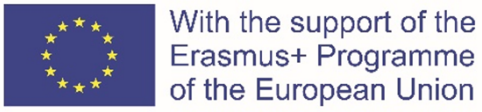
Master thesis : Vessel Performance Analysis using High Frequency Operational Data

Auteur : Abbas, Farhan
Promoteur(s) : 14957; 18526
Faculté : Faculté des Sciences appliquées
Diplôme : Master : ingénieur civil mécanicien, à finalité spécialisée en "Advanced Ship Design"
Année académique : 2021-2022
URI/URL : http://hdl.handle.net/2268.2/16491



Vessel Performance Analysis using High Frequency Operational Data

Submitted on July 28, 2022

by

ABBAS Farhan | Max Planck Street | 18059 | farhanabbas@outlook.com

Student ID No.: 221 200 017

First Reviewer:
Prof. Dr.Eng. Patrick Kaeding
Ship Structure / University of Rostock
Universitätsplatz 1
18055 Rostock
Germany

Second Reviewer:
Dipl.-Ing. Lutz Kleinsorge
Mecklenburger Metallguss GmbH
Teterower Str. 1
17192 Waren (Müritz)
Germany

Contents

List of Figures

List of Tables

DECLARATION OF AUTHORSHIP

I declare that this thesis and the work presented in it are my own and have been generated by me as the result of my own original research.

Where I have consulted the published work of others, this is always clearly attributed.

Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.

I have acknowledged all main sources of help.

Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

This thesis contains no material that has been submitted previously, in whole or in part, for the award of any other academic degree or diploma.

I cede copyright of the thesis in favour of the University of Rostock.

Date: July 28, 2022 Signature

ABSTRACT

The increase in Greenhouse Gas (GHG) emissions due to shipping and its impact on climate change is a concern for the maritime industry. Urgent steps are required to achieve the International Maritime Organization (IMO) target of 50% GHG reduction by 2050. The recent boom in digitalization of the shipping industry can be vital to improving the energy efficiency of vessels. The high frequent operational data obtained from the onboard sensors can be used to analyse the decrease in ship's performance over time due to aging and sufficient measures can be suggested timely to achieve the maximum possible efficiency.

This master thesis investigates the performance of an intercontinental 14000 TEU container vessel using high-frequency operational data. The operational data has been obtained from University of Rostock server. Statistical outliers are removed using ISO19030 standards. Then this data is further filtered for different environmental and operational conditions using simple filtering techniques and Machine Learning algorithms. Outputs of both methods are compared with the design data to analyze the vessel's performance and the most suitable method is proposed. Moreover, the Energy Efficiency Operational Indicator (EEOI) of the vessel is investigated over time. Results show that the performance of the vessel is constant over the investigated time period. A good agreement has also been found between the operational data with the design data.

Furthermore, this study also compares the performance of different machine learning algorithms; Random Forrest, Decision Tree, Gradient Boosting, Multilayer Perceptron and Least square methods over the filtered high-frequency data set. Blind testing results show that the Random Forest algorithm has the best performance.

Keywords: Greenhouse Gas Emissions, Vessel Performance Analysis, Big Data Analytics, Machine Learning

1 INTRODUCTION

The international regulations on shipping have made the maritime transport industry competitive. The restrictions on gas emissions by International Maritime Organization (IMO) are becoming more strict. IMO is developing different strategies to reduce the air pollution and Greenhouse Gas (GHG) emissions caused by ships. According to MEPC (2018), the long-term goal of IMO is to reduce GHG emissions by 50% by 2050 and 40% by 2030 as compared to the 2008 numbers. Short-term regulations are regularly enforced by IMO to achieve the goal of 2050. On 1st January 2020, the emissions of sulphur oxides (SOx) are limited to 0.50% mass by mass. Amendments are adopted to MARPOL Annex VI in this regard in the 76th session of the Marine Environment Protection Committee (MEPC) in June 2021. MEPC is inforcing the Carbon Intensity Indicator (CII) and the Energy Efficiency Design Index (EEXI) in January 2023. This will demand immediate measures for the shipping industries to keep gas emissions and air pollution in check.

International Maritime Organization suggests a variety of measures to improve the efficiency of a ship such as the installation of renewable power sources (e.g. solar panels, wind energy etc) for on-board power requirements, hull cleaning, usage of energy-efficient equipment etc. A mixture of operational and technical solutions is required to achieve the IMO Greenhouse gas goal. Figure 1 shows some of these solutions with the approximate GHG reduction potential proposed by IMO.

According to the UN Global Pulse Report (2012), it is the era of the Internet of Things (IoT) and Industry 4.0. The companies which use data to their advantage have 5-6 % of increased productivity according to the study conducted by Brynjolfsson, Hitt, and H. H. Kim (2011) on 179 big companies. The Internet of Things (IoT) deals with modern technologies such as wired/wireless sensors, networks pervasive computing etc. The inclusion of IoT in the modern industry has resulted in an enormous amount of data collected from various sources. This data has to be processed and analyzed in an efficient way. This processing and analysis require big data analytics (BDA). The maritime industry is also experiencing a big data revolution with the recent introduction of digitalization. This digital transformation presents opportunities and challenges to the researchers to use this data for data-driven decision-making and increase shipping productivity. The introduction of IoT and BDA can be vital in meeting the IMO goals on GHG emissions.

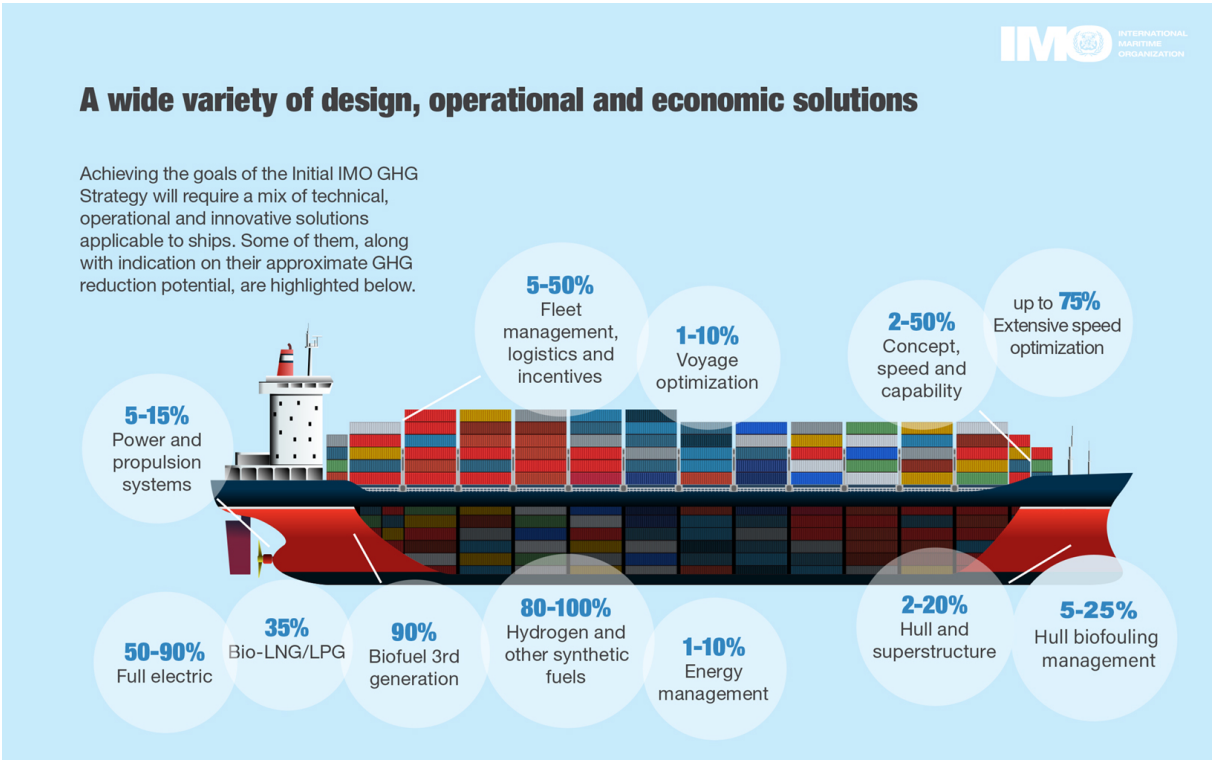


Figure 1: GHG reduction potential of various operational solutions provided by IMO (*Online Source* 2021)

Research has been carried out to optimize vessel efficiency by taking measures such as fuel cost and route optimization. These measures significantly impact the operational costs and the shipping company's finances. Modern ship operators use sensor technology to collect extensive data on ships' operational parameters from onboard and shore-based sources. Operational performance of the ship is monitored using this data. Ship operational data analysis helps the shipping companies to take timely measures such as hull cleaning, propeller retrofit, engine maintenance etc. to optimize the running costs and decrease GHG emissions. So, IoT and BDA are the solutions for shipping companies to meet the IMO requirements and improve vessel efficiency.

1.1 Literature Review

Many types of research have been carried out to monitor and analyse vessel performance. Some shipping companies provide in-house built software tools to voyage planning, onboard route optimization, fleet performance management and route advisory services (*StormGeo Optimziation Tools* 2022). Researchers have given much attention to the prediction of engine power or fuel consumption to improve the energy efficiency. Various statistical and empirical models have been developed to solve this problem (Erto et al. 2015, Sasa et al. 2015). However, these models have not succeeded to deal with non-linearity of data. Therefore, researchers have implemented the Machine Learning

algorithms in their research to correctly predict the ship's fuel consumption and engine power. Karagiannidis and Themelis (2021) have utilized Artificial Neural networks for training predictive models. Beşikçi et al. (2016) and Brandsæter and Vanem (2018) have used regression models for the same purpose. Soner, Akyuz, and Celik (2018) and Gkerekos, Lazakis, and Theotokatos (2019a)) have studied the implementation of ensemble models.

Researchers have targeted the optimization approach such as speed optimization and trim optimization to improve the vessel's operational efficiency. For this approach, Big Data Analytics (BDA) has been implemented. Wang et al. (2017) have used Hadoop platform architecture for improving the vessel's energy efficiency using BDA. In this research, route was divided based on the environmental factors and vessel's speed optimization was investigated for different portions of the route. Yan et al. (2018) have applied the clustering algorithm to find out a suitable route and the calculated the optimum engine speed for inland navigation. Coraddu et al. (2017) have used the BDA approach to optimize the trim and predict fuel consumption of a tanker. In this research, predictive models are developed for fuel consumption predictions and trim optimization method is developed based on these predictions. Lee et al. (2018), predicted fuel consumption using the big data involving the weather archive. A decision support system was developed for optimizing the fuel consumption in this study.

It can be noticed that the researchers are more inclined towards the application of Machine Learning (ML) to predict fuel consumption through the optimization of engine speed, vessel speed, route and trim. However, high-frequent data has been analysed by researchers to judge the ship performance. Some studies have implemented ISO 19030 to investigate the vessel's performance by calculating the speed and power performance values. Baumfalk (2018) has studied the application of ISO 19030 on a intercontinental container vessel.

1.2 Problem Statement

The propulsion performance of a ship is checked using calm water model tests. Furthermore, sea trials are performed to verify its performance. However, during these sea trials, the operating environmental conditions are assumed ideal. These tests are performed at a few speeds and drafts. Moreover, the additional loads on the propeller due to fouling, severe weather, shallow water effect and unfavourable loading conditions are neglected in these tests.

An overwhelming amount of data for the ship's real operating conditions is available due to the digitalization of the maritime industry. Apart from noon to noon reports logged by the authorized crew, high-frequency sensors are also installed on board the vessels to measure and report various operational parameters frequently (up to 5 Hz).

Ship's operational performance can be measured using this high frequency data but it is very difficult to analyse the enormous amount of data and draw conclusions about the propulsion performance. Additionally, filtering and cleaning of invalid data points are required to correctly assess the vessel's performance.

1.3 Objective and Scope

This study deals with Big Data Analytics of a 14000 TEU containership. The raw data obtained from the ship is stored, filtered and analyzed to check the propulsion performance of the vessel. The objectives of this study are:

- Study the various data sources (i.e. onboard and external sources) and their sampling rates. Understand the data logging methodology and data model structure used for storing high-frequency data.
- Access the required data from the University of Rostock online server using Python programming language (PyCharm Professional). Store the accessed data on the local hard disk for further analysis.
- Study various data filtering techniques. Apply filtering techniques to the stored operational data. Compare and analyse different data block lengths for the application of ISO 19030 and propose suitable block lengths for the available data.
- Train different machine learning (ML) models on the filtered data set. Perform the feature engineering. Compare and evaluate the performance of the trained ML models and propose the best model for the available data set.
- Develop operational propeller curve from filtered data set, machine learning models and compare them with the design curves to evaluate the vessel performance.

This thesis discusses a procedure to handle big data obtained from a container vessel. It includes the analysis of different statistical and environmental filters applied to the available data. Furthermore, it discusses the comparison and evaluation of the performance of Machine Learning (ML) algorithms such as polynomial regression, Decision Trees, Random Forest and Multilayer Perceptron (Artificial Neural Network). Finally, it provides a comparison between the operational propeller curve and the CFD based design curve.

In this thesis, the performance indicators (PIs) from ISO 19030 are not calculated because the information for dry docking or maintenance is not available in the data set. Furthermore, ISO 19030 suggests the calculation of speed and power performance values which are calculated after defining a reference period and evaluation period. This method is useful when the time duration of available data is more than two years (one

year for the reference period and the evaluation period each) but the present data set contains data for 8 months only. Moreover, it is unlikely to observe the degradation in the vessel's performance in this short period of time. Therefore, it is intuitive to compare the operational data with the design stage data rather than the operational data itself. The design data is based on Computational Fluid Dynamic (CFD) simulations because the sea trial test data is not available. This comparison will also validate the CFD powering predictions using real-world operational data.

1.4 Structure of the Report

The outline of thesis is given as:

- **Section 1** This section presents the introduction, problem statement, objective and scope of this study.
- **Section 2:** This section discusses the data sources for this research project. Furthermore, it also includes the data logging methodology and data model structure used for storing high-frequency data.
- **Section 3:** This section contains various data filtering techniques. Some of the data filtering methods are directly taken from the ISO 19030 standard and some modifications are made according to the available operational data. Furthermore, this chapter also discusses statistical data filtering methods for the removal of outliers which are widely used in Data Science.
- **Section 4:** In this section, available ship data is analyzed to understand its normal operational behaviours in the investigated time frame.
- **Section 5:** In this section, different machine learning (ML) models are trained on the filtered data set. The input features for training the ML models are also analyzed. The performance of the trained ML models is compared and the best model for the available data set is proposed.
- **Section 6:** This section discusses the comparison of filtered operational data and Machine Learning predictions with the data obtained during the ship design stage to evaluate the vessel performance.
- **Section 7:** This section presents the conclusion of the report.
- **Section 8:** This section presents the recommendations for the future work.

2 DATA SOURCE

This chapter discusses the data sources for this research project. Furthermore, it also includes the data logging methodology and data model structure used for storing high-frequency data.

The data used in this project are based on two different sources i.e. onboard data and external data.

2.1 Onboard Data

This type of data is measured by the sensors on board the ship. It relates to all the data obtained by sensors on board the vessel. The measured parameters include; engine power, engine shaft speed, draft, trim etc. This type of data also contains status reports (noon reports) which are logged by the crew on board the vessel.

2.2 External Data

It is the data related to the vessel, which independent third parties collect. It includes navigational data and environmental data. Navigational data and environmental data is collected through the Automatic Identification System (AIS) and the weather stations respectively.

The recorded data can be further classifed into high frequency and low frequency data. Status reports and loading conditions are included in the low frequency data. Motion and sensor data is the high frequency data.

2.3 Low Frequency Data

Noon reports logged by the crew on board the vessel and the ship's loading condition recorded while leaving a port are included in the low frequency data. The data sampling rate of such reports is low.

2.3.1 Noon Reports Data

Noon reports are traditionally recorded at 12 o clock at noon but in this research project noon data is recorded on the event basis. Events are classifed as below:

1. Beginning of the sea passage (BOSP)
2. End of the sea passage (EOSP)

3. Arrival at the port (ARRI)
4. Departure from the port (DEPA)
5. Noon at the sea (SEA)
6. Noon while manoeuvering (MAN)
7. Noon while drifting (PRT)
8. Noon while anchoring (ANC)
9. Begining of the anchorage (BANCH)
10. Ending of the anchorage (EANCH)
11. Begining of the drifting (BDRIFT)
12. Ending of the drifting (EDRIFT)

The infromation about the aforementioned events are logged in the noon reports by the authorized members of the crew on board the vessel. However, if the vessel is equipped with modern technology, such reports are updated regularly by the technical monitoring service providers via online connections. For the current research project, more than 200 parameters are logged which includes information about the administrative, weather, vessel's condition and other on board systems.

The events in the noon reports are further classified into ship operation states. The ship should be in one of the following operation state at any time:

1. In port
2. At sea
3. Drifting
4. At anchor
5. Manoeuvring

The data measured by the sensors is used to check the plausibility of the manually reported data by the crew. The sensors measure high frequency data and this data is averaged over the relevant time period and compared with the manually reported values. The difference between the manually recorded data and the sensor data is not significant. Furthermore, high frequency sensor data is also stored separately.

2.3.2 Loading Condition Data

Loading condition of the ship is recorded during the 'DEPA' event (Departure of ship from port). This data is acquired from the loading calculator of ship through a web interface. This data consists of following information:

Stability: It contains data about the stability parametres of the vessel such as rolling time period, rolling amplitude, angle of maximum righting lever (GZ), metacentric height (GM) etc

Hydrostatics: Hydrostatics includes information about center of gravity of vessel, drafts, trim, list angle etc.

Tanks: It contains data about the center of gravity of different tanks and the filling levels.

Container Cargo: It contains complete information about every container on board the vessel. For example: cargo type, container serial number, position, weight, length, width, height etc.

2.4 High Frequency Data

The sensor system installed in the vessel collects a large amount of data internally. The data is collected in short intervals and stored. Its sampling rate is as high as 5 Hz. It means that it can record upto 5 data points each second. A single parameter recorded over a period of one month has 12.9 million data points. This high frequency data is stored separately and also averarged over time to compare and validate it with low frequency data obtained by other means.

Furthermore, vessel's motion in 6 degrees of freedom (DoF) is measured using the installed the Inertial Measurement Unit (IMU) and the Global Navigation Satellite System (GNSS). The data is collected using ship's internal system and the additionally installed sensors at different frequencies with the help of a on board logging system. Internal data and data collected from additional sensors is discussed in the following sections.

2.4.1 Internal Data

Internal measurement system has been installed on the vessel and it collects more than 400 paramater values in total. These values are either directly measured by the sensors or

calculated from the combinition of other measured values. For example, engine power can be calculated from parameters like open water diagram, torque, propeller speed, propeller diameter, thrust deduction factor, wake fraction, ship speed and efficiency.

The available data contains following information:

- **Navigational Information:** Heading angle, rudder angle, true and relative wind speed, ship speed over ground and through water, distance covered, GPS coordinates, water depth etc
- **Tanks information:** Filling level and mass content for fuel tanks, ballast tanks and other tanks.
- **Engine information:** Power, torque, shaft speed, temperature, fuel consumption etc for main and auxialiary engines.
- **Cargo information:** Power system status for reefers etc.

For this thesis, data related to the vessel's propulsion performance is important. This includes engine power and all the parameters which have an effect the power demand. These paramters can be environmental factors like wind speed and direction, water depth, wave height etc. or navigational factors like speed over ground, speed through water, draft, trim etc.

2.4.2 Additional Sensor Data

The Octans V surface inertial measurement unit (IMU) has been installed in the engine control room of the vessel. It measures the roll, pitch and yaw rotations using three fiber optic gyroscopes (FOG). Furthermore, it can also measure the surge, sway and heave motions using Micro Electro-Mechanical System accelerometers. This Octans V surface is a highly accurate surface gyrocompas and motion sensor and it is certified for marine applications. Table 1 shows the measurement accuracy of this system for different parameters.

Table 1: Accuracy of Octans V surface inertial measurement unit (iXblue 2014)

Measured Parameter	Accuracy
Heading angle	0.05 deg
Roll / Pitch angle	0.01 deg RMS
Heave / Surge / Sway motion	5 cm or 5% (whichever is greater)

The GPS input is required to reach the accuracy of 0.05 degrees for heading angle. For this purpose, a dedicated Global Navigation Satellite System (GNSS) is installed in

the vessel's bridge and it is connected to the IMU unit. The GNSS system sends inputs of GPS cordinates and ship speed in time domain to the IMU system each second. It allows IMU system to reach its highest accuracy level. Figure 2 shows the additional sensors (GNSS and IMU systems) on board the vessel and their coupling.

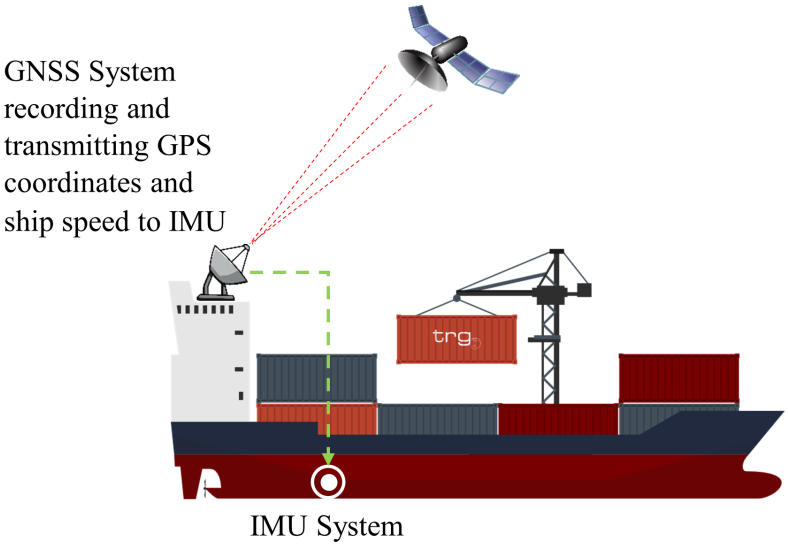


Figure 2: Additional sensors on board the vessel

Following data is provided by the IMU system:

- Roll, pitch and heading angle
- Surge sway and heave motion speeds
- Coordinated Universal Time (UTC) of each measurement
- Heading rate

2.5 Data Logging

Internal data and additional sensor data are stored separately on a ruggedized computer on board the ship. Data is stored at different acquisition rates depending on its level of importance. Important parameters that change rapidly have a higher sampling rate. For example, the ship's motion in waves has a higher rate of variability and it is recorded at a high sampling rate of 5 Hz. On the other hand, the ballast filling level of tanks changes less rapidly and it is acquired at a 15-minute interval. This 15-minute interval allows a sufficient detection of change in hydrostatics. However, the Inertial Measurement Unit (IMU) system logs data at a frequency of 50 Hz.

The data is stored in the database files present in two Solid State Drives (SSD) on board the ship in a special RAID configuration. This configuration is required to transfer

data from the ship to the shore. These SSD drives are taken out and replaced by new drives during the port calls. The data from the drives is integrated with the final database.

2.6 External Data Sources

Data provided by external sources is also used for analysis in addition to the data recorded by the on board sensors. This data includes the Automatic Identification System (AIS) information as well as the environmental conditions at the current location of the ship. The information about the environmental conditions is referred to as weather data.

2.6.1 Automatic Identification System (AIS) Data

AIS system is an automatic tracking system used to identify and track a vessel's position by communicating with AIS Base stations or nearby ships. The vessel's identity, position, speed, course, cargo information etc can be exchanged using the AIS system. As the data recorded by AIS is already reported and logged by on-board measuring systems so the data provided by AIS is redundant. In this project, the AIS data is recorded and updated every 15 minutes. The data recorded by this system can be categorized as follow:

- **Voyage information:** Destination, route plan, ETA, cargo type, ship's draft etc
- **Static information:** IMO number, length, beam, call sign, name etc
- **Dynamic information:** Course, speed, heading, navigational status, position-time stamp etc

2.6.2 Weather Data

The environmental conditions should be known to conduct a logical and meaningful vessel performance analysis. The effects of environmental conditions on the resistance and power requirement of the engine are discussed in the Section 3.3. Such conditions are recorded as weather data. The weather data is provided by a series of different weather stations. Then this data is transferred to the ship locations with a spatial resolution of 0.5 degrees latitude/longitude using mathematical algorithms. This data is recorded and updated every 120 minutes. It includes; wind, water, current, air and sea state information.

2.7 Data Model

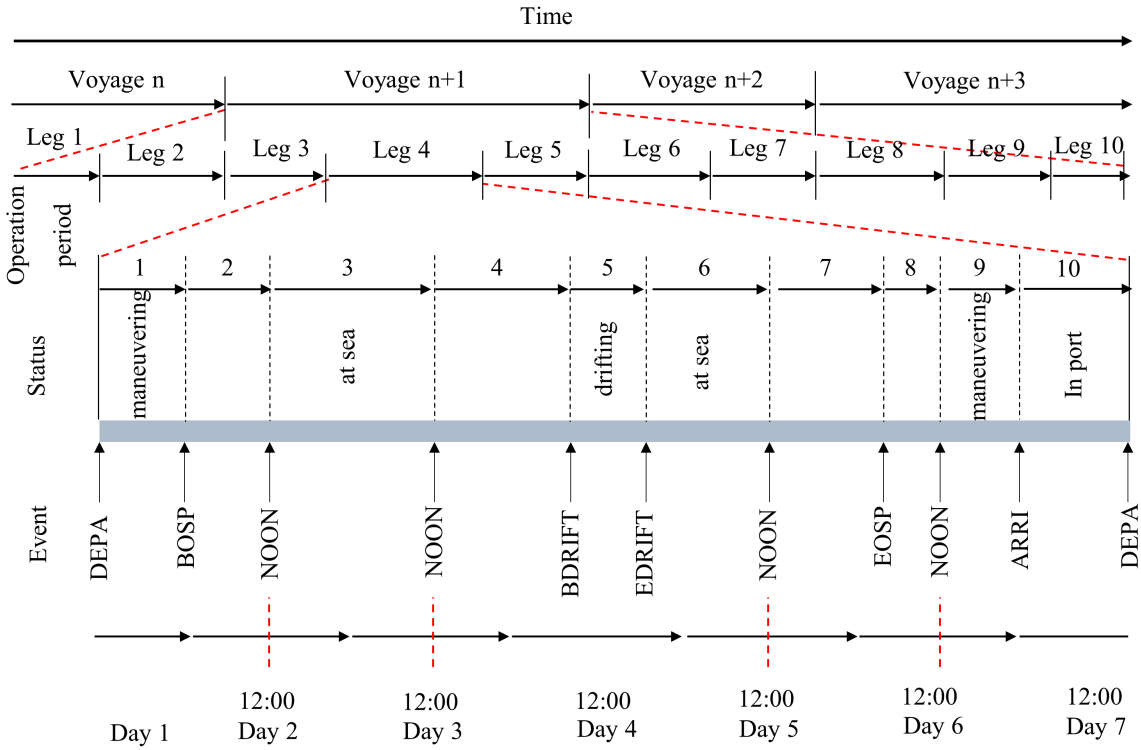


Figure 3: Data structure (Modified from Wisam et al. (2022))

The recorded data is stored in the final database in a time-dependent structure. The journey timeline is categorized into individual sections which relate to each other. Figure 3 shows the complete schematic of stored data structure and different terminologies related to it.

The top-level of the structured data is categorized as voyages. Voyages are divided into legs. Each leg contains data about the journey of the vessel from one port to another port. For example, leg 0 is the journey of the vessel from Antwerp (Belgium) to Felixstowe (England) for the data investigated in the current thesis. The journey from one port to another port is further divided into operational periods. Each operational period has an individual operation status (i.e. manoeuvring, at anchor, drifting, in port, at sea). As shown in the schematic, leg 4 has 10 operation periods. An operation period starts and ends with an event (i.e. DEPA, BOSP, BDRIFT, EDRIFT etc) which are already discussed in Section 2.3.1. Events provide additional information about the vessel's journey on the time axis. It can be observed in Figure 3, leg 4 starts after the ship departs from the port. It starts maneuvring and then begins the sea passage. It spends 5 days at the sea and arrives at the next port on day 7. When the ship departs from the port new leg 5 starts and this sequence is continued.

The time duration of a voyage is defined by the schedule of the vessel and it is a common practice for a transcontinental container vessel. However, the duration of a leg is defined by the time taken by the ship to travel between two ports of call. It depends on the travelling distance and the ship's speed. Furthermore, it also depends on the actual weather conditions and the modifications in the route plan due to environmental or operational reasons. (Wisam et al. 2022)

In this masterthesis, the data used for analysis is accessed from the University of Rostock online server using the Pycharm Professional module of Python programming language. The relevant data is accessed and stored in the form of Microsoft Excel file format. Then data is imported to Python and different built-in Python libraries such as Pandas, Numpy, Scikit, Matplotlib, Seaborn, Pyplot etc are used to analyse it.

3 DATA FILTERING

In this chapter, various data filtering techniques are discussed. Some of the data filtering methods are directly taken from the ISO 19030 standard and some modifications are made according to the available operational data. Furthermore, this chapter also discusses statistical data filtering methods for the removal of outliers which are widely used in Data Science.

3.1 Outlier Visualization

There are several definitions of data outliers. However an outlier is a data point which lies far outside the data cluster norm (Jarrell, 1994). The presence of outliers in a data set can cause deleterious effects on the data analysis. The outliers can be caused by multiple mechanisms. Anscombe (1960) categorizes the main cause of their occurrence into two parts; Outliers from data errors and outliers from intentional misreporting. However, in this thesis we have data directly logged by the sensors so we are dealing with the first type of outliers as categorized by Anscombe (1960).

Before examining the outliers in the available data set, a few methods to physically observe outliers are discussed on the following pages.

Scatter plot is one of the simple method to observe outliers. The methodology is to scatter plot the data and observe data points lying outside the main cluster of data. (Kannan and Manoj 2015). Figure 4 shows a scatter plot of main engine power with the shaft speed. In this figure, we can observe some points marked with red boxes which lie outside the cluster of data points. These points can be classified as outliers. Although, a scatter plot is a good practical approach to visually observe the outliers but does not provide a sound mathematical foundation to point out and remove them.

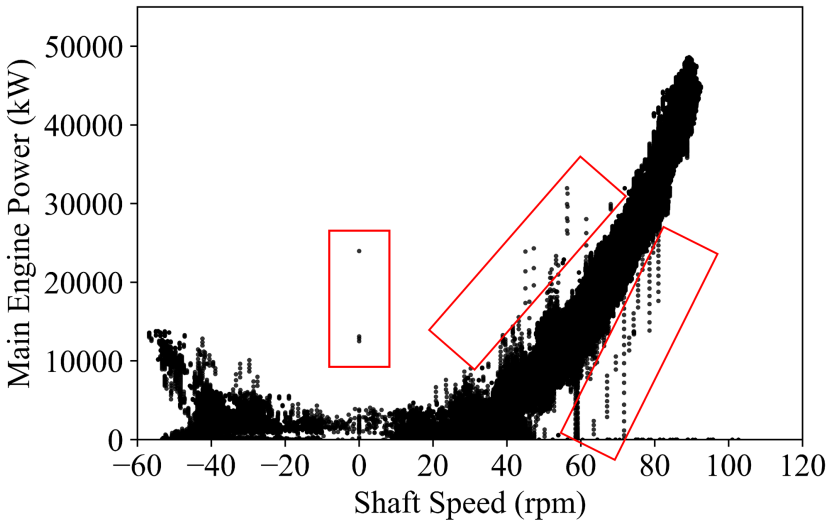


Figure 4: Main Engine Power vs Shaft Speed Scatter Plot

Distribution of data can also be observed using the plotting the distribution plots such as normal distribution plot, normalized density plot, Kernel density estimation plot etc. Kernel density estimation (KDE) estimates the probability density function by smoothing out the fundamental data problem (Latecki, Lazarevic, and Pokrajac 2007). It is a smooth histogram which estimates the probability of data distribution. It helps to visualize the data distribution. Any data point lying outside the estimated kernel density function can be classified as an outlier.

Another method of visualizing outliers is by using the box whisker plot. A box whisker plot represents the physical distribution of data through minima, maxima and interquartile ranges (Williamson, Parker, and Kendrick 1989). Description of a box plot is given in the Figure 5. Minima and maxima of the data are represented by the whiskers of the box plot. A data point lying outside these whiskers is an outlier. (ibid.)

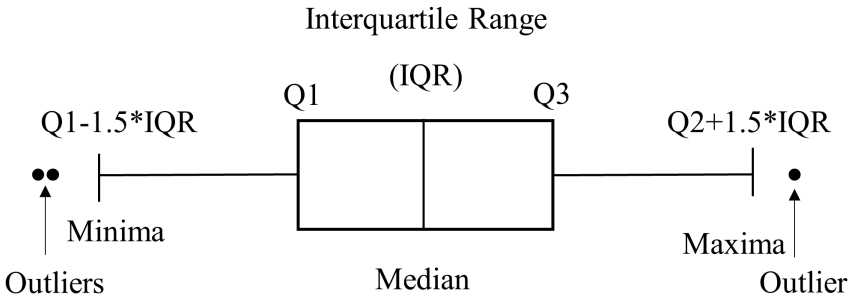


Figure 5: Box Plot Description

Removal of these outliers are absolute necessity for effective and correct data analysis.

A data without outliers ensures a controlled data analysis (Kwak and J. H. Kim 2017).

3.2 Outlier Removal Methods

There are a number of outlier removal methods which can be implemented on the a available data set. ISO 19030 suggests following two methods for statistical removal of outliers:

1. Chauvenet's Criterion
2. Validation

According to ISO 19030, data set should be divided into non overlapping blocks of 10 minute before detecting and removing the outliers. The number of values in each block depend on the frequency of available data set. Initially the parameters are measured at 5 Hz frequency but then they are averaged over one second. So, the available data set has 1 Hz frequency (one value every second). The number of data points in a 10 min data block will be 600.

3.2.1 Chauvenet's Criterion

It is a statistical filtering technique which judges a data based on the probability of its occurrence in a block of data. First of all complete data set is divided into a number of smaller 10 minute data blocks. Then mean and standard deviation of each block is calculated.

For the data which is not measured in angles the computation of mean is done using equation 1.

$$Mean = \mu = \frac{1}{N}\sum_{i}^{N} d_i \tag{1}$$

Where,

N = Number of data points in a block

d = Value of a data point

Computation of mean is different for data measured in angles. To calculate the mean for such a case, equation 2 is used.

$$Mean = \mu = atan2\left(\frac{1}{N}\sum_{i=1}^{N} sin(d_i), \frac{1}{N}\sum_{i=1}^{N} cos(d_i)\right) \tag{2}$$

Where,

$atan2$ = two argument arctangent function

Then difference between the mean of a data block and all of its individual values are calculated. If the data is not in angles, equation 3 is used. However, if the data is measured in angles equation (4) is applied.

$$\Delta_i = |(d_i - \mu)| \tag{3}$$

$$\Delta_i = \begin{cases} 360\,\text{deg} - r_i & r_i = mod(|(d_i - \mu)|, 360) > 180deg \\ r_i & otherwise \end{cases} \tag{4}$$

Standard error of the mean for each data point in a data block is computed using the equation 5.

$$\sigma = \sqrt{\frac{1}{N}\sum_{i}^{N} \Delta_i^2} \tag{5}$$

Complementary error function (erfc) is then used to calculate the probability of occurrence (P(d_i)) of each data point using equation 6.

$$P(d_i) = erfc\left(\frac{\Delta_i}{\sigma.\sqrt{2}}\right) \tag{6}$$

A data point is considered as an outlier if the following condition is satisfied.

$$P(d_i).N < 0.5 \tag{7}$$

If a data point in a block of data happens to be an outlier then whole block of data is marked invalid. If the data for one parameter does not satisfy the Chauvenet's criterion, the whole data point is marked invalid. According to ISO 19030, data for every parameter should be filtered using Chauvenet's criterion. There are 400 different parameters in the available data set. Application of this criteria on every parameter is computationally expensive. So, only those parameters are considered which effect the engine power. In this study, outliers from the engine power are removed using Chauvenet's criteria. Validation is used for filtering other parameters. Figure 6 shows the graphical representation of Chauvenet's Criteria.

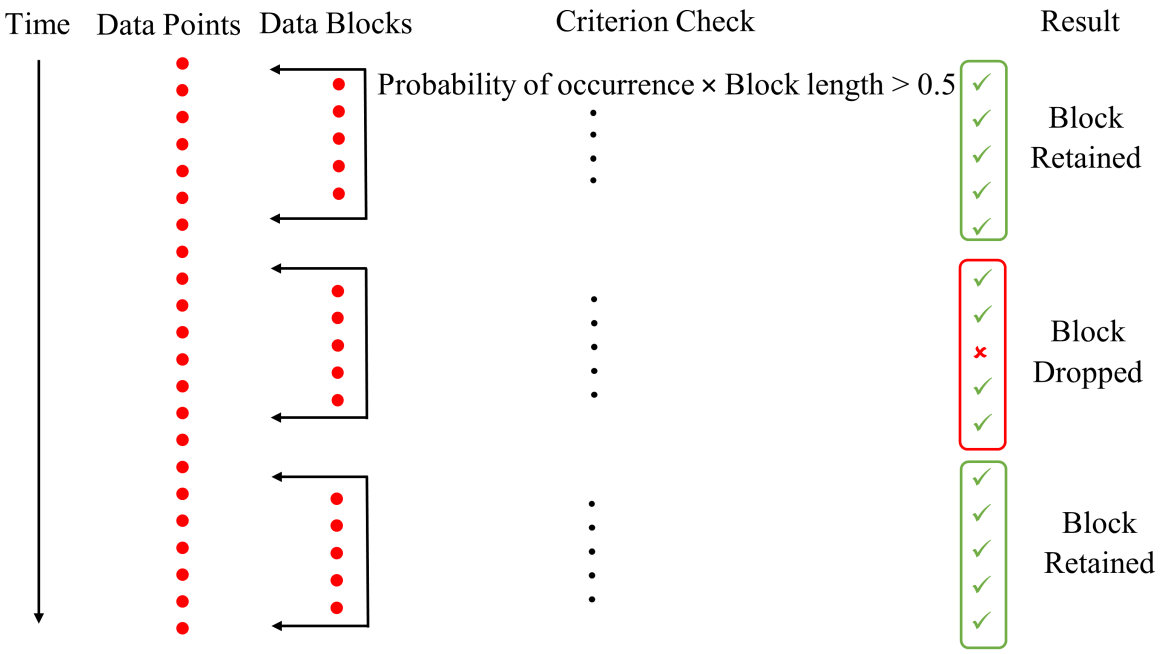


Figure 6: Graphical Representation of Chauvenet's Criteria

3.2.2 Validation

The data set is divided into data blocks with N data values. Then Mean μ and standard error of the mean Δ for each data block is computed for shaft speed, speed through water, speed over ground and rudder angle using equations mentioned in section 3.2.1. The standard error of the mean for each data block is checked against the following criteria: (ISO 19030-2:2016)

- Shaft Speed (rpm) < 3 min^{-1}
- Speed through water < 0.5 knots
- Speed over ground < 0.5 knots
- Rudder angle $< 1°$

If the standard error of the mean for a data block becomes larger than the criteria mentioned above, the whole data block is marked as invalid. Figure 7 shows the graphical representation of validation procedure applied on high frequency data.

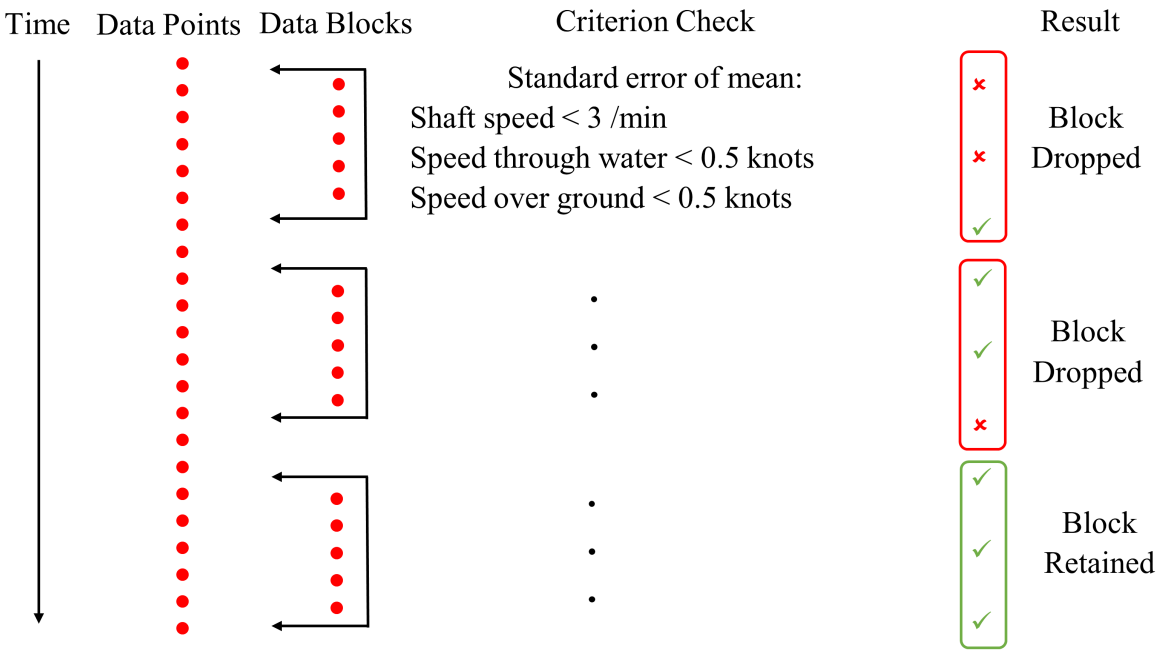


Figure 7: Graphical Representation of Validation

There are some additional rules of data validation process for twin screw vessel but we are dealing with a single screw vessel. Furthermore, rudder angle is not available in the data set so validation process for rudder angle entries is not taken into account in further analysis. Moreover, ISO 19030 also suggest the analysis for sensor drifting but no methods to analyze this phenomenon are given in the ISO rules.

3.2.3 Interquartile Range filter

It is one of the most used filter in data science is Interquartile range filter. It can filter out outliers from big data with less computational time as compared to the filtering methods discussed in section 3.2.1 and 3.2.2. (Vinutha, Poornima, and Sagar 2018)

First interquartile range (IQR), minimum and maximum values for the available data set should be calculated using equation 8, 9 and 10. (Wan et al. 2014)

$$IQR = Q_3 - Q_1 \tag{8}$$

$$Min = Q_1 - 1.5 \times IQR \tag{9}$$

$$Max = Q_3 + 1.5 \times IQR \tag{10}$$

Where,

Q_1 = First Quartile

Q_3 = Third Quartile

According to Wan et al. (2014), first quartile (Q_1) places the cut on the data at the 25% and divides it into two parts. One part represents 25% and second part represents 75% of the data. Similarly, third quartile (Q_3) cuts the data at 75% and divides it into two parts. In between the Q_1 and Q_3 is the interquartile range which composes of 50% of the data. To find first and third quartile the data is first arranged in the ascending order. Equation 11 and 12 are used to calculate Q_1 and Q_3.

$$Q_1 = \left(\frac{n+1}{4}\right)^{th} \text{Term} \tag{11}$$

$$Q_1 = 3 \times \left(\frac{n+1}{4}\right)^{th} \text{Term} \tag{12}$$

Where,

n = Total number of term

Figure 8 shows the physical representation of quartiles with normal distribution curve. The data points lying outside the minimum and maximum data range are termed as outliers . These points are removed to filter the data from outliers. One of the advantage of such a method is its fast computational times because this method is usually applied over the whole data set without splitting it into blocks. However, this method can only be applied on the data which is distributed normally (Vinutha, Poornima, and Sagar 2018). It can be observed in Figure 20 that the some operational parameters in the available data set are not distributed normally. Especially, mean draft is distributed in patches representing different loading conditions onboard the vessel. So, the application of this method will not be useful in this case.

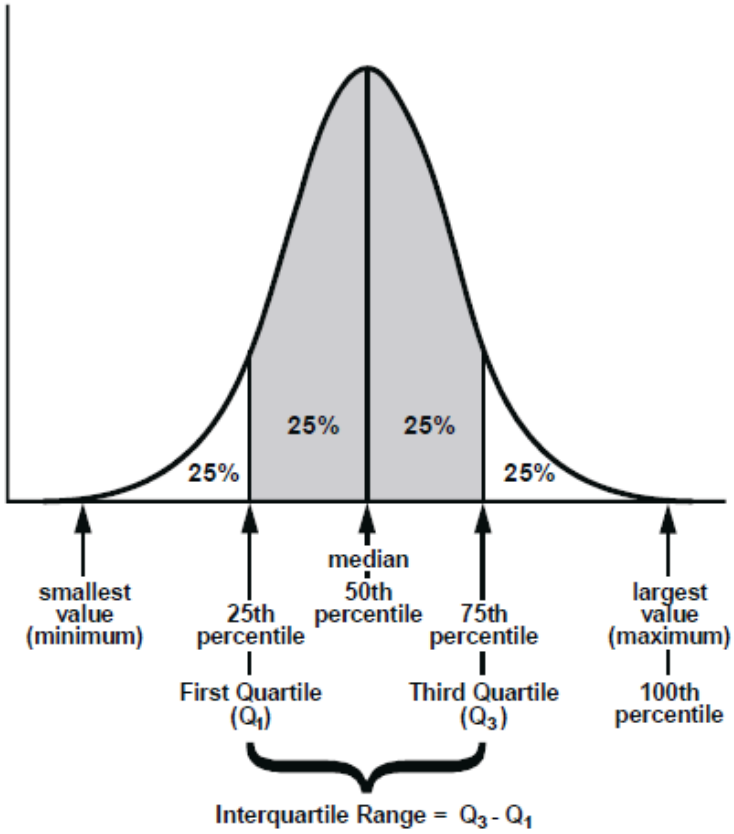


Figure 8: Representations of quartiles with normal distribution curve (Wan et al. 2014)

3.3 Filter for Environmental Conditions

Data set is filtered for environmental conditions after the removal of outliers. This type of filter can be set according to the requirement of data analysis being carried out. However, most of the time operational data has to be compared with the design data to analyze the ships performance during trail conditions and operational conditions. So, only those data points are taken out of the data set which mimic the calm water environmental conditions as they were during the trail test phase.

According to ISO 19030, following criterion should be satisfied to mimic the calm water environmental conditions:

Water Temperature The Water temperature should be higher than 2 °C. It ensures that ship is not treading in the icy waters and hull is not facing any additional resistance due to ice.

The temperature of water has direct effect on its density and viscosity. The variation in density of sea water for a temperature change from 2 °C to 15 °C is only 0.2% (Fofonoff 1985). The change in density does not increase the hull resistance significantly. However, the change in the viscosity for temperature change from 0.1 °C to 15°C is 35%. This

can change the frictional resistance of a tanker up to 4% (Hasselaar 2011). For a container vessel, the overall resistance increase can change up to 2.5% for this variation of temperature (Larsson 2010). However, this change in density and viscosity are not taken into account in ISO 19030 rules but one should filter out large temperature variation for accurate analysis.

Wind Speed As given in ISO 19030, the true wind speed should be between 0 m/s (BF0) and 7.9 m/s (BF4). According to Luo, Ma, and Hirakawa (2016), main engine power data will have irregularities at higher wind speeds. The wind speed and waves generated due to wind are interconnected with each other. There is a major contribution of sea state towards the total resistance of the container ship. Layout of the containers on the vessel also plays an important role in the additional wind resistance. The contribution of weather condition in total resistance of container vessel is 16.7% at the design speed (ibid.). So, the weather condition cannot be neglected while analysing the ship performance.

The delivered power should be within the range of the power values that are covered in the design power curves for a good comparison.

Water Depth The Water depth should be greater larger value obtained from the equation 13 and equation 14.

$$h = 3\sqrt{B.T_M} \tag{13}$$

$$h = 2.75\frac{V_s^2}{g} \tag{14}$$

where,
h = Water depth
B = Ship beam
T_M = Mean draft
V_s = Ship speed
g = Gravitational acceleration = 9.80665 m/s^2

Water depth filter ensures that there is no shallow water effect. Shallow water effect can increase the main engine power requirements Rotteveel and Hekkenberg 2015. While filtering for water depth, height of the water depth sensor mounted on the ship should also be taken into account for accurate analysis. It is mentioned in ISO 19030 that if the water depth values are out of the range of the measurement sensor this condition can not be applied. This can also be visualized in the available data set. When the water depth values reach a certain point, data entries in the water depth column go missing. This is due to the limitation of sensor to record depth values up to a certain water depth. This phenomenon can be seen in Figure 9. The keel clearance increases up to 250m and then sensor stops recording values and when keel clearance comes back to 150m range, sensor

starts to log values. We cannot say that 250m or 150m is the limit of the sensor. Sensor limit can be higher than that but due to physical features of sea floor at that location.

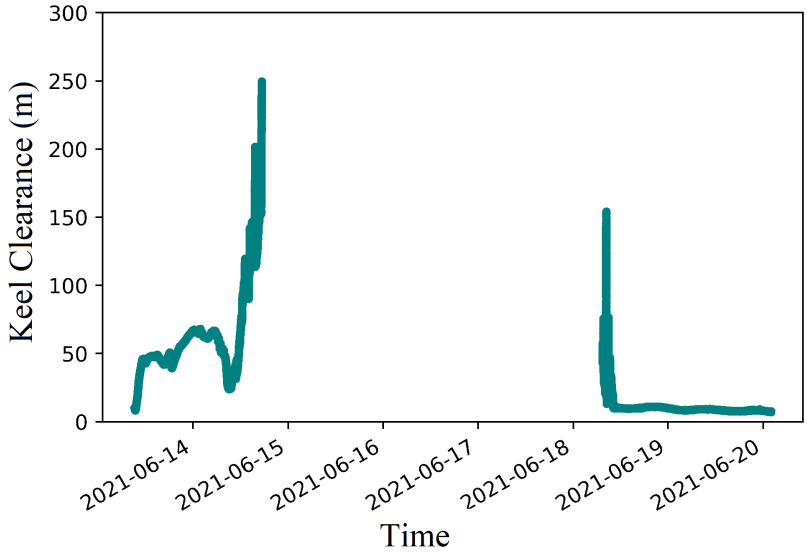


Figure 9: Variation of Keel Clearance with Time

Apart from the environmental condition filters mentioned in the ISO rules, further filters can also be applied on the available data set depending on the data analysis purpose. For example, following parameters can be limited to some specific values.

- **Wind direction**: It can be filtered on the type of data analysis. If upwind operating conditions and engine power curves in the upwind conditions have to be investigated, data can be filtered according to the required wind angle. Similarly, downward wind conditions and side wind conditions can be filtered according to the data analysis requirement.

- **Vessel Speed**: Data can be filtered for different speed ranges to study the behaviour of vessel. For example, turning rate of vessel at different speed ranges can be studied with the variation of rudder angle.

- **Heading Angle**: Heading angle can also be filtered for some specific angles to study the behaviour of vessel during specific portion of the journey. This type of filter is mostly used for vessels which travel more or less in a straight line between the same two ports. One way of the return journey can be filtered out using the heading angle. For example, a database of a passenger vessel which operates between Rostock and Gedser can be filtered to only those journeys for which the vessel travels from Rostock to Gedser or vice versa using heading angle,

- **Location based on GPS coordinates**: To study the behaviour of vessel in some specific oceans, data can be filtered according to the GPS coordinate readings. For

example, if the behaviour of the vessel has to be investigated in the South China sea, data can be filtered for GPS coordinates of South China sea.

The available data set is filtered for wind speed and water depth. Vessel speed filter is also used in this study. Water temperature data is not available in the available data set. . So, this filter is not applied. Moreover, wind direction, heading angle and location based filters are not applied because the current data analysis does not require such filters to be applied.

3.4 Filter for 'Zero' and 'NAN' values

The 'Zero' and 'NAN' (not a number) data points should be removed from the data set before performing the data analytics. Such data points are sometimes filtered out by outlier filtering methods discussed later. However, these data points should be removed separately by applying necessary procedures. In the present case, these values can be removed easily using 'Pandas' library in Python. These 'NAN' values occur when the sensors are not logging data into the data base. Similarly, sensors can also return a series of 'Zero' values due to malfunction. Presence of these logically incorrect values change the distribution of the whole data (Kwak and J. H. Kim 2017).

3.5 Methodology

In this section, the applied methodology to filter out the raw data is discussed. At first, Chauvenet's critieria is applied on the available data set. Then, validation procedures are applied as discussed in Section 3.2.2. After applying validation, data is filtered out for various environmental conditions. At last, zero and not a number values are removed. The complete methodology is presented in Figure 10.

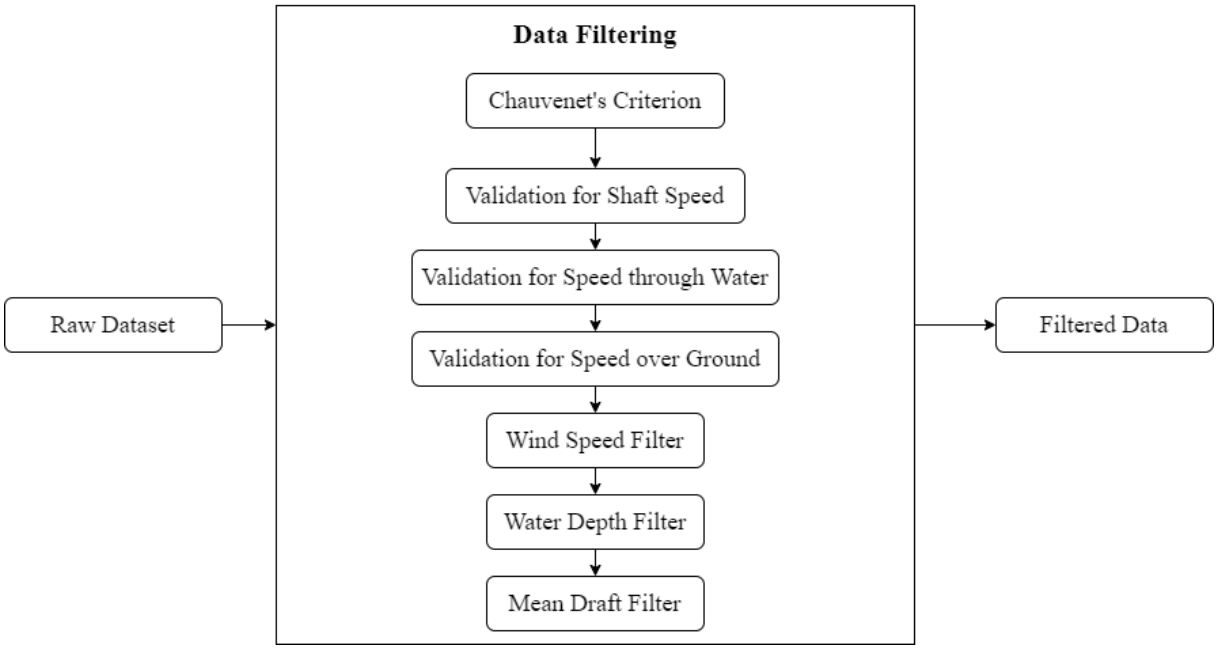


Figure 10: Applied filtering methodology

3.6 Results and Discussion

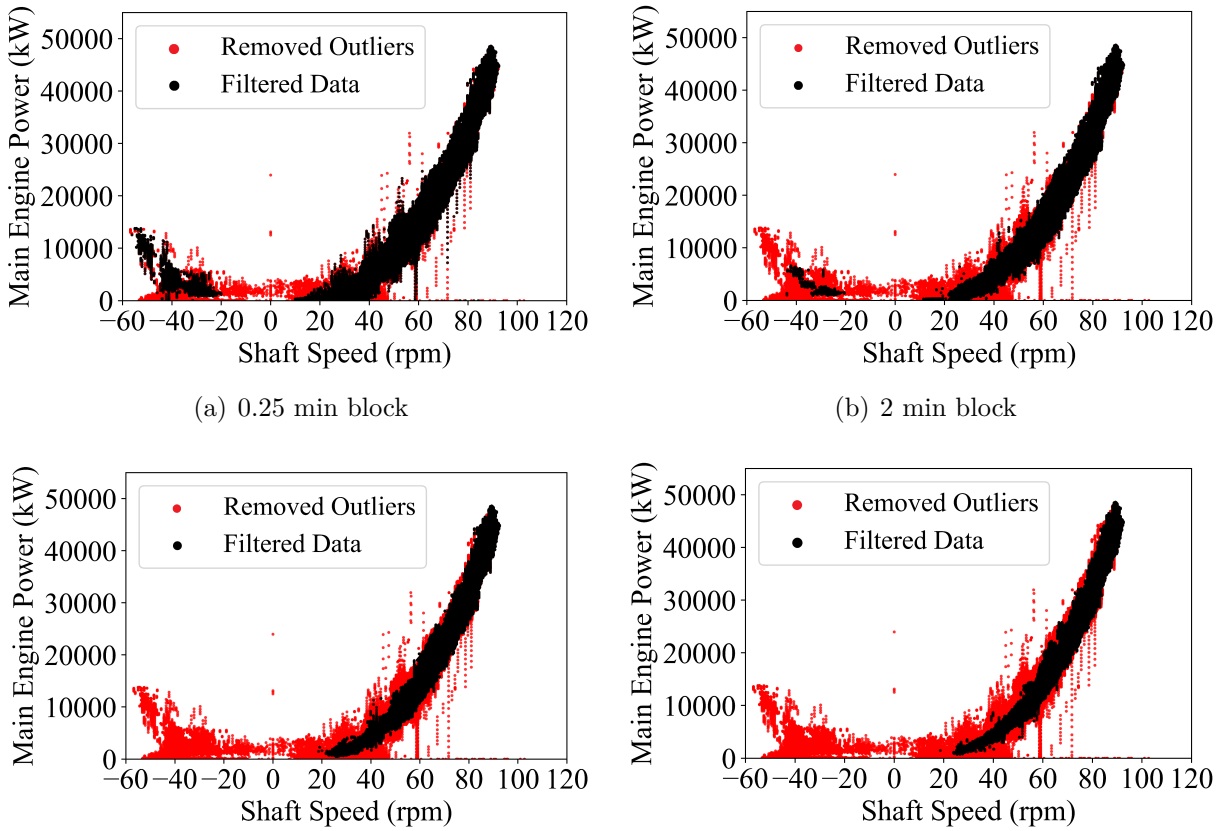
In this section, the results of applying the data filtering methods on the available data set will be presented and discussed.

3.6.1 Block Length

According to ISO 19030 rules, 10 minute blocks should be made and checked for the Chauvenet's criteria and validation procedure but we can change the length of the block to check which length is best suitable to filter the data without removing too many data points.

Figure 11 shows the scatter plot of main engine power with shaft speed before and after applying ISO 19030 filtering procedure with different block lengths. As the block length is increased, data scanning while filtering gets more rough and more data points are removed. To understand, why more points are removed we have to look at the schematic plot of the filtering criteria presented in Figure 6 and 7. As the length of the data block is increased the variation in readings also increases. Consequently, the mean and standard deviation of the data block increases which decreases the probability of occurrence of a single data point in a data block. Hence, more points are dropped at longer block lengths. Furthermore, when one data point in a block breaches the Chauvenet's criteria the whole block is eliminated. If the block length is bigger then a bigger chunk of data will be removed.

It can be observed in Figure 11, that as the data block length is increased, the data points at lower shaft speeds are completely eliminated. The data points at the negative shaft speeds are valid points. These points are recorded when the ship is reversed and shaft is rotated in the opposite direction. As the ship is reversed less frequently and the probability of occurrence of such data points are less in a bigger data block length as compared to smaller block length that is why these negative shaft speed data points are eliminated as the block length is increased. This phenomenon can be observed in Figure 11a, data points at the lower shaft speeds are not removed completely. On the other hand, at 5 min block almost all of the data points at lower shaft speeds are removed as shown in Figure 11c . The data density in the higher shaft speed filtered data cluster starts to get low at longer block lengths.



(a) 0.25 min block

(b) 2 min block

(c) 5 min block

(d) 10 min block

Figure 11: Comparison of ISO 19030 filtering procedure applied on different time blocks

The raw data set shows a vague trend between main engine power and shaft speed. The main purpose of the data filtering is to have a clear trend between two entities (Kwak and J. H. Kim 2017). As data block length is increased the trend becomes clear as shown in Figure 11. But, as the trend becomes clear by increasing the block length, more number of data points are dropped. Furthermore, computational time required to filter out raw data set is also a factor while selecting the data block length. So, a compromise should be

made between the percentage of removed data points, computational time and a filtered data set with clear trend.

The percentage of remaining data points after filtering the raw data and computational time for different block lengths are presented in Figure 12 and Table 2. The computational time will be different for central processing units (CPUs) with different specifications. In this case, computations are performed on a 2.7GHz (16 CPUs) computer with 32768 MB of RAM. Efficiency of code is also a factor in the running time of the written code. However, the computational time of same code on the single computer for different block lengths is a good way to compare them. It can be seen in Figure 12 that for smaller block lengths the computational time is high. It decays exponentially as the block length is increased. This trend was expected because as the blocks get smaller there are more means and standard deviations to be computed by the code. Hence, the computational time grows as block length decreases.

The percentage of remaining data points from filtering increases as the block length is increased until 2 minute block length. After that it shows a decreasing trend as shown in Figure 12. When data is filtered using the validation procedure, a term 'Standard Error' is computed which is calculated using equation 15.

$$\text{Standard Error} = \frac{\sigma}{\sqrt{n}} \tag{15}$$

Where,

σ = Standard deviation

n = Number of data points in a data block

When the length of data block is shorter, the standard deviation is small but n is also small. This causes the standard error value to get bigger than the validation criteria. Hence, more data points are marked as invalid for data block length less than 2 min. On the other, when the data block length is longer, standard deviation gets bigger. As a result, standard error value gets bigger. This causes more data blocks to violate the validation criteria and hence more data points are removed as outliers.

Table 2: Comparison of ISO 19030 outlier removal procedure applied on different time blocks

Block Length (min)	Entries Per Block	Computation Time (sec)	Remaining Data Points Number	%
0.25	15	2282	11055855	56.72
0.5	30	1451	11503650	59.01
1	60	992	12065295	61.90
2	120	782	12344460	63.33
5	300	586	12115583	62.15
7	420	550	11899495	61.04
10	600	566	11499627	58.99
20	1200	522	9938424	50.98
60	3600	458	6714771	34.45

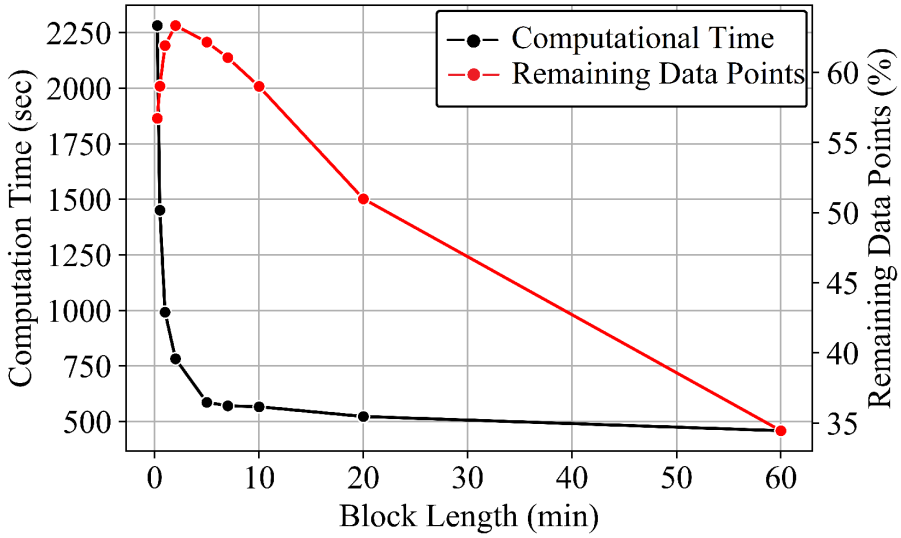


Figure 12: Computational time and remaining data points after filtering for different block lengths

A block length of 5 minute will be most suitable for the analysis of the available data set as it retains 62% of data points and computational time is also less in comparison with shorter block length filtering. Furthermore, the filtered data scatter of 5 min block length is less as compared to lower block lengths as shown in Figure 11.

Moreover, it should be noted that ISO 19030 recommends 10 minute block length but here in this case 5 minute block length is more suitable. According to ISO 19030, 1 data point for every 15 seconds is recorded for a 10 min block. Hence, there are 40 data points in each data block. But, in this case 1 data point for every 1 second is recorded. So, there are 300 data points for a 5 minute block. The data block is smaller in this case but the data density is greater than the recommended 10 min block. So, data analysis will be carried on by using a 5 minute data block.

3.6.2 Investigation of Chauvenet's Criteria with 5-min Block Length

First of all, Chauvenet's criteria has been applied on the main engine power. As engine power is analyzed against the shaft speed so the main focus of the filtering is to filter out data with respect to these two parameters. The complete procedure for this criteria has been already mentioned in Section 3.2.1. The length of block is chosen as 5 min as discussed in Section 3.6.1.

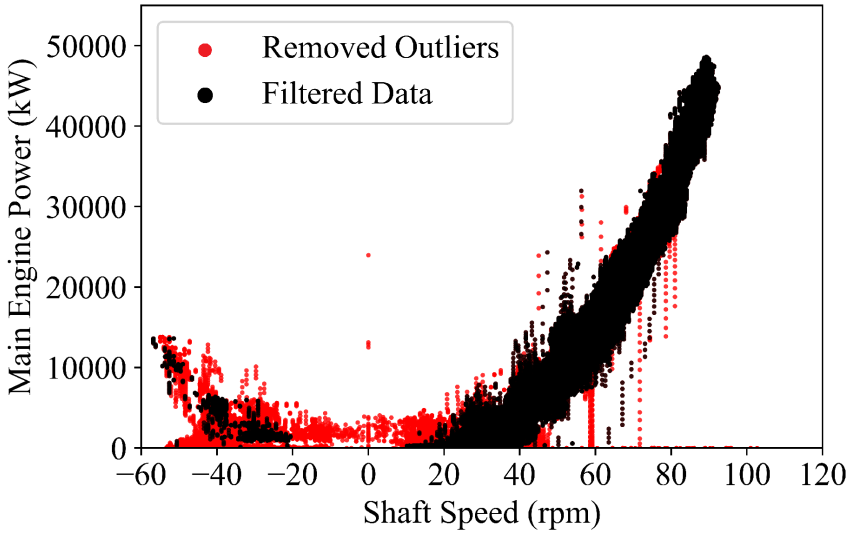


Figure 13: Application of Chauvenet's Criteria on Raw data set

Figure 13 shows the removed outliers after applying the Chauvenet's criteria on the raw data set. Most of the outliers that have been removed by this criteria lie around the 0 rpm shaft speed. Furthermore, the results show that most of the data points are filtered out from those areas where the data density or the probability of occurrence is low especially below 0 rpm shaft speed. The main engine power curve against the shaft speed becomes clear after applying this criteria but this data still requires more filtering as evident from the scatter plot.

3.6.3 Application of Validation

The validation method to remove outliers is implemented on the given data set in three steps. In the first step, validation for shaft speed is checked. After that, validation for ship speed through water is applied. At last, validation for ship speed over ground is checked.

Shaft Speed After applying the Chauvenet's criteria on raw data set, the data is further filtered using the validation procedure on shaft speed as dicussed in Section 3.2.2. Figure 14 shows the removed outliers after applying the validation of shaft speed procedure on the already filtered data obtained after the application of Chauvenet's criteria on the raw data set. Results show that most of the points in the lower side of the shaft speed are marked as outliers. Furthermore, some data points which were not located inside the data cloud in the shaft speed 10 rpm to 80 rpm are also removed. This filtered data will be further filtered using the remaining procedures in the validation method.

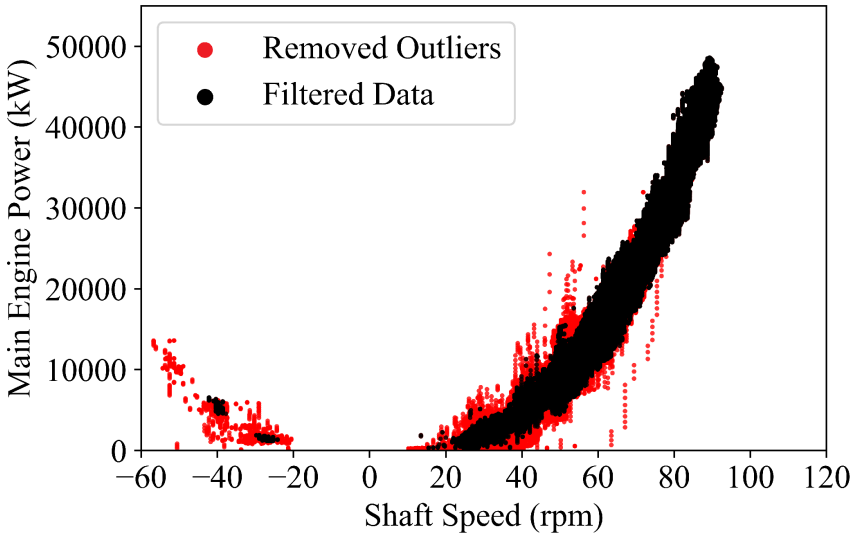


Figure 14: Application of Validation (shaft speed) on data filtered with Chauvenet's criteria

Speed Through Water The validation procedure is continued and data is further filtered using validation criteria on speed through water. Figure 15 shows the result obtained after applying the validation of speed through water on the already filtered data obtained after the validation of shaft speed. Results show that the data points located below the 0 rpm shaft speed are completely removed by this filter. Moreover, the thickness of the data cloud in the 30 rpm to 80 rpm is also trimmed down. The data is now starting to show clear trend.

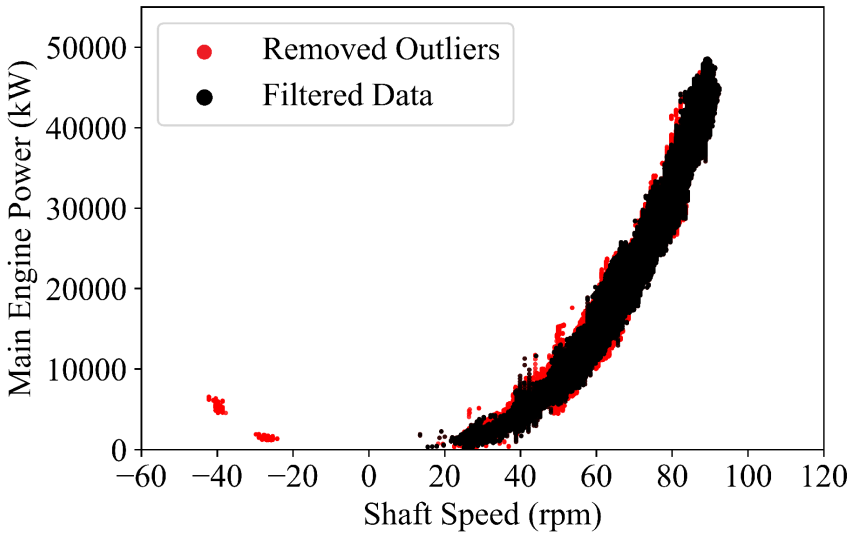


Figure 15: Application of Validation (speed through water) on data filtered with Validation (shaft speed)

Speed Over Ground Figure 16 shows the filtered data after the application of validation of speed over ground on the already filtered data obtained after the application of validation of speed through water. It should be noted that very few outliers are being marked here. Few data points at the lower shaft speed are removed. This is due to the fact that validation of speed through water and validation of speed over ground are very similar criteria. Both of them filter out the data blocks when the standard deviation of the block is more than 0.5 knots. This is the reason why very few points are removed after applying the validation of speed over ground criterion after the application of validation of speed through water criterion.

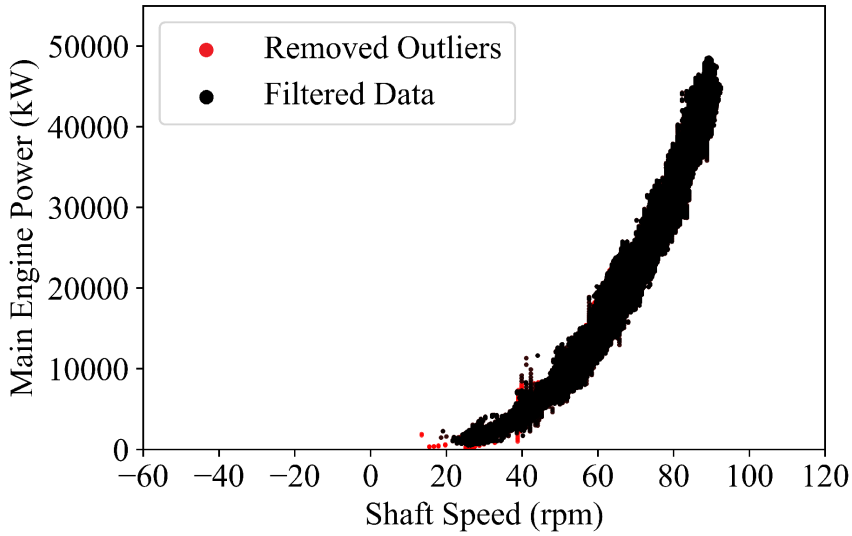


Figure 16: Application of Validation (speed over ground) on data filtered with Validation (speed through water)

Now the filtered data is showing a clear trend. We can use this data for our analysis but to in order to have a uniform trends among parameters and mimic the sea trail test conditions. This data further filtered for different environmental conditions as mentioned in Section 3.3.

3.6.4 Application of Environmental Filters

The environmental filters that are being applied on the data are wind speed and water depth filter. The water temperature filter is not being applied on the data set because there is no high frequency data available for the water temperature. Furthermore, the ship sailed in the waters where there is normally no ice during the whole time period of the available data set.

Water Depth Water depth filter is applied on the data obtained after Chauvenet's criterion and validation procedure. Its methodology is already discussed in Section 3.3. Figure 17, shows the removed data points after applying the water depth filter. It can be observed that most of the data points that are being removed are located on the upper side of the curve. This shows that at low water depths, shallow water effect comes into play and more engine power is required to propel the ship as suggested by Rotteveel and Hekkenberg (2015).

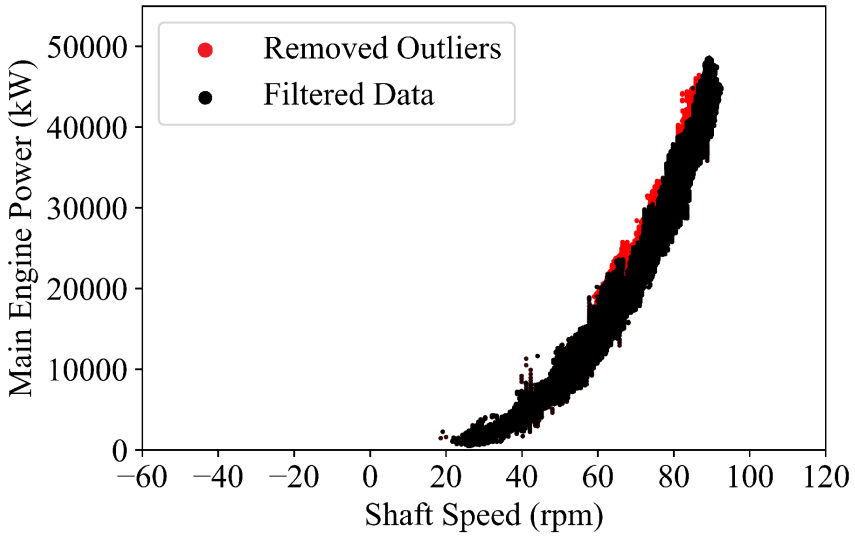


Figure 17: Application of water depth filter on data filtered with validation procedure

Wind Speed After the application of water depth filter, wind speed filter is applied. Figure 18 shows the result obtained after applying the wind speed filter is applied. It can be observed that data from both side of the curve is removed if we remove the data points when the wind speed is high. This is due to the fact that higher wind speed can either increase or decrease the engine power requirement depending on the wind direction. If the

ship is experiencing upwind condition, more engine power will be required. On the other hand, less engine power will be required if the ship is experiencing downwind conditions.

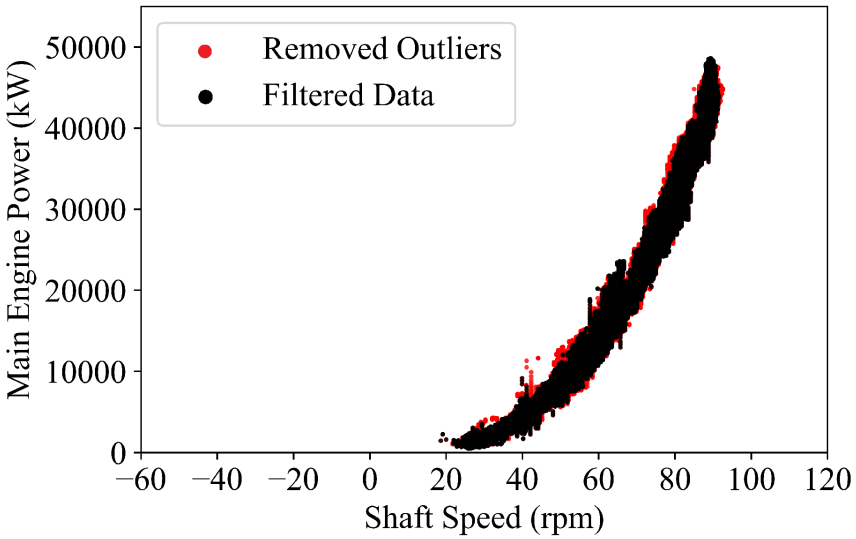


Figure 18: Application of wind speed filter on data filtered with water depth filter

3.6.5 Filtering for 'Zero' and 'NAN' values

The data has been filtered for zero and 'NAN' (not a number) values. The removal of 0 values is important in some cases. For example, physically the ship cannot have a 0 mean draft value. So, logically this filter is applied on mean draft values. On the contrary, ship can have '0' engine power, shaft speed, speed through water and speed over ground. Moreover, this filter should be applied on relevant operational parameters (i.e. the operational parameters which are directly involved in the data analysis). However, if these filters are applied on irrelevant filters which does not have an effect on the main engine power curve then filtering can become computationally expensive and useful data points can be eliminated.

Figure 19 shows the result of applying this filter on the data obtained after applying the environmental filter. Small percentage of data is removed from the data set as most of the zero and 'NAN' data points are already removed by the outlier removal methods.

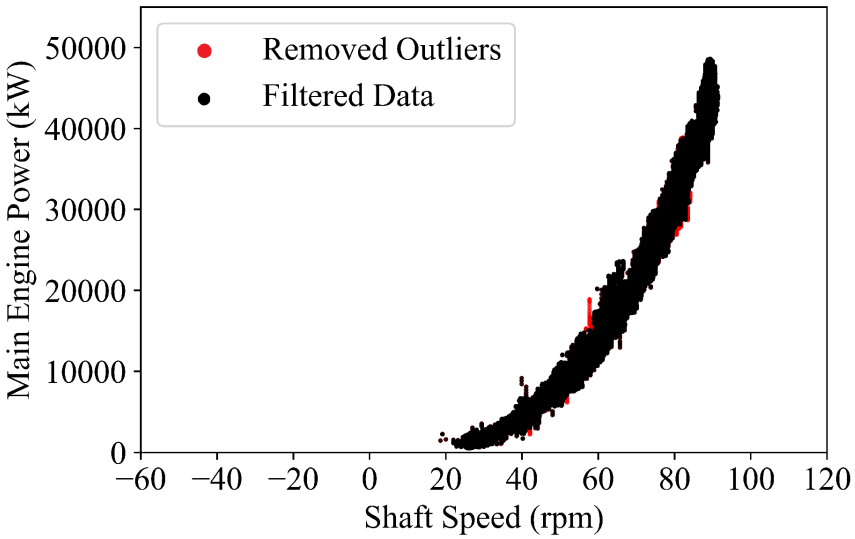


Figure 19: Filtering for 'Zero' and 'NAN' values

However, in the complete data set, almost 8.5% of the total data points have mean draft value as 0 as given in Table 5. This can also be observed in Figure 20, which shows the distribution plot of mean draft.

Table 3: Invalid data points percentage

Description	Count	Percentage of Total Data
Total Data	19493012	-
Mean Draft = 0m	1654477	8.487 %

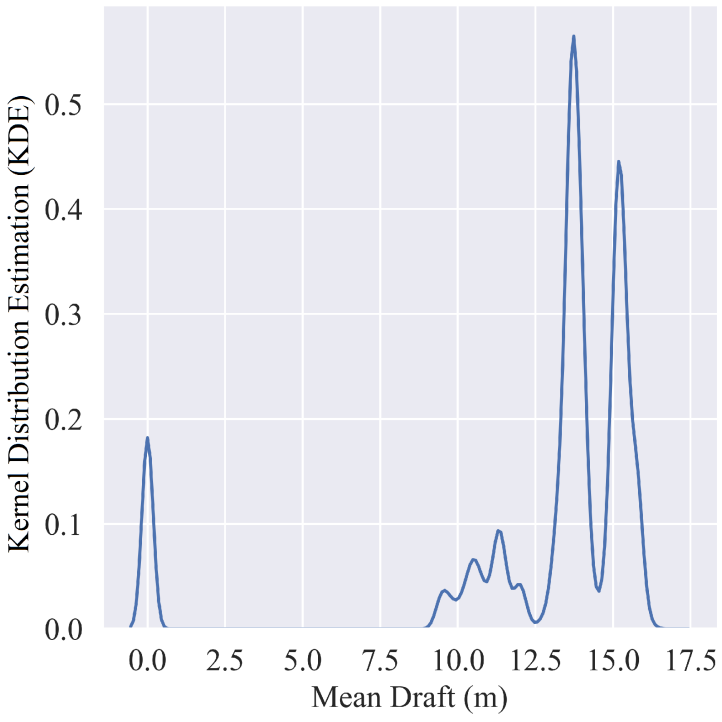


Figure 20: Distribution of Mean Draft in Unfiltered Data

Distribution of data points before and after removal of 0m mean draft value is tabulated in Table 4. The mean value of the data set is increased from 11.9m to 13.1m. Moreover, the minimum value is also changed from 0m to 8.17m. This change in the distribution of parameters can be very significant in the later stages of data analysis.

Table 4: Distribution of data points before and after removal of 0m mean draft values

Description	Before Removal	After Removal
Count	19493012	17834618
Mean (m)	11.9	13.1
Minimum (m)	0	8.17
25% (m)	11.23	11.6
50% (m)	13.41	13.56
75% (m)	14.04	14.31
Maximum (m)	16.9	16.9

3.7 Summary

After the analysis of different block lengths, 5 min block length has been chosen for the application of statistical filters on the high frequency raw data set. Chauvenet's criterion and validation procedures were applied for filtering outliers. These methods collectively eliminated approximately 39% of data. The majority of data points ($\sim$ 36%) were eliminated by Chauvenet's criterion.

Table 5: Data filtering applied to High Frequency Data set

Method	Remaining points		Difference	
	N	%	N	%
Raw dataset	19493012	-	-	-
Chauvenet's Criterion	12502936	64.14	6990076	35.86
Validation (shaft speed)	12304614	63.12	198322	1.59
Validation (speed through water)	12153389	62.35	151225	1.23
Validation (speed over ground)	12115583	62.15	37806	0.31
Water depth filter	11808285	60.58	307298	2.54
Wind speed filter	6130294	31.45	5677991	48.08
Mean draft filter	5619213	28.83	511081	8.34

After the application of statistical filters for the removal of outliers, environmental filters were applied to obtain the data recorded when ship was in calm waters and sea trail test conditions were mimicked. In environmental filters, the majority of data points (29% of the total raw data set) are eliminated by the application of wind speed filter. This depicts that one third of the time during its journey, the ship was experiencing wind speed $>$ 7.9 m/s. However, only 2.54% of data points were eliminated by water

depth filter. This shows that the ship was sailing mostly in the deep seas and it was not experiencing shallow water effect. Figure 21 shows the percentage of total data points removed by different filters applied on raw data set.

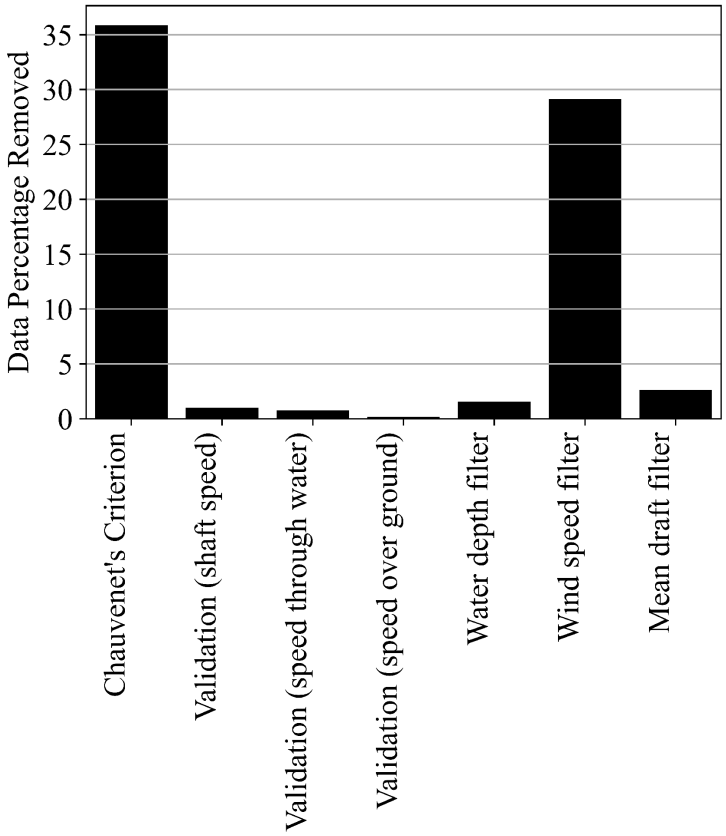


Figure 21: Percentage of data points removed by the applied filtering methods

Figure 22 shows the comparison of raw data set with the filtered data. Only 28.82 % of the total data points are left. A clear trend is observed in the graph between main engine power and engine shaft speed after filtering.

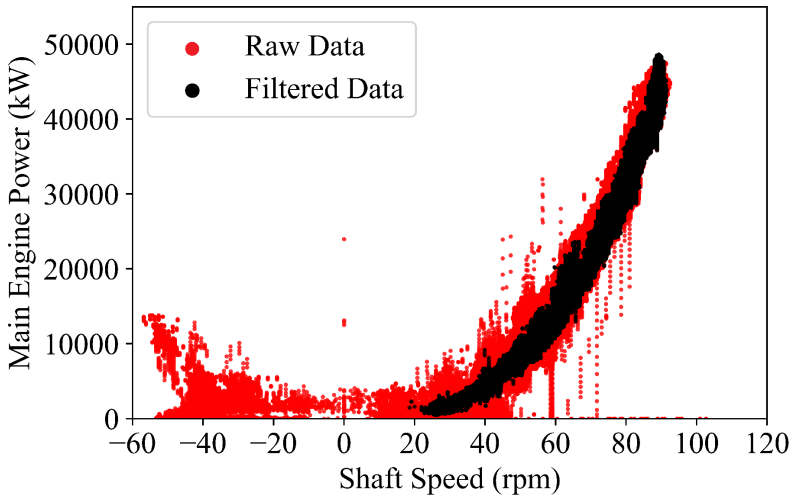


Figure 22: Filtered raw data set

4 DATA ANALYSIS

In this section, available ship data is analyzed to understand it's normal operational behaviours in the investigated time frame. This analysis is performed on the filtered data set.

4.1 Ship's Information

Figure 23: An Intercontinental 14000 TEU container vessel (FleetMon 2020)

An intercontinental 14000 TEU container vessel has been investigated in this study. It has a length of 361m with 51m beam. It has a gross tonnage of over 15000 tons. Table 6 contains the some of the general information of the investigated container vessel obtained through static Automatic Identification System (AIS).

Table 6: Ship's general information

Ship type	Container vessel
Gross tonnage	153115 tons
Carrying capacity	14000 TEU
Length	361 m
Beam	51 m
Flag	Liberia
Year of built	2011

4.2 Operating Route

The data set available from May 7^{th}, 2021 to January 18^{th}, 2022 shows that the vessel travelled in Europe, Africa and Asia. Figure 24 shows the travel route of the vessel. While

travelling from Europe to Asia, it's journey starts from North Sea. Then moving through the English channel it goes to Mediterranean Sea via Strait of Gibraltar. It crosses Suez anal and reaches Asian ports through Indian Ocean. Then it connects Asian ports of Malaysia, Singapore and China in the South China Sea. Most frequently visited port is Rotterdam (Netherlands), followed up by Felixstowe (England) and Tanjung Pelepas (Malaysia). It visited Rotterdam 5 times in 8 months.

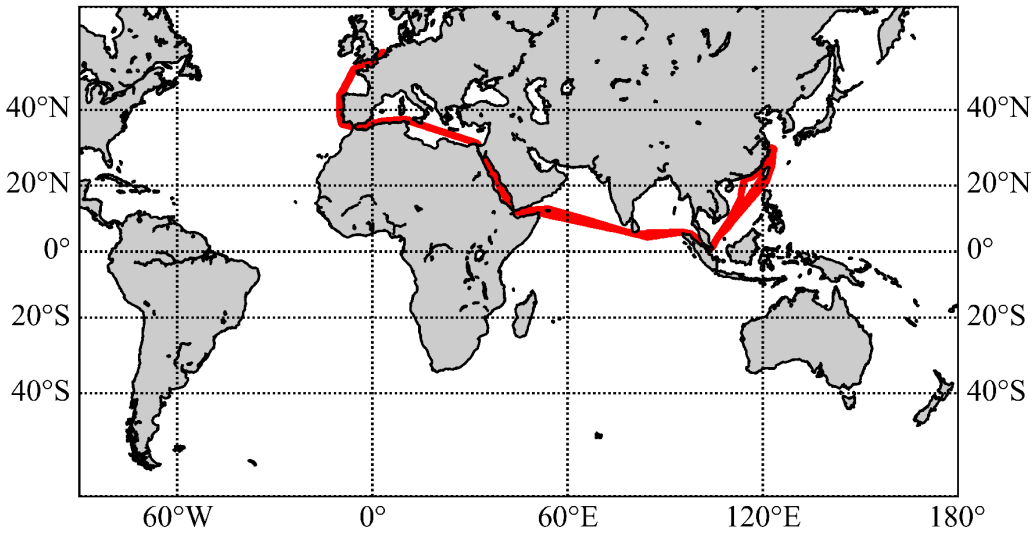


Figure 24: The vessel's travel route from May 2021 to Jan 2022

4.3 Cargo Carrying

The carrying capacity of the vessel has been defined as 14000 TEU but the maximum number of containers it carried during the 8 months are 12488. Some of the containers carried by the ship are empty. On average 24.6% of the total containers onboard the vessel are empty. This average rises to 48.8% when the ship is travelling from Europe to Asia and drops to 3.3% when the ship is travelling from Asia to Europe. This shows that most of the cargo is being exported to European market from Asian Market.

Figure 25 shows a typical filled container carriage behaviour between different ports. It can be observed in the Figure 25 that the ship starts to get loaded from Singapore, Yangshan, Ningbo and Tanjung Pelepas ports in Asia and then sails to Europe and gradually unloads the filled containers at Rotterdam, Felixstowe and Le Havre during this particular journey. Then it loads up to almost half of it's capacity at Rotterdam before going back to Asian market.

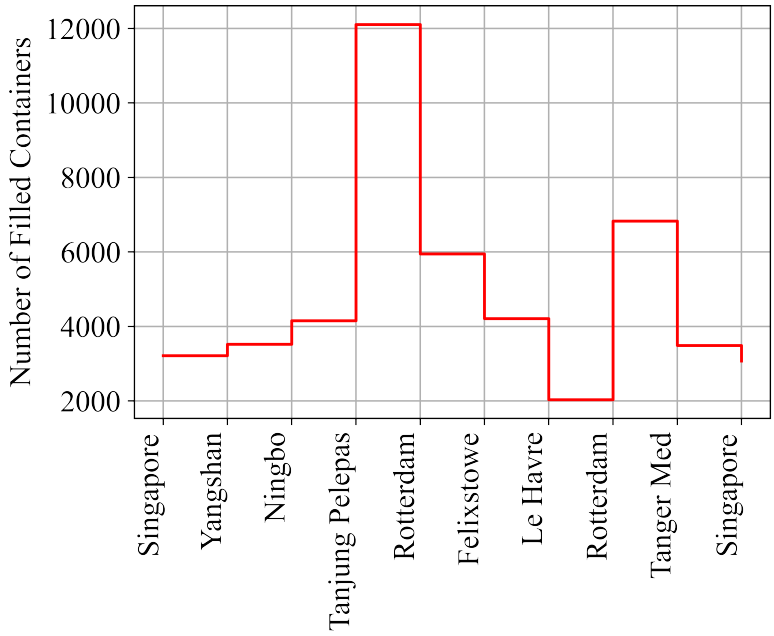


Figure 25: Filled containers carried for one complete journey between Asia and Europe

4.4 Sailing Speeds and Time

Ship has three operation status while it is navigating i.e. At Sea, Drifting and Manoeuvring. Sailing speeds for these operating status are different and are given in Table 7. Note that these speeds are taken from noon reports and not from high frequency data.

Table 7: Sailing speeds for different operating conditions

Operation Status	Speed (knots)		
	Min	Average	Max
Drifting	0	1.58	4.5
At Sea	9.25	16.12	21.42
Manoeuvring	0	7.88	15.18

The longest time period the ship has spent on sea during these 8 months is 22 days 19 hours while travelling from Rotterdam to Singapore. Table 8 shows the average sailing time between different ports. It takes the ship about 20 days to travel between Asia and Europe through Suez canal. These times does not include any stops during the journey (e.g. the time when ship is anchored and waiting for its turn to cross Suez canal). Ship crossed Suez canal 4 times during the 8 months and on average it waited 6 hours 50 minutes for its turn to cross Suez canal.

Table 8: Average sailing time between different ports

Journey	Average Sailing Time (days)
Rotterdam - Singapore	21.15
Tanjung Pelepas (Malaysia) - Rotterdam	20.7
Tanjung Pelepas - Le Havre (France)	19.28
Tanger Med (Morocco) - Singapore	16.7
Singapore - Yangshan (China)	5.4

4.5 Draft and Trim

The operating draft highly depends on the displacement of the ship which varies due to loading conditions. Figure 26 shows the distribution of operational mean draft values of the vessel.

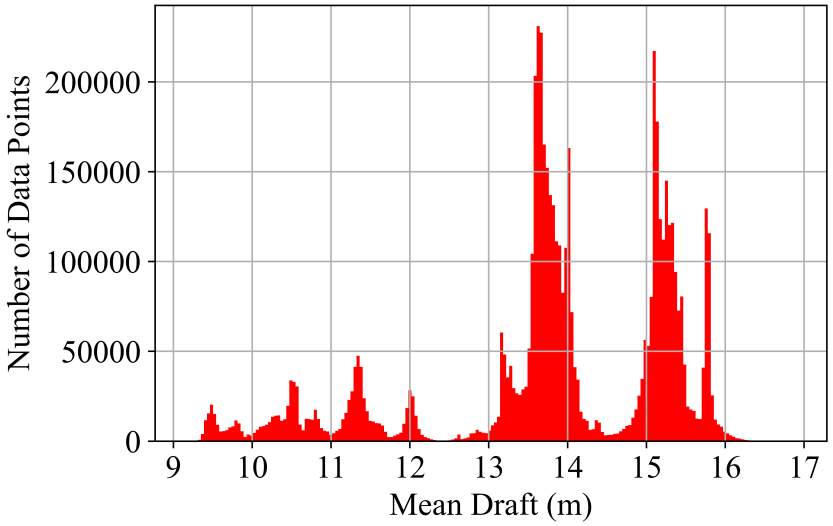


Figure 26: Operational mean draft data distribution in high frequency filtered data

It can be observed that mostly vessel operated around two operational drafts i.e. 13.7m and 15.2m. This mean draft values depend on the loading condition of the vessel. When the ship is loaded with filled containers, its displacement is increased. As a result, its mean draft value increases. On the other hand, when the ship is not loaded to its full capacity, it has lower displacement values. Consequently, it has smaller mean draft. It can also be observed in Figure 26 that ship has also sailed with mean draft values as low as 9.5m. Small peaks of mean draft values are seen from 9.5m to 12.5m. These peaks depict that the ship travelled with less cargo tonnage.

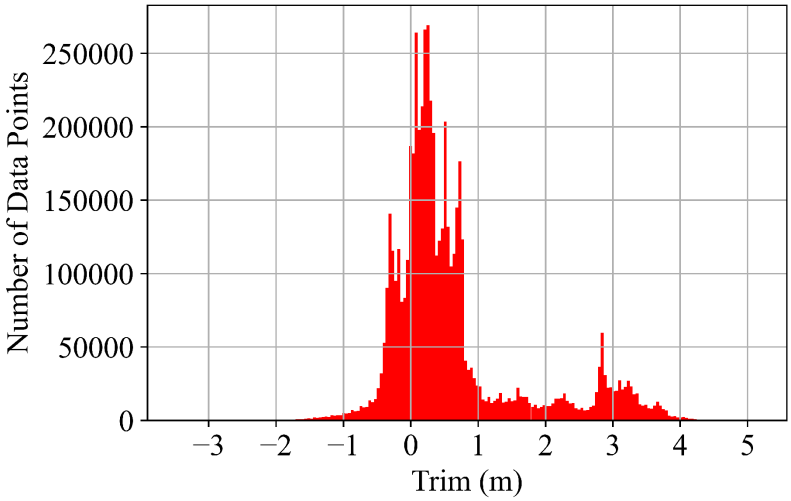


Figure 27: Trim data distribution in high frequency filtered data

The ship trim has a great impact on the energy efficiency (discussed in Section 4.6). According to the filtered high frequency data, the mean trim of the vessel during these 8 months is +0.61m. Positive trim means that ship is trimmed by the stern. It can be seen in the Figure 27, that most of the data set is concentrated around even keel condition. It can also be observed that some data points lie around +3m trim. One of the possibility for these data points to exist is bad sensors values due to rough seas. Trim is calculated from the fore and aft draft values. These draft values are measured by the sensor mounted on the ship. The sensor measures the distance from itself and the free water surface to calculate the ship draft at that location. If the sea condition is rough, water surface can move up and down relative to the sensor and wrong values will be recorded. However, higher trim values are not eliminated after filtering out the data for rough weather conditions.

So, it can be concluded that these trim values are not due to bad weather. Furthermore, higher trim values are validated by the authorised crew of the ship in the noon reports. It is found that when the ship had lower displacement and draft values, it had higher trim. In other words, when ship was sailing with less number of filled containers, it had trim values around +3.5m. This phenomenon can be observed in the Figure 28.

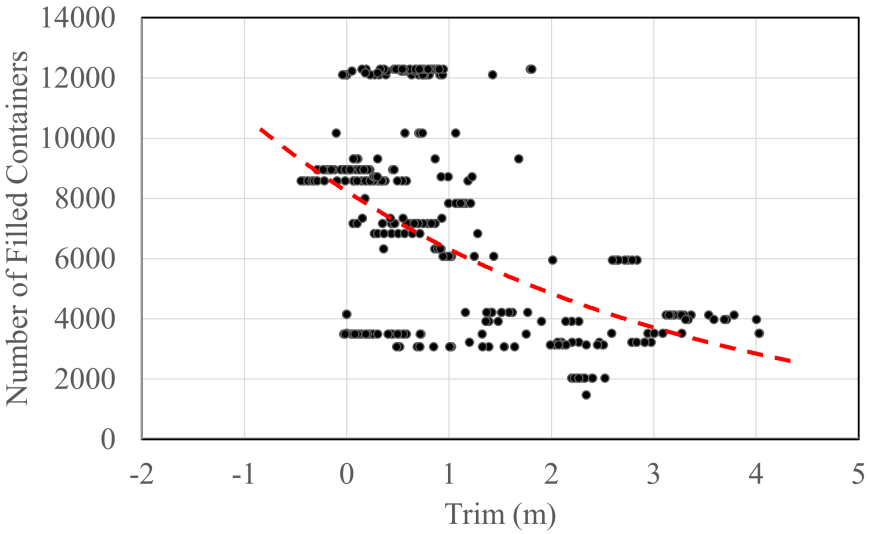


Figure 28: Relation of Trim (m) with the number of filled containers on board ship

The distribution of containers on board the ship can change the trim values significantly. In this case, the ship has higher trim values when it has a low number of filled containers and it is travelling between two ports (usually in Asia) with a small travelling distance between. The reason could be that at lower tonnage mean draft is less. But, the propeller has to stay submerged. So, the vessel is trimmed more by the stern to ensure the proper submergence of propeller and avoid ventilation. Furthermore, It can also be speculated that when this ship travels small distances and time is of great importance, the arrangement of containers on board the vessel and trim values hold less importance. These higher trim values cause a drop in vessel performance.

4.6 Transport Efficiency

The variation Energy Efficiency Operational Indicator (EEOI) and other key performance indices over time can be investigated to study the vessel performance. To limit carbon emissions, the International Maritime Organization (IMO) has adopted ship energy efficiency tools (e.g. EEOI, EEDI etc). The ship's fuel energy efficiency in a specific operational cycle or voyage can be represented by EEOI. It is the ratio of the mass of carbon dioxide emitted per useful work done.

$$EEOI = \frac{\sum_j FC_j \times C_{Fj}}{m_{cargo} \times D_i} \tag{16}$$

Where,

j = Fuel type

i = Voyage number

FC = Fuel Consumption (mass)

C = Fuel Consumption to mass of CO_2 conversion factor

m_{cargo} = Mass of cargo

D = Cargo carried distance in Nautical Miles

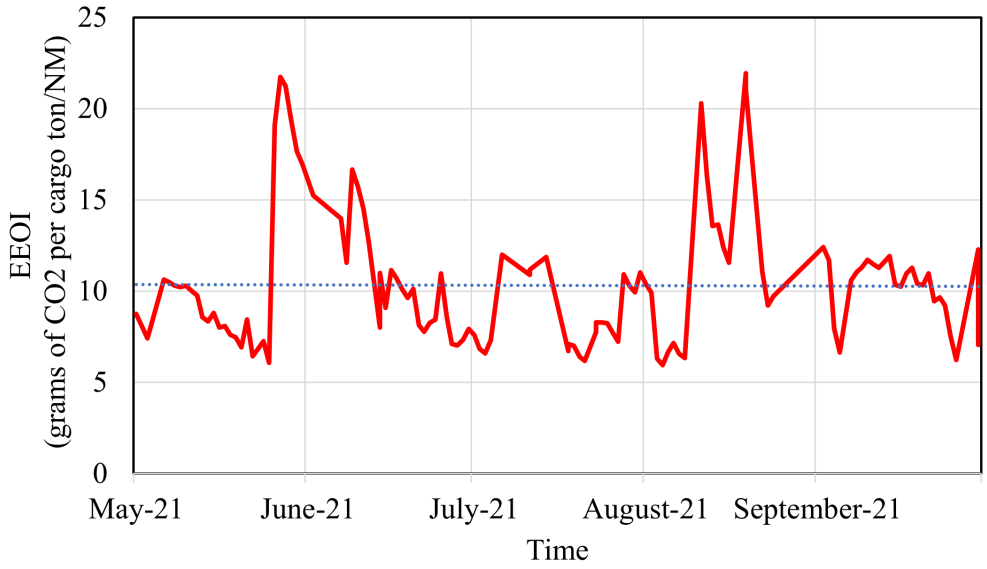


Figure 29: EEOI of the investigated vessel from May 2021 to October 2021

Figure 29 shows the variation of EEOI over a 5-month period for the vessel. Only those EEOI values are taken for which the operation status is 'at sea'. On average 10.3 grams of carbon dioxide per one ton of cargo carried over a distance of one nautical is emitted by the ship. Two peaks can be observed in the graph during the month of June and August. These peaks relate to higher trim values discussed in Section 4.5. Fuel efficiency in the months of June and August is dropped due to higher trim values. Similar trends are also obtained for fuel oil consumption per transport work and EEOI (TEU - grams CO2 per TEU and nautical mile).

The trend line in the Figure 29 shows that the fuel efficiency of the vessel remained constant throughout the investigated time period. In terms of EEOI, no degradation in the vessel performance is observed from May 2021 to October 2021. Similarly, an increase in performance levels is also not observed. Furthermore, additional information about dry-docking and maintenance during this period is not available.

4.7 Noon Data vs High Frequency Data

Noon data is recorded when a new event is started. Complete description of when and how the data is recorded and logged is already given in the Section 2. Figure 30 shows the visualization of noon data recording in comparison with high frequency data cluster for the same noon data.

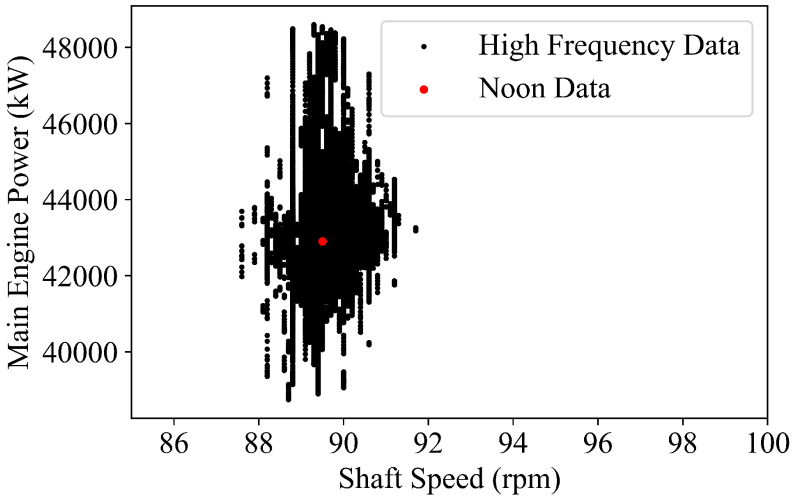


Figure 30: Comparison of noon data with high frequency data

This data is recorded from 10am on May 14th, 2021 to 10am on May 15th, 2021. The vessel travelled from Rotterdam to Singapore through Mediterranean Sea during this time frame. It can be noticed in Figure 30 that the noon data point lies at the mean position of the high frequency data recorded for the 24 hour time. Furthermore, it is observed in Table 9 that noon data point is very close to the mean value of the high frequency data.

Table 9: Comparison of noon data with high frequency data

	Engine Shaft Speed (rpm)	Engine Power (kW)
Noon data	89.504	42904
Mean value of high frequency data	89.504	42911

5 COMPARATIVE ANALYSIS OF MACHINE LEARNING PREDICTION MODELS

In this section, different machine learning (ML) models are trained on the filtered data set. The input features for training the ML models are also analyzed. The performance of the trained ML models are compared and the best model for the available data set is proposed.

5.1 Machine Learning

The Scientific community has been researching and developing machine learning algorithms to analyze and predict the different matters related to efficiency of the ship. Most of the work been done is related to the route optimization and fuel consumption. Now a days, the operational data has been used frequently due to the digital transformation of the marine industry and different methods are being proposed for the calculation of real time performance of the ship (Tsujimoto and Orihara 2019). The introduction of digital twin concept has brought the attention of data scientists and machine learning experts toward the naval architect field to great extent. Data scientists have investigated simple regression techniques to complex neural networks to analyze and predict the vessel performance.

In this study, the purpose of applying machine learning models is to predict the main engine power and engine speed curve for those operating conditions for which the vessel was not being operated on. In other words, machine learning models are trained for predicting the propeller curves using the operational data.

Following machine learning methods are applied on the filtered data and a performance comparison is made.

1. Linear Regression Model
2. Polynomial Regression Models
3. Gradient Boosting Model
4. Decision Tree Model
5. Random Forest Model
6. Multilayer Perceptron Model

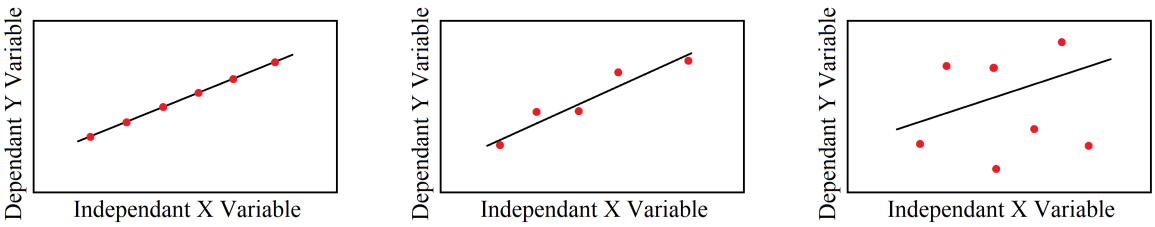
Although different ML models have been trained but there are three principle theories behind all of these methods: polynomial regression, decision trees and artificial neural network.

A polynomial regression method works on simple least square technique. It plots a regression line through the data set. It calculates and minimizes the distance between the each data point and the regression line through various techniques. A decision tree follows a flow chart like structure. It has nodes which represent branches and leafs of a tree. At each branch node it makes a decision by minimizing the error function and moves on to the next branch node. Each leaf node represents an optimized objective problem after computing all the decisions. An Artificial Neural Network (ANN) simulates the working of human brain using different layers of neurons. It consists of the three different types of layers (i.e. input layer, hidden layers, output layer). These layers have 'n' number of neurons. An activation function is associated to each neutron in the layers which introduces the non linearity in the relation between input and output.

5.1.1 Blind Testing

Before the application of ML algorithms on the available data set, the methods to test and evaluate the performance of a ML model is discussed. Blind testing is used to compare the results obtained by the different machine learning algorithms with the testing data set. To choose and compare which predictive machine learning techniques works best for the available data set some metrics are calculated. These metrics include coefficient of determination (R^2), mean absolute error (MAE), mean square error (MSE) and root mean square error (RMSE).

Coefficient of determination ***(R^2)*** It shows how good a model fits the test data. Its value is always between 0 and 1. A model with a value closer to 1 is a good model. Physical representation of different R^2 with the model fit is shown in Figure 31. In this figure, black line represents the trained model predicted line and round red markers represent the actual test data.



(a) $R^2 = 1$ (b) $R^2 = 0.83$ (c) $R^2 = 0$

Figure 31: Visualization of R^2 scores with model fit

R^2 is calculated using equation 17.

$$R^2 = 1 - \frac{\Sigma(y_i - \hat{y}_i)^2}{\Sigma(y_i - \bar{y}_i)^2} \tag{17}$$

where,

y_i = actual y-value

$\hat{y}_i$ = predicted y-value

$\bar{y}_i$ = mean y-value

Mean Absolute Error (MAE) It is a measure of mean absolute error between the predicted value by the machine learning algorithm and the actual value (i.e. test data). This also shows how good a machine learning (ML) model fits the test data. MAE value should be as minimum as possible for an accurate ML model. Its value is calculated using equation 18.

$$MAE = \frac{\Sigma_{i=1}^{n} \mid y_i - \hat{y}_i \mid}{n} \tag{18}$$

where,

n = total number of values

Mean Squared Error (MSE) This parameters also shows how good a ML model fits the test data. It is a measure of mean squared error between the predicted value by the machine learning algorithm and the actual value (i.e. test data). An accurate model should have low MSE value. It is calculated using equation 19.

$$MSE = \frac{\Sigma_{i=1}^{n}(y_i - \hat{y}_i)^2}{n} \tag{19}$$

Root Mean Square Error (RMSE) This parameter is also used to check the accuracy of a trained ML model against the test data. It is the square root of the mean squared error. RMSE value should be low for an accurate ML model. It is calculated using equation 20.

$$MSE = \sqrt{\frac{\Sigma_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \tag{20}$$

All of the parameters discussed above are used to test and compare the accuracy of trained machine learning model against the test data. It is possible that two ML models have same R^2, that is why multiple parameters are used to have a complete comparison. The best ML models will minimize the MAE, MSE and RMSE values while making R^2 to approach 1.

5.2 Methodology

The methodology of the comparative analysis of machine learning prediction models is shown in Figure 32. With the data already filtered and discussed in Section 3, data is partitioned in two groups (i.e. training data set and testing data set). Training data set contains 80% of the total filtered data and the testing data set contains 20% of the filtered data set. This partitioning of data is done randomly. Machine learning models are trained on the training data set. After training process, these ML models are tested on the unseen test data and performance metrics discussed in Section 5.1.1 are calculated. The performances of the trained ML models are compared and best performing model for the available data set is chosen for output prediction and further analysis.

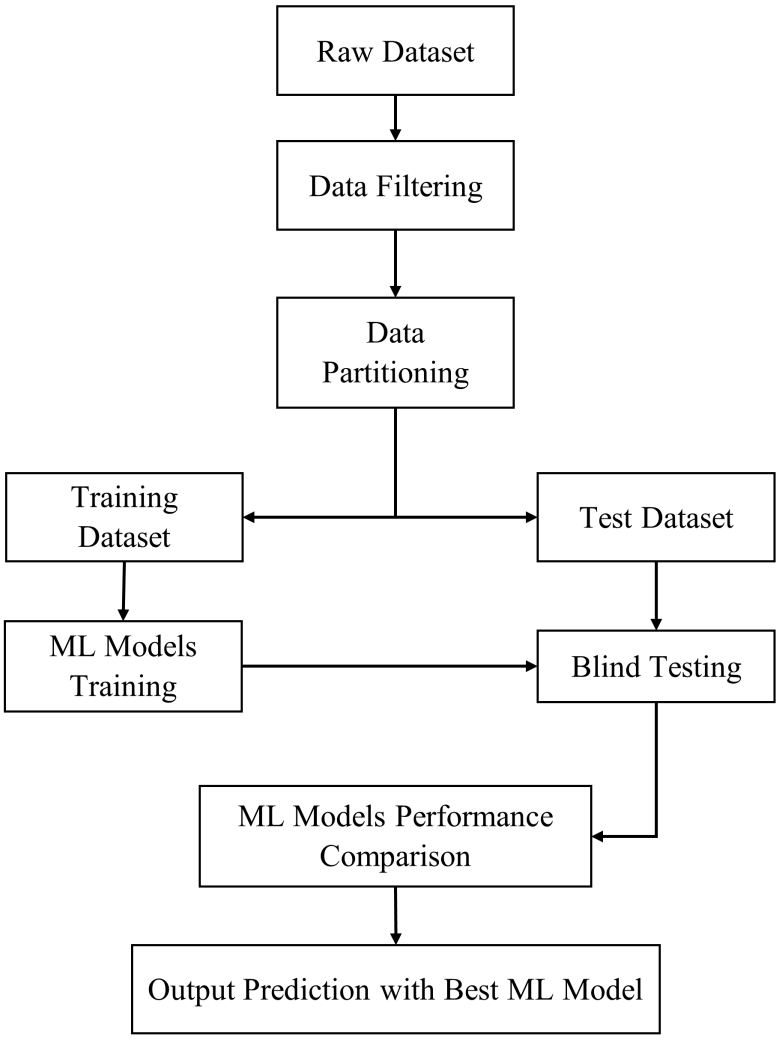


Figure 32: Machine learning prediction models comparative analysis methodology

5.3 Feature Engineering

To train machine learning models, some key operating parameters (i.e. draft, trim, speed etc) should be defined as training features on which the model can be trained to predict the delivered power values. The selection of these features are discussed in this section.

The input features should have high effect on the delivered power. One of the method to select the relevant input features is by using critical thinking and naval architecture literature. Shigunov (2018) discusses the important features which affect the power of the main engine during navigation. According to Ferreira, Lima, and Caprace (2022), main engine power is affected by many parameters including environmental conditions (wind speed, wind direction, current speed, current direction, wave height), navigation conditions (speed over ground, speed through water, rudder angle, rudder angle of attack) and also propulsion system. It is also suggested that those features should be used which are commonly recorded by the vessels. This ensures that the developed method would be applicable to majority of the ships. The suggested input features by Ferreira, Lima, and Caprace (ibid.) are speed over ground, course angle, speed through water, engine shaft speed, mean draft and trim.

A correlation study can be performed to investigate the relation of the recorded operational parameters on the main engine power. This study can help to define which features highly effect the engine power.

Table 10: Correlation coefficient of different parameters with engine power

	Engine Power
Shaft speed	0.979
Speed over ground	0.920
Speed through water	0.942
Mean Draft	0.286
Trim	-0.195
Wind Speed	0.402

The correlation coefficients of different input features with engine power are given in Table 10. A correlation coefficient close to 0 shows no effect on the engine power. A correlation coefficient close to +1 shows a strong direct relation with the main engine power. In this data set, shaft speed, speed over ground and speed through water show a strong direct relation with the engine power. Mean draft also has a direct relation with the engine power but this relation is not as strong as shaft speed, speed over ground and speed through water. A negative correlation indicates a inverse relation but in this correlation study no strong inverse relation has been found. Trim shows a weak inverse relation with the main engine power.

So according to this correlation study following input features should be used to train the machine learning models:

1. Shaft speed
2. Speed over ground
3. Speed through water
4. Mean draft
5. Trim
6. Wind Speed

By looking at the correlation coefficient of input features at different main engine power range, we can predict how the trained machine learning models will perform at these engine power ranges.

The input features used to train the machine learning models should also be input to the trained model to get the predictions. Shaft speed, speed over ground and speed through water depend on each other in the chosen input features. In this case, a series of shaft speed values can be given as an input to the trained model independently in the operating range of the engine (i.e. 19 rpm to 91 rpm). But speed over ground and speed through water depend on this shaft speed and cannot be given as an input independently. However, mean draft, wind speed and trim values do not depend on any other parameter and can be input to the trained model independently. The exclusion of highly correlated operational parameters with engine power (i.e. speed over ground and speed through water) as input features can decrease the accuracy of the machine learning models.

5.4 ML Model Comparison and Analysis

The performance of the selected machine learning models for predicting the main engine power using the shaft speed, draft, wind speed and trim values as input features are compared.

The simple regression methods including linear and polynomial regression were investigated in this study. These are the basic predictive methods. They require less computational time and memory as compared to required to decision tree and artificial neural network (Makridakis, Spiliotis, and Assimakopoulos 2018). The linear regression method assumes that the relation between the input features and the output feature is linear in nature. However, the polynomial regression method predicts the output using a polynomial relation of different degrees. This study investigates the polynomial regression of degree 2 and degree 5.

With regards to the decision tree methods, three different models; Random Forest, simple regression tree and gradient boosted tree were trained and tested. Gradient boosted tree algorithm is based on the Friedman (2001). The hyperparameter values for gradient boosted tree were chosen as default values provided in the Scikit library version 1.1.1: 3 for the maximum depth of tree, 100 for the number of estimators or models and 0.1 for the learning rate. The hyperparameters for Random forest ML model algorithm are also chosen as default from the Scikit library version 1.1.1: 100 for the number of trees in the forest. Hyperparameters for decision tree methods include terms like number of leafs, branches etc.

At the end, Multilayer Perceptron algorithm based on artificial neural network was implemented using Python. For this algorithm, 100 for the hidden layer size, 25 for the maximum number of iterations until solver reaches convergence and 0.01 as the initial learning rate were chosen. Its performance can be increased by fine tuning the hyper-parameters like number of hidden neurons, layers, and iterations. Due to limited graphic processing unit (GPU) resources maximum number of iterations and learning rate were set to 25 and 0.01 respectively.

Table 11: Machine learning algorithms accuracy comparison

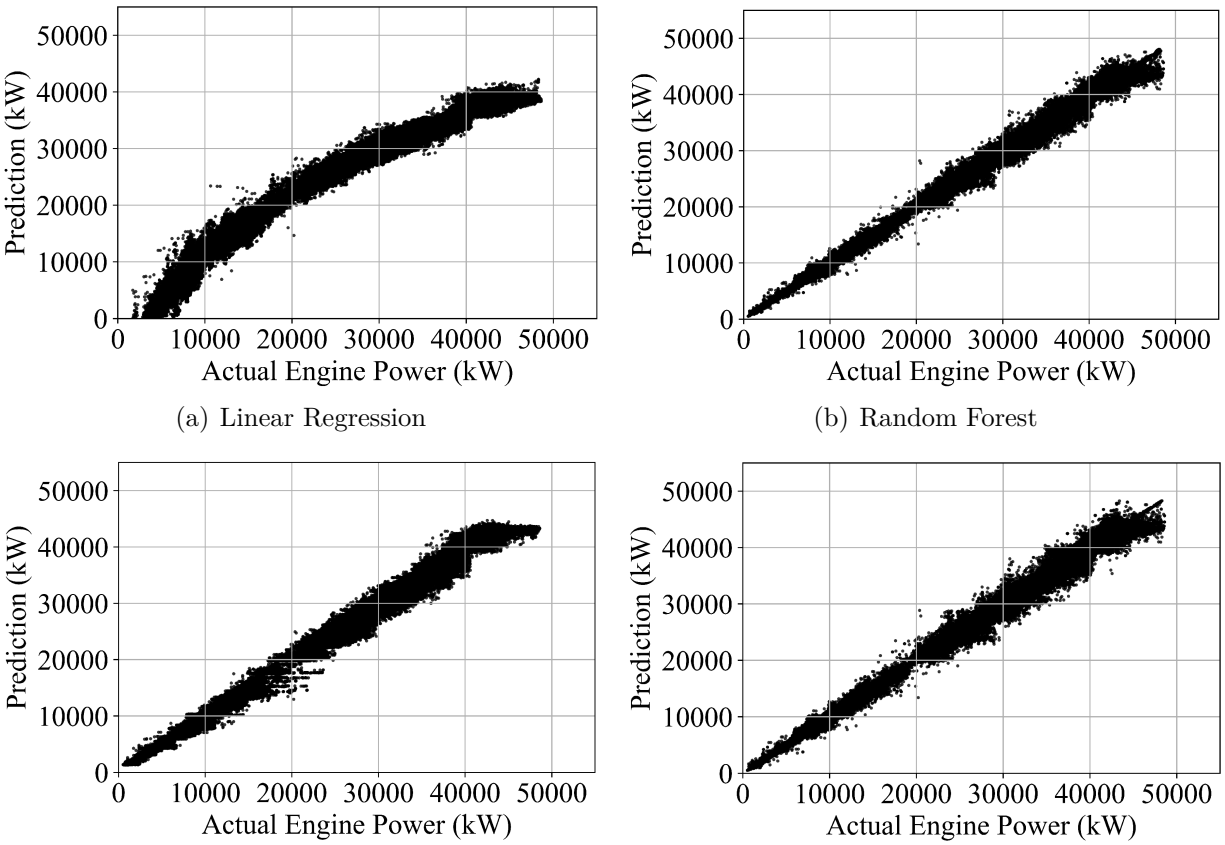
	R^2	MAE	MSE	RMSE
Random Forrest Regressor	0.999	215	135722	368
Decision Tree Regression	0.998	205	143343	378
Gradient Boosting Regression	0.998	372	311527	558
Multilayer Perceptron Regression	0.997	439	394456	628
Polynomial Regression (Degree 2)	0.997	481	449476	670
Polynomial Regression (Degree 5)	0.998	370	301061	548
Linear Regression	0.949	1895	6852290	2617

Blind testing of the ML models was performed on the unseen test data. Table 11 shows the computed R^2, MAE, MSE and RMSE values. The results show that R^2 of all the algorithms are closer to 1. However, the values of other blind testing parameters differ for the trained ML models. It can be identified from these results that the ML models which are based on the decision tree algorithm perform better as compared to other ML models which implement the simple polynomial regression and artificial neural network. Furthermore, results indicate that Random forest has the best performance among the implemented decision tree algorithm and it is the best fit for carrying out the main engine power predictions for different operating conditions.

These results are supported by the previously conducted studies by the researchers on the main engine power prediction using ML algorithms. Study conducted by Gkerekos, Lazakis, and Theotokatos (2019b) also shows that Random Forest model shows the best performance while predicting the fuel oil consumption of the vessel. Similarly, Ferreira,

Lima, and Caprace (2022) came to the same conclusion. So, further analysis of the available data will be carried out using the Random Forest ML model.

Figure 33 shows the performance comparison of the machine learning models. It shows the relation between the predicted power and the actual engine power. For an accurate ML model, this graph should be a straight line with slope equals to 1. It can be observed that the scatter for Random Forrest is less as compared to the other models. The scatter for all of the models increases as the engine power is increased. Hence, the performance of the trained ML models decreases as engine power is increased. The investigation on the decreasing trend of performance is reported in Section 5.4.1.



(a) Linear Regression

(b) Random Forest

(c) Gradient Boost Regressor

(d) Multilayer Perceptron Regressor

Figure 33: Machine learning models performance comparison

5.4.1 Model Performance with variation in Engine Speed

Investigations on decreasing ML model performance with the increase in engine speed values are conducted on the Random Forest algorithm. Blind testing parameters; MAE, MSE and RMSE are calculated at different shaft speeds to investigate the variation in Random Forest algorithm performance at different engine powers. The decreasing trend in Random Forest ML model performance is shown in Table 12.

Table 12: Random Forrest model performance variation with engine shaft speed

Shaft Speed Range (rpm)	MAE	MSE	RMSE
20 - 30	37.28	3879	62.28
30 - 40	55.97	10292	101.45
40 - 50	105.32	30048	173.34
50 - 60	143.28	49735	223.01
60 - 70	194.26	87796	296.30
70 - 80	253.53	164843	406.01
80 - 90	364.57	320073	565.75
90 - 100	385.78	392939	626.85

Figure 34 shows a decreasing trend in model performance with increase in shaft speed.

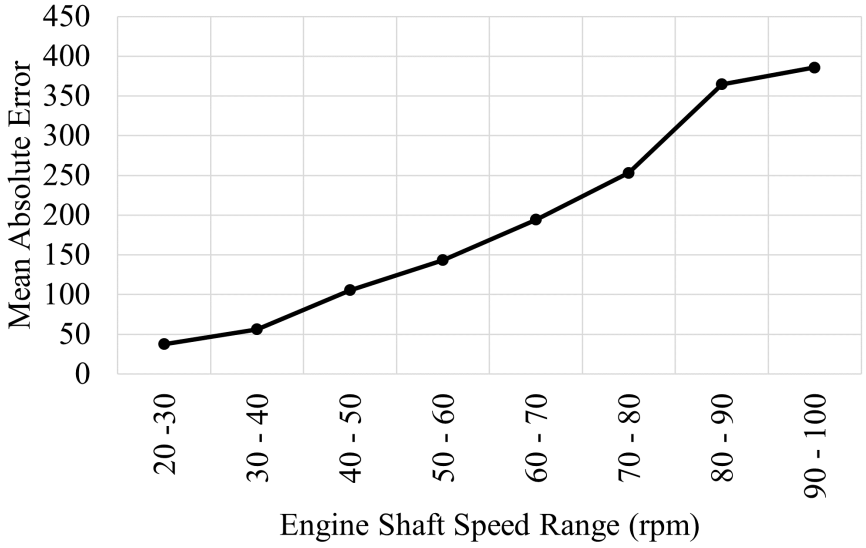


Figure 34: Random Forrest model performance variation with engine shaft speed

This decrease in ML model performance can be attributed to the variation of correlation coefficient between the input and output features at different engine power range.

5.4.2 Effect of Input Features on Model Performance

The performance of Random Forest model can be improved by increasing the number of highly correlated input features during the training stage. Two different Random Forest models are trained with the same hyperparameters but different input features. Table 13 shows the input and output features of the trained Random Forest model. Model-II has the additional input features; speed through water, speed over ground and wind speed. The correlation coefficient of speeds through water and over ground are already discussed in Section 5.3.

Table 13: Random Forest models input and output features

	Model - I	**Model - II**
Input Features	Engine shaft speed Mean draft Trim Wind Speed	Engine shaft speed Mean draft Trim Speed through water Speed over ground Wind Speed
Output Features	Engine power	Engine Power

Random Forest model performance parameters are given in Table 14. It can be observed that the model performance improves by the inclusion of the speed through water and speed over ground in the input features in addition to the engine shaft speed, mean draft, trim and wind speed. The MAE, MSE and RMSE values improves by 117%, 218% and 78% respectively.

Table 14: Random Forest Model-I and Model-II performance comparison

	Model I	Model II
R^2	0.999	0.9997
MAE	215	99
MSE	135722	42619
RMSE	368	206

The performance of Model II is presented in the Figure 35 are given in Table 14. It can be observed that the scatter for Model-II is less as compared to the Model-I (Figure 33b) and the performance of Model-II is better. This result show the importance of highly correlating parameters in the machine learning model training and its performance.

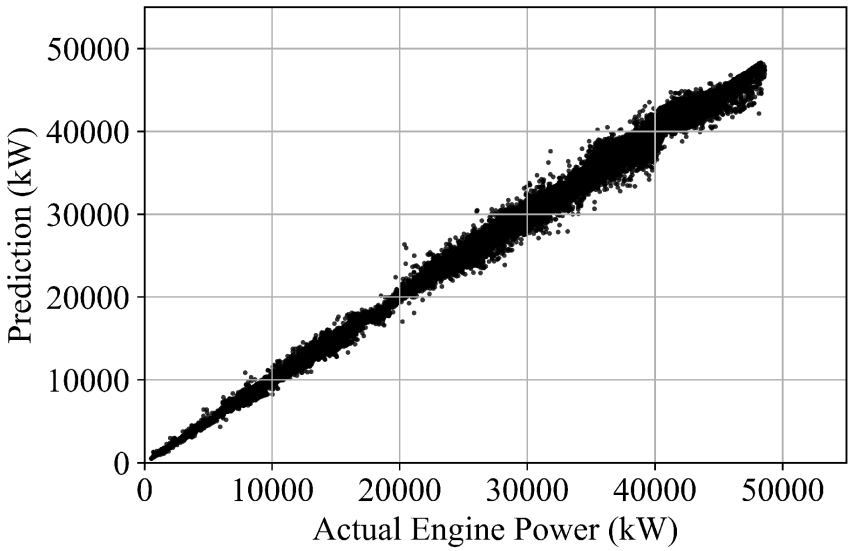


Figure 35: Random Forrest performance with speeds included in input features

5.5 Effect of Wind speeds on Predicted Results

Figure 36 shows a comparison of three predicted engine speed vs engine power curves at different wind speeds. These curves are predicted using the Model-I discussed in Section 5.4.2. The input given to the Model-I for these predictions are given in Table 15.

Table 15: Model-I inputs

Input Features	**Values**
Shaft Speed	50 - 90 RPM
Mean Draft	14 m
Wind Speeds	0 m/s, 20 m/s, 40 m/s
Trim	0 m

Results show that as the wind speed is increased, power demand also increases. The difference in power demands for different curves is reduced at higher shaft speeds. This reduced difference is due to the fact that the performance of the trained random forest also reduces at higher shaft speeds as already discussed in Section 5.4.1 . That is why the RPM-Power curves seem to converge at higher shaft speeds.

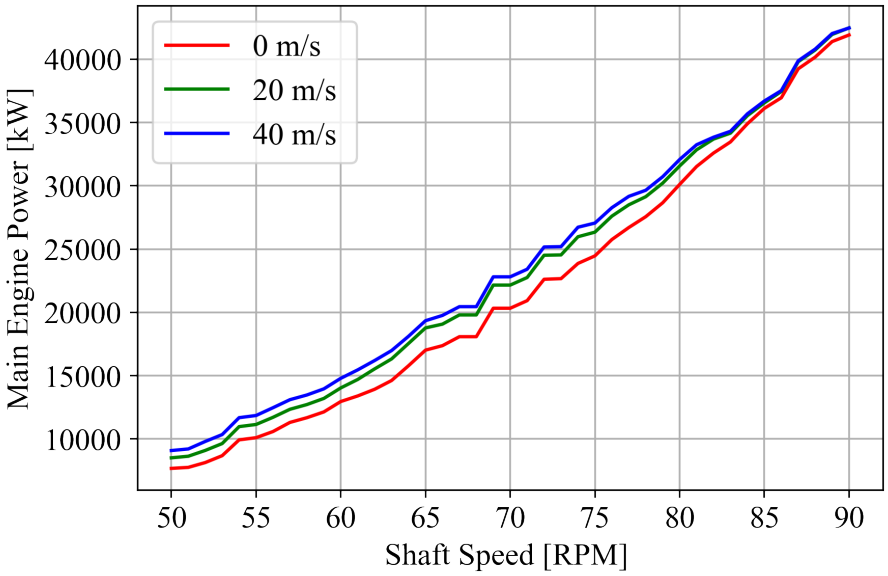


Figure 36: Comparison of predicted RPM-Power curve at different wind speeds

5.6 Summary

This chapter discusses the different machine learning models to predict the output feature of main engine power using operational parameters as input features. A correlation study has been conducted to investigate the importance of available operational parameters as input features. The study performed on operational parameters show that engine shaft speed, speed through water, speed over ground, wind speed, draft and trim are important input features for training a machine learning model. Filtered data obtained after the study performed in Section 3 is split into training and testing data set. Training data set is composed of 80% of the total filtered data and it is used to train machine learning models based on polynomial regression, decision tree and artificial neural network techniques. Then trained models are tested on the remaining 20% of the test data which is unseen by the ML models previously and a performance analysis is performed using blind testing.

The comparison results of the implemented machine learning techniques show that algorithms based on the decision tree had a better performance as compared to polynomial regression and neural network techniques. According to the results presented in Section 5.4, random forest shows the best performance in the blind testing results among the decision tree based algorithms. These results are supported by the previous studies conducted by the researchers.

Moreover, the predicted RPM-Power curves for varying wind speeds show that the power demand increases as the wind speed is increased. Furthermore, this study also shows that the performance of the machine learning model can be improved by increasing the number of highly correlated input features during the training stage.

6 COMPARISON OF ENGINE OPERATIONAL DATA WITH DESIGN DATA

This chapter discusses the comparison of engine operational data with the data obtained during the ship design stage. The design stage data can be obtained from sea trail test, model testing or Computational Fluid Dynamics (CFD) simulations. However, for a better comparison between sea trail test data should be compared with the operational data. But sea trail test data was not available so CFD simulations data has been compared with the operational data.

The importance of numerical flow simulations during the design stage is increasing as the computation power of computers is improving in recent years. However, the full-scale simulations for industrial applications are still computationally expensive and time-consuming. During the design stage, such simulations are the most suited for obtaining flow effects. So, CFD simulations are frequently used for propulsion calculations.

The accuracy of results from a computational fluid dynamics simulation depends on a number of factors. The discrepancy in the Reynold's number due to Froude similitude can have an effect on the calculation of coefficient of friction (C_f). In this case, CFD simulations are performed on a full scale model so there will be no scaling effects in the results. Furthermore, the effect of surface roughness of real ship hull in the CFD simulations is difficult to capture. CFD results also depends on the wave conditions such as calm water condition or different waves of different spectrum. Simplification in the modelling can also lead to inaccurate results. For example, propeller model is simplified to a virtual disk. In such a case the wake effects cannot be captured effectively. So, comparison of CFD simulation results with the real-world operational data will also provide a good validation basis.

6.1 Results and Discussion

In this analysis, engine power for operational and design data is compared. The engine power should be compared for the same operational parameters (i.e. ship speed, engine shaft speed, draft and trim). The majority of the data points in the operational data have the draft values around 14m and 15.5m, as already discussed in Section 4.5. So, it is intuitive to compare the operational and design data at these draft values for a better comparison. Furthermore, the ship operational speed and the speeds of design data should be within the same range.

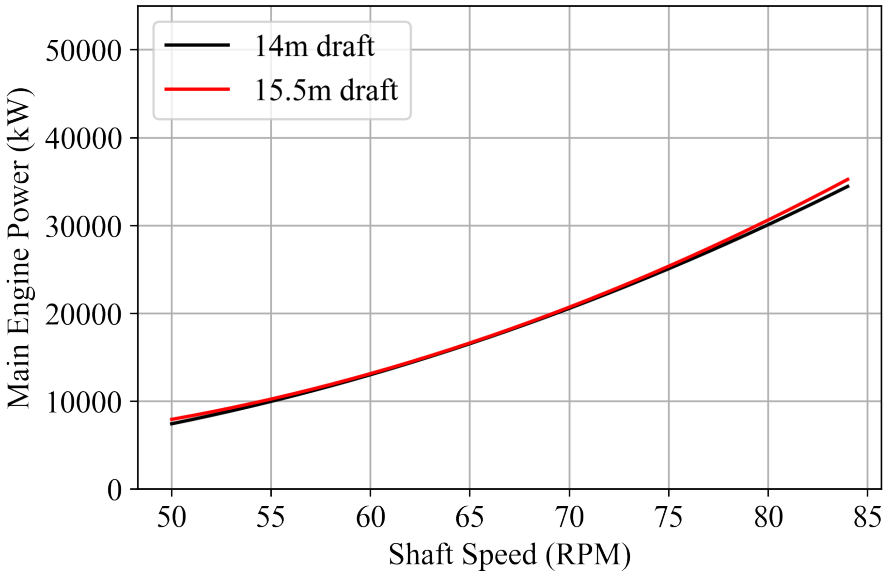


Figure 37: Comparison of operational data curves at different drafts

Figure 37 compares the operational curves at two drafts (i.e. 14m and 15.5m). These curves are obtained by simple curve fitting of filtered operational data at respective drafts. It can be seen that these curves overlap each other and there is very small difference between them. So, it was found that the change in draft of the vessel does not show sufficient contribution to the engine power requirement in the RPM-Power curve. Furthermore, to analyze the operational data it would be better to analyze the engine power against the speed of the vessel in order to investigate the effect of different drafts at engine power requirements.

Figure 38 shows a comparison between the operational data, design data and the random forest predicted curve for 14m mean draft condition. The trained random forest model has speed, draft, wind speed and trim as an input. The inputs given to the model for prediction of main engine power are speed range from 14 knots to 22 knots, 14m draft, 0 m/s wind speed and 0 trim. The operational data shown in orange color has been filtered for 14m mean draft, 0 trim condition, wind speed less than 7.9 m/s and deep water conditions. The operational data is not continuous rather it shows patches of data. Moreover, the operational data has been missing for higher speeds. A polynomial curve fit for the operational data is also shown in the graph. This curve has the exponent 3 as suggested by IMO (2020). The design data is consists of discrete points at 14 knots to 22 knots with 2 knots increment. The design data plot in between the points has been interpolated. It can be observed that from 14 knots, there is negligible difference between the predicted power value and the design data. However, for the rest of points there is a significant difference between the design data and the ML predicted main engine power values. This difference is due to the fact that the operational data is missing for these speed values and machine learning model is extrapolating in this region without any

training examples. The performance of machine learning algorithm is comparatively less during extrapolation as indicated by Hooker (2004).

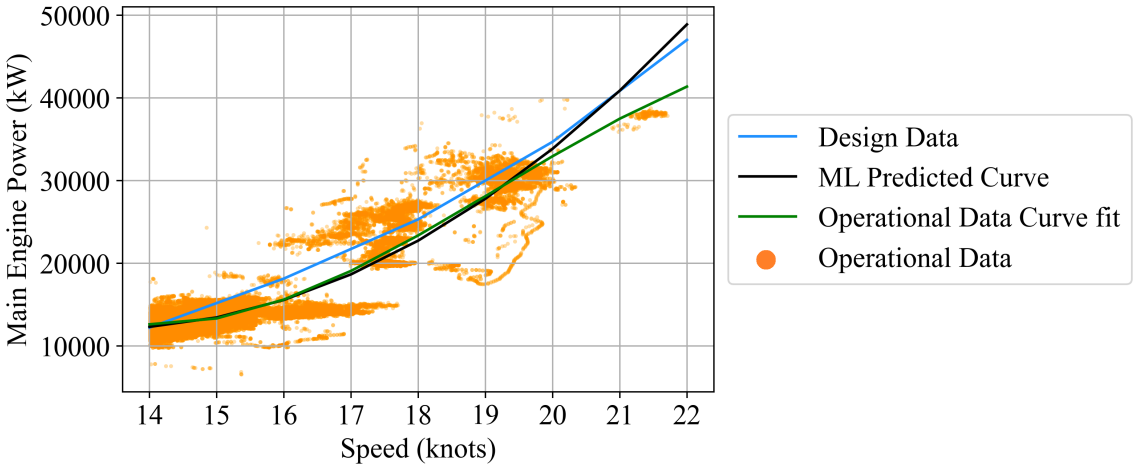


Figure 38: Comparison of filtered operational data, design data and ML predictions at 14m mean draft

Figure 39 shows a similar comparison between the operational data, design data and the ML predictions at 15.5m mean draft. Same random forest model has been used for the predictions. However, the input of mean draft has been changed to 15.5m. Other inputs remained same as discussed above. The operational data has been filtered for 15.5m mean draft, 0 trim condition, wind speed less than 7.9 m/s and deep water conditions. The operational data for 15.5m draft is continuous as compared to 14m draft data. However, some patches of data can be observed at lower speeds. The design data is consists of discrete points at 14 knots to 22 knots with 2 knots increment and it is interpolated in between the point. It can be observed that there is a difference between the design data and the ML predictions 14 knots. This difference is due to the missing data at this speed value. The ML model is extrapolating. In other words, machine learning model does not have sufficient data at these input values for training and to predict the output with great accuracy. Furthermore, it can be observed that the ML output and the filtered operational data curve with exponent 3 are in agreement. This also justifies the IMO (2020) discussion on speed power relation.

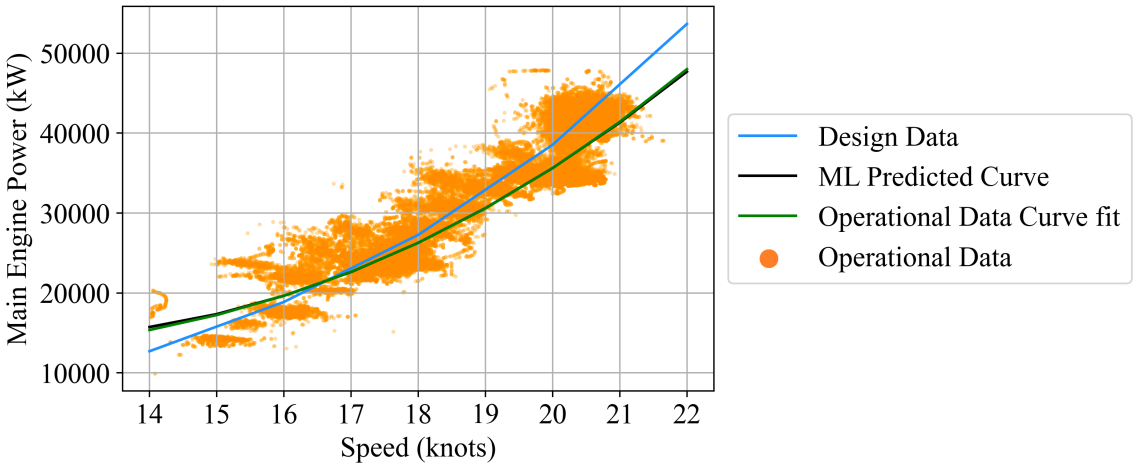


Figure 39: Comparison of filtered operational data, design data and ML predictions at 15.5m mean draft

It can be concluded from Figure 38 and Figure 39 that when the data is complete for some operating condition, ML shows a good accuracy. However, if the data is incomplete and it is in patches (as in case of Figure 38) then results of ML are not accurate. Furthermore, ML is computationally more costly than simple data filtering. So, it is proposed that if the operational data is available for some operating condition it should be used for the analysis rather than using ML predictions for analysis.

Now a comparison between the engine layout curve, design data and operational data has been made. Complete data is considered without the filtering specific draft values. This complete high frequency data has already been filtered for outliers and calm water conditions as discussed in Section 3.

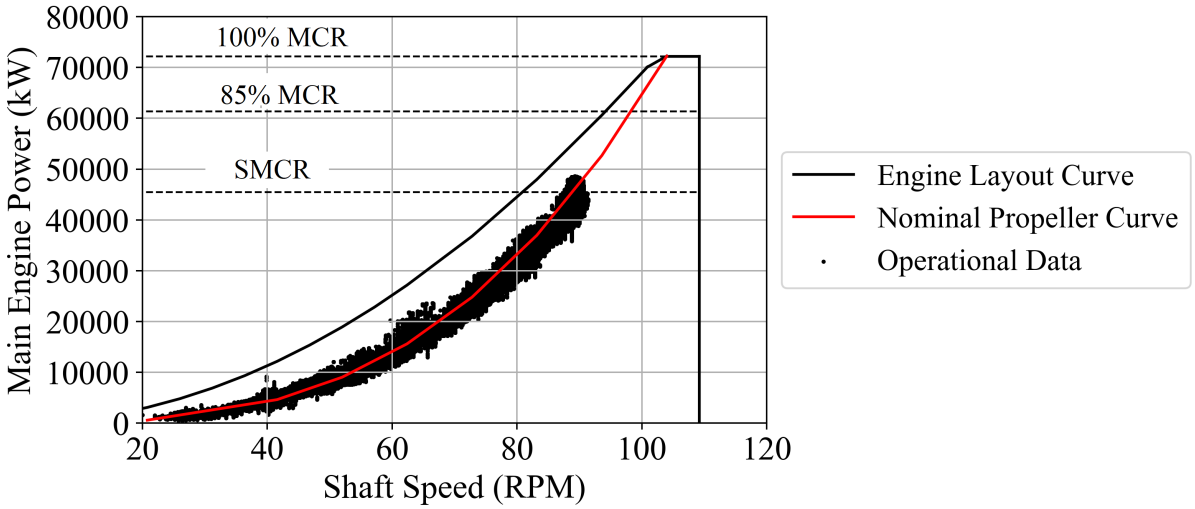


Figure 40: Comparison of filtered operational data with Nominal Propeller Curve and Engine Layout Curve

Figure 40 shows a comparison of complete operational data with the nominal propeller

curve. It can be seen that the vessel is being operated well within the maximum continuous power range of the engine. The specified maximum continuous rating (SMCR) is 42,511 kW which is 63% of the maximum continuous rating (MCR). The maximum operational engine power of the vessel in these 8 months was recorded 48,599 kW. The nominal propeller curve represents the 0% light running margin. Light running margin is calculated using following formula:

$$LRM = \frac{N_o - N_{\text{nominal propeller curve}}}{N_{\text{nominal propeller curve}}} \times 100 \tag{21}$$

where,

N_o = Shaft speed at SMCR

$N_{\text{nominal propeller curve}}$ = Shaft speed of nominal propeller curve at SMCR

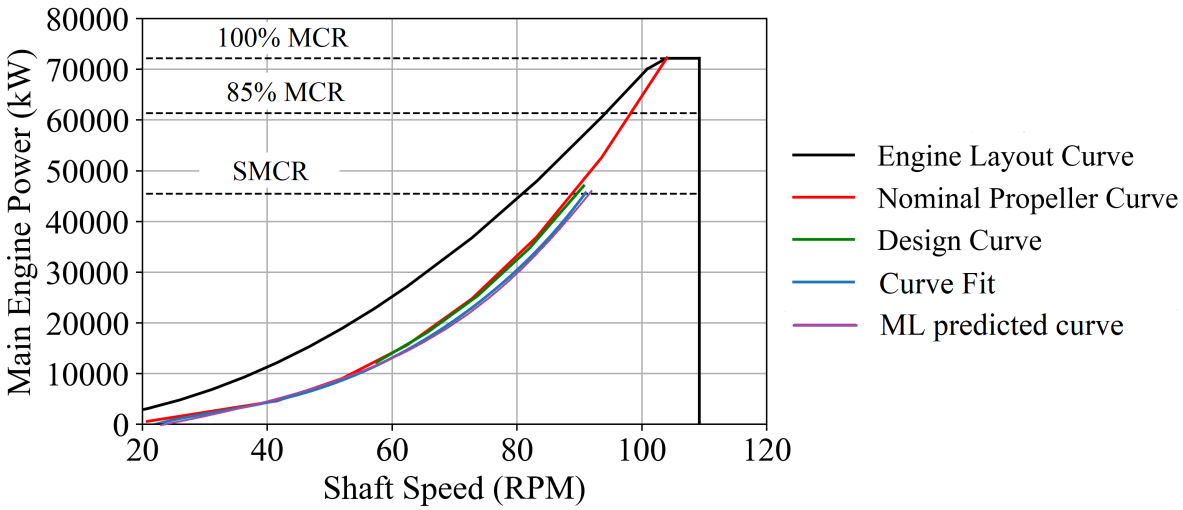


Figure 41: Comparison of operational data curves with design data curves

Figure 41 shows a comparison between the operational data curve fit, CFD power predictions and ML predictions. The inputs to the ML model are 20-93 rpm shaft speed range, 0 m/s wind speed and 0m trim. A polynomial curve of third order has been used to represent the filtered operational data as suggested by IMO (2020). There is a good agreement between the filtered operational data, ML predictions and the CFD calculations. Table 16 shows the calculated light running margin. The difference between the light running margin is very small.

Table 16: Light Running Margin comparison of Design data and Operational Data

	Light Running Margin
Design Data	1.92 %
Filtered Operational Data	2.88 %
ML predictions	3.21 %

These small difference in the light running margin show that the heavy running operation state of the propeller due to rough environmental conditions has been filtered out from the high frequency data set. Only calm water conditions are left behind in the operational data set. Furthermore, the results also show that the current operational vessel efficiency is also comparable to the design stage. Hull fouling or the marine growth can increase the added resistance. This increases the engine operational power requirement. In such a case, operational curve moves upward and reaches the engine maximum engine layout curve and the available engine power is reduced. However, in this case added resistance due to the aforementioned reasons is minimal.

6.2 Summary

Filtered operational data, design data (CFD simulations), and ML predictions are compared in this chapter. Results indicate a good agreement between the ML predictions and the filtered operational data. The ML predictions show good accuracy for filtering out operational and environmental conditions. However, if some of the data is missing for some operational conditions in the training data set then the accuracy of ML predictions is reduced. So, it is proposed that if the operational data is available for required operational condition it should be used for the analysis rather than using ML predictions for analysis.

A comparison between the engine layout curve, design data and filtered operational data show that the difference between the light running margin of design data, filtered operational data and ML predictions is small. This means that the raw data has been filtered efficiently for the different environmental conditions. The ML predictions are also comparable to the filtered data set. However, it is suggested that normal data filtering should be preferred when the operational data for available rather than ML.

Furthermore, results show minimal added resistance due to hull fouling or marine growth and no degradation in the vessel propulsion performance has been observed.

7 CONCLUSIONS

This study deals with Big Data Analytics (BDA) of high-frequency sensor data obtained from an intercontinental 14000 TEU containership. This data has been stored on the University of Rostock server and is accessed using Python programming language. The data includes vessel information from May 2021 to Jan 2022 with a sampling rate of 5 Hz.

The accessed data has been filtered using statistical filters and environmental filters mentioned in ISO 19030. A 5 min block length has been selected after a sensitivity analysis, for the application of Chauvenet's criterion and validation statistical filters for the removal of outliers. These outlier removal methods collectively eliminated approximately 39% of data. The majority of data points ($\sim$ 36%) were eliminated by Chauvenet's criterion. After the application of statistical filters, environmental filters have been applied to obtain the data recorded when the ship was in calm waters and sea trail test conditions were mimicked. Application of environmental filters shows that one-third of the time during its journey, the ship was experiencing wind speed greater than 7.9 m/s and it was mostly sailing (>97.5% of the total journey) in the deep seas without experiencing shallow water effect.

Using the filtered data set, a general data analysis is conducted to better understand the voyage and trading patterns of the vessel. An EEOI anlaysis show that the fuel efficiency of MSC Bari remained constant throughout the investigated time period. In terms of EEOI, no degradation in the vessel performance is observed from May 2021 to October 2021.

Then high-frequency data set filtered for statistical outliers is used to train different machine learning models to predict the output feature of the main engine power using operational parameters and environmental conditions as input features. A correlation study has been conducted to investigate the importance of available operational parameters as input features. The study performed on operational parameters show that engine shaft speed, speed through water, speed over ground, wind speed, draft and trim are important input features for training a machine learning model. The filtered data set is split into training (80%) and testing (20%) data set. Training data set is used to train machine learning models based on polynomial regression, decision tree and artificial neural network techniques. Then trained models are tested on the remaining 20% of the test data which is unseen by the ML models previously and performance analysis is performed using blind testing.

The comparison results of the implemented machine learning techniques show that algorithms based on the decision tree had better performance as compared to polynomial regression and neural network techniques. Random forest shows the best performance in the blind testing results among the decision tree-based algorithms. These results are

supported by the previous studies conducted by the researchers. Furthermore, this study also shows that the performance of the machine learning model can be improved by increasing the number of highly correlated input features during the training stage.

At the end, a comparison between the design data, ML predictions and the filtered operational data has presented. It is concluded that there is a good agreement between the operational data and the design data. Results depict a good agreement between the ML predictions and the filtered operational data. Although, the ML predictions show good accuracy for filtering out operational and environmental conditions but if some of the data is missing for some operational conditions in the training data set then the accuracy of ML predictions is reduced. So, it is proposed that if the operational data is available for required operational condition it should be used for the analysis rather than using ML predictions for analysis.

Results show only a marginal difference between the light running margin between the operational data and design data. It depicts that the vessel efficiency has not degraded over time due to hull fouling, marine growth or other some means.

8 FUTURE WORKS

This chapter discusses the proposed future work which can be carried out to further investigate the high frequency vessel operational data. Future recommendations are given below:

- Future work should consider the investigation of vessel performance over a longer period of time. In this study, only 8 month operational data was investigated in which no degradation in the vessel performance was observed. In addition the dry docking and maintenance should also be reported which will be vital in the implementation of ISO 19030. Operational data for longer period of time will also increase the amount of data. This will also prove vital for the training of Machine Learning algorithm.
- As discussed in the Section 5, the performance of the trained ML algorithms decreases at higher engine powers. It can be due to the decrease in the dependence of output feature on the input features with the increase in engine power. Input features should be further investigated and future research should also focus to solve this problem.

ACKNOWLEDGEMENTS

I want to express my gratitude to the University of Rostock and Mecklenburger Metallguss GmbH which provided me with the opportunity to conduct this research. I would like to acknowledge Mr. Lutz Kleinsorge and Mr. Hauke Baumfalk for supervising this thesis. I would like to extend my gratitude towards University of Liege and Emship faculty especially Professor Philippe Rigo.

I am grateful to my parents and siblings for their support throughout my academic career. I could not have completed this thesis without insights from Dr. Asad Abbas. Furthermore, I cannot forget the moral support from my colleague Mr. Ammar Muhammad Khan.

At the end, I want to thank Stackoverflow and Youtube community without whom I could not have written a single line of Python code.

References

Baumfalk, Hauke J. (2018). "Assessment of ISO 19030 based on operational data of a container vessel". In: *University of Rostock*.

Beşikçi, E Bal et al. (2016). "An artificial neural network based decision support system for energy efficient ship operations". In: *Computers & Operations Research* 66, pp. 393–401.

Brandsæter, Andreas and Erik Vanem (2018). "Ship speed prediction based on full scale sensor measurements of shaft thrust and environmental conditions". In: *Ocean Engineering* 162, pp. 316–330.

Brynjolfsson, Erik, Lorin M Hitt, and Heekyung Hellen Kim (2011). "Strength in numbers: How does data-driven decisionmaking affect firm performance?" In: *Available at SSRN 1819486*.

Coraddu, Andrea et al. (2017). "Vessels fuel consumption forecast and trim optimisation: a data analytics perspective". In: *Ocean Engineering* 130, pp. 351–370.

Erto, Pasquale et al. (2015). "A procedure for predicting and controlling the ship fuel consumption: its implementation and test". In: *Quality and Reliability Engineering International* 31.7, pp. 1177–1184.

Ferreira, Ricardo dos Santos, João Victor Padilha de Lima, and Jean-David Caprace (2022). "Comparative analysis of machine learning prediction models of container ships propulsion power". In: *Ocean Engineering* 255.

FleetMon (2020). *MSC Bari*. [Online; accessed June 27, 2022]. URL: https://www.fleetmon.com/vessels/msc-bari_9461441_2129731/?language=de.

Fofonoff, NP (1985). "Physical properties of seawater: A new salinity scale and equation of state for seawater". In: *Journal of Geophysical Research: Oceans* 90.C2, pp. 3332–3342.

Friedman, Jerome H. (2001). "Greedy function approximation: A gradient boosting machine." In: *The Annals of Statistics* 29.5, pp. 1189–1232. DOI: 10.1214/aos/1013203451. URL: https://doi.org/10.1214/aos/1013203451.

Gkerekos, Christos, Iraklis Lazakis, and Gerasimos Theotokatos (2019a). "Machine learning models for predicting ship main engine Fuel Oil Consumption: A comparative study". In: *Ocean Engineering* 188, p. 106282.

— (2019b). "Machine learning models for predicting ship main engine Fuel Oil Consumption: A comparative study". In: *Ocean Engineering* 188, p. 106282. ISSN: 0029-8018. DOI: https://doi.org/10.1016/j.oceaneng.2019.106282. URL: https://www.sciencedirect.com/science/article/pii/S0029801819304561.

Hasselaar, Thijs Willem Frederik (2011). "An investigation into the development of an advanced ship performance monitoring and analysis system". PhD thesis. Newcastle University.

Hooker, Giles (2004). *Diagnostics and extrapolation in machine learning*. stanford university.

IMO (2020). *Fourth IMO Greenhouse Gas Study 2020*.

iXblue (2014). *Octans Fifth Generation Survey-Grade Surface Gyrocompass and Motion Sensor Datasheet*.

Kannan, K Senthamarai and K Manoj (2015). "Outlier detection in multivariate data". In: *Applied Mathematical Sciences* 47.9, pp. 2317–2324.

Karagiannidis, Pavlos and Nikos Themelis (2021). "Data-driven modelling of ship propulsion and the effect of data pre-processing on the prediction of ship fuel consumption and speed loss". In: *Ocean Engineering* 222, p. 108616.

Kwak, Sang Kyu and Jong Hae Kim (2017). "Statistical data preparation: management of missing values and outliers". In: *Korean journal of anesthesiology* 70.4, p. 407.

Larsson, Lars (2010). "Ship resistance and flow". In: *Published by The Society of Naval Architects and Marine Engineers, SNAME, The Principles of Naval Architecture Series, ISBN: 978-0-939773-76-3*.

Latecki, Longin Jan, Aleksandar Lazarevic, and Dragoljub Pokrajac (2007). "Outlier detection with kernel density functions". In: *International Workshop on Machine Learning and Data Mining in Pattern Recognition*. Springer, pp. 61–75.

Lee, Habin et al. (2018). "A decision support system for vessel speed decision in maritime logistics using weather archive big data". In: *Computers & Operations Research* 98, pp. 330–342.

Luo, Shaoze, Ning Ma, and Yoshiaki Hirakawa (2016). "Evaluation of resistance increase and speed loss of a ship in wind and waves". In: *Journal of Ocean Engineering and Science* 1.3, pp. 212–218.

Makridakis, Spyros, Evangelos Spiliotis, and Vassilios Assimakopoulos (2018). "Statistical and Machine Learning forecasting methods: Concerns and ways forward". In: *PloS one* 13.3, e0194889.

MEPC, IMO (2018). "304 (72) Initial IMO Strategy on Reduction of GHG Emissions from Ships". In: *IMO: London, UK*.

Online Source (2021). URL: `https://www.imo.org/en/MediaCentre/HotTopics/Pages/Cutting-GHG-emissions.aspx`.

Report, UN Global Pulse (2012). "UN Global Pulse (May 2012) Big Data for Development: Challenges and Opportunities". In.

Rotteveel, Erik and Robert Hekkenberg (May 2015). "The Influence of Shallow Water and Hull Form Variations on Inland Ship Resistance". In.

Sasa, Kenji et al. (2015). "Evaluation of ship performance in international maritime transportation using an onboard measurement system-in case of a bulk carrier in international voyages". In: *Ocean Engineering* 104, pp. 294–309.

Shigunov, Vladimir (2018). "Numerical Prediction of Added Power in Seaway". In: *Journal of Offshore Mechanics and Arctic Engineering* 140.5.

Soner, Omer, Emre Akyuz, and Metin Celik (2018). "Use of tree based methods in ship performance monitoring under operating conditions". In: *Ocean Engineering* 166, pp. 302–310.

StormGeo Optimziation Tools (Jan. 2022). URL: `https://www.stormgeo.com/`.

Tsujimoto, Masaru and Hideo Orihara (2019). "Performance prediction of full-scale ship and analysis by means of on-board monitoring (Part 1 ship performance prediction in actual seas)". In: *Journal of Marine Science and Technology* 24.1, pp. 16–33.

Vinutha, HP, B Poornima, and BM Sagar (2018). "Detection of outliers using interquartile range technique from intrusion dataset". In: *Information and Decision Sciences*. Springer, pp. 511–518.

Wan, Xiang et al. (2014). "Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range". In: *BMC medical research methodology* 14.1, pp. 1–13.

Wang, Kai et al. (2017). "Study on route division for ship energy efficiency optimization based on big environment data". In: *2017 4th International Conference on Transportation Information and Safety (ICTIS)*. IEEE, pp. 111–116.

Williamson, David F, Robert A Parker, and Juliette S Kendrick (1989). "The box plot: a simple visual method to interpret data". In: *Annals of internal medicine* 110.11, pp. 916–921.

Wisam, Jabary et al. (June 2022). "Development of a Unified Data Model to Improve Ship Operational Performance Analyses". In.

Yan, Xinping et al. (2018). "Energy-efficient shipping: An application of big data analysis for optimizing engine speed of inland ships considering multiple environmental factors". In: *Ocean Engineering* 169, pp. 457–468.