Ocean Engineering 235 (2021) 109392

Contents lists available at ScienceDirect

Ocean Engineering

journal homepage: www.elsevier.com/locate/oceaneng

Advanced data analytics for ship performance monitoring under localized operational conditions

Khanh Q. Bui [a,b,*], Lokukaluge P. Perera [a]

[a] Department of Technology and Safety, UiT The Arctic University of Norway, Tromsø, Norway
[b] Faculty of Navigation, Vietnam Maritime University, Hai Phong, Viet Nam

ARTICLE INFO

Keywords:
Big data analytics
Machine learning
Ship performance monitoring
Energy efficiency
Emission control
Data anomaly detection

ABSTRACT

Improving the operational energy efficiency of existing ships is attracting considerable interests to reduce the environmental footprint due to air emissions. As the shipping industry is entering into Shipping 4.0 with digitalization as a disruptive force, an intriguing area in the field of ship's operational energy efficiency is big data analytics. This paper proposes a big data analytics framework for ship performance monitoring under localized operational conditions with the help of appropriate data analytics together with domain knowledge. The proposed framework is showcased through a data set obtained from a bulk carrier pertaining the detection of data anomalies, the investigation of the ship's localized operational conditions, the identification of the relative correlations among parameters and the quantification of the ship's performance in each of the respective conditions. The novelty of this study is to provide a KPI (i.e. key performance indicator) for ship performance quantification in order to identify the best performance trim-draft mode under the engine modes of the case study ship. The proposed framework has the features to serve as an operational energy efficiency measure to provide data quality evaluation and decision support for ship performance monitoring that is of value for both ship operators and decision-makers.

1. Introduction

International shipping is an indispensable sector for the facilitation of global economy since it is responsible for about 80% of the total volume of global trade (UNCTAD, 2019). Furthermore, seaborne transportation is recognized as the most energy-efficient mode of cargo transport as regards energy use per unit transported. Nonetheless, considering its scale and current growth rate, the shipping industry is a major catalyst for global ecological change (Balcombe et al., 2019). According to the Fourth Greenhouse Gas (GHG) Study published by the International Maritime Organization (IMO), global anthropogenic emissions from shipping increased by approximately 10% from 2012 to 2018 (IMO, 2020). It is envisaged that shipping emissions will rise between 90% and 130% by 2050 relative to 2008 for long-term economic and energy scenarios. Therefore, shipping CO_2 emissions are increasing. By way of illustration, if the maritime sector had been treated as a country, it would have been the sixth largest CO_2 emitter in 2015 (Olivier et al., 2016).

Such environmental concerns have been acknowledged in a number of regulatory frameworks established by the IMO. GHG emissions from shipping are addressed by energy efficiency measures under Annex VI of the International Convention for the Prevention of Pollution from Ships (MARPOL). In response to the Paris agreement, the IMO set out an Initial IMO Strategy on reducing GHG emissions from ships, aiming at reducing the total annual GHG emissions by at least 50% by 2050, compared to 2008 levels. These increasingly stringent regulations have exerted pressure on the shipping industry to pursue possible avenues of reducing its environmental footprint (Perera et al., 2021). In order to achieve this, finding alternative fuel sources has been paid much attention in the industry. The search for the right future fuel is challenging since it is a multi-faceted problem where the evaluation of a pallet of different alternative options is influenced by multiple criteria, such as technical, economic, environmental, and social criteria (Bui and Perera, 2019; Bui et al., 2020). In addition to fuel changes, it is an orthodox norm that reducing fuel consumption or improving energy efficiency is an effective solution to reduce ship emissions due to the fact that GHG emissions from internal combustion engines are directly related to ship fuel consumption.

Energy efficiency improvement solutions are generally divided into technical and operational measures. The former refers to improvements made throughout the ship design phase, such as hull form optimization, air lubricant, propulsion efficiency devices, waste heat recovery

* Corresponding author at: Department of Technology and Safety, UiT The Arctic University of Norway, Tromsø, Norway.
E-mail addresses: khanh.q.bui@uit.no (K.Q. Bui), prasad.perera@uit.no (L.P. Perera).

https://doi.org/10.1016/j.oceaneng.2021.109392
Received 21 February 2021; Received in revised form 19 May 2021; Accepted 25 June 2021

technology (Brynolf et al., 2016); the latter refers to measures including optimal handling of ships (e.g., trim and ballast optimization), voyage optimization (e.g., weather routing, slow steaming, just-in-time arrival), and good maintenance practices for engine, hull and propeller (Ölçer, 2018). It has been observed from the literature that there is still a large potential for increasing energy efficiency from operational practices, thereby reducing CO_2 emissions. For example, voyage optimization has the potential effect on CO_2 emissions reduction at the figure of up to 48% (Bouman et al., 2017). Nonetheless, technical support systems, ship performance monitoring systems are required to facilitate this practice (IMO, 2014; Viktorelius and Lundh, 2019).

It is a widely held view that the shipping industry is on its way to the fourth industrial revolution (as known as Shipping 4.0) (Rødseth et al., 2016). The transformational role of digitalization and the rise of Artificial Intelligence (AI) together with Machine Learning (ML) will exert tremendous impacts on all of the aspects of the industry. Internet of things (IoT) with the utilization of sensor technologies as well as data acquisition systems can produce a massive amount of sensor data, referred to as big data, which can be used for analysis and further insights on ship performance monitoring. Therefore, proper techniques are required to leverage big data to support increased energy efficiency during ship operation (Zaman et al., 2017; Sullivan et al., 2020). In this respect, big data analytics have emerged as a disruptive technology that can be an operational energy efficiency measure under the ship performance monitoring systems.

The last few years have witnessed a considerable growth in the number of data-driven studies on improving ship energy efficiency. Despite this interest, scant studies have applied big data analytics approach. In addition, several studies have failed to demonstrate significant advantages of domain knowledge in every step of data analysis workflow. The term "domain knowledge" means the domain-specific expertise of the field and it plays an important role in each step of a data analysis project, ranging from problem formulation, data collection, data pre-processing, modeling, to result interpretation. Therefore, the accuracy of data-driven models based on ML can be increased if domain knowledge is incorporated into such models.

Furthermore, concerns have arisen which call into question the quality of ship performance and navigation data. This problem is related to data veracity, which is one of the characteristics of big data, as known as 'the four V's of big data', including volume, velocity, variety, and veracity (Perera and Mo, 2017; Zaman et al., 2017). It should also be noted that knowledge and awareness of ship operators have been recognized as one of the energy efficiency gaps from the operational side (Kitada and Ölçer, 2015; Rasmussen et al., 2018).

Given the above-mentioned background, this paper aims to develop an advanced data analytics framework for ship performance monitoring under localized operational conditions, where domain knowledge is taken into account. The proposed framework will be able to serve as an operational energy efficiency measure to provide data quality evaluation and decision support for ship performance monitoring under the digitalization of the maritime industry.

The structure of this paper is organized as follow. Section 2 reviews the literature on ship's operational energy efficiency and data anomaly detection. The proposed methodology is described in Section 3. Results of the proposed methodology are reported in Section 5. The conclusions are drawn in Section 6.

2. Literature review

2.1. Ship's operational energy efficiency

On the question of improving operational energy efficiency, more attention in the literature has been given to the prediction of ship fuel consumption or engine power. In this regard, statistical models were deployed in several studies (Erto et al., 2015; Sasa et al., 2015). However, these parametric methods may have bias problems due to their assumptions on data distributions. Additionally, they have failed to cope with complicated and non-linear data (Yan et al., 2020; Soner et al., 2018). Therefore, ML models have been widely developed to overcome these problems. In this context, a number of studies implemented ML models such as artificial neural networks (ANNs) (Petersen et al., 2012a,b; Bal Beşikçi et al., 2016; Farag and Ölçer, 2020; Karagiannidis and Themelis, 2021), regression models (Brandsæter and Vanem, 2018; Yan et al., 2020; Wang et al., 2018) and ensemble models (Soner et al., 2018; Gkerekos et al., 2019).

Engine speed optimization and trim optimization have also been gained attention in the literature in terms of improving operational energy efficiency. In this respect, there has been considerable interest in using big data analytics approach. Wang et al. (2017) made an attempt to achieve ship energy efficiency through a big data analysis based on Hadoop platform architecture. In this study, route division with regard to environmental factors was examined and speed optimization in different navigational segments of a route was investigated. Yan et al. (2018) proposed a big data analytics platform to analyze environmental factors for the purpose of optimizing engine speed for inland ships. This study applied the distributed parallel k-means clustering algorithm to obtain an elaborate route division and then find the optimal engine speed for the selected inland ship. Coraddu et al. (2017) employed a data analytics approach for fuel consumption prediction and trim optimization of a tanker. In this study, two gray box models were proposed as predictive models for the prediction of the fuel consumption. Based on these models, a trim optimization method of the tanker was developed. Lee et al. (2018) utilized weather archive big data to estimate the fuel consumption function for speed optimization in maritime logistics. In this study, they developed a decision support systems for minimizing fuel consumption while maintaining service level agreement by applying an optimization method called Particle Swarm Optimization.

It is probable that these ML-based studies have become the means to provide better prediction and decision support towards energy efficiency. Nonetheless, several studies have not treated domain knowledge in much detail. In this regard, Man et al. (2020) proposed an ethnographic method to identify operational challenges on using fuel monitoring systems. One of these challenges is the lack of effective analytical approaches for ship performance evaluation. This leads to a need for utilizing big data analytics in order to gain understanding of actual fuel consumption to achieve energy efficiency.

It has also been observed that many studies hold the view that ship speed is the major determinant for ship fuel consumption. Nonetheless, other factors including, among others, displacement, trim-draft conditions, loading conditions, environmental conditions, and navigation conditions also have impacts on ship fuel consumption (Tran, 2020; Yuan et al., 2017; Soner et al., 2019). It should be noted that these factors may pose a high dimensional challenge for data visualization as pointed out by Perera and Mo (2020).

2.2. Data anomaly detection

It is a self-evident fact that data collected from real-world sources are often impure. The so-called "garbage in – garbage out" (GIGO) refers to the fact that poor quality data input is associated with untrustworthy output (Pyle, 1999). This leads to the needs for methods that can be used for preparing quality data (i.e. data preprocessing) as a fundamental step during data analysis workflow (Zhang et al., 2003). Nevertheless, data quality awareness has yet to be reached its maturity in the maritime industry and a call for the industry to value and improve data quality. In addition, it is worth bearing in mind that the practicality of data quality cannot be done without considering domain knowledge.

In the literature, several taxonomies for data anomaly detection have been developed such as fault diagnosis, fault detection, and fault-tolerant control. Such taxonomies can be treated under decision support

systems and condition monitoring, aiming at enhancing reliability, safety, and energy efficiency of ship systems. Different approaches for the detection of possible faults in decision support systems of a container ship were proposed, i.e., the deployment of residuals and the generalized likelihood ratio (GLR) algorithm (Lajic and Nielsen, 2010), the deployment of Volterra theory (Lajic et al., 2009) and the deployment of a frequency domain-based model (Nielsen et al., 2012). Raptodimos and Lazakis (2018) proposed a method based on the integration of ANNs and Self Organising Maps (SOM) along with inter-clustering for data clustering and fault diagnosis of measurement data of physical parameters of a ship main engine cylinder. Vanem and Brandsæter (2019) deployed unsupervised learning techniques for data anomaly detection for sensor-based condition monitoring for a marine diesel engine.

Capezza et al. (2019) developed a model based on the combination of partial least squares (PLS) regression and prediction error control charts for monitoring of fuel consumption and diagnosis of faults. Lazakis et al. (2019) investigated the utilization of Support Vector Machine (SVM) for the detection of deviant and abnormal ship machinery conditions. Dalheim and Steen (2020) developed a data preparation toolbox for time series data in order to improve the quality of ship operation and performance analysis. Cheliotis et al. (2020) proposed a method based on Expected Behavior (EB) models in combination with Exponentially Weighted Moving Average (EWMA) control charts for early faults detection in the main engine of a ship. Karagiannidis and Themelis (2021) demonstrated that their proposed algorithms for replacing and cleaning data were be able to increase the accuracy of their produced ANN models.

2.3. Research contribution

The research studies reviewed in the previous section point to the following research drawbacks. First, the use of ANNs have been observed in several studies albeit its shortcomings. The most fundamental shortcoming of this approach has been clearly recognized as a 'black-box' approach and it is challenging to interpret behavior of the network. Second, the contribution of domain knowledge has received little attention within the context of maritime applications. Third, most of the studies reviewed have not been able to take into account correlations between factors contributing to ship fuel consumption in a high-dimensional data space. Fourth, data quality for ship operation and performance is still a neglected area in the maritime domain and few researchers have addressed this issue in the literature.

In order to overcome aforementioned drawbacks, this study proposes an advanced data analytics framework for ship performance monitoring. The novelty of this study is to utilize the proposed framework in order to quantify the performance of a selected ship in the context of its localized operational conditions (i.e., engine and trim-draft modes). As the novel contributions of this study, the proposed framework is be able to: (i) detect and isolate data anomalies existing in a given data set, (ii) investigate the ship's localized operational conditions, (iv) deal with numerous factors that have influences on the ship's performance in a high-dimensional data space, and (iii) to provide a KPI (i.e. ship performance indicator) for ship performance quantification.

3. Method

The proposed framework presented in this paper has been built upon a preliminary work as described in Bui and Perera (2020). Fig. 1 illustrates the overall architecture of the proposed framework with key aspects that can be listed as follows: domain knowledge, descriptive analytics, diagnostic analytics, visual analytics, and prescriptive analytics:

- Domain knowledge: is embedded in every step of the proposed framework. It refers to an understanding of the ship's localized operational conditions (i.e., engine and trim-draft modes), the reasoning behind conclusions in each data analytics. It also refers to the knowledge obtained from interactions with experts in the field of maritime transport (e.g., ship owners, ship operators, engine manufacturers).
- Descriptive analytics: this attempts to answer the question of 'What happened?'; it provides an understanding of what happened to the system. From this perspective, two anomaly detectors are proposed to detect and isolate data anomalies. Furthermore, digital modeling is proposed for the investigation of certain patterns of the ship's operational conditions through data clustering.
- Diagnostic analytics: this attempts to answer the question of 'Why did it happen?'; it reflects an understanding of why something happened to the system. From this perspective, the causes of data anomalies are identified.
- Visual analytics: this visualizes the improved data in order to identify the relative relationships or correlations among ship performance and navigation parameters under each of the localized operational conditions.
- Prescriptive analytics: this attempts to answer the question of 'What do we do?'. From this perspective, a selected KPI (i.e. key performance indicator) for ship performance quantification is provided.

3.1. Descriptive analytics

3.1.1. Digital modeling

What follows is an account of digital modeling which aims to provide insights into data properties with respect to the ship's localized operational conditions. For this reason, a digital model is formulated to gain a better understating of discrete data distributions in a high-dimensional space. Fig. 2 depicts the digital model which is an extended version of Perera and Mo (2020). The digital model is represented in the right-handed coordinate system of three parameters (i.e., X_1, X_2, X_3) of a selected data set. It is assumed in the digital model that there is an existence of several data clusters, i.e., A, B, C, which represent engine localized operational conditions. These data clusters are represented by vectors with their respective mean values, i.e., μ_1, μ_2, μ_3. Moreover, each of the data clusters contains several structural vectors in the form of singular vectors (SVs). For example, SVs of cluster $i, i = \{A, B, C\}$ are denoted as $Z_{i,1}$, $Z_{i,2}$, $Z_{i,3}$. Furthermore, Fig. 2 also pinpoints some arrows between the data clusters in the digital model. A probable explanation is that there are certain transient regions, representing the transition modes from an operational condition to another. It is necessary to be borne in mind that data outliers and data anomaly clusters can also be represented under the digital model. This is attributed to the data veracity that should be properly addressed. It is also of interest to further investigate other operational conditions, i.e. trim and draft conditions, within the respective data clusters. In this respect, the projection of the data cluster A onto another high dimensional space is shown in the window on the right-handed side of Fig. 2, where sub-data clusters with respect to trim and draft condition, i.e., A_1, A_2, A_3, can be explored.

A more detailed account of the ship's localized operational conditions can be observed in Fig. 3. In this regard, there are hierarchical relationships between engine operational conditions and trim-draft operational conditions. It can be assumed that there are several engine modes, e.g., engine mode A, B, C, etc. Such engine modes can be demonstrated by cluster A, B, C, etc. Several trim-draft modes can be further explored under each of these engine modes. For example, trim-draft modes A_1, A_2, A_3, etc. can be found under engine mode A. These trim-draft modes can be demonstrated by sub-clusters (e.g., sub-cluster A_1, A_2, A_3, etc.).

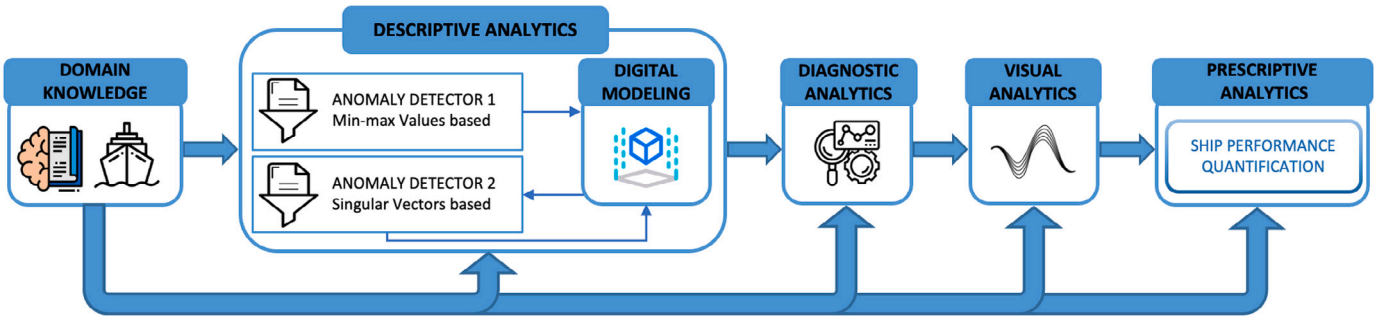


Fig. 1. A representation of the proposed framework.

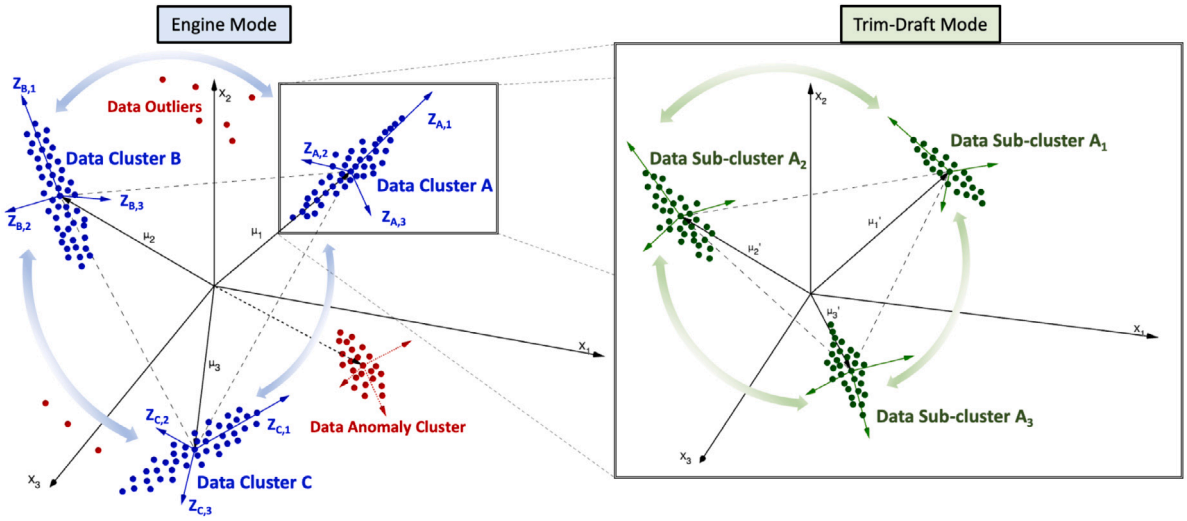


Fig. 2. A representation of the digital model.

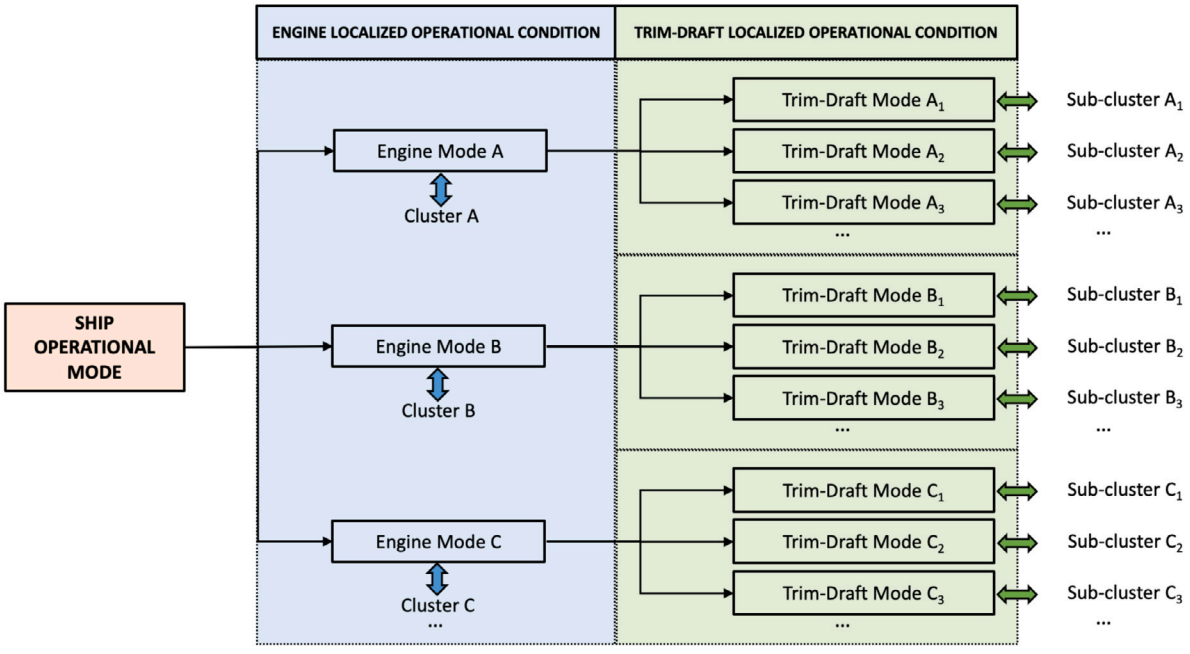


Fig. 3. Ship's localized operational conditions.

3.1.2. Kernel Density Estimation (KDE)

The investigation on the properties of the data can be done by KDE, a non-parametric density estimation method, which yields a smooth representation of the underlying probability density function of the data. Supposing a data set of observations $x = [x_1, x_2, x_3, \ldots, x_N]$ with N samples are being drawn from an unknown probability density $p(x)$. We wish to estimate the shape of $p(x)$, the kernel density estimation at x is defined as follows (Bishop, 2006)

$$\hat{p}(x) = \frac{1}{Nh}\sum_{i=1}^{N}\phi\left(\frac{x - x_i}{h}\right) \tag{1}$$

where ϕ is a kernel function, which specifies the shape of the distribution placed at each point, h is a smoothing parameter called the bandwidth, which controls the size of the kernel at each point. The choice for ϕ in this study is the Gaussian kernel.

3.1.3. Gaussian Mixture Models (GMMs)

The following is a brief description of an unsupervised learning technique for data clustering. The technique is based on probability density estimation using GMMs and the Expectation–Maximization (EM) algorithm for distributing data into different clusters. The Gaussian distribution of a d-dimensional vector x is defined as (Bishop, 2006)

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

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

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

$$p(x) = \sum_{k=1}^{K}\pi_k\mathcal{N}(x|\mu_k, \Sigma_k) \tag{3}$$

where each Gaussian density $\mathcal{N}(x|\mu_k, \Sigma_k)$ is called a component of the mixture with its mean vector μ_k and covariance Σ_k for the k^{th} Gaussian component, π_k is the prior probability of the k^{th} Gaussian; π_k is also defined as the mixing coefficients with the constraint that $\sum_{k=1}^{K}\pi_k = 1$

EM algorithm for Gaussian Mixtures

Fitting a mixture of Gaussians to data can be done by using the maximum likelihood and the EM algorithm. From Eq. (3), the log of the likelihood function is expressed as

$$ln\ p(X|\pi, \mu, \Sigma) = \sum_{n=1}^{N} ln\left(\sum_{k=1}^{K}\pi_k\mathcal{N}(x_n|\mu_k, \Sigma_k)\right) \tag{4}$$

Given a Gaussian mixture model, the EM algorithm is a powerful technique for maximizing this likelihood function with respect to the parameters, i.e., the means μ_k, the covariances of the components Σ_k and the mixing coefficients π_k.

- 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

$$\gamma(z_{nk}) = \frac{\pi_k\mathcal{N}(x_n|\mu_k, \Sigma_k)}{\sum_{j=1}^{K}\pi_j\mathcal{N}(x_n|\mu_j, \Sigma_j)} \tag{5}$$

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

$$\mu_k^{new} = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})x_n \tag{6}$$

$$\gamma_k^{new} = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})(x_n - \mu_k^{new})(x_n - \mu_k^{new})^T \tag{7}$$

$$\pi_k^{new} = \frac{N_k}{N} \tag{8}$$

where

$$N_k = \sum_{n=1}^{N}\gamma(z_{nk}) \tag{9}$$

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

- Step 4: Evaluate the log likelihood

$$ln\ p(X|\pi, \mu, \Sigma) = \sum_{n=1}^{N} ln\left(\sum_{k=1}^{K}\pi_k\mathcal{N}(x_n|\mu_k, \Sigma_k)\right) \tag{10}$$

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

3.1.4. Finding the optimal number of clusters

The GMMs for data clustering is an unsupervised learning technique in which the ground true class labels are not given in the data set. Consequently, the performance of the GMMs is constrained by finding the number of components K. In order to do this, several techniques exist. It may not possible to use the silhouette metric because it may not reliable if the clusters are not spherical or have different sizes, shapes and orientations. Instead, finding the model that minimizes a theoretical criterion information such as the Bayesian Information Criterion (BIC) or the Akaike Information Criterion (AIC) is considered. The BIC and the AIC are expressed as follows (Schwarz, 1978; Akaike, 1974).

$$BIC = ln\ (n)q - 2\ ln\ (\hat{L}) \tag{11}$$

$$AIC = 2q - 2\ ln\ (\hat{L}) \tag{12}$$

where n is the number of observations, q is the number of parameters learned by the model, $\hat{L}$ is the maximized value of the likelihood function of the model. The optimal number of components K (i.e. the number of clusters) is likely with the lowest BIC and AIC value.

3.1.5. Data anomaly detectors

In the section that follows, it is critical to investigate the quality of the data set before proceeding to deploy the digital modeling with further data analysis. For the purpose of such investigation, two data anomaly detectors are proposed, as illustrated in Fig. 4. First of all, the data set needs to go through the first data anomaly detector based on minimum–maximum values. In this regard, a limit check approach, as discussed by Isermann (2006) and Perera (2016), is adopted for the detection of data anomalies and/or outliers. The domain knowledge is required to define the minimum and maximum values of the parameters of the data set. These values represent the general range that the parameters can exist. If data points stay beyond one of the given minimum and maximum thresholds, they are indicating data outliers and will then be removed.

The second data anomaly detector will be executed when the digital modeling is constructed. If there are any anomalies detected, flag alarms will be given. Afterwards, these anomalies are isolated. It is noted that these outliers and anomalies acquired from the two data anomaly detectors are then stored in a data anomaly database for data recovery. However, dealing with the recovery process is beyond the scope of this study.

The second data anomaly detector is based on Singular Value Decomposition (SVD) (Brunton and Kutz, 2019). This is a numerically stable matrix decomposition method with versatile applications. Considering the following data set $X \in \mathbb{R}^{n\times m}$ where n is the number of

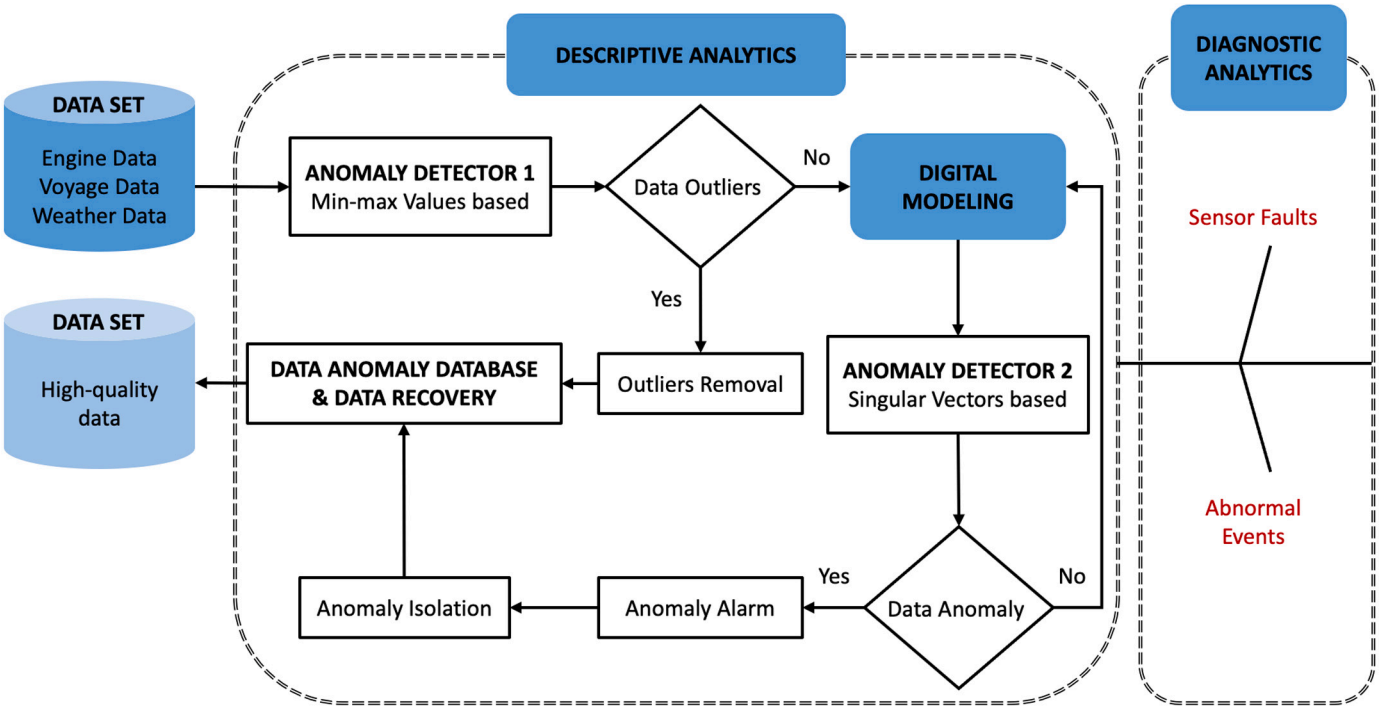


Fig. 4. Descriptive analytics architecture.

observations and m is the number of features (i.e. parameters) ($n > m$). The SVD formula can be expressed as follows.

$$X = U\Sigma V^T \tag{13}$$

where $U \in \mathbb{R}^{n\times n}$ is a square matrix with its column vectors are called the left-singular vectors. $V \in \mathbb{R}^{m\times m}$ is a square matrix with its column vectors are called the right-singular vectors. $\Sigma \in \mathbb{R}^{n\times m}$ is called the singular value matrix, consisting of singular values $\sigma_i, i = 1, \ldots, m$. These singular values are ordered as $\sigma_1 \geqslant \sigma_2 \geqslant \ldots \sigma_m \geqslant 0$.

An elegant interpretation of the SVD can be observed in the correlation matrix $X^T X$ (i.e. the normalized covariance matrix) as follows.

$$X^T X = V\hat{\Sigma}^2 V^T \implies X^T XV = V\hat{\Sigma}^2 \tag{14}$$

where $\hat{\Sigma} \in \mathbb{R}^{m\times m}$ is the square diagonal matrix with the singular values.

This interpretation provides some important advantages in using the SVD in this study. First, the SVD is able to construct optimal orthogonal expansions for projecting the original data set onto a linear subspace. In this respect, the columns of V (i.e. the right-singular vectors) can be used as principal axes for data projections. Therefore, the representation of the original data set can be constructed intuitively and meaningfully. Second, with the help of the SVD, the most important information of the data set can be extracted based on the hierarchical order of importance of the dominant features. This information can be observed in the top SVs. On the contrary, the least important information of the data set are accommodated in the bottom SVs. For this reason, data anomalies can be perceived in such bottom SVs. These anomalies can be understood as the parameter relationships that are deviated from the existing physical relationships of the parameters.

3.2. Visual analytics

As indicated previously, high-dimensional data may cause a difficulty for data visualization. In other words, it is not easy to have intuition about the structure of data clusters in a high-dimensional space. The visual analytics is therefore proposed in order to identify the relative correlations or relationships among parameters under the respective data clusters. In this regard, the SVD is performed and the

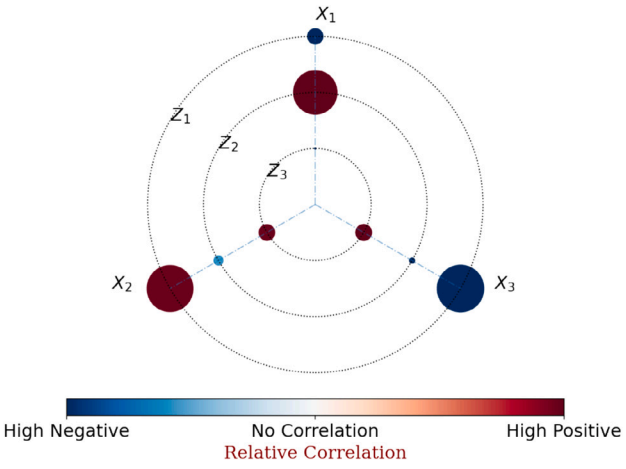


Fig. 5. Visual analytics on a high dimensional singular vector space.

structure of each data cluster is denoted by SVs. Fig. 5 illustrates this approach with three parameters (i.e., X_1, X_2 and X_3) in a high dimensional space as a general representation. Presumably, there are three SVs Z_1, Z_2 and Z_3, sorted in descending order (i.e. from the outermost circle to the innermost circle) associated with their singular values which represent the descending variance directions. Such variance information can be used to extract relevant correlations among parameters, as represented by colored circles in the SVs. The size of each colored circle expresses the significance (i.e. the strength) of the parameter correlation. The color of each colored circle expresses the positive/negative sign of the parameter correlation. When the colored circle is denoted in a large red circle, it means that there is a high positive correlation. When it is denoted in a large blue circle, it means that there is a high negative correlation. Taking the top singular vector Z_1 in Fig. 5 as an example, there is a significant increase in the parameter X_2 while there is a significant decrease in the parameter X_3. A decrease in the parameter X_1 can also be seen in this condition.

It should be noted that the top singular vector Z_1 represents the largest variance direction (i.e. the most information) of the data cluster while the bottom singular vector Z_3 represents the smallest variance direction (i.e. the least information) of the respective data cluster. For this reason, several correlations among parameters might be unclear in the bottom singular vector.

3.3. Prescriptive analytics

The section below proposes a selected KPI (i.e key performance indicator) for ship performance quantification. It is important to stress that this KPI is derived with respect to the availability of the ship performance and navigation parameters in the respective data set. The KPI is attached to each of the ship's localized operational conditions (i.e. represented by a cluster or a sub-cluster) in order to evaluate its performance. The resulting KPI for ship performance quantification can be expressed as

$$SPI_i = \frac{FC_i}{D_i} \tag{15}$$

where

$$FC_i = FC_{avg,i} \times t_i \tag{16}$$

$$D_i = STW_{avg,i} \times t_i \tag{17}$$

here SPI_i is the ship performance index of the ship's localized operational condition i, 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 $STW_{avg,i}$ is the average speed through water (STW) [NM/h] under the respective localized operational condition i, correspondingly. For the sake of unit consistency, Eq. (15) can be rewritten as follows.

$$SPI_i = \frac{FC_{avg,i}}{24\ STW_{avg,i}} \tag{18}$$

It is noted that SPI_i [Ton/NM] is a representation of the ship's average ME fuel cons per nautical mile.

4. Data description and experimental settings

As an exemplification for the application of the proposed method, a ship performance and navigation data set was obtained from a bulk carrier. This is a time-series data set of 3 years with a sampling rate of 15 minutes. Table 1 shows several principal particulars of the selected ship while Table 2 demonstrates twelve parameters with respect to ship performance and navigation along with their minimum–maximum values.

The programming language used to analyze the data was Python with Jupyter Notebook 6.0.3 interface. It was running on a macOS computer, consisting of Intel Core i7 CPU 2.2 GHz with 6 Cores and 32 GB RAM. The computational complexity of training the GMMs depends on the number of observations n, the number of parameters m, the number of clusters K, and the constraints on the covariance matrices. Regarding the settings of the GMMs, it needs to be run several times in order to end up converging to the best solution. The number of initializations was set in this study is 10.

Table 1
Ship particulars.

Feature	Value [Unit]
Ship length	225 [m]
Beam	33 [m]
Gross tonnage	38.889 [N/A]
Deadweight at max draft	72.562 [Ton]
A 2-stroke main engine with maximum continuous rating (MCR)	7564 [kW]
Main engine - shaft rotational speed	105 [rpm]
Two auxiliary engines with MCR	850 [kW]
Auxiliary engines - shaft rotational speed	800 [rpm]
Fixed pitch propeller with 6.20 [m] in diameter and four blades	

Table 2
Ship performance and navigation parameters and their minimum–maximum values.

Parameter	Unit	Min value	Max value
Auxiliary (Aux) fuel consumption (cons)	[Ton/day]	1	8
Main Engine (ME) fuel consumption (cons)	[Ton/day]	1	40
Auxiliary (Aux) power	[kW]	100	850
Main Engine (ME) power	[kW]	3000	8000
Shaft speed	[rpm]	80	120
Relative (Rel) wind speed	[m/s]	0	25
Relative (Rel) wind direction (dir)	[deg]	0	360
Course	[deg]	0	360
Speed over ground (SOG)	[Knots]	3	20
Speed through water (STW)	[Knots]	3	20
Trim	[m]	−2	4
Average (Avg) draft	[m]	0	15

5. Results and discussion

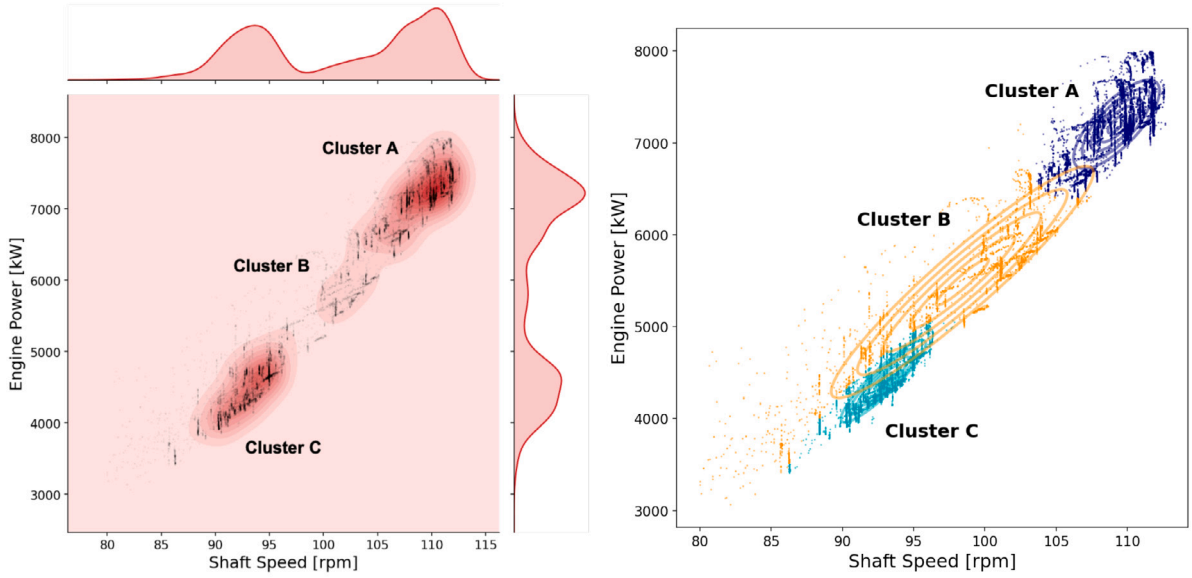
5.1. Descriptive analytics

Regarding the deployment of the first data anomaly detector, it was found that several data points were unreasonable. For example, the values of trim were around −10 [m] or the values of the ME fuel (cons) were around 118 [Ton/day]. Based on the domain knowledge, these values were characterized as outliers and should be omitted. Therefore, threshold values, i.e. the minimum and maximum values of the navigation and performance parameters, were accordingly identified based on the domain knowledge, as shown in Table 2. The ranges for the engine power and the shaft speed were given by the engine manufacturer.

In the case of the digital modeling, Fig. 6 exemplifies the implementation of the KDE and the GMMs for engine data (i.e. shaft speed and engine power). In the first place, the KDE was constructed to gain insights into the number of components K for the GMMs. In this respect, the density estimation of the engine data can be approximately perceived as three components (i.e., cluster A, B, and C), as shown in Fig. 6a. Among of these, cluster A and C are the two main modes of the engine in operation while other data points are belonging to cluster B which could be attributed to a transient state of the engine. Therefore, by using the KDE as a representation guidance together with the domain knowledge, the number of components (i.e. the number of clusters) $K = 3$ was then be suggested for the GMMs. Fig. 6b illustrates the results of the deployment of the GMMs, capturing these three clusters as ellipsoid-shaped clusters, denoted in dark blue, orange and turquoise respectively. Therefore, it arrived at a conclusion that the selected ship was operating in three engine modes. The GMMs was further investigated in three-dimensional space where the ME fuel consumption, the shaft speed and the engine power were taken into consideration. As presented in Fig. 7, there are three clusters in relation to engine modes existing in the digital modeling.

When the digital modeling had been constructed, the deployment of the second data anomaly detector was carried out. As mentioned earlier, the bottom singular vector (i.e. Z_{12}) carries the least important information of the data set. Hence, it was used to detect anomalies for the second anomaly detector. Fig. 8 shows that data cluster A is projected onto a new subspace represented by Z_{12}. It should be noted that -3σ and 3σ (here σ is the standard deviation of the respective data distribution) were chosen as appropriate threshold values. If data points exceed these values, they are flagged as anomalies for this detector. In this regard, a number of anomalies are detected, as shown in the middle and the bottom plot of Fig. 8.

The identification of such anomalies was further investigated in Fig. 9, where all parameters are presented in a time-series format with respect to the number of data points. What stands out in this figure is that several anomalies are detected, denoted by the red pulses and



(a) KDE implementation. (b) GMMs implementation.

Fig. 6. Engine data clustering.

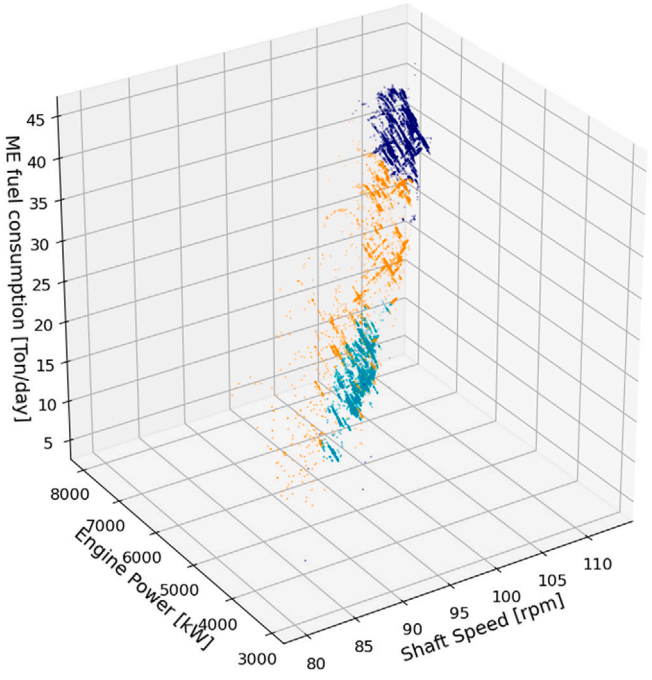


Fig. 7. Engine data clustering in three-dimensional space.

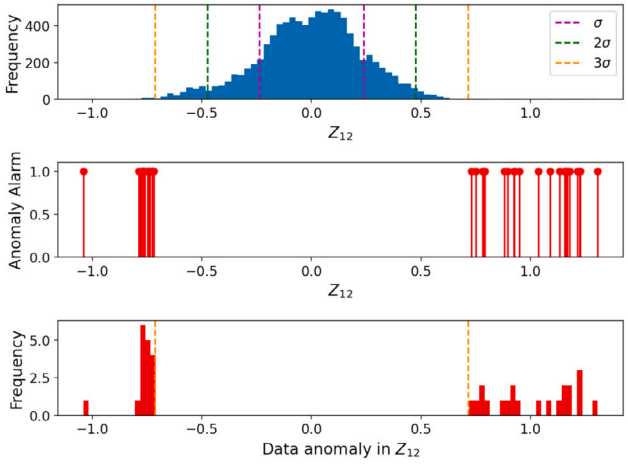


Fig. 8. Data anomaly detection in the bottom singular vector Z_{12}.

Table 3
Number of identified data anomalies using the SVs-based detector.

Cluster	No of identified data anomalies	Ratio[a] (%)
A	38	0.41
B	37	1.08
C	48	0.81

[a]The ratio (%) indicates the number of identified anomalies per the number of data points in the respective cluster.

the anomaly alarms are accordingly raised. It should be borne in mind that the shaft speed, arranged in ascending order in the first plot, is a basis for the detection of such anomalies. It is also important to bear in mind that the operation of other on-board systems, including hotel systems, is completely independent of the main engine in some situations. Therefore, any variations in the Auxiliary (Aux) fuel cons or the Aux power do not have any effects on the actual ME fuel consumption in such situations. In the first anomaly point (DA 1), there are some sudden changes with respect to the ME power and the STW. In the second anomaly point (DA 2), there are falling points in the ME fuel cons and the STW. Similar strange behaviors can also be observed with respect to the ME fuel cons and the STW in the third anomaly point (DA 3). This approach was further adopted to cluster *B* and cluster *C*. Table 3 presents the number of anomalies identified by this detector in the respective clusters. The most likely causes of identified data anomalies existing in the data set are sensor faults and/or abnormal events. These causes draw conclusions for the diagnostic analytics.

Perhaps the most interesting aspect of the descriptive analytics is the exploration of the ship's localized operational conditions. As was pointed out previously, the selected ship was operating in three engine modes, represented by cluster *A*, *B* and *C*. Each of these engine modes has different trim-draft modes which can be represented by sub-clusters. To further examine this, the deployment of the KDE and the

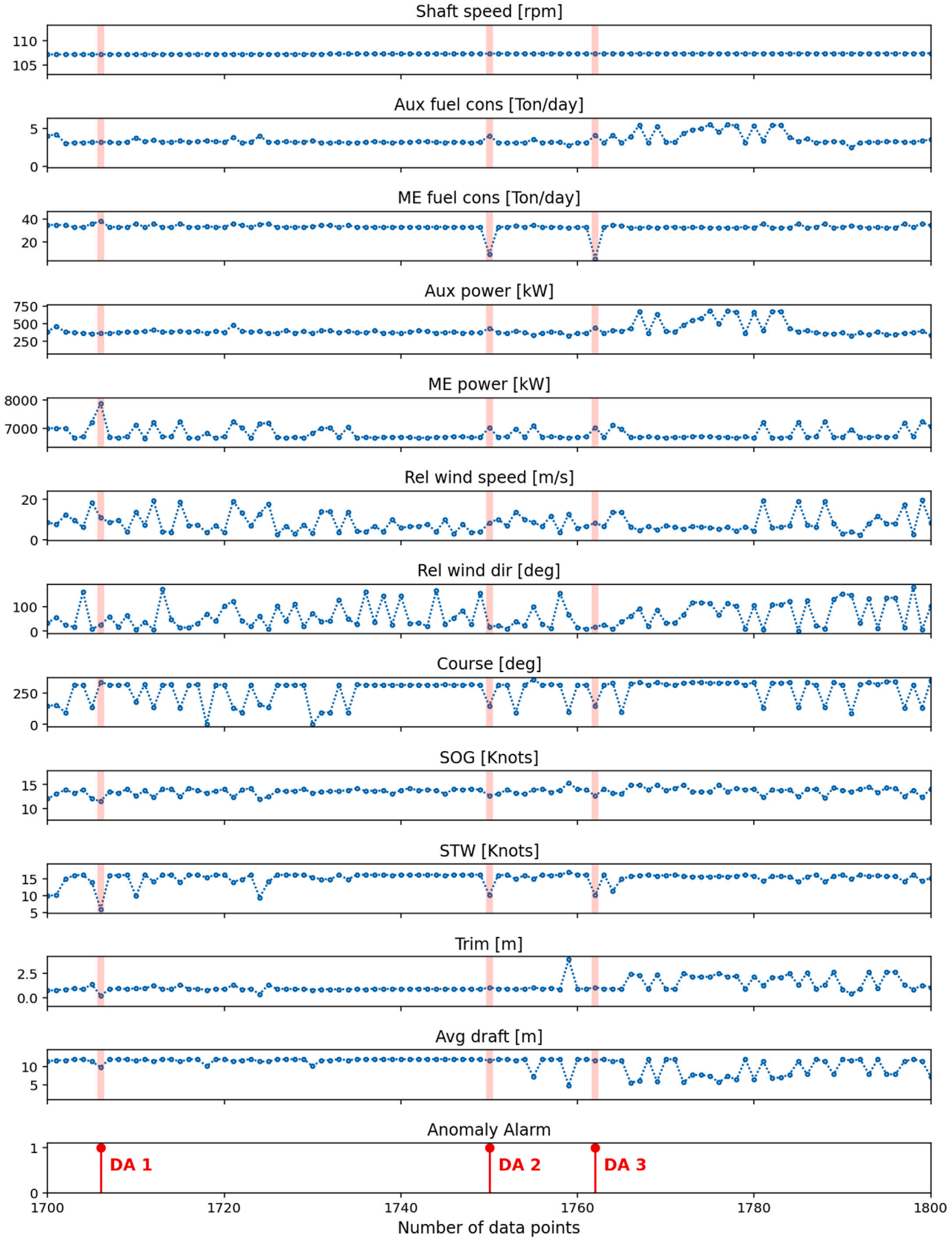
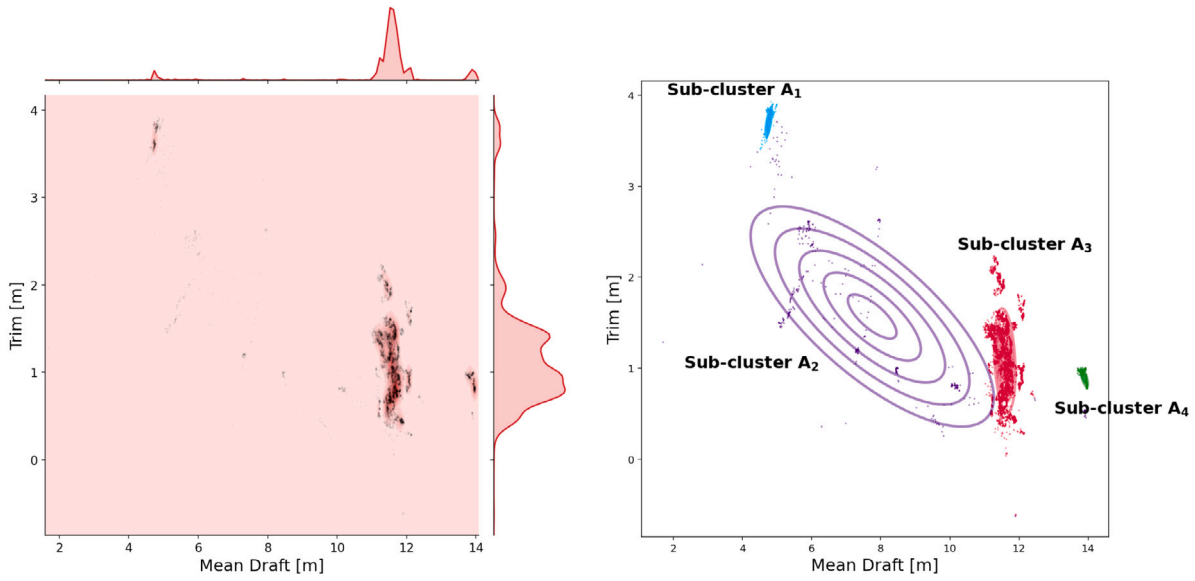
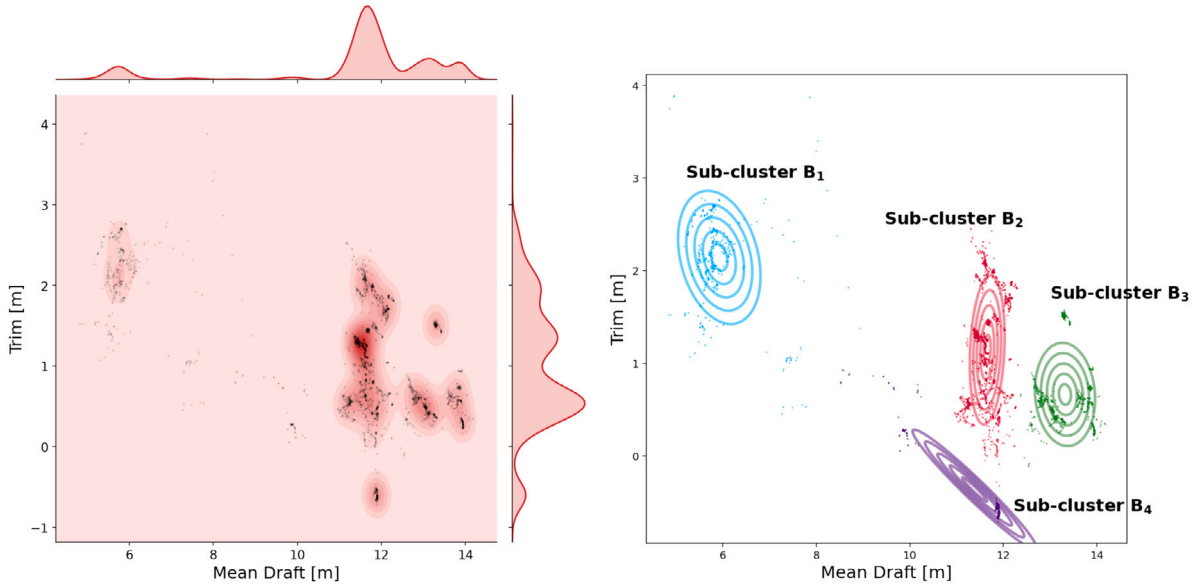


Fig. 9. Data anomaly detection in the time-series plot.



(a) KDE implementation. (b) GMMs implementation.

Fig. 10. Trim-draft data clustering with respect to data cluster *A*.



(a) KDE implementation. (b) GMMs implementation.

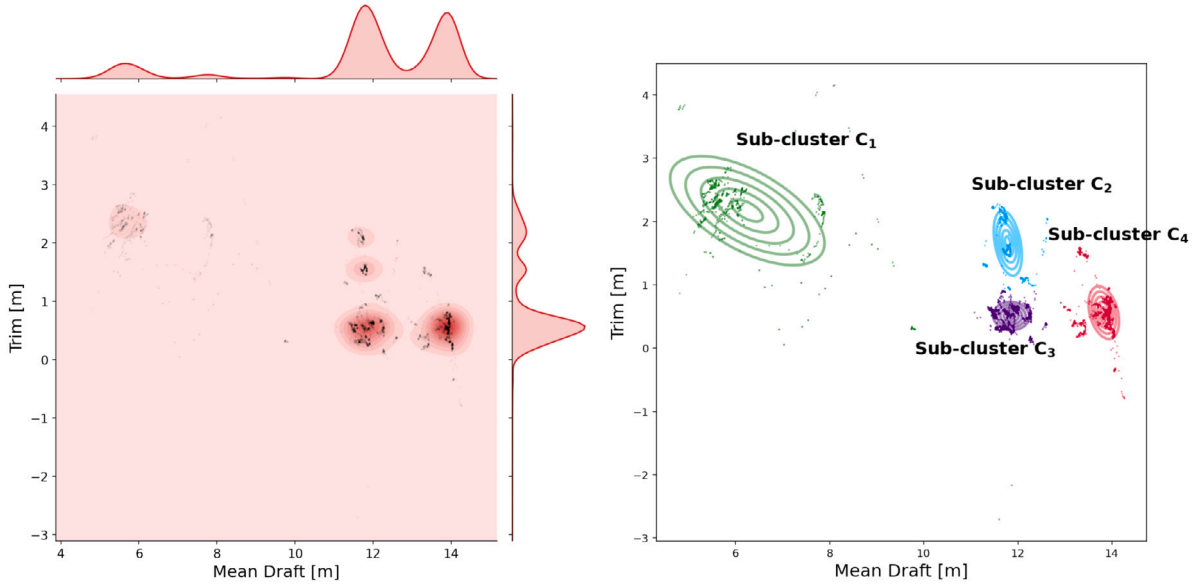
Fig. 11. Trim-draft data clustering with respect to data cluster *B*.

GMMs for trim-draft data was carried out under each of these engine modes. The domain knowledge also occupied a role in determining the number of sub-clusters in these cases. Fig. 10 indicates there are four sub-clusters A_1, A_2, A_3, and A_4, representing trim-draft modes with respect to cluster A. It can be seen from Fig. 11 that there are four sub-cluster B_1, B_2, B_3, and B_4, (i.e. trim-draft modes) with respect to cluster B. Fig. 12 reports four sub-clusters C_1, C_2, C_3, and C_4 (i.e. trim-draft modes) with respect to cluster C.

5.2. Finding the optimal number of clusters

As mentioned earlier, an important factor by which the GMMs can be evaluated is finding the optimal number of clusters K. This can be done by calculating the BIC and the AIC. The results on the BIC and the AIC of the engine data (i.e., shaft speed and engine power) are shown in Fig. 13. It can be seen from this figure that the BIC and the AIC results do not give an optimal position for K. If there are many components K in the GMMs, it will increase the probability of over-fitting. Therefore, in this case, the BIC and the AIC results are inconclusive. The domain knowledge can play a crucial role in this case. After consulting with the ship owner who provided us the data set, they confirmed that the ship was operating in three engine modes. This is in line with what we determined before.

Further experiments on the BIC and the AIC were also performed for trim-draft data, as shown in Figs. 14, 15, 16. It can be observed from these figures that the results on the BIC and the AIC in all cases are also inconclusive. Therefore, in this study, the domain knowledge regarding the engine operational modes and the trim-draft operational modes should be directly embedded into the GMMs in order to identify



(a) KDE implementation. (b) GMMs implementation.

Fig. 12. Trim-draft data clustering with respect to data cluster C.

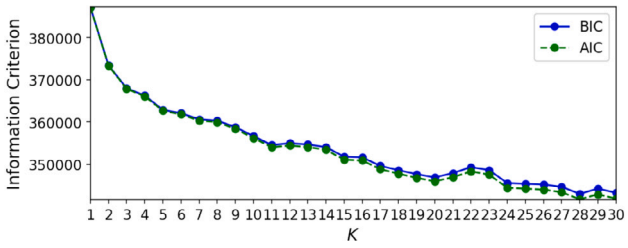


Fig. 13. BIC and AIC results for engine data clustering.

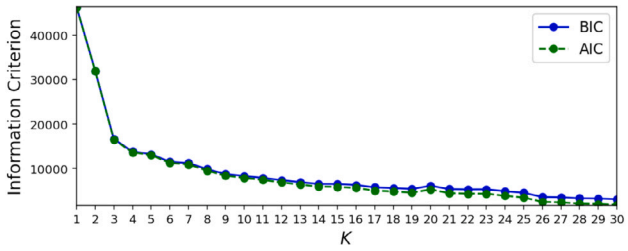


Fig. 14. BIC and AIC results for trim-draft data clustering with respect to cluster A.

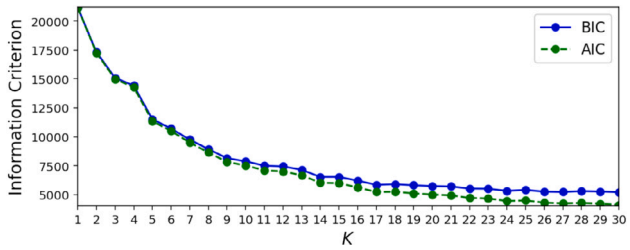


Fig. 15. BIC and AIC results for trim-draft data clustering with respect to cluster B.

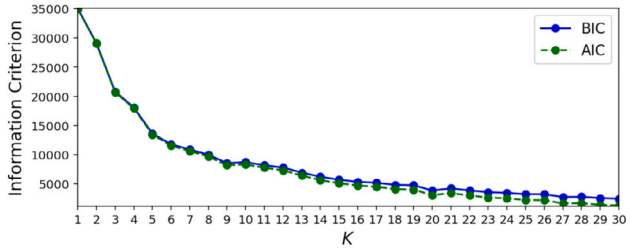
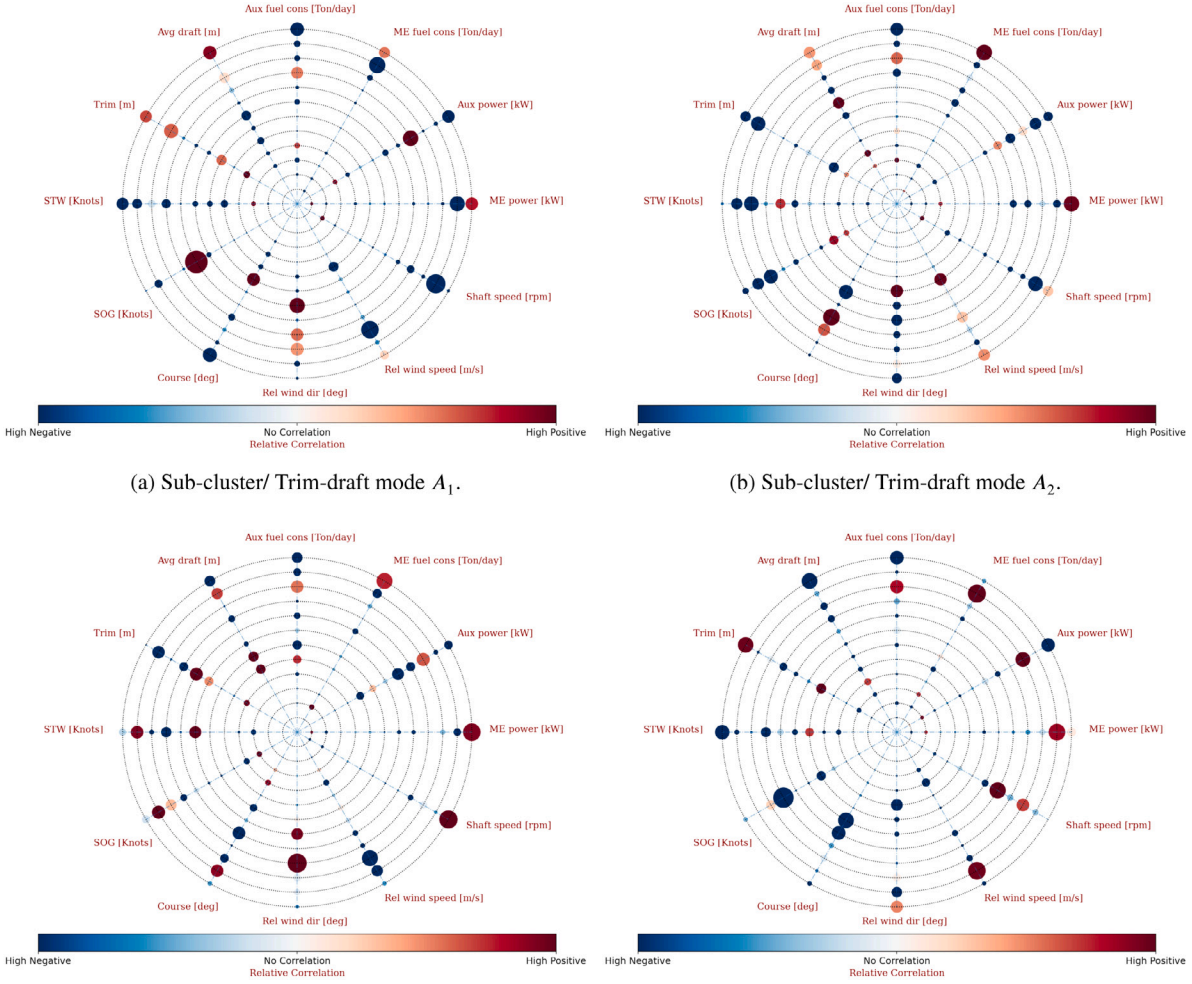


Fig. 16. BIC and AIC results for trim-draft data clustering with respect to cluster C.

possible data clusters. By doing this, the accuracy of the digital model can be improved.

5.3. Visual analytics

As explained earlier, the purpose of the visual analytics is to reveal the relative correlations among parameters under a cluster or a sub-cluster. The results on the visual analytics of data cluster A under trim-draft modes (i.e. represented by sub-cluster A_1, A_2, A_3, and A_4) are illustrated in Fig. 17. The results on this analytics of sub-cluster A_3 was selected for the purpose of illustration. Fig. 17c is revealing in several ways. The top singular vector shows an increase in the Shaft speed and an increase in the ME power, thus the ME fuel cons also increases. It can also be found that a decrease in the Aux power leads to a decrease in the Aux fuel cons. The Average (Avg) draft is also decreased in this condition. Turning to the second singular vector, there is an adjustment of the Trim and the Avg draft, thereby increasing the STW and the speed over ground (SOG). The third singular vector demonstrates that an increase in the Aux fuel cons is attributed to an increase in the Aux power. It can be observed from the fourth singular vector that an decrease in the STW may cause a considerable increase in the Rel wind direction. The fifth singular vector indicates that a decrease in the Aux fuel cons stems from a decrease in the Aux power. It can also be seen in this situation that the Trim is increased. The sixth singular vector shows that there is an increase in the STW along with an increase in the Rel wind direction. Besides, a trim-draft adjustment can be observed. It is noted that the bottom singular vectors have low singular values. As a result of this, the correlations among parameters are unclear or there are no realistic correlations that can be observed in



(a) Sub-cluster/ Trim-draft mode A_1.

(b) Sub-cluster/ Trim-draft mode A_2.

(c) Sub-cluster/ Trim-draft mode A_3.

(d) Sub-cluster/ Trim-draft mode A_4.

Fig. 17. Visual analytics of data cluster A under trim-draft modes.

these bottom singular vectors. The remaining results on this analytics of other sub-clusters (e.g., A_1, A_2, and A_4) can be explained in the same manners.

5.4. Prescriptive analytics

As mentioned previously, the prescriptive analytics is proposed to provide the KPI, expressed by the ship performance index SPI, in order to quantify the ship's performance under the identified localized operational modes. Table 4 compares the SPI results of trim-draft modes under the respective engine modes, as defined in Eq. (18). Considering engine mode A, trim-draft mode A_1 appears to be the best performance mode because of its lowest SPI value ($SPI = 0,0797$ [Ton/NM]). Looking at engine mode B, the SPI value of trim-draft mode B_4 ($SPI = 0,0805$ [Ton/NM]) indicates that this trim-draft mode is the best performance mode. Turning to engine mode C, based on the SPI value of trim-draft mode C_1 ($SPI = 0,0699$ [Ton/NM]), this is the best performance trim-draft mode. It is apparent from this table that, overall, trim-draft mode C_1 has the lowest SPI value among other trim-draft modes. It can thus be suggested that this is the best performance mode of the selected ship. However, with the ship performance and navigation parameters available in the data set were considered, caution should be applied. The lack of the loading conditions in the data set added the caution regarding the generalizability of these results for energy efficiency quantification. For this reason, 'ship performance quantification' was addressed rather than 'energy efficiency quantification'.

Table 4
SPI value for ship performance quantification.

Cluster (Engine Mode)	Sub-cluster (Trim-draft Mode)	SPI [Ton/NM]
A	A_1	0,0797
	A_2	0,1030
	A_3	0,1121
	A_4	0,1468
B	B_1	0,0936
	B_2	0,0992
	B_3	0,1080
	B_4	0,0805
C	C_1	0,0699
	C_2	0,0728
	C_3	0,0748
	C_4	0,0753

6. Conclusion

Prior studies have only focused on predicting ship fuel consumption or optimizing engine speed/trim as regards the improvement of operational energy efficiency. To the best of our knowledge, no other authors have studied the ship's performance in a local scale with respect to its operational conditions. The novelty of this study is to quantify the performance of a selected ship under localized operational conditions (i.e., engine and trim-draft modes) by developing an advanced data analytics framework. It was demonstrated through a data set collected

from a bulk carrier. The research findings obtained from the proposed framework have been summarized as follows.

- Descriptive analytics has proposed two data anomaly detectors that were be able to detect and isolate a number of data anomalies existing in the data set. Furthermore, it has offered a better understanding of the ship's localized operational conditions. This can be perceived by the engine and the trim-draft modes. With the help of the KDE and the GMMs, the investigation of the digital model has shown that the selected ship was operating in three engine modes, represented by clusters A, B and C. The digital model was further examined for trim-draft data with respect to these clusters. The findings of this examination have shown that several trim-draft modes were identified, represented by sub-clusters.
- Diagnostic analytics has suggested two main reasons why there are data anomalies in almost data sets collected from data acquisition systems. In this regard, sensor faults and/or abnormal events were identified as the causes strongly associated with these data anomalies.
- Visual analytics has revealed the relative relationships or correlations among the ship performance and navigation parameters in relation to the respective engine modes and trim-draft modes.
- Prescriptive analytics has provided a KPI in order to quantify the ship's performance under the respective engine modes and trim-draft modes. The KPI was expressed by the ship performance index SPI (i.e. the average ME fuel cons per nautical mile). Based on the SPI findings, it is likely that sub-cluster C_1 was the best performance trim-draft mode of the selected ship.

Taken together, the findings suggest a role for the domain knowledge in every step of the proposed framework. Moreover, the findings have the potential to serve as an operational energy efficiency measure that is of value for both ship operators (captains, chief-engineers, ship officers) and decision-makers (ship owners, fleet managers, technical divisions) for improving energy efficiency through operational practices. In this respect, the findings can be integrated into the ship performance monitoring systems. Specifically, they can be simulated and displayed on the on-board user interfaces. Therefore, ship operators are equipped with meaningful visualizations and indicators in order to evaluate their practices and raise their awareness with respect to energy efficiency. By considering the KPI proposed in this study, ship operators could know in which engine/trim-draft mode they should facilitate the eco-maneuvering, e.g. operating the engine under the load range with the lowest specific fuel oil consumption. This KPI will change depending on system's operational conditions and hull fouling conditions. Hence, ship operators can also consult with technical divisions ashore in order to trouble-shoot their operational problems via remote communication. Furthermore, such visualizations and indicators can assist ship owners/fleet managers in achieving performance improvement across their fleet.

Looking ahead towards Shipping 4.0, the ship performance monitoring systems can be transformed into digital platforms by the Digital Twin technology. The Digital Twin is a virtual representation which serves as the real-life counterpart of the ship. The digital model proposed in the study has the potential for exploiting the Digital Twin. In this way, the Digital Twin has the capabilities to become an automated, self-aware anomaly detection, self-visualization platform that enables ship operators and fleet managers to monitor the instantaneous performance of the ship in real-time.

Nonetheless, the findings in this study are subject to a limitation. The SPI findings on the account of ship performance quantification maybe somewhat limited by the absence of the loading conditions parameter in the data set. Therefore, these findings need to be interpreted with caution. This is the main reason why 'ship performance quantification' was concerned in this study, rather than 'energy efficiency quantification'. Further studies, which take the loading conditions and other factors into account for energy efficiency quantification, will need to be undertaken. Moreover, the issue of data quality is an intriguing one which could be usefully explored in the further studies.

CRediT authorship contribution statement

Khanh Q. Bui: Methodology, Software, Formal analysis, Writing - original draft, Visualization. **Lokukaluge P. Perera:** Conceptualization, Methodology, Writing - review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Automat. Control 19 (6), 716–723. http://dx.doi.org/10.1109/TAC.1974.1100705.

Bal Beşikçi, E., Arslan, O., Turan, O., Ölçer, A.I., 2016. An artificial neural network based decision support system for energy efficient ship operations. Comput. Oper. Res. 66, 393–401. http://dx.doi.org/10.1016/j.cor.2015.04.004.

Balcombe, P., Brierley, J., Lewis, C., Skatvedt, L., Speirs, J., Hawkes, A., Staffell, I., 2019. How to decarbonise international shipping: Options for fuels, technologies and policies. Energy Convers. Manage. 182, 72–88. http://dx.doi.org/10.1016/j.enconman.2018.12.080.

Bishop, C., 2006. Pattern Recognition and Machine Learning. In: Information Science and Statistics, Springer-Verlag, New York.

Bouman, E.A., Lindstad, E., Rialland, A.I., Strømman, A.H., 2017. State-of-the-art technologies, measures, and potential for reducing GHG emissions from shipping – A review. Transp. Res. D 52, 408–421. http://dx.doi.org/10.1016/j.trd.2017.03.022.

Brandsæter, A., Vanem, E., 2018. Ship speed prediction based on full scale sensor measurements of shaft thrust and environmental conditions. Ocean Eng. 162, 316–330. http://dx.doi.org/10.1016/j.oceaneng.2018.05.029.

Brunton, S.L., Kutz, J.N., 2019. Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control. Cambridge University Press, Cambridge, http://dx.doi.org/10.1017/9781108380690.

Brynolf, S., Baldi, F., Johnson, H., 2016. Energy efficiency and fuel changes to reduce environmental impacts. In: Andersson, K., Brynolf, S., Lindgren, J.F., Wilewska-Bien, M. (Eds.), Shipping and the Environment : Improving Environmental Performance in Marine Transportation. Springer, Berlin, Heidelberg, pp. 295–339. http://dx.doi.org/10.1007/978-3-662-49045-7_10.

Bui, K.Q., Ölçer, A.I., Kitada, M., Ballini, F., 2020. Selecting technological alternatives for regulatory compliance towards emissions reduction from shipping: An integrated fuzzy multi-criteria decision-making approach under vague environment. Proc. Inst. Mech. Eng. M http://dx.doi.org/10.1177/1475090220917815.

Bui, K.Q., Perera, L.P., 2019. The compliance challenges in emissions control regulations to reduce air pollution from shipping. In: OCEANS 2019 - Marseille. pp. 1–8. http://dx.doi.org/10.1109/OCEANSE.2019.8867420.

Bui, K.Q., Perera, L.P., 2020. A decision support framework for cost-effective and energy-efficient shipping. In: ASME 2020 39th International Conference on Ocean, Offshore and Arctic Engineering. American Society of Mechanical Engineers Digital Collection, http://dx.doi.org/10.1115/OMAE2020-18368.

Capezza, C., Coleman, S., Lepore, A., Palumbo, B., Vitiello, L., 2019. Ship fuel consumption monitoring and fault detection via partial least squares and control charts of navigation data. Transp. Res. D 67, 375–387. http://dx.doi.org/10.1016/j.trd.2018.11.009.

Cheliotis, M., Lazakis, I., Theotokatos, G., 2020. Machine learning and data-driven fault detection for ship systems operations. Ocean Eng. 216, 107968. http://dx.doi.org/10.1016/j.oceaneng.2020.107968.

Coraddu, A., Oneto, L., Baldi, F., Anguita, D., 2017. Vessels fuel consumption forecast and trim optimisation: A data analytics perspective. Ocean Eng. 130, 351–370. http://dx.doi.org/10.1016/j.oceaneng.2016.11.058.

Dalheim, Ø.Ø., Steen, S., 2020. Preparation of in-service measurement data for ship operation and performance analysis. Ocean Eng. 212, 107730. http://dx.doi.org/10.1016/j.oceaneng.2020.107730.

Erto, P., Lepore, A., Palumbo, B., Vitiello, L., 2015. A procedure for predicting and controlling the ship fuel consumption: Its implementation and test. Qual. Reliab. Eng. Int. 31 (7), 1177–1184. http://dx.doi.org/10.1002/qre.1864.

Farag, Y.B.A., Ölçer, A.I., 2020. The development of a ship performance model in varying operating conditions based on ANN and regression techniques. Ocean Eng. 198, 106972. http://dx.doi.org/10.1016/j.oceaneng.2020.106972.

Gkerekos, C., Lazakis, I., Theotokatos, G., 2019. Machine learning models for predicting ship main engine fuel oil consumption: A comparative study. Ocean Eng. 188, 106282. http://dx.doi.org/10.1016/j.oceaneng.2019.106282.

IMO, 2014. Third IMO GHG Study 2014. International Maritime Organization (IMO), London, UK.

IMO, 2020. Fourth IMO GHG Study 2020. International Maritime Organization (IMO).

Isermann, R., 2006. Fault detection with limit checking. In: Isermann, R. (Ed.), Fault-Diagnosis Systems: an Introduction from Fault Detection to Fault Tolerance. Springer, Berlin, Heidelberg, pp. 95–110. http://dx.doi.org/10.1007/3-540-30368-5_7.

Karagiannidis, P., Themelis, N., 2021. Data-driven modelling of ship propulsion and the effect of data pre-processing on the prediction of ship fuel consumption and speed loss. Ocean Eng. 222, 108616. http://dx.doi.org/10.1016/j.oceaneng.2021.108616.

Kitada, M., Ölçer, A., 2015. Managing people and technology: The challenges in CSR and energy efficient shipping. Res. Transp. Bus. Manag. 17, 36–40. http://dx.doi.org/10.1016/j.rtbm.2015.10.002.

Lajic, Z., Blanke, M., Nielsen, U.D., 2009. Fault detection for shipboard monitoring – Volterra kernel and Hammerstein model approaches. IFAC Proc. Vol. 42 (8), 24–29. http://dx.doi.org/10.3182/20090630-4-ES-2003.00004.

Lajic, Z., Nielsen, U.D., 2010. Fault detection for shipboard monitoring and decision support systems. In: ASME 2009 28th International Conference on Ocean, Offshore and Arctic Engineering. American Society of Mechanical Engineers Digital Collection, pp. 679–686. http://dx.doi.org/10.1115/OMAE2009-79367.

Lazakis, I., Gkerekos, C., Theotokatos, G., 2019. Investigating an SVM-driven, one-class approach to estimating ship systems condition. Ships Offshore Struct. 14 (5), 432–441. http://dx.doi.org/10.1080/17445302.2018.1500189.

Lee, H., Aydin, N., Choi, Y., Lekhavat, S., Irani, Z., 2018. A decision support system for vessel speed decision in maritime logistics using weather archive big data. Comput. Oper. Res. 98, 330–342. http://dx.doi.org/10.1016/j.cor.2017.06.005.

Man, Y., Sturm, T., Lundh, M., MacKinnon, S.N., 2020. From ethnographic research to big data analytics—A case of maritime energy-efficiency optimization. Appl. Sci. 10 (6), 2134. http://dx.doi.org/10.3390/app10062134.

Nielsen, U.D., Lajic, Z., Jensen, J.J., 2012. Towards fault-tolerant decision support systems for ship operator guidance. Reliab. Eng. Syst. Saf. 104, 1–14. http://dx.doi.org/10.1016/j.ress.2012.04.009.

Ölçer, A.I., 2018. Introduction to maritime energy management. In: Ölçer, A.I., Kitada, M., Dalaklis, D., Ballini, F. (Eds.), Trends and Challenges in Maritime Energy Management. In: WMU Studies in Maritime Affairs, Springer International Publishing, pp. 1–12. http://dx.doi.org/10.1007/978-3-319-74576-3_1.

Olivier, J.G., Janssens-Maenhout, G., Muntean, M., Peters, J.A., 2016. Trends in Global CO2 Emissions: 2016 Report. Technical Report 2315, PBL Netherlands Environmental Assessment Agency, The Hague, p. 86.

Perera, L.P., 2016. Marine engine centered localized models for sensor fault detection under ship performance monitoring. IFAC-PapersOnLine 49 (28), 91–96. http://dx.doi.org/10.1016/j.ifacol.2016.11.016.

Perera, L.P., Mo, B., 2017. Machine intelligence based data handling framework for ship energy efficiency. IEEE Trans. Veh. Technol. 66 (10), 8659–8666. http://dx.doi.org/10.1109/TVT.2017.2701501.

Perera, L.P., Mo, B., 2020. Ship performance and navigation information under high-dimensional digital models. J. Mar. Sci. Technol. 25 (1), 81–92. http://dx.doi.org/10.1007/s00773-019-00632-5.

Perera, L., Ventikos, N., Rolfsen, S., Öster, A., 2021. Advanced data analytics towards energy efficient and emission reduction retrofit technology integration in shipping. In: 31st International Ocean and Polar Engineering Conference (ISOPE2021). Rhodes, Greece.

Petersen, J.P., Jacobsen, D.J., Winther, O., 2012a. Statistical modelling for ship propulsion efficiency. J. Mar. Sci. Technol. 17 (1), 30–39. http://dx.doi.org/10.1007/s00773-011-0151-0.

Petersen, J.P., Winther, O., Jacobsen, D.J., 2012b. A machine-learning approach to predict main energy consumption under realistic operational conditions. Ship Technol. Res. 59 (1), 64–72. http://dx.doi.org/10.1179/str.2012.59.1.007.

Pyle, D., 1999. Data Preparation for Data Mining, first ed. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Raptodimos, Y., Lazakis, I., 2018. Using artificial neural network-self-organising map for data clustering of marine engine condition monitoring applications. Ships Offshore Struct. 13 (6), 649–656. http://dx.doi.org/10.1080/17445302.2018.1443694.

Rasmussen, H.B., Lützen, M., Jensen, S., 2018. Energy efficiency at sea: Knowledge, communication, and situational awareness at offshore oil supply and wind turbine vessels. Energy Res. Soc. Sci. 44, 50–60. http://dx.doi.org/10.1016/j.erss.2018.04.039.

Rødseth, Ø.J., Perera, L.P., Mo, B., 2016. Big data in shipping - Challenges and opportunities. In: Proceedings of the 15th International Conference on Computer Applications and Information Technology in the Maritime Industries (COMPIT 2016). Lecce, Italy.

Sasa, K., Terada, D., Shiotani, S., Wakabayashi, N., Ikebuchi, T., Chen, C., Takayama, A., Uchida, M., 2015. Evaluation of ship performance in international maritime transportation using an onboard measurement system - in case of a bulk carrier in international voyages. Ocean Eng. 104, 294–309. http://dx.doi.org/10.1016/j.oceaneng.2015.05.015.

Schwarz, G., 1978. Estimating the dimension of a model. Ann. Statist. 6 (2), http://dx.doi.org/10.1214/aos/1176344136.

Soner, O., Akyuz, E., Celik, M., 2018. Use of tree based methods in ship performance monitoring under operating conditions. Ocean Eng. 166, 302–310. http://dx.doi.org/10.1016/j.oceaneng.2018.07.061.

Soner, O., Akyuz, E., Celik, M., 2019. Statistical modelling of ship operational performance monitoring problem. J. Mar. Sci. Technol. 24 (2), 543–552. http://dx.doi.org/10.1007/s00773-018-0574-y.

Sullivan, B.P., Desai, S., Sole, J., Rossi, M., Ramundo, L., Terzi, S., 2020. Maritime 4.0 – Opportunities in digitalization and advanced manufacturing for vessel development. Procedia Manuf. 42, 246–253. http://dx.doi.org/10.1016/j.promfg.2020.02.078.

Tran, T.A., 2020. Effect of ship loading on marine diesel engine fuel consumption for bulk carriers based on the fuzzy clustering method. Ocean Eng. 207, 107383. http://dx.doi.org/10.1016/j.oceaneng.2020.107383.

UNCTAD, 2019. Review of Maritime Transport 2019. UNITED NATIONS, New York, NY, USA.

Vanem, E., Brandsæter, A., 2019. Unsupervised anomaly detection based on clustering methods and sensor data on a marine diesel engine. J. Mar. Eng. Technol. 1–18. http://dx.doi.org/10.1080/20464177.2019.1633223.

Viktorelius, M., Lundh, M., 2019. Energy efficiency at sea: An activity theoretical perspective on operational energy efficiency in maritime transport. Energy Res. Soc. Sci. 52, 1–9. http://dx.doi.org/10.1016/j.erss.2019.01.021.

Wang, S., Ji, B., Zhao, J., Liu, W., Xu, T., 2018. Predicting ship fuel consumption based on LASSO regression. Transp. Res. D 65, 817–824. http://dx.doi.org/10.1016/j.trd.2017.09.014.

Wang, K., Yan, X., Yuan, Y., Jiang, X., Lodewijks, G., Negenborn, R.R., 2017. Study on route division for ship energy efficiency optimization based on big environment data. In: 2017 4th International Conference on Transportation Information and Safety (ICTIS). pp. 111–116. http://dx.doi.org/10.1109/ICTIS.2017.8047752.

Yan, R., Wang, S., Du, Y., 2020. Development of a two-stage ship fuel consumption prediction and reduction model for a dry bulk ship. Transp. Res. E 138, 101930. http://dx.doi.org/10.1016/j.tre.2020.101930.

Yan, X., Wang, K., Yuan, Y., Jiang, X., Negenborn, R.R., 2018. Energy-efficient shipping: An application of big data analysis for optimizing engine speed of inland ships considering multiple environmental factors. Ocean Eng. 169, 457–468. http://dx.doi.org/10.1016/j.oceaneng.2018.08.050.

Yuan, Y., Li, Z., Malekian, R., Yan, X., 2017. Analysis of the operational ship energy efficiency considering navigation environmental impacts. J. Mar. Eng. Technol. 16 (3), 150–159. http://dx.doi.org/10.1080/20464177.2017.1307716.

Zaman, I., Pazouki, K., Norman, R., Younessi, S., Coleman, S., 2017. Challenges and opportunities of big data analytics for upcoming regulations and future transformation of the shipping industry. Procedia Eng. 194, 537–544. http://dx.doi.org/10.1016/j.proeng.2017.08.182.

Zhang, S., Zhang, C., Yang, Q., 2003. Data preparation for data mining. Appl. Artif. Intell. 17 (5–6), 375–381. http://dx.doi.org/10.1080/713827180.