UNIVERSITÄT
DUISBURG
ESSEN

Offen im Denken

Master Thesis

on the topic of

Modelling and optimization of ship's fuel consumption using Random Forest Regression (RFR)

Submitted to the Faculty of Engineering
of University Duisburg Essen

by

Hibatul Wafi
3021919

Betreuer:	M. T. Muhammad Fakhruriza Pradana
1. Gutachter:	Prof. Dr.-Ing. B. Noche
2. Gutachter:	Dr.-Ing. Alexander Goudz
Studiengang:	ISE General Mechanical Engineering
Studiensemester:	Summer semester 2023
Datum:	04.05.2023

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Marine industry stakeholders are actively pursuing research on efficient ship operation. This research direction is motivated by the increasing price of fuel oil and stricter environmental regulations. The fuel aboard a ship is referred to as "bunkers" and accounts for a substantial portion of the vessel's operational expenses (OPEX). It is known that bunker fuel takes up more than 50% of voyage costs and constitutes up to 75% of the ship's total operating cost. It can be inferred that energy efficient ship operations that could reduce fuel consumption translate to an increase in profitability (**Stopford**, **2009**; **Ronen**, **2011**; **Bialystocki and Konovessis**, **2016**). Furthermore, efficient operation also means reducing Greenhouse Gas Emissions (GHG). The most recent report by International Maritime Organisation indicated that GHG emissions from shipping make up 2.51% of global emissions (**IMO**, **2020**). This alignment in motivation implies that through energy efficient ship operation, marine industry stakeholders gain economic benefits while adhering to stringent environmental regulations.

With that, maritime industry stakeholder actively searches for methods to ensure energy efficient operation. Two approaches are considered, namely technical solutions and operational solutions. Technical solutions involve modification to the vessel's structure and power system. But these solutions are expensive, and it requires engineering innovations (**Yan et al.**, **2021**; **Li et al.**, **2022**). Because of this, stakeholders look for cheaper solutions to achieve energy efficient operation. The answer for an inexpensive approach lies in optimisation of operational measures, it carries less cost, and it does not require initial investments. Several recommended solutions can be found in Ship Energy Efficiency Management Plan (SEEMP).

However, greater focus will be given in this thesis towards optimising ship speed as reduction of ship speed has the greatest impact on fuel consumption. Different studies indicated that fuel consumption is correlated through a third-order, non-linear function of the ship speed (**Wang and Meng**, **2012**; **Ronen**, **2011**; **Du et al.**, **2019**). The significant impact of ship speed on fuel consumption is further supplemented by reports and studies stating that reducing ship speed by about 2 – 3 knots could halve the operating cost of shipping companies (**Stopford**, **2009**; **Wijnolst et al.**, **2009**). For these reasons, slow steaming is the measure that is most widely adopted by shipping operator.

While inexpensive, optimising operational measures is not an easy and trivial task. Several factors ranging from vessel operational performance to varying weather conditions make it challenging to model the ship speed. Some fuel consumption models, which are based on historical data and ship parameters, lack generalisation capabilities, and it is sensitive towards noisy data. To address this problem, recent research turns towards data-driven approach i.e. machine learning approach to predict ship speed and fuel consumption. These studies reported success in their modelling, citing good generalisation capability and low prediction errors. Despite these successes, maritime experts find it difficult to accept models based on data driven approach, as some data-driven models are complex as well as unintuitive and in some cases can violate basic physical knowledge of the vessel. The performance of the data-driven model is also greatly dependent on both data quantity and quality (**Yan et al.**, **2021**; **Gkerekos et al.**, **2019**).

As such, prompted by volatility and ever-increasing bunker fuel price, developing a model that could accurately predict Fuel Oil consumption (FOC) could prove to be useful to maritime industry stakeholders. As stakeholders could make critical economical decisions at the most opportune moment without violating the stringent environmental regulations.

1.1 Thesis Objective

This thesis proposes an intuitive, data-driven modelling approach that considers varying ship state and environment conditions to predict fuel consumption. To ensure the abundance of data during modelling, this thesis utilise data fused between Automatic Identification System (AIS) and weather data.

To achieve this, Grey Box Model (GBM) approach is selected. Machine learning approach using tree-based regressor is considered to provide a certain degree of intuitiveness to predict ship over ground (SOG) over different journey periods using fused AIS and weather data. Predicted SOG is then converted to actual ship speed i.e. Speed Through Water (STW). STW will be used as the input for modelling of Fuel Oil Consumption (FOC), which is carried out through Holtrop-Mennen estimation method (**Holtrop and Mennen**, **1978**, **1982**; **Holtrop**, **1984**), a power estimation method based on hydrodynamic laws which consider resistance forces exerted by environmental conditions.

The following Research Questions (RQs) could be raised during the development of the model :

- **RQ1**: What are the steps that should be taken to optimise the predictive performance of the model?

- **RQ2**: Is it feasible to fuse AIS data and meteorological data to accurately predict the ship's SOG and subsequently FOC of the ship?
- **RQ3**: Which approximations and empirical equations are suitable to estimate the resistance forces required to estimate the power required by the ship?

1.2 Thesis Boundaries

The following research boundaries are set throughout this thesis:

- Due to the continuous nature of the SOG, only the regression aspect of Random Forest (RF) will be considered.
- The focus of this work is a detailed study of the performance and possible optimisation configuration of different tree-based predictors for SOG. As such, an exhaustive comparison study between different types of machine learning models will not be performed.
- In the case study, the approximation for ship parameters and dimensions is based on a similar type of ship with nearly identical dimensions.

1.3 Thesis Contributions

The GBM approach using the fusion of AIS data and weather data provides the following contributions :

- Economical and independent data source.
- Robust modelling approach that requires minimal data pre-processing and minimal model configuration.
- Comprehensible model that adheres to physical principles and hydrodynamic laws of the vessel.

1.4 Thesis Structure

The thesis is organised with the following structure:

Chapter 1 introduces the problem statement and described the objective and boundaries of the thesis. The novelty of this thesis is declared in this chapter.

Chapter 2 The fundamental aspects of the methodologies used to develop the model will be explained in this chapter.Section 2.1 including literature review of relevant past and present research. The fundamentals of the tree-based model will be discussed in Section 2.2, basic explanation of the parameters used in AIS and weather data will be given in Section 2.3 and Section 2.4. Section 2.5.2 presents the empirical formulas and parameters used to estimate fuel consumption used by the ship based on various literature studies.

Chapter 3 discuss the methodology used to develop tree-based model used for SOG prediction. The discussion comprises analysis of training data, feature selection and reduction and selection of tuning parameters of the model. The methodology to estimate resistance for ship power estimation will be discussed in this chapter as well.

??, the GBM model will be evaluated using appropriate performance metrics and their effectiveness will be discussed. The review of the strength and limitations concerning the GBM method will be discussed here.

Chapter 5 The summary of this study and reflections on the research process will be presented here.

Chapter 2

Theoretical Background

2.1 Literature Review

The literature review in Section 2.1 presents past and present research on utilisation of machine learning methods to achieve energy efficient operation. The concept of different modelling approaches for ship operation will be discussed in Section 2.1.1. Short summary of the data source in the modelling of FOC is given in Section 2.1.2. The review of popular machine learning model used to predict FOC is presented in Section 2.1.3. The performance of tree-based model, which include random forest and extra trees in various research will be discussed in Section 2.1.4. Brief summary of the literature review is presented in Section 2.1.6.

2.1.1 Modelling Approach for Ship Operation

According to Haranen et al. (2016) and Coraddu et al. (2017), the modelling strategies to predict fuel consumption are classified into three categories:

White Box Models (WBM)

Based on *a priori* mechanistic knowledge and physical principles of the vessel's system. This means that the dimensions of the vessel's structure, design parameters, and propulsion plant configuration are known.

Black Box Models (BBM)

Purely data driven, and it is developed using data from different sailing journey and historical observations. Contrary to WBM, this approach does not require detailed information on the vessel. This modelling approach can be further split into two categories. *Statistical Modelling* aims to find explanations for relationships between fuel consumption and different factors that affect it. *Machine Learning (ML) Modelling* focuses on the predictive capabilities of the model that could predict fuel consumption at different points in time.

Grey Box Models (GBM)

Fuse WBM and BBM into a single model that considers both *a priori* knowledge of the vessel and historical sailing data, This method aims to complement the performance of WBM and BBM.

Each of these strategies possesses its strength and limitations. WBMs are developed based on physical and hydrodynamics laws as well as theories of naval architecture, it is transparent and comprehensible, making them the preferred model used by various shipping industries. However, the deterministic nature of WBMs causes them to have poor suitability and generalisability. This is mainly caused due to limited *a priori* knowledge of different vessel dimensions, parameters, and narrow application limits of principle dimensions and form parameters of the vessel. Subsequently, the inability of WBMs to add randomness makes it rigid and restrictive. (Haranen et al., 2016; Yan et al., 2021)

BBMs in general have a good fitting ability for training data and good predictive accuracy for unseen data. BBMs developed using machine learning approach can generalise better compared to BBMs that are based on statistical modelling (Petersen et al., 2012a). BBMs are purely data driven, which means BBMs do not require former knowledge of vessel principle dimensions and form parameters. With increasing amount of data, better generalisation performance and handling of noisy data should be expected in a BBM. However, for the same reason, the quality of BBM model is highly dependent on data quantity and quality. For BBMs based machine learning approach, the amount of data is a major factor in determining the effectiveness of machine learning (Halevy et al., 2009). Data driven approach means that BBMs neglect basic vessel physical knowledge and are generally complex making it challenging to analyse and explain. For these reasons, experts in shipping industries are critical of models that do not include basic vessel knowledge and those that violate concepts of the domain knowledge in serious ways (Yan et al., 2021).

Hence, GBMs are introduced to address the limitations of both WBMs and BBMs by combining the mechanistic knowledge of the ship and physical principles of the vessel's system with BBM models, which possess good predictive capability. Despite these advantages, Yan et al. (2021) noted that GBM approach is not a common approach, recent research to predict fuel consumption are mainly dominated by BBM approach, specifically BBM based on machine learning approach.

2.1.2 Review of data source used for FOC model

The modelling of FOC using GBM requires both components of WBM and BBM. For the BBM modelling part using machine learning approach, it is especially important to ensure sufficient amount of good quality data to be available for model training to ensure precise and accurate training of the model (Halevy et al., 2009). It summarised by Yan et al. (2021), that the modelling of FOC use the following types of data source:

(Daily) Noon Report

Daily reports manually filed by ship's chief engineer and sent by the ship's masters to the shipping company and shore management. The reports include informations on types of daily fuel consumption, basic voyage information (e.g. ship location, load condition), sailing behaviour information e.g. (average sailing speed, average engine revolution per minute (RPM)), as well as sea and weather conditions. While it provides relevant information regarding the ship operation, the inherent problem of daily and manual data entry means that the quality and quantity of data cannot be guaranteed.

Sensor Data

Data obtained from installed sensor onboard the vessel. This may include fuel flow sensors, Global Positioning System (GPS) receiver and wind speed sensors are among the possible sensors that can be installed onboard a vessel. Sensor data address the issues of data quantity from noon report, as pointed out in the study by Gkerekos et al. (2019) for the prediction of daily FOC. The machine learning models, which are produced by the Automated Data Logging and Monitoring (ADLM) system outperforms the models that used noon data for their training by 5 – 7% for a collection period of 3 months of the ADLM system and 2.5 years for the noon data. However, installing onboard sensors may be complex and costly (Petersen, 2011) and the resulting sensor data will need to be handled properly to account for error in the sensors.

AIS Data

Apart from its intended use as collision avoidance system, AIS data have seen potential usage in ship behaviour analysis and environmental analysis. The Green House Gas (GHG) study by IMO (IMO, 2020; Smith et al., 2015), uses AIS to estimate global shipping emission inventories. Rakke (2016) proposed a methodology termed ECAIS to calculate ship emissions based on the fuel consumption from AIS data. Through Holtrop-Mennen approximation and literature approximation, the ship's power propulsion can be determined which is subsequently used to predict specific fuel consumption. Kim et al. (2020) used publicly accessible AIS data, ship static data, and environmental data to estimate EEOI using big data technology. Generally, the study using AIS data is done to achieve independence from the need to use commercial databases. The details of AIS data will be discussed in Section 2.3

2.1.3 Review of ML approach to predict FOC

Modelling of FOC using *machine learning* approach generally focus on prediction of unseen data. The general framework usually include collection and preprocessing of ship operational data, training and validation of the model, and evaluation and selection of the most appropriate model. Some machine learning models allow further hyperparameter tuning of the model and in case of data rich environment, the data can be further split into test data to further validate the performance of machine learning model.

The study by Yan et al. (2021) indicated that the majority of recent research that uses machine learning approach employ ANN as the model to predict FOC. ANN models are powerful models capable of modelling nonlinear data which are based on theories on how the brain works. The outcome is modelled by intermediate set of unobserved variables known as hidden layer. (Kuhn and Johnson, 2013). Back propagation neural networks, Multi Level Perceptron (MLP), and wavelet neural networks are some examples of ANN model subclasses.

ANN has shown respectable performance in its attempt to predict FOC. Petersen et al. (2012b) reported Root Mean Square Error of 47.2 L/h for the fuel flow i.e. FOC. To put this into context, the fuel flow in their case study fluctuates between $1000-2500$ L/h. Bal Beşikçi et al. (2016) considered sailing speed, trim, wind, sea effects, propeller pitch, and engine rotation speed as input variables to predict FOC per hour and achieved model fit score of $R^2 = 0.759$ in test set. Other studies also reported similar range of results using ANNs (Yan et al., 2021).

However, the development of ANN models is a challenging task. ANN models tend to overfit when there is shortage of data, as such, regularisation is necessary to improve model performance. The balancing process during regularisation is a demanding task and unsuitable regularisation may lead to counterintuitive prediction results. Adding layers is computationally expensive, and it does not always guarantee promising results (Hastie et al., 2009). Additionally, in machine learning terms, ANN is classified as a black box model, which makes it unintuitive and lacking in interpretability (Géron, 2019), this particular limitation cause shipping industry expert generally reluctant to accept the model generated using machine learning approach.

2.1.4 Tree-Based Model as FOC model

Concerning interpretability, modelling approaches such Linear Regression (LR), KNN and tree-based models have shown superior interpretability in comparison to ANNs. LR can explain the effect of each input variable on the output through the coefficients. KNN searches for the nearest neighbour and their closeness is evaluated through distance measurement algorithms such as Euclidean distance. Additionally, LRs and KNNs also offer easy implementation and adequate explainability. However, both approaches suffer from sensitivity to outliers and noise in data.

This brings us to tree-based model, a supervised, highly interpretable machine learning modelling approach capable of performing classification tasks for discrete data and regression tasks for continuous data. According to summary of Yan et al. (2021), it is not as popular as ANN, however some literature work and studies have indicated its benefits and performance superiority over other machine learning modelling approaches:

Soner et al. (2018) used the ferry dataset from Petersen et al. (2012b) to predict FOC using tree-based model, which includes bagging, random forest (RF), and bootstrap. From the test dataset, the random forest model achieved RMSE of 43.5 L/h for the fuel consumption. Which suggested improvement from ANN model from the study of Petersen et al. (2012b).

Yan et al. (2020) used random forest (RF) model to predict FOC for a voyage of a dry bulk ship using ship operational data i.e. ship noon data and sea and weather data from noon report and EMCWF. The prediction model considered ship sailing speed, total cargo weight and meteorological conditions and RF model obtained mean absolute percentage error (MAPE) of 7.91% for the FOC. The RF model displayed superior result in comparison to Decision Tree Regressor (DTR), ANN, LASSO, and SVR.

The advantage of tree-based model is further highlighted by Gkerekos et al. (2019). The study compared the performance of different machine learning models to predict ship's FOC per day using both noon data and automated data logging and monitoring (ADLM) system from a bulk carrier. This research concludes that tree-based model displayed good prediction performance on both noon data and sensor-based data. ETR achieved remarkable model fit score of 89% using the noon data and 97% when using the data from ADLM system, outperforming ANN, SVR, and RFR models.

Li et al. (2022) performed more extensive research on the effects of data fusions between meteorological data, ship voyage data, and AIS data on different machine learning models to predict the ship's FOC. The study classified ETR and RFR as tree-based model which is produced by *bagging ensemble strategy*. While AdaBoost (AB), Gradient Tree Boosting (GB), XGBoost(XG) and LightGBM (LB) are classified as tree-based models produced by *boosting ensemble strategy*. The study recommends all tree-based models that are produced by *boosting ensemble strategy* and ETR to be used to model energy efficient operation. Additionally, RFR shows the best robustness among the proposed model in the study.

Abebe et al. (2020) attempted to use machine learning approach to predict SOG of the ship. In this study, AIS data and noon-report weather data from 14 tracks and 62 ships are used for model training. The model considered the ship draught, ship dynamic information, tonnage, and environmental conditions. The result of this study exhibited the feasibility of using AIS data and meteorological data to predict SOG of the ship. The results also further indicated the strength of tree-based model, on test dataset, ETR achieved the best result with model fit of 98.47% and RMSE of 0,234 knots. It is also reported that ETR achieved better performance with about half of the computational cost of RFR.

2.1.5 Review of WBM for FOC prediction

To predict the FOC of a ship, WBMs usually calculate the resistances encountered by the vessel based on physics and hydrodynamic laws. The total resistance is summed from resistance of calm water resistances and additional resistance from wind, wave, and other external factors. The corresponding engine power at a particular speed can be calculated, and consequently the FOC can be calculated.(Haranen et al., 2016)

The methods from Guldhammer and Harvald (1974), Hollenbach (1999), and Kristensen and Lützen (2012) use different formulations, assumptions, and input variables for engine power estimation. For this thesis, the main focus will be the use of the estimation method from Holtrop-Mennen (Holtrop and Mennen, 1978, 1982; Holtrop, 1984). Holtrop-Mennen estimate method allowable application range is suitable in most of the cases. This is indicated by studies from Rakke (2016) and Kim et al. (2020). Rakke (2016) used ship operational and mechanical data from various works of literature and AIS data for input variables to estimate the engine power using Holtrop-Mennen method to subsequently calculate FOC. The FOC is then used to estimate GHG emissions for different ships and the study reported about 5% error rate during model testing. Kim et al. (2020) successfully estimated Energy Efficiency Operational Index (EEOI) without actual FOC. The study used AIS data as well as publicly accessible weather data and ship static information. The approach in this study used Holtrop-Mennen method to estimate engine power which is consequently used to calculate FOC for EEOI estimations.

2.1.6 Conclusion of Literature Review

As termed by Yan et al. (2021), the GBM model in this thesis falls under the category of sequential GBM, where the BBM and the WBM will be developed in series and combined to form a single GBM. The BBM will be developed to perform initial prediction and the resulting prediction will be passed into the WBM. The use of tree-based regressor, which will be used to predict SOG, provides solution to the problem of poor interpretability of some machine learning models. Furthermore, tree-based models can outperform most of the available machine learning models while providing added benefits of little requirement for data preprocessing and relatively cheap computational cost. The selection of Holtrop-Mennen as engine power estimation method is justified by the application range of the methodology and successes from previous studies.

2.2 Tree-based model

Decision Tree, Random Forest and Extra-Tree are classified as tree-based model, which is supervised machine learning model capable of classification tasks for discrete variables and regression tasks for continuous variables. In this section, the theory of Decision Tree (DT), Random Forest (RF) and Extra Tree (ET) will be discussed in detail in Section 2.2.1, Section 2.2.2 and Section 2.2.3.

2.2.1 Decision Tree

The principle of decision tree as a predictor can be defined as one or more nested `if-then` statements based on a rule that partitions the data into partition space as shown in Figure 2.1. Alternatively, the partition space generated from `if-then` statements can be represented using binary tree representation, which is more interpretable as multiple input response can be represented by a single tree.**(Kuhn and Johnson, 2013; Hastie et al., 2009)**

A decision tree consists of the following type of nodes, ***Root node*** defines the topmost node. ***Leaf nodes*** are also termed as terminal nodes, it is the node that will give the final prediction output. The ***Internal Node*** is defined as the nodes between the root node and leaf node. The process of dividing a node into successive nodes is called ***splitting***. The node that is being split is called ***parent node*** and the successive nodes that are created are called ***child nodes***. To grow a tree in a regression task, the splitting process is commonly regulated by Mean Square Error (MSE). The tree growth algorithm are based on Classification and Regression Tree (CART).

To understand the principle of selection for the feature, k_t , of the parent node and splitting rule, t_k , for data partition, the following example will be presented:

For the selection of the optimal splitting rule t_k: Given a case with single feature k and response y with m data points present. The algorithm starts by looking for possible splits between two distinct data points y. This split results in two distinct partition spaces. For each partition space S_1 and S_2, the mean is calculated by dividing the sum of response y with the amount of data points m for each respective partition space S_1 and S_2.

This step is then followed by calculating the sum of squared error (SSE) of each data point in partition space S_1 and S_2 and dividing it by the number of data points m_{s_1} and m_{s_2} respectively to obtain the MSE. Subsequently, the MSE from the respective partition space S_1 and S_2 is summed. The process is then recursively repeated until a threshold t_k that produces minimum sum of MSE is found, this threshold will be selected as splitting rule for the parent node and correspond to the threshold that minimise the cost function $J(k, t_k)$, with $\hat{y}_{S_i}$, being the mean of the response, y_{S_i}, in partition space S_i. **(Géron, 2019; Kuhn and Johnson, 2013)**:

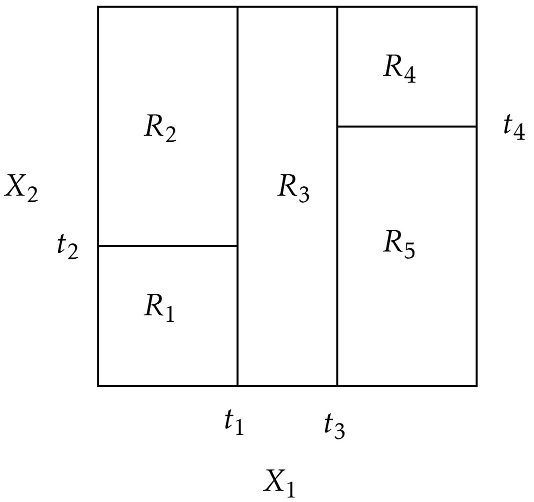


Figure 2.1: Example of partition space (Hastie et al., 2009)

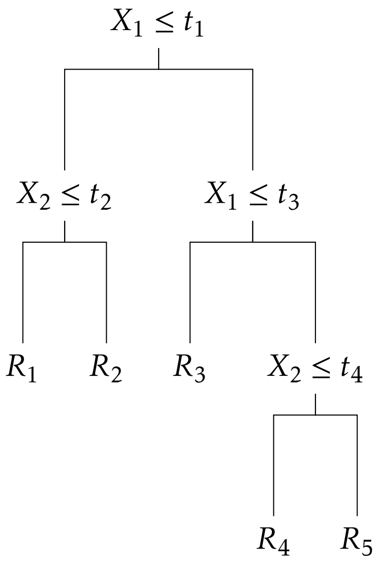


Figure 2.2: Example of partition tree (Hastie et al., 2009)

$$\mathrm{MSE}_{S_i} = \frac{1}{m_{S_i}}\mathrm{SSE}_{S_i} \quad \textbf{where} \quad i = (1,2) \tag{2.2.1}$$

$$J(k,t_k) = \frac{1}{m_{S_1}}\mathrm{SSE}_{S_1} + \frac{1}{m_{S_2}}\mathrm{SSE}_{S_2} \begin{cases} \mathrm{SSE}_{S_i} = \sum\limits_{i\in S_i}(\hat{y}_{S_i} - y_{S_i})^2 \\ \hat{y}_{S_i} = \frac{1}{m_{S_i}}\sum\limits_{i\in S_i} y \end{cases} \tag{2.2.2}$$

For the selection of the most optimal feature for parent node t_k: Similar principle is also applied for the selection of the most optimal feature for the parent node. Consider there are k_t features, then for each respective feature $k_1, k_2, \ldots, k_t$, The MSE for each of the features is calculated following the cost function $J(k,t_k)$. The feature that can best ***minimise*** the cost function will be selected as the root node of the tree. The subsequent selections of the feature for the parent node follow the same principle. **(Hastie et al., 2009; Géron, 2019).**

Once complete, then the partition space is further split into two more regions to look for the next possible split that minimise the cost function $J(k,t_k)$. This process is recursively continued until the number of samples to split falls under a certain threshold or when it cannot find a split that can further reduce MSE.

The resulting decisions for the best possible splits can be represented using binary tree, this makes decision tree highly interpretable and easy to implement. The inherent logic structure from `if-then` statements means that it can handle various types of data (sparse, skewed, continuous, categorical, etc.) without the need for data pre-processing. Decision tree implicitly conducts feature selection which is a desirable trait for many modelling problems **(Kuhn and Johnson, 2013).**

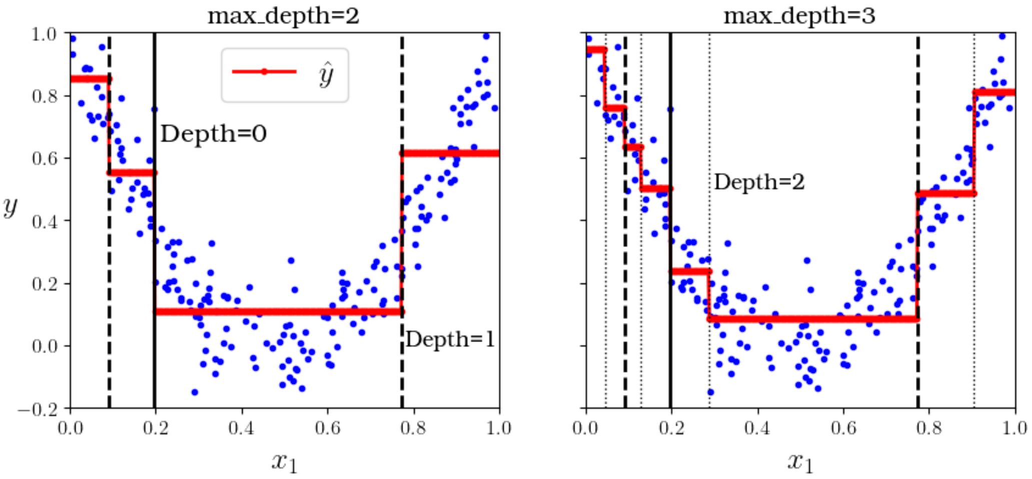


Figure 2.3: Prediction of two Decision tree regression models (Géron, 2019)

However, a single decision tree suffers from overfitting when the model is unconstrained. The logical principle of `if-then` statements means that decision tree will attempt to fit the training data as closely as possible. Furthermore, a single decision tree model tends to be unstable, altering the data will cause drastic changes in the structure of the tree, there exist possibilities where completely different sets of splits might be found resulting in different interpretations (Hastie et al., 2009; Kuhn and Johnson, 2013).

From Figure 2.1, it can be implied that each decision boundaries are orthogonal to an axis i.e. all splits are perpendicular to an axis and this form rectangular subspaces for each predicted value. If the relationship between predictors and response cannot be adequately defined by the rectangular subspaces, then tree based models will suffer from larger prediction error than other kinds of models (Kuhn and Johnson, 2013).

Therefore, it is necessary to regularise i.e., restrict the decision tree's freedom to grow during model training. Overfitting could be reduced by controlling how deep the tree can grow through the `max_depth` parameter. Additionally, setting the amount of minimum number of samples a leaf node has, through `min_samples_leaf` can alleviate overfitting as well, as shown in Figure 2.4. Other regularisation techniques will be discussed in Section 3.3.2.

Regularisation of decision tree will help to address the overfitting issues and improve the robustness of the model, this may result in better generalisation capability. Nonetheless, in order to attain significant improvements in the performance of decision tree model, it is necessary to seek alternative solutions.

2.2.2 Random Forest

Ensemble learning is one of the possible solutions to improve the performance of DT regressors. The main idea of ensemble learning is combining the strengths of a

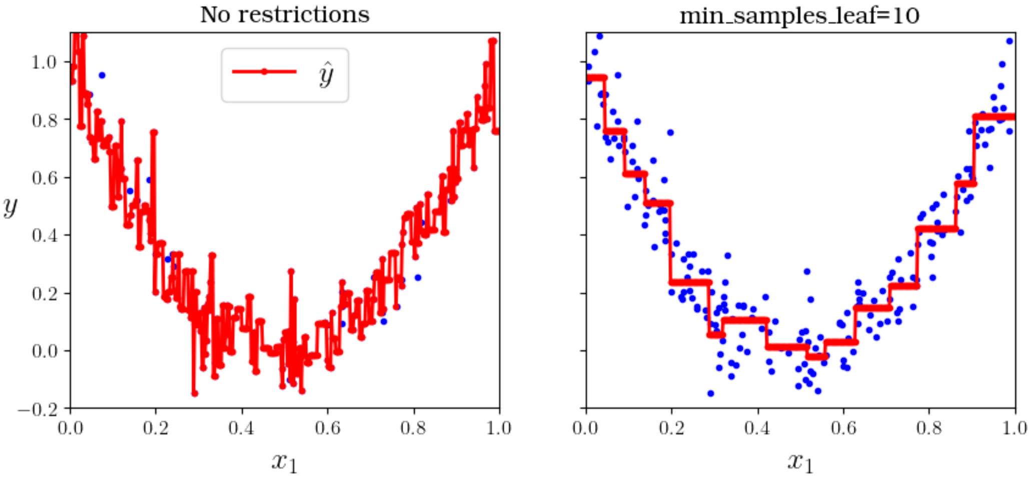


Figure 2.4: Regularising a Decision Tree regressor (Géron, 2019)

collection of simpler base models (Hastie et al., 2009). The algorithm, involving the creation of bootstrap samples, random selection of splitting feature and aggregation of the prediction is termed by Breiman (2001) as ***Random Forest***. It involves combination of multiple learning algorithms, known as weak learners. In random forest, each of these learners are individual decision tree.

The most common ensemble methods are *boosting* and *bagging*. In boosting, the learner evolves over time, where successive trees are dependent on the earlier trees. In bagging (short for *bootstrap aggregating*) each tree is trained using bootstrap sample of the training set i.e. this means that a sample of the training dataset is randomly selected and allowed to appear more than once[1]. Each model in the ensemble then generates a prediction from the bootstrapped sample and the predictions are aggregated across the learners (Tin Kam Ho, 1995; Breiman, 2001).

The performance of bagging can be further improved by reducing correlation between trees i.e. de-correlating trees. This can be achieved by adding randomness during tree construction process. Dietterich (2000) introduced the idea of random split selection, which means that a feature k will be selected from a random subset of feature. From this random subset, the assignment of the feature for the parent node follows the CART algorithm described in Equation (2.2.2). Further randomness is added by exploiting the instability of single decision tree mentioned in Section 2.2.1.

The methodology introduced in random forest address the tendency of decision tree to overfit and the issue of lack of robustness. De-correlating trees means that each learner is independent of each other, and the combination of many independent, strong learners yields an improvement in error rates i.e. reduction in variance and robustness against noisy response. It is also proven by Breiman (2001) that random forest cannot overfit, that means growing more trees should not affect the performance of random forest, albeit with a greater computational burden. Both Kuhn

[1] This sampling technique is referred to as sampling *with* replacement

and Johnson (2013) and Hastie et al. (2009) reported that remarkable prediction results can be obtained without extensive tuning of tree parameter.

However, random forest loses the benefit of interpretability of tree-based model. Due to ensemble nature of random forest, it is not possible to gain an understanding between the feature and the prediction. Nevertheless, it is still possible to quantify the impact of each feature in the ensemble (Kuhn and Johnson, 2013)[2]. Random forest also tends to perform poorly with small number of samples (Hastie et al., 2009). Nevertheless, it is possible to traverse through a single tree to see the path taken to reach the predicted value.

2.2.3 Extra-Trees (Extremely Randomised Trees)

Extra-trees (Extremely Randomised Trees) is introduced by Geurts et al. (2006) to further randomise random forest and further de-correlate the trees in the forest. Unlike random forest, which selects the optimal split by selecting the best feature among randomly selected subset of features, Extra-trees selects a split at random. Extra-trees also does not bootstrap the sample[3] and uses the whole training dataset. The random selection of split means that it saves computational power and the increase in variance caused by tree de-correlation can be countered by increasing the number of trees in the ensemble.

2.3 AIS Data

2.3.1 Overview of AIS

Automatic Identification System (AIS) is an automated tracking system onboard ships to automatically transmit information about the ship to other ships and coastal authorities, AIS was developed to avoid ship collision accidents. As part of the revised new chapter V of SOLAS[4] regulation, International Maritime Organization (IMO) requires all international voyage ships of 300 gross tonnage (GT) and upwards, cargo ships with 500 GT not engaged on international voyage, and all passenger ships irrespective of size to be equipped of AIS class A equipment (Yang et al., 2019; IMO, 2015).

AIS uses Very High Frequency (VHF) with special protocol for communication system for information exchange between the ships. This information will be received by either ships directly, buoys, Land based AIS transceivers (T-AIS) and satellites (S-AIS). The information transmitted by AIS is distinguished into three different types.

[2] Known as feature importances in `Scikit-Learn`

[3] This sampling technique is referred to as sampling *without* replacement

[4] International Convention for the Safety of Lives at Sea

Information Item	**Description**
Static	
MMSI	MMSI number of vessel
Callsign	Callsign of vessel
Name	Name of the vessel
IMO	IMO number of the vessel
Length	Length of vessel
Width	Width of vessel
Ship Type	Describes the AIS ship type of this vessel
Dynamic	
Ship's position	Automatically updated from position sensor connected to AIS. Longitude and Latitude.
Position time stamp in UTC	Automatically updated from ship's main position sensor. Format: DD/ MM/YYYY HH:MM:SS
Course over Ground (COG)	***If available***, automatically updated from ship's main position sensor connected to AIS.
Speed Over Ground (SOG)	***If available***, automatically updated from the position sensor connected to AIS.
Heading	Automatically updated from the ship's heading sensor connected to AIS
Navigational status	Navigational status information has to be manually entered by the Officer on Watch (OOW) and changed as necessary. For example : "*underway by engines*","*engaged in fishing*","*at anchor*".
Rate of Turn (ROT)	***If available***, Automatically updated from the ship's ROT sensor or derived from the gyro.
Voyage Related	
Ship's draught	To be manually entered at the start of the voyage using the maximum draft for the voyage and amended as required
(Hazardous) Cargo Type	Type of cargo from AIS message.
Destination and ETA	To be manually entered at the start of the voyage and kept up to date as necessary.

Table 2.1: Structure of AIS data (IMO, 2015)

Static information which is entered into the AIS on installation. **Dynamic information**, which is automatically updated from the ship's sensors connected to AIS and **voyage-related information**, which might need to be manually entered and updated during the voyage. The structure of the AIS data that is relevant to this thesis is summarised in Table 2.1(IMO, 2015).

AIS is also further differentiated by its equipment class. The classification is based on the reporting interval and the type of information that is conveyed. **Class A** autonomously report their position within 2-10 seconds interval, depending on the state of ship's movement. The reporting interval is less frequent at 3 minutes, When the ship is at anchor or moored and moving slower than 3 knots. Class A AIS is also capable of sending safety related information, meteorological and hydrological data, electronic broadcast to mariners and marine safety messages. **Class B** reports at longer interval and at a lower power. They can only receive safety related messages, not send them. (Rakke, 2016; IMO, 2015)

It is also stated by Yang et al. (2019) that AIS data can be combined with data from other databases to provide additional information such as:

- Port to port average speed, the voyage time can be calculated from the time stamps reported by AIS data; the voyage distance can be found from corresponding navigation distance tables.
- Cargo weight which can be estimated from draught and ship size.
- Technical ship specification from fleet database which can be derived from IMO number.
- Port to port bunker consumption which can be estimated based on the speed, technical ship specification and distance between two ports.

2.3.2 Speed Correction

The speed that is shown in AIS is the speed over ground (SOG). However, the ship actual speed i.e. speed through water (STW) will be required to calculate the bunker fuel consumption. Therefore, the SOG will need to be corrected for STW. This correction is performed by considering the current speed V_c and the direction of the current γ *with respect to True North*. In principle, STW will be greater than SOG, when the current is moving against the current as the ship tries to compensate for the current to maintain the SOG. Whereas, the STW will be greater than the SOG when the current is moving in the same direction of the ship movement.

To calculate the correction, this study will adopt the methodology proposed by Kim et al. (Kim et al., 2020) and Yang et al. (Yang et al., 2020). The x and y component of SOG can be obtained through vector decomposition using the ship's heading angle α *with respect to True North*. Similar vector decomposition is also performed for current speed V_{current}, it is resolved with current direction γ *with respect to True North*:

$$V_{\text{SOG}}^{x} = V_{\text{SOG}} \cdot \sin(\alpha) \tag{2.3.1}$$

$$V_{\text{SOG}}^{y} = V_{\text{SOG}} \cdot \cos(\alpha) \tag{2.3.2}$$

$$V_{\text{current}}^{x} = V_{\text{current}} \cdot \sin(\gamma) \tag{2.3.3}$$

$$V_{\text{current}}^{y} = V_{\text{current}} \cdot \cos(\gamma) \tag{2.3.4}$$

Then the resulting equation to determine STW, including the current compensation, is given by:

$$V_{\text{STW}}^{x} = V_{\text{SOG}}^{x} - V_{\text{current}}^{x} \tag{2.3.5}$$

$$V_{\text{STW}}^{y} = V_{\text{SOG}}^{y} - V_{\text{current}}^{y} \tag{2.3.6}$$

$$V_{\text{STW}} = \sqrt{(V_{\text{STW}}^{x})^2 + (V_{\text{STW}}^{y})^2} \tag{2.3.7}$$

2.3.3 Source of error in AIS

Errors and inaccuracies may still exist in AIS data. The main source of errors is caused by data that requires manual entry such as static information and voyage related information which include estimated time of arrival (ETA) and draught. There exist cases where MMSI is shared by different ships even though it is supposed to be unique. The data that is automatically connected by sensors can be erroneous, this may happen when the sensors are faulty or when it is not properly installed (Yang et al., 2019). Therefore, data preprocessing of AIS data is an important step to ensure correct representation of the ship state.

2.4 Weather data

During voyage, a vessel may encounter winds and waves from different directions with varying degree of magnitude. This may affect the vessel's path taken during the voyage and also ship performance such as speed and engine power, furthermore it may also affect the seakeeping capability of a vessel (Molland, 2011). It is important to consider different weather conditions to ensure accurate and precise estimation of required engine power by the vessel. With that in mind, the discussion in this section will focus on definition of wind and wave effects, as well as the relation between some of these parameters.

2.4.1 Definitions of weather parameters

Wind Waves and Swell

Wind Waves are also known as wind sea, wind waves are irregular and short-crested waves generated by local wind. **Swell** are waves that travel outside the wave generation area and are no longer the result of wind, they take on regular and long-crested appearance (Holthuijsen, 2007)

Significant Wave Height, $H_{1/3}$

It is defined as the mean of the highest one-third of waves in the wave record. The distribution of wave heights can be represented by probability density function. Hence, the term "highest one-third of waves" here means the region of wave heights that belong in the upper one-third of a probability density function, this is illustrated in Figure 2.5. From this distribution, the relation between significant wave height $H_{1/3}$, the highest ten percent of waves H_{10}, maximum wave height H_{max} and average wave height $\overline{H}$ can be summarised as follows (Bretschneider, 1965; Holthuijsen, 2007):

$$\overline{H} = 0.625 \cdot H_{1/3} \tag{2.4.1}$$

$$H_{10} = 2.03 \cdot \overline{H} = 1.27 \cdot H_{1/3} \tag{2.4.2}$$

$$H_{\max} = 2 \cdot H_{1/3} \tag{2.4.3}$$

Additionally, Bitner-Gregersen (2005) and Nielsen and Dietz (2020) described the relation between the significant wave height, wind wave height and swell height through following equation:

$$H_{1/3} = \sqrt{(H_{\text{swell}})^2 + (H_{\text{windwave}})^2} \tag{2.4.4}$$

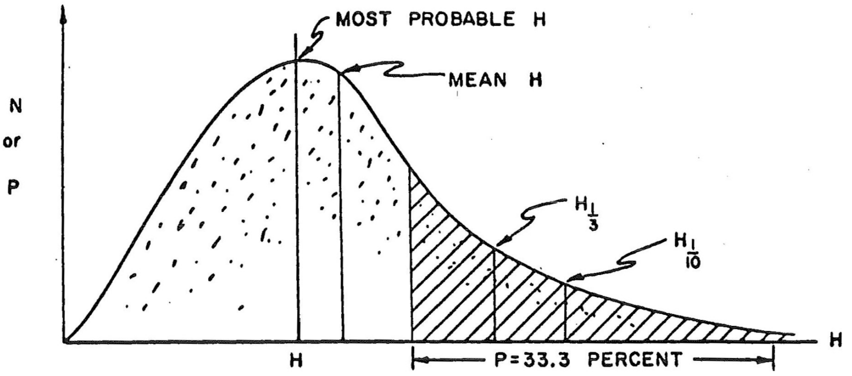


Figure 2.5: Statistical distribution of wave heights (Bretschneider, 1965)

Wave Period

Defined as the time interval between the start and the end of a wave. Some characteristics of wave period can be derived to define wave spectrum.

Wave Spectrum

The most important form in which ocean waves are described. Wave spectrum characterises all possible observations of the waves which include wave heights, frequencies i.e. period and wave direction. For example, Bitner-Gregersen (2005) stated that the state of the sea can be described through the significant height $H_{1/3}$ and spectral peak T_p with the help of Torsethaugen peak, given average wave spectrum T_f and constant $a_f = 6.6$ (Torsethaugen et al., 2004).

$$T_p = a_f \cdot H_{1/3} \tag{2.4.5}$$

$$\text{Sea State (SS)} = \begin{cases} \text{Swell dominated} & \textbf{if} \quad T_p > T_f \\ \text{Wind sea dominated} & \textbf{if} \quad T_p \leqslant T_f \end{cases} \tag{2.4.6}$$

2.5 General concept of ship propulsion

A ship's bunker fuel consumption in actual operating conditions is affected by several factors including the operating parameter of the ship's engine, propeller efficiency, and encountered resistance by the ship. Furthermore, a ship's propulsion power is correlated to the sailing speed (SOG) and meteorological conditions (**Lang, 2020**). Therefore, in addition to the calm water resistance R_{CALM}, the additional resistance caused by wind R_{AA} and wave R_{AW} should be considered to estimate the total resistance of the ship R_{TOTAL}. The power needed to propel a ship forward at a given ship STW v_S, to overcome R_{TOTAL} is defined as **effective power** P_e:

$$R_{TOTAL} = R_{TOTAL} + R_{AW} + R_{AA} \tag{2.5.1}$$

$$P_e = R_{TOTAL} \cdot v_S \tag{2.5.2}$$

The effective power P_e is transmitted through the shaft connected to the main engine of the ship which generates power to rotate the propeller of the ship, which is termed as **brake power of the engine,** P_b. The brake power can be calculated through effective power by considering the **shaft efficiency** η_s**, hull efficiency** η_h**, relative rotative efficiency** η_r **and open water efficiency** η_o:

$$P_b = \frac{P_e}{\eta_s \cdot \eta_h \cdot \eta_r \cdot \eta_o} \tag{2.5.3}$$

The bunker fuel consumption can then be calculated by multiplying the brake power P_b with the Specific Fuel Oil Consumption (SFOC) and the operation time:

$$FOC = P_b \cdot SFOC \cdot \mathcal{T}_{operation} \tag{2.5.4}$$

2.5.1 Ship dimensions and form coefficients

Principal Dimension of a vessel

The summary of important ship dimensions and parameters are shown in Figure 2.6 and Figure 2.7 (**Biran et al., 2014**):

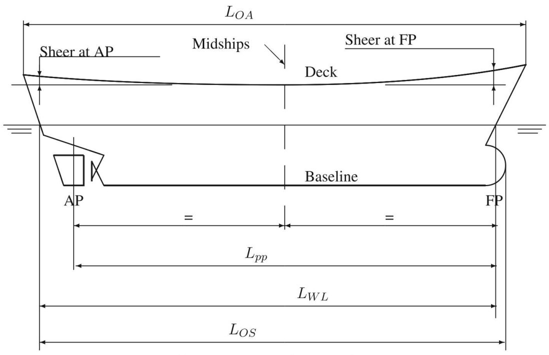


Figure 2.6: Side view of a vessel

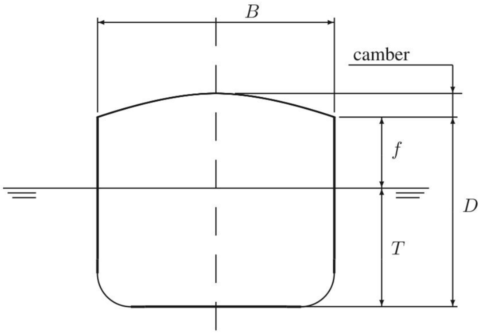
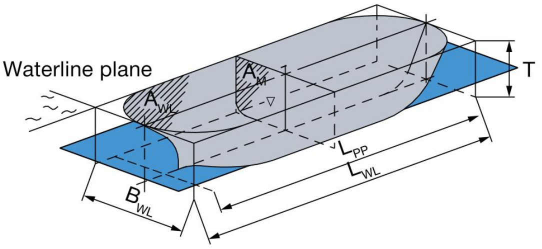


Figure 2.7: Front view of a vessel

The outer surface of the ship is usually not uniform as not all plates have the same thickness, Therefore the hull surface is measured with respect to the inner surface of the plating which is termed as *moulded surface* of the hull. All dimensions measured to this surface are defined as *moulded* dimensions whereas dimensions measured to the outer surface of the hull or of an appendage are qualified as *extreme* dimensions (**Biran et al., 2014**).

Coefficients of form

The form coefficients are non-dimensional numbers required to classify the hulls and to find relationships between forms and their properties, the summary of some important form coefficients are summarised in Figure 2.8



Volume of displacement : ∇

Waterline area : A_{WL}

Block coefficient, L_{WL} based : $C_{B,WL} = \dfrac{\nabla}{L_{WL} \times B_{WL} \times T}$

Midship section coefficient : $C_M = \dfrac{A_M}{B_{WL} \times T}$

Longitudinal prismatic coefficient : $C_P = \dfrac{\nabla}{A_M \times L_{WL}}$

Waterplane area coefficient : $C_{WL} = \dfrac{A_{WL}}{L_{WL} \times B_{WL}}$

Figure 2.8: Form coefficients (MAN, 2011)

Block Coefficient

Block Coefficient C_B is defined as the ratio of moulded displacement volume to the volume of parallelepiped (rectangular block) with dimensions L, B and T. Alternatively, **Schneekluth and Bertram** (**1998**) provided an estimation of the value using the Froude number within the range of $0.15 < Fr < 0.32$

$$C_b = -4.22 + 27.8\sqrt{Fr} - 39.1Fr + 46.6Fr^3 \tag{2.5.5}$$

The Froude number Fr is defined with the following equation:

$$Fr = \frac{v}{\sqrt{gL_{WL}}} \tag{2.5.6}$$

Midship Coefficient

Midship Coefficient C_M is defined as the ratio of the midship-section area A_M to the product of breadth and draught, BT. According to **Schneekluth and Bertram** (**1998**), changing C_M value will have an effect on separation resistance and wave resistance. **Jensen** (**1994**) presented a method based on regression equation on a graph to calculate C_M:

$$C_M = \frac{1}{1 + (1 - C_B)^{3.5}} \tag{2.5.7}$$

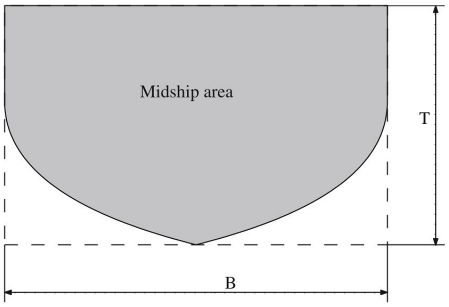


Figure 2.9: Definition of C_M (Biran et al., 2014)

Prismatic Coefficient

Prismatic Coefficient C_P is defined as the ratio of moulded displacement volume[5] V. It is an indicator on how much of a cylinder with constant section A_M and length L is filled with submerged hull as shown in Figure 2.10.

$$C_P = \frac{C_B}{C_M} \tag{2.5.8}$$

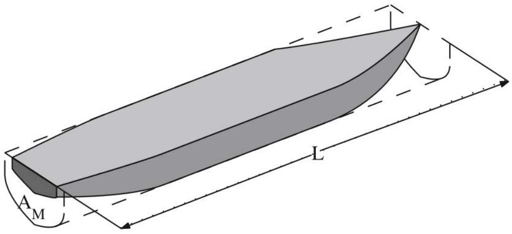


Figure 2.10: Definition of C_P (Biran et al., 2014)

Waterplane area coefficient

Waterplane area coefficient C_{WP} is defined as the ratio between the ship's waterline area A_W and the product of L and B. In ship design, C_{WP} significantly impacts resistance and stability (Schneekluth and Bertram, 1998). MAN (2011) approximated that C_{WP} is 0.10 higher than C_B, alternatively Schneekluth and Bertram (1998) provided the following formulation for C_{WP}:

$$C_{WP} = \frac{1 + 2C_B}{3} \tag{2.5.9}$$

2.5.2 Holtrop & Mennen's Method

This power prediction method was applied in the late 1970s and early 1980s by J. Holtrop and G.G.J Mennen and it was based on regression analysis of vast model tests and trial data of MARIN, the model basin in Wageningen, The Netherlands.

[5] In some notations it is denoted as ∇

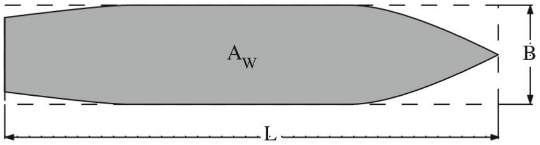


Figure 2.11: Definition of C_{WP} (Biran et al., 2014)

This gives Holtrop-Mennen method a wide applicability range and the only method that adopted the use of the ITTC form factor k. The resistances in this method are calculated as dimensional force. Furthermore, the method also gives estimates of hull-propeller interaction, thrust deduction, full-scale wake fraction and relative rotative efficiency (Birk, 2019).

Parameter	**Symbol**	**Remarks**
Required Parameters		
Length in waterline	L_{WL}	
Moulded breadth	B	
Moulded mean draught	T	typically $T = \frac{1}{2}(T_A + T_F)$
Moulded draught at aft perpendicular	T_A	
Moulded draught at forward perpendicular	T_F	
Volumetric displacement (molded)	V	alternatively use the block coefficient as $C_B = V/BTL_{WL}$
Prismatic coefficient (based on L_{WL})	C_P	
Midship section coefficient	C_M	or use $C_M = C_B/C_P$
Waterplane area coefficient	C_{WP}	may have to be estimated in early design stages
Longitudinal Centre of buoyancy	ℓ_{CB}	positive forward; with respect to $L_{WL}/2$ in percent of L_{WL}
Area of ship and cargo above waterline	A_V	projected in direction of v_S
Immersed transom area	A_T	measured at rest
Transverse area of bulbous bow	A_{BT}	Measured at forward perpendicular
Height of centre A_{BT} above basis	h_B	has to be smaller than $0.6T_F$
Propeller Diameter	D	
Propeller expanded area ratio	A_E/A_0	
Stern shape parameter	C_{stern}	
Optional Parameters		
Wetted surface (hull)	S	
Wetted Surface of appendages	S_{App}	bilge keels, stabiliser fins, etc.
Half angle of waterline entrance	i_E	
Diameter of bow thruster tunnel	d_{TH}	

Table 2.2: Required and optional input parameters for Holtrop & Mennen's method according to Birk (2019)

Application Range

The publication from Holtrop and Mennen (1978, 1982); Holtrop (1984) does not provide explicit information regarding the application range of the method. However, from the experience of Birk (2019), reasonable estimates from the method can be achieved for the following conditions:

$$\begin{gathered} Fr \leqslant 0.45 \\ 0.55 \leqslant C_p \leqslant 0.85 \\ 3.9 \leqslant \frac{L}{B} \leqslant 9.5 \end{gathered} \tag{2.5.10}$$

2.5.2.1 Calm water resistance

The calm water resistance R_{CALM} is broken down into several components and can be approximated using the following relation:

$$R_{CALM} = R_F(1+k_1) + R_{APP} + R_W + R_B + R_{TR} + R_A \tag{2.5.11}$$

Frictional Resistance R_F

R_F is calculated using the ITTC-1957 frictional resistance correlation line C_F as the basis of a representation of a resistance plate with a wetted surface area S of bare hull.

$$R_F = \frac{1}{2}\rho v_S^2 S C_F \tag{2.5.12}$$

The frictional coefficient C_F can be calculated through the Reynold number Re for a given ship speed v_S and kinematic viscosity ν:

$$C_F = \frac{0.075}{[\log_{10}(Re) - 2]^2} \quad \textbf{where} \quad Re = \frac{v_S L_{WL}}{\nu} \tag{2.5.13}$$

If not known, then the wetted surface area of bare hull S can be estimated by the following formula:

$$S = c_{23} L_{WL}(2T + B)\sqrt{C_M} + 2.38\frac{A_{BT}}{C_B} \tag{2.5.14}$$

with the factor c_{23} given as :

$$c_{23} = \left[0.453 + 0.4425C_B - 0.2862C_M - 0.003467\frac{B}{T} + 0.3696C_{WP}\right] \tag{2.5.15}$$

The flat plate resistance is subsequently adjusted by including a form factor k during the calculation of total resistance. The constant c_{14} must be determined first to calculate form factor k, which serves the purpose of capturing the impact of the aft body shape.

$$c_{14} = 1.0 + 0.011C_{stern} \quad \textbf{with} \tag{2.5.16}$$

Aft body shape	C_{stern}
Pram with gondola	−25
V-shaped sections	−10
Normal sections	0
U-shaped sections	+10

To complete the required input for the calculation of $(1 + k_1)$, the length of run L_R can be estimated from the following equation:

$$L_R = L_{WL}(\frac{1 - C_P + 0.06 C_P \ell_{CB}}{4C_P - 1}) \tag{2.5.17}$$

The formula by **Guldhammer and Harvald (1974)** can be used if ℓ_{CB} is not known:

$$\ell_{CB} = -(0.44Fr - 0.094) \tag{2.5.18}$$

Then, the form factor $(1 + k_1)$ can be determined with the constant c_{14}, the length of run L_R and input values from Table 2.2.

$$1 + k_1 = 0.93 + 0.487118 c_{14} \left[\left(\frac{B}{L_{WL}}\right)^{1.06806} \left(\frac{T}{L_{WL}}\right)^{0.46106} \left(\frac{L_{WL}}{L_R}\right)^{0.121563} \left(\frac{L_{WL}}{V}\right)^{0.36486} (1 - C_p)^{-0.604247} \right] \tag{2.5.19}$$

Appendage Resistance

An appendage is defined as addition to the main part or main structure of a vessel (**Molland, 2011**). Examples of appendages include rudders, shaft brackets, skeg and bilge keels. The form factors associated with these appendages, denoted as k_{2_i} are presented in Table 2.3. In practice, reasonable estimates can be made based on these form factors, as model tests are not the most suitable method for accurately quantifying appendage resistance. Furthermore, effects of appendages are typically considered as a whole and not as individual unit (**Birk, 2019**).

Appendage	k_{2_i} value
rudder behind skeg	0.2 – 0.5
rudder behind stern	0.5
twin screw rudder (slender)	1.5
twin screw rudder (thick)	2.5
shaft brackets	2.0 – 4.0
skeg	0.5 – 1.0
strut bossing	2.0 – 3.0
hull bossing	1.0
exposed shafts (angle with buttocks about 10 degrees)	1.0
exposed shafts (angle with buttocks about 20 degrees)	4.0
stabiliser fins	1.8
dome	1.7
bilge keels	0.4

Table 2.3: Approximate values for appendage form factors k_{2_i}

The equivalent form factor for multiple appendages, $(1 + k_{2_i})_{eq}$ is given by:

$$(1 + k_{2_i})_{eq} = \frac{\sum_i (1 + k_{2_i}) S_{APP_i}}{\sum_i S_{APP_i}} \tag{2.5.20}$$

If bow thruster is present, the resistance due to the bow thruster tunnel R_{TH} can be obtained through:

$$R_{TH} = \rho v_S^2 \pi \mathrm{d}_{TH}^2 C_{D_{TH}} \quad \textbf{where} \quad C_{D_{TH}} = 0.003 + 0.003\left(\frac{10_{d_{TH}}}{t} - 1\right) \tag{2.5.21}$$

The coefficient $C_{D_{TH}}$ defines the drag coefficient for the tunnel, and it ranges between 0.003 and 0.012. Smaller values indicate thrusters which are in the cylindrical part of bulbous bow. The coefficient can also be estimated using the equation by **Hollenbach (1999)** in Equation (2.5.21).

With that, the appendage resistance R_{APP} can be calculated using:

$$R_{APP} = \frac{1}{2}\rho v_S^2 (1 + k_{2_i})_{eq} C_F \sum_i S_{APP_i} + \sum R_{TH} \tag{2.5.22}$$

Wave Resistance

The estimation of wave resistance R_W is dependent on Froude number Fr, and it is subdivided into three categories.[6]:

$$R_W(Fr) = \begin{cases} R_{W_a}(Fr) & \textbf{if} \quad Fr \leqslant 0.4 \\ \text{Interpolation} & \textbf{if} \quad 0.4 < Fr \leqslant 0.55 \\ R_{W_b}(Fr) & \textbf{if} \quad Fr > 0.5 \end{cases} \tag{2.5.23}$$

The wave resistance for $R_{W_a}(Fr)$ can be calculated using:

$$R_{W_a}(Fr) = c_1 c_2 c_5 \rho g V \exp\left[m_1 Fr^d + m_4 \cos(\lambda Fr^{-2})\right] \tag{2.5.24}$$

And consequently for $R_{W_b}(Fr)$:

$$R_{W_b}(Fr) = c_{17} c_2 c_5 \rho g V \exp\left[m_3 Fr^d + m_4 \cos(\lambda Fr^{-2})\right] \tag{2.5.25}$$

The remaining range of Froude number between $0.4 < Fr \leqslant 0.55$ are calculated by mean of interpolation between equation Equation (2.5.24) and Equation (2.5.25). However, this range of Froude number is considered uneconomical and ship does not operate in this speed range for extended durations **(Birk, 2019)**.

$$R_W(Fr) = R_{W_a}(0.4) + \frac{20Fr - 0.8}{3}\left[R_{W_b}(0.55) - R_{W_a}(0.4)\right] \tag{2.5.26}$$

To compute each of the constants in Equation (2.5.24), The following equations are presented, note that the calculations for the $\cos(\lambda Fr^{-2})$ are in **Radians**:

[6] Considering the length of the equations, only the scenario where $Fr \leqslant 0.4$ will be thoroughly examined in this thesis. The formulations of R_W for other ranges of Froude number can be referenced in the studies by **Holtrop (1984)** and **Birk (2019)**

$$c_7 = \begin{cases} 0.229577\left(\dfrac{B}{L_{WL}}\right)^{\frac{1}{3}} & \textbf{if} \quad \dfrac{B}{L_{WL}} \leqslant 0.11 \\ \dfrac{B}{L_{WL}} & \textbf{if} \quad 0.11 < B/L_{WL} \leqslant 0.25 \\ 0.5 - 0.0625\dfrac{L_{WL}}{B} & \textbf{if} \quad B/L_{WL} > 0.25 \end{cases} \tag{2.5.27}$$

$$c_1 = 2223105 c_7^{3.78613}\left(\frac{T}{B}\right)^{1.07961} (90 - i_e)^{1.37565} \tag{2.5.28}$$

i_e is defined as half angle of the waterline entrance and the estimation can be calculated by:

$$i_e = 1 + 89e^a \tag{2.5.29}$$

and a can be obtained through:

$$a = -\left[\left(\frac{L_{WL}}{B}\right)^{0.80856}\left(1 - C_{WP}\right)^{0.30484}\left[1 - C_P - 0.0225\ell_{CB}\right]^{0.6367} \left(\frac{L_R}{B}\right)^{0.34574}\left(\frac{100V}{L_{WL}^3}\right)^{0.16302}\right] \tag{2.5.30}$$

$$c_3 = 0.56\frac{A_{BT}}{\left[BT\left(0.31\sqrt{A_{BT}} + T_F + h_B\right)\right]} \tag{2.5.31}$$

$$c_2 = e^{(-1.89\sqrt{c_3})} \tag{2.5.32}$$

$$c_{15} = \begin{cases} -1.69385 & \textbf{if} \quad \dfrac{L_{WL}^2}{V} \leqslant 512 \\ -1.69385 + \dfrac{\frac{L_{WL}}{V^{(1/3)}} - 8}{2.36} & \textbf{if} \quad 512 < \dfrac{L_{WL}^2}{V} \leqslant 1726.91 \\ 0 & \textbf{if} \quad \dfrac{L_{WL}^2}{V} > 1726.91 \end{cases} \tag{2.5.33}$$

$$c_{16} = \begin{cases} 8.07981C_P - 13.8673C_P^2 + 6.984338C_P^3 & \textbf{if} \quad C_P \leqslant 0.8 \\ 1.73014 - 0.7067C_P & \textbf{if} \quad C_P > 0.8 \end{cases} \tag{2.5.34}$$

$$d = -0.9 \tag{2.5.35}$$

$$\lambda = \begin{cases} 1.446C_P - 0.03\dfrac{L_{WL}}{B} & \textbf{if} \quad \dfrac{L_{WL}}{B} \leqslant 12 \\ 1.446C_P - 0.36 & \textbf{if} \quad \dfrac{L_{WL}}{B} > 12 \end{cases} \tag{2.5.36}$$

$$m_1 = 0.0140407C_P - 0.03\frac{L_{WL}}{B} - 1.75254\frac{V^{(1/3)}}{L_{WL}} - 4.79323\frac{B}{L_{WL}} - c16 \tag{2.5.37}$$

$$m_4 = 0.4c_{15}\exp\left(-0.034Fr^{-3.29}\right) \tag{2.5.38}$$

Resistance of bulbous bow

The approximation of the resistance due to bulbous bow R_B can be obtained through the immersion Froude number F_{r_i} for the bulbous bow and the constant P_B which is a measure of the emergence of the bow:

$$Fr_i = \frac{v_S}{\sqrt{g(T_F - h_b - 0.25\sqrt{A_{BT}}) + 0.15v_S^2}} \tag{2.5.39}$$

$$P_B = 0.56\frac{\sqrt{A_{BT}}}{T_F - 1.5h_B + h_F} \tag{2.5.40}$$

$$R_B = 0.11\rho g(\sqrt{A_{BT}})^3\frac{Fr_i^3}{1 + Fr_i^2}e^{(-3.0P_B^{-}2)} \tag{2.5.41}$$

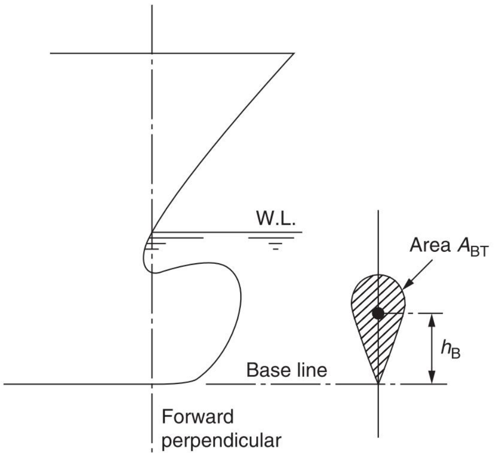


Figure 2.12: Bulbous bow definition (Molland, 2011)

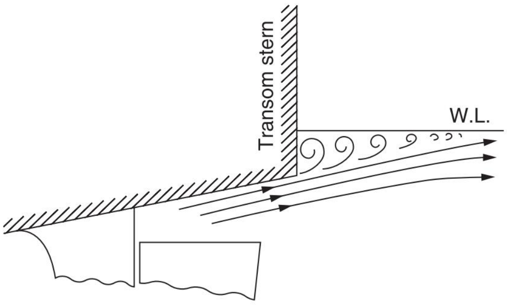


Figure 2.13: Flow around immersed transom stern (Molland, 2011)

(Immersed) Transom Resistance

The flat section at ship stern is termed as transom and the immersion of the transom causes pressure loss, the resulting resistance from the pressure loss are considered using the term R_{TR} for immersed transom area $A_T > 0$. The transom resistance is a function of the depth Froude number Fr_T:

$$Fr_T = \frac{v_S}{\sqrt{\frac{2gA_T}{(B+BC_{WP})}}} \tag{2.5.42}$$

The expression $A_T/(B+BC_{WP})$ is a measure for the average draught of the transom. When the average draught is smaller than the speed, there will be clean separation of the flow at transom edge and the resistance due to transom vanishes. Immersion resistance R_{TR} is considered if $Fr_T > 5$

$$c_6 = \begin{cases} 0.2(1-0.2Fr_T) & \textbf{if} \quad Fr_T < 5 \\ 0 & \textbf{if} \quad Fr_T > 5 \end{cases} \tag{2.5.43}$$

$$R_{TR} = \frac{1}{2}\rho v_S^2 A_T c_6 \tag{2.5.44}$$

Correlation allowance resistance

The resistance term R_A considers other effects that are not captured by other resistance components.

$$c_4 = \begin{cases} \dfrac{T_F}{L_{WL}} & \textbf{if} \quad \dfrac{T_F}{L_{WL}} \leqslant 0.04 \\ 0.04 & \textbf{if} \quad \dfrac{T_F}{L_{WL}} > 0.04 \end{cases} \tag{2.5.45}$$

The correlation allowance coefficient C_A and subsequent correlation resistance is defined as:

$$C_A = 0.00546(L_{WL}+100)^{-0.16} - 0.00205 + 0.003\frac{L_{WL}}{7.5}C_B^4 c_2(0.04-c_4) \tag{2.5.46}$$

$$R_A = \frac{1}{2}\rho v_S^2 C_A(S + \sum S_{APP}) \tag{2.5.47}$$

2.5.2.2 Added resistance due to wind

The magnitude of added resistance caused by wind, R_{AA}, is determined by the area of the ship superstructure and relative wind. Therefore, for a ship with large lateral areas above the water level, this added resistance due to wind can be significant. The estimation of added resistance due to wind in this thesis consider the method by Blendermann (1994):

$$R_{AA} = \frac{\rho_{air}}{2}u^2 A_L CD_l \frac{\cos(\varepsilon)}{1-\frac{\delta}{2}(1-\frac{CD_l}{CD_t}\sin^2(2\varepsilon))} \tag{2.5.48}$$

Where u is the apparent wind velocity, A_L the lateral plane area, ε the apparent wind angle ($\epsilon = 0$ in headwind). δ the cross-force parameter, and coefficients CD_t and CD_l the non-dimensional drag in beam wind and headwind. For given true wind velocity u_{TW} and true wind angle TWA, The calculation for the apparent wind u and apparent wind angle ϵ is performed using the following equations:

$$u = \sqrt{u_{\mathrm{TW}}^2 + v_S^2 + 2\cdot u_{\mathrm{TW}}\cdot v_S \cdot \cos(\mathrm{TWA})} \tag{2.5.49}$$

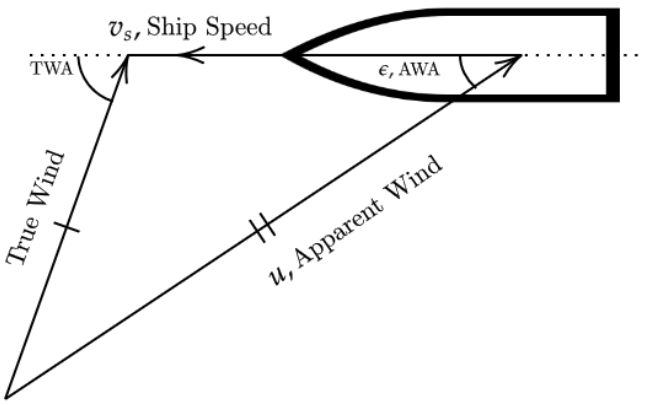


Figure 2.14: Apparent and true wind (Knudsen, 2013)

$$\frac{u_{TW}}{\sin(\varepsilon)} = \frac{u}{\sin(\mathrm{TWA})} \tag{2.5.50}$$

According to Schneekluth and Bertram (1998), maximum wind resistance is encountered when $0° < \varepsilon < 20°$ and it is more convenient to express the longitudinal drag with respect to the frontal area A_F. Typical values for the constants are summarised in Table 2.4

$$CD_{lAF} = CD_l \frac{A_L}{A_F} \tag{2.5.51}$$

	CD_t	CD_{lAF}	δ
Car carrier	0.95	0.55	0.8
Cargo ship, container on deck, bridge aft	0.85	0.65/0.55	0.40
Containership, loaded	0.90	0.55	0.40
Ferry	0.90	0.45	0.80
LNG Tanker	0.70	0.60	0.50
Passenger liner	0.90	0.40	0.80
Speed boat	0.90	0.55	0.60
Tanker, loaded	0.70	0.90	0.40
Tanker, in ballast	0.70	0.75	0.40

Table 2.4: Coefficients to estimate wind resistance

2.5.2.3 Added resistance due to wave

The added resistance due to wave, R_{AW} is estimated using the STAWAVE-1 method recommended by ITTC (2014). This method only considers waves encountered within the bow sector i.e. within ±45° off bow and does not consider wave correction for other encounters. Also, STAWAVE-1 is valid for the following condition:

$$\text{Significant wave height:} \quad H_{1/3} = 2.25 \leqslant \sqrt{L_{PP}/100} \tag{2.5.52}$$

$$R_{AWL} = \frac{1}{16}\rho g H_{1/3}^2 B \sqrt{\frac{B}{L_{BWL}}} \tag{2.5.53}$$

In which, L_{BWL} is the length of bow on the water line to 95% of maximum breadth.

2.5.2.4 Efficiencies affecting brake power

Open water efficiency

The open water efficiency η_O, can be understood as the propeller working in open water conditions i.e. the propeller operates in a homogenous wake field with no hull in front of it. The curve of different propulsion devices with its respective efficiencies is summarised in the work of Breslin and Andersen (1994):

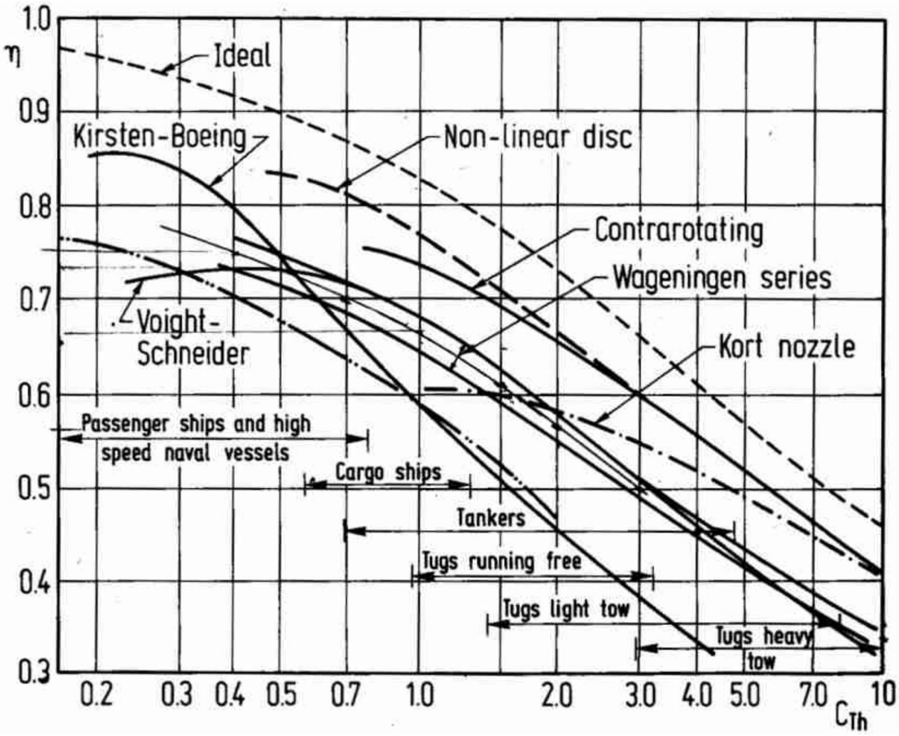


Figure 2.15: Efficiencies of various propulsion devices (Breslin and Andersen, 1994)

Hull efficiency

The Hull efficiency η_H can be calculated using the following equation:

$$\eta_H = \frac{1-t}{1-w_S} \tag{2.5.54}$$

The term t refers to the thrust deduction fraction, which represents the thrust force required to overcome the towing resistance of the ship R_{TOTAL} and the additional resistance caused by the propeller's interaction with the hull. On the other hand, the term w_S corresponds to the wake fraction, characterising the influence of the ship's hull on the water flow into the propeller (MAN, 2011; Birk, 2019). The following equations are presented for twin-screw vessels[7] to calculate w_S and t.

$$w_S = 0.3095C_B + 10C_V C_B - 0.23\frac{D}{\sqrt{BT}} \tag{2.5.55}$$

$$t = 0.325C_B - 0.1885\frac{D}{\sqrt{BT}} \tag{2.5.56}$$

where C_V is the viscous resistance coefficient, which combines all friction-related components of the resistance and the correlation resistance:

$$C_V = \frac{(1+k_1)R_F + R_{APP} + R_A}{\frac{1}{2}\rho v_S^2 (S + \sum_i S_{APP_i})} \tag{2.5.57}$$

Relative rotative efficiency

The relative rotative efficiency η_R can be expressed by the following ratio, with V_A defined as the arriving water velocity to propeller (MAN, 2011):

$$\eta_R = \frac{\text{Power absorbed in open water at } V_A}{\text{Power absorbed in wake behind the ship at } V_A} \tag{2.5.58}$$

According to Holtrop and Mennen (1982), η_R for twin screw vessels can be estimated using the following formula, with P/D defined as the propeller pitch to diameter ratio:

$$\eta_R = 0.9737 + 0.111(C_P - 0.0225\ell_{CB}) - 0.06325\frac{P}{D} \tag{2.5.59}$$

Shaft efficiency

The shaft efficiency η_S is defined as the ratio between the power delivered to the propeller P_D and the brake power of the main engine P_B, with values ranging from $\eta_S = 0.95 - 0.99$ depending on shaft design and gear configuration.

[7] Considering the length of the equations, the equations for single screw vessels can be obtained from Holtrop and Mennen (1982) and Birk (2019)

Chapter 3

Research Methodology

The methodology used to develop the grey box model will be discussed in this chapter. The grey box modelling approach in this thesis falls under the category of sequential GBM. Hence, the development process is divided into two stage. The first stage of the modelling focus on machine learning modelling i.e. BBM using tree-based model using python with help of `Scikit-Learn` (Pedregosa et al., 2011). This includes the process of data acquisition, data preprocessing, hyperparameter optimisation and model evaluation. The training of the models utilises a fusion of T-AIS data and weather data, then, relevant features are selected to predict the SOG. These processes are visually presented in Figure 3.1.

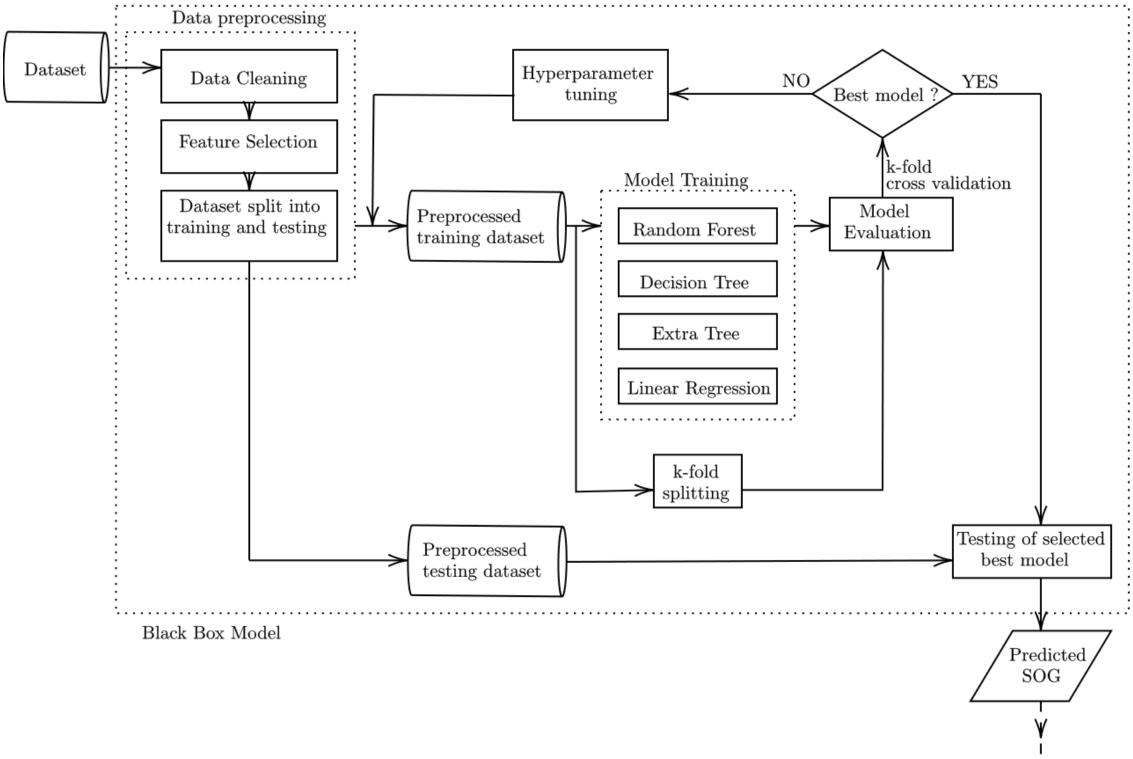


Figure 3.1: Scheme of proposed BBM methodology

The second stage of the modelling focus on WBM aspect of the GBM. The predicted SOG will be fed into the WBM to predict the required brake power required to propulse the ship. The SOG will need to be first converted into STW for the resistance calculations and consequently power calculations. The framework presented in Figure 3.2 refers is a graphical summary of Section 2.5.

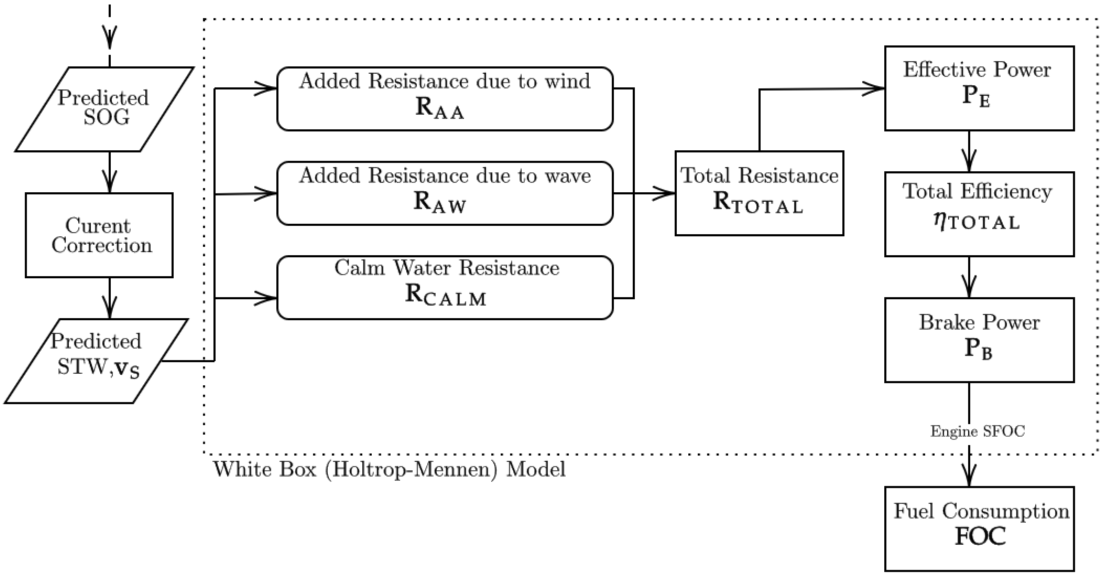

Predicted SOG
Curent Correction
Predicted STW,$\mathbf{v_S}$
Added Resistance due to wind $\mathbf{R_{AA}}$
Added Resistance due to wave $\mathbf{R_{AW}}$
Calm Water Resistance $\mathbf{R_{CALM}}$
Total Resistance $\mathbf{R_{TOTAL}}$
Effective Power $\mathbf{P_E}$
Total Efficiency $\eta_{\mathbf{TOTAL}}$
Brake Power $\mathbf{P_B}$
Engine SFOC
Fuel Consumption FOC
White Box (Holtrop-Mennen) Model

Figure 3.2: Scheme of proposed WBM methodology adopted from Lang (2020)

The development process of GBM is summarised in Figure 3.3. Detailed discussion regarding the development of BBM and WBM model will be discussed in the following sections of this chapter.

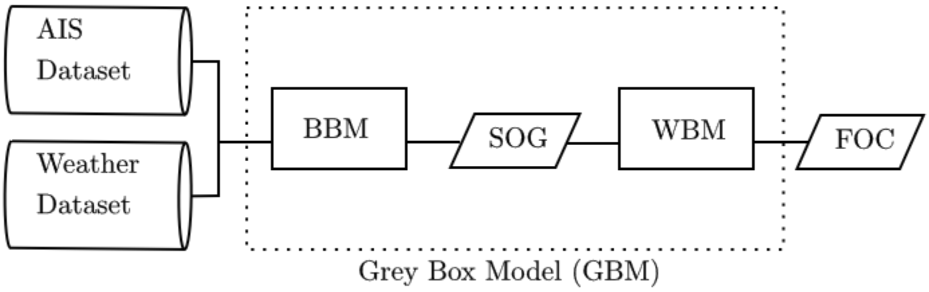


Figure 3.3: Scheme of proposed GBM methodology

3.1 Data Acquisition

The data is collected from a ferry serving between ports of Køge, Rønne, Ystad and Sassnitz, as shown in Figure 3.5 (Commons, 2010). The trip between Køge, Rønne takes about 5 h 30 minutes and it sails between Rønne and Sassnitz for 3 h and

20 minutes. The journey is tracked by T-AIS system of Danish Maritime Authority (DMA). The weather data along her sailing path are acquired from ECMWF[1] with temporal resolution of 1 hour at granularity of 0.25° (longitude) x 0.25° (latitude), data from ECMWF provides information for wind, waves and seawater temperature. The information for current is obtained from CMEMS[2] with temporal resolution of 3 hours at granularity of 0.25° (longitude) x 0.25° (latitude).

IMO	9812107
Type & Service	Passenger ferry
L_{OA}	158.00 m
L_{WL}	144.80 m
B (moulded)	24.5 m
T_{DESIGN}	5.70 m
T_{MAX}	5.85 m
Gross Tonnage (GT)	18,009
Deadweight (dwt)	4,830 t
Main Engines	Wärtsillä 8V31 2 x 4,880 kW
SFOC	169.4 g/kWh
Service Speed	17.7 knots
Bow Thrusters	2 x 1500 kW

Figure 3.4: Particular of M/S Hammershus

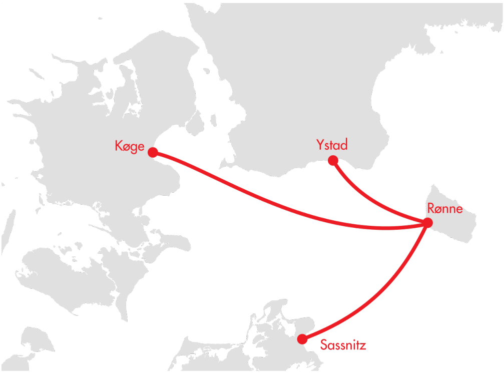


Figure 3.5: Journey of the ferry

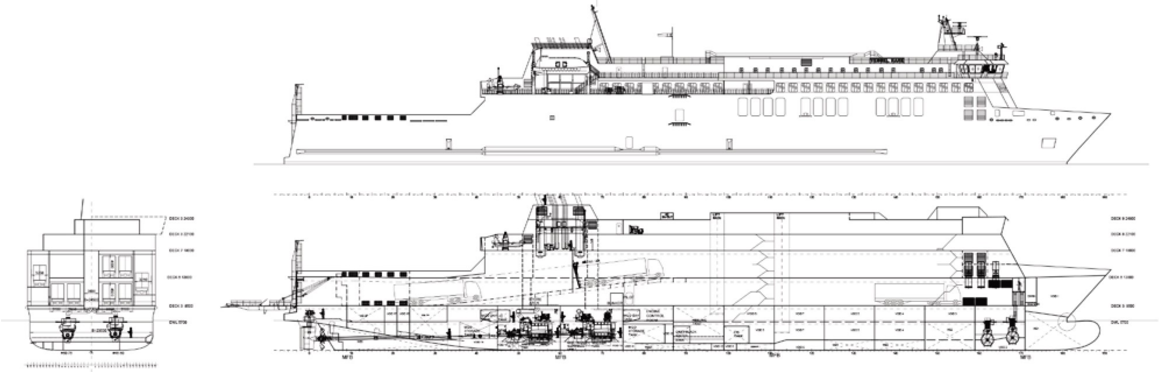


Figure 3.6: Schematics of M/S Hammershus

The resulting fused dataset has a temporal resolution of 1 hour. Due to the difference in temporal resolution of the data from CMEMS and ECMWF, the weather information is synchronised so that the wind, waves, seawater temperature and sea current belongs to the same weather grid with same temporal resolutions. The features **wind direction**,**swell direction**, and **wind wave direction** are oriented to true north. However, to reflect the actual direction of weather effects that are acting on the ship, these features are converted to true direction; where true direction is defined as the direction of weather effect with respect to the bow of the ship. The value ranges between 0° and 180°. Subsequently, through vector decomposition, the northward and eastward wind velocity is converted to absolute wind speed and wind direction *with respect to True North*,φ:

[1] European Centre for Medium-Range Weather Forecast

[2] Copernicus Marine Environment Monitoring Service

$$V_{\text{wind}} = \sqrt{(V^N_{\text{wind}})^2 + (V^E_{\text{wind}})^2} \tag{3.1.1}$$

$$\varphi = \begin{cases} 360 - \arctan(\frac{V^E_{\text{wind}}}{V^N_{\text{wind}}}) & \textbf{if} \quad V^E_{\text{wind}} > 0 \quad \wedge \quad V^N_{\text{wind}} < 0 \\ 180 - \arctan(\frac{V^E_{\text{wind}}}{V^N_{\text{wind}}}) & \textbf{if} \quad V^E_{\text{wind}} < 0 \quad \wedge \quad V^N_{\text{wind}} > 0 \\ 270 - \arctan(\frac{V^E_{\text{wind}}}{V^N_{\text{wind}}}) & \textbf{if} \quad V^E_{\text{wind}} > 0 \quad \wedge \quad V^N_{\text{wind}} > 0 \\ \arctan(\frac{V^E_{\text{wind}}}{V^N_{\text{wind}}}) & \textbf{otherwise} \end{cases} \tag{3.1.2}$$

Similarly, information of Northward and Eastward current Velocity is converted to absolute current speed and current direction *with respect to True North γ.*

$$V_{\text{current}} = \sqrt{(V^N_{\text{current}})^2 + (V^E_{\text{current}})^2} \tag{3.1.3}$$

$$\gamma = \begin{cases} 360 - \arctan(\frac{V^E_{\text{current}}}{V^N_{\text{current}}}) & \textbf{if} \quad V^E_{\text{current}} < 0 \quad \wedge \quad V^N_{\text{current}} > 0 \\ 180 - \arctan(\frac{V^E_{\text{current}}}{V^N_{\text{current}}}) & \textbf{if} \quad V^E_{\text{current}} > 0 \quad \wedge \quad V^N_{\text{current}} < 0 \\ 270 - \arctan(\frac{V^E_{\text{current}}}{V^N_{\text{current}}}) & \textbf{if} \quad V^E_{\text{current}} < 0 \quad \wedge \quad V^N_{\text{current}} < 0 \\ \arctan(\frac{V^E_{\text{current}}}{V^N_{\text{current}}}) & \textbf{otherwise} \end{cases} \tag{3.1.4}$$

The initial dataset provides information to the true directions, which indicate the encounter direction of weather conditions with respect to the ship bow. Static information from the AIS data which only indicated the ship's identity and navigational status are discarded in the dataset. The initial structure have 27 features, 9 AIS features and 18 weather features. The structure of the initial dataset i.e. before data preprocessing and feature selection, is summarised in Table 3.1

3.2 Data Preprocessing

This section presents the steps taken during data preprocessing. The dataset will be subjected to data cleaning which include identification of anomalies and missing values. SOG threshold is applied to ensure that the model represent operating condition at steady state. Domain knowledge based, feature selection is performed to ensure the model does not violate vessel domain knowledge. The datasets then will be split to training, validation and test dataset.

3.2.1 Data Cleaning

The plot of the journey indicates that the journey between Rønne and Sassnitz is not represented completely. This might be caused due to the limitation of T-AIS system which is caused by the due to poor coverage in the area between Sassnitz and

Feature	**Feature Name**
AIS data	
Position Time Stamp [DD/MM/YYYY HH:MM:SS]	`Time`
Latitude [°]	`LAT`
Longitude [°]	`LON`
Width [m]	`width`
Length [m]	`length`
SOG [Knots]	`sog`
COG [m/s]	`cog`
Heading [°]	`heading`
Draught [m]	`draught`
Weather Data (0.5° Granularity)	
Wind Speed [m/s]	`windspeed`
True North Wind Direction, φ [°]	`truenorthcurrentdir`
Air Temperature Above Oceans [K]	`oceantemperature`
Maximum Wave Height [m]	`waveheight`
Swell Period [s]	`swellperiod`
Wind Wave Period [s]	`windwaveperiod`
Wave Period [s]	`waveperiod`
Sea Surface Temperature [K]	`surftemp`
Combined Wind Wave Swell Height [m]	`windwaveswellheight`
Swell Height [m]	`swellheight`
Wind Wave Height [m]	`windwaveheight`
Current Speed [m/s]	`curspeed`
True North Current Direction γ [°]	`truenorthcurrentdir`
True Wind Direction [°]	`truewinddir`
True Current Direction [°]	`truecurrentdir`
True Swell Direction [°]	`trueswelldir`
True Wind Wave Direction [°]	`truewindwavedir`
True Wave Direction [°]	`truewavedir`

Table 3.1: Structure of fused dataset

Rønne. This is shown by the plot shown in Figure 3.7. Therefore, the data plot for the journey between Sassnitz and Rønne will be excluded. Basic threshold of decimal degrees of 55.04° N for latitude is applied, this threshold will exclude the journey between Sassnitz and Rønne.

In its initial state, the dataset contains 7453 data points which described the journey of the ship in one year. The initial data points represented all navigational status of the ship, which include “mooring”, “anchoring” and “underway using engine”. This is clearly observed in the histogram for the SOG Figure 3.8.

To ensure that the dataset represents the actual operating condition of ship in steady state, a threshold for SOG must be applied. SOG can vary due to changing sea state, but it can also be reduced by the ship’s operator around the port when it departs from port of origin or arriving at port of arrival. Therefore, any data points with SOG less than 5 knots will be discarded which is considered as manoeuvring (Abebe et al., 2020; Yan et al., 2020). Post filtering, the amount of data points decrease significantly from 7453 data points to 3828 data points. This indicated that about half of the total data points represented the ship’s stationary behaviour.

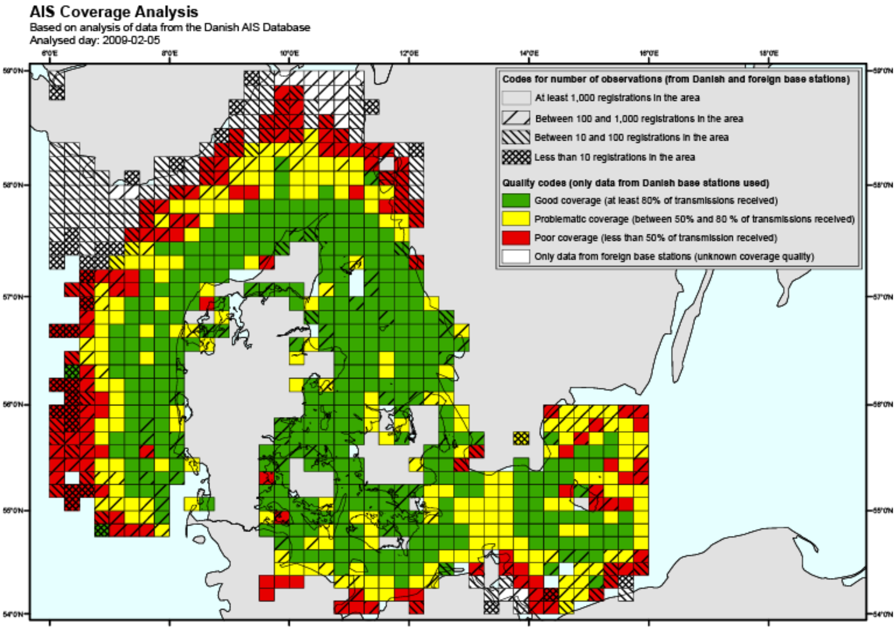

AIS Coverage Analysis
Based on analysis of data from the Danish AIS Database
Analysed day: 2009-02-05
6°0'E
8°0'E
10°0'E
12°0'E
14°0'E
16°0'E
18°0'E
59°0'N
58°0'N
57°0'N
56°0'N
55°0'N
54°0'N
Codes for number of observations (from Danish and foreign base stations)
At least 1,000 registrations in the area
Between 100 and 1,000 registrations in the area
Between 10 and 100 registrations in the area
Less than 10 registrations in the area
Quality codes (only data from Danish base stations used)
Good coverage (at least 80% of transmissions received)
Problematic coverage (between 50% and 80 % of transmissions received)
Poor coverage (less than 50% of transmission received)
Only data from foreign base stations (unknown coverage quality)

Figure 3.7: Shore based AIS Coverage based on data from AIS database Danish Maritime Authority (2023)

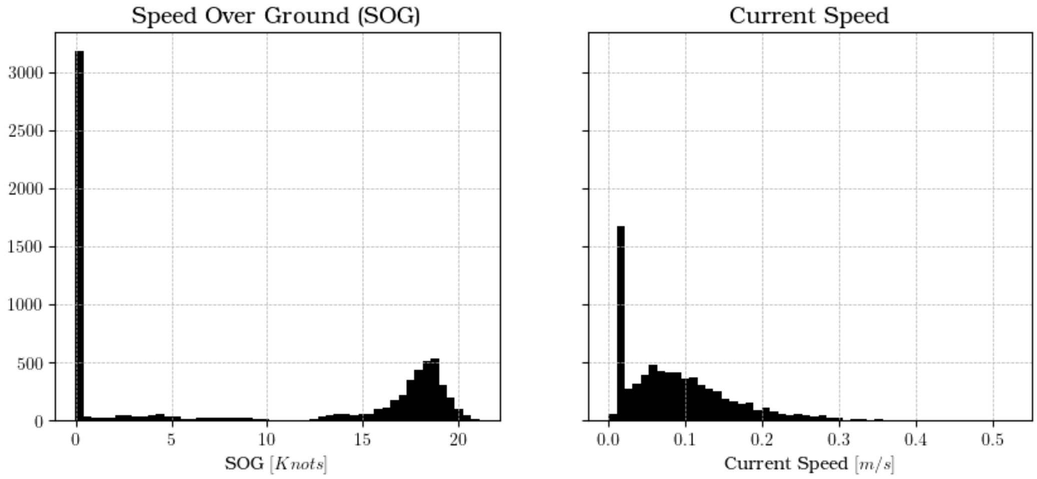


Figure 3.8: Histogram plot of pre-filtered SOG and current speed

From preliminary analysis, possible source of error is identified for data points representing current speed. In range of current speed between 0.01 and 0.03 m/s, noticeable peak in data points is observed as shown in Figure 3.8. This peak attributed to missing information on northward and eastward current speed in some data points from the provided dataset. This resulted in single random error value for current speed which resulted in the peak observed in the histogram.

To address the missing values, the missing values for eastward current and northward current are imputed using `KNNImputer` feature from `Scikit-Learn`. This is necessary as modelling package by `Scikit-Learn` cannot handle missing values. During

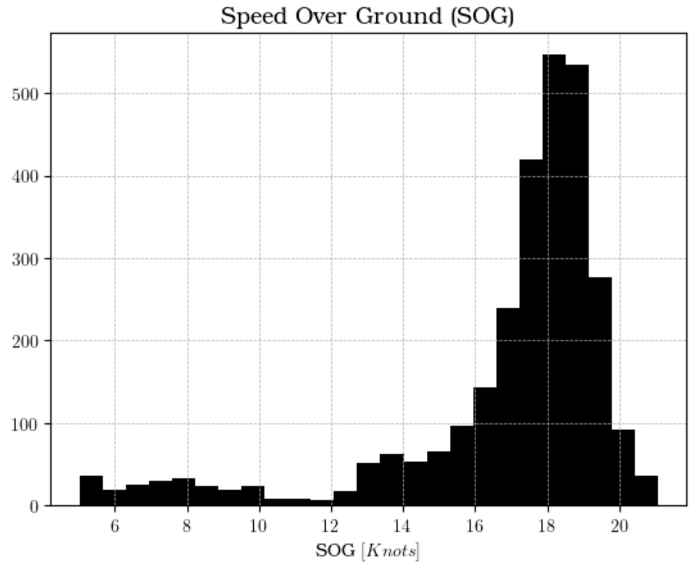


Figure 3.9: Histogram plot of SOG after threshold

imputing, each sample's missing values are imputed using the mean of nearest neighbour found in training dataset (Pedregosa et al., 2011). Imputing strategy using k-nearest neighbour is considered as it should reflect the weather conditions within the region of missing values. Once the missing values of northward and southward current are imputed, the current speed for the missing values will be recalculated. The imputing approach using k-nearest neighbour is also applied to other weather features that contained missing values i.e. `NaN` values.

3.2.2 Feature Selection

To select appropriate features for the model, correlation between the features is first studied. Feature selection is necessary to simplify the model and subsequently save computing cost during training. Selection of features is based on statistical approach of High Correlation Filter proposed by Abebe et al. (Abebe et al., 2020). This approach considers pairs of features with correlation features higher than 0.7 as one entity. However, the selection of highly correlated features must not violate physical knowledge. Therefore, the feature selection in this study is based on physical justification and this takes priority over purely statistical reasoning.

From AIS data, the information on *time, latitude, longitude, width and length* are not included for training. Time, latitude and longitude only describe the location of the ship at a particular position and the width and length of ship are constant dimensions. As discussed in Section 2.4.1, the features *combined wind wave swell height, swell height, maximum wave height and wind wave height* are physically correlated. The combined wind wave swell height defines the significant wave height $H_{1/3}$ and can be described using Equation (2.4.4), Equation (2.4.6) shows that the significant wave height also can be used to identify weather the sea is swell or wind sea dominated.

With that, it is clear that significant wave height should be retained for modelling, as many wave properties can be derived from it. The features swell height, wind wave height and maximum wave height will be dropped as it can be defined through significant wave height $H_{1/3}$. This decision is also statistically supported through the high correlation filter method. As shown in Figure 3.10, high correlation are obtained between the $H_{1/3}$, swell height, wind wave height and maximum wave height.

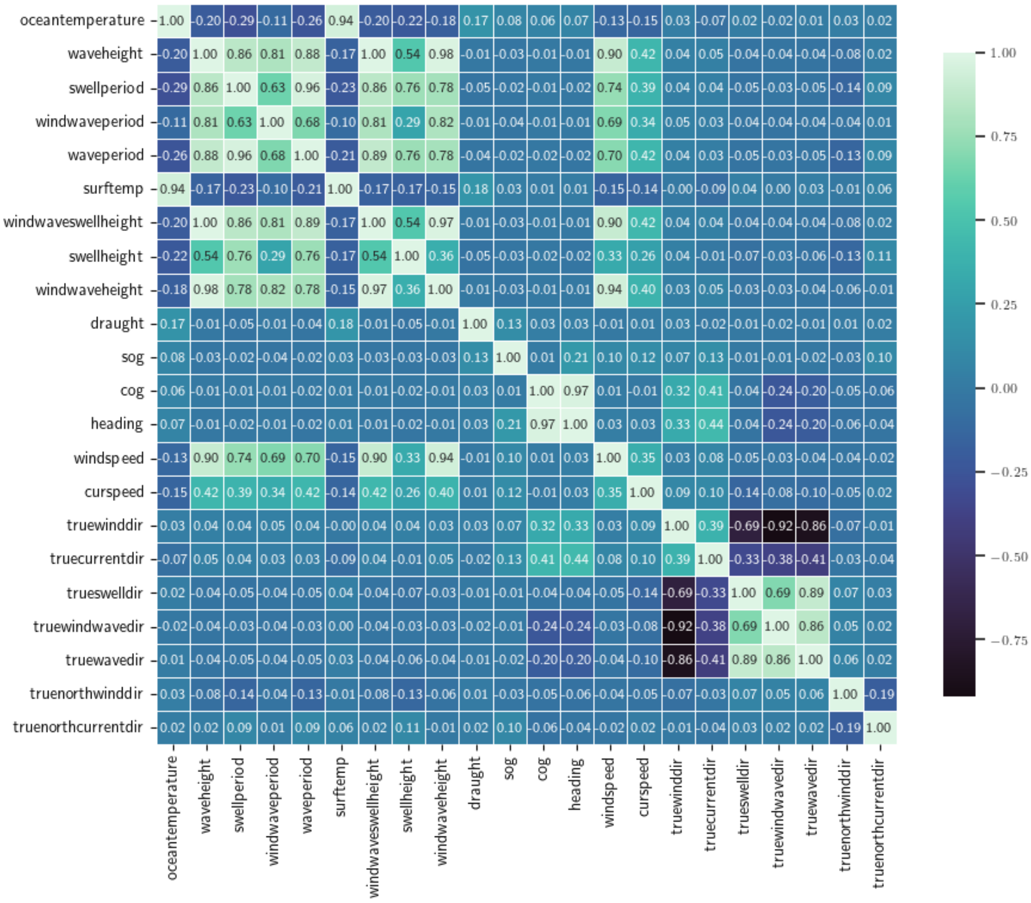


Figure 3.10: Correlation Heat Map

From Figure 3.10, high correlation is observed between wave period, swell period and wind wave period. As discussed in Section 2.4.1, the sea state can be described through the significant height $H_{1/3}$ and spectral peak T_p with help of Torsethaugen peak (**Torsethaugen et al., 2004**). Hence, the features swell period and wind wave period are discarded as it only distinguish whether the sea is dominated by swell or by wind. The feature wave period will still be retained. As a result, the features “true wind wave direction” and “true swell direction” will be excluded from consideration since the features that account for their magnitude have been discarded.

Statistically, the heading and COG are highly correlated, but both features are retained as it explain two different parameters of the ship. Course Over Ground reflects the ship course heading while heading represented the actual heading of the

Training Label	
SOG [Knots]	sog
Training Features	
COG [°]	cog
Heading [°]	heading
Draught [m]	draught
Wind Speed [m/s]	windspeed
Air Temperature Above Oceans [K]	oceantemperature
Wave Period [s]	waveperiod
Sea Surface Temperature [K]	surftemp
Combined Wind Wave Swell Height [m]	windwaveswellheight
Current Speed [m/s]	curspeed
True Wind Direction [°]	truewinddir
True Current Direction [°]	truecurrentdir
True Wave Direction [°]	truewavedir

Table 3.2: Structure of training dataset

Features	Count	Mean	Std.	Min	25%	50%	75%	Max
sog	2871	16.91	3.18	5.03	16.56	17.94	18.72	21.07
cog	2871	196.47	85.93	69.77	102.58	188.01	282.26	355.07
heading	2871	187.88	88.47	67.90	100.86	124.65	279.19	355.07
draught	2871	5.22	0.18	4.74	5.11	5.28	5.38	5.67
windspeed	2871	6.42	2.97	0.25	4.13	6.15	8.36	16.01
oceantemperature[3]	2871	282.71	6.49	264.08	277.13	282.64	288.82	296.83
waveperiod	2871	3.66	0.82	1.86	3.07	3.57	4.14	7.05
surftemp[4]	2871	283.40	5.73	273.05	278.13	282.83	288.86	294.75
windwaveswellheight[5]	2871	0.75	0.51	0.07	0.38	0.64	0.95	3.70
curspeed	2871	0.10	0.07	0.00	0.05	0.08	0.13	0.53
truewinddir	2871	87.14	55.96	0.00	34.19	84.79	140.58	179.77
truecurrentdir	2871	89.15	57.53	0.25	31.01	86.78	143.32	179.99
truewavedir	2871	91.74	55.53	0.13	39.12	92.28	143.33	179.92

Table 3.3: Descriptive statistics of preprocessed dataset

ship at a particular point of time. Same principle also apply between air temperature above ocean and sea surface temperature. Air temperature above oceans represents the temperature of wind while sea surface temperature represents current temperature of current. From feature selection, 5 features from AIS data are discarded while 11 features are removed from the weather data. To predict the ship speed, The SOG will be selected as the label to train the model. The remaining attributes will be selected as training features. This is summarised in Table 3.2.

3.3 Black Box Modelling

In this section, the modelling of ship speed through SOG using selected features will be performed using tree-based regressor model. The tree-based regressor model considered are decision tree regressor (DTR), random forest regressor (RFR) and

[3] Air temperature above oceans

[4] Sea Surface Temperature

[5] Significant wave height

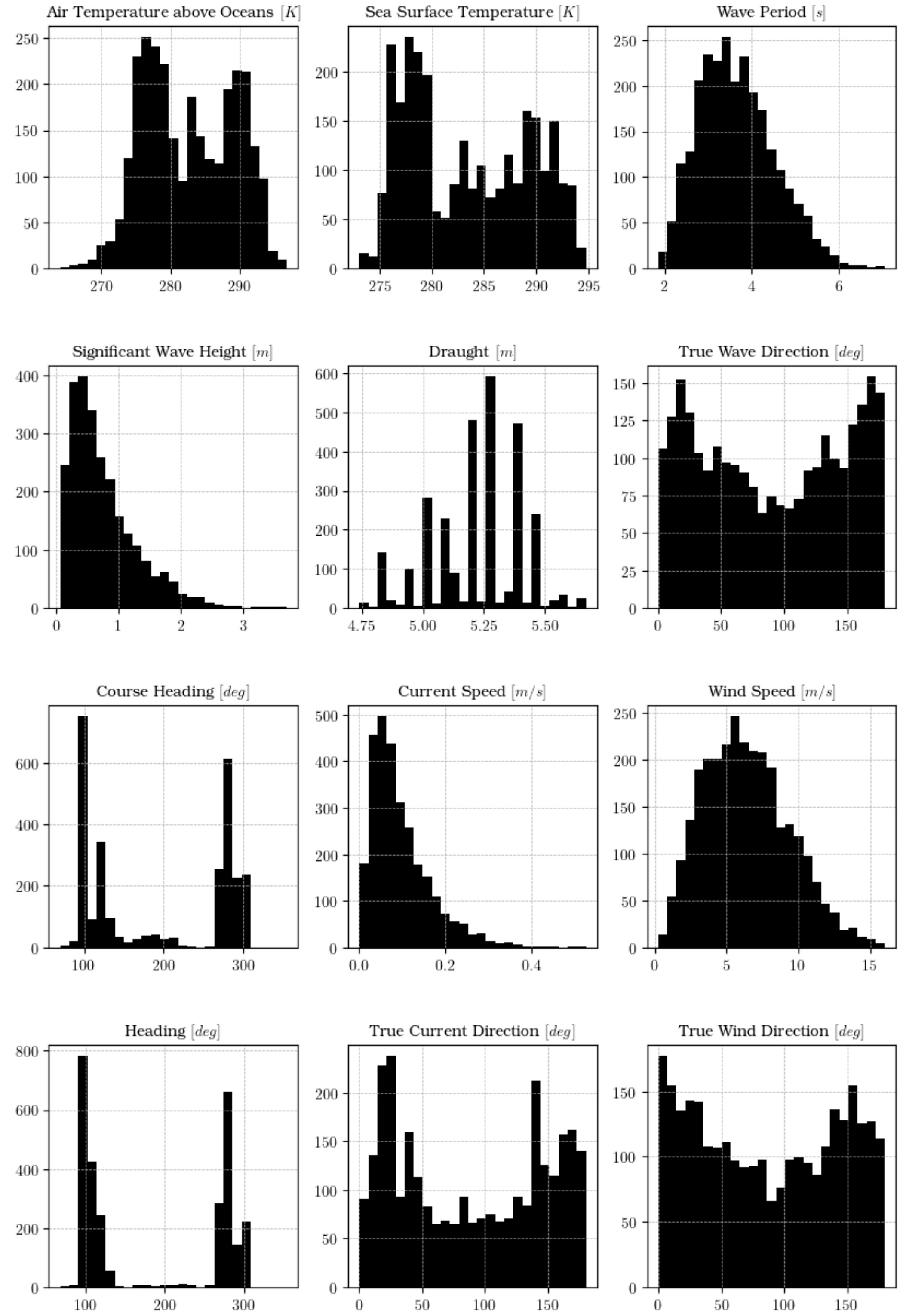


Figure 3.11: Histogram of training features

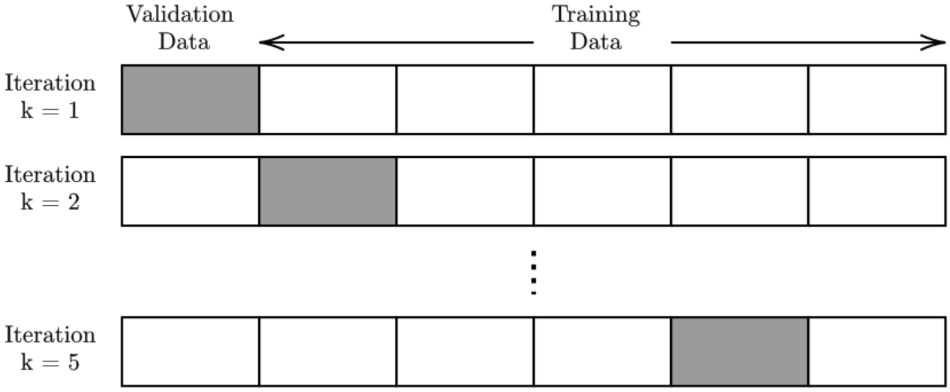


Figure 3.12: Visual illustration of k-folding, Grey shaded box represents the validation data while white box represents the training data

extra-tree regressor (ETR). The tree-based models are compared against multiple linear regressor (MLR) for benchmarking. For training, the dataset is split into training and test dataset in ratio of 75:25. This corresponds to 2871 data points for training and 957 data points for testing. The training dataset will be subjected to 10-fold splitting, this k-fold splitting is performed as the model will be evaluated using k-fold cross validation. The hyperparameter of the tree-based regressor will be iteratively tuned until no further improvements of the model can be made.

3.3.1 Performance Metrics for Validation

To gain sensible estimate of model performance and how precise a model is, the model will be cross validated by means of k-folding. K-fold cross validation split the training set into k subsets which is called *folds*, then the model will be trained k times using k-1 subsets and remaining one for validation, this process is illustrated in Figure 3.12. For each iteration, each model is evaluated using different performance metrics such as **Coefficient of Determination (R^2), Explained Variance (EV), Mean Absolute Error (MAE),Root Mean Square (RMSE), Median Absolute Deviation (MAD) and Mean Absolute Percentage Error (MAPE)**. The results from each iteration is then averaged, where the information on model precision can be gained from the standard deviation. Performing k-fold cross validation checks model robustness against different datasets. The properties of each performance metric will be discussed in the following sections.

Coefficient of Determination (R^2)

The coefficient of determination R^2 gives a measure on prediction quality, R^2 quantifies the ability of the regression model to approximate the actual values. R^2 is defined by Equation (3.3.1), where y represents true target output, $\hat{y}$ represents the predictor output and $\overline{y}$ represents the mean. R^2 score range between 0 and 1, higher values i.e. $R^2 \rightarrow 1$ indicate better model fit and score of 1 indicate perfect prediction.

$$R^2(y,\hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_\mathrm{i} - \hat{y}_\mathrm{i})^2}{\sum_{i=1}^{n}(y_\mathrm{i} - \overline{y}_\mathrm{i})^2} \quad \textbf{where} \quad \overline{y} = \frac{1}{n}\sum_{1}^{n} y_\mathrm{i} \tag{3.3.1}$$

Explained Variance (EV)

Explained variance indicate how well a model can capture variance from a dataset. It is defined by Equation (3.3.2), where σ_x represents standard deviation of parameter x. EV score range between 0 and 1, where the best score of $EV = 1$ can be obtained if $\sigma^2_{(y-\hat{y})} \rightarrow 0$.

$$EV(y,\hat{y}) = 1 - \frac{\sigma^2_{(y-\hat{y})}}{\sigma^2_y} \tag{3.3.2}$$

Mean Absolute Error (MAE)

MAE indicated the expected value of absolute (L^1 norm) error, and it can be calculated by:

$$MAE(y,\hat{y}) = \frac{1}{n}\sum_{i=1}^{n}|y_\mathrm{i} - \hat{y}_\mathrm{i}| \tag{3.3.3}$$

Root Mean Square Error (RMSE)

The RMSE describe the expected value of quadratic error. RMSE place large penalty on large deviation between true and estimated values and for this reason, it can be used to as a metric to indicate model performance against outliers. Ideal score is observed when RMSE → 0. RMSE can be considered as absolute measure of model fitness. Omitting the root term, RMSE becomes MSE, which is the loss function of Equation (2.2.2) that is used to determine the most optimal split in a regression decision tree.

$$RMSE(y,\hat{y}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_\mathrm{i} - \hat{y}_\mathrm{i})^2} \tag{3.3.4}$$

Median Absolute Deviation (MAD)

MAD is a performance metrics that considers the median of the absolute errors. It is robust to outlier as it only consider median performance

$$MAD(y,\hat{y}) = \mathrm{median}(|y_1 - \hat{y}_1|, \ldots, |y_\mathrm{i} - \hat{y}_\mathrm{i}|) \tag{3.3.5}$$

Mean Absolute Percentage Error (MAPE)

Is an alternative to MAE, which provide easier interpretation, the result of MAPE can be interpreted according to Equation (3.3.6) (Montaño Moreno et al., 2013). The usage of MAPE in model evaluation is to get initial estimate, as MAPE comes with some drawback such as instability when $y_i = 0$ and it may lead to biased forecast (Gkerekos et al., 2019). As such, the evaluation of the model performance will be mainly based on MAE and RMSE.

$$MAPE(y,\hat{y}) = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_\mathrm{i}-\hat{y}_\mathrm{i}}{y_i}\right|\cdot 100\% \quad \textbf{with} \quad \begin{array}{lc} \text{MAPE} & \text{Interpretation} \\ \hline < 10 & \text{Highly accurate forecasting} \\ 10-20 & \text{Good Forecasting} \\ 20-50 & \text{Reasonable forecasting} \\ > 50 & \text{Inaccurate forecasting} \end{array} \tag{3.3.6}$$

3.3.2 Model Hyperparameter Optimisation

The subject of parameter tuning was briefly discussed in Section 2.2.1. In Section 2.2.1 parameter tuning was applied to decision tree regressor to avoid overfitting by changing the minimum amount of samples a leaf node has. This example implies that altering model hyperparameter will affect the model performance. However, the optimisation of the hyperparameter cannot be performed *a priori* and as such iterative process will be performed until best hyperparameter value is found.

`Scikit-Learn` offers `GridSearchCV` and `RandomizedSearchCV` to help search for the most optimal hyperparameter. Both solutions operate with similar principle: The selected hyperparameters to be tuned with its value range is evaluated using cross validation to evaluate the best possible combination between the selected hyperparameters. The difference between `GridSearchCV` and `RandomizedSearchCV` lies in how it searches for the best value for the selected hyperparameters: `GridSearchCV` involves construction of grids containing all possible combinations of hyperparameter value in specified range.`RandomizedSearchCV` randomly samples hyperparameter values.

The exhaustive nature of `GridSearchCV` means that it is computationally costly to perform, especially when there are multiple hyperparameters to be considered and value search space is large. `RandomizedSearchCV` gives more control to computing budget by setting the number of iteration and usually produces more accurate results than `GridSearchCV` approach. (Géron, 2019; Bergstra and Bengio, 2012).

For this reason, the `RandomizedSearchCV` will be employed to search for best possible hyperparameter. However, the limitation of *a priori* knowledge of hyperparameter value still exists. In spite of `RandomizedSearchCV` ability to control the computational budget, it is still takes considerable time to obtain the best hyperparameter value. The computational budget may be spent on searches in unpromising search space. With that, initial exploration on the effect of each hyperparameter on model performance will be performed to give better overview on which search space that should be considered during hyperparameter optimisation. In the next subsections, the effect of tunable hyperparameter of tree-based model from `Scikit-Learn` will be explored to give baseline numbers for the search space. RMSE is used as performance metrics as the hyperparameter parameter optimisation done in this thesis aims to reduce the error during prediction.

Number of features

Defined with default value as `max_features=None` in `Scikit-Learn`. This hyperparameter controls the number of features to be considered when looking for the best split, the default `None` option means it will consider all features. This parameter tuning is available for Decision Tree Regressor, Random Forest Regressor and Extra-Tree Regressor. Initial exploration indicated Random Forest Regressor and Extra Tree Regressor benefit from considering more features. Decision Tree Regressor also benefits from considering all features for determining the split.

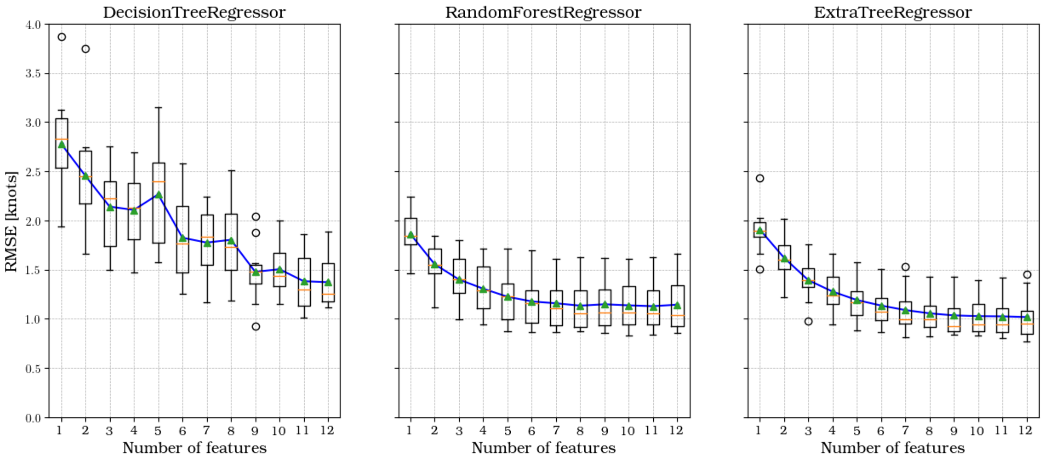


Figure 3.13: Hyperparameter tuning of `max_features`

Minimum samples to split a node

Defined with default value as `min_samples_split=2` in `Scikit-Learn`. This hyperparameter controls the minimum number of samples i.e. data points required to split a node. The default value of `2` is the least number of sample required to split a node i.e. 1 sample is split to the left and right branch respectively. The plot at Figure 3.15 indicates that tuning this hyperparameter will not have any major impact on any model's performance.

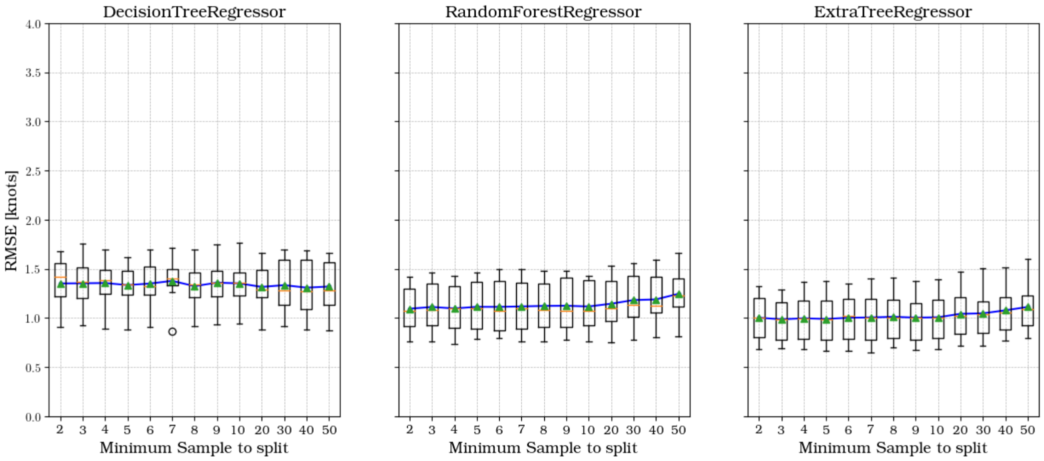


Figure 3.14: Hyperparameter tuning of `min_samples_split`

Number of sample in a leaf node

Defined with default value as `min_samples_leaf=1` in `Scikit-Learn`. This parameter controls number of samples required to be at leaf node, where split point will be considered if the leaf contains at least `min_samples_leaf=n` training samples in each left and right branch. As shown in Figure 2.4, tuning this hyperparameter to higher values helps to smoothen the model and avoid overfitting. However, this may lead to underfitting as the model is unable to capture the trend within the data. This is supported by the findings shown in Figure 3.15, the DTR benefits from regularisation at certain breakeven point. But, after this breakeven point, the model's performance degrades. It is also observed that tuning this parameter will have a negative impact on RFR and ETR model's performance-

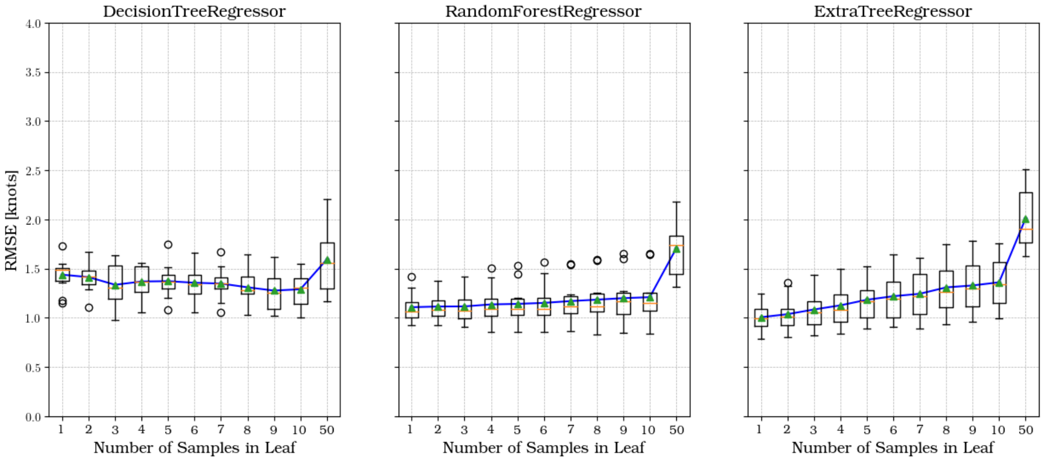


Figure 3.15: Hyperparameter tuning of `min_samples_leaf`

Depth of Tree

Defined with default value as `max_depth=None` in `Scikit-Learn`. This hyperparameter controls the growth of the tree. Leaving it at `max_depth=None` means the tree will grow until all leaves are pure i.e. until minimum MSE is obtained or when the number of samples is less than the minimum number of samples required to split an internal node. Similar to `min_samples_leaf`, DTR shows improvement until a certain breakeven point. RFR performance seems to stabilise at certain depth while ETR benefits from allowing full growth of the tree. It can also be observed that the model's performance are identical for `max_depth=1`, which is evident as shown in Figure 3.16

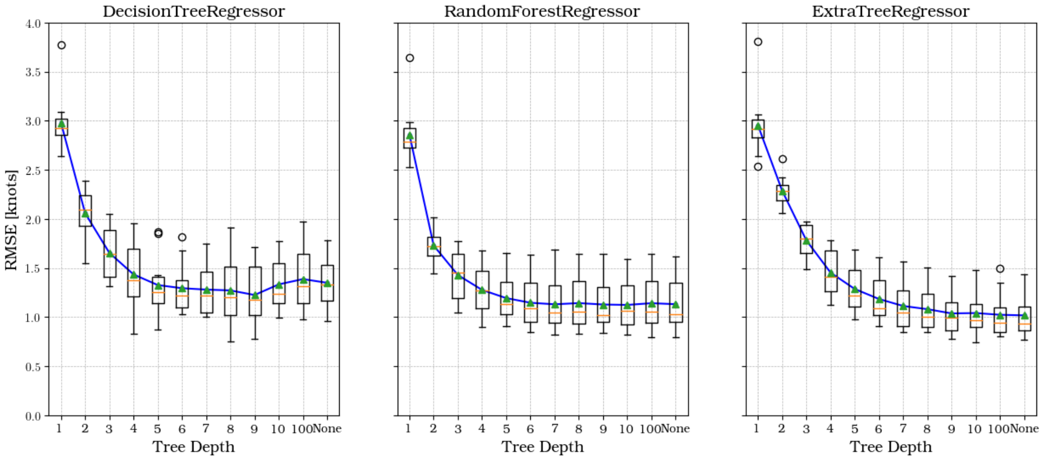


Figure 3.16: Hyperparameter tuning of `max_depth`

Number of Trees

Defined with default value as `n_estimators=100`. This hyperparameter controls the amount of trees i.e. predictors in a forest. Tuning of number of trees will have an effect on the training time, and it is only available to RFR and ETR. The default value seems to yield satisfactory result, as the performance for both RFR and ETR stabilise after in this case stabilise after 100 trees, as seen in Figure 3.13.

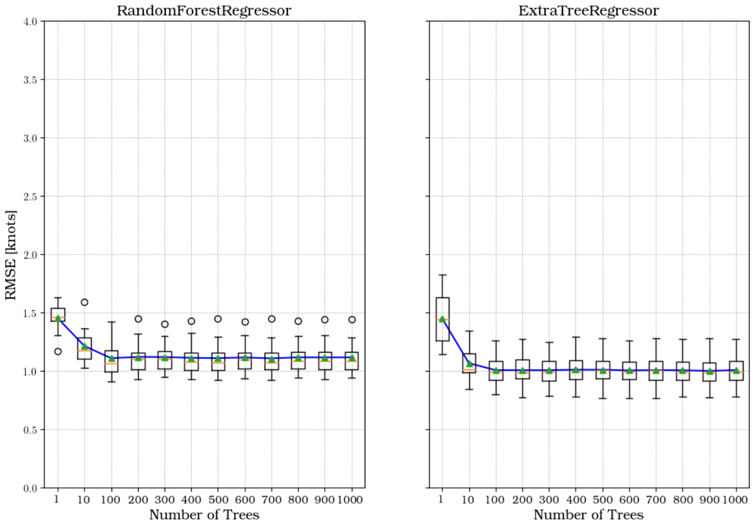


Figure 3.17: Hyperparameter tuning of `n_estimators`

3.4 White Box Modelling

In this section, the predicted SOG from the BBM will be fed into the WBM to predict the fuel consumption (FOC). Using Holtrop-Mennen approximation method, the resistance encountered by the ship will be estimated to find the total resistance which will be used to calculate the power required to propel the ship. However, during resistance calculation, some form coefficients and ship parameters are not readily available and may need to be assumed based on other literature studies, this assumption will be indicated throughout this section. The resulting brake power P_B is plotted against the STW v_S to obtain a power speed curve model which can be used to predict FOC.

3.4.1 Calculation of Total Resistance

The formula used to calculate total resistance R_{TOTAL} in this section are presented in Section 2.5.2 with principle dimension of the ship which was given in Figure 3.4. Some assumptions are made for the sea state and the values of some form coefficients are calculated based on empirical formulas presented in Section 2.5.1. In this case study, Holtrop-Mennen method can be used as it fulfills the condition set in Equation (2.5.10):

$$\begin{gathered} Fr = 0.2417 \leqslant 0.45 \\ 0.55 \leqslant C_P = 0.6707 \leqslant 0.85 \\ 3.9 \leqslant \frac{L}{B} = 5.91 \leqslant 9.5 \end{gathered} \tag{3.4.1}$$

The calculations of resistance are dynamics as the Froude number Fr is based on v_S, the design Froude number Fr_{DESIGN} is used to check use cases for some equations.

Calculation of frictional resistance R_F

For the calculation of surface area of bare hull S, it is assumed that the aft has a U-shaped section. Then the appropriate C_{stern} can be calculated to obtain the constant c_{14} Equation (2.5.16). For the calculation of length of run L_R, approximation for ℓ_{CB} are made based on Equation (2.5.18). The constants c_{14} and L_R are used to calculate the form factor $1 + k_1$ which will be used to calculate R_F.

Calculation of appendage resistance R_{APP}

From known ship information and ship schematics shown in Figure 3.6, it can be deducted that the ship consists of the following appendages:

- Two high-lift flap rudders
- Single centre skeg
- Twin shafts supported by two brackets

Appendage type	Value	$(1+k_{2_i})$
Two high-lift flap rudders		3
h_{RUDDER}	$4.06\ m$	
B_{RUDDER}	$1.99\ m$	
S_{RUDDER}	$16.16\ m^2$	
Single centre skeg		1.5
h_{SKEG}	$4.41\ m$	
B_{SKEG}	$26.23\ m$	
S_{SKEG}	$115.67\ m^2$	
Twin shafts supported by two brackets		3
D_{SHAFT}	$0.55\ m$	
L_{SHAFT}	$13.54\ m$	
S_{SHAFT}	$46.79\ m^2$	
$S_{APP_{tot}}$	$\mathbf{178.62}\ m^2$	
$(1+k_{2_i})_{eq}$		**2.03**

Table 3.4: Assumed appendage values

The assumptions for the appendage area are made by scaling the schematics to known measurements (e.g. the L_{WL}). From here, the appropriate k_{2_i} constants for individual appendages will be selected to obtain the $(1+k_{2_i})_{eq}$ from Equation (2.5.20). Additionally, there are two bow thrusters installed with approximated diameter of $d_{TH} = 2.15m$, from here, the constant C_{D_TH} can be approximated using Equation (2.5.21). Hence, the appendage resistance R_{APP} can be calculated using Equation (2.5.22).

Calculation of wave resistance R_W

The calculation of wave resistance is based on the case for $Fr \leq 0.4$ using equation Equation (2.5.24). The estimation is done by adding the constants presented between Equation (2.5.27) and Equation (2.5.38). There are some use cases for some equations, which is summarised in Table 3.5.

Constant	Use Case	Equation
c_7	$0.11 < \frac{B}{L_{WL}} \leq 0.25$	Equation (2.5.27)
c_{15}	$\frac{L_{WL}^2}{V} \leq 512$	Equation (2.5.33)
c_{16}	$C_P \leq 0.8$	Equation (2.5.34)
λ	$L_{WL} \leq 12$	Equation (2.5.36)

Table 3.5: Use case of constants for R_W

Calculation of bulbous bow resistance R_B

The area A_{BT} used to calculate is approximated based on **Kracht** (**1978**) with:

$$A_{BT} = 0.085 A_M \tag{3.4.2}$$

For the height h_B, the upper limit of $h_B = 0.6 T_{DESIGN}$ is selected, and it is assumed that $T_F = T_{DESIGN}$.

Calculation of (immersed) transom resistance R_{TR}

Since the immersed transom area is unknown, Rakke (2016) approximated the immersed transom area based on correlation of ship dimension from literature review of Holtrop and Mennen (1982):

$$A_{TR} = 0.051 A_M \tag{3.4.3}$$

This approximation must be used with caution as it is only based on case study of Holtrop and Mennen (1982). However, to author's best knowledge, there are no other literature that provide empirical estimation of A_{TR}. Therefore, this estimation will be selected in this case study. The selection for the value of constant c_6 is dependent on the value of Fr_T, which is a function of v_S. From there, the transom resistance can be calculated using Equation (2.5.44)

Calculation of correlation allowance resistance R_A

The selection of constant c_4 in equation Equation (2.5.45) is based on the T_F, then the correlation resistance R_A can be calculated using Equation (2.5.47).

Calculation of added resistance due to wind R_{AA}

Two assumptions are made during the calculation of R_{AA}, since the information of lateral area A_L and A_F are not readily available, these values are assumed based on the dimension of similar ferry in the case study of Blendermann (1994). It is assumed that the ferry has an A_L of **2125.80** m^2 and A_F of **325.30** m^2. From Table 2.4, the case for ferry ship is taken to get the necessary constants for the calculation of R_{AA}.

Calculation of added resistance due to wave R_{AW}

This part of the equation is relatively straightforward, L_{BWL} will be approximated to about **43.75** m. The calculation of R_{AWL} will be based on the data of significant wave height $H_{1/3}$ from the dataset.

3.4.2 Calculation of total efficiency η_{TOT}

Calculation of open water efficiency η_O

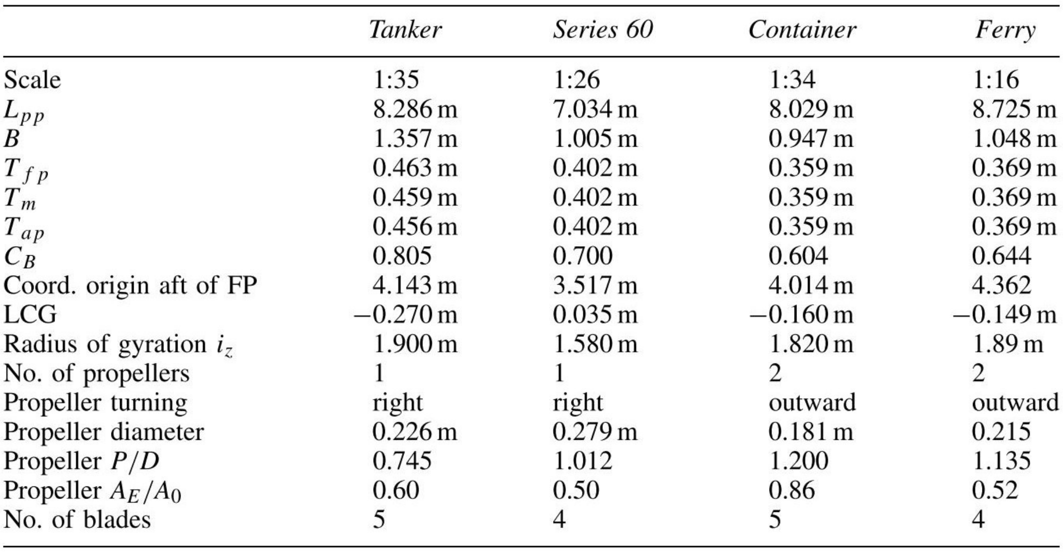
This value is approximated based on the line of Wageningen series in Figure 2.15 (Breslin and Andersen, 1994). The case will be for "Passenger ships and high speed naval vessels". Since the value of C_Th is not available, the value of η_O is approximated as **0.7**.

Calculation of hull efficiency η_H

For the calculation of η_H, the value of the propeller diameter D is approximated as **4 m**, which is based on the schematics of the ship shown in Figure 3.6.

Calculation of relative rotative efficiency η_R

The missing value required to compute Equation (2.5.59) is the pitch-diameter propeller ratio. This value will be estimated as $P/D = \mathbf{1.135}$, which is obtained from the work of **Bertram (2000)**.



	Tanker	*Series 60*	*Container*	*Ferry*
Scale	1:35	1:26	1:34	1:16
L_{pp}	8.286 m	7.034 m	8.029 m	8.725 m
B	1.357 m	1.005 m	0.947 m	1.048 m
T_{fp}	0.463 m	0.402 m	0.359 m	0.369 m
T_m	0.459 m	0.402 m	0.359 m	0.369 m
T_{ap}	0.456 m	0.402 m	0.359 m	0.369 m
C_B	0.805	0.700	0.604	0.644
Coord. origin aft of FP	4.143 m	3.517 m	4.014 m	4.362
LCG	−0.270 m	0.035 m	−0.160 m	−0.149 m
Radius of gyration i_z	1.900 m	1.580 m	1.820 m	1.89 m
No. of propellers	1	1	2	2
Propeller turning	right	right	outward	outward
Propeller diameter	0.226 m	0.279 m	0.181 m	0.215
Propeller P/D	0.745	1.012	1.200	1.135
Propeller A_E/A_0	0.60	0.50	0.86	0.52
No. of blades	5	4	5	4

Figure 3.18: Estimated value of propeller dimensions **(Bertram, 2000)**

Calculation of shaft efficiency η_S

The value of shaft efficiency is estimated as $\eta_S = \mathbf{0.99}$ based on **MAN (2011)** and **Holtrop and Mennen (1982)**.

3.4.3 Calculation of FOC

Once R_{TOTAL} and η_{TOTAL} is obtained, then the brake power of the ship P_B can be calculated using Equation (2.5.3). The resulting P_B will be plotted against v_S to obtain a power-speed curve, then regression line will be constructed to fit the data points. Each of the BBM model will generate different prediction for SOG, the resulting equation of the regression line represents the characteristic of different BBM model. To compare the performance, the regression model from each BBM will be compared against the regression model generated from real data to assess the predictive performance of each model.

The FOC can be calculated using Equation (2.5.4). The information of the SFOC can be obtained from Figure 3.4. Without the multiplication with operation time, $\mathcal{T}_{operation}$, The fuel consumption of the fuel in T/h can be obtained by dividing the value by $1 \cdot 10^6$.

Parameter	Value	Remarks
g	9.805 kg/ms^2	
ρ_{sea}	1025 kg/m^3	
ν_{sea}	0.00000118 m^2/s	
ρ_{air}	1.25 kg/m^3	
1 m/s	1.9438 knots	
Required Parameters for Holtrop-Mennen		
L_{WL}	144.80 m	From Figure 3.4
B	24.50 m	From Figure 3.4
T	5.85 m	Assume $T_A = T_F = T$ for initial phase, otherwise use T from dataset, also assume maximum draught
V	13592.1413 m^3	$V = C_B \cdot L_{WL} \cdot T_{MAX}$
Fr_N	0.2417	From Equation (2.5.6)
C_B	0.6549	From Equation (2.5.5)
C_M	0.9764	From Equation (2.5.7)
C_P	0.6707	From Equation (2.5.8)
C_{WP}	0.7700	From Equation (2.5.9)
ℓ_{CB}	-0.0123	Equation (2.5.18)
A_{TR}	7.3581 m^2	Equation (3.4.3)
A_{BT}	12.2634 m^2	Equation (3.4.2)
h_B	3.5100 m	Assume upper limit $h_B = 0.6T_F$
D	4 m	Approximated from schematics Figure 3.6
A_E/A_0	1.135	Value assumed from Figure 3.18
C_{stern}	10	Assume u-shaped section Equation (2.5.16)
Optional Parameters for Holtrop-Mennen		
S	3881.0231 m^2	approximated from Equation (2.5.14)
S_{APP}	178.62m^2	approximated from schematics Figure 3.6
i_E	21.6014°	Equation (2.5.29)
d_{TH}	2.15 m	Approximated from schematics Figure 3.6

Table 3.6: Assumed value for some constants

Chapter 4

Result and Discussion

The performance of GBM will be evaluated by means of a case study using the test dataset. The test dataset comprises journey data from the whole year of 2021. The first part will focus on performance evaluation of BBM, where the trained model will be used to predict the SOG. The second part focus on the power estimation method using Holtrop-Mennen method. The output of BBM, which is the ship SOG, will be fed to the WBM to estimate the power. For further clarity regarding the methodology, the following steps are taken which are based on the proposed methodology shown in Figure 3.1 and Figure 3.2. For generation of the BBM, the steps taken are:

1. Dataset is loaded.
2. Identify and remove any anomalies.
3. Remove static and unneeded features.
4. Apply speed threshold of 5 knots.
5. Highly correlated features are combined/removed based on physical and statistical reasoning.
6. Impute missing values using `KNNImputer`.
7. Split the dataset into training and testing.
8. Train the model using the whole dataset with default hyperparameter.
9. Evaluate model performance using k-fold cross-validation.
10. Tune the model until the best model is obtained.
11. For the case study, the best models will be used to predict the SOG using the test dataset.

Subsequently, for FOC calculation, the following steps are taken:

1. The test dataset is split into seasonal data. Summer-Fall season and Winter-Spring season which correspond to data for 6 months respectively.
2. Impute missing values using `KNNImputer`.
3. SOG is converted to STW.
4. Calculate calm water resistance R_{CALM}.

5. Calculate added resistance due to wave R_{AW}.
6. Calculate added resistance due to wind R_{AA}.
7. Calculate total effective power P_E using total resistance R_{TOTAL}.
8. Calculate brake power P_B from total efficiencies.
9. Plot resulting regression line for Power-Speed curve from all models and actual case.
10. Calculate the FOC by considering the engine SFOC and operation time.
11. Plot resulting regression line for FOC-Speed curve from all models and actual case.
12. Evaluate the performance of the model generated from the regression lines.

4.1 Evaluation of BBM

Model Training and Selection of Optimal Parameter

As mentioned in Section 3.3. There are 2871 data points available for training. To help narrow the search range of the hyperparameters for the tree-based model, RMSE plots against different values of hyperparameters will be performed. This method was presented in Section 3.3.2. The hyperparameter will be iteratively tuned until the best model is obtained. The result of the optimal parameter is found in Table 4.1. The model training is executed using **AMD Ryzen 7 2700X, Eight-Core Processor @ 3.7 GHz processor with 16384 MB installed RAM**.

Model	Training time [s]	Optimal Hyperparameter
DTR	0.044	None

DTR$_{OPT}$	0.021	`min_samples_split = 7`
		`min_samples_leaf = 10`
		`max_features = 12`
		`max_depth = 8`

RFR	4.112	None

RFR$_{OPT}$	3.431	`min_samples_split = 2`
		`min_samples_leaf = 1`
		`max_features = 10`
		`max_depth = 120`
		`n_estimators = 100`

ETR	0.944	None

ETR$_{OPT}$	4.390	`min_samples_split = 9`
		`min_samples_leaf = 1`
		`max_features = 12`
		`max_depth = 120`
		`n_estimators = 800`

MLR	0.004	None

Table 4.1: Optimal hyperparameter with training time of each model

With the default hyperparameter, RFR takes the longest training time followed by ETR and DTR. This is expected as RFR uses greedy algorithm i.e. it looks for the best possible feature when splitting the node. ETR takes significantly shorter time to train as ETR randomly select for features when splitting the node. DTR takes the shortest training time as it only generates a single tree. However, in the case of optimised model, ETR takes a longer time to train compared to RFR. This is caused by the number of trees in the optimised model which is controlled by the parameter `n_estimator`, the optimised ETR model has 800 trees in comparison to 100 trees of RFR. It is also observed that the training time of optimised DTR model is halved as pruning the tree resulted in a simpler model to train.

To further investigate the effect of hyperparameter optimisation, the learning curve of each tree-based model is plotted. For DTR, generated model with default parameter will result in a model that heavily overfits the training data, which is evident from the large gap between the training error and validation error which indicated a high variance as shown in Figure 4.1. Regularisation i.e. parameter tuning of the DTR model helps balance between bias and variance by trading bias for variance. This is observed from the substantial reduction in the gap between the training and validation error from Figure 4.1. Additionally, the learning curve indicates that the most notable improvement in model performance occurs until around 1000 data points. Beyond this point, the enhancement in model performance becomes less substantial.

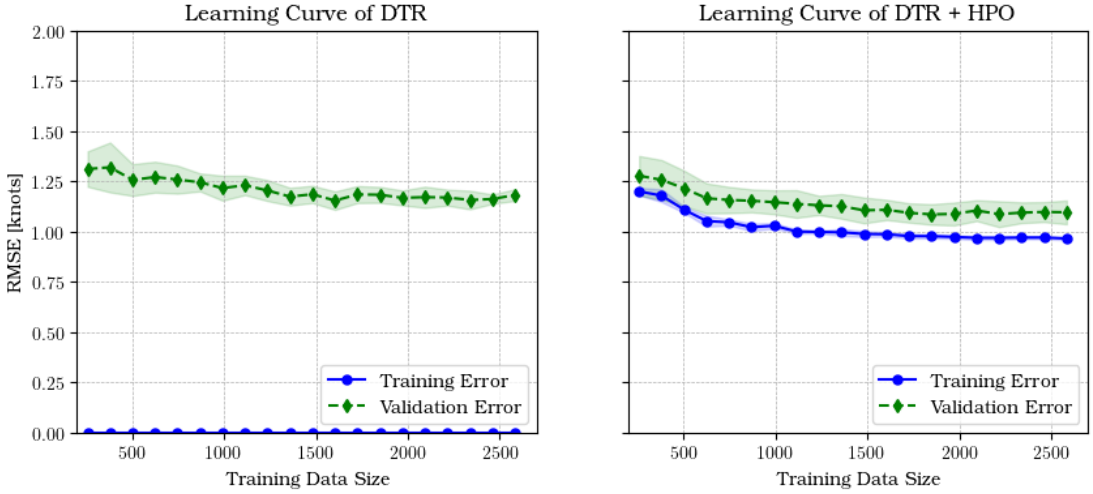


Figure 4.1: Learning curve of DTR

The process of hyperparameter tuning for the Random Forest Regressor (RFR) model did not show any significant improvement in model performance. This outcome aligns with the findings of Kuhn and Johnson (2013) and Hastie et al. (2009) which was discussed in Section 2.2.2. The most notable improvement on model performance is observed until around 750 points, after which the model appears to reach a plateau. Furthermore, there is noticeable variance in the RFR model, which indicates that the model will have a slight tendency to overfit.

Hyperparameter tuning helps to reduce variance in the ETR model. But it does not have any major impact on model's performance. The ETR model reaches plateau

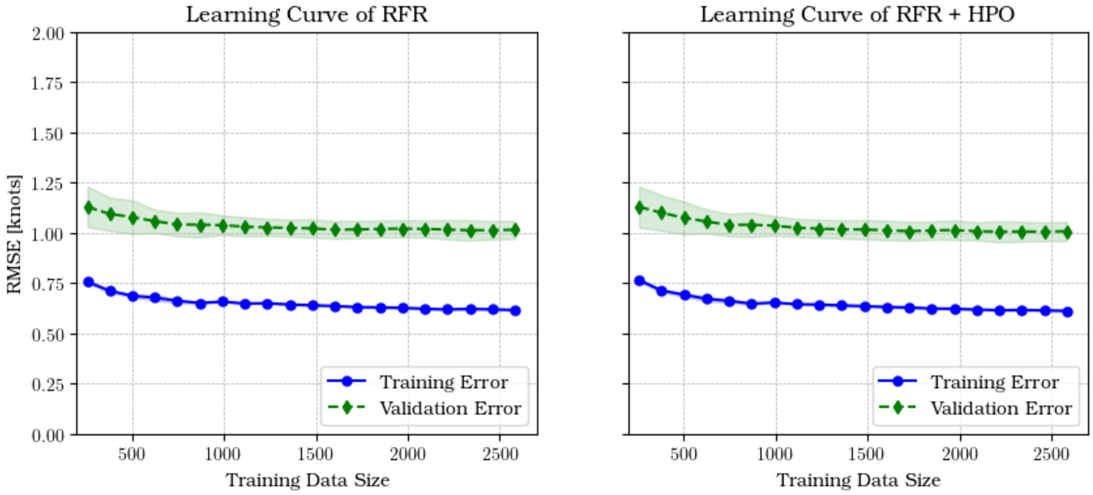


Figure 4.2: Learning curve of RFR

beyond 1000 data points. Suggesting that adding more data points will not result in any significant increase in model performance.

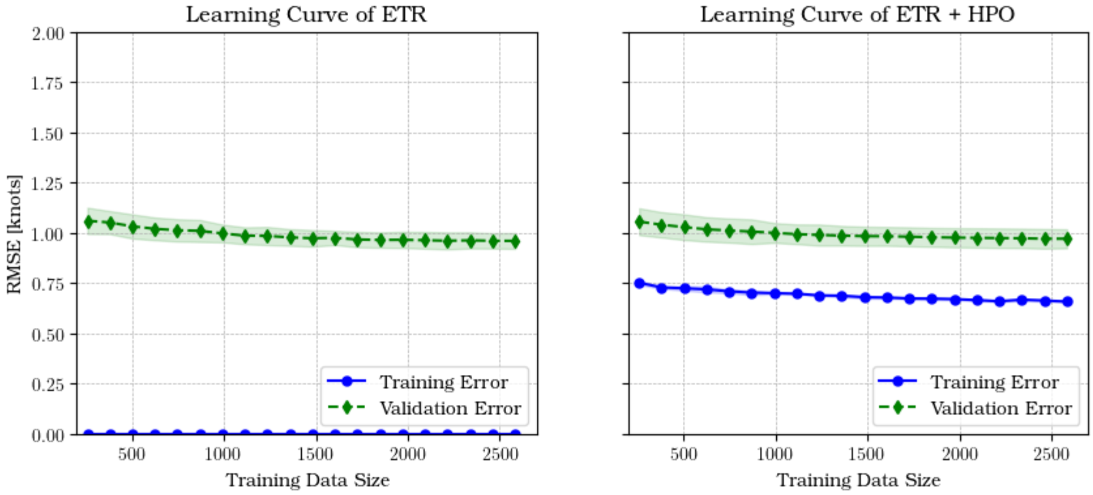


Figure 4.3: Learning curve of ETR

In addition to the initial exploration in Section 3.3.2, it can be concluded that hyperparameter tuning for number of features and tree depth will have the biggest impact in affecting the model's performance. To improve training time, lower number of trees should be considered for RFR and ETR model.

Analysis of trained model

Feature Importance

As discussed in Section 2.2.2, tree-based models are able to quantify the impact of each feature during the split process, this is performed using `feature_importances_` feature in `Scikit-Learn` (Kuhn and Johnson, 2013). According to documentation by Pedregosa et al. (2011), it is computed as the mean and the standard deviation of accumulation of the impurity decrease within each tree i.e. total reduction of the criterion brought by a feature. Alternatively, it can be defined as how much a feature is used in each tree.

DTR_{OPT}		RFR_{OPT}		ETR_{OPT}	
Feature	Importance	Feature	Importance	Feature	Importance
`heading`	0.6563	`heading`	0.4927	`cog`	0.6410
`cog`	0.3183	`cog`	0.4183	`heading`	0.2707
`draught`	0.0105	`draught`	0.0210	`truecurrentdir`	0.0200
`truewinddir`	0.0047	`curspeed`	0.0104	`draught`	0.0144
`oceantemperature`	0.0029	`waveperiod`	0.0093	`windwaveswellheight`	0.0112
`surftemp`	0.0025	`truecurrentdir`	0.0092	`curspeed`	0.0110
`waveperiod`	0.0019	`windwaveswellheight`	0.0084	`waveperiod`	0.0095
`truecurrentdir`	0.0010	`surftemp`	0.0075	`windspeed`	0.0053
`windwaveswellheight`	0.0008	`truewinddir`	0.0075	`surftemp`	0.0046
`curspeed`	0.0004	`truewavedir`	0.0058	`truewavedir`	0.0045
`windspeed`	0.0004	`oceantemperature`	0.0057	`oceantemperature`	0.0044
`truwavedir`	0.0001	`windspeed`	0.0056	`truewinddir`	0.0033

Table 4.2: Feature importance of different models

The feature importances for all tree-based models shown in Table 4.2 indicated that the structure of the model is significantly influenced by the features `heading` and `cog`. This finding indicated that the models predicted the SOG by basis of ship movement direction i.e. heading and COG for a particular location. However, in a physical sense, it will be more insightful to consider the ship state and weather conditions that affect the prediction of the SOG.

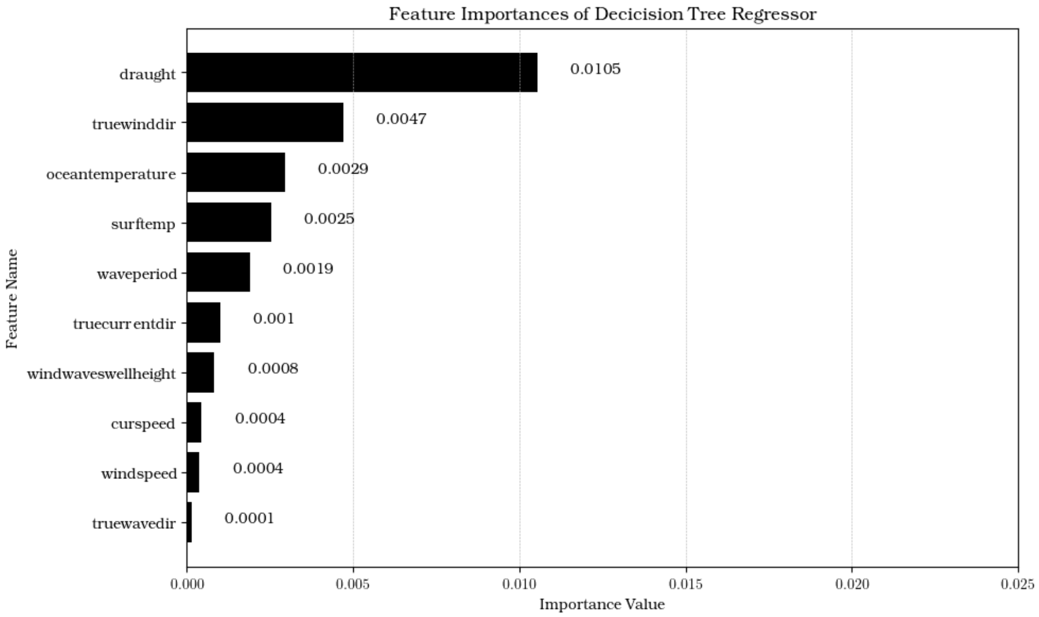


Figure 4.4: Feature importance of DTR

Excluding ship heading and COG. The ship draught T is regarded as the significant factors that affect the SOG prediction. This aligns with the theory of frictional resistance R_F encountered by the ship, which is discussed in Section 2.5.2.1. Equation (2.5.12) is a function of wetted surface area of bare hull S. Deeper draught T will result in more submerged area of the hull and this will consequently increase the frictional force R_F of the ship. Given a constant supply of power to the ship propulsion system, the speed of the ship will decrease which is shown in Equation (2.5.2).

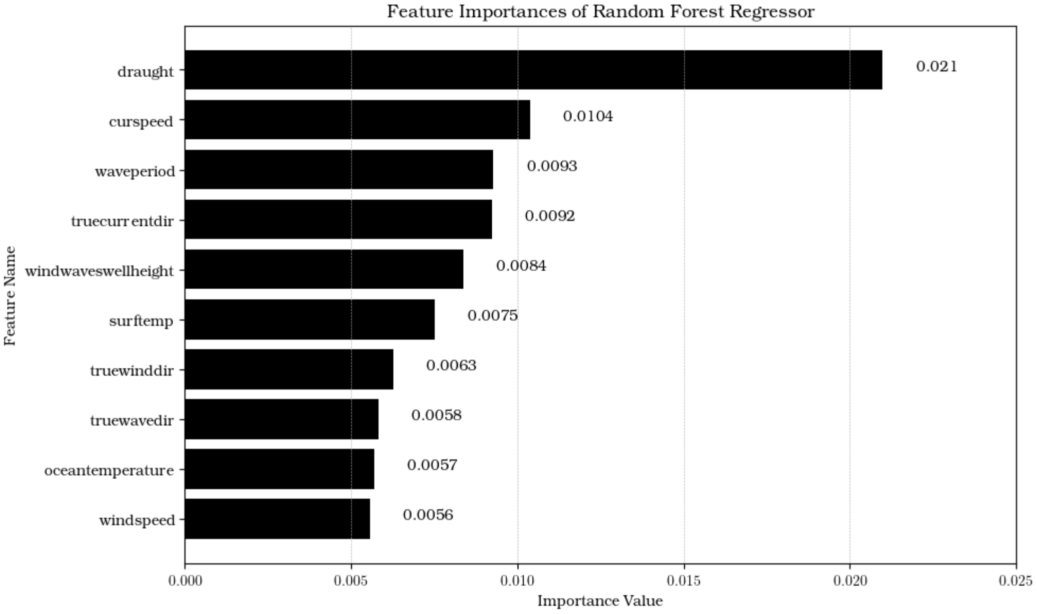


Figure 4.5: Feature importance of RFR

For weather states, both RFR and ETR considered current based information such as current speed and true current direction as the most significant weather factor that affect SOG prediction. This aligns to the proposed current correction methodology presented in Section 2.3.2 which states that the process of current correction to convert SOG to STW requires both the magnitude and direction of the current. The next influencing feature ranked by RFR and ETR model are wave related features which are significant wave height $H_{1/3}$, true wave direction, and the wave period `waveperiod`. This corresponds to the added resistance due to wave R_{AW} in the calculation of total resistance R_{TOTAL} encountered by the ship. The wind related features, which are wind speed and its true direction, corresponds to added resistance due to wind force R_{AA} and is found to be the least influential in SOG prediction in RFR and ETR model.

Based on the behaviour of the Random Forest Regressor (RFR) and Extra Trees Regressor (ETR) models, it can be inferred that waves have a more significant impact on the Speed Over Ground (SOG) compared to the influence of wind during the ship's journey. However, the Decision Tree Regressor (DTR) model demonstrates that temperature-related features, such as Sea Surface Temperature (SST) and air temperature above the ocean, have a more significant effect on SOG predictions than most other features. While the importance of temperature is not as pronounced as in RFR or ETR models, this finding suggests that the ship's SOG is implicitly influenced

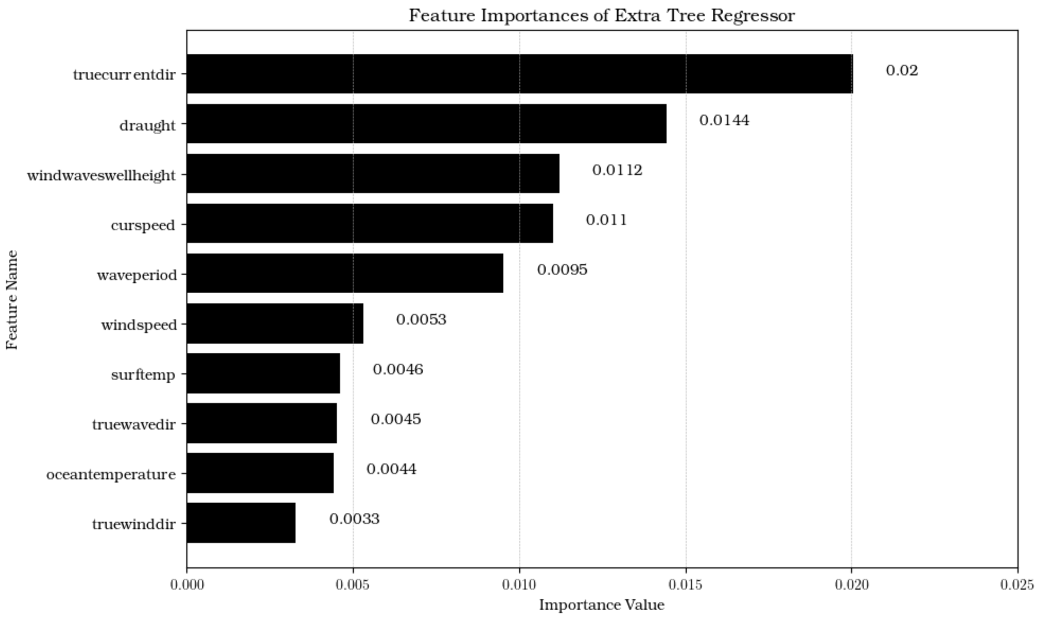


Figure 4.6: Feature importance of ETR

by the time of the travel or the season in which the journey takes place.

Structure of generated tree-based model

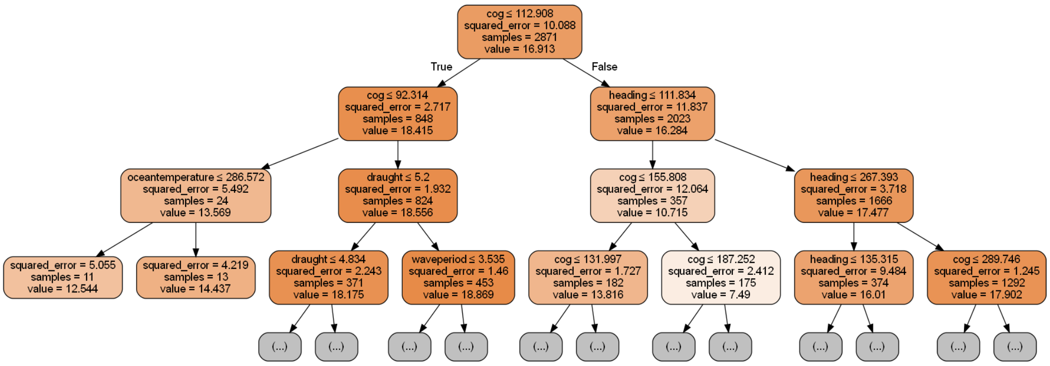

cog ≤ 112.908 squared_error = 10.088 samples = 2871 value = 16.913
True
False
cog ≤ 92.314 squared_error = 2.717 samples = 848 value = 18.415
heading ≤ 111.834 squared_error = 11.837 samples = 2023 value = 16.284
oceantemperature ≤ 286.572 squared_error = 5.492 samples = 24 value = 13.569
draught ≤ 5.2 squared_error = 1.932 samples = 824 value = 18.556
cog ≤ 155.808 squared_error = 12.064 samples = 357 value = 10.715
heading ≤ 267.393 squared_error = 3.718 samples = 1666 value = 17.477
squared_error = 5.055 samples = 11 value = 12.544
squared_error = 4.219 samples = 13 value = 14.437
draught ≤ 4.834 squared_error = 2.243 samples = 371 value = 18.175
waveperiod ≤ 3.535 squared_error = 1.46 samples = 453 value = 18.869
cog ≤ 131.997 squared_error = 1.727 samples = 182 value = 13.816
cog ≤ 187.252 squared_error = 2.412 samples = 175 value = 7.49
heading ≤ 135.315 squared_error = 9.484 samples = 374 value = 16.01
cog ≤ 289.746 squared_error = 1.245 samples = 1292 value = 17.902
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)

Figure 4.7: Structure of DTR

To understand the effect of hyperparameter optimisation and feature importance, analysis on the structure of generated tree-based models will be performed. The shading in the nodes indicated the likelihood of the decision, where darker shading means greater likelihood. Each node indicates information on the splitting feature with the threshold, the SSR score, amount of samples and the predicted SOG value. Even after pruning, the tree can grow relatively large, therefore, the illustration of the trees is limited to a depth of `max_depth = 3` and for RFR and ETR, only the illustration of a specific tree in the forest will be shown.

The structure of the optimised decision tree shown in Figure 4.7 show the effect of regularisation at the leaf node. For example, the leaf nodes that splits the feature, ocean temperature, does not completely minimise the SSR. This is caused by the hyperparameter tuning of the minimum samples at the leaf node, which was set at m̂in_samples_leaf = 10, splitting these nodes further will cause subsequent leaf nodes that have less than 10 samples. In this figure, the significance of COG and ship heading can be clearly seen, as it is used to split many of the internal nodes in the tree.

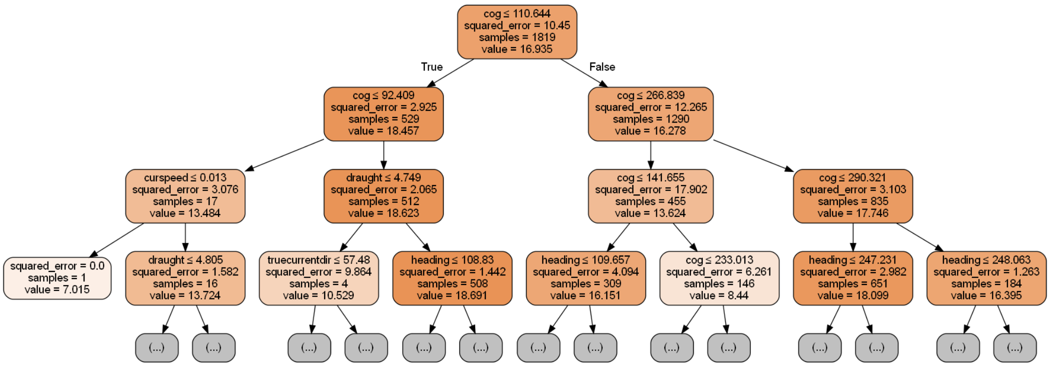


Figure 4.8: Structure of RFR

The illustration of the first RFR tree is shown in Figure 4.8. Similar to DTR, both COG and ship heading are regarded as the best features to split the internal node. In this tree of the forest, the effect of allowing full tree growth can be observed in the leaf node when splitting the feature current speed. This tree is able to minimise the SSR to its possible minimum value 0, and the leaf node cannot be further split as there are no more available samples. The effect of bagging for the dataset and feature selection in RFR can also be observed in this tree as the structure of this tree is completely different to DTR tree shown in Figure 4.7.

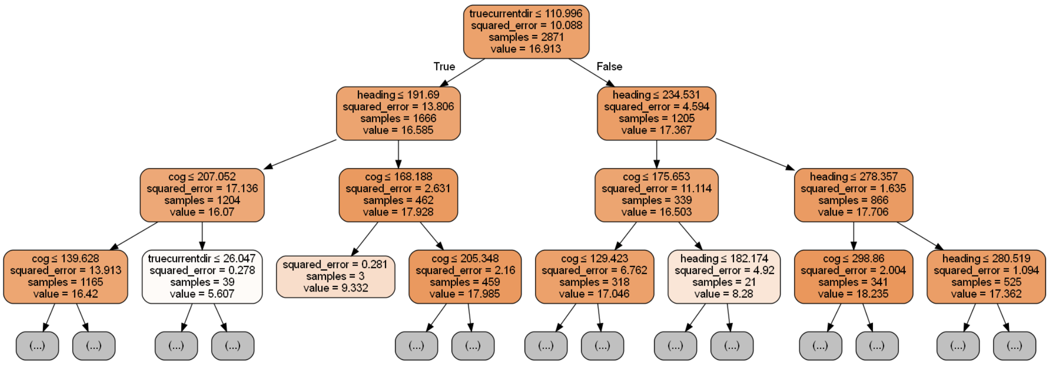


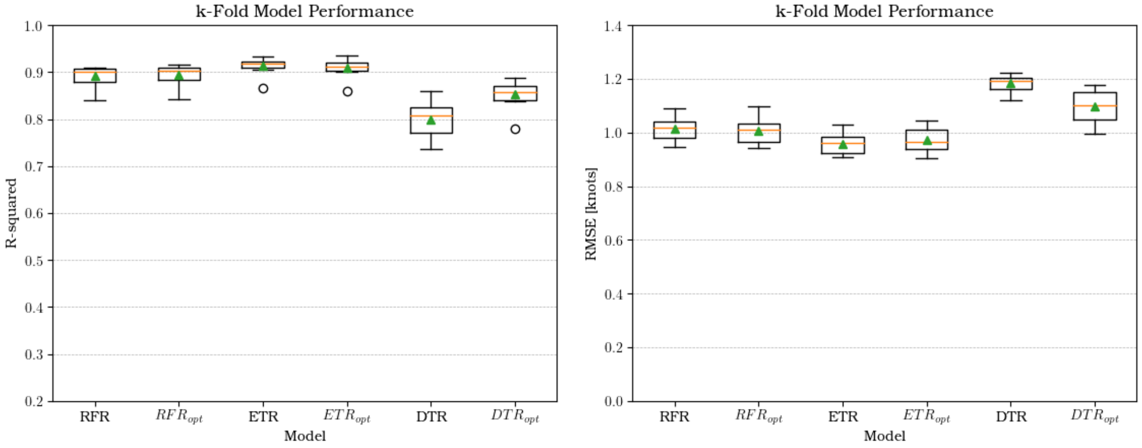
Figure 4.9: Structure of ETR

The random selection of the feature to split of ETR can be seen in the structure of the first tree in an ETR in Figure 4.9. Since both DTR and RFR uses greedy algorithm, i.e. it finds the possible splits which minimise the cost function, both model selected COG as the parent node. However, the randomness in feature selection of ETR can be clearly observed in this illustration, the model selects the true current direction as the parent node. Also, due to regularisation of ETR, the leaf node when splitting

COG does not completely minimise the SSR. The split is not allowed since it does not meet the tuning criteria of `min_samples_split = 9`.

Evaluation of k-fold cross-validation

The performance of the model is evaluated using the training dataset using 10-fold cross validation. This means that the training will be repeated 10 times using 9 of the folds as training dataset, the remaining fold will be used as validation dataset. The results from k-folding validation process is shown in Figure 4.10. The inside (orange) line represents the median i.e. 50% of the score in k-folding. The top and the bottom of the box correspond to the first i.e. 25% and third quartile i.e. 75% respectively. The whiskers represent the lowest data point within the 1.5 Interquartile Range (IQR) of the lowest quartile and the highest point of data within 1.5 IQR of the upper quartile. The mean is indicated by the (green) triangle. Data points beyond the whisker range is shown as hollow circle.



(a) k-fold R^2 validation performance (b) k-fold RMSE validation performance

Figure 4.10: Box plots of different models with default and optimised parameter in k-folding for training dataset

The box plots indicated that ETR achieved the best performance, the model is able to achieve R^2 score of around 91% and RMSE of around 0.96 knots. The model is also relatively stable, which is indicated by the narrow box plots. RFR also achieved similar performance, achieving R^2 score of about 89% and RMSE of approximately 1.00 knots and slightly worse stability. This behaviour may be caused due to the high variance shown from the learning curves shown in Figure 4.2 and Figure 4.3. This means that the model will have slight tendency to overfit.

DTR greatly benefits from regularisation, the model achieves an increase of about 5% for the R^2 score and a reduction from about 1.2 knots to 1.1 knots for the RMSE. To summarise, all tree-based models exhibited good fit performance with mean/median R^2 scores above 80%. However, the value of RMSE range is quite significant,

Features	Count	Mean	Std.	Min	25%	50%	75%	Max
sog	957	16.99	3.10	5.10	16.68	18.05	18.72	21.00
cog	957	196.73	86.72	56.02	102.32	185.22	282.18	319.85
heading	957	187.88	88.47	67.90	100.86	124.65	279.19	319.85
draught	957	5.23	0.19	4.74	5.11	5.29	5.38	5.66
windspeed	957	6.45	3.04	0.40	4.11	6.13	8.21	15.85
oceantemperature	957	282.27	6.48	267.25	276.80	281.91	288.42	295.70
waveperiod	957	3.40	0.88	1.67	3.06	3.62	4.22	7.01
surftemp	957	283.22	5.72	273.15	277.98	282.73	288.82	294.92
windwaveswellheight	957	0.77	0.54	0.08	0.37	0.66	0.94	3.24
curspeed	957	0.09	0.07	0.00	0.05	0.07	0.13	0.50
truewinddir	957	91.39	56.23	0.03	38.80	95.25	142.83	179.86
truecurrentdir	957	90.75	57.76	0.26	31.52	90.44	144.65	179.95
truewavedir	957	86.90	55.74	0.06	36.24	81.54	138.04	179.81

Table 4.3: Descriptive statistics of DS_{year}

the values lies between 1.00 to 1.20 knots across the models. To put this into scale, the mean SOG of the training data is at 16.91 knots as shown in Table 3.3.

Analysing the testing dataset

Once the best model is determined, the model will be tested against the testing dataset. The testing dataset correspond to 957 datasets across 2021, the dataset for the whole year is indicated as DS_{year}. To investigate the effect of data points on the model performance, the dataset is split into two seasons, DS_{summer}, which corresponds to summer datasets and it represents data between May 2021 and October 2021, there are 454 data points between this period. Winter dataset, DS_{winter}, correspond to testing datasets between January 2021 and April 2021 as well as November 2021 and December 2021 which correspond to 503 data points. Any missing values which are present in the testing dataset will be using KNNImputer.

Evaluation of testing datasets

To evaluate model performance, the testing datasets DS_{year}, DS_{summer} and DS_{winter} will be passed through the optimised model with hyperparameter values presented in Table 4.1. The model performance is summarised in Table 4.4

From Table 4.4, it can be observed that all tree-based model is able to achieve good results on testing datasets. All tree-based model obtained R^2 scores above 90% and is also able to obtain RMSE as low as 0.409 knots. In general, all tree based models performs better when DS_{winter} datasets are used for testing. The results also show that ETR and RFR have nearly identical performance for SOG prediction. Both of the model achieved model fit score R^2 of about 96% and RMSE of about 0.5 – 0.6 knots. While the fit performance of DTR is comparable to both RFR and ETR, DTR makes more substantial errors during predictions, this can be seen from the MAE of DTR from figure Table 4.4, DTR model makes larger error during the prediction, this pattern is consistent for all error metrics such as MAE, RMSE, MAD and MAPE,

Model	Dataset	R^2	expVar	MAE	RMSE	MAD	MAPE
		[%]	[%]	$[kn]$	$[kn]$	$[kn]$	[%]
DTR$_{OPT}$	DS_{year}	90.10	90.12	0.629	0.975	0.420	4.21
	DS_{winter}	93.18	93.19	0.561	0.846	0.390	3.92
	DS_{summer}	85.69	84.90	0.704	1.100	0.451	4.52
RFR$_{OPT}$	DS_{year}	96.59	96.60	0.335	0.572	0.187	2.29
	DS_{winter}	**98.41**	**98.42**	**0.265**	**0.409**	**0.173**	**1.94**
	DS_{summer}	94.02	94.14	0.412	0.710	0.215	2.68
ETR$_{OPT}$	DS_{year}	96.82	96.82	0.347	0.553	0.214	2.35
	DS_{winter}	98.40	98.40	0.287	0.410	0.196	2.03
	DS_{summer}	95.49	94.68	0.413	0.676	0.239	2.70
MLR	DS_{year}	69.75	69.85	1.139	1.704	0.908	7.64
	DS_{winter}	68.16	68.17	1.129	1.828	0.871	7.94
	DS_{summer}	71.43	71.87	1.150	1.554	0.951	7.32

Table 4.4: Performance indices for different modelling approach and different testing datasets

where DTR model exhibits errors that are twice as large as those of other models with exception to MLR.

Chapter 5

Summary and Outlook

In this chapter the summary of this research will be discussed. This section includes reflections of the research process and presents any possible suggestions and recommendations in this line of research. This chapter concludes this thesis.

Bibliography

Misganaw Abebe, Yongwoo Shin, Yoojeong Noh, Sangbong Lee, and Inwon Lee. Machine learning approaches for ship speed prediction towards energy efficient shipping. *Applied Sciences*, 10(7):2325, 2020. doi:10.3390/app10072325.

E. Bal Beşikçi, O. Arslan, O. Turan, and A. I. Ölçer. An artificial neural network based decision support system for energy efficient ship operations. *Computers & Operations Research*, 66:393–401, 2016. ISSN 03050548. doi:10.1016/j.cor.2015.04.004.

J Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *J. Mach. Learn. Res.*, 2012. URL https://www.semanticscholar.org/paper/Random-Search-for-Hyper-Parameter-Optimization-Bergstra-Bengio/188e247506ad992b8bc62d6c74789e89891a984f.

Volker Bertram. *Practical ship hydrodynamics*. Elsevier Butterworth-Heinemann, Oxford, 2000. ISBN 0750648511.

Nicolas Bialystocki and Dimitris Konovessis. On the estimation of ship's fuel consumption and speed curve: A statistical approach. *Journal of Ocean Engineering and Science*, 1(2):157–166, 2016. ISSN 24680133. doi:10.1016/j.joes.2016.02.001.

Adrian Biran, Rubén López-Pulido, and Javier de Juana Gamo, editors. *Ship hydrostatics and stability*. Butterworth-Heinemann, Amsterdam, second edition edition, 2014. ISBN 978-0-08-098287-8.

Lothar Birk. *Fundamentals of ship hydrodynamics: Fluid mechanics, ship resistance and propulsion / Lothar Birk*. John Wiley & Sons, Hoboken, New Jersey, 1st edition, 2019. ISBN 1118855515. doi:10.1002/9781119191575. URL https://onlinelibrary.wiley.com/doi/book/10.1002/9781119191575.

Elzbieta M. Bitner-Gregersen. Joint probabilistic description for combined seas. In *24th International Conference on Offshore Mechanics and Arctic Engineering: Volume 2*, pages 169–180. ASMEDC, 2005. ISBN 0-7918-4196-0. doi:10.1115/OMAE2005-67382.

Werner Blendermann. Parameter identification of wind loads on ships. *Journal of Wind Engineering and Industrial Aerodynamics*, 51(3):339–351, 1994. ISSN 0167-6105. doi:10.1016/0167-6105(94)90067-1. URL https://www.sciencedirect.com/science/article/pii/0167610594900671.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. ISSN 1573-0565. doi:10.1023/A:1010933404324. URL https://link.springer.com/article/10.1023/a:1010933404324.

John P. Breslin and Poul Andersen. *Hydrodynamics of ship propellers*, volume 3 of *Cambridge ocean technology series*. Cambridge University Press, Cambridge, 1994. ISBN 9780521413602. doi:10.1017/CBO9780511624254. URL https://www.cambridge.org/core/books/hydrodynamics-of-ship-propellers/60B96C78A2B5CC3FADC96C9512A522F6.

Charles L. Bretschneider. *Generation of waves by wind. State of the art*. 1965. URL https://apps.dtic.mil/sti/citations/ad0612006.

Wikimedia Commons. Routen der bornholmslinjen, 2010. URL https://de.wikipedia.org/wiki/Datei:Bornholmerf%C3%A6rgen_route_map.svg.

Andrea Coraddu, Luca Oneto, Francesco Baldi, and Davide Anguita. Vessels fuel consumption forecast and trim optimisation: A data analytics perspective. *Ocean Engineering*, 130:351–370, 2017. ISSN 00298018. doi:10.1016/j.oceaneng.2016.11.058. URL https://www.sciencedirect.com/science/article/pii/S0029801816305571.

Danish Maritime Authority. Safety at sea, navigational information, AIS data, 2023. URL https://dma.dk/safety-at-sea/navigational-information/ais-data.

Thomas G. Dietterich. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine Learning*, 40(2):139–157, 2000. ISSN 1573-0565. doi:10.1023/A:1007607513941. URL https://link.springer.com/article/10.1023/A:1007607513941.

Yuquan Du, Qiang Meng, Shuaian Wang, and Haibo Kuang. 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, 2019. ISSN 01912615. doi:10.1016/j.trb.2019.02.004.

Aurélien Géron. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems / Aurélien Géron*. O'Reilly, Sebastopol, CA, second edition edition, 2019. ISBN 978-1-492-03264-9.

Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine Learning*, 63(1):3–42, 2006. ISSN 1573-0565. doi:10.1007/s10994-006-6226-1. URL https://link.springer.com/article/10.1007/s10994-006-6226-1.

Christos Gkerekos, Iraklis Lazakis, and Gerasimos Theotokatos. Machine learning models for predicting ship main engine fuel oil consumption: A comparative study. *Ocean Engineering*, 188:106282, 2019. ISSN 00298018. doi:10.1016/j.oceaneng.2019.106282.

H. E. Guldhammer and S. A. Harvald. Ship resistance - effect of form and principal dimensions. (revised). *Danish Technical Press, Danmark, Danmarks Tekniske Hojskole, kademisk Forlag, St. kannikestrade 8, DK 1169 Copenhagen*, 1974. URL https://repository.tudelft.nl/islandora/object/uuid:4a6f2694-a3ab-4a90-beac-7f38c41d4e40.

Alon Halevy, Peter Norvig, and Fernando Pereira. The unreasonable effectiveness of data. *IEEE Intelligent Systems*, 24(2):8–12, 2009. ISSN 1941-1294. doi:10.1109/MIS.2009.36.

Michael Haranen, Pekka Pakkanen, Risto Kariranta, and Jouni Salo. White, grey and black-box modelling in ship performance evaluation. 2016. URL https://www.researchgate.net/publication/301355727_White_Grey_and_Black-Box_Modelling_in_Ship_Performance_Evaluation.

Trevor Hastie, Robert Tibshirani, and J. H. Friedman. *The elements of statistical learning: Data mining, inference, and prediction / Trevor Hastie, Robert Tibshirani, Jerome Friedman*. Springer series in statistics. Springer, New York, 2nd ed. edition, 2009. ISBN 9780387848570. doi:10.1007/b94608.

U. Hollenbach. Estimating resistance and propulsion for single-screw and twin-screw ships in the preliminary design. In Chryssostomos Chryssostomidis and Kaj. Ed Johansson, editors, *10th international conference on computer applications in shipbuilding*, International conference on computer applications in shipbuilding, pages 237–250. 1999. ISBN 1561720240.

Leo H. Holthuijsen. *Waves in oceanic and coastal waters*. Cambridge University Press, Cambridge, 2007. ISBN 9780521860284.

J. Holtrop. A statistical re-analysis of resistance and propulsion data. *Published in International Shipbuilding Progress, ISP, Volume 31, Number 363*, 1984. URL https://repository.tudelft.nl/islandora/object/uuid%3Aca12a502-fc85-45e4-a078-db7284127e3c.

J. Holtrop and G.G.J. Mennen. A statistical power prediction method. *Netherlands Ship Model Basin, NSMB, Wageningen, Publication No. 603, Published in: International Shipbuilding Progress, ISP, Volume 25, Number 290, October 1978*, 1978. URL https://repository.tudelft.nl/islandora/object/uuid%3A62c40df8-18cc-4225-a65a-54ff5c1609fb.

J. Holtrop and G.G.J. Mennen. An approximate power prediction method. *Netherlands Ship Model Basin, NSMB, Wageningen, Publication No. 689, Published in: International Shipbuilding Progress, ISP, Volume 29, Nr 335, 1982*, 1982. URL https://repository.tudelft.nl/islandora/object/uuid%3Aee370fed-4b4f-4a70-af77-e14c3e692fd4.

IMO. Revised guidelines for the onboard operational use of shipborne Automatic Identification Systems (AIS), 2015. URL https://www.imo.org/en/OurWork/Safety/Pages/AIS.aspx.

IMO. Fourth IMO GHG study 2020. *International Maritime Organization London, UK*, 2020.

ITTC. Analysis of speed/power trial data. *ITTC Recommended Procedures and Guidelines*, pages 25–33, 2014.

G Jensen. Moderne schiffslinien. *Handbuch der Werften*, 22:93, 1994.

Seong-Hoon Kim, Myung-Il Roh, Min-Jae Oh, Sung-Woo Park, and In-Il Kim. Estimation of ship operational efficiency from ais data using big data technology. *International Journal of Naval Architecture and Ocean Engineering*, 12:440–454, 2020. ISSN 2092-6782. doi:10.1016/j.ijnaoe.2020.03.007. URL https://www.sciencedirect.com/science/article/pii/S2092678220300091.

Stig Staghøj Knudsen. *Sail Shape Optimization with CFD*. Master, 2013. URL https://www.researchgate.net/profile/stig-knudsen/publication/257143505_sail_shape_optimization_with_cfd.

A. M. Kracht. Design of bulbous bows. *Publication of: Society of Naval Architects and Marine Engineers*, (Paper No. 7), 1978. URL https://trid.trb.org/view/81194.

Hans Otto Kristensen and Marie Lützen. Prediction of resistance and propulsion power of ships. *Clean Shipping Currents*, 1(6):1–52, 2012. ISSN 2242-9794.

Max Kuhn and Kjell Johnson. *Applied predictive modeling*. Springer, New York, 2013. ISBN 9781461468486.

Xiao Lang. *Development of Speed-power Performance Models for Ship Voyage Optimization*. PhD thesis, 2020. URL https://www.researchgate.net/publication/347977212_Development_of_Speed-power_Performance_Models_for_Ship_Voyage_Optimization.

Xiaohe Li, Yuquan Du, Yanyu Chen, Son Nguyen, Wei Zhang, Alessandro Schönborn, and Zhuo Sun. Data fusion and machine learning for ship fuel efficiency modeling: Part i – voyage report data and meteorological data. *Communications in Transportation Research*, 2:100074, 2022. ISSN 27724247. doi:10.1016/j.commtr.2022.100074.

Diesel MAN. Turbo. basic principles of ship propulsion. *MAN Diesel & Turbo, Copenhagen*, 2011.

Anthony F. Molland. *The maritime engineering reference book: A guide to ship design, construction and operation / edited by Anthony F. Molland*. Butterworth-Heinemann, 2011. ISBN 9780080560090. doi:10.1016/B978-0-7506-8987-8.X0001-7.

Juan José Montaño Moreno, Alfonso Palmer Pol, Albert Sesé Abad, and Berta Cajal Blasco. Using the r-mape index as a resistant measure of forecast accuracy. *Psicothema*, 25(4):500–506, 2013. ISSN 1886-144X. doi:10.7334/psicothema2013.23. URL https://www.researchgate.net/publication/257812432_Using_the_R-MAPE_index_as_a_resistant_measure_of_forecast_accuracy.

Ulrik D. Nielsen and Jesper Dietz. Ocean wave spectrum estimation using measured vessel motions from an in-service container ship. *Marine Structures*, 69:102682, 2020. ISSN 09518339. doi:10.1016/j.marstruc.2019.102682.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011. URL http://jmlr.org/papers/v12/pedregosa11a.html.

Joan P. Petersen, Ole Winther, and Daniel J. Jacobsen. A machine-learning approach to predict main energy consumption under realistic operational conditions. *Ship Technology Research*, 59(1):64–72, 2012a. ISSN 0937-7255. doi:10.1179/str.2012.59.1.007.

Jóan Petur Petersen. Mining of ship operation data for energy conservation. 2011. URL https://orbit.dtu.dk/en/publications/mining-of-ship-operation-data-for-energy-conservation.

Jóan Petur Petersen, Daniel J. Jacobsen, and Ole Winther. Statistical modelling for ship propulsion efficiency. *Journal of Marine Science and Technology*, 17(1): 30–39, 2012b. ISSN 0948-4280. doi:10.1007/s00773-011-0151-0.

Stian Glomvik Rakke. *Ship emissions calculation from AIS*. PhD thesis, NTNU, 2016. URL https://ntnuopen.ntnu.no/ntnu-xmlui/handle/11250/2410741.

D. Ronen. The effect of oil price on containership speed and fleet size. *Journal of the Operational Research Society*, 62(1):211–216, 2011. ISSN 0160-5682. doi:10.1057/jors.2009.169.

H. Schneekluth and Volker Bertram. *Ship design for efficiency and economy*. Butterworth-Heinemann, Oxford, 2nd ed. edition, 1998. ISBN 9780080517100. URL https://ebookcentral.proquest.com/lib/kxp/detail.action?docID=313756.

T.W.P. Smith, J.P. Jalkanen, B.A. Anderson, J. J. Corbett, J. Faber, S. Hanayama, E. O'Keeffe, S. Parker, L. Johansson, L. Aldous, C. Raucci, M. Traut, S. Ettinger, D. Nelissen, D. S. Lee, S. Ng, A. Agrawal, J. J. Winebrake, M. Hoen, S. Chesworth, and A. Pandey. Third imo greenhouse gas study 2014. 2015. URL https://research.manchester.ac.uk/en/publications/third-imo-greenhouse-gas-study-2014.

Omer Soner, Emre Akyuz, and Metin Celik. Use of tree based methods in ship performance monitoring under operating conditions. *Ocean Engineering*, 166: 302–310, 2018. ISSN 00298018. doi:10.1016/j.oceaneng.2018.07.061. URL https://www.sciencedirect.com/science/article/pii/S0029801818314446.

Stopford. The organization of the shipping market. page 47, 2009.

Tin Kam Ho. Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, pages 278–282 vol.1, 1995. doi:10.1109/ICDAR.1995.598994.

K Torsethaugen, S. Haver, and S. Norway. Simplified double peak spectral model for ocean waves. 2004. URL https://www.semanticscholar.org/paper/Simplified-Double-Peak-Spectral-Model-For-Ocean-Torsethaugen-Haver/0f1b1509791d441861ff6c2940dd13b1f939f149.

Shuaian Wang and Qiang Meng. Sailing speed optimization for container ships in a liner shipping network. *Transportation Research Part E: Logistics and Transportation Review*, 48(3):701–714, 2012. ISSN 13665545. doi:10.1016/j.tre.2011.12.003. URL https://www.sciencedirect.com/science/article/pii/S1366554511001554.

N. Wijnolst, Tor Wergeland, and Kai Levander. *Shipping Innovation*. IOS Press, 2009. ISBN 9781586039431.

Ran Yan, Shuaian Wang, and Yuquan Du. Development of a two-stage ship fuel consumption prediction and reduction model for a dry bulk ship. *Transportation Research Part E: Logistics and Transportation Review*, 138:101930, 2020. ISSN 13665545. doi:10.1016/j.tre.2020.101930.

Ran Yan, Shuaian Wang, and Harilaos N. Psaraftis. Data analytics for fuel consumption management in maritime transportation: Status and perspectives. *Transportation Research Part E: Logistics and Transportation Review*, 155:102489, 2021. ISSN 13665545. doi:10.1016/j.tre.2021.102489. URL https://www.sciencedirect.com/science/article/pii/S1366554521002519.

Dong Yang, Lingxiao Wu, Shuaian Wang, Haiying Jia, and Kevin X. Li. How big data enriches maritime research – a critical review of automatic identification system (ais) data applications. *Transport Reviews*, 39(6):755–773, 2019. ISSN 0144-1647. doi:10.1080/01441647.2019.1649315. URL https://www.researchgate.net/publication/334738291_How_big_data_enriches_maritime_research_-_a_critical_review_of_Automatic_Identification_System_AIS_data_applications.

Liqian Yang, Gang Chen, Jinlou Zhao, and Niels Gorm Malý Rytter. Ship speed optimization considering ocean currents to enhance environmental sustainability in maritime shipping. *Sustainability*, 12(9):3649, 2020. doi:10.3390/su12093649.

Declaration in lieu of oath

I hereby solemnly declare that I have independently completed this work or, in the case of group work, the part of the work that I have marked accordingly. I have not made use of the unauthorised assistance of third parties. Furthermore, I have used only the stated sources or aids and I have referenced all statements (particularly quotations) that I have adopted from the sources I have used verbatim or in essence.
I declare that the version of the work I have submitted in digital form is identical to the printed copies submitted.
I am aware that, in the case of an examination offence, the relevant assessment will be marked as ‘insufficient’ (5.0). In addition, an examination offence may be punishable as an administrative offence (Ordnungswidrigkeit) with a fine of up to €50,000. In cases of multiple or otherwise serious examination offences, I may also be removed from the register of students.
I am aware that the examiner and/or the Examination Board may use relevant software or other electronic aids in order to establish an examination offence has occurred
I solemnly declare that I have made the previous statements to the best of my knowledge and belief and that these statements are true and I have not concealed anything.
I am aware of the potential punishments for a false declaration in lieu of oath and in particular of the penalties set out in Sections 156 and 161 of the German Criminal Code (Strafgesetzbuch; StGB), which I have been specifically referred to.

Section 156 False declaration in lieu of an oath

Whoever falsely makes a declaration in lieu of an oath before an authority which is competent to administer such declarations or falsely testifies whilst referring to such a declaration incurs a penalty of imprisonment for a term not exceeding three years or a fine.

Section 161 Negligent false oath; negligent false declaration in lieu of oath

(1) Whoever commits one of the offences referred to in Sections 154 to 156 by negligence incurs a penalty of imprisonment for a term not exceeding one year or a fine. (2) No penalty is incurred if the offender corrects the false statement in time. The provisions of Section 158 (2) and (3) apply accordingly.

______________________	______________________
Place,date	Signature