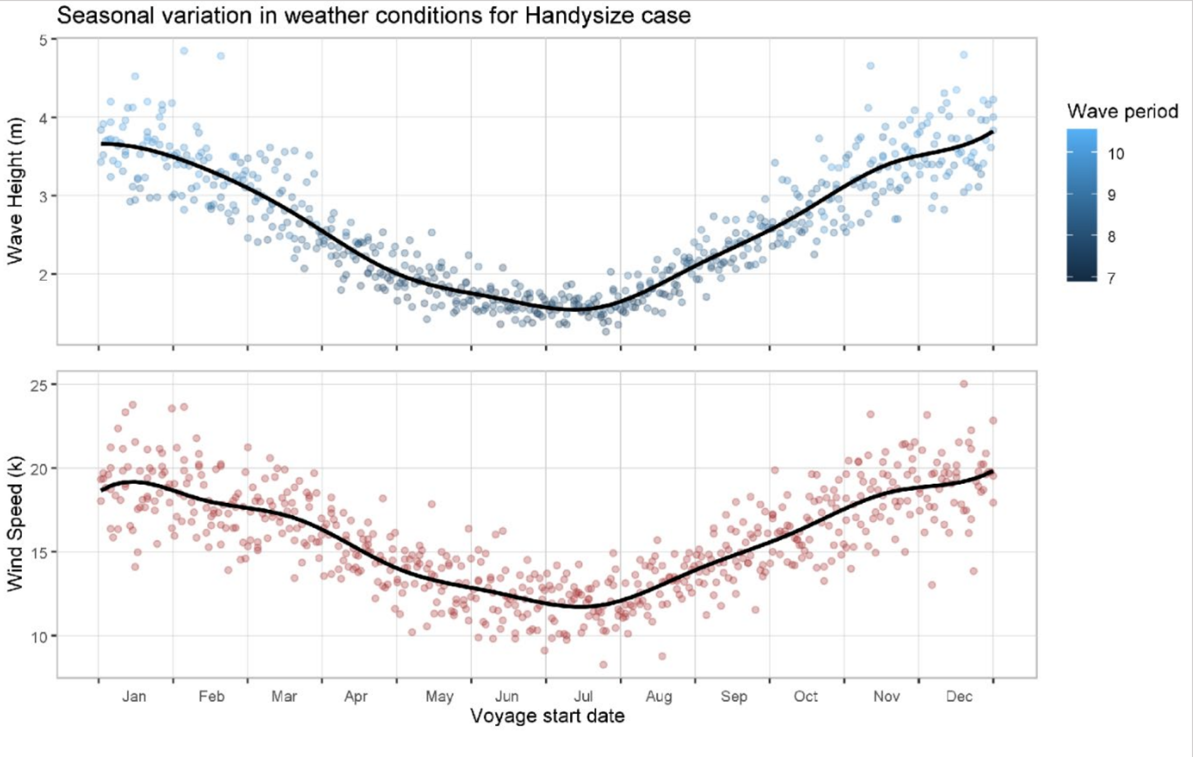

Figure 14. Weather conditions experienced along the route at different times of the year. Each observation shows the median weather encountered on a given voyage. The black line shows the estimated means.

Figure 15 shows the time of year, the total consumption, and the total duration predicted by our model for each generated dataset. As the plot indicates, the weather margin is, on average, much lower during summer than during winter. In addition, there appears to be a much higher variance in consumption during winter. We also see from the color scaling that higher fuel consumption is correlated with longer trip durations.

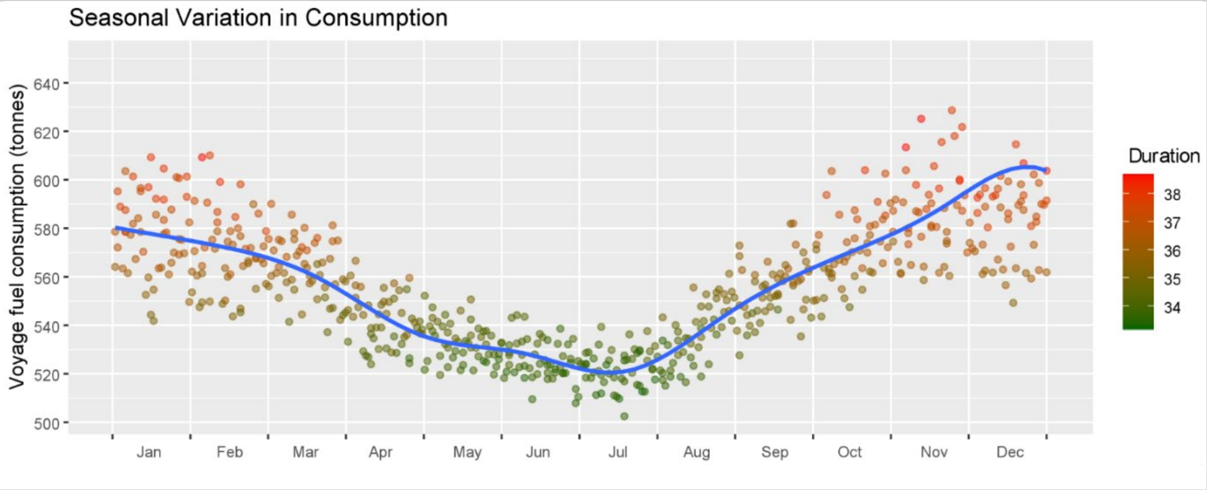

Figure 15. Start time, total consumption, and total duration for each simulated voyage. The blue line shows the estimated mean of our fuel consumption predictions for the Handysize case route.

Figure 16 shows in more detail the monthly distribution in fuel consumption for four selected months. The plot indicates that the average consumption and distribution of consumption are relatively similar between March and September, but July and November are clearly different. July has a much narrower distribution that is also situated at the lower fuel consumption levels. Conversely, consumption tends to be higher in November, but is also much less densely distributed, with predictions stretching from below 560 to above 620 tonnes.

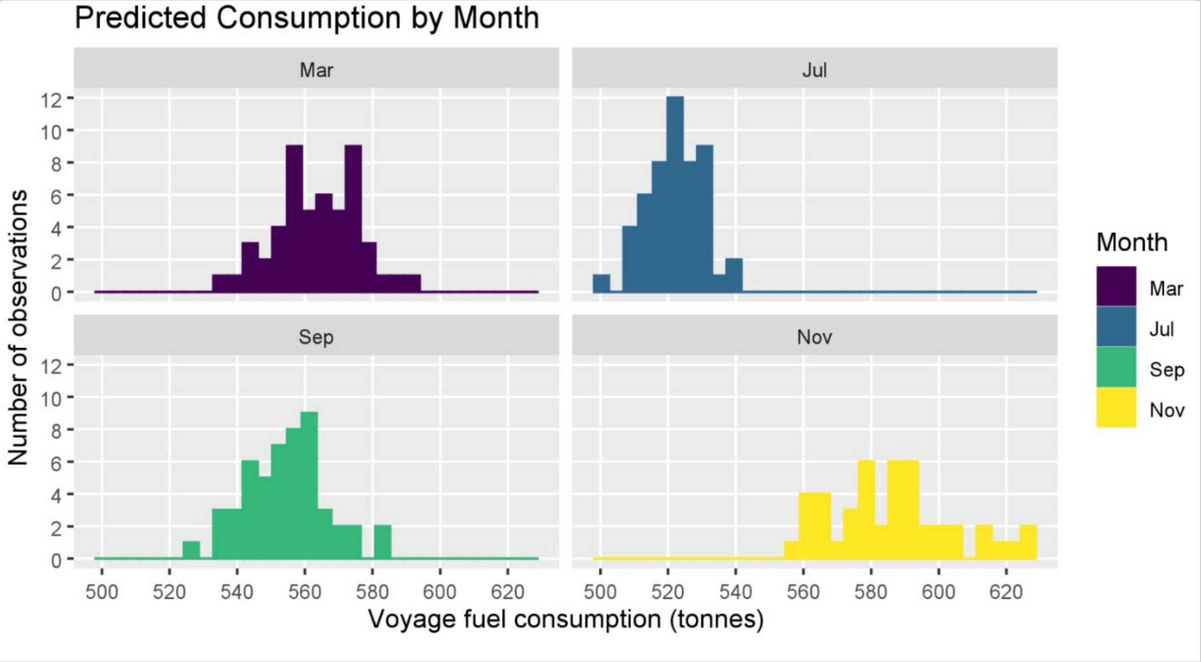

Figure 16. Variation in predicted fuel consumption for four selected months.

In addition to the figures, numerical descriptive statistics are available in Appendix E. Voyages starting in July are predicted to have the lowest fuel consumption, with a mean prediction of 522 tonnes. Fuel consumption is highest in November with 586 tonnes, a difference of 12.3%. However, some of the highest estimates are based partly on predictions for weather conditions not present in the test data and may therefore not be accurate. We will discuss this issue in greater detail in section 6.4. The 3-day difference in duration between July and November corresponds to a decrease in average SOG from 12.0 knots to 11.0 knots. The results also showed that variance was greater in the winter months. From Appendix E, we find that the largest difference in standard deviation is between July with 7.8 tonnes and November with 18.5 tonnes, an increase of 137%. Expressed as Coefficients of Variance (CoV), the standard deviations equate to a CoV of 1.49% for July and a CoV of 3.16% for November. The CoV expresses the standard deviation as a ratio of the mean (Brown, 1998).

To summarize the Handysize case route between Japan and the North American west coast, we found that the seasonal variation in fuel consumption was around 12.3% between winter and summer, with standard deviations varying from 7.8 tonnes in July to 18.5 in November. Operators on this route may use our estimated fuel consumption for the various months shown in Appendix E to price their forward cargo contracts more accurately. In addition, variances throughout the seasons indicate that the financial risk related to forward pricing is much higher during the winter than during the summer.

6.3 Supramax case voyage

The selected route for the Supramax vessel is from the Gulf of Mexico to Gibraltar, with the entire route being sailed in laden condition. Originally the vessel held an average SOG of around 11 knots, and the total fuel consumption was 323 tonnes. From the 631 generated datasets, 24 were discarded due to missing weather data.

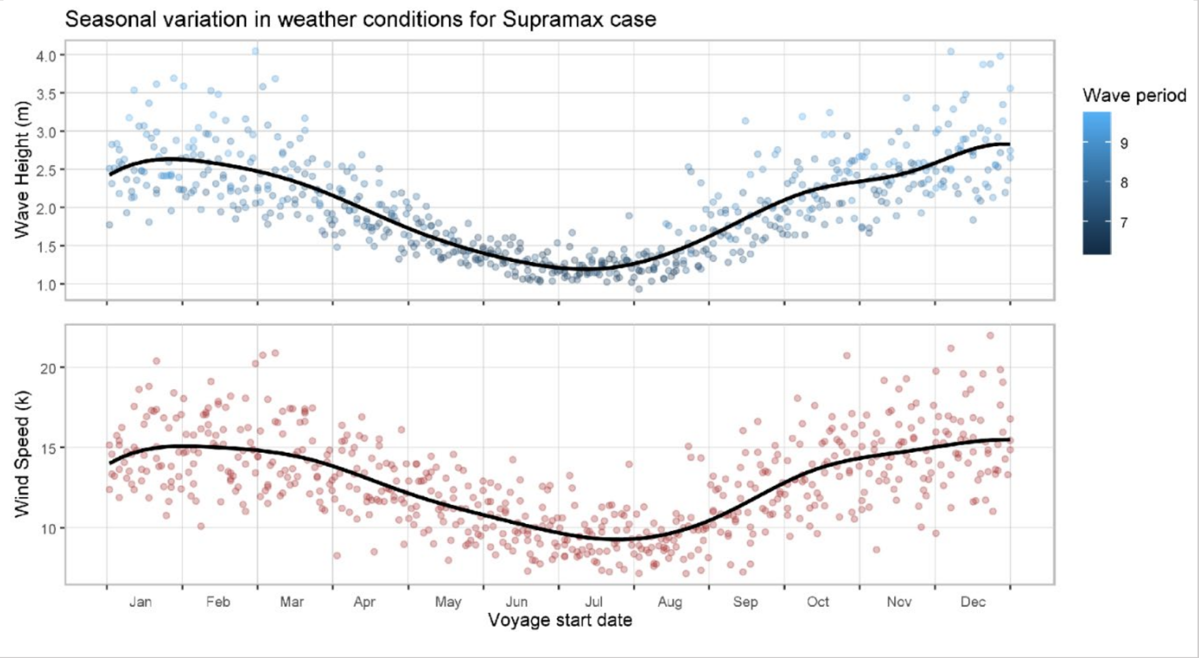

Figure 17. Weather conditions experienced along the route at different times. Each observation shows the median weather encountered on a given voyage. The black lines show the estimated means.

Figure 17 shows the weather conditions experienced along the route at different times. This route also shows a significant increase in the variance of weather conditions during the winter months, which in turn results in a greater increase in the uncertainty of consumption and duration estimates. Additional but less influential weather variables can be seen in Figure 23 (Appendix B).

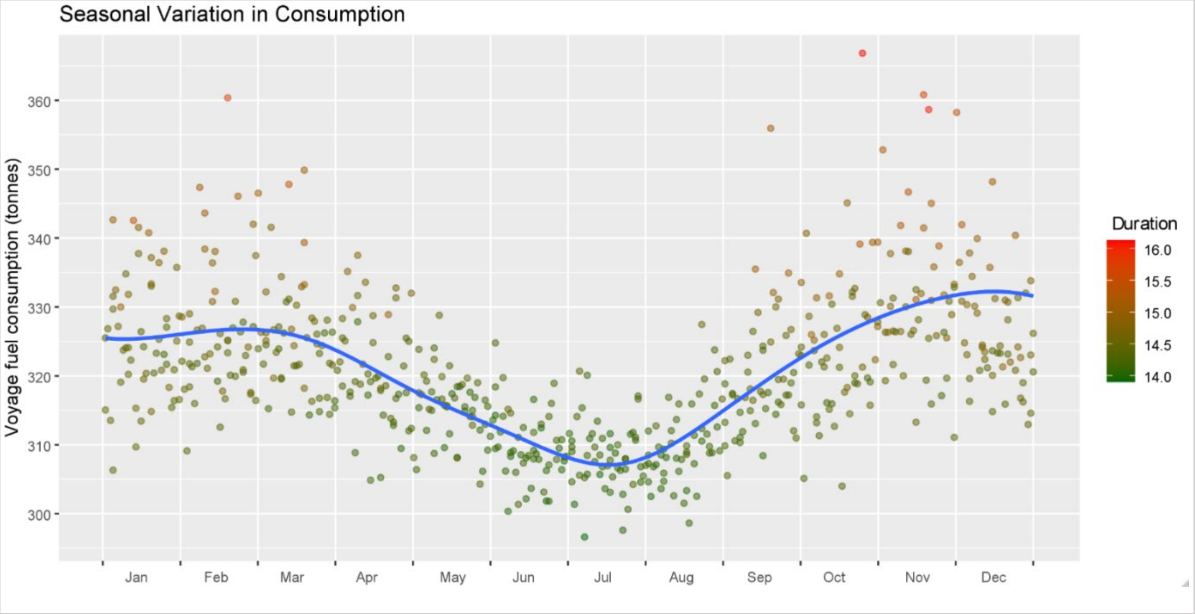

Figure 18. Start time, total consumption, and total duration for each simulated voyage. The blue line shows the estimated mean of our fuel consumption predictions for the Supramax case route.

As shown in Figure 18, Voyages starting in July are predicted to have the lowest fuel consumption, with a mean prediction of around 309 tonnes. Consumption is highest in November with 329 tonnes, a difference of 6.4%. The average voyage duration increased by 14 hours from July to November as the average SOG decreased from 12.2 to 11.8 knots. The consumption variance more than doubles during winter compared to summer. The highest difference in standard deviation is found between July with 4.8 tonnes and October with 10.8 tonnes, an increase of 125%. The CoV is 1.6% in July and 3.3% in November. Numerical descriptive statistics areas are available in Appendix E.

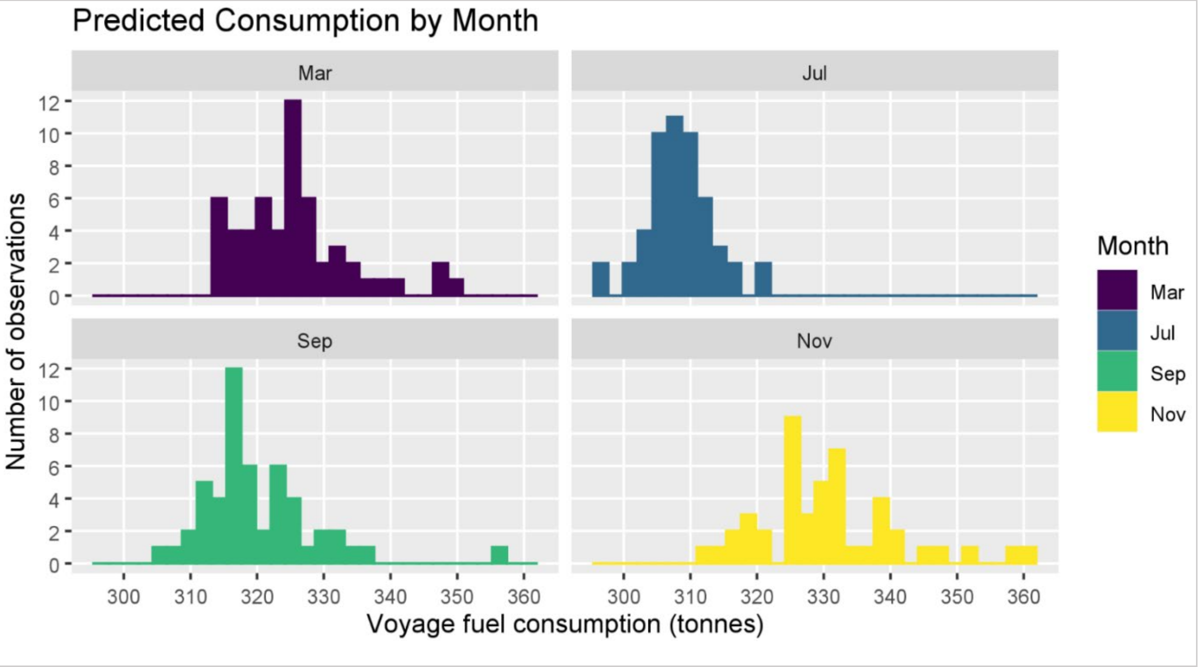

Figure 19. Variation in predicted fuel consumption for four selected months.

As with the Handysize case, some predictions in some voyages are based on wave heights higher than seen in the dataset the model trained on, and the accuracy of these predictions are therefore uncertain. There are fewer instances of such predictions in this case, however, with only 5.2% of the 9,712 individual predictions based on wave heights above 4 meters and 2.0% above 5 meters.

To summarize the Supramax case route between the Gulf of Mexico and Gibraltar, we found that the seasonal variation in fuel consumption was around 6.4% between winter and summer, with standard deviations varying from 4.8 tonnes in July to 10.8 in October. Similarly, as for the last case, operators may use the fuel estimates available in Appendix E to improve the pricing of their forward cargo contracts. In addition, our estimates show that the financial risk associated with routes during the winter is much higher than during the summer.

6.4 Extrapolation of training data

Section 6.1 briefly mentioned that we used our second-best Cubist models to make fuel consumption predictions in these case studies rather than our best-performing Extra Trees models. This section will describe the reasoning behind this and discuss some limitations in the presented results.

When combined, the results in these case studies were based on more than 1,000 simulated voyages and 30,000 noon reports spanning 25 years. In comparison, our original dataset contains around 6,500 noon reports after pre-processing, spanning just under five years. The lower volume means that we, in our simulations, run the risk of encountering weather conditions rarely or never observed in our training data. Our analysis showed that this is indeed the case, particularly for high wave heights and, to a lesser extent, high wind speeds. This is illustrated in Figure 20.

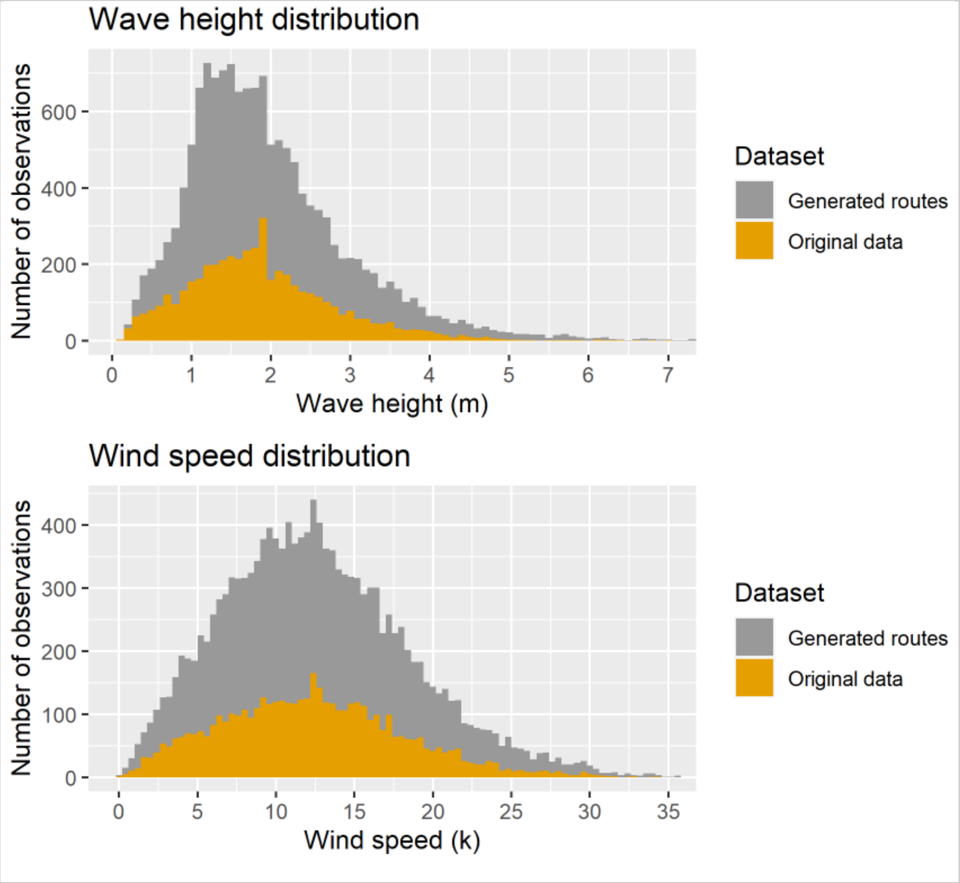

Figure 20. The number of observations at different weather conditions in the generated Supramax datasets and the original noon report dataset. Twenty observations with wave heights above 7m and 10 observations with wind speeds above 35k are not shown.

The distribution of observed wave heights in the original dataset looks comparable to our dataset, but the higher number of observations also leads to more extreme observations. The same is also true for the Handysize case. Consequently, our models need to extrapolate to make predictions outside the boundaries of the training data. Such extrapolation can be problematic for black-box models. Since physical and hydrodynamic principles do not constrain them, there is no guarantee that predictions outside the training data boundaries will be reasonable. The superior extrapolation capacity is, in fact, an area where white-box models and gray-box models are superior to black-box models (Coraddu et al., 2015). Gray-box models are able to combine the modeling of physical features from white-box models with the data-driven approach of black-box models.

There are also considerable differences between different black-box models in their capacity to extrapolate. Decision trees are inferior in this regard, as they can only interpolate (Malistov & Trushin, 2019). Linear regression is, on the other hand, much better suited for extrapolation, as the estimated predictor coefficients will ensure that the prediction is based on the estimated relationship between the dependent and independent variables even outside the boundaries of the training data. This difference is illustrated in Figure 21.

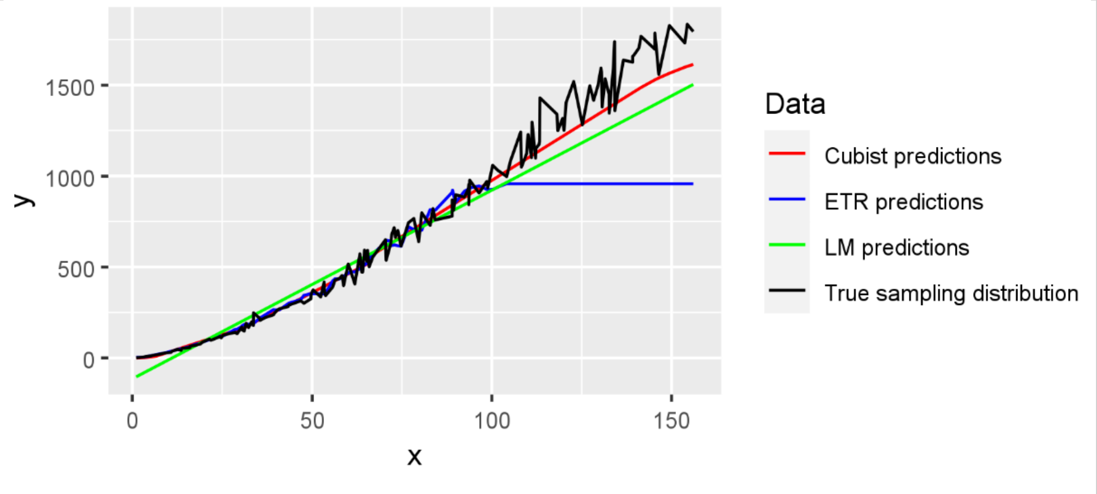

Figure 21. The sampling distribution shown is a simple exponential function with some added noise. The three models shown have been trained on values in the range $0 < x \leq 100$.

The Extra Trees (ETR) predictions in Figure 21 show a simplified example of decision trees' predictions outside the training data boundaries. When ETR is applied to the case studies presented, this will result in predictions that underestimate consumption in strong adverse weather. For this reason, we opted to use our second-best performing algorithm, Cubist, for providing predictions in our case studies. As a combined tree- and linear regression-based algorithm, Cubist is especially well-suited for this purpose because it combines the excellent accuracy of decision trees with the favorable extrapolation capability of linear regression models.

Nevertheless, the small number of observations in the training data will lower prediction accuracies for more unusual values. In our two case studies, this will primarily affect voyages where the median wave height for the voyage approaches 4 meters. For reference, wave heights encountered along the voyages are shown in Figure 14 and Figure 17. Due to the lack of data, neither bias nor accuracy can be quantified for predictions in this range.

6.5 Cost estimate example

This section will provide a numeric example showing how our findings in the Supramax case voyage translate into changes in costs associated with the voyage. For this simplified example, we will assume that the vessel is hired on a time charter contract where the charterer pays a fixed daily rate, in addition to voyage costs that include fuel costs and port charges. The charterer's total costs are then:

$$\text{Costs} = \text{Fuel consumption} \cdot \text{Fuel Price} + \text{Duration} \cdot \text{Charter Rate} + \text{Port Charges} \qquad (9)$$

The vessel carried 52,760 tonnes of cargo on this voyage, giving a profit function as follows:

$$\text{Profit} \;=\; 52{,}760 \cdot \text{Freight Rate} - \text{Costs} \qquad (10)$$

In the following, we will assume a daily charter rate of \$22,500 per day, a bunker (fuel) price of \$500 per tonne, and fixed port charges of \$50,000 for the voyage. The variable component of the costs for the voyage is then $\$500 \cdot \text{Fuel Consumption} + \$22{,}500 \cdot \text{Duration}$. Figure 22 shows the result of applying this cost function to the results previously presented in the Supramax case study.

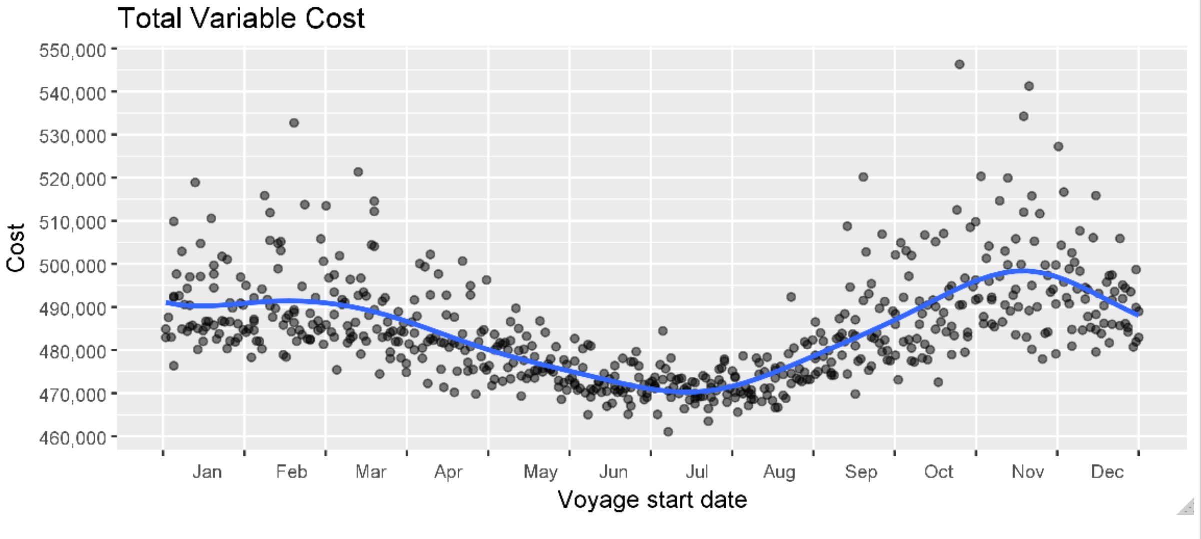

Figure 22. Distribution of variable costs for Supramax case route assuming a daily charter rate of \$22,500 per day and a fuel price of \$500 per tonne.

Although similar to the consumption distribution presented in Figure 18, this plot clearly shows the economic impact of the predicted variation. These results can further be used to find the minimum required freight rate for the expected profit to be positive, as shown below.

Using the mean variable cost of $\$470{,}000$ in July:

$$52{,}760 \cdot \text{Freight Rate} - 470{,}000 - 50{,}000 > 0 \Rightarrow \text{Freight rate} > 9.86 \qquad (11)$$

Using the mean variable cost of $\$500{,}000$ in November:

$$52{,}760 \cdot \text{Freight Rate} - 500{,}000 - 50{,}000 > 0 \Rightarrow \text{Freight Rate} > 10.42 \qquad (12)$$

6.6 Prediction uncertainty in case studies

When predicting consumption for a future voyage without knowing the weather conditions, there are two sources of uncertainty. First, there is uncertainty caused by the inherent variance in the weather conditions, and how these weather conditions will impact consumption. The variances in our case study predictions are our best estimates of this uncertainty. For instance, we found a CoV of 1.6% for the Supramax case voyage in July, and a CoV of 3.3% in November. The second source of uncertainty stems from our models' prediction errors in the given weather conditions. Our initial results, detailed in Section 5.1.4, quantifies this second source of uncertainty: the models achieved mean absolute prediction errors of around 2% on the voyage level.

When making predictions for a specific voyage based on the case study results, it is also useful to know the expected consumption. This is simply the estimated mean of the predictions for the given time of year, as shown by the blue lines in Figure 15 and Figure 18. There is also uncertainty related to the estimated means due to the models' prediction errors.

There are two ways to improve upon the prediction accuracy we achieve with our framework. The first would be to reduce the uncertainty of our estimates. The model prediction errors of 2% can be further reduced through better-quality data and improved modeling approaches. In addition, the added uncertainty we introduced through extrapolation may be reduced through increased quantities and diversity of data, or by using white or grey-box models.

The second means of improvement would be to reduce the uncertainty caused by the inherent variance in the weather conditions by using weather forecasts. The estimated CoV provides an indication of the potential value this can have. We previously described how weather forecasts are only correct 40% of the time when forecasting weather 6-10 days ahead. However, the weather forecasts only need to provide slightly more accurate forecasts than the average weather conditions to improve prediction accuracy.

7. Limitations and further work

While prediction errors can be measured on a test set, it is not possible to measure the achieved accuracy of the case study predictions, given that they are based on simulated voyages. The lack of a test set makes it difficult to quantify the uncertainty related to our weather margin estimates accurately. As previously discussed in greater detail, our models must sometimes extrapolate predictions outside the dataset range to estimate weather margins. The uncertainty related to the extrapolation increases the further the model must extrapolate, and it is difficult to measure precisely how much uncertainty is introduced as a result. Further work may negate some of this uncertainty by using gray-box models, which are more suitable for extrapolation.

Furthermore, the data we used contained many missing values. Many were accurately imputed based on their relationship with other values, but the remaining 2,247 values were imputed based on the average value of the feature. This imputation technique means our models were trained on data where the predictors had lower variances than the actual variances of these features, while the dependent variable they were trained to predict had unchanged variance. The models may thus have compensated by increasing the marginal effects of the predictors to achieve accurate predictions. Our weather margin estimates are computed based on varying values of predictors that have unchanged variance, which means the models may at times have overestimated the actual variation in fuel consumption. One possible extension that could have reduced the tendency to overestimate would have been using a more advanced imputation algorithm. An example of this is a forest-based method that aims to minimize the impact of imputed values on the final result while accounting for all other predictors (e.g., Stekhoven & Buhlmann, 2012).

Another limitation relates to the way we set SOG and trim in the case studies. As mentioned in the Procedure section, there are many potential reasons why speed might change in different weather conditions, both voluntary and non-voluntary. In our case studies, we took a rather simplistic approach to set SOG, primarily relying on a data-driven approach to ensure that the values used seemed representative for the given weather. The same procedure was then performed for trim. In this area, it seems likely that a closer familiarity with the exact policies employed for ships in our dataset could make the results more representative. Another issue with adjusting SOG based on weather is that our results no longer show how weather impacts consumption. Instead, the results show the combined effect of, in most cases, higher SOG and better weather or lower SOG and worse weather. In choosing to adjust SOG based on weather, we have prioritized showing how the case voyages tend to unfold rather than showing the isolated effect of weather on consumption. Nevertheless, the framework we have described can easily accommodate any arbitrary policy for setting SOG and trim, not least using a constant SOG and trim setting.

8. Conclusion

This thesis proposes a data-driven modeling framework for estimating weather margins in the shipping industry. The study was based on noon report data from Handysize and Supramax vessels, weather data from Copernicus (CDS and CMEMS), and Clarksons' World Fleet Register data. For the first part of the study, we developed a predictive model for fuel consumption and applied several machine learning algorithms in the process. We found that Extra Tree models gave the most accurate predictions, with an R squared of 87.6% for Handysize vessels and 88.7% for Supramax vessels. The accuracy increased to 99.5% and 98.6%, respectively, for total fuel consumption on a voyage level. We also found that Cubist, RF, ANN, and variants of SVM and GP with radial kernels achieved accurate predictions, while the linear models, SVM and GP with polynomial kernels and shrinkage-based models were less accurate.

For the second part of our thesis, we used the trained models to generate predictions using historical weather conditions from the last decades and studied the seasonal patterns of weather margins. We applied this methodology on two real-world case routes, one for Handysize vessels across the North Atlantic and one for Supramax vessels across the North Pacific. Our model predictions suggest a seasonal variation in fuel consumption of 12.3% and 6.4% for the Handysize and Supramax case routes, respectively. In addition, we found the standard deviations for weather margins to be more than twice as high during winter as during summer for both cases.

The weather margin estimates are, however, computed under some degree of uncertainty. Complicating factors include the imputations' effect on predictor weighting, the uncertainty related to extrapolation outside the range of our dataset, and the behavioral patterns related to weather avoidance and route optimization measures. The combination of model uncertainty and generally high variance in weather conditions also make accurate point predictions unfeasible, even for voyages planned for a couple of weeks into the future. Weather forecast integration may reduce this uncertainty in future work. Nevertheless, we have shown how the vast amounts of historical weather data freely available can be used to estimate averages and variances for seasonal patterns in weather margins, and by extension, fuel consumption. The framework may be used by chartering managers to determine the expected weather margin and variance given any route and load configuration. This information can indicate the cost and risk associated with a route, thus facilitating improved forward pricing of cargo.

References

Abebe, M., Shin, Y., Noh, Y., Lee, S., & Lee, I. (2020). Machine Learning Approaches for Ship Speed Prediction towards Energy Efficient Shipping. *Applied Sciences*, *10*(7), 2325. https://doi.org/10.3390/app10072325

Abraham, N. (2017, July 13). Looking for Alternatives in Validation for Machine Learning. *Dummies*. https://www.dummies.com/programming/looking-alternatives-validation-machine-learning/

Adland, R., Cariou, P., Jia, H., & Wolff, F.-C. (2018). The energy efficiency effects of periodic ship hull cleaning. *Journal of Cleaner Production*, *178*, 1–13. https://doi.org/10/gc5h5t

Adland, R., Cariou, P., & 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*, 101972. https://doi.org/10/ghr3qs

Aggarwal, C. C. (2015). *Data Mining: The textbook*. Springer International Publishing. https://doi.org/10.1007/978-3-319-14142-8

Aldous, L., Smith, T., & Bucknall, R. (2013). *Noon report Data Uncertainty*. 13.

Anish. (2019, June 11). What is Noon Report On Ships And How Is It Prepared? *Marine Insight*. https://www.marineinsight.com/guidelines/what-is-noon-report-on-ships/

Arribas, F. P. (2007). Some methods to obtain the added resistance of a ship advancing in waves. *Ocean Engineering*, *34*(7), 946–955. https://doi.org/10/b4ms58

Bergstra, J., & Bengio, Y. (2012). Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research*, *13*, 281–305.

Bohlander, J. (2009). *Review of options for in-water cleaning of ships*. MAF Biosecurity New Zealand. http://www.biosecurity.govt.nz/files/pests/salt-freshwater/options-for-in-water-cleaning-of-ships.pdf#14

Brown, C. E. (1998). Coefficient of Variation. In C. E. Brown, *Applied Multivariate Statistics in Geohydrology and Related Sciences* (pp. 155–157). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-80328-4_13

Brownlee, J. (2020a, August 14). A Tour of Machine Learning Algorithms. *Machine Learning Mastery*. https://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/

Brownlee, J. (2020b, August 26). Train-Test Split for Evaluating Machine Learning Algorithms. *Machine Learning Mastery*. https://machinelearningmastery.com/train-test-split-for-evaluating-machine-learning-algorithms/

Chugh, A. (2020, December 8). *MAE, MSE, RMSE, Coefficient of Determination, Adjusted R Squared—Which Metric is Better?* Medium. https://medium.com/analytics-vidhya/mae-mse-rmse-coefficient-of-determination-adjusted-r-squared-which-metric-is-better-cd0326a5697e

Clarksons Research Services Limited. (2021). *The Bulk Carrier Register* (World Fleet Register). https://www.clarksons.net/wfr/

Copernicus Climate Change Service (C3S) Climate Data Store (CDS). (2018). *ERA5 hourly data on single levels from 1979 to present*. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels?tab=overview

Coraddu, A., Oneto, L., Baldi, F., & Anguita, D. (2015). Ship efficiency forecast based on sensors data collection: Improving numerical models through data analytics. *OCEANS 2015 - Genova*, 1–10. https://doi.org/10/f3m3pq

Du, Y., Meng, Q., Wang, S., & 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. https://doi.org/10/ghzm77

Dürr, S., & Thomason, J. (Eds.). (2010). *Biofouling* (1st ed). Wiley-Blackwell.

Eide, E. (2015). *Calculation of Service and Sea Margins* [Norwegian University of Science and Technology]. https://ntnuopen.ntnu.no/ntnu-xmlui/bitstream/handle/11250/2350635/13896_FULLTEXT.pdf?sequence=1

Erto, P., Lepore, A., Palumbo, B., & Vitiello, L. (2015). *A Procedure for Predicting and Controlling the Ship Fuel Consumption: Its Implementation and Test*. 8. https://doi.org/10/f7wzcm

E.U. Copernicus Marine Service Information. (2018). *GLORYS12V1—Global Ocean Physical Reanalysis Product*. E.U. Copernicus Marine Service Information. https://resources.marine.copernicus.eu/?option=com_csw&view=details&product_id=GLOBAL_REANALYSIS_PHY_001_030

E.U. Copernicus Marine Service Information. (2019). *Global Sea Physical Analysis and Forecasting Product*. E.U. Copernicus Marine Service Information. https://resources.marine.copernicus.eu/?option=com_csw&view=details&product_id=GLOBAL_ANALYSIS_FORECAST_PHY_001_024

Fernando, J. (2020, November). *R-Squared Definition*. Investopedia. https://www.investopedia.com/terms/r/r-squared.asp

Fidan, M. C. (2019, October 28). *3 Reasons Why Freight Forwarding Will Not See Disruption from Technology*. More Than Shipping. https://www.morethanshipping.com/3-reasons-why-the-freight-forwarding-industry-will-not-experience-a-big-disruption-from-technology/

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, *63*(1), 3–42. https://doi.org/10/frqxsw

Gkerekos, C., Lazakis, I., & Theotokatos, G. (2019). Machine learning models for predicting ship main engine Fuel Oil Consumption: A comparative study. *Ocean Engineering*, *188*, 106282. https://doi.org/10.1016/j.oceaneng.2019.106282

Goldsworthy, L., & Goldsworthy, B. (2015). Modelling of ship engine exhaust emissions in ports and extensive coastal waters based on terrestrial AIS data – An Australian case study. *Environmental Modelling & Software*, *63*, 45–60. https://doi.org/10/f6vx82

Hu, Q. S., & Skaggs, K. (2009). Accuracy of 6-10 Day Precipitation Forecasts and Its Improvement in the Past Six Years. *NOAA Annual Climate Prediction Application Science Workshop*, *7th*, 1.

Jaitley, U. (2018, August 10). Why Data Normalization is necessary for Machine Learning models. *Medium*. https://medium.com/@urvashilluniya/why-data-normalization-is-necessary-for-machine-learning-models-681b65a05029

Jalkanen, J.-P., Brink, A., Kalli, J., Pettersson, H., Kukkonen, J., & Stipa, T. (2009). A modelling system for the exhaust emissions of marine traffic and its application in the Baltic Sea area. *Atmospheric Chemistry and Physics*, *9*(23), 9209–9223. https://doi.org/10/cv5bd6

Jalkanen, J.-P., Johansson, L., Kukkonen, J., Brink, A., Kalli, J., & Stipa, T. (2012). Extension of an assessment model of ship traffic exhaust emissions for particulate matter and carbon monoxide. *Atmospheric Chemistry and Physics*, *12*(5), 2641–2659. https://doi.org/10/gb8q7f

Jeon, M., Noh, Y., Shin, Y., Lim, O.-K., Lee, I., & Cho, D. (2018). Prediction of ship fuel consumption by using an artificial neural network. *Journal of Mechanical Science and Technology*, *32*(12), 5785–5796. https://doi.org/10.1007/s12206-018-1126-4

Jia, H., Prakash, V., & Smith, T. (2019). *Estimating vessel payloads in bulk shipping using AIS data*. 16. https://doi.org/10.1504/IJSTL.2019.096864

Kiil, S., Dam-Johansen, K., Weinell, C. E., Pedersen, M. S., & Codolar, S. A. (2002). Dynamic simulations of a self-polishing antifouling paint exposed to seawater. *Journal of Coatings Technology*, *74*(6), 45–54. https://doi.org/10/cmpjn7

Lewinson, E. (2020, November 1). *Choosing the correct error metric: MAPE vs. sMAPE*. Medium. https://towardsdatascience.com/choosing-the-correct-error-metric-mape-vs-smape-5328dec53fac

Lindholdt, A., Dam-Johansen, K., Olsen, S. M., Yebra, D. M., & Kiil, S. (2015). Effects of biofouling development on drag forces of hull coatings for ocean-going ships: A review. *Journal of Coatings Technology and Research*, *12*(3), 415–444. https://doi.org/10/gj7hvv

Magnussen, A. K. (2017). *Rational calculation of sea margin* [Norwegian University of Science and Technology]. https://ntnuopen.ntnu.no/ntnu-xmlui/bitstream/handle/11250/2453425/17296_FULLTEXT.pdf?sequence=1

Malistov, A., & Trushin, A. (2019). Gradient Boosted Trees with Extrapolation. *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, 783–789. https://doi.org/10/gj6ttx

MAN Diesel & Turbo. (2015). *Basic principles of ship propulsion*. https://spain.mandieselturbo.com/docs/librariesprovider10/sistemas-propulsivos-marinos/basic-principles-of-ship-propulsion.pdf?sfvrsn=2

Meng, Q., Du, Y., & Wang, Y. (2016). Shipping log data based container ship fuel efficiency modeling. *Transportation Research Part B: Methodological*, *83*, 207–229. https://doi.org/10/f77qhc

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2018). *Foundations of machine learning* (Second edition). The MIT Press.

Nabergoj, R., & Prpi, J. (2007). *A comparison of different methods for added resistance prediction*. 4.

Otto, S. A. (2019, January 7). *How to normalize the RMSE.* https://www.marinedatascience.co/blog/2019/01/07/normalizing-the-rmse/

Pedersen, B. P., & Larsen, J. (2009). Prediction of Full-Scale Propulsion Power using Artificial Neural Networks. *In Proceedings of the 8th International Conference on Computer and IT Applications in the Maritime Industries*, 537–550.

Pedregosa et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Petersen, J. P., Jacobsen, D. J., & Winther, O. (2012). Statistical modelling for ship propulsion efficiency. *Journal of Marine Science and Technology*, *17*(1), 30–39. https://doi.org/10.1007/s00773-011-0151-0

Rakke, S. G. (2016). *Ship emissions calculation from AIS* [Norwegian University of Science and Technology]. https://ntnuopen.ntnu.no/ntnu-xmlui/handle/11250/2410741

ScienceX. (2008, July 9). *Ocean Wind Power Maps Reveal Possible Wind Energy Sources* [Newsletter]. Ocean Wind Power Maps Reveal Possible Wind Energy Sources. https://phys.org/news/2008-07-ocean-power-reveal-energy-sources.html

Soner, O., Akyuz, E., & Celik, M. (2019). Statistical modelling of ship operational performance monitoring problem. *Journal of Marine Science and Technology*, *24*(2), 543–552. https://doi.org/10/ggqf4x

Stekhoven, D. J., & Buhlmann, P. (2012). MissForest—Non-parametric missing value imputation for mixed-type data. *Bioinformatics*, *28*(1), 112–118. https://doi.org/10/dhxth8

Stopford, M. (2009). *Maritime economics* (3rd ed). Routledge.

Swalin, A. (2018, April 7). *Choosing the Right Metric for Evaluating Machine Learning Models—Part 1*. Medium. https://medium.com/usf-msds/choosing-the-right-metric-for-machine-learning-models-part-1-a99d7d7414e4

Tillig, F., & Ringsberg, J. W. (2019). A 4 DOF simulation model developed for fuel consumption prediction of ships at sea. *Ships and Offshore Structures*, *14*(sup1), 112–120. https://doi.org/10/gh83hb

Tipping, M. E. (2004). Bayesian Inference: An Introduction to Principles and Practice in Machine Learning. In O. Bousquet, U. von Luxburg, & G. Rätsch (Eds.), *Advanced Lectures on Machine Learning* (Vol. 3176, pp. 41–62). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-28650-9_3

UNCTAD. (2020). *Review of maritime transport 2020* (p. 159). United Nations. https://unctad.org/system/files/official-document/rmt2020_en.pdf

Uyanık, T., Arslanoğlu, Y., & Kalenderli, O. (2019). Ship fuel consumption prediction with machine learning. *In Proceedings of the 4th International Mediterranean Science and Engineering Congress*, 757–759.

Wang, B. J. S., Zhao, J., Wei, L., & Xu, T. (2018). Predicting ship fuel consumption based on LASSO regression. *Transportation Research Part D*, *65*, 817–824. https://doi.org/10/ghtpmv

Wang, S., & 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. https://doi.org/10/fx5pqb

Xiong, Y., Wang, Z., & Qi, W. (2013). Numerical study on the influence of boss cap fins on efficiency of controllable-pitch propeller. *Journal of Marine Science and Application*, *12*(1), 13–20. https://doi.org/10/gj73zm

Yilmaz, S., Erdem, D., & Kavsaoglu, M. (2013, January 7). Effects of Duct Shape on a Ducted Propeller Performance. *51st AIAA Aerospace Sciences Meeting Including the New Horizons Forum and Aerospace Exposition*. 51st AIAA Aerospace Sciences Meeting including the New Horizons Forum and Aerospace Exposition, Grapevine (Dallas/Ft. Worth Region), Texas. https://doi.org/10/gj73x4

Appendices

Appendix A - Cargo weight - draft relationship

We here provide an estimated regression formula for both vessel classes that accurately estimates the relationship between draft in meters and cargo weight in tonnes. Equation (13) for the Handysize vessels achieved an R squared of 94%.

$$D_M = \exp(1.7024 + 0.000017268 \cdot C_{tonnes}) \tag{13}$$

where D_M denotes draft in meters, and C_{tonnes} denotes cargo weight in tonnes.

For the Supramax vessels, a simple linear regression function becomes imprecise when cargo weight is below 5000 tonnes, and for this configuration, we instead provide the mean draft as a good approximation. The regression on cargo weights above 5000 tonnes, shown in Equation (14), achieved an R squared of 98%.

$$D_M = \begin{cases} 5.89, & C_{tonnes} < 5000 \\ \exp(1.6960 + 0.000014836 \cdot C_{tonnes}), & C_{tonnes} \geq 5000 \end{cases} \tag{14}$$

Appendix B - Seasonality in sea state for case routes

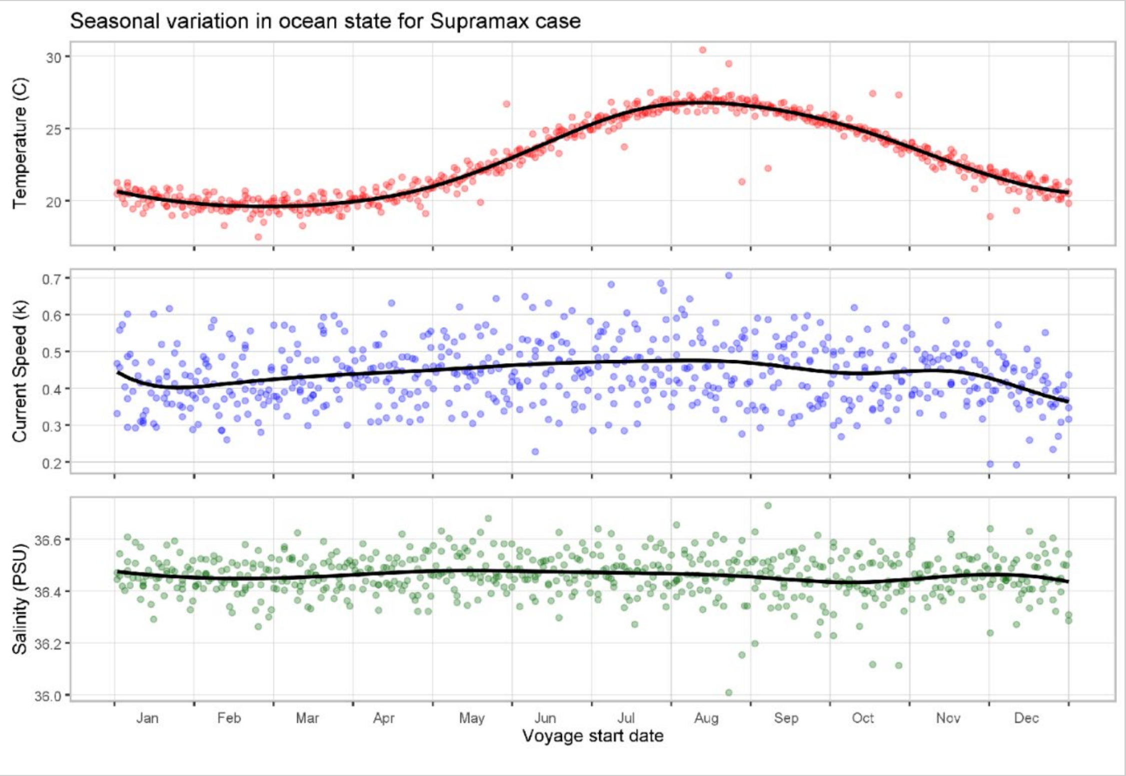

Figure 23. Seasonal variation in weather variables along the Supramax case route.

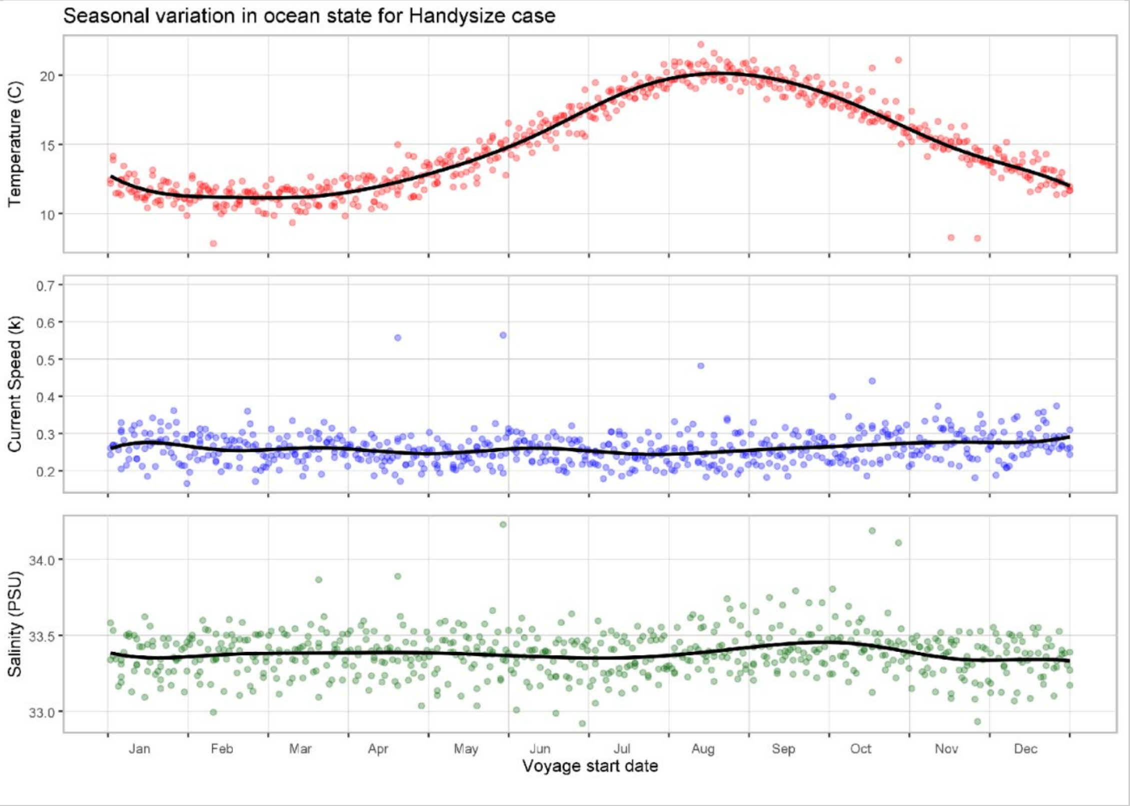

Figure 24. Seasonal variation in weather variables along the Handysize case route.

Appendix C - Result comparison for different predictors

Log/level comparison for speed - Supramax		
Transformation	Log-transformed	Level
Model	sMAPE (%)	sMAPE (%)
Linear Regression	7.89	7.77
Neural Network	6.50	6.42
Extra Trees Regression	3.65	3.62
Random Forest	6.49	6.47
LASSO	7.93	7.81
Ridge	7.96	7.83
SVM Poly	5.85	6.01
SVM Radial	5.23	5.20
GP Poly	5.66	5.81
GP Radial	15.31	15.31

Table 12. Impact of log-transforming speed variable on prediction accuracy.

Longitude and latitude comparison for Supramax		
Transformation	Long/lat included	Long/lat not included
Model	RMSE	RMSE
Linear Regression	2.241	2.235
Neural Network	1.946	1.879
Extra Trees Regression	1.331	1.383
Random Forest	1.875	1.915
LASSO	2.251	2.250
Ridge	2.261	2.256
SVM Poly	2.253	2.182
SVM Radial	1.766	1.814
GP Poly	1.811	2.247
GP Radial	3.830	3.830
Cubist	1.568	1.617

Table 13. Impact of including latitude and longitude on prediction accuracy.

Dry docking comparison for Supramax		
Transformation	Dry docking included	Dry docking not included
Model	RMSE	RMSE
Linear Regression	2.241	2.242
Neural Network	1.946	2.003
Extra Trees Regression	1.331	1.382
Random Forest	1.875	1.928
LASSO	2.251	2.253
Ridge	2.261	2.261
SVM Poly	2.253	1.959
SVM Radial	1.766	1.696
GP Poly	1.811	2.367
GP Radial	3.830	3.830

Table 14. Impact of including dry docking variable on prediction accuracy.

Temperature and salinity comparison for Supramax		
Transformation	Variables included	Variables not included
Model	RMSE	RMSE
Linear Regression	2.237	2.241
Neural Network	1.791	1.946
Extra Trees Regression	1.343	1.331
Random Forest	1.882	1.875
LASSO	2.249	2.251
Ridge	2.256	2.261
SVM Poly	1.774	2.253
SVM Radial	1.655	1.766
GP Poly	2.246	1.811
GP Radial	3.830	3.830

Table 15. Impact of including temperature and salinity on prediction accuracy.

Appendix D - Prediction error distributions for Handysize vessels

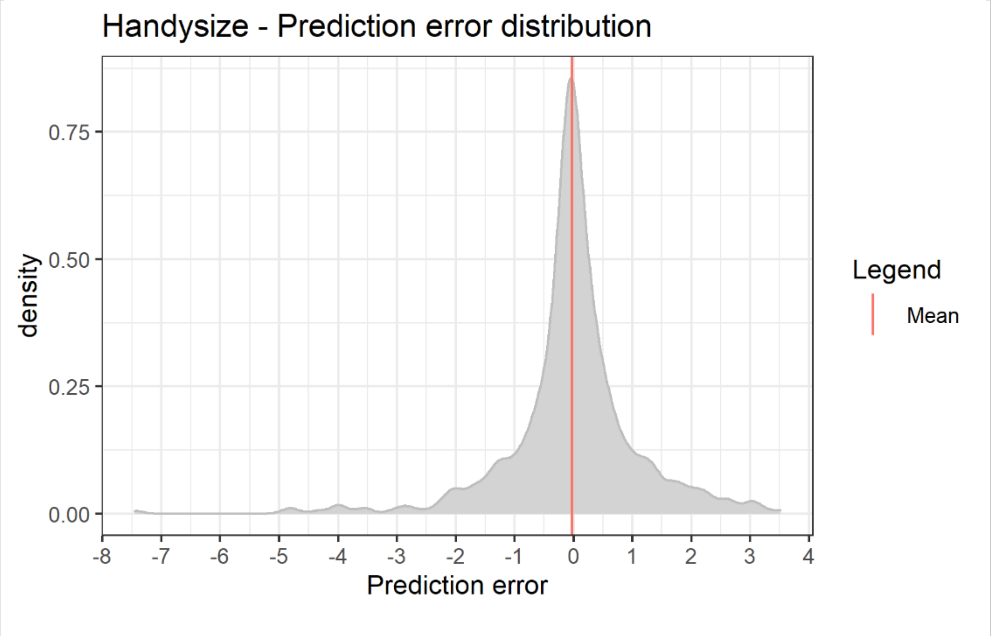

Figure 25. Prediction error distribution on fuel consumption reported in Handysize noon reports.

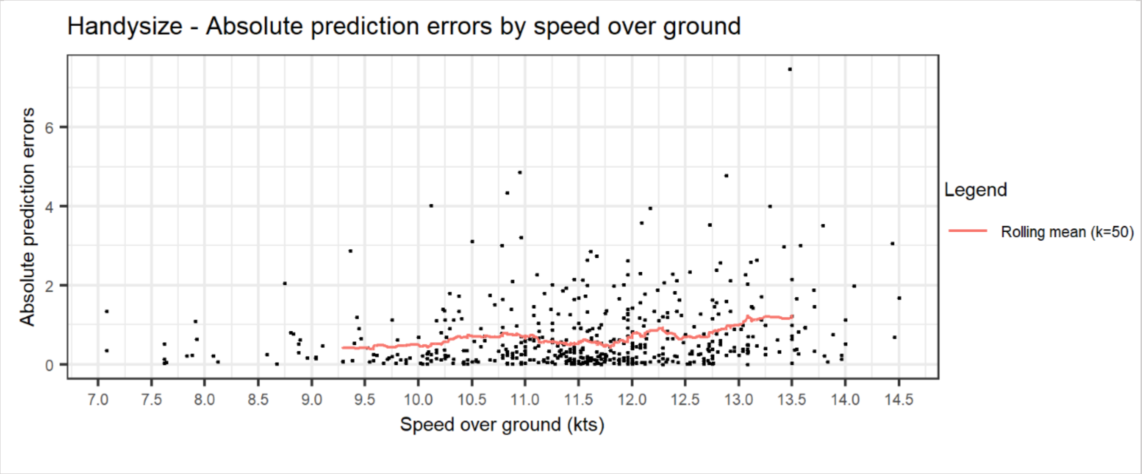

Figure 26. Absolute prediction errors on fuel consumption reported in Handysize noon reports, sorted by speed over ground. Plotted with a rolling mean of $k = 50$.

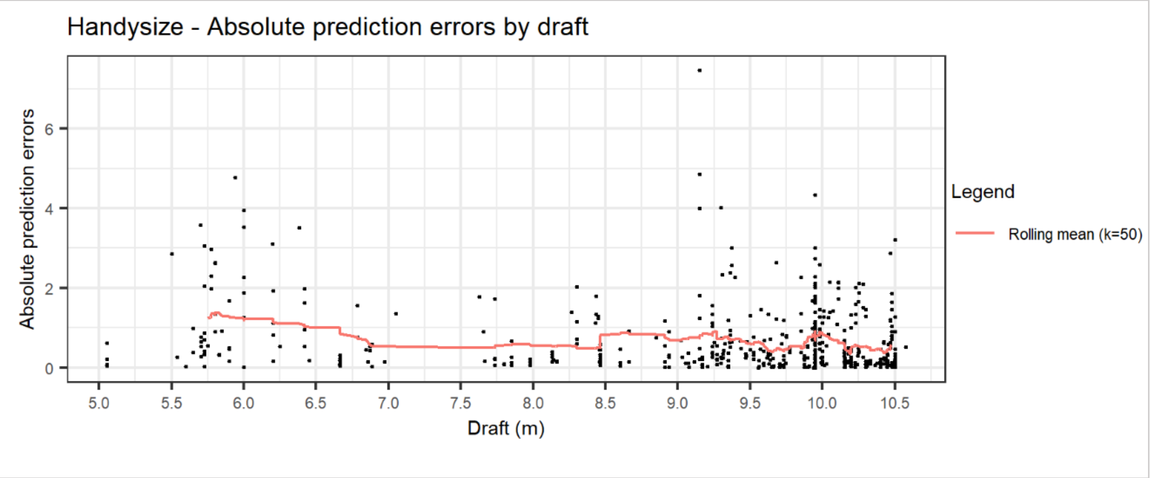

Figure 27. Absolute prediction errors on fuel consumption reported in Handysize noon reports, sorted by draft. Plotted with a rolling mean of $k = 50$.

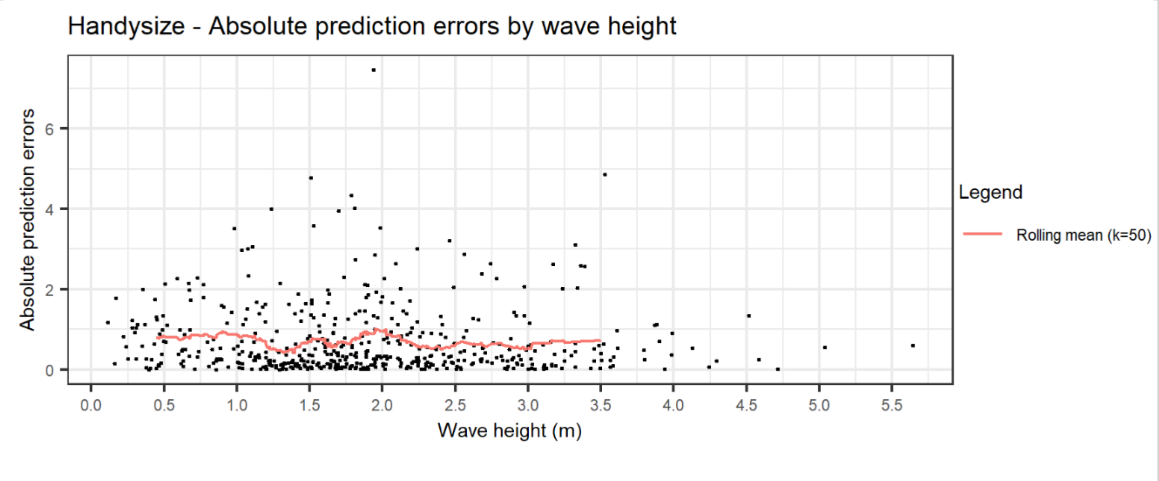

Figure 28. Absolute prediction errors on fuel consumption reported in Handysize noon reports, sorted by wave height. Plotted with a rolling mean of $k = 50$.

Appendix E - Descriptive statistics of fuel consumption estimates from case routes

Handysize – Descriptive statistics for fuel consumption estimates

Statistic	Mean fuel	Std. fuel	5% fuel	25% fuel	75% fuel	95% fuel	Mean duration	Std. duration	5% duration	25% duration	75% duration	95% duration
January	577	16.7	551	563	591	601	36.9	0.8	35.6	36.4	37.5	38.0
February	570	16.4	546	561	579	599	36.6	0.8	35.3	36.1	37.1	37.8
March	563	12.0	543	556	574	580	36.1	0.8	34.8	35.5	36.6	37.3
April	542	10.8	526	534	549	559	34.9	0.6	34.0	34.4	35.2	35.8
May	533	8.0	521	527	537	546	34.2	0.4	33.7	33.9	34.5	34.9
June	526	7.9	515	520	532	540	33.9	0.3	33.5	33.7	34.0	34.4
July	522	7.8	510	517	528	534	33.8	0.3	33.3	33.6	34.0	34.4
August	536	12.4	517	527	545	554	34.5	0.6	33.7	33.9	34.9	35.3
September	554	11.8	536	545	562	572	35.4	0.6	34.5	35.1	35.8	36.3
October	571	14.7	549	561	580	599	36.3	0.8	35.2	35.8	36.9	37.5
November	586	18.5	560	573	595	621	37.2	0.7	35.9	36.7	37.8	38.3
December	582	15.5	560	569	593	603	37.1	0.7	36.0	36.5	37.6	38.1

Table 16. Handysize - Descriptive statistics of consumption and duration estimates.

Supramax – Descriptive statistics for fuel consumption estimates

Statistic	Mean fuel	Std. fuel	5% fuel	25% fuel	75% fuel	95% fuel	Mean duration	Std. duration	5% duration	25% duration	75% duration	95% duration
January	323	8.4	311	318	330	337	14.6	0.2	14.3	14.4	14.6	15.0
February	324	9.9	314	318	326	342	14.6	0.3	14.2	14.4	14.7	15.1
March	324	8.2	312	318	327	341	14.5	0.3	14.2	14.3	14.7	15.1
April	320	7.7	308	315	325	331	14.4	0.2	14.1	14.2	14.5	14.8
May	315	5.3	306	313	318	323	14.2	0.1	14.0	14.1	14.3	14.5
June	310	5.1	302	307	313	319	14.1	0.1	14.0	14.0	14.2	14.3
July	309	4.8	302	306	311	316	14.1	0.1	14.0	14.0	14.2	14.2
August	310	5.0	302	307	312	318	14.2	0.1	14.0	14.1	14.2	14.3
September	319	7.9	310	314	321	332	14.4	0.3	14.1	14.2	14.5	15.0
October	323	10.8	311	317	328	339	14.6	0.4	14.2	14.4	14.8	15.1
November	329	10.5	314	324	333	348	14.7	0.4	14.3	14.5	14.9	15.4
December	325	8.8	314	319	329	340	14.7	0.3	14.3	14.4	14.8	15.1

Table 17. Supramax - Descriptive statistics of consumption and duration estimates.