National Technical University of Athens
School of Naval Architecture and Marine Engineering

Diploma Thesis

Investigation of data preprocessing techniques for ship performance analysis

---

Panagiotis Georgios Iliopoulos

Supervisor: Nikolaos Themelis, Assistant Professor

Athens, December 2022

Abstract

The environmental impact of air emissions produced by the maritime industry is being reduced by increasing the operating energy efficiency of existing ships. An increasing number of vessels are equipped with sensors and devices for monitoring operational behavior, and the amount and access to operational data is gradually increasing. Big data analytics can drastically improve the ship's performance. With the use of proper data preprocessing techniques as well as domain expertise, this research provides an extensive data analytics framework for tracking ship performance under localized operational conditions. A data set from a containership is used to demonstrate the proposed framework. Due to various reasons described in this thesis, the operational data may contain erroneous data points that are critical to assess before performing data analysis or building mathematical and statistical models. The presented investigation relates to detecting data anomalies, identifying the ship's localized operational conditions, calculating the relative correlations among the ship's operational parameters, quantifying the ship's performance in each of the respective conditions, and the visual representation and analysis of the results. The innovative aspect of this study is the provision of a KPI (i.e., key performance indicator) for ship performance quantification in order to determine the optimal performance trim-draft mode under the engine modes of the case study ship. The suggested framework can be used as an operational energy efficiency measure to provide data quality evaluation and decision support for ship performance monitoring that is valuable to both ship operators and decision-makers.

Περίληψη

Οι περιβαλλοντικές επιπτώσεις των ατμοσφαιρικών ρίπων , οι οποίοι παράγονται από τη ναυτιλιακή βιομηχανία, μπορούν να μειωθούν με την αύξηση της ενεργειακής αποδοτικότητας των υφιστάμενων πλοίων. Πολλά υπάρχοντα πλοία είναι εξοπλισμένα με αισθητήρες και συσκευές που στοχεύουν στην συνεχή παρακολούθηση και καταγραφή του επιχειρησιακού προφίλ τους. Γι΄αυτό τον λόγο η ποσότητα και η πρόσβαση σε επιχειρησιακά δεδομένα εν-λειτουργία πλοίων αυξάνεται σταδιακά. Με την ανάλυση επιχειρησιακών δεδομένων μπορεί να βελτιωθεί δραστικά η ενεργειακή απόδοση του πλοίου κατά την διάρκεια του κύκλου ζωής του. Στην συγκεκριμένη μελέτη παρουσιάζεται μια καινοτόμος μεθοδολογία ανάλυσης δεδομένων ,με σκοπό την παρακολούθηση και καταγραφή της ενεργειακής απόδοσης του υπό μελέτη πλοίου ,υπό συγκεκριμένες επιχειρησιακές συνθήκες, μέσω της εφαρμογής κατάλληλων τεχνικών προετοιμασίας και ανάλυσης δεδομένων καθώς και την αξιοποίηση της εμπειρικής γνώσης στον τομέα της ναυτιλιακής βιομηχανίας. Ένα σύνολο λειτουργικών δεδομένων από ένα υπάρχον πλοίο μεταφοράς εμπορευματοκιβωτίων μελετήθηκε για την παρουσίαση και την αξιολόγηση της προτεινόμενης μεθοδολογίας. Για διάφορους λόγους που περιγράφονται στην παρούσα εργασία, τα δεδομένα των υπό εξέταση λειτουργικών παραμέτρων του πλοίου μπορεί να περιέχουν επισφαλείς μετρήσεις, και γι΄αυτό είναι κρίσιμη η διερεύνηση και η αξιολόγηση των διαθέσιμων μετρήσεων πριν την εφαρμογή εξεζητημένων μαθηματικών και στατιστικών μοντέλων ανάλυσης δεδομένων. Στο υπό μελέτη σύνολο δεδομένων εξετάζετε η ανίχνευση και απομόνωση επισφαλών μετρήσεων , διερευνάτε το επιχειρησιακό προφίλ του πλοίου υπό τοπικές λειτουργικές συνθήκες, προσδιορίζετέ η αλληλεπίδραση και η συσχέτιση μεταξύ συγκεκριμένων λειτουργικών παραμέτρων, προσδιορίζετε ποσοτικά η απόδοση του πλοίου σε κάθε μία από τις αντίστοιχες τοπικές λειτουργικές συνθήκες και γίνεται η γραφική αναπαράσταση και ανάλυση των τελικών αποτελεσμάτων. Η καινοτόμος πτυχή αυτής της μελέτης είναι ο υπολογισμός ενός KPI (δηλαδή, καίριου δείκτης απόδοσης) για τον ποσοτικό προσδιορισμό της βέλτιστης απόδοσης του πλοίου με σκοπό τον εντοπισμό του καταλληλότερου συνδυασμού των λειτουργικών παραμέτρων τρίμ και βυθίσματος. Η ανάλυση αυτή μπορεί να αποτελέσει χρήσιμο οδηγό στην διαδικασία λήψης κρίσιμων αποφάσεων που αφορούν τις συνθήκες λειτουργίας ,συντήρησης και φόρτωσης ενός πλοίου και να οδηγήσει στην βελτίωση της λειτουργικής αποδοτικότητα τους.

Acknowledgments

The present work is the final requirement for completing my studies at the School of Naval Architecture and Marine Engineering of the National Technical University of Athens.

To begin with, I would like to express my sincere gratitude to all the academic personnel of the school and, especially, to my supervisor, Assistant Professor Nikolaos Themelis, for giving me the opportunity and inspiration to explore this topic. This diploma thesis would not have been possible without his support. Furthermore, I am grateful for his excellent cooperation, patience, and guidance throughout the project.

I want to express my gratitude to my family and friends for their constant encouragement, support, and motivation throughout my studies. In addition, I want to thank my colleagues for their encouragement and support throughout my studies, with whom I worked side-by-side and who inspired me to evolve into a better person and engineer.

I am also grateful for the data provided by Prisma Electronics.

Table of contents

List of Figures

List of Tables

1. Introduction

1.1 Data analysis in shipping

The marine industry is one of many in today's digital age, where competition is severe, and businesses are continuously investing in solutions that can help them enhance efficiency while cutting overall costs. As a result, commercial shippers and other end users are increasingly in demand for cutting-edge solutions like marine data analysis. Big data is utilized in the shipping sector to perform predictive analysis, enhance overall ship performance, and increase the ship's productivity. In addition, big data analytics is being actively used to improve decision-making and can be used for the duration of a ship's life to prevent and predict further costs. [1]

By enhancing overall shipping operations, enhancing ship safety, and protecting the environment, predictive analytics technologies have the potential to transform the shipping sector. The great degree of customization that these solutions provide, depending on the particular requirements of any port or shipping company, is also anticipated to support demand over the projected period. The demand for freight transport will rise dramatically in the upcoming years due to the expansion of globalization. As a result, maritime enterprises will increasingly need advanced data processing and predictive analytics to improve productivity and cost savings. These elements are fueling global demand for marine analytics. [2]

Big Data in maritime and marine data analytics can essentially be classified into three sections based on the type and volume of data generated:

- Utilizing information from numerous logs, manifests system parameters, bunker statistics, etc., to manage vessels. This will involve effective bunkering, better staff management, and improved vehicle maintenance using digital twins.
- Using information kept by port authorities, freight forwarders, trading houses, etc., for port and cargo management. This will entail effective cargo handling, commodities tracking, port infrastructure improvement, etc.
- Data from location monitoring systems like AIS and LRIT, photos from ships, coastal and space radars, optical sensors, etc., are used in the analysis of spatial imaging. Soon, this will also encompass effective routing, fleet tracking, traffic pattern analysis, anomaly detection, etc.

Even if the shipping industry acknowledges the importance of data analysis for decision-making in areas like reduction in fuel consumption and improving the vessels' environmental footprint, there are many challenges they have yet to overcome in order to establish a data-driven culture as it is moving towards digitalization. [3]

Some of them are summarized down below:

- **Data transfer:** Ships frequently have a huge number of sensors inside. Data transfer from the sensors is a significant source of uncertainty. Every sensor has a unique connection bandwidth requirement, so it's critical to have appropriate data transmission for each sensor to send its data to the database. High-tech communication technologies can help to accelerate the rate of data transport.

- **Data Accuracy:** Interpretation mistakes could arise from insufficient data. All new entries will be too many for the database to keep up with. Therefore, the data should preferably be free from errors. A major issue for the sector will be data quality.
- **Data Integration:** The marine sector currently uses inconsistent and frequently erroneous data collection methods. For analysis, it will be necessary to integrate data from several sources. To monitor the performance of the ship, for instance, fuel usage, GPS data, and engine data would need to be combined.

The shipping industry must address these challenges to drastically improve the quality of available data. This will require critical investments in technological equipment. Remote networks should be used to collect data from ships autonomously using sensors. A robust wireless network with high transmission capabilities is required for the shipping industry. The real-time sensor data will come to the database and be distributed to the interested parties giving them up-to-date information on what is happening onboard. Data management needs to be considered as well. It is necessary to store and structure the data effectively after data acquisition, especially when dealing with big data sets. Sophisticated data storage systems and technologies need to be implemented for the storage and distribution of the data for further and more detailed analysis. After improving the quality of the available data, advanced data analytics techniques need to be considered. A comprehensive description of data analytics is presented in the next paragraphs.

1.2 Introduction to data analytics

Data analytics has just recently been incorporated into the maritime industry. So, there is a lot of research yet to be done before engineers conclude effective and easily interpreted ways to manage and analyze vessel data. The main phases of data analysis, followed by the four most fundamental data analytics categories, are described below. A short description of data analytics techniques is given afterward. [4]

Data analysis involves **various phases**, including:

- Identifying the data requirements or how the data is grouped is the first stage. Data might be divided based on the ship's operational profile, loading conditions, weather conditions, or other factors. Data values could be categorical or numerical.
- The process of gathering data is the second phase in data analytics. Multiple tools, including speed sensors, pressure sensors, mass flow meters, anemometers, and others, can be used to accomplish this.
- Data must first be structured so that it may be studied after it has been collected. A spreadsheet or other software tool that can handle statistical data may be used for this.
- After then, the data is cleaned up for analysis. This indicates that it has been cleaned up and double-checked to make sure there is no duplicate, errors, or missing information. Before the data is sent to a data analyst for analysis, this stage aids in the correction of any inaccuracies, and it's called the data preparation process.

Four fundamental **categories** of data analytics are distinguished.

- **Descriptive analytics**: This explains what has occurred over a specific time period.
- **Diagnostic analytics**: It reflects an understanding of why something happened to the system. This requires more varied data inputs as well as some speculation.

- **Predictive analytics**: This shifts to what is most likely going to occur soon.
- **Prescriptive analytics**: This offers suggestions on how to proceed.

Nowadays, due to enormous volumes of data, new data analytics methods are being investigated. A worth-mentioned category is **visual analytics** which aims to Synthesize information and derive insight from massive, dynamic, ambiguous, and often conflicting data. Research in visual analytics is very diverse and combines a number of related fields, including statistics, data fusion, data management, visualization, data mining, and cognitive science (among others). There is more to visual analytics than just visualization. Instead, it can be thought of as a comprehensive strategy for making decisions that integrates data analysis, human factors, and visualization. The difficulty lies in determining the best-automated algorithm for the analysis task at hand, identifying its limits where further automation is not possible, and then developing a tightly integrated solution that effectively integrates the best-automated analysis algorithms with suitable visualization and interaction techniques. Although some of this research has been done in the visualization field in the past, there hasn't been much application of sophisticated knowledge discovery techniques. To radically alter that is the goal of visual analytics. This will assist in keeping the emphasis on the proper aspect of the issue and offer solutions to issues that we have previously been unable to address.

1.2.1 Data analytics techniques

Data analysts can process data and extract information using a variety of analytical methodologies and techniques. Below is a list of some of the most widely used techniques. [4]

- **Regression analysis** comprises examining the relationship between dependent variables to determine whether a change in one may have an impact on a change in another.
- In order to perform **factor analysis**, a huge data collection must be reduced to a smaller data set. This strategy is to look for hidden trends that would have been more challenging to spot otherwise.
- The practice of dividing a data collection into groups of related data is known as **cohort analysis**. This enables data analysts and other data analytics users to go deeper into the statistics relevant to certain subsets of data.
- **Monte Carlo simulations** simulate the likelihood that various events will occur. These simulations, which frequently include many values and variables and frequently have better-predicting abilities than other data analytics techniques, are frequently utilized for risk reduction and loss prevention.
- **Time series analysis** examines data across time and establishes a connection between a data point's value and its occurrence. This method of data analysis is frequently employed to identify cyclical patterns or to forecast financial outcomes.

As illustrated in Figure 1 below, before moving to more detailed data analysis after the data structure, a data preprocessing framework shall be implemented. This aims to improve the quality of the data by removing outliers, handling missing values, etc. In the next paragraph, we will discuss more details about the Data Preprocessing framework.

Figure 1: Data preparation process bar.

1.3 Data preprocessing

Data preprocessing, a part of data preparation, refers to any type of processing done on raw data to get it ready for another data processing technique. It has historically been a crucial first stage in the data mining process. Data preprocessing approaches have lately been modified for the training of AI and machine learning models as well as for gaining insights about them. Data mining, machine learning, and other data science tasks can more quickly and effectively analyze data that has undergone data preprocessing. In order to get correct results, the techniques are typically applied early in the machine learning and AI development pipeline.[5]

Real-world data is unorganized and frequently generated, processed, and stored by a variety of people, business procedures, and software programs. A data set may therefore be incomplete, contain manual input errors, contain duplicate data, or use several names to represent the same thing. That's why data used to train machine learning or deep learning algorithms it's important to be automatically preprocessed.

The main steps in data preprocessing are described as follows:

- **Data profiling**: Data profiling is the process of looking at, evaluating, and reviewing data to gather statistics on its quality. It begins with an analysis of the qualities of the currently available data. Data scientists find the data sets that are relevant to the issue at hand, list their important characteristics, and make a hypothesis about which properties would be pertinent for the suggested analytics or machine learning assignment. Additionally, they think about which preprocessing libraries might be employed and tie data sources to the necessary business principles.
- **Data cleansing:** Here, the goal is to identify the simplest way to address quality issues, such as removing inaccurate data, completing data gaps, or generally ensuring that the raw data is appropriate for feature engineering.
- **Data transformation**: Here, data scientists consider how various data elements might be arranged to best serve the objectives. This may entail actions like structuring unstructured data, grouping significant variables where it makes sense, or choosing crucial ranges to focus on.
- **Data integration:** the process of putting together data from diverse sources into a single, unified view for effective data management, acquiring insightful understanding, and making effective decisions.
- **Data compression:** Raw data collections typically contain redundant data that results from categorizing occurrences in many ways, as well as data that is unrelated to a certain ML, AI, or analytics application. Principal component analysis and other data reduction techniques are used to simplify the raw data so that it is more acceptable for specific use cases.

- **Data validation:** Data validation refers to the process of making sure that data is accurate and of high quality. In order to assure the logical consistency of input and stored data, it is implemented by including a number of checks in a system or report.

The main techniques of data preprocessing are summarized in Figure 2.

Figure 2: Data preprocessing techniques.

1.4 Purpose and study structure

Data availability in shipping is continuously increasing. However, fleet managers and marine superintendents often fail to identify complex data patterns and gain useful insights from the available data. It is also challenging for them to integrate data analysis into the decision-making process. The neglected area of data quality for the ship's operation and performance analysis is the main reason for this issue, and little attention has been given to this matter in the scientific literature. Even though there are many available studies in detailed data analysis with complex statistical and machine-learning models, little attention has been given to the prepossessing of the available data prior to their analysis. This study aims to investigate the available data preprocessing techniques, with the main goal of improving the quality of the available data before their analysis.

Most scientific research concerning ship performance analysis mainly focuses on ship performance quantification through consecutive time intervals, neglecting the effect that other operational variables, such as weather or loading conditions, may have on the ship's performance. Additionally, the presented study aims to quantify the ship's performance under specific localized operational conditions.

Domain knowledge is incorporated into every step of the presented study since the little emphasis given to the existing literature failed to highlight its importance in data preprocessing and ship performance analysis.

To sum up, the overall purpose of the diploma thesis is to present a data preprocessing framework for ship performance quantification under localized operational conditions concerning domain knowledge. The proposed framework is capable of: a) detecting and isolating existing data anomalies in the available data set, b) investigating the ship's localized operational profile, and c) measuring the ship's performance under specific localized operational conditions.

The thesis structure is organized as follows. In the first chapter, an introduction to data analysis and data preprocessing in the maritime industry is given, followed by a brief literature review in chapter 2. An in-depth analysis of the proposed methodology is presented in chapter 3, accompanied by the corresponding mathematical equations and the constructed algorithm flowchart. The results of the suggested methodology are reviewed in chapter 4, along with the available interpretation of the corresponding outcome in each step of the proposed framework. The final conclusions are summed up in chapter 5.

2. Literature review

In Perera and Mo's publication [6], a marine engine-centered data analytics framework is presented. Gaussian mixture models appliance is collectively proposed as part of the framework with the expectation-maximization algorithm (EM). This study includes the data set's three parameters (Shaft speed, ME power, and fuel consumption). Data points that stand for slow-moving conditions are extracted from data analysis. Statistical analysis is used for the identification of engine propeller operating regions. The combinator diagram is used in considering engine-propulsion interactions. ME power and Shaft speed variables are used for developing the combinator diagram in a high-dimensional space concerning other navigation variables.

Also, Perera and Mo [7] proposed a structure for identifying ship power performance under relative wind conditions using statistical data analysis and visualization approaches. The selective data are investigated with the purpose of data outliers and data anomalies detection. These anomalies are isolated and cleaned to further measure vessel performance and navigation conditions. For predicting accurate ship speed and motion conditions along a voyage, relative wind profiles along the shipping route and ship model tests are proposed. Also, multiple visualization methods are presented in this study, targeting the extraction of applicable information for ship performance quantification and data anomaly detection. Statistical data analysis is presented with the ship's main navigation parameters. The combined plots of the individual parameters are utilized for pattern recognition throughout each data set and form an effective technique for data cleaning, resulting in a better representation of speed-power performance.

In their research [8], Perera presented a framework for sensor and data acquisition (DAQ) fault detection based on statistical filters. An analysis of the principal components is used to determine hidden paths in the data set and anomaly values in DAQ concerning a particular operating region. Perera reports that a model learning approach can efficiently deal with large data sets. The respective principal components are presented in descending order. The most important information is shown on the top PCs, and the least important information is on the bottom PCs. In this study, two parameters monitor condition is presented. Firstly, the respective thresholds used for sensor and DAQ fault identification are discussed. Then, the PCA method is implemented for the title of the sensor and DAQ faults detection, which are within the thresholds. Then, the data set is presented with two new parameters based on two principal components. It's worth noting that the bottom PC identifies all fault situations. Ship performance and navigation data cluster are further analyzed by considering the vessel's main engine. A normalized data set consists of ten parameters associated with ship performance and navigational data. Statistical distributions are used to present the PCs derived from PCA. The value of three standard deviations (3σ) is defined as the threshold limit, so every data point beyond this range is considered an outlier. Sensor and DAQ faults are subdivided from A to S windows for a more detailed investigation. The parameter variation under each fault window and the separate filter that each fault situation is noted are presented. Results are summarized in a table that displays the filter numbers versus the fault windows.

Dalheim and Steen [9] proposed a data preparation framework for ship operation and performance investigation. Therefore, this paper offers a structure for unifying the maximum physical interpretable data preprocessing strategies that are easy to use, identifies the order in which the respective method should be implemented, and logically connects all relevant

facts. Feature selection is the first stage in the proposed data preparation framework. As the next step, time vector jumps, and synchronization should be considered. A mathematical approach for coping with time synchronization is also presented. The primary purpose of this approach is to select the most representative time reference for the whole data set. It also ensures that the final time reference tries to obtain maximum overlap with its parent time vector. Finally, a brief introduction to signal synchronization when the time vector is unavailable is presented. An outlier detection method is also presented in this study, and the outlier detection method is separated into four blocks, each used to identify a specific type of outlier. In the first block, domain knowledge is implemented. In the second block, repeated values are determined. And in the third and fourth blocks, dropouts and spikes are respectively placed. Data validation is the last step in the presented data preparation framework. Using this method ensures that erroneous data are removed from ship performance analysis. Two methods for data extraction are presented in this study. The first one is port-to-port trips to avoid low-speed maneuvering conditions in the port. The second method involves a steady-state detector to identify stationary portions of the in-service data measurement.

3. Methodology

The proposed framework is aimed at preprocessing and analyzing data effectively to monitor ship performance under specific operational conditions. Domain knowledge has been incorporated into every step of this process. The techniques that have been examined can be briefly described as follows: Data pattern recognition, Data Clustering, Outlier detection, Visual analytics, and Ship performance analysis.

A flowchart illustrating the steps that must be implemented to achieve the final goal of ship performance quantification under localized operational conditions is provided in Figure 3.

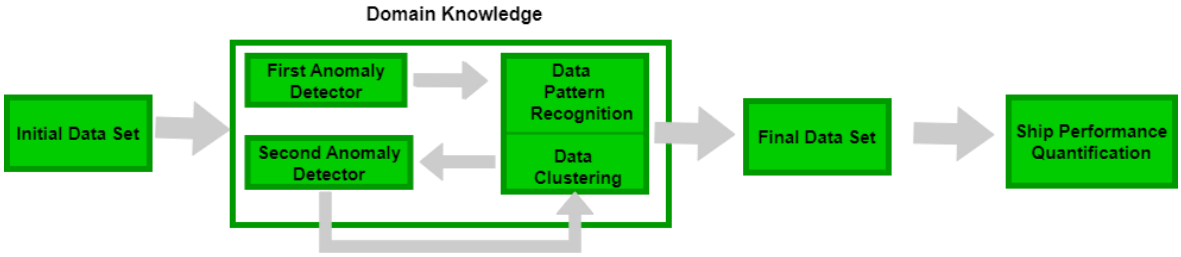

Figure 3: Abstract flowchart of the proposed framework.

3.1 Domain knowledge

It is crucial to pinpoint the importance of domain knowledge in our investigation. In data science, the term domain knowledge refers to the general background knowledge of the field or environment in which data science methods are applied. Our case involves knowing the ship's operational profile. Every vessel operates in a specific range, so knowing our ship's behavior can help us better understand the outcomes of our implemented Statistics and machine-learning models. For example, the ship's speed and propeller shaft power are non-linearly related. The total ship's resistance is proportional to the square of the ship's velocity, and the propeller's shaft power is proportional to the cube of the propeller's shaft speed (propeller law). So, domain knowledge guide and intervenes in the data preparation process when necessary. Domain knowledge is required to specify the minimum and maximum values of parameters. These numbers represent the parameters' typical range of values. Whenever the data points exceed the minimum and maximum thresholds, data anomalies are detected and eliminated.

3.2 Data pattern recognition

Various data density estimation methods are explored in this framework. For example, Histograms, scatter plots, density scatter plots, and the interesting case of the Kernel Density Estimation method, which is based on the Probability Density Function. Examples of various data estimation methods and plots are presented in Figure 4. The aforementioned methods are assigned in the research literature as exploratory data analysis. These techniques aim to identify structures and regularities in data, which can later be classified based on statistical information or knowledge gained from patterns and their representation.

3.2.1 Histograms

A frequency distribution shows how often each different value in a set of data occurs. A histogram is the most used graph to show frequency distributions. There are many types of histograms. We mainly use bivariate histograms, which allow us to group data in 2-d bins.

3.2.2 Scatter plots

In a scatter plot, dots are used to show the values of two different numerical variables. The placement of each dot on the horizontal and vertical axes indicates an individual data point's values. To see how other variables relate to one another, utilize scatter plots. Finding different data patterns can be done using a scatter plot. We can categorize data points into groups based on how closely sets of points cluster together. If there are any unexpected gaps in the data or any outlier points, scatter plots might also expose them.

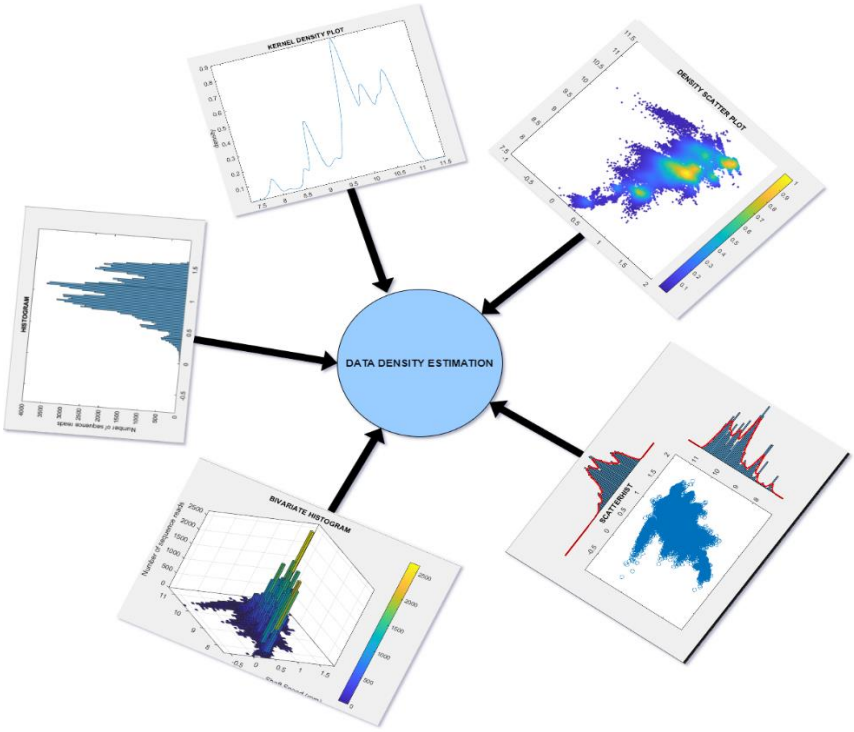

Figure 4: Data density estimation plots.

3.2.3 Density scatter plots

A sort of two-dimensional histogram that displays the number of points in each plot section is called a density scatterplot. It is mainly used when the plotted data in the scatter plot are too dense to get a good impression of the distribution of the data.

3.2.4 Kernel Density Estimation method

The Kernel Density Estimation is a mathematical process for estimating the probability density function of a random variable. It produces a smooth empirical pdf based on individual locations of all sample data. Such a pdf estimate seems to better represent the "true" pdf of a continuous variable. In kernel estimating, two ideas are fundamental: kernel function form and coefficient of smoothness, the latter of which is essential to the approach. In this application, Kernel Density Estimation is used to gain insights into the properties of the data. Let (x1, x2, ..., xn) be independent and identically distributed samples drawn from some univariate distribution with an unknown density ƒ at any given point x. We are interested in estimating the shape of this function ƒ. Its kernel density estimator is

$$\mathrm{f_h}(\mathrm{x}) = \frac{1}{h}\sum_{i=1}^{n} k_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x - x_i}{h}\right) \qquad (3.1)$$

where K is the kernel — a non-negative function — and $h > 0$ is an explored parameter called the bandwidth, $K(x) = \phi(x)$, where ϕ is the standard normal density function. In principle, the Kernel can be any valid probability density function but the usual choice is the gaussian one. [10]

An example of a kernel density plot along with the individual gaussian kernels is presented in Figure 5.

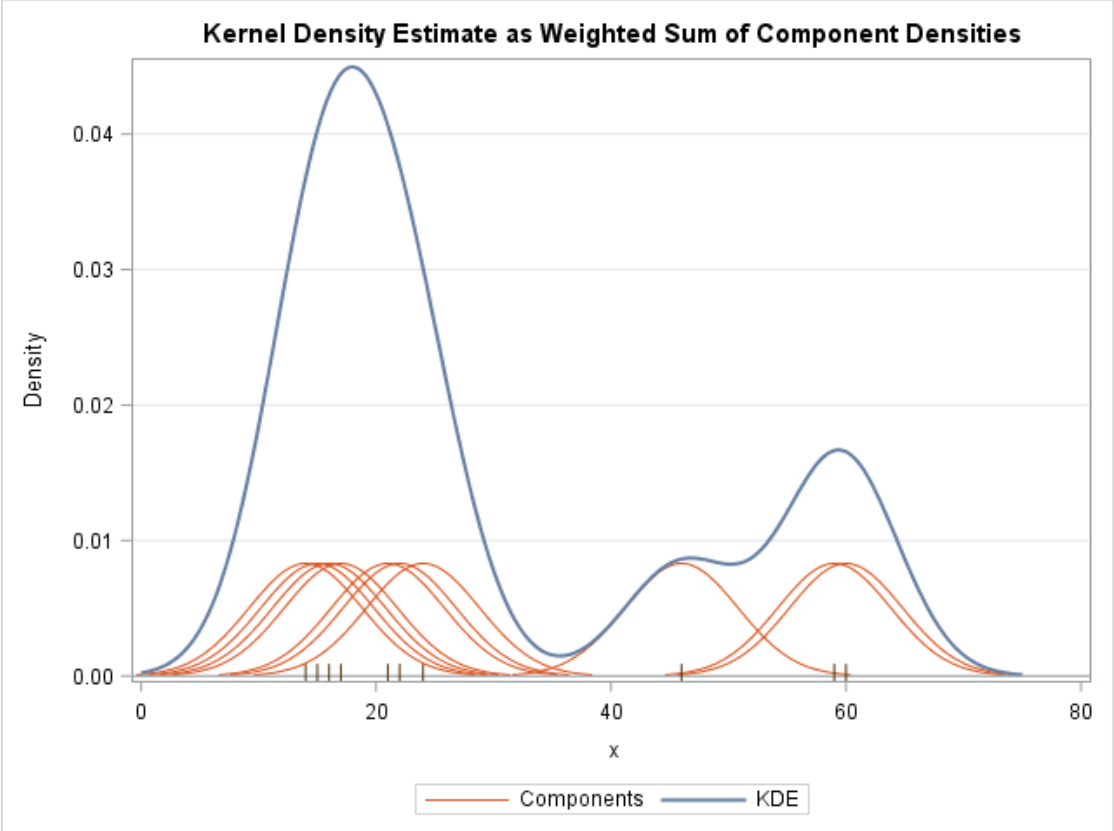

Figure 5: Kernel Density Estimation Plot. [11]

3.3 Data clustering

The goal of cluster analysis or clustering is to organize a collection of objects into groups that are more similar (in some ways) to one another than to objects in other groups (clusters). Clustering only required the data set to have data points without being provided with the labels (unsupervised learning). Clustering is one of the most common exploratory data analysis techniques used for pattern recognition and will provide us with useful findings about our data's behavior. A clustering example is presented in Figure 6.

3.3.1 Types of clustering

In general, there are two clustering types:

- **Hard Clustering**: Data points in hard clustering either belong to a cluster completely or not.
- **Soft Clustering**: In soft clustering, a probability or likelihood of each data point being in those clusters is assigned rather than placing each data point into a separate cluster.

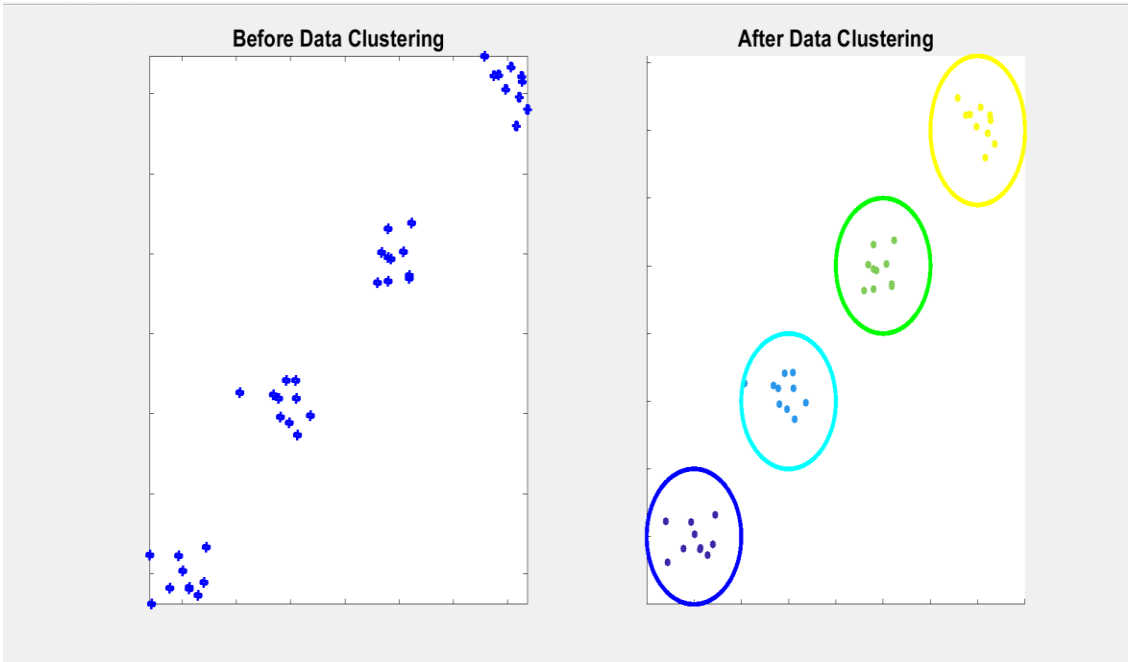

Figure 6: An example of a data set before clustering and after clustering.

3.3.2 Types of clustering algorithms

Connectivity models: These models, as their names imply, are based on the idea that data points that are closer to one another in a data space show greater similarity than those that are farther apart. There are two possible approaches for these models. In the first method, they begin by grouping every data point into a different cluster and then aggregate them as the distance grows less. The second method partitions the data as the distance grows after classifying all the data points into a single cluster. Choosing a distance function is also an individual decision. Although fairly simple to understand, these models cannot handle large datasets due to their lack of scalability. These models include the hierarchical clustering technique and its variations. [12]

Centroid models: These techniques for iterative clustering get the idea of similarity from how near a data point is to the centroid of the clusters. One well-known algorithm that fits this description is the k-means clustering algorithm. It is crucial to have prior knowledge about the dataset in these models since the number of clusters needed at the conclusion must be specified beforehand. In order to locate the local optimum, these models run iteratively.[12]

Distribution models: The idea behind these clustering models is how likely it is for all the data points in the cluster to belong to the same distribution (For example: Gaussian). Overfitting affects these models frequently. Gaussian Mixture Model and the Expectation-Maximization algorithm, which employs multivariate normal distributions, are well-known examples of these models. [12]

Density models: These models look for regions in the data space where there is a variety in the density of data points. The data points are assigned to the same cluster after the system isolates different density zones. DBSCAN and OPTICS are common examples of density models. [12]

3.3.3 Investigated clustering methods

In this framework, two data clustering techniques are proposed, the K-Means algorithm and Gaussian Mixture Model. Although K-Means is an easy and quick clustering method, it might not accurately represent the heterogeneity in the data set. Complex patterns can be found using Gaussian Mixture Models, which can then be sorted into cohesive, homogenous components that closely resemble the data set's actual patterns. Therefore, it is of actual interest to implement both algorithms and evaluate their results based on similarities and differences. By using K-Means and GMM's clustering methods, Ship's localized operational profiles are further investigated. A key step in this process is clustering the examined data by engine modes, utilizing domain knowledge and data density estimation plots. The identified engine data clusters are further divided into sub-clusters regarding the identified trim-draft modes of each cluster. The spotted sub-clusters represent the ship’s localized operational conditions. They are particularly useful for quantifying the ship's performance under its localized operational conditions, which can provide many insights into its operational profile. [12]

3.3.4 K-MEANS Algorithm

The k-Means algorithm establishes the presence of clusters by finding their centroid points. A centroid point is the average of all the data points in the cluster. By iteratively assessing the Euclidean distance between each point in the dataset, each one can be assigned to a cluster. The centroid points are random, to begin with, and will change each time as the process is carried out. K-means is commonly used in cluster analysis, but it has a limitation in being mainly useful for scalar data.[13], [14]

K-Means Algorithm [14]:

1. Indicate K, the number of clusters.

2. Choose K random data points for the centroids without replacing them after shuffling the dataset.

3. Continue iterating until the centroids remain unchanged. In other words, the clusters to which the data points are assigned remain the same.

- Calculate the total of the squared distances between each centroid and the data points.
- Assign each data point to the nearest cluster (centroid).
- Calculate the centroids for each cluster by averaging all the data points that make up that cluster.

K-Means uses the Expectation-Maximization strategy to resolve the iteration problem. The data points are assigned to the closest cluster in the E-step. The centroid of each cluster is calculated in the M-step. Here is how we can resolve it mathematically, in detail. [13], [14]

The objective function is:

$$J = \sum_{i=1}^{m} \sum_{k=1}^{K} w_{ik} \left\| x^i - \mu_k \right\|^2 \quad (3.2)$$

Where $w_{ik} = 1$ for data point x^i if it belongs to cluster *k*, otherwise, $w_{ik} = 0$. Also, μ_k is the centroid of x^i's cluster.

It's a two-part minimization problem. First, we treat μ_k as fixed and minimize J with respect to w_{ik}. Then, we assume w_{ik} as fixed and minimize J with respect to μ_k. Technically, we differentiate J regarding w_{ik} first and update cluster assignments (E-step). After the cluster assignments from the previous phase, we differentiate J with respect to μ_k and recompute the centroids (M-step). As a result, E-step is:

$$\frac{\vartheta J}{\vartheta w_{ik}} = \sum_{i=1}^{m} \sum_{k=1}^{K} w_{ik} \left\| x^i - \mu_k \right\|^2 \quad (3.3)$$

$$\rightarrow w_{ik} = \begin{cases} 1 & if\ k = argmin_j \left\| x^i - \mu_j \right\|^2 \\ 0 & otherwise. \end{cases} \quad (3.4)$$

Assign the data point x^i to the cluster that is closest to it, as determined by its sum of squared distances from its centroid.

And M-step is:

$$\frac{\vartheta j}{\vartheta \mu_\kappa} = 2 \sum_{i=1}^{m} w_{ik} \left(\chi^i - \mu_\kappa \right) = 0 \quad (3.5)$$

$$\rightarrow \mu_\kappa = \frac{\sum_{i=1}^{m} w_{ik} x^i}{\sum_{i=1}^{m} w_{ik}} \quad (3.6)$$

Which corresponds to recalculating each cluster's centroid to account for the new assignments.

3.3.5 Gaussian Mixture Models

Gaussian mixture models are a type of machine learning algorithm that is commonly used in data science. It is also a probabilistic model that assumes all the data points are generated from a mix of Gaussian distributions with unknown parameters. A Gaussian mixture model can be used for clustering. GMMs can be used to find clusters in data sets that may not be clearly defined.[13], [15]

The Gaussian distribution of a d-dimensional vector x is defined as :

$$N(\chi/\mu, \Sigma) = \frac{1}{\sqrt{2\pi^{\,d}\,|\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(\chi-\mu)\right) \quad (3.7)$$

where μ is a mean vector and Σ is a covariance matrix.

The probability given in a mixture of K Gaussians is defined as:

$$p(x)=\sum_{K=1}^{K} \pi_\kappa N(X / \mu_\kappa, \Sigma_\kappa) \quad (3.8)$$

where each Gaussian density $N(\chi/\mu_\kappa, \Sigma_\kappa)$ is called a component of the mixture with its mean vector μ_κ and covariance Σ_κ for the k^{th} Gaussian component,π_κ is the prior probability of the k^{th} Gaussian, π_κ is also defined as the mixing coefficients with the constraint that $\sum_{\kappa=1}^{\kappa} \pi_\kappa$=1.(3.9)

Let us now illustrate these parameters graphically in Figure 7. In which three Gaussian functions are presented. Each one interprets the data contained in each of the three existing clusters. As mentioned before the mixing coefficients, are themselves probabilities and must meet the condition $\sum_{\kappa=1}^{\kappa} \pi_\kappa$=1.

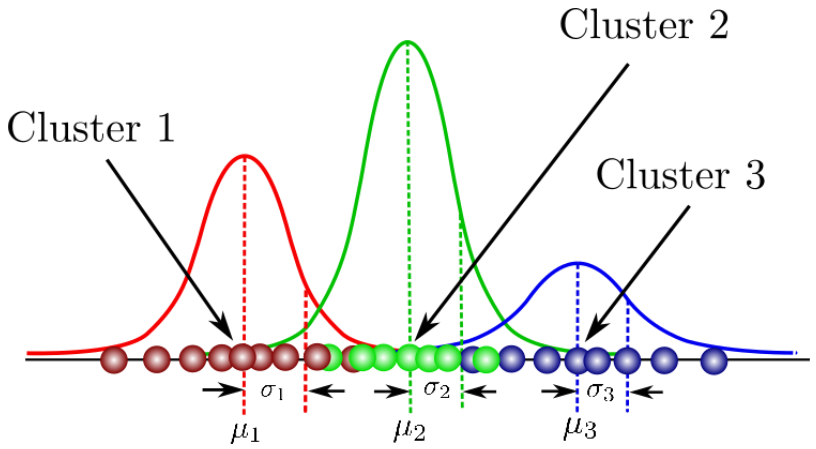

Figure 7: Gaussian mixture model parameters explained graphically. [16]

3.3.6 Expectation-Maximization algorithm

The **expectation-maximization algorithm** is an approach for performing maximum likelihood estimation in the presence of latent variables. It does this by first estimating the values for the latent variables, then optimizing the model, then repeating these two steps until convergence. It is an effective and general approach and is mostly used for density estimation with missing data, such as clustering algorithms like the Gaussian Mixture Model. [15], [16]

Step 1: Initialize μ_k, Σ_k, π_k, and evaluate the initial value of the Log-likelihood.

Step 2: (Expectation step): Use the current values for parameters to evaluate the posterior probabilities, or the responsibilities $\gamma(z_{nk})$ which is taken by component k for explaining the observation of data point x_n:

$$\gamma(z_{nk}) = \frac{\pi_\kappa N(X_n \mid \mu_\kappa, \Sigma_\kappa)}{\sum_{j=1}^{K} \pi_j N(X_n \mid \mu_j, \Sigma_j)} \tag{3.10}$$

Z is a latent variable that takes only two possible values. It is one when x came from Gaussian k, and zero otherwise.

Step 3: (Maximization step): Re-estimate the parameters using the current responsibilities:

$$\mu_\kappa^{new} = \frac{1}{N_K} \sum_{n=1}^{N} \gamma(Z_{nk}) x_n \tag{3.11}$$

$$\Sigma_\kappa^{new} = \frac{1}{N_K} \sum_{n=1}^{N} \gamma(Z_{nk}) (x_n - \mu_\kappa^{new}) (x_n - \mu_\kappa^{new})^T \tag{3.12}$$

$$\pi_\kappa^{new} = \frac{N_K}{N} \tag{3.13}$$

Where

$$N_\kappa = \sum_{n=1}^{N} \gamma(Z_{nk}) \tag{3.14}$$

N_k can be interpreted as the effective number of points assigned to cluster k.

Step 4: Evaluate the log-likelihood:

$$\ln p(X/\pi, \mu, \Sigma) = \sum_{n=1}^{N} ln \left(\sum_{k=1}^{K} \pi_k \, N(\chi_n / \mu_\kappa, \Sigma_\kappa) \right) \quad (3.15)$$

and check for convergence of either the parameters or the log-likelihood. If the convergence criterion is not satisfied, get back to Step 2.

3.3.7 Clustering evaluation criteria

As already stated, in order to perform clustering analysis with the K-means algorithm and GMM model, the number of clusters needs to be predefined. The best number of clusters in the specific application is determined based on data density plots and domain knowledge. The available literature has also proposed mathematical approaches for identifying the optimum number of clusters and evaluating the clustering results. Some of them will be implemented in the specific framework to examine the effectiveness of data density plots in identifying the optimum number of clusters.

Elbow method

The first method, which will be examined, is the most well-known method in cluster validation literature, and it's called the "elbow" method. The "elbow" method uses the sum of squared distance (SSE) to choose an ideal number of clusters (k) based on the distance between the data points and their assigned clusters. We would select a value of k where the SSE begins to flatten out, and we see an inflection point. When visualized, this graph would look somewhat like an elbow, hence the method's name.

An example of an elbow plot is presented in Figure 8:

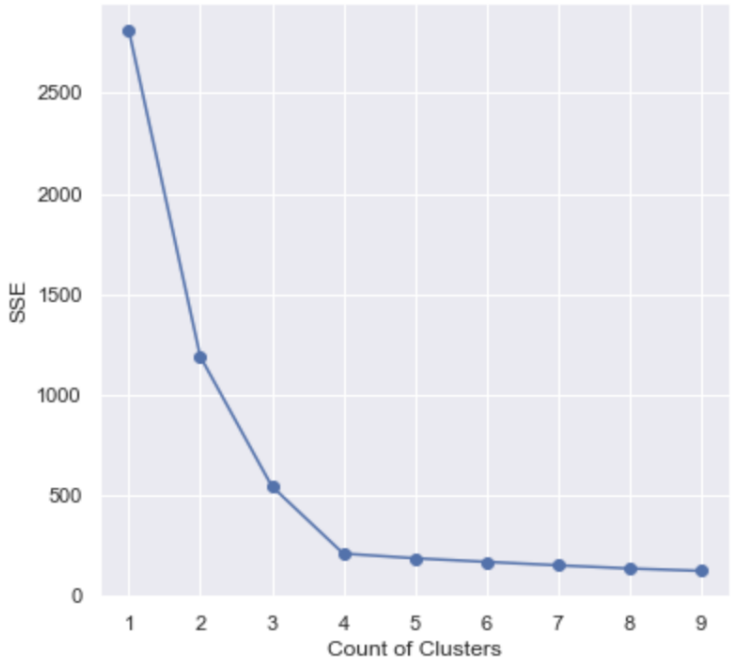

Figure 8: Elbow plot. [17]

This plot shows that the number of optimal clusters (k) is four. Initially, the SSE, within-cluster variance, decreases with an increase in cluster number. However, after a particular point, k=4, the SSE value starts flattening. So, there is no added value in increasing the number of clusters. Therefore, the number of clusters corresponding to that point, k=4, should be considered the optimal number of clusters. The ”Elbow” method is an easily applicable measure for identifying the optimum number of clusters, but sometimes it can be hard to interpret the plotted results. That's why it's considered a subjective measure of clustering validation. [17]

AIC/BIC Criterion

Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are considered to evaluate the clustering results of the implemented GMM model. The simplest way to choose the best model that fits the data is to compare all the competing models and select the one with the highest likelihood. However, the maximization of likelihood can lead to an overfitting of the model to the data with additional degrees of freedom. That's why a more robust and accurate criterion, which penalizes the use of extra free parameters, is needed. That can be achieved by using the Information criterion tests, such as Akaike Information Criteria (AIC) and the Bayesian Information Criteria (BIC), commonly used in Astrophysics literature. Information criteria are likelihood-based measurements of model fit with a penalty for complexity (specifically, the number of parameters). Different information criteria can favor various models and can be recognized by how the penalty is formed.

The AIC compares models from the perspective of information entropy, as measured by Kullback-Leibler divergence. The AIC for a given model is :

$$\mathrm{AIC}=2q - 2\ln(L) \tag{3.16}$$

The BIC compares models from the decision theory perspective, as measured by expected loss. The BIC for a given model is

$$\mathrm{BIC} = \ln(n)q - 2\ln(L) \tag{3.17}$$

Where n is the number of observations, q is the number of parameters learned by the model, and L is the maximized value of the likelihood function of the model. In both cases, a lower value denotes the better model.

Although information criteria penalize the models with a large number of clusters, there are many cases in which the greater the number of clusters, the lower the AIC/BIC value. So overfitting cannot always be avoided. In such cases, an additional technique is implemented to calculate the gradient of the AIC/BIC scores curve. Comparing the gradient values of the AIC/BIC score curve can reveal the optimum number of clusters. Such an example is given in Figure 9, in which we notice that the greater the number of clusters, the lower the BIC score. That can cause overfitting to the model.

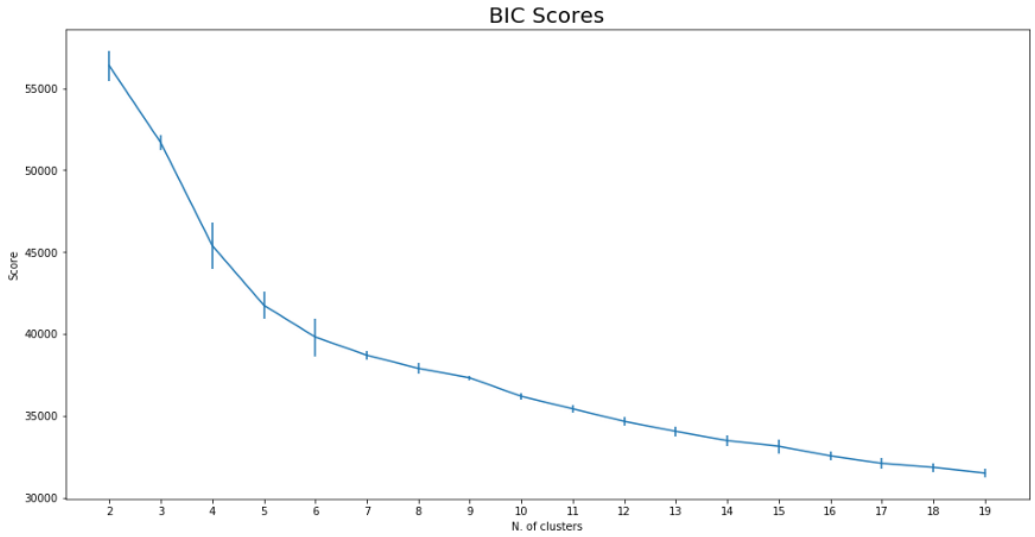

Figure 9: BIC score curve. ***[18]***

That's why the gradient value of the BIC score curve is calculated and plotted in Figure 10, and it can be noticed that after the cluster size of seven, the gradient becomes almost constant. So, the BIC scores in the original function are decreasing much gentler. That leads to the conclusion that there is little gain in increasing the number of clusters. So, according to this technique, the ideal number of clusters is six (6).

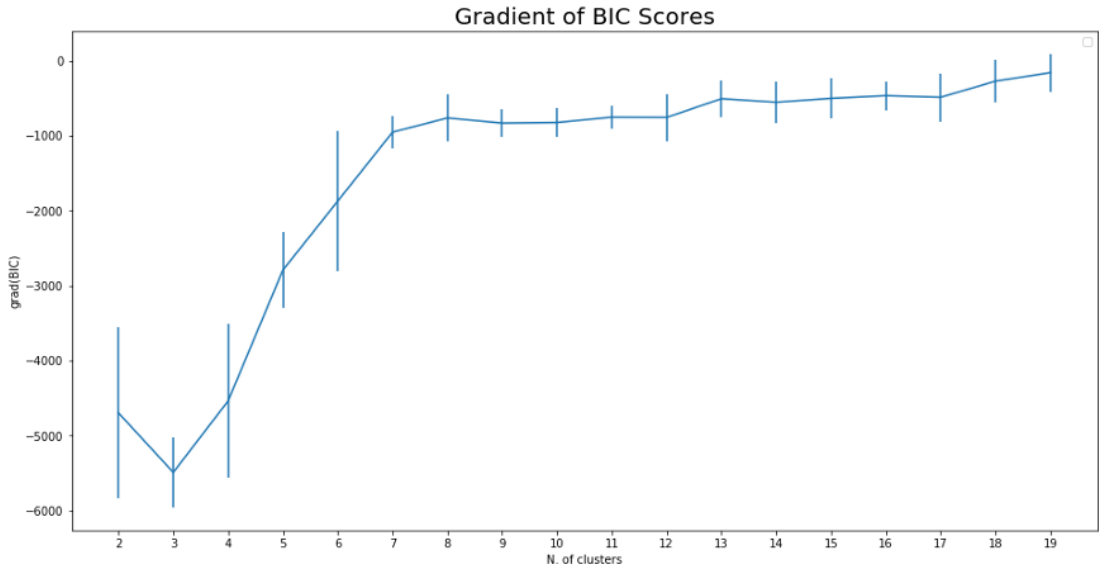

Figure 10: Gradient plot of the BIC scores. ***[18]***

3.4 Outlier detection

In actual data sets, it frequently occurs that some observations deviate from the norm. These observations are referred to as outliers. Observations that are outliers may be errors, or they may have been acquired under unusual circumstances. They, therefore, do not accurately suit the model. Understanding how to spot these outliers is crucial for the final quality of the data set. [19], [20]

3.4.1 Types of outliers

Relevant to machine learning models, there are three basic categories of outliers. Each type differs in terms of the ability to detect anomalous data and the characteristics that set the data point apart from the rest of the data collection. For outlier analysis, types are crucial since each type has a unique pattern to look for.

The three main types of outliers are:

- **Point outliers**

A point outlier is a single data point that falls beyond the dataset's normal operational range. Within the dataset, there may be a definite pattern, trend, or grouping; an outlier as a data point will be very distinct from this. Point outliers are frequently the result of a measurement or data entry error.

- **Contextual outliers**

A data point that significantly deviates from the dataset only when seen in a particular context is known as a contextual outlier. A dataset's context may change seasonally or shift in response to larger trends or human activity. When the context of the dataset changes, a contextual outlier will become apparent. This could be due to seasonal variations in the weather, the state of the economy, modifications in consumer behavior around important holidays, or simply the time of day. A contextual outlier may therefore appear to be a typical data point in other situations.

- **Collective outliers**

A group of data points that deviates significantly from the overall dataset's trends is referred to as a collective outlier. A collective outlier's individual data points might not appear to be outliers in terms of point or context. Anomaly patterns are only visible when the data points are seen as a group. Because of this, collective outliers may be the most challenging kind of outlier to spot. Collective outliers play a crucial role in machine learning's notion of drift monitoring. A series of data has deviated from the model's predicted behavior.

There are machine learning methods for a wide variety of applications. The type of data and potential outlier will change depending on the model, whether it is trained to cluster engine data or to identify operational regions with the lowest fuel oil consumption. Outlier detection is fundamentally divided into three broad categories based on how outliers are detected.

3.4.2 Outlier detection methods

- **Statistical Methods**

Finding the extreme values in the data can be done by simply beginning with a visual study of the Univariate data using Boxplots, Scatter plots, Whisker plots, etc. Calculate the z-score, which represents how far away from the sample mean the standard deviation (σ) times the data point is, presuming a normal distribution. We can recognize data points that are more than three times the standard deviation as outliers because we know from the Empirical Rule that 68% of the data falls within one standard deviation, 95% percent within two standard deviations, and 99.7% percent within three standard deviations from the mean. Another approach would be to treat outliers that are more than 1.5 times the first or third interquartile range (IQR) as a criterion.

- **Proximity Methods**

Proximity-based methods utilize clustering techniques to locate each cluster's centroid and identify the clusters in the data. If an object's closest neighbors are far away from it in feature space, deviating noticeably from the proximity of the majority of the other objects to their neighbors in the same data set, they are assumed to be outliers. The typical method is as follows: Set a threshold, measure each data point's separation from the cluster centroid, then exclude any outliers before continuing with the modeling. As obvious as it may seem, the metric chosen to measure distance has a significant impact on the effectiveness of these models. The disadvantages include the possibility of difficulty in determining the appropriate distance measure for some particular problem types. Another issue is that it is less reliable when the outliers are clustered together. Two categories of proximity-based approaches are recognized: Data points are evaluated using distance-based approaches based on how far they are from their neighbors. The density-based approach interprets the behavior of data groups based on their local density. Proximity-based outlier detection approaches include DBScan, k-means, and hierarchical clustering.

- **Projection Methods**

By exploiting linear correlations, projection methods map the data into a lower-dimensional subspace using techniques like Principal Component Analysis. Post that, the distance of each data point to a plane that fits the sub-space is calculated. Then, the outliers can be identified using this distance. Methods for projecting values are intuitive, simple to use, and can draw attention to minor values. The PCA-based method analyzes the features that are currently accessible to decide what constitutes a "normal" class. The module then applies distance metrics to identify cases that represent anomalies.

3.4.3 Common outlier causes

Machine learning algorithms and models are trained using a variety of several sorts of data. Particularly if data needs to be prepared and labeled, as, in supervised machine learning, a human mistake can frequently be the cause of outliers. However, there can be outliers as a result of measurement or data extraction problems in all kinds of datasets and machine learning use cases.

Common causes of outliers in machine learning include:

- Human error when entering or labeling data.
- Errors in measuring or collecting the data.
- Errors in data extraction, processing, or manipulation.
- Man-made outliers for testing outlier detection processes.
- Natural occurrences of outliers that aren't errors, which can be called dataset novelties.

3.4.4 Challenges of outlier detection

1. Effective Identification

An outlier can be defined in a very specific way, depending on the application scenario and the domain. Often, the difference between normal observations and outliers is quite narrow, and even a little misinformation can result in an outlier being treated as a regular observation. As a result, we must be extremely careful while choosing the outlier identification approach to handle the outliers.

2. Application-Specific Challenges

As previously said, selecting the similarity or distance metric as well as the relationship model to represent data objects is crucial for outlier detection. Sadly, they are frequently application-specific. As an illustration, datasets from the medical industry may have outliers that are even slightly off from the rest of the dataset. Different applications may have quite different needs. Consequently, it is necessary to create specialized outlier identification techniques for a particular application.

3. Handling Noise

Noise in the data tends to be similar to the actual outliers and hence is difficult to distinguish and remove them from harmful outliers. Since noise in the data often resembles true outliers, it can be challenging to separate it from harmful outliers. We must comprehend that outliers and noise are two distinct things that differ from one another. Additionally, because noise can frequently and clearly be present in all types of data collected, it can pose several difficulties for outlier detection by distorting the distinction between regular observations and outliers. Outlier objects are hidden by noise, which reduces the effectiveness of the method used to find them.

3.4.5 Outlier detection method selection

Effective outlier detection can be a real challenge in many real-world applications. This can happen due to various types of outliers that are difficult to identify and due to the complexity of the data, making it harder to distinguish regular data points from outliers. However, many methods are available to deal with any outlier point. The technique used in this case for outlier detection is Principal Component Analysis (PCA). This simple method can be applied in a machine learning model and is effective in identifying complex outliers.

3.4.6 Principal Component Analysis

Principal component analysis (PCA) is a technique for reducing the dimensionality of large datasets, increasing interpretability while minimizing information loss. It does so by creating new uncorrelated variables that successively maximize variance.

Accordingly, "preserving as much variability as possible" entails identifying new uncorrelated variables, linear functions of the original dataset's variables, and successively maximizing variance. Solving an eigenvalue/eigenvector problem is how the principal components (PCs), these new variables, are identified.

PCA implementation is essential in our data set because smaller data sets are easier to explore and visualize and make analyzing data much easier and faster for machine learning algorithms without extraneous variables to process. [21]

PCA Algorithm

- **Standardization**

This stage standardizes the range of the continuous initial variables with the intention of ensuring that each one contributes equally to the analysis. Standardization must be done before PCA because the latter is very sensitive to variations in the initial variables. That is, if there are significant differences in the initial variable ranges, the variable with the larger range will take precedence over the variable with the smaller range.

Mathematically, this can be done by subtracting the mean and dividing it by the standard deviation for each value of each variable.

$$Z = \frac{value - mean}{standard\ deviation}$$

Once the standardization is done, all the variables will be transformed to the same scale.

- **Covariance matrix computation**

In other words, this step's goal is to determine whether there is any link between the variables in the input data set and how they differ from the mean relative to one another. Because variables can occasionally be highly connected in a way that causes them to contain redundant information. Therefore, we compute the covariance matrix to find these relationships.

- **Compute the eigenvectors and eigenvalues of the covariance matrix to identify the principal components**

Principal components are new variables constructed as linear combinations or mixtures of the initial variables. As a result of these combinations, the new variables (i.e., principal components) are uncorrelated, and most of the information in the initial variables is squeezed into the first components. Principal components are the lines that, geometrically speaking, encompass most of the information in the data and reflect the directions of the data that account for the greatest amount of variance. The eigenvectors are computed and sorted by

their eigenvalues in descending order to find significant principal components. Simply defined, principal components are new axes that offer the greatest perspective for seeing and analyzing the data, making the differences between the observations easier to see.

- **Feature vector**

In this step, we decide whether to keep all of these components or discard any that have low eigenvalues and create a matrix of vectors that we refer to as the "Feature vector" using the ones that are left. Therefore, the feature vector is just a matrix with the eigenvectors of the components that we choose to maintain as columns. This makes it the first step towards dimensionality reduction since the final data set will only have p dimensions if we decide to keep only p eigenvectors (components) out of n.

- **Recast the data along the principal components' axes**

The goal of this final step is to reorient the data from the original axes to those represented by the principal components using the feature vector created using the eigenvectors of the covariance matrix. This can be done by multiplying the transpose of the original data set by the transpose of the feature vector.

The steps of the Principal Component analysis are summarized in

Figure 11:

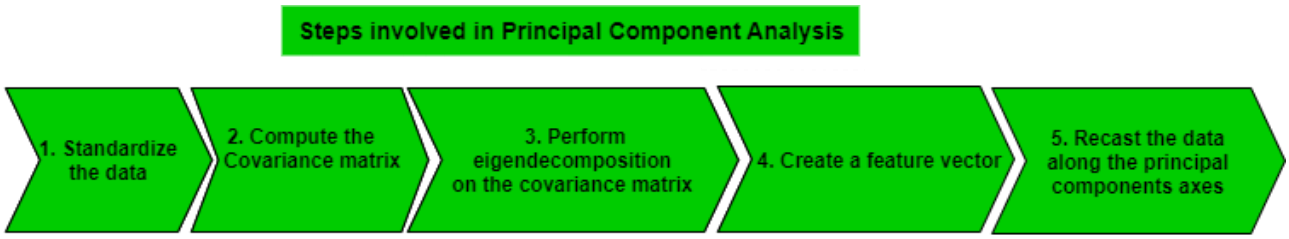

Figure 11: Steps involved in Principal Component Analysis.

3.5 Visual analytics

Visual analytics combines automated analysis techniques with interactive visualizations for a practical understanding, reasoning, and decision-making based on extensive and complex data sets.

The goal of visual analytics is the creation of tools and techniques to enable people to:

- Synthesize information and derive insight from massive, dynamic, ambiguous, and often conflicting data.
- Detect the expected and discover the unexpected.
- Provide timely, defensible, and understandable assessments.
- Communicate assessment effectively for action.

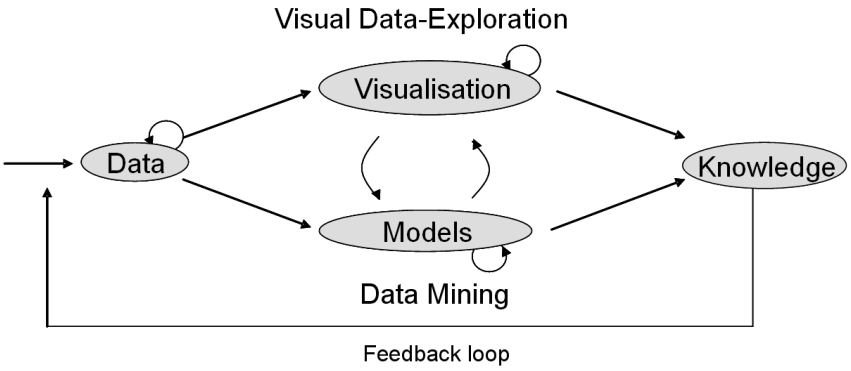

Figure 12: Tight integration of visual and automatic data analysis methods with database technology for scalable interactive decision support. [22]

Visual analytics is more than just visualization. It can rather be seen as an integral approach to decision-making, combining visualization, human factors, and data analysis. A challenge lies in determining the best-automated algorithm for the analysis task at hand, identifying its limitations that cannot be further automated, and developing an integrated solution that integrates the best-automated analysis algorithms with appropriate interaction and visualization techniques. Visual analytics can help us better understand our data. By visualizing the enhanced data, we can determine the relationships or correlations among ship performance and navigation characteristics under localized operational conditions. MATLAB has extended capabilities for data visualization and analysis. This analysis uses interactive plots and more sophisticated charts to extract as much information as possible.[22]

3.6 Ship performance quantification

Evaluating the Ship's performance requires a good amount of qualitative data obtained in our case in the data preprocessing framework as mentioned above. The interpreted results of ship performance analysis can be a practical guide for the Ship's captain, crew, and operators. For quantifying a ship's performance considering localized operational conditions, two selected key performance indicators (KPIs) are proposed. The proposed KPIs are calculated in each subcluster (trim-draft) mode concerning the identified clusters (engine modes).

The resulting KPI for ship performance quantification can be expressed as:

- $KPIa_i = \frac{P_i}{n_i^3}$: It corresponds to the propeller curve coefficient. When the KPI's value decreases, the vessel's performance increases since less engine power is required to maintain constant shaft revolutions, and greater rpm can be achieved for the same level of engine power.

Where: P_i, is the propeller's shaft power [KW] and n_i is the propeller's shaft speed [RPM].
i corresponds to the ship's localized operational conditions.

- $KPIb_i = \frac{FC_i}{D_i}$: Is the representation of the ship's main engine fuel consumption per nautical mile. When the KPI's value decreases, the vessel's performance increases since less fuel oil is consumed for the same traveled distance, and greater distance is covered for the same fuel oil consumption.

Where: $FC_i = FC_{avg,i} \times t_i$ [Ton/day], $D_i = SOG_{avg,i} \times t_i$ [NM].

FC_i is the main engine (ME) fuel consumption (cons) [ton], $FC_{avg,\,i}$ is the average ME fuel cons [ton/day], D_i is the traveled distance [NM], t_i is the time traveled [day], and $SOG_{avg,I}$ is the average speed over ground (SOG) [NM/h] under the respective localized operational condition i, correspondingly. For the sake of unit, consistency can be rewritten as follows.

$$\mathrm{KPIb_i} = \frac{FC_{avg,i}}{24\ SOG_{avg,i}}$$

3.7 Presentation of the calculation framework

After presenting all the methods which are investigated in this analysis, a brief introduction to the structure of the proposed algorithm is going to be given.

Once the investigation vessel's operational data has been collected, the first anomaly detector is applied. The first anomaly detector defines the minimum-maximum values that each investigated parameter can obtain. The used thresholds are based on domain knowledge, specifically in identifying the physical range of the investigated parameters and the ship's operational limitations. Additional limitations are imposed in the propeller shaft speed and power parameters to exclude measurements that apply to the ship's maneuvering conditions from the investigating data set. Afterward, the data density plots are utilized to identify regularities and patterns in the engine data, i.e. (propeller shaft speed and power). K-means algorithm and Gaussian Mixture model clustering methods are applied in the engine data set. The number of clusters needs to be predefined in both clustering methods. The optimum number of clusters is selected based on domain knowledge and the implemented data density plots. Note that the "Elbow" method and AIC/BIC information criteria are utilized to evaluate KMEANS and GMMS clustering implementation, respectively. Once the data set is divided into clusters, the second anomaly detector is deployed to detect outlier points in each data set. The second anomaly detector is based on principal component analysis since the engine data are transformed and projected into principal axes. The detected outliers are isolated and omitted from the respective engine data cluster. Data density plots are presented again in order to gain insights into the behavior of trim-draft parameters and identify the optimum number of clusters under the respective engine mode. Trim-draft data are classified into subclusters concerning each engine mode cluster. These sub-clusters represent the ship's localized operational conditions. A KPI index is calculated under each trim-draft mode in order to quantify the ship's performance. The optimum trim-draft mode is identified based on the KPI analysis. Identifying and analyzing the ship's localized operational conditions is vital since hidden data patterns and underlying parameter correlations may be distinguished. It is also a robust and versatile way to identify the optimum operating range of the investigated ship's parameters. Finally, two extra algorithms are presented to evaluate the second data anomaly detector results. The first is based on assessing each outlier point individually, and the second is in evaluating the detected outliers in groups. There is also an attempt to explain why these outliers occur.

In the flowchart down below (Figure 13), we see every step of this algorithm in sequential order.

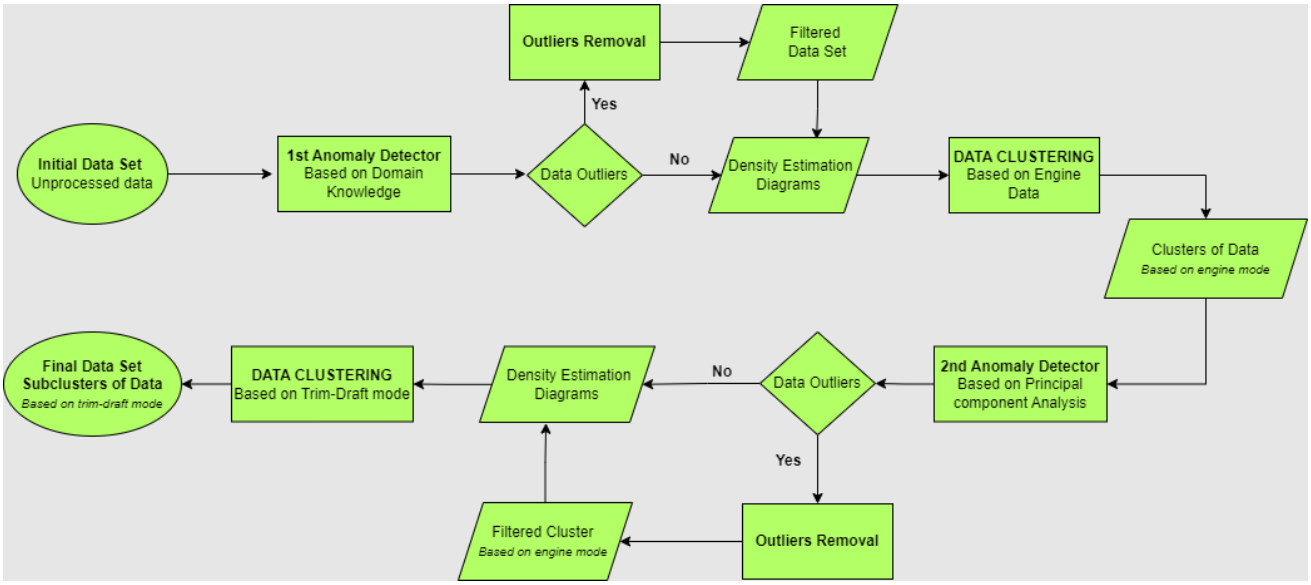

Figure 13: Graphical representation of the constructed algorithm.

3.8 Outlier evaluation algorithms

The first and the second anomaly detector isolate and exclude the detected outliers in the data set. As already stated, the second anomaly detector is based on Principal Component Analysis to identify the existing outliers effectively and comprehensively. After integrating the presented framework, the results of the second anomaly detector are further investigated to evaluate if the algorithm marked the emerged outlier points correctly. Also, an attempt to explain what causes these outliers is based on their time position and the correlation between the investigated operational variables, and the number of identified outliers. For the specific application, we construct two algorithms. The first is called “Outlier evaluation 1”, and the second is called “Outlier evaluation 2”.

3.8.1 “Outlier evaluation 1”

The “Outlier evaluation 1” algorithm examines the detected outliers individually in a time series order.

Let $DATA_{nXm}$ be a matrix, where n represents the number of observations and m represents the parameters in the investigated data set after the first data anomaly detector implementation.

Let d be a list of the identified outlier data points of a particular parameter’s signal, which are contained in the $DATA_{nXm}$ matrix, and t a list of the data points’ time stamps.

Then di, i=1,…,n is the i-th element of the list d, and ti, i=1,…,n is the corresponding i-th timestamp, and n is the number of the identified outlier points by the second anomaly detector. Also, di-1 and di+1 are the consecutive points of the di outlier point.

The absolute difference between the identified outlier value and its previous measured value, plus the absolute difference between the identified outlier value and its occurring measured

value, is calculated. The corresponding values are marked as adi, i=1,…,n, and are stored in a vector called $DIFF_{1Xn}$, DIFF= [ad1,..,adi,…,adn]

Then the maximum value of the $DIFF_{1Xn}$ vector is identified, and the corresponding outlier point, along with its timestamp, is presented in a time-series plot, along with ten previous points and ten successive points, so as to give the end-user a sense of the particular parameters' signal local behavior. Next, the identified outlier is plotted with green color and considered a "reasonable outlier" point since the difference between the specific outlier and its consecutive points is the maximum observed within the cluster. Finally, the regular data points are plotted with blue color.

The exact process is followed for the minimum value of the $DIFF_{1Xn}$ vector. But with the difference that, the identified outlier point is plotted with red color and considered an "unreasonable outlier" point since the difference between the specific outlier and its consecutive points is the minimum that can be observed in the cluster.

At that point, we need to highlight that the outlier designations are objective, and the main goal of the presented algorithm is to give the end user a sense of the behavior of the outlier data points in a time-series plot. Furthermore, the principal component analysis is well known for detecting outliers based on the mean and standard deviation of a whole set of data, which includes many different operational parameters rather than just the consecutive measurement of a single parameter.

In addition, it is essential to note that the algorithm has been designed so that the end user can choose from a variety of cluster types, such as (slow, transient, and service), along with the corresponding clustering method (KMEANS, GMMS). The algorithm can also be adjusted to simultaneously examine more than one outlier data point. As a result, the final plots can include several "reasonable" or "unreasonable" consecutive outliers. This modification can give the end user a greater sense of the local behavior of the identified outliers in a time-series format. The algorithm can also be modified to examine outlier points in between the minimum ad_{min} and maximum ad_{max} values.

3.8.2 "Outlier evaluation 2"

The "Outlier evaluation 2" algorithm collectively evaluates the detected outliers based on a time series order.

Let $DATA_{nXm}$ be a matrix, where n represents the number of observations and m represents the parameters in the investigated data set after the first data anomaly detector implementation.

Let d be a list of the identified outlier data points of a particular parameter's signal, which are contained in the $DATA_{nXm}$ matrix, and t a list of the data points' time stamps.

Then di, i=1,…,n is the i-th element of the list d, and ti, i=1,…,n is the corresponding i-th timestamp, and n is the number of the identified outlier points by the second anomaly detector.

The data set is sorted into eleven equally spaced time groups, and the sum of the detected outliers in each group is calculated. Afterward, a time-series plot is presented, split into eleven sections containing the corresponding data group. Red lines delimit the sections. The

distinctiveness of this plot is that the sections that contain a sufficient number of detected outliers are marked with dark red color, and the sections which contain very few outliers are marked with light red color. The color intensity depends on the number of identified outlier data points. So, it becomes easier for the end user to decide visually the groups that contain a lot of outliers and those that contain fewer outliers and consequently to identify the time period when these outliers occur. Finally, a histogram presents the calculated sums of the detected outliers in each group. Also, the detected outliers are presented in ascending time order by blue pulses in a signal plot.

After visualizing the detected outliers in a time series plot, the “Outlier evaluation 2” algorithm examines what caused these outlier points to occur. In order to achieve that, the mean values for the main operational parameters in each of the eleven date-time groups of the data set are calculated. Then, the correlation between the mean values and the number of the identified outlier data points of the respective groups is calculated and plotted for every investigated parameter. In this way, the algorithm reveals any possible connections between the number of the identified outliers and the measured values of each parameter. For example, if the mean value of the trim parameter is higher in the groups that contain a lot of outliers and lower in the groups that include only a few outliers, the measured correlation between these two variables will be close to 1. That can strongly indicate that higher trim values cause more outlier points.

The “Outlier evaluation 2” algorithm can also investigate the connection between the ship’s main operational parameters and the number of the identified outlier data points to a greater degree. That can be done by calculating the correlation between the variability in each main operational parameter and the number of the identified outliers for the eleven date-time groups.

The following variability measures are being used in the specific application:

- Range: the difference between the highest and lowest values
- Interquartile range: the range of the middle half of a distribution
- Standard deviation: average distance from the mean
- Variance: average of squared distances from the mean

This way, the algorithm can reveal the connection between the ship’s operational parameters variability and the identified number of outlier points. For example, groups with many outlier points display high variability in wind speed measurements. That possibly means that sudden changes in weather conditions can spark more outlier points in the final data set.

4. Results

4.1 Introduction

In this chapter, the results of the proposed framework are presented. Below are the minimum-maximum values of the navigation parameters imposed by the first anomaly detector, followed by data density plots of the engine data (i.e., propeller shaft speed and propeller shaft power). Afterward, we present the plots of the K-Means and GMM clustering methods. A scatter plot and a consecutive table highlight both clustering methods' second anomaly detector outcome. Next, data density plots and clustering implementation plots represent the ship's localized operational modes. Afterward, Ship performance quantification results and visual analytics are displayed with the respective tables and figures. Finally, the two constructed algorithms, outlier evaluation 1 and outlier evaluation 2, are presented to assess the efficiency of the second anomaly detector.

For the specific application, a data set from a containership was obtained. The main characteristics of the investigated containership are presented in Table 1.

Table 1: Main ship's particulars.

Main Ship's particulars

Length B.P	199.00 (m)
Breadth	30.20 (m)
Depth	16.70 (m)
Draft	11.50 (m)
TEU	2550
Engine type: Hyundai-Wärtsilä 7RTA72U-B	M.C.R : B.H.P (kW):21560 / 99.0 R.P.M

4.2 Data description

This application obtained a data set of fifteen ship operational parameters (navigation, engine data, etc.). This time series data set contains data measurements from mid-December 2016 till late December 2017, with a sampling rate of 1 minute. Table 2 below presents the physical quantities that were examined in the respective data set.

Table 2: Examined parameters description.

Measured physical quantity	Parameter name	Units
Speed over ground	SOG	Knots (kn)
Propeller shaft speed	PSS	revolutions per minute (rpm)
Propeller shaft power	PSP	kilowatts (kW)
Draft mean	DM	meters(m)
Trim	T	meters(m)
Main engine's fuel oil consumption	MEFOC	Tons/day
Propeller shaft torque	PST	kilonewton meters (kN*m)
Wind Speed	WS	Meters per second (m/s)

4.3 First data anomaly detector

It was discovered that numerous data points were erroneous during the deployment of the first data anomaly detector, which is based on domain knowledge. For example, many Main Engine (ME) fuel consumption measurements are identified with zero values while their physical value should be different. These points were characterized as outliers and removed from the data set. In addition, all data points with propeller shaft Power and Shaft Speed measurements less than 3000kW and 60 rpm were excluded from the study because they were considered to correspond to not an open-sea operation (e.g., port maneuvers). These values were isolated and omitted from the data set. A percentage of 24.33 % of the data points were omitted from the original data set during the first anomaly detector implementation.

The minimum and maximum values of the navigation parameters, which are based on domain knowledge, are shown in Table 3.

Table 3: Ship operational parameters and their minimum–maximum values.

Parameter	Unit	Min-value	Max-value
Propeller shaft power	*[kW]*	*3000*	*21560*
Propeller shaft speed	*[rpm]*	*60*	*100*
Main Engine (ME) fuel oil consumption (cons)	*[Ton/day]*	*1*	*100*
Speed over ground (SOG)	*[Knots]*	*2*	*25*
Average (Avg) draft	*[m]*	*0*	*11.5*
Trim	*[m]*	*-2*	*4*
Propeller shaft torque	*[kN*m]*	*0*	-
Wind Speed	*[m/s]*	*0*	-

Figure 14 shows the propeller shaft power - Speed diagram before and after the First Anomaly Detector implementation. There are two main regions in this plot where outliers have been identified. Outlier points are detected in the first region, which is in the bottom-left of the diagram, and are prompted by thresholds applied to propeller shaft power and propeller Shaft Speed measurements. There is a second smaller region on the right, which is prompted by some zero values in fuel consumption measurements, as identified by the respective thresholds. The interesting case of zero values in fuel consumption measurements is further investigated in Figure 15. In which the identified zero fuel consumption values are presented in the propeller shaft power-speed diagram on the left and in the propeller shaft power-time diagram on the right in red color. According to the right plot in Figure 15, the data acquisition system acquired the second region measurements in a relatively short time. As a result, we assume that there was an error in the specific system during that period.

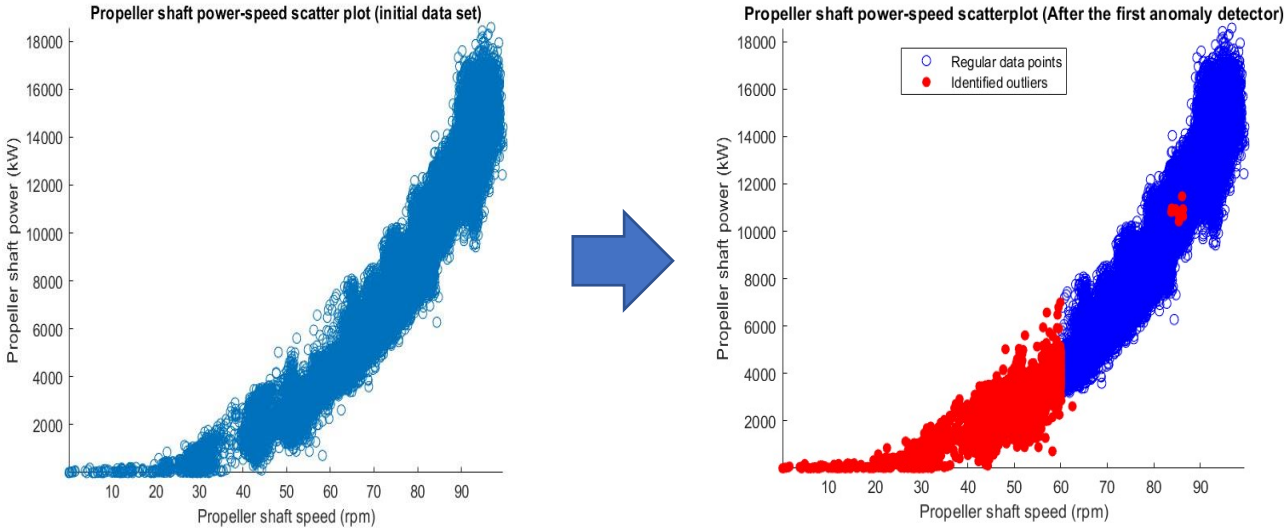

Figure 14: Propeller shaft power-speed diagram before and after first anomaly detector implementation.

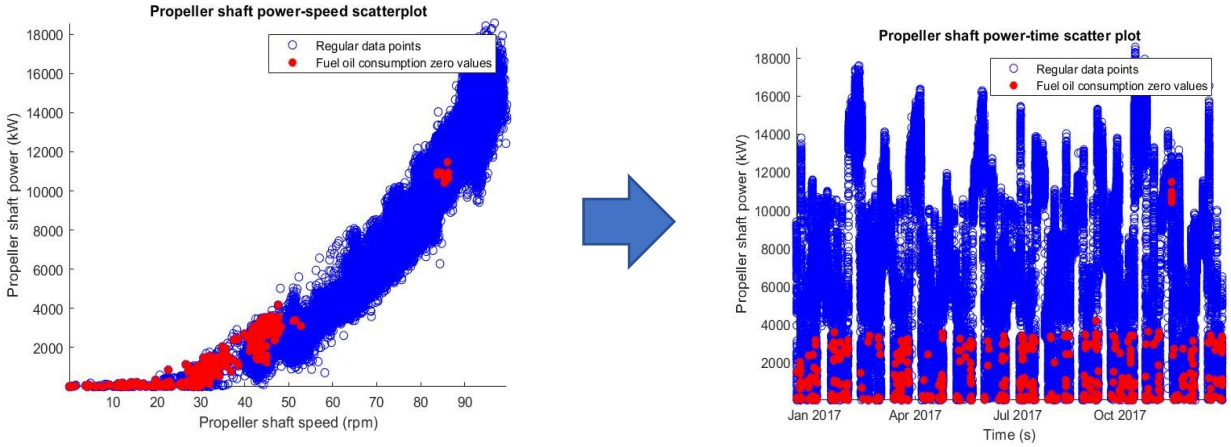

Figure 15: Propeller shaft power - speed diagram and propeller shaft power - time diagram concerning main engine's fuel consumption zero values.

The histograms for each of the examined parameters before and after the first anomaly detector implementation are presented in the left and the right plot, respectively. The black dotted lines represent the applied thresholds of the first anomaly detector, which are based on domain knowledge. The propeller shaft speed, propeller shaft power, speed over ground, main engine fuel oil consumption, mean draft, and trim histogram plots are presented in Figure 16, Figure 17, Figure 18, Figure 19, Figure 20, and Figure 21 correspondingly.

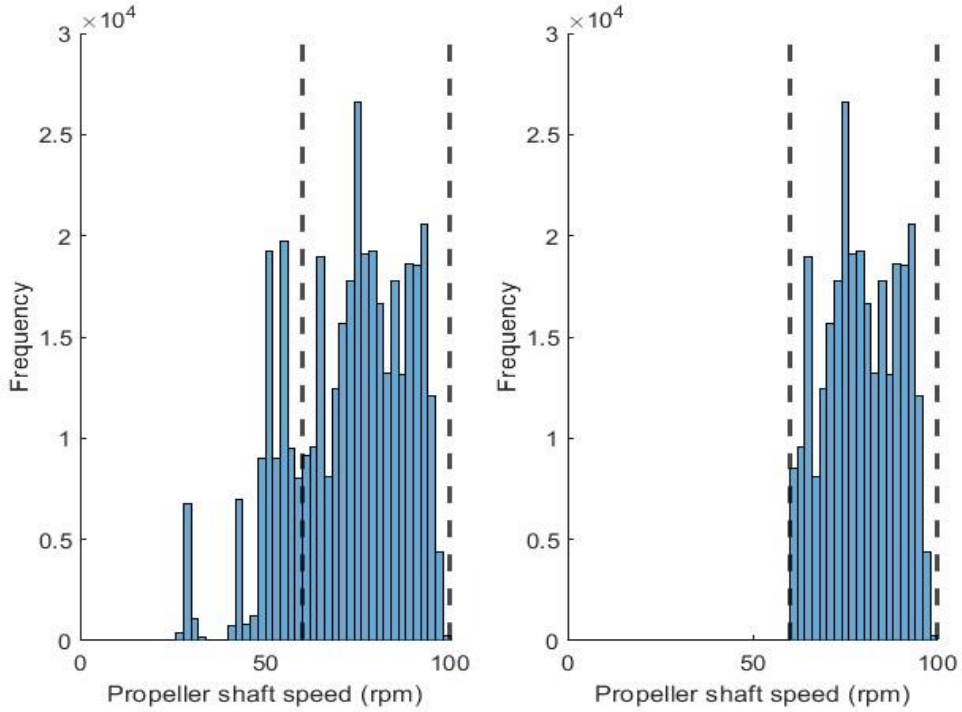

Figure 16: Propeller shaft speed histograms before and after the first anomaly detector implementation.

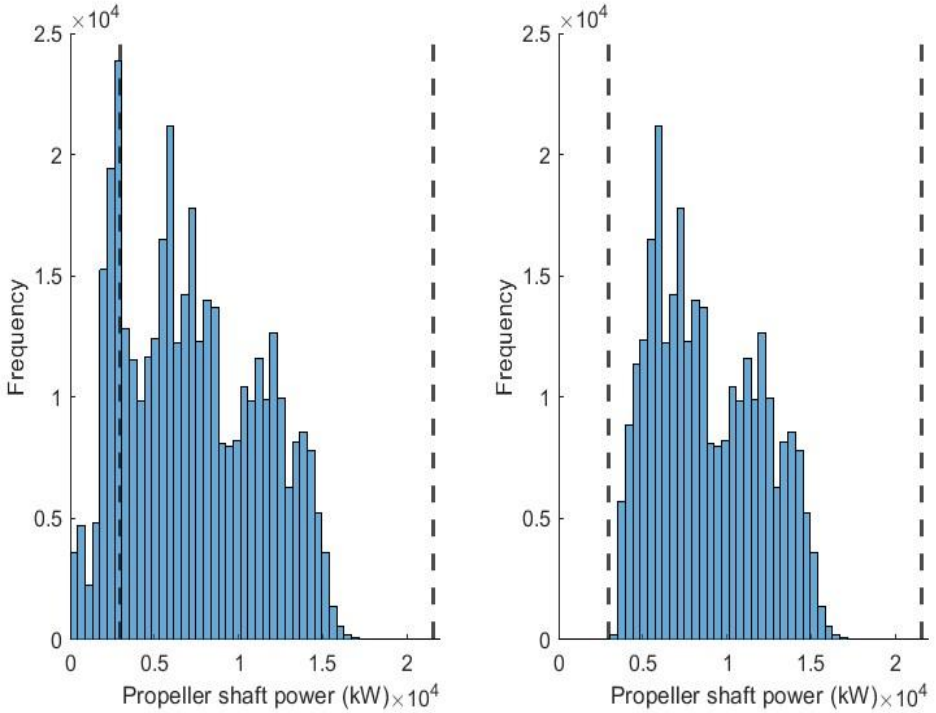

Figure 17: Propeller shaft power histograms before and after the first anomaly detector implementation.

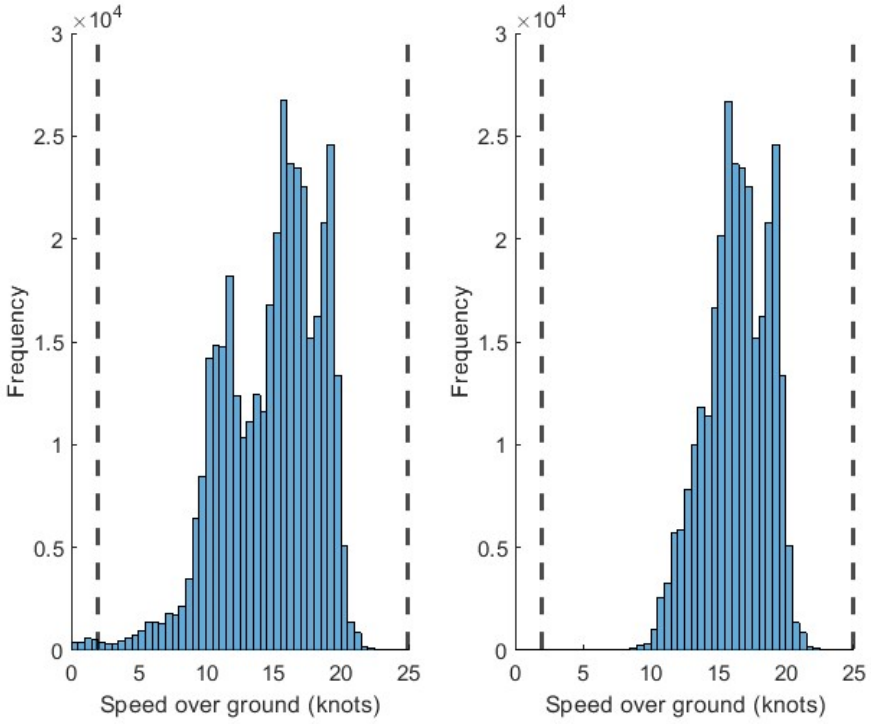

Figure 18: Speed over ground histograms before and after the first anomaly detector implementation.

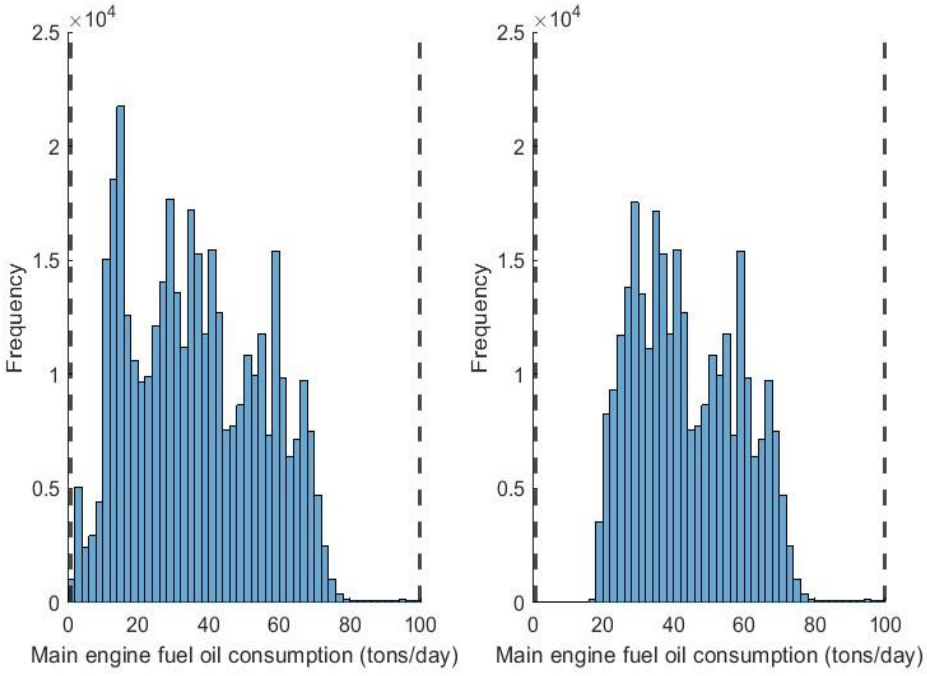

Figure 19: Main engine fuel oil consumption histograms before and after the first anomaly detector implementation.

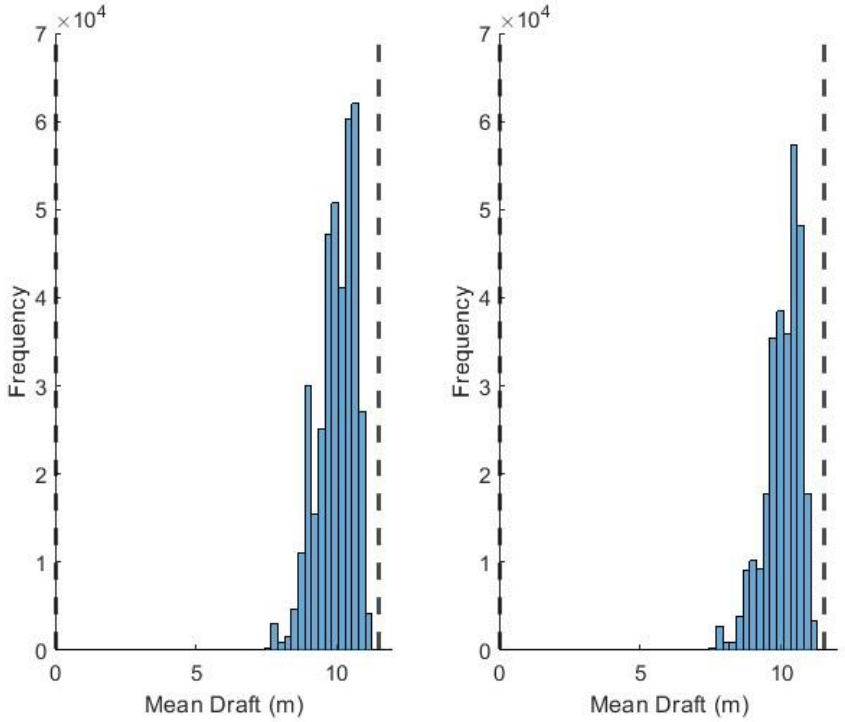

Figure 20: Mean draft histograms before and after the first anomaly detector implementation.

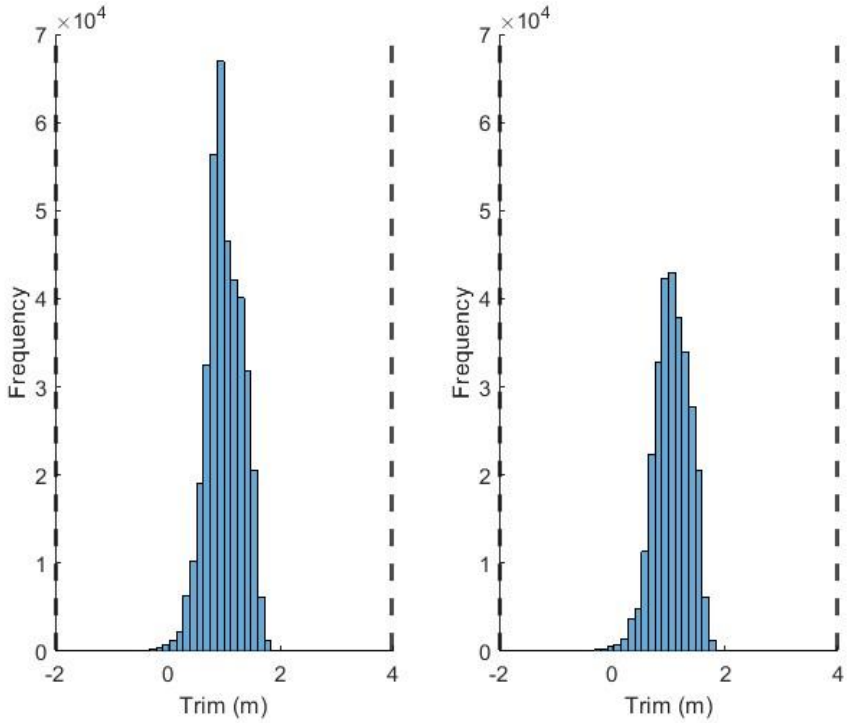

Figure 21: Trim histograms before and after the first anomaly detector implementation.

It is worth noting that most measurements of the respected parameters fall within the applied thresholds, except for the lowest data points of the propeller shaft power and shaft speed parameters, which are set to exclude maneuvering conditions from the investigated data set. So, the main volume of the excluded data points is marked as maneuvering measurements. It is also visible in the presented histograms that the observed values of the speed over ground measurements are approximately between 8 and 22 knots, higher than the 2 knots lower limit we imposed during the first anomaly detector.

4.4 Data pattern recognition

The bivariate histogram is utilized to visualize the behavior of our data about the propeller shaft power-Speed variables. The Kernel Density Estimation Function is considered afterward with the main goal of giving us insights into the number of clusters in our data combined with univariate histograms into a scatterhist plot. Lastly, a density scatter plot is regarded to give a better insight into the density of our data. The results are presented in Figure 22, Figure 23, and Figure 24, respectively.

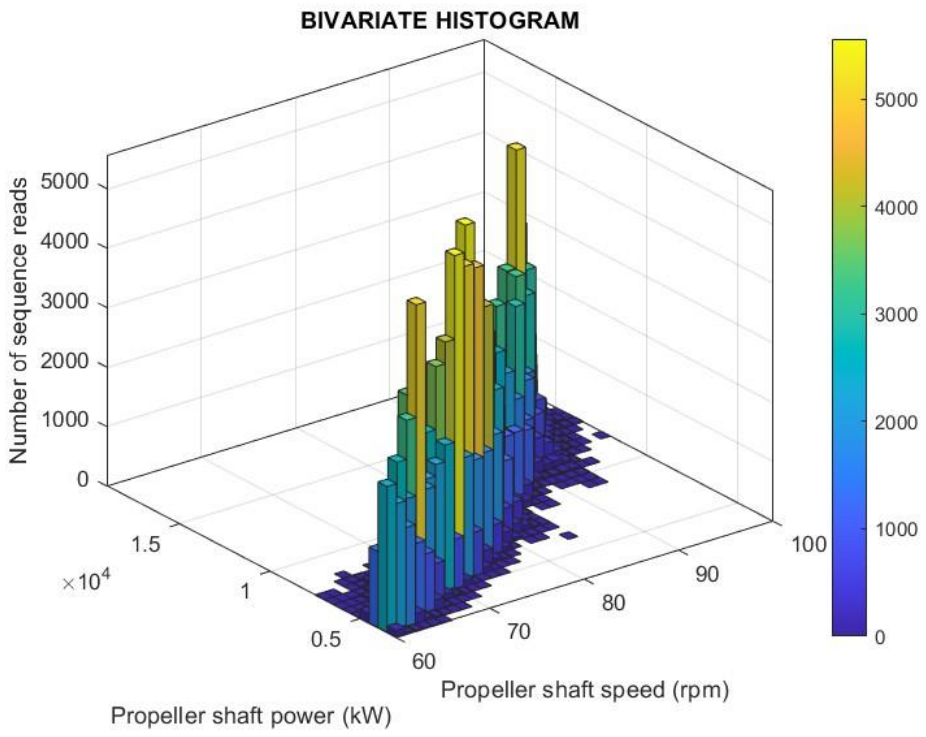

Figure 22: Bivariate colored histogram based on engine data density.

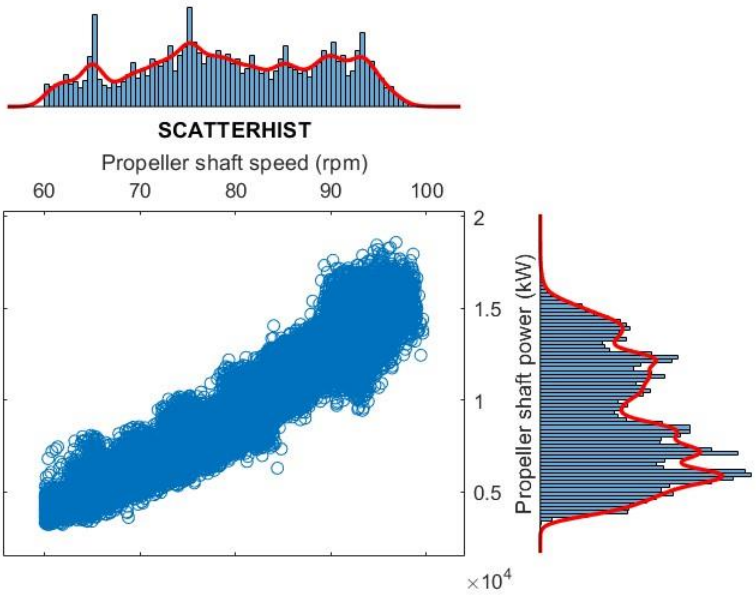

Figure 23: Scatterplot combined with univariate histograms and kernel Density Estimation plots.

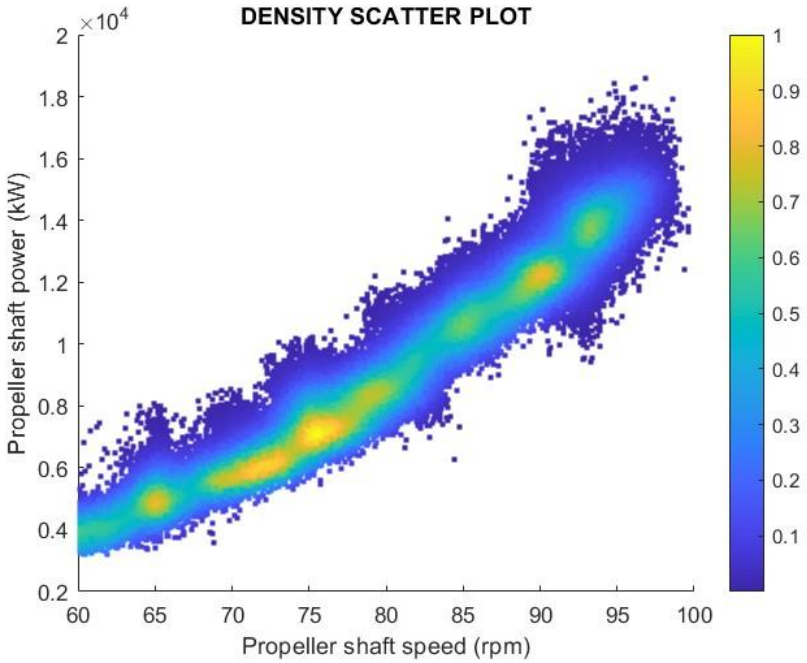

Figure 24: Data density scatter plot based on engine data (i.e., Propeller shaft power - speed).

In the above figures, it is noticed that there are specific areas with high-density data and some others with low-density data. The bivariate histogram is very insightful about the frequency of the examined data. In the scatterhist plot, it is worthwhile to note that the histogram plot and kernel density estimation function follows the same trend, so the insights provided by these plots are nearly identical. It is primarily due to MATLAB's automated binning algorithm

that reveals the shape of the underlying distribution with high accuracy in histogram plots. Lastly, the Data Density plot depicts very accurately the density of our data in a 2-D space making this plot informative and comprehensible at the same time.

4.5 Data clustering

Approximately three components (clusters A, B, and C) may be distinguished from the density estimation of the engine data, as illustrated in Figure 22, Figure 23 and Figure 24. Among these, clusters A and C represent the two primary engine operating modes, the slow speed mode and the service speed mode. Whereas the data points which are included in cluster B represent a transient engine condition. Therefore, K = 3 was recommended as the number of components (i.e., the number of clusters) for the K-Means algorithm and GMMs. Clustering results are presented in Figure 25 and Figure 26.

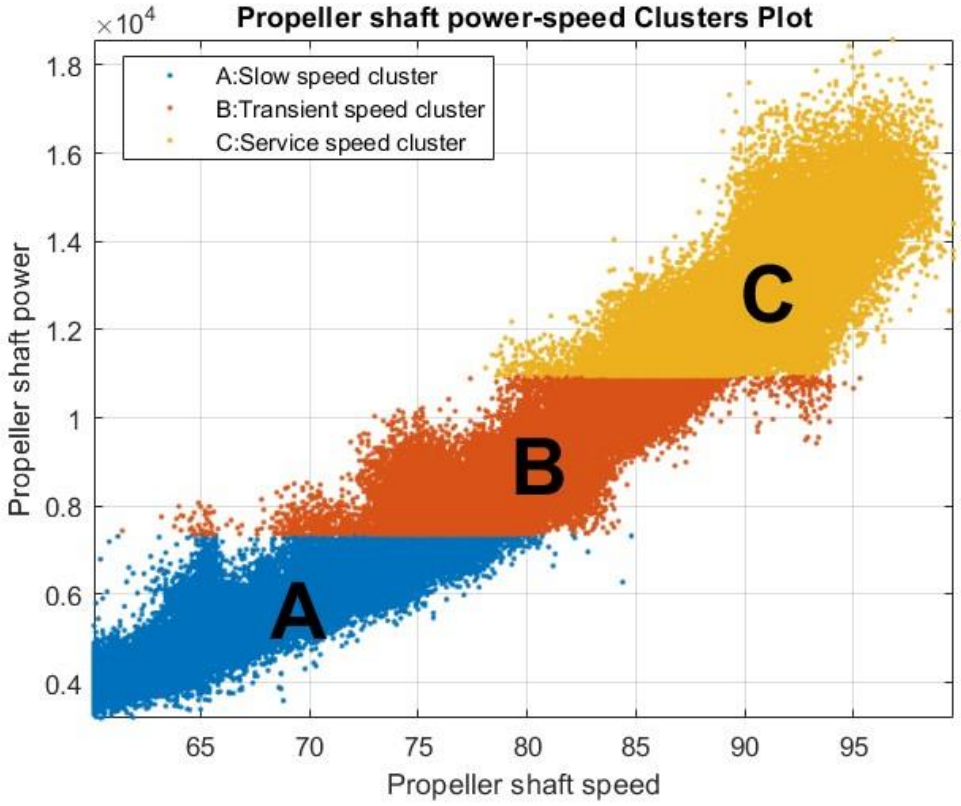

Figure 25: K-MEANS Clustering plot based on engine data (i.e., Propeller shaft speed and power).

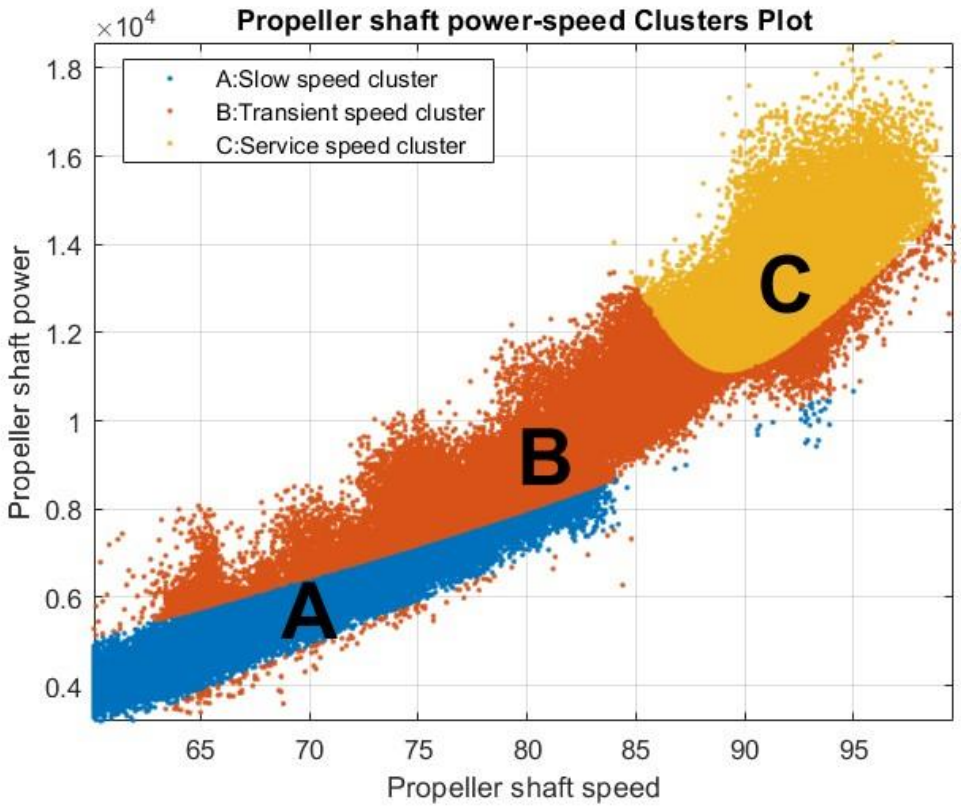

Figure 26: GMM'S Clustering plot based on engine data (i.e., Propeller shaft speed and power).

It is observed in Figure 25 and Figure 26 that the clustering results in implemented methods are similar but not identical. That may be due to differences in the mathematical approaches of these methods. The Euclidean distance, the within-cluster similarity measure in K-Means, cannot detect complex non-linear usage patterns in the data set. The main difference between these methods is that K-means uses a deterministic approach and assigns each data point to a unique cluster. This is referred to as the hard clustering method. GMM uses a probabilistic approach and gives the probability of each data point belonging to any of the clusters. This is referred to as the soft clustering method. Two clustering evaluation techniques are deployed for clustering assessment as presented in chapter 4.6. Also, a time series plot is shown in Appendix A: to manifest the behavior of the speed over ground parameter after the clustering implementation to the engine data.

4.6 Clustering evaluation criteria

The clustering results are further investigated as described in chapter 3.3.7. The optimum number of clusters is also identified based on heuristic and information criteria for the K-means algorithm and gaussian mixture models, respectively.

4.6.1 Evaluation of K-means algorithm clustering results

The elbow plot method is implemented to evaluate the clustering results of the k-means algorithm, as presented in Figure 27. The investigated range of clusters for that specific case varies from 1 to 10. It is evident that when the number of clusters rises, the sum of squared distances within clusters is reduced. But after the third cluster, the respective value gradually decreases. As a result, according to this method, the k-means algorithm works best for three clusters. Also, at this point, the ''elbow shape'' is presented.

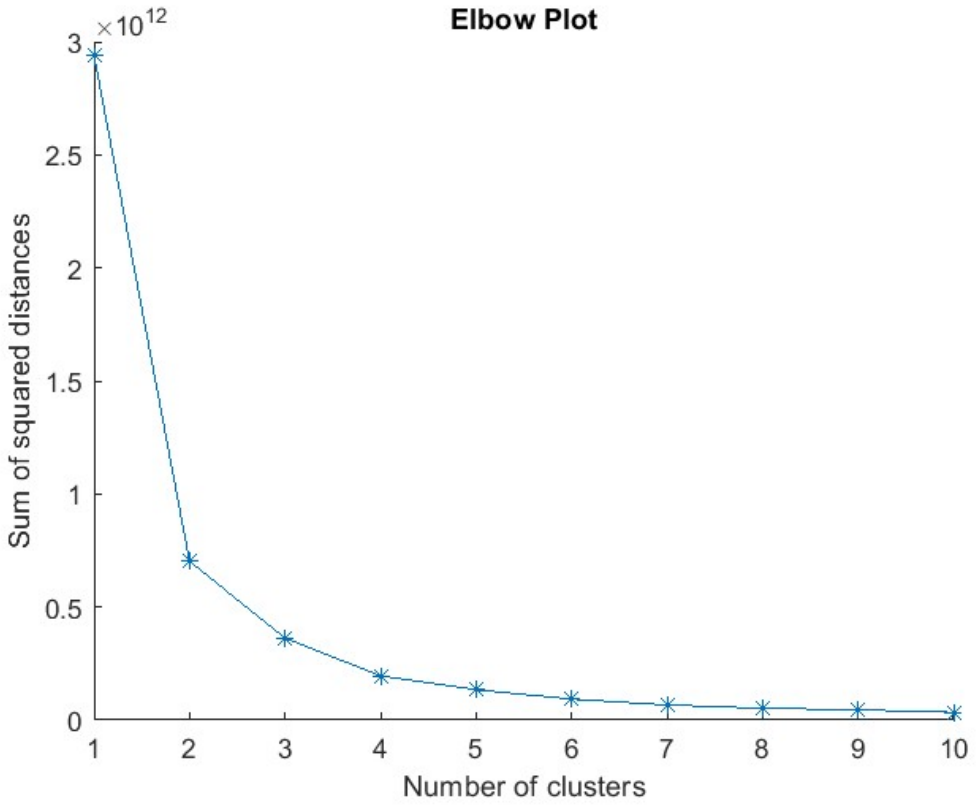

Figure 27: Elbow plot for k-means clustering results evaluation.

Another way to identify the optimum number of clusters is to calculate the gradient values of the elbow plot so as to be informed about the magnitude difference of two consecutive points. The individual results are presented in Figure 28.

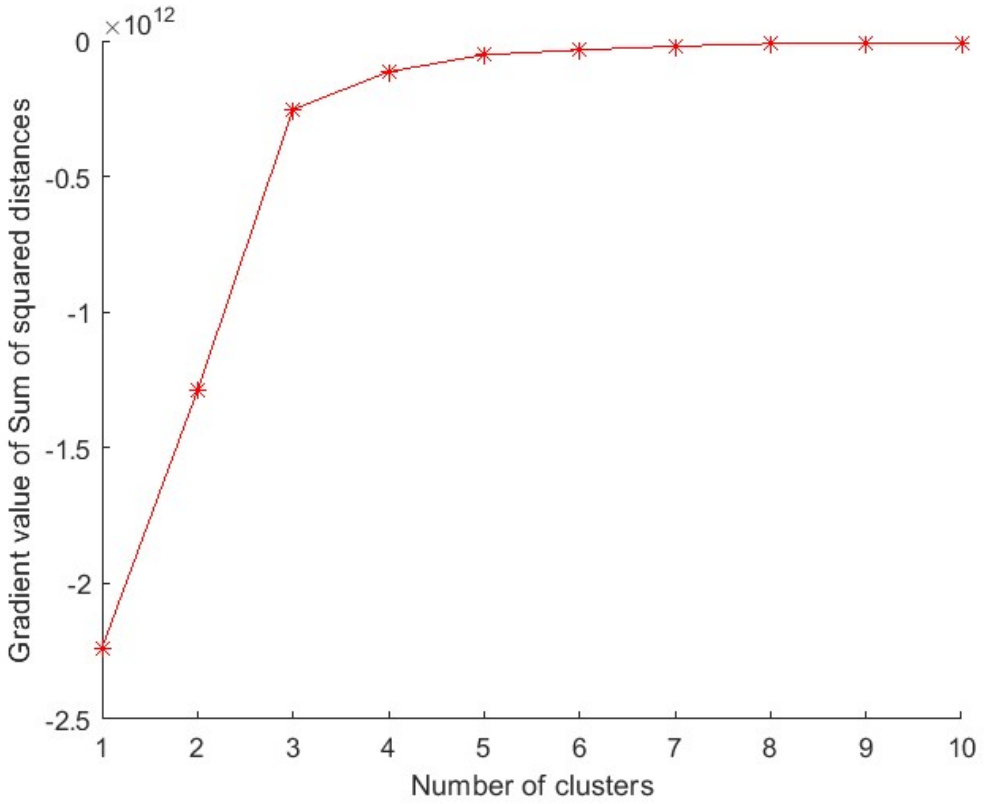

Figure 28: Gradient values of elbow plot.

As expected, all the gradients have negative values. But We can notice that after the fourth cluster, the gradient becomes almost constant, so the ideal number of clusters for this data set is three.

4.6.2 Evaluation of gaussian mixture models clustering results

For Gaussian mixture model clustering, the AIC/BIC criterion is investigated. The lowest calculated AIC/BIC score for a range of clusters depicts the best fit for the model.

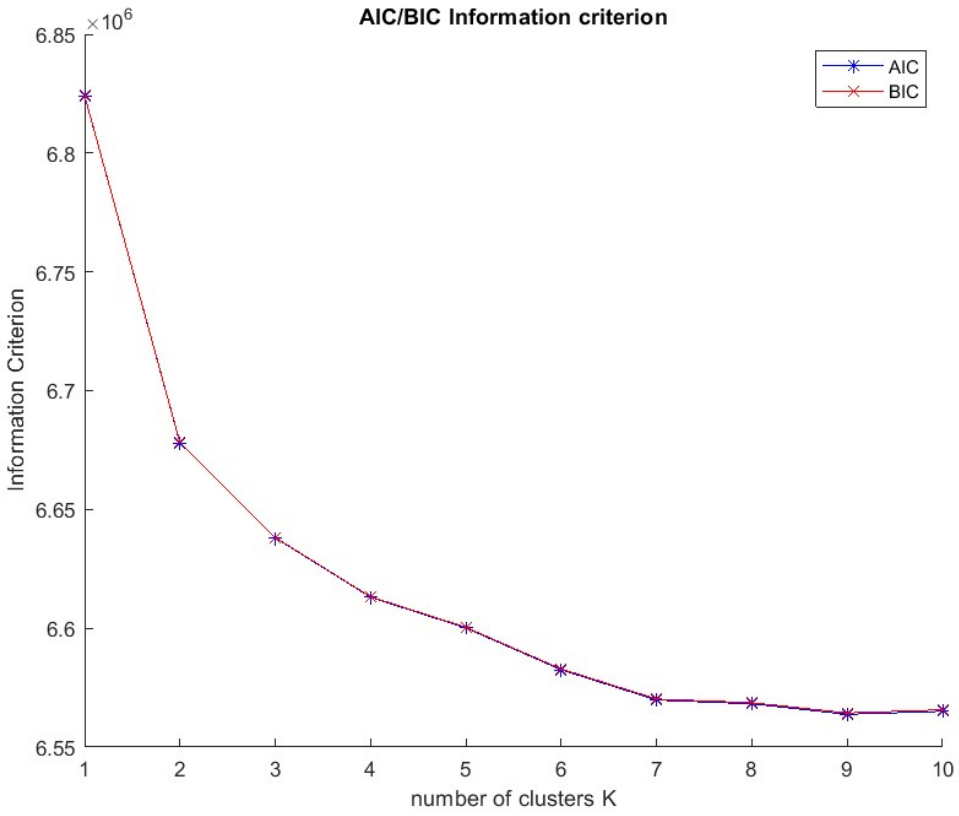

Figure 29: AIC/BIC information criterion plot.

In Figure 29, it is evident that the greater the number of clusters, the better the model should be. Unfortunately, that causes overfitting to our data, and it's one of the major disadvantages of these criteria. One way to overcome this problem is to calculate the Gradient of the AIC/BIC score curves. The calculated Gradient of the AIC/BIC score is depicted in Figure 30.

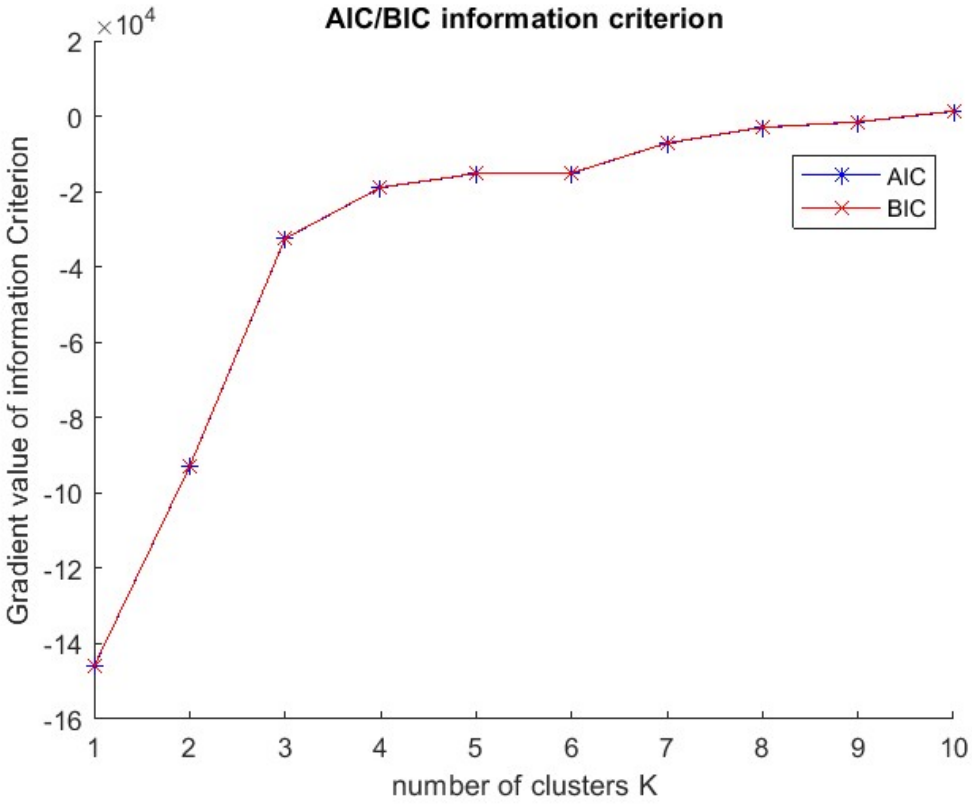

Figure 30: Gradient value of AIC/BIC score.

As expected, all the gradients have negative values. But we see more clearly that starting from a cluster size of four, the gradient becomes almost constant, i.e., the original function has a gentler decrease, i.e., there is not much gain in increasing the number of clusters. In short, this technique suggests we use three clusters.

4.7 Second data anomaly detector

After dividing the data into clusters, the second data anomaly detector was set up and used. The results of the Second anomaly detector implemented in the Service Speed cluster after the K-MEANS clustering algorithm are first analyzed. As already stated, the second anomaly detector was based on Principal Component Analysis, where the first components contain the most information and the last components contain the least information. Then our data were projected along the principal component axes. It should be noted that the proper threshold values were -3σ and 3σ, where σ is the standard deviation of the corresponding data distribution. Data points that are greater than specified values are marked as anomalies by this detector. Several abnormalities are found in this regard, as demonstrated in Figure 31.

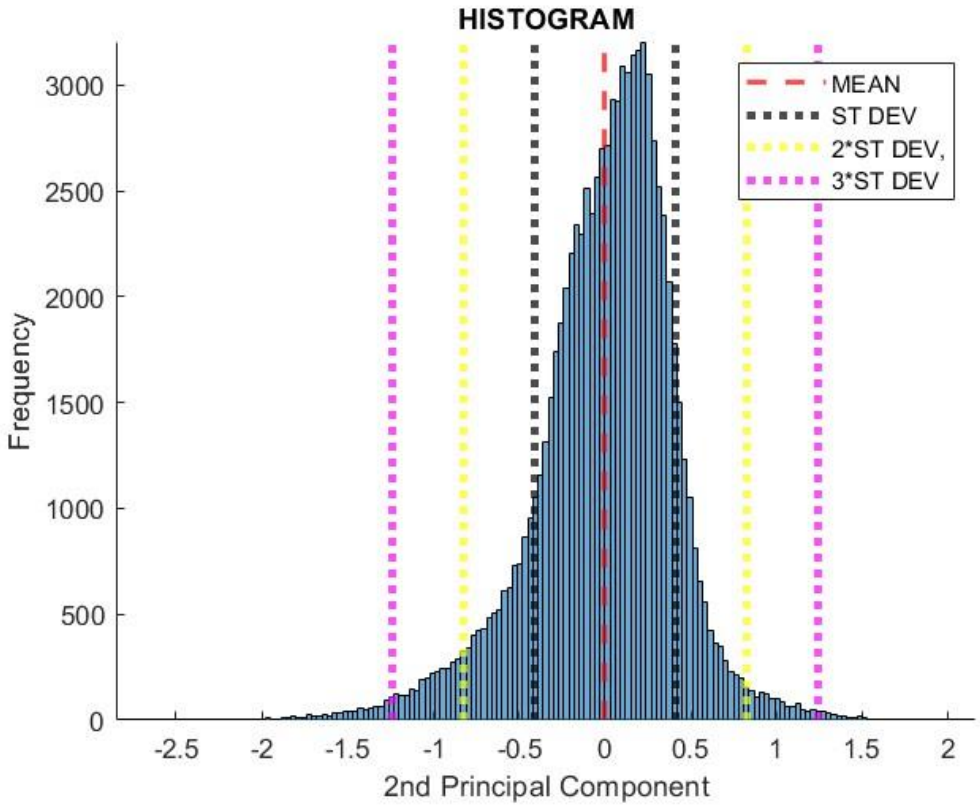

Figure 31: Histogram of Service Speed Cluster Data represented by the Second Principal Component.

A graph is also presented in Figure 32, in which pulses represent outlier data points. Our analysis shows that outliers are highly concentrated in some areas, whereas in other areas, there are only a few detected outlier points. That can be justified by the fact that sensor measurements are more vulnerable in unsteady and harsh weather operational conditions (areas with high concentrations of outliers) in comparison with more steady weather operating conditions (areas with low concentrations of outliers). This is also supported by Figure 33, which displays a histogram of detected outliers in relation to the time-series format of our data set.

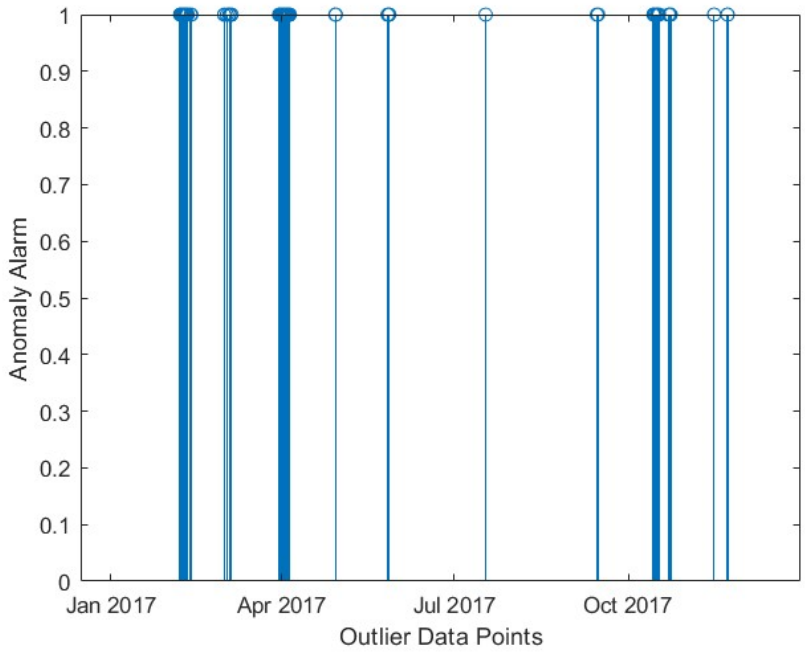

Figure 32: Detected data anomalies presented in a discrete-time signal plot.

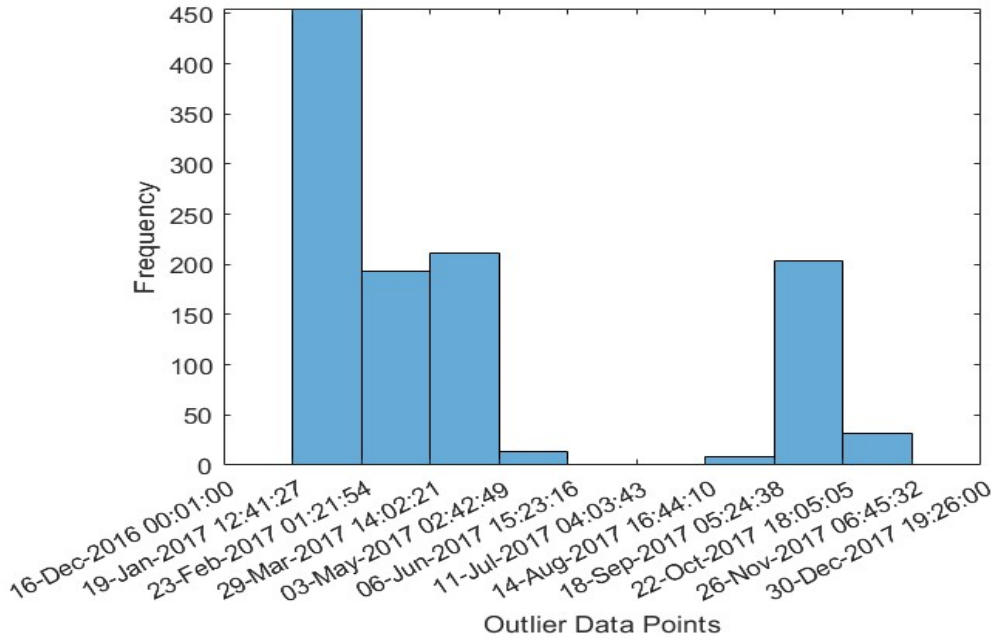

Figure 33: Frequency of detected outliers concerning the time-series format of our data set.

In Figure 34, the Service Speed Cluster after the K-MEANS algorithm and Second Anomaly Detector implementation is presented. The identified outliers are marked with red color and the regular data points with green color. It is observed that the identified outliers are located on the edge of the cluster where the detected data density is low. That's one of the main characteristics of outlier points, so we conclude that the Second Anomaly Detector efficiently identified the existing outlier points.

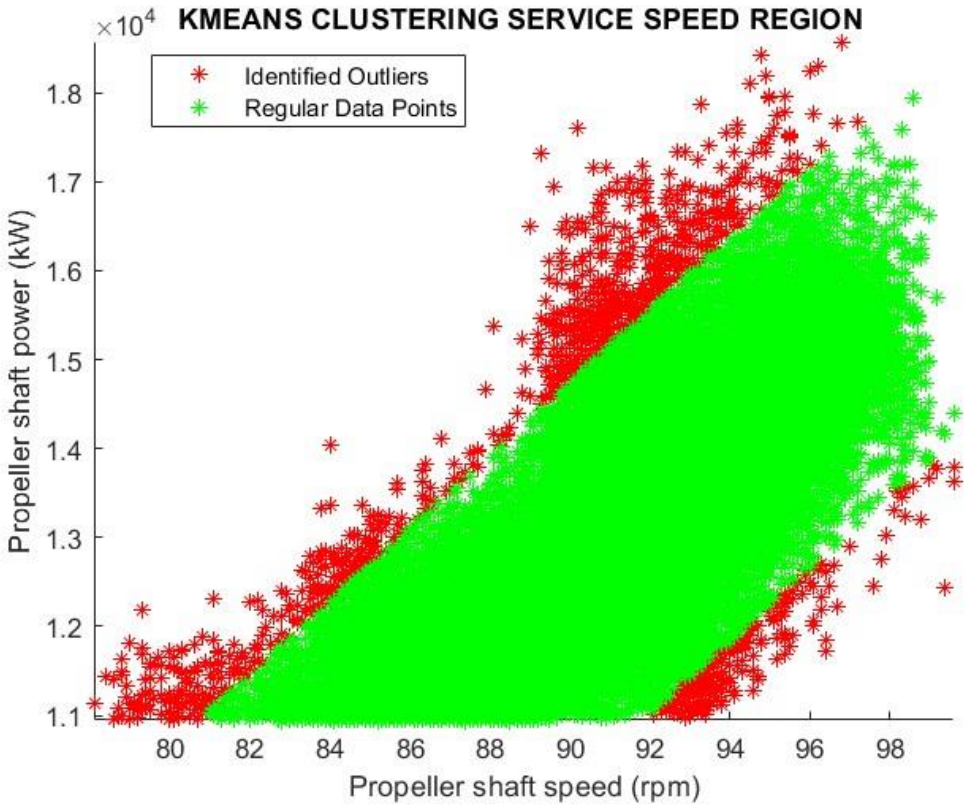

Figure 34: Graphical representation of Service speed cluster after k-means clustering regarding inlier and identified outlier data points.

In Figure 35, the Service Speed Cluster after the GMMS model and Second Anomaly Detector implementations are presented. The identified outliers are marked with red color and the regular data points with green color. It is observed that the identified outliers are located on the top edge of the cluster where the detected data density is low. Most of the detected outlier points identified in the KMEANS Service Speed Cluster and GMM'S Service Speed Cluster are identical. The existing differences are attributed to the differences in the structure of the clusters.

The graphical representations of Slow and Transient Speed Clusters for K-MEANS and GMMS clustering methods regarding inlier and identified outlier data points are presented in Appendix B:.

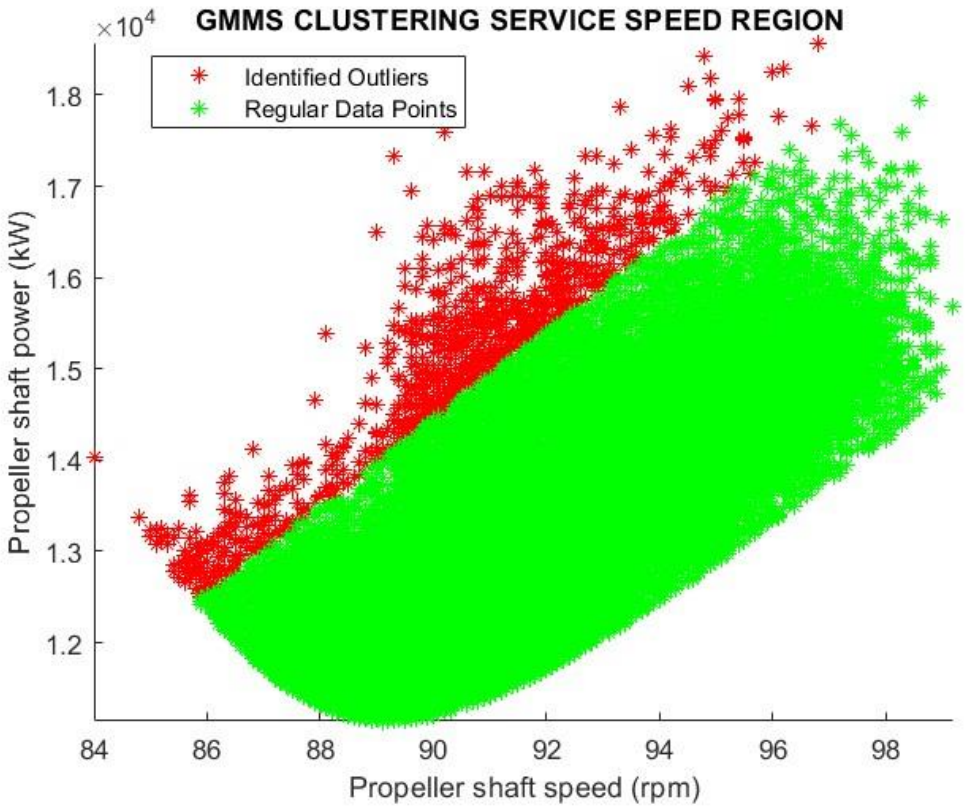

Figure 35: Graphical representation of Service speed cluster after GMM'S clustering regarding inlier and identified outlier data points.

The total number and the percentage of the identified outliers for each of the three engine modes for both Clustering methods are summarized in Table 4. Finally, the detected outlier points are omitted from the respective cluster to improve the quality of the data set and proceed with further analysis of the ship's performance under specific localized operational conditions.

Table 4: Number and percentage of identified anomalies based on the second anomaly detector.

Clustering Method / *Engine Operating Region*	*K-Means Algorithm*		*Gaussian Mixture Model*	
	Identified Anomalies	*Percentage (%)*	*Identified Anomalies*	*Percentage (%)*
Slow Speed Region	*1709*	*1.48 %*	*698*	*0.60 %*
Transient Speed Region	*779*	*0.88 %*	*936*	*0.92 %*
Service Speed Region	*1117*	*1.28 %*	*955*	*1.3 %*

The percentage of identified outliers varies depending on the clustering method. A significant percentage difference is noticed in the Slow Speed Region, as expected. In this cluster, the results of clustering deviate the most. In the other two clusters, the percentages of identified outliers converge because the structure of the identified clusters also converges. In total, 3605 outlier points are detected in K-MEANS Service Speed Cluster, 39.2% more than 2589 detected outlier points in GMM'S Service speed cluster. That difference is attributed to the inadequacy of the K-MEANS algorithm to detect complex data clusters in comparison with GMM'S model, which can identify hidden data patterns and return robust results in complex data sets. So, we consider the results of the second anomaly detector more reliable in GMM clustering.

To further explore the efficiency of the Second Anomaly Detector and the outlier's behavior, Visual analytics are applied. The detected anomaly points for the KMEANS Service speed Cluster are presented in a time-series plot, as shown in Figure 36 and in Figure 37 for GMMS Service Speed Cluster.

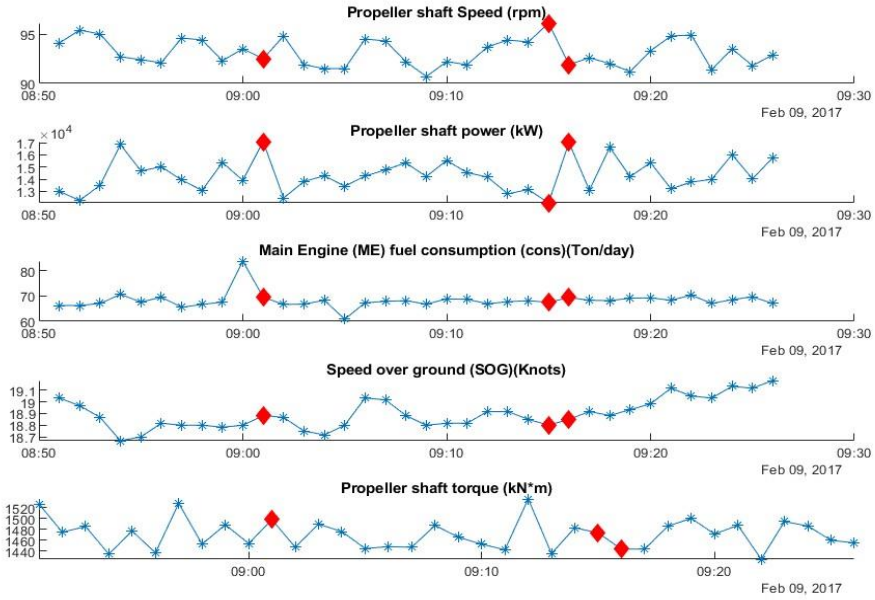

Figure 36: Time series plot of the Main engine operational variables regarding the KMEANS Service Speed Cluster identified outlier points.

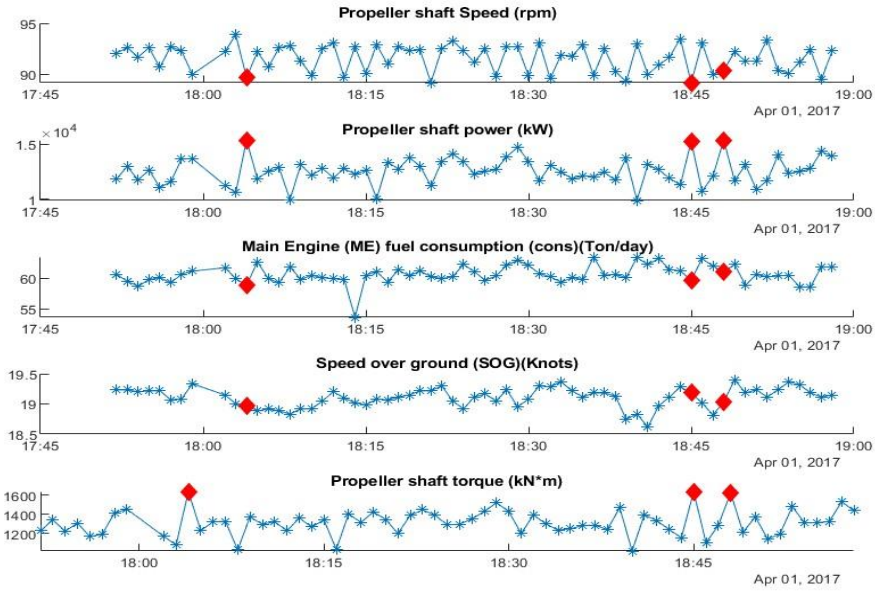

Figure 37: Time series plot of the Main engine operational variables regarding the GMMS Service Speed Cluster identified outlier points.

In the aforementioned plots, the first three detected outliers are marked in a time series plot with red color, and the regular data points between them are marked with blue color. Five operational parameters are included in this time-series format. The first one is Shaft Speed (rpm), followed by propeller shaft power (kW), Main Engine fuel consumption (Ton/day), Speed over ground (knots), and Propeller Shaft torque (kN*m). We notice that in these detected anomalies, sudden changes are observed with respect to Shaft Speed, Main Engine Power, and Propeller shaft torque. Due to the extreme challenge of visually investigating all the detected outliers in the specific cluster, two different algorithms were constructed, with the primary goal of providing us some more insights into the behavior of the detected outliers. More details about these algorithms are discussed in 4.11.

4.8 Exploration of the ship's localized operational conditions

To better understand the ship's localized operational conditions, we investigate the trim-draft modes under which the vessel operates in each engine mode. To achieve that, the deployment of data density plots is determined. Domain knowledge combined with KDE provides us with the most information to decide under how many trim-draft modes our ship operates concerning each engine mode. Subclusters can represent these regions.

The bivariate histogram is utilized to visualize the behavior of our data in relation to the Trim–Draft variables. The Kernel Density Estimation Function was considered afterward with the main goal of giving us insights into the number of subclusters in our examined cluster, combined with univariate histograms into a scatterhist plot. Lastly, a density scatter plot is regarded to give a better insight into the density of our data in the respective cluster.

Based on GMMS and K-MEANS clustering, the Slow Speed Cluster (Cluster A) is being analyzed. The results of the implemented data density plots are shown in Figure 38.

The respective results of bivariate Histogram, Kernel Density Estimation plot are presented in Appendix C:

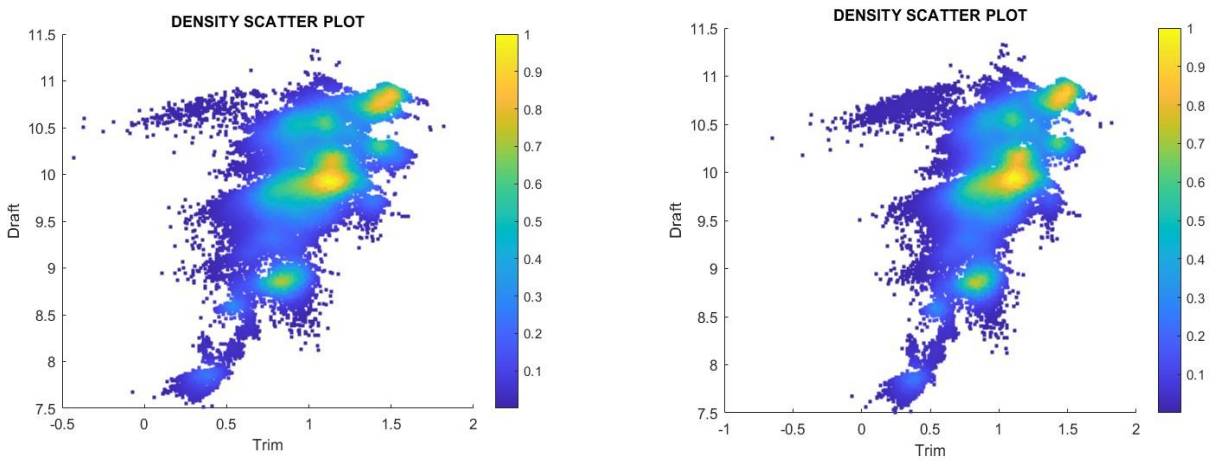

Figure 38: Data density scatter plot of trim/draft variables with respect to Slow Speed Cluster. After GMMS (on the left) and K-MEANS clustering (on the right).

Based on GMMS and K-MEANS clustering, the Transient Speed Cluster (Cluster B) is being analyzed. The results of the implemented data density plots are shown in Figure 39.

The respective results of Histograms, Kernel Density Estimation plots are presented in Appendix C:

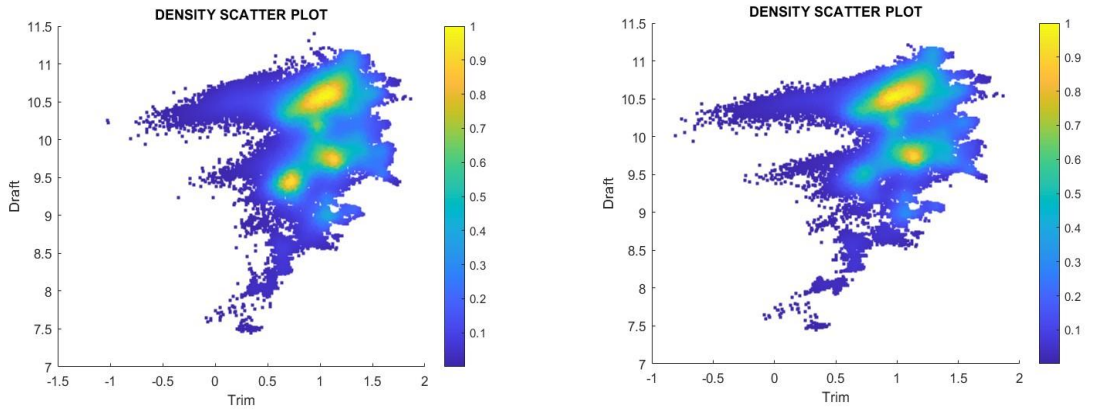

Figure 39: Data density scatter plot of trim/draft variables with respect to Transient Speed Cluster. .After GMMS (on the left) and K-MEANS clustering (on the right).

Based on GMMS and K-MEANS clustering, the Service Speed Cluster (Cluster C) is being analyzed. The results of the implemented data density plots are shown in Figure 40.

The respective results of Histograms, Kernel Density Estimation plots are presented in Appendix C:

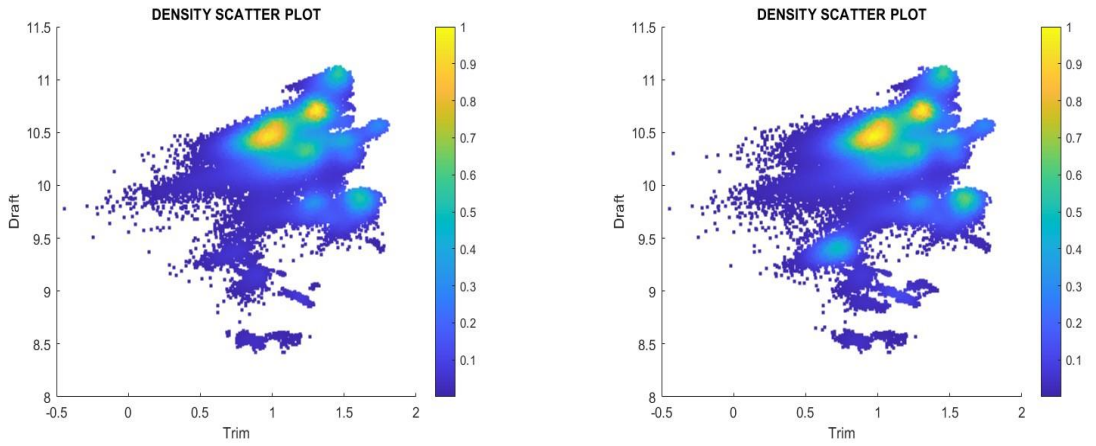

Figure 40: Data density scatter plot of trim/draft variables with respect to Service Speed Cluster. After GMMS (on the left) and K-MEANS clustering (on the right).

4.9 Data sub-clustering

From the previous analysis, we conclude that the number of sub-clusters varies depending on the Engine mode (i.e., Cluster A, Cluster B, Cluster C). Therefore, three Trim-Draft modes are identified under the Slow Speed Region (Cluster A), two Trim-Draft modes are identified under the Transient Speed Region (Cluster B), and one Trim-Draft mode is identified under Service Speed Region (Cluster C). The respective Sub-clusters are presented in Figure 41, Figure 42, and, Figure 43 regarding GMMS and K-MEANS clustering implementation.

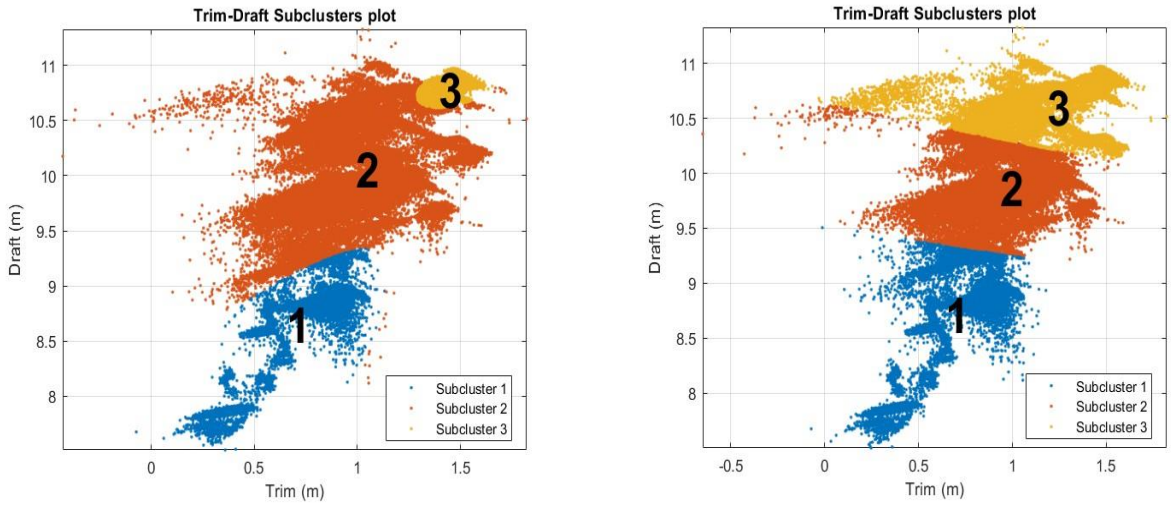

Figure 41: Subclusters plot of Trim-Draft variables concerning Slow Speed Region. After GMMS (on the left) and K-MEANS clustering (on the right).

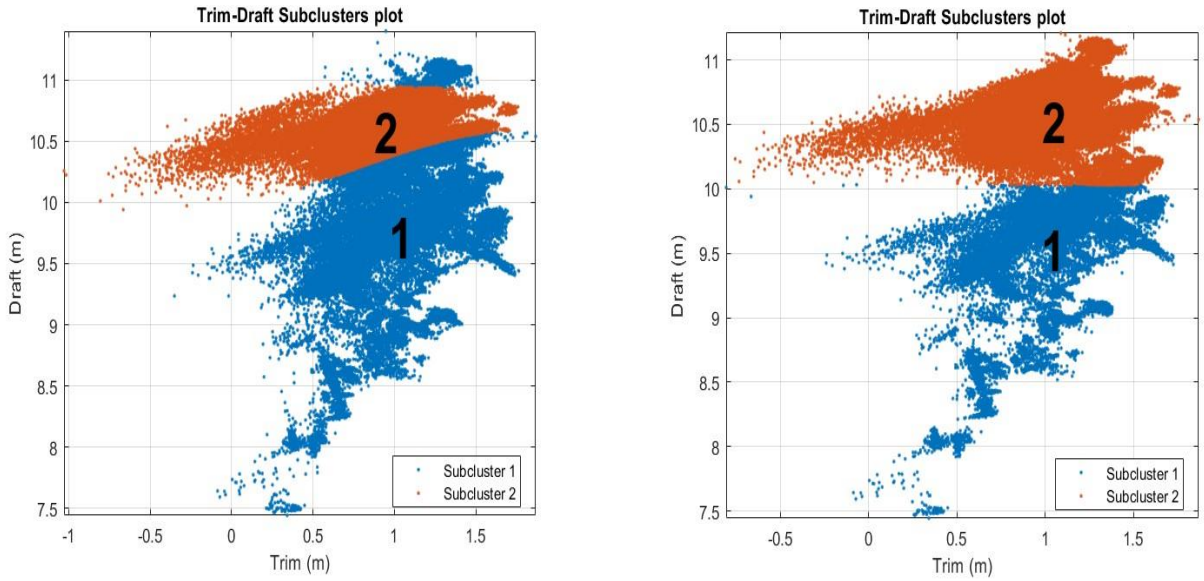

Figure 42: Subclusters plot of Trim-Draft variables concerning Transient Speed Region. After GMMS (on the left) and K-MEANS clustering (on the right).

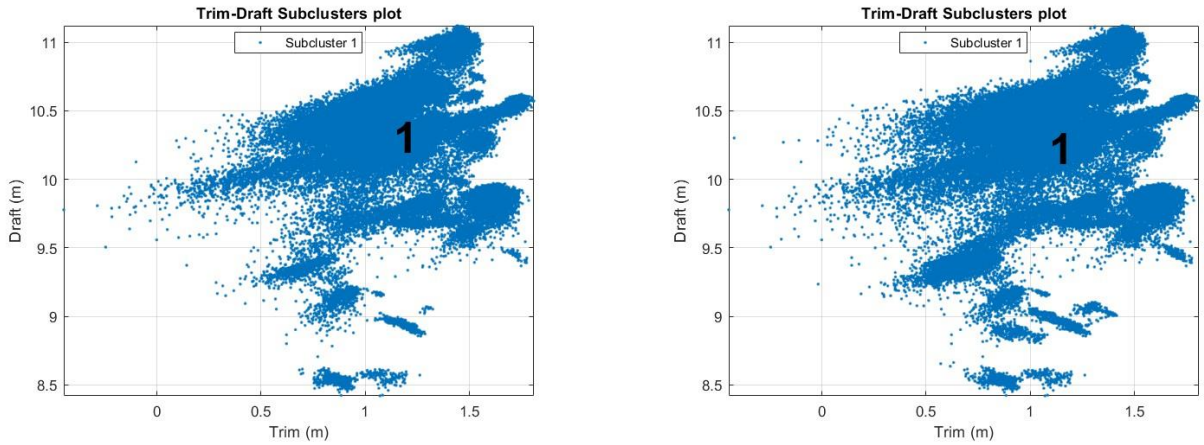

Figure 43: Subcluster plot of Trim-Draft variables concerning Service Speed Region. After GMMS (on the left) and K-MEANS clustering (on the right).

4.10 Ship performance quantification

The final goal is to provide a Ship performance index to measure the ship's performance under localized operational conditions. The calculation is being made concerning each trim-draft mode under the respective engine mode. The results are summarized in the table below for both K-Means and GMMS clustering methods. By comparing the results for each cluster, we can identify the best performance mode with the lowest KPI value. As described in chapter 3.6, the mean values of two different KPIs are calculated, KPIa and KPIb. The mathematical expression of these two key performance indexes is given down below:

$$KPIa_i = \frac{P_i}{n_i^3} \ , \left[\frac{Kw}{rpm^3}\right]$$

$$KPIb_i = \frac{FC_i}{D_i} \ , \ \left[\frac{Ton}{NM}\right]$$

i corresponds to the respective subcluster under the concerning engine mode.

Table 5: Ship performance quantification results.

		K-MEANS		GMMS	
		KPIa $\frac{Kw}{rpm^3}$	KPIb $\frac{Ton}{NM}$	KPIa $\frac{Kw}{rpm^3}$	KP1b $\frac{Ton}{NM}$
CLUSTER A	Subcluster 1	0.0175	0.0818	0.0173	0.0811
	Subcluster 2	0.0168	0.0860	0.0166	0.0855
	Subcluster 3	0.0166	0.0845	0.0161	0.0826
CLUSTER B	Subcluster 1	0.0173	0.1146	0.0175	0.1175
	Subcluster 2	0.0171	0.1128	0.0174	0.1157
CLUSTER C	Subcluster 1	0.0170	0.1374	0.0169	0.1385

After completing the investigation, it turns out that each key performance indicator results converge for K-Mean and GMM'S clustering concerning the ship's localized operational mode.

In relation to KPIa, subcluster 3 in the slow speed cluster, cluster A, is the most efficient in total and in the specific cluster since it has the lowest KPI value. For the transient speed region, cluster B, subcluster2, is the best performance trim-draft mode. As we pointed out before, for cluster C, there is only one trim-draft mode with a KPIa value (0.0170/0.0169) for KMEANS and GMM clustering, respectively.

According to KPIb, for cluster A, the slow speed region, subcluster 1, is the best performance mode and the best performance mode in total since it has the lowest KPI value. For cluster B, the transient speed region, subcluster 2, is the best performance trim-draft mode. Finally, as we pointed out before, for cluster C, there is only one trim-draft mode with a KPIb value (0.1374/0.1385) for KMEANS and GMM clustering, respectively.

4.11 Outlier evaluation algorithms

4.11.1 Outlier evaluation 1 algorithm

The first algorithm aims to examine the behavior of each detected outlier individually. A brief description of the construction and the visualization of the presented algorithm is given in chapter 3.8.1. Three different cases of the proposed algorithm are presented here for two variables. Firstly, the outlier evaluation 1 algorithm results are displayed for one outlier point, followed by the respective outcome when three consecutive outliers are investigated. Finally, the results of 5 successive detected outliers are presented to highlight the flexibility the specific algorithm offers when examining the detected outliers in a time-series format. The propeller shaft power variable is investigated in this case. The presented algorithm also evaluates outliers in other variables efficiently. Note that the investigated outliers emerged in the service speed cluster after GMM'S model implementation.

Firstly, the reasonable and unreasonable outliers of the aforementioned cases for the propeller shaft power variable are presented in Figure 44/Figure 45 for an individual outlier, Figure 46/Figure 47 for three successive outliers, and Figure 48/Figure 49 for five successive outlier points.

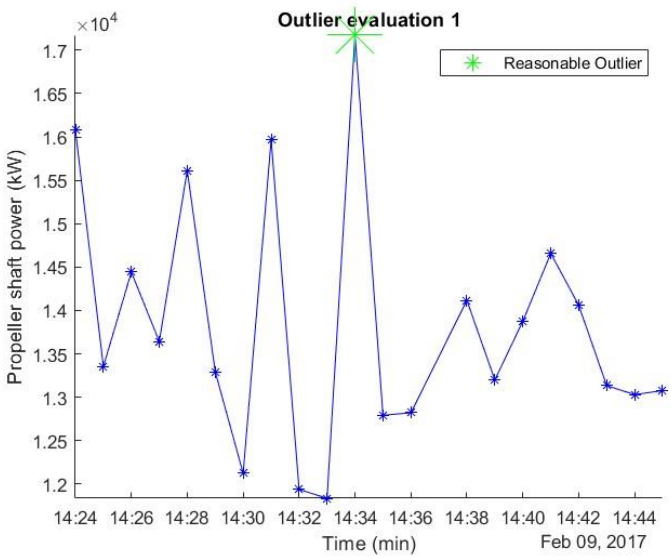

Figure 44: Time series plot of a reasonable individual outlier concerning propeller shaft power measurements.

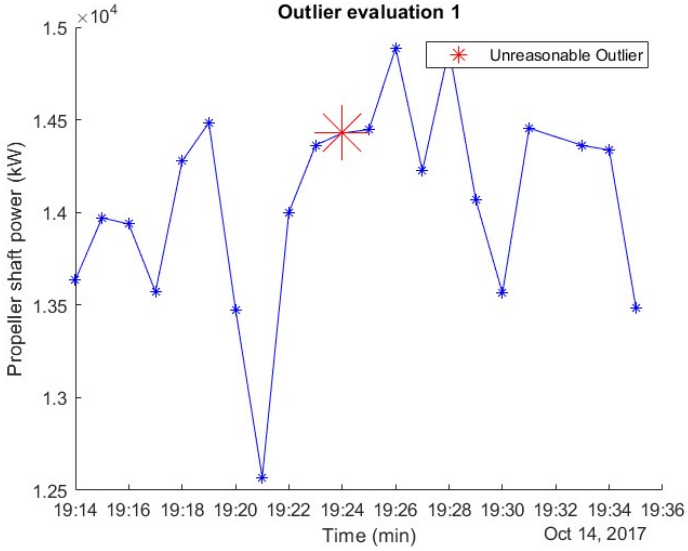

Figure 45: Time series plot of an unreasonable individual outlier concerning propeller shaft power measurements.

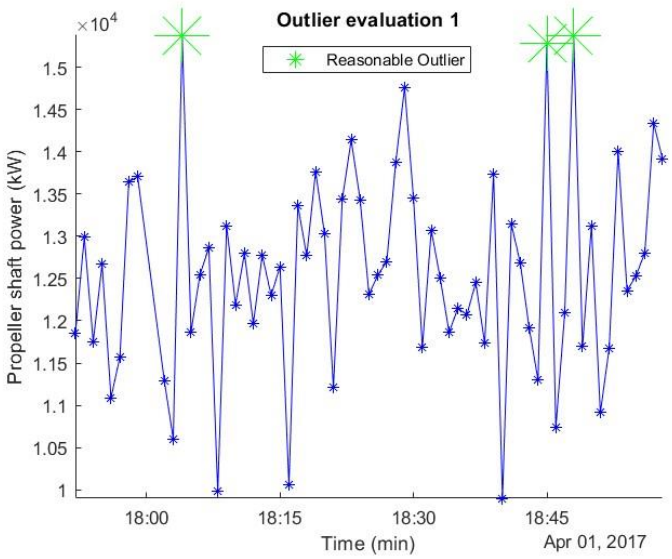

Figure 46: Time series plot of three reasonable successive outliers concerning propeller shaft power measurements.

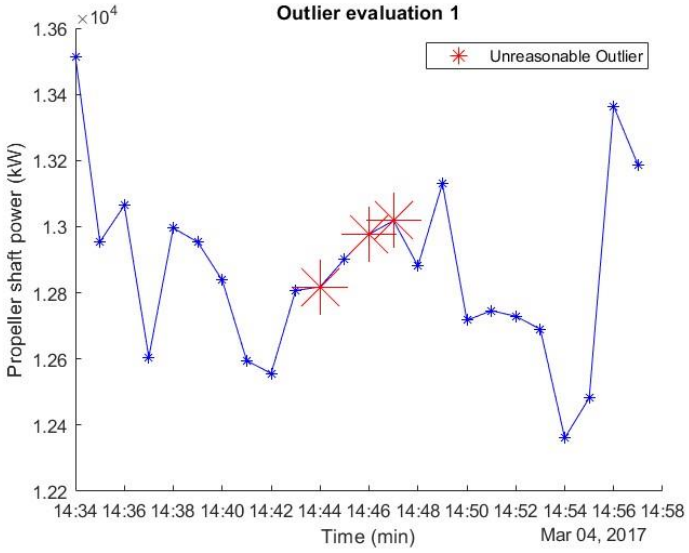

Figure 47: Time series plot of three unreasonable successive outliers concerning propeller shaft power measurements.

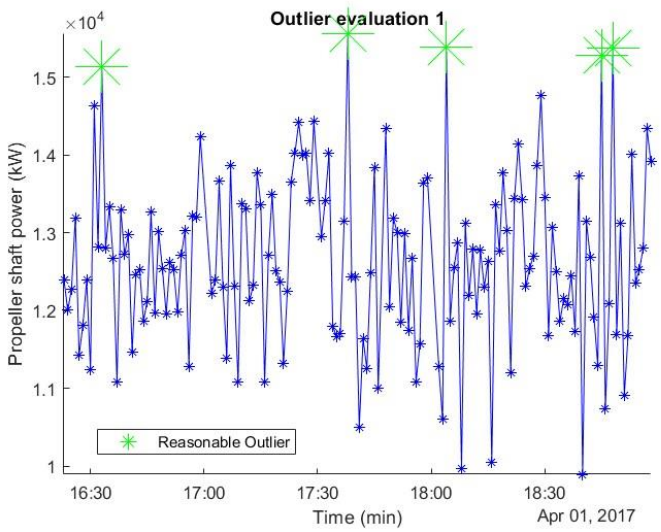

Figure 48: Time series plot of five reasonable successive outliers concerning propeller shaft power measurements.

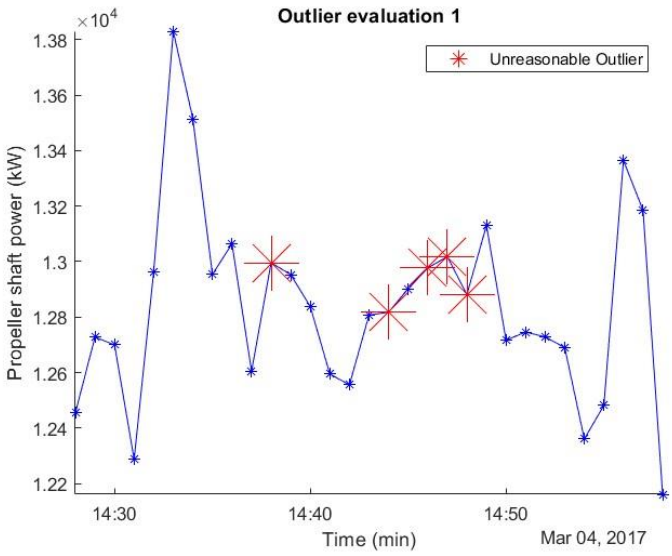

Figure 49: Time series plot of five unreasonable successive outliers concerning propeller shaft power measurements.

In Figure 44, Figure 46, and Figure 48, the ''reasonable'' outlier data points are presented in green color. We notice sudden changes between the identified anomalies and their consecutive points on these occasions. So, the second anomaly detector correctly characterized these points as outliers. In Figure 45, Figure 47, and Figure 49, we present the ''unreasonable'' outlier data points in red. It is obvious that even though they were marked as outliers, there are no sudden changes between the identified anomalies and their consecutive points. So, the second anomaly detector possibly wrongly indicated these measurements as outliers. Please note that the unreasonable outliers must be further investigated before we can draw any conclusions. As the second anomaly detector examines and characterizes the data points combined for propeller shaft speed-power variables, it projects them onto the principal component axis to determine the anomalies.

4.11.2 Outlier evaluation 2 algorithm

As noted in chapter 3.8.2, the outlier's evaluation 2 algorithm points to a broader visual analysis and investigation of the second anomaly detector efficiency. It focuses more on identifying the general trend of the outlier data and comparing it with the general trend of the regular data points. As illustrated in the figure below, the total data set is plotted in a time series format and divided into eleven groups of data regarding the propeller shaft speed and the propeller shaft power variables. The goal is first to identify the time position of the detected outliers and bin the data points that belong to the same group. Then the binning results are plotted in a histogram, as presented in Figure 50. The identified outliers are shown as discrete signal points. Note that the groups with a high frequency in outlier points are marked with red color. The red color gets more intense as the outliers rise in frequency in the specific group. In Figure 50, the identified outliers in the service speed cluster after GMM'S implementation are investigated.

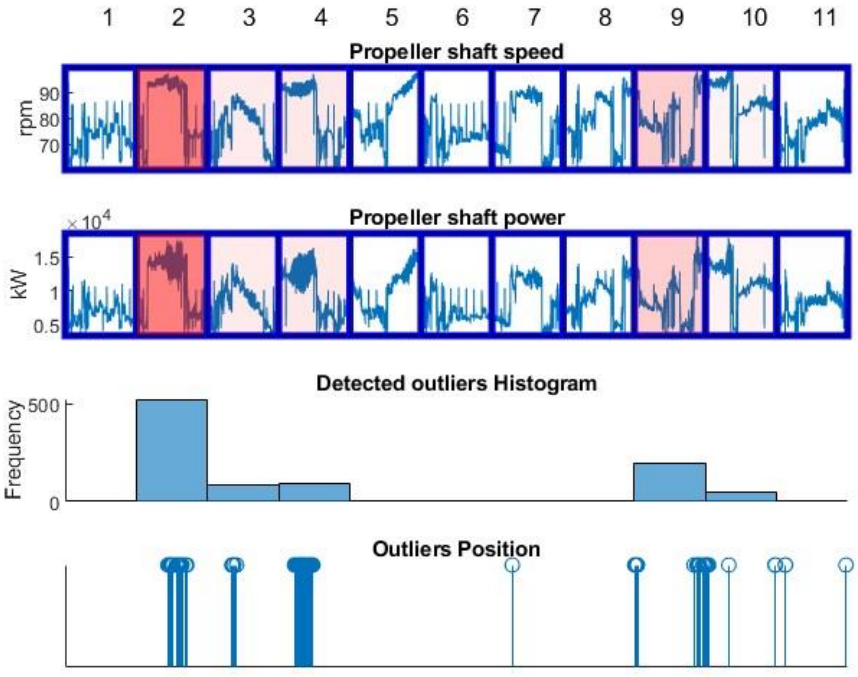

Figure 50: Graphical representation of the outlier evaluation 2 algorithm in the service speed cluster after GMMS implementation.

Most outliers are detected in group 2, while some are in groups 3, 4, 9, and 10. The same results arise when the outlier evaluation 2 algorithm is utilized in the service speed cluster after the K-means algorithm implementation. So the majority of the identified outliers emerged under a specific time period. That strongly indicates that the marked outliers didn't come up by chance, but possibly there is a systemic failure in the data collection system under these periods. The reason for that systemic failure will be investigated in PART C, in which the correlation between the ship's main operational parameters and the number of the identified outlier data points concerning each data group is calculated.

4.11.3 Correlation matrices

Firstly the correlation matrix between the number of the identified outliers and the measurements of seven operational variables concerning each data group is presented. The presented operational variables are wind speed (W.SPE), draft (DRAFT), trim (TRIM), speed over ground (SOG), fuel oil consumption (FOC), propeller shaft power (PSP), propeller shaft speed (PSS). In the presented matrix, we notice a moderate positive correlation between the number of identified outliers and the wind speed measurements, a negligible correlation between the number of identified outliers and the draft, trim, and speed over ground variables, and a low positive correlation between the number of identified outliers and the fuel oil consumption, propeller shaft power-speed variables. In this case, we can not make any safe conclusions about the connection between the number of identified outliers and the measurements of the presented operational variables.

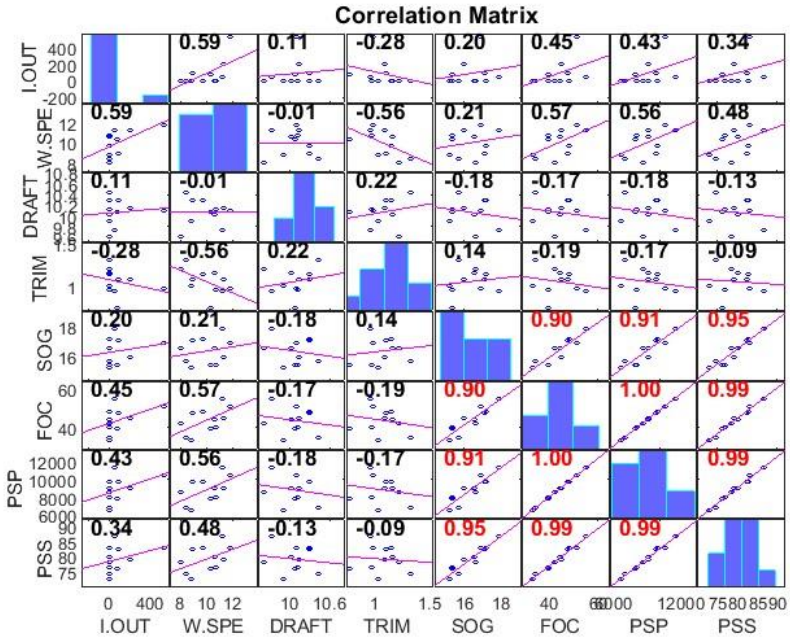

Figure 51: Correlation matrix between the number of identified outlier and seven operational parameters of the investigated data set.

The investigation extends to a greater degree in which the correlation between the variability in each main operational parameter and the number of the identified outliers is calculated for the eleven date-time groups. As described in chapter 3.8.2, four variability measures are utilized in the specific application, including variance (VAR), standard deviation (ST.DE), range (RANGE), and Interquartile range (IQR).

The results are presented down below:

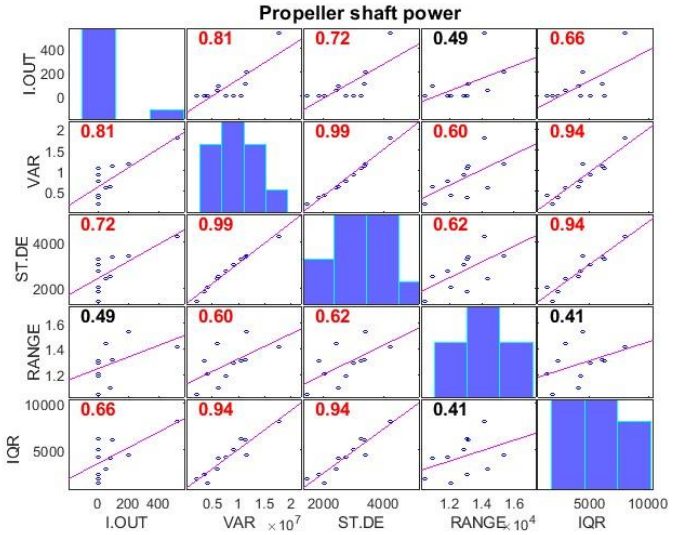

Figure 52: Correlation matrix between the number of identified outliers and the measured variability in the propeller shaft power values.

In Figure 52, we notice a moderate positive to a high positive correlation between the number of identified outliers and the calculated variability measures of the propeller shaft power values. That's a strong indication that higher variability to the propeller shaft power measurements leads to more identified outlies by the second anomaly detector.

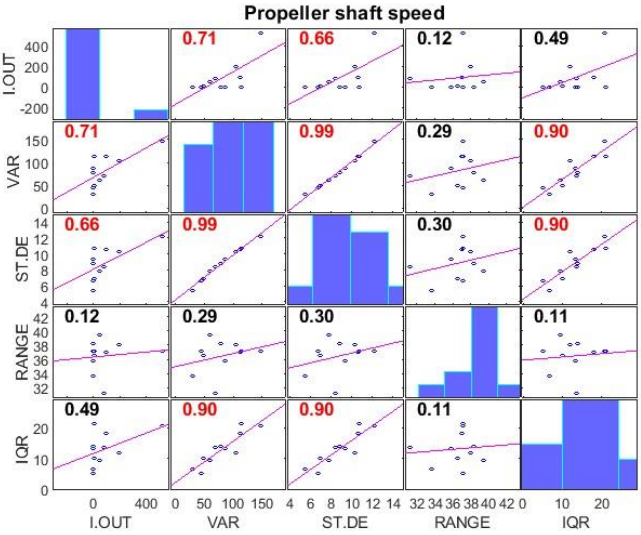

Figure 53: Correlation matrix between the number of identified outliers and the measured variability in the propeller shaft speed values.

In Figure 53, we notice a low to moderate positive correlation between the number of identified outliers and the calculated variability measures of the propeller shaft speed values. That may suggest that higher variability to the propeller shaft speed measurements leads to more identified outlies by the second anomaly detector.

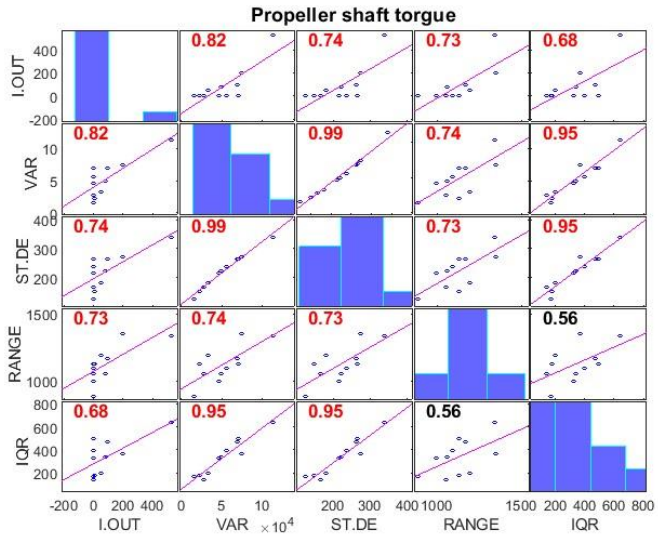

Figure 54: Correlation matrix between the number of identified outliers and the measured variability in the propeller shaft toque values.

In Figure 54, we notice a high positive correlation between the number of identified outliers and the calculated variability measures of the propeller shaft torque values. That strongly indicates that higher variability to the propeller shaft torque measurements leads to more identified outlies by the second anomaly detector.

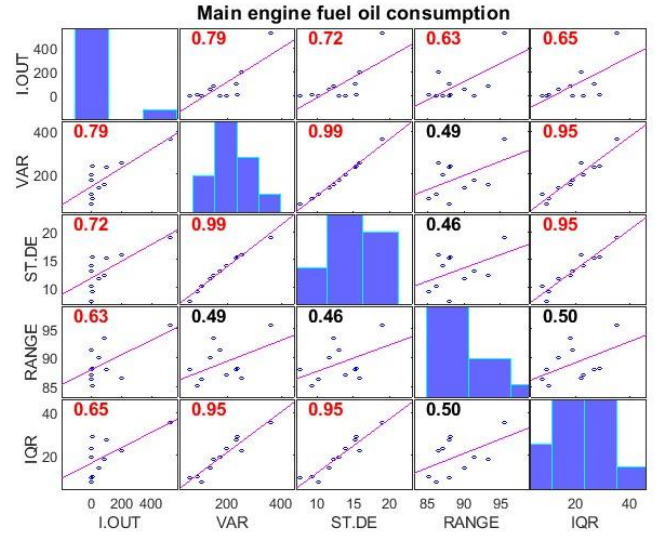

Figure 55: Correlation matrix between the number of identified outliers and the measured variability in the Main engine's fuel oil consumption values.

In Figure 55, we notice a moderate positive to a high positive correlation between the number of identified outliers and the calculated variability measures of the Main engine's fuel oil consumption values. That's a strong indication that higher variability to main engine fuel oil consumption measurements leads to more identified outliers by the second anomaly detector.

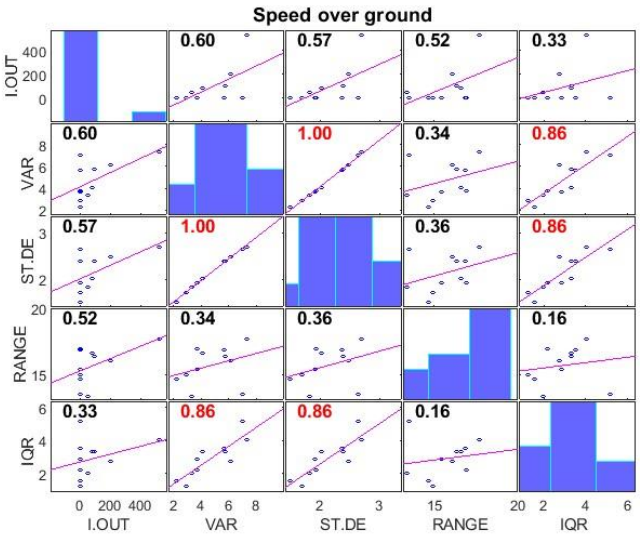

Figure 56: Correlation matrix between the number of identified outliers and the measured variability in speed over ground values.

In Figure 56, we notice a low positive to moderate positive correlation between the number of identified outliers and the calculated variability measures of the Speed over ground values. That may suggest that there is an underlying connection between the number of identified outliers and the variability in the Speed over ground measurements.

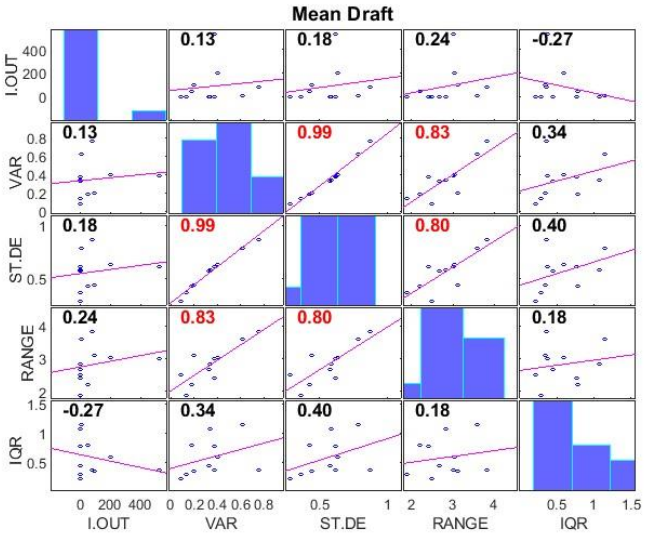

Figure 57: Correlation matrix between the number of identified outliers and the measured variability in Mean draft values.

In Figure 57, we notice a negligible correlation between the number of identified outliers and the calculated variability measures of the Mean Draft values. That strongly indicates that there is not an actual connection between the number of identified outliers and the variability in the mean draft measurements.

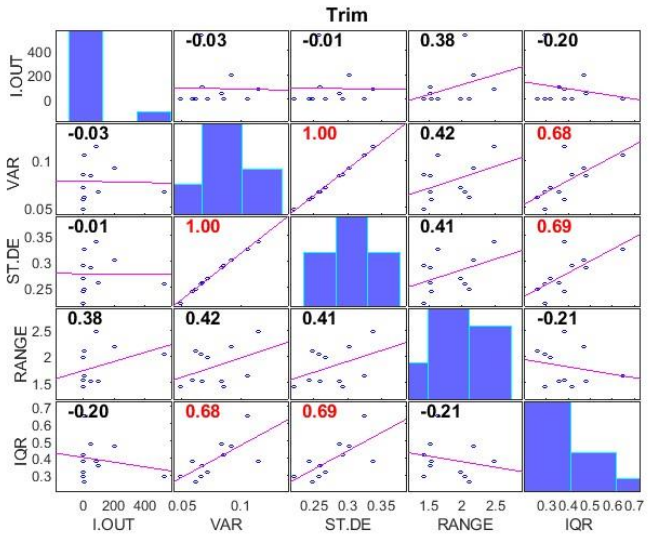

Figure 58: Correlation matrix between the number of identified outliers and the measured variability in Trim values.

In Figure 58, we notice a negligible to low positive correlation between the number of identified outliers and the calculated variability measures of the Trim values. Consequently, we can not set up any connection between the number of identified outliers and the variability of the Trim measurements.

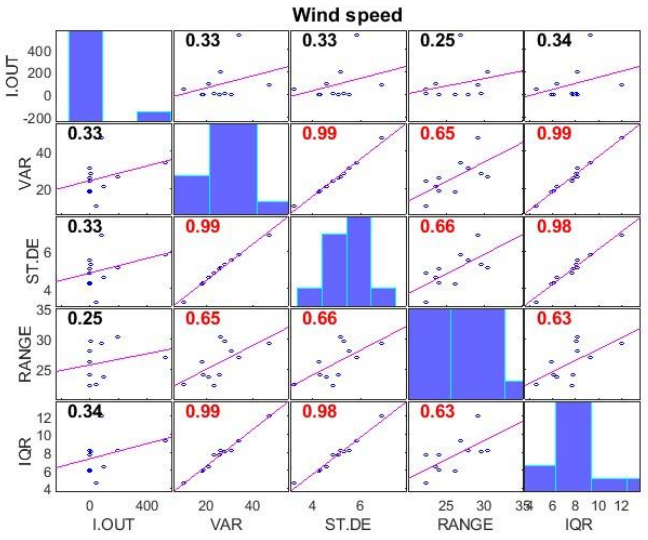

Figure 59: Correlation matrix between the number of identified outliers and the measured variability in wind speed values.

In Figure 59, we notice a negligible to low positive correlation between the number of identified outliers and the calculated variability measures of the wind speed values. Consequently, we can not set up any connection between the number of identified outliers and the variability of the wind speed measurements.

After presenting all the correlation matrices between the number of identified outliers and the variability of the ship's main operational parameters, we lead to some final conclusions. Firstly there is a strong indication that the number of identified outliers largely depends on the variability of the main engine's operational parameters, such as propeller shaft power, propeller shaft speed, propeller shaft torque, and main engine fuel oil consumption. There is a possible connection between the number of identified outliers and the varibility in speed over ground measurements, which is closely correlated with the main engine's operational parameters. Finally, we can not establish any actual connection between the number of identified outliers and the variability in wind speed, trim, and draft measurements. The conclusions mentioned above are desirable and foreseeable since the second anomaly detector is based on engine data measurements to mark and isolate the outliers points.

5. Conclusions

The intention and purpose of this thesis was twofold. First, we constructed a data preprocessing framework able to detect and isolate erroneous data points and identify underlying data patterns concerning the ship's localized operational conditions in the investigated data set. To achieve this, data density plots were deployed along with two different clustering techniques and domain knowledge to quantify the ship's operational behavior. Also, two distinct clustering evaluation methods were presented to determine the efficiency of the applied clustering techniques. Furthermore, domain knowledge coupled with principal component analysis was utilized to detect and isolate erroneous data measurements aiming to improve the quality of the investigated data set. Also, two outlier evaluation algorithms were constructed to assess the results of the applied outlier detection technique and to identify possible connections between the identified outliers and the behavior of the ship's main operational parameters. Secondly, two key performance indicators were proposed to quantify the ship's performance under the specified operational conditions. The following are the main conclusions derived from this analysis:

- Domain knowledge utilization in every step of the process is decisive to the construction and evaluation of the presented framework and should gain more attention in the scientific literature.
- The applied minimum-maximum thresholds in the data set's variables should be defined carefully since they have a significant impact on the final results of the applied framework.
- Histograms, scatter plots, and data density plots can be insightful in identifying underlying data patterns and quantifying a ship's operational behavior.
- The k-means algorithm is a relatively easily implemented and time-efficient clustering technique since the computational cost is low and the execution time is short. But the k-means algorithm can not identify complex data patterns that may exist in the examined data set.
- Gaussian mixture models can identify complex data patterns and provide more accurate and realistic results. But it can be hard to implement and time-consuming, especially in large data sets, since the computational cost is high and the execution time is long.
- The k-means algorithm can be used in cases where the time limits are strict, the computational capacity is low, and there is no demand for highly accurate results. Gaussian mixture models can be used when there are extended time limits, high computational capacity, and a need for high-quality, accurate results.
- Clustering evaluation methods can be very insightful about the final clustering results and the existing number of clusters in the examined dataset.
- The principal component analysis is an effective and easily interpretable technique for identifying outlier data points in the respected data set.
- The two constructed outlier evaluation algorithms are very informative about the behavior of the detected outliers in a time series plot and capable of identifying possible causes of the occurring outlier points.

- The ship performance quantification under the ship's localized operational conditions, utilizing Key performance indexes, drastically increases our awareness about the ship's total and side performance efficiency.

Concerning the aforementioned, it is concluded that the utilization of proper data preprocessing techniques, coupled with domain knowledge, drastically improves the quality of the examined data sets and provides powerful insights about underlying data patterns and possible connections between the examined variables. Furthermore, ship performance quantification can also enhance our understanding of the investigated ship's operational behavior and performance efficiency. Also, the outlier evaluation algorithms efficiently validate the identified outliers and identify possible outlier causes. The constructed framework is, therefore, a versatile and functional tool for ship operators and managers.

Based on the present thesis and the above conclusions, some suggestions for further research are listed.

- Try to define the minimum - maximum thresholds of the examined parameters that apply to maneuvering conditions concerning the engine propeller combinator diagram and ship's position data.

- Consider adding more operational variables than the two in the specific investigation in order to extend the clustering analysis to a higher dimensional space.

- Perform principal component analysis to identify outlier points in a more detailed data set by adding extra variables in the respective study.

- Utilize the ship's position data to perform data preprocessing and ship performance quantification techniques between voyages.

In conclusion, data preprocessing and analysis can play a key role in improving a ship's operational performance. So, more research needs to be done in that direction. In addition, more applicable and flexible algorithms and tools need to be constructed to serve each investigation's application-specific needs.

References

[1] A. Shameer, S. P. Sankar S, and A. Athilafsal, “Predictive Analysis Using Big Data In Marine Industry,” 2022, doi: 10.22541/au.166299557.79115471/v1.

[2] “Data analytics in maritime/shipping.” https://marine-digital.com/article_data_analytics_in_maritime (accessed Oct. 15, 2022).

[3] I. Zaman, K. Pazouki, R. Norman, S. Younessi, and S. Coleman, “Challenges and opportunities of big data analytics for upcoming regulations and future transformation of the shipping industry,” in *Procedia Engineering*, 2017, vol. 194, pp. 537–544. doi: 10.1016/j.proeng.2017.08.182.

[4] FRANKENFIELD JAKE, “Data Analytics: What It Is, How It’s Used, and 4 Basic Techniques.” https://www.investopedia.com/terms/d/data-analytics.asp (accessed Oct. 15, 2022).

[5] Lawton George, “Data Preprocessing: Definition, Key Steps and Concepts.” https://www.techtarget.com/searchdatamanagement/definition/data-preprocessing (accessed Oct. 15, 2022).

[6] L. P. Perera and B. Mo, “Marine engine-centered data analytics for ship performance monitoring,” *Journal of Offshore Mechanics and Arctic Engineering*, vol. 139, no. 2, Apr. 2017, doi: 10.1115/1.4034923.

[7] L. P. Perera and B. Mo, “Ship speed power performance under relative wind profiles in relation to sensor fault detection,” *Journal of Ocean Engineering and Science*, vol. 3, no. 4, pp. 355–366, Dec. 2018, doi: 10.1016/j.joes.2018.11.001.

[8] L. P. Perera, “Statistical Filter based Sensor and DAQ Fault Detection for Onboard Ship Performance and Navigation Monitoring Systems,” in *IFAC-PapersOnLine*, 2016, vol. 49, no. 23, pp. 323–328. doi: 10.1016/j.ifacol.2016.10.362.

[9] Ø. Ø. Dalheim and S. Steen, “Preparation of in-service measurement data for ship operation and performance analysis,” *Ocean Engineering*, vol. 212, Sep. 2020, doi: 10.1016/j.oceaneng.2020.107730.

[10] S. Węglarczyk, “Kernel density estimation and its application,” *ITM Web of Conferences*, vol. 23, p. 00037, 2018, doi: 10.1051/itmconf/20182300037.

[11] Rick Wicklin, “How to visualize a kernel density estimate - The DO Loop.” https://blogs.sas.com/content/iml/2016/07/27/visualize-kernel-density-estimate.html (accessed Oct. 12, 2022).

[12] Sauravkaushik8 Kaushik, “Clustering | Types Of Clustering | Clustering Applications.” https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/ (accessed Oct. 12, 2022).

[13] E. Patel and D. S. Kushwaha, “Clustering Cloud Workloads: K-Means vs Gaussian Mixture Model,” in *Procedia Computer Science*, 2020, vol. 171, pp. 158–167. doi: 10.1016/j.procs.2020.04.017.

[14] Imad Dabbura, “K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks | by Imad Dabbura | Towards Data Science.” https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a (accessed Oct. 12, 2022).

[15] K. Q. Bui and L. P. Perera, “Advanced data analytics for ship performance monitoring under localized operational conditions,” *Ocean Engineering*, vol. 235, Sep. 2021, doi: 10.1016/j.oceaneng.2021.109392.

[16] “Gaussian Mixture Models Explained | by Oscar Contreras Carrasco | Towards Data Science.” https://towardsdatascience.com/gaussian-mixture-models-explained-6986aaf5a95 (accessed Oct. 12, 2022).

[17] “K-Means Explained. Explaining and Implementing kMeans… | by Vatsal | Towards Data Science.” https://towardsdatascience.com/k-means-explained-10349949bd10 (accessed Nov. 22, 2022).

[18] Lavorini Vincenzo, “Gaussian Mixture Model clustering: how to select the number of components (clusters) | by Vincenzo Lavorini | Towards Data Science.” https://towardsdatascience.com/gaussian-mixture-model-clusterization-how-to-select-the-number-of-components-clusters-553bef45f6e4 (accessed Nov. 22, 2022).

[19] “Outlier Detection - Outlier Detection Techniques, Definition & Examples.” https://www.mygreatlearning.com/blog/what-is-outlier-detection/ (accessed Oct. 12, 2022).

[20] “Outliers explained: a quick guide to the different types of outliers | by Ira Cohen | Towards Data Science.” https://towardsdatascience.com/outliers-analysis-a-quick-guide-to-the-different-types-of-outliers-e41de37e6bf6 (accessed Oct. 12, 2022).

[21] Zakaria Jaadi, “Principal Component Analysis (PCA) Explained | Built In.” https://builtin.com/data-science/step-step-explanation-principal-component-analysis (accessed Oct. 12, 2022).

[22] D. Keim, G. Andrienko, J.-D. Fekete, C. Görg, J. Kohlhammer, and G. Melançon, “Visual Analytics : Definition, Process, and Challenges”, [Online]. Available: http://nbn-resolving.de/urn:nbn:de:bsz:352-opus-68555

Appendix A: Engine data clustering investigation

A time series plot of the speed over ground variable concerning the identified engine mode clusters is presented. Each point is plotted with a different color based on the cluster assignment. A polynomial fitting is also applied to the specific investigation to give us a greater sense of the behavioral patterns of the respective data.

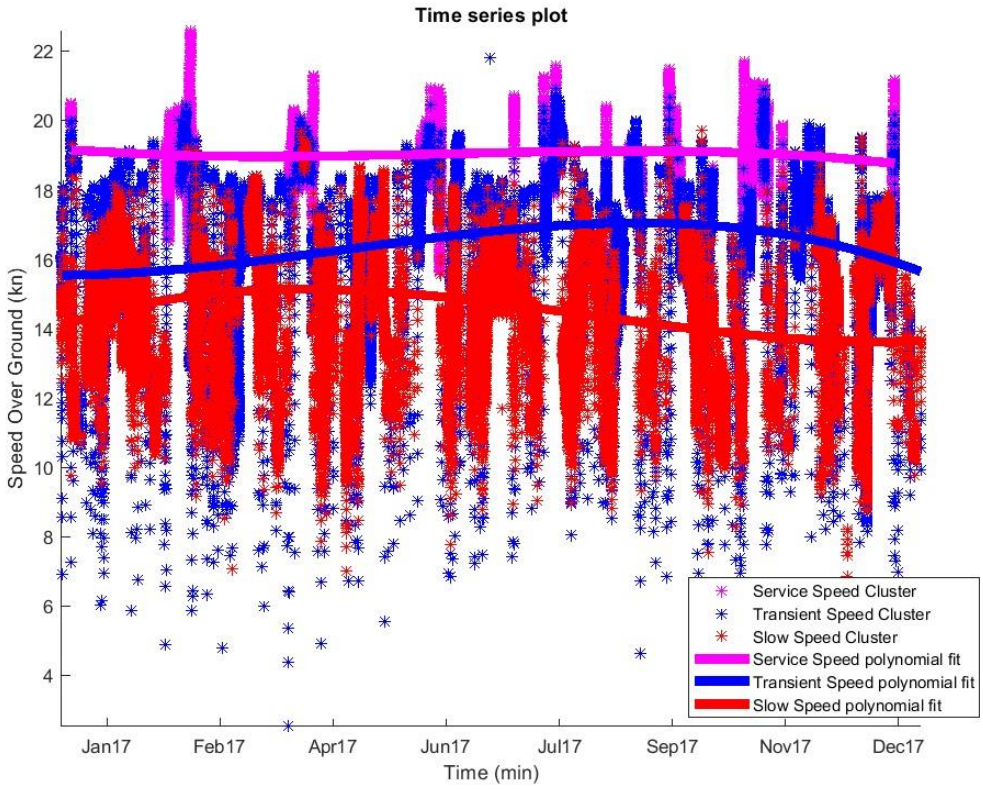

Figure 60: Time series plot of speed over ground variable concerning the identified engine data clusters.

Appendix B: Cluster plots after the second anomaly detector implementation.

The graphical representations of Slow and Transient Speed Clusters for K-MEANS and GMMS clustering methods regarding inlier and identified outlier data points are presented next. The outlier points are marked with red color and the inlier points are marked with green color.

PART A: Cluster plots based on k-means algorithm

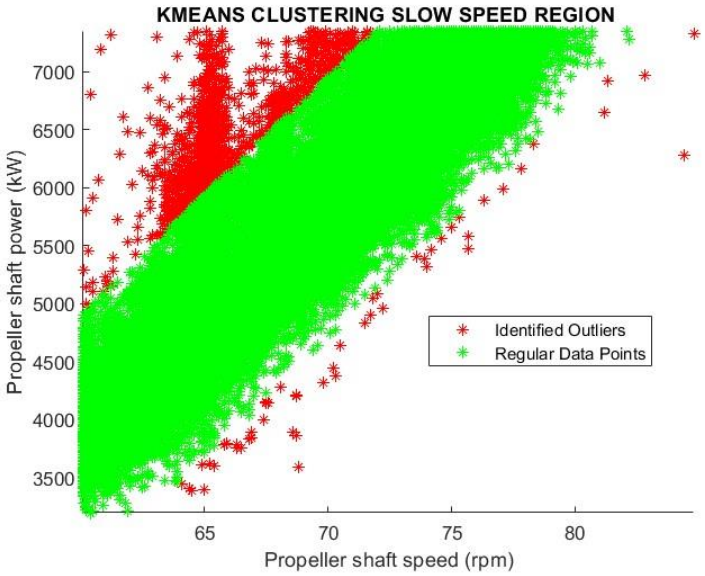

Figure 61: Graphical representation of Slow Speed Cluster after K-MEANS clustering regarding inlier and identified outlier data points.

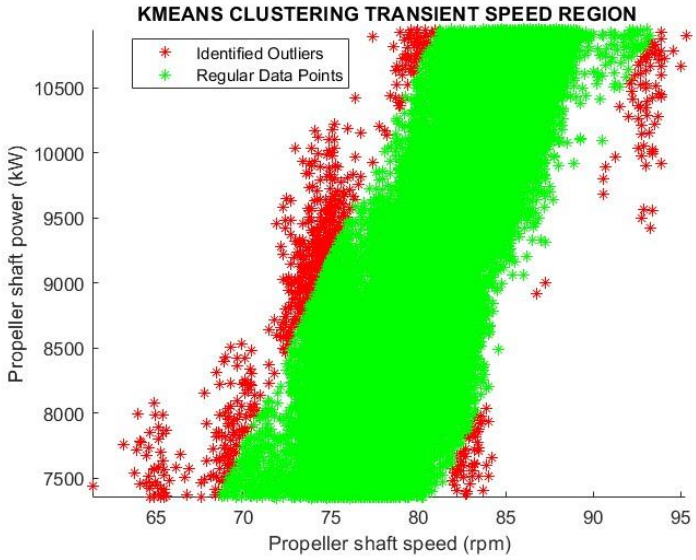

Figure 62: Graphical representation of Transient Speed Cluster after K-MEANS clustering regarding inlier and identified outlier data points.

PART B: Cluster plots based on gaussian mixture models

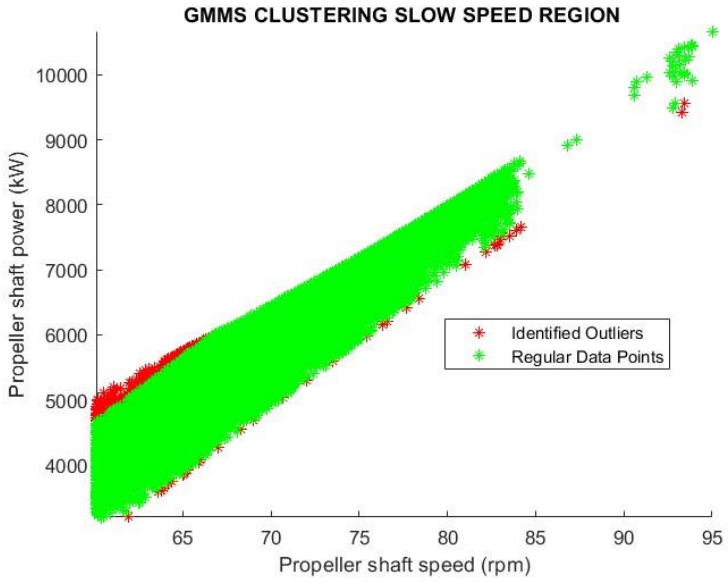

Figure 63: Graphical representation of Slow Speed Cluster after GMM'S clustering regarding inlier and identified outlier data points.

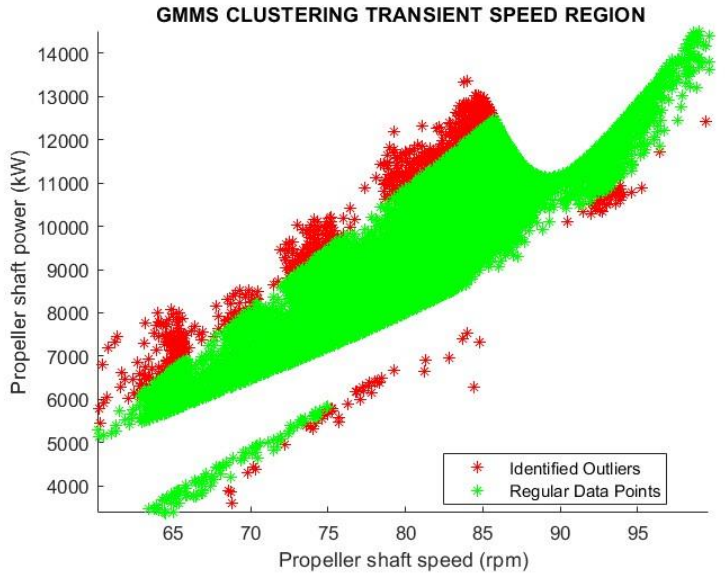

Figure 64: Graphical representation of Transient Speed Cluster after GMM'S clustering regarding inlier and identified outlier data points.

Appendix C: Exploration of the ship's localized operational conditions.

Based on K-MEANS and GMMS clustering, the engine mode clusters (cluster A, cluster B, cluster C) are being analyzed. The results of the implemented data density estimation methods are shown in the following figures. Bivariate Histograms and Kernel Density Estimation plots are presented.

PART A: Slow speed cluster (cluster A) investigation.

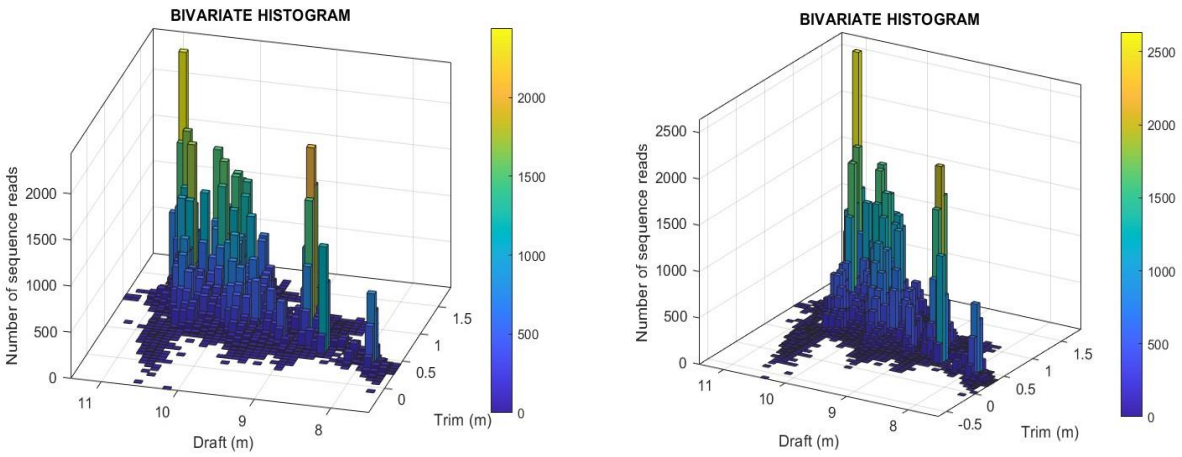

Figure 65: Bivariate histogram of trim/draft variables in Slow Speed Cluster. After GMMS (on the left) and K-MEANS clustering (on the right).

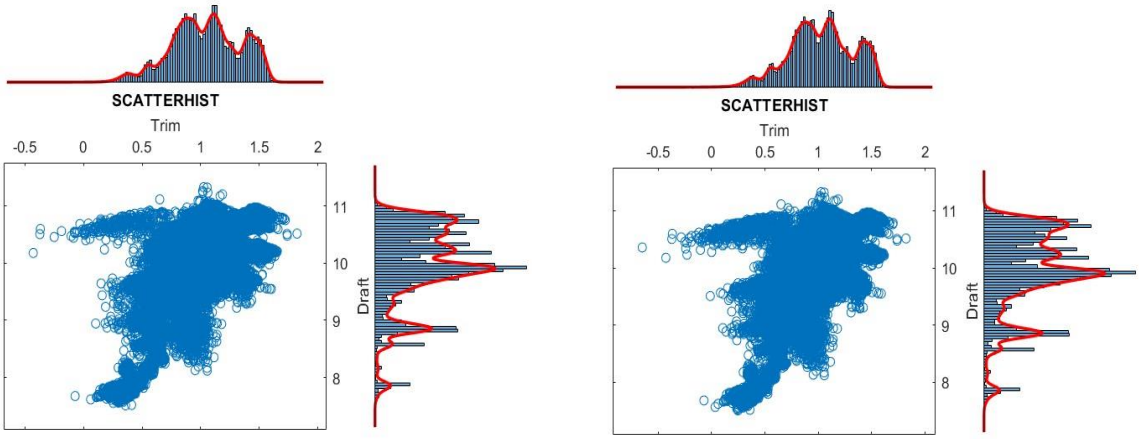

Figure 66: Scatterplot Combined with univariate Histograms and kernel Density Estimation plots for trim/draft variables of Slow Speed Cluster. After GMMS (on the left) and K-MEANS clustering (on the right).

PART B: Transient speed cluster (cluster B) investigation

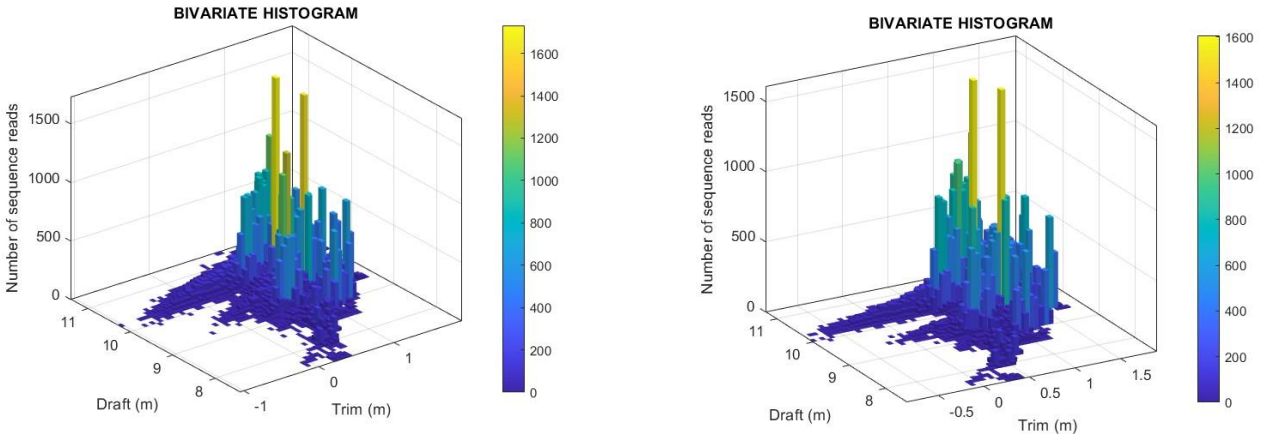

Figure 67: Bivariate histogram of trim/draft variables in Transient Speed Cluster. After GMMS (on the left) and K-MEANS clustering (on the right).

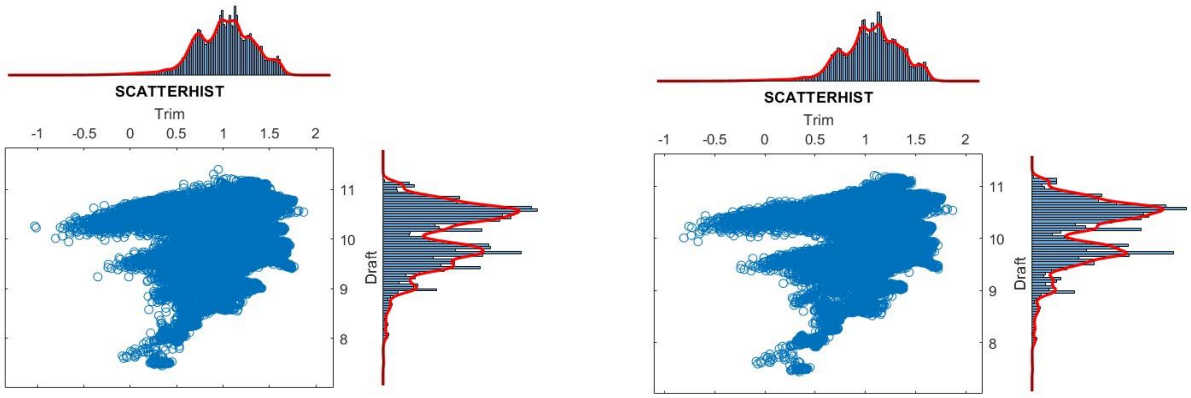

Figure 68: Scatterplot Combined with univariate Histograms and kernel Density Estimation plots for trim/draft variables of Transient Speed Cluster. After GMMS (on the left) and K-MEANS clustering (on the right).

PART C: Service speed cluster (cluster C) investigation.

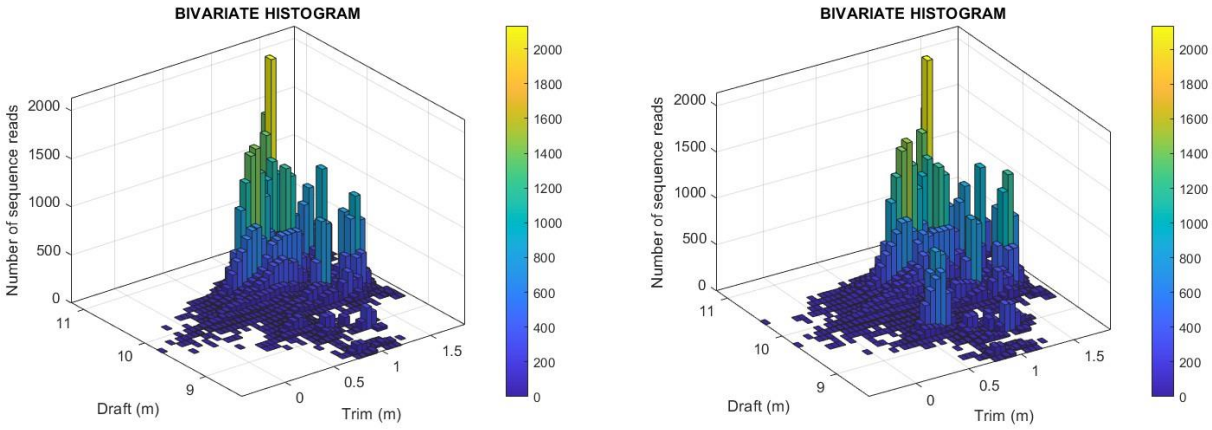

Figure 69: Bivariate histogram of trim/draft variables in Service Speed Cluster. After GMMS (on the left) and K-MEANS clustering (on the right).

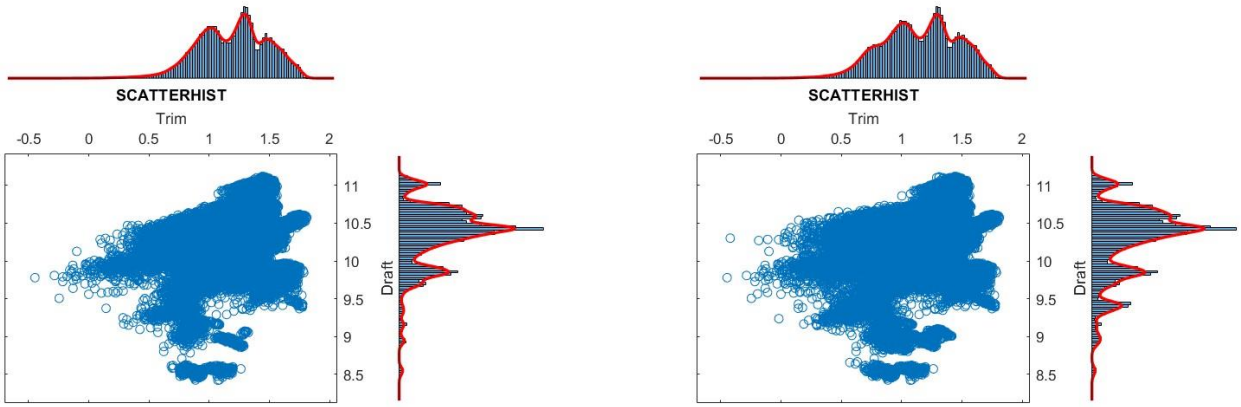

Figure 70: Scatterplot combined with univariate histograms and kernel Density Estimation plots for trim/draft variables of Service Speed Cluster. After GMMS (on the left) and K-MEANS clustering (on the right).