Contents lists available at ScienceDirect

Ocean Engineering

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

Research paper

Advanced data cluster analyses in digital twin development for marine engines towards ship performance quantification

Mahmood Taghavi[a,*], Lokukaluge P. Perera[b]

[a] UiT the Arctic University of Norway, Norway
[b] UiT the Arctic University of Norway and SINTEF Digital, Norway

ARTICLE INFO

Keywords:
Ship energy efficiency
Digital twin
Clustering
Machine learning
Performance prediction
Gaussian mixture models (GMM)

ABSTRACT

Due to the growing rate of energy consumption, it is necessary to develop frameworks for enhancing ship energy efficiency. This paper proposes a solution for this issue by introducing a digital twin framework for quantifying ship performance. For this purpose, extensive low-level clustering is performed using Gaussian Mixture Models (GMM) with the Expectation Maximization algorithm on a dataset of a selected vessel to detect the vessel's most frequent operating regions. Then, a regression analysis is performed in each operating region, to identify their shapes using Singular Value Decomposition (SVD). The results of SVD make the basis for model development in digital twin applications. For this reason, a low-level clustering is performed so that a more accurate model can be developed in future. Moreover, based on the resulting cluster analysis, an energy efficiency index is developed, and the energy efficiency of each cluster has been evaluated to identify the most efficient operating condition. Hence, the main contribution of this research is to develop a digital twin framework of a marine engine which can be utilized for green ship operations. The same contribution can facilitate the shipping industry to meet the International Maritime Organization energy efficiency requirements.

1. Introduction

The fourth industrial revolution, called Industry 4.0, is characterized by the integration of digital and physical systems in various industrial processes through data sets, which can help commercial companies improve their industrial performance (Coraddu et al., 2019). This revolution is taking place by utilizing state-of-the-art technologies and innovative applications such as robotics, Artificial Intelligence (AI), big data analytics, 3D printing, Internet of Things (IoT), cloud computing, cyber-physical systems and communication infrastructure into traditional industrial applications (Li et al., 2017). Such technologies can offer several benefits, such as improved product quality while having higher productivity, energy-efficient performance, improved supply chain management, and reduced emission in the respective industries, also include decentralization of decision-making, enhanced flexibility, and increased system interconnectivity (Mohamed, 2018). For example, Namazi and Taghavipour (2021) developed a control strategy utilizing vehicle-to-vehicle as well as vehicle-to-infrastructure communication to improve traffic flow and reduce emissions in smart intersections that have intelligent capabilities by utilizing traffic data. The benefits of Industry 4.0 are introduced into industrial processes and products due to the advancements in computational power, communication capabilities that are equipped with information technology, and intelligent operational capabilities (Xu et al., 2018). The benefits offered by the integration of various technologies into industrial processes under the framework of Industry 4.0 contribute to economic sustainability by utilizing the available resources and materials, more effectively. This, in turn, leads to social and environmental sustainability dimensions (Bai et al., 2020; Jamwal et al., 2021). This industrial revolution has the potential to transform the world, significantly on a larger scale than the previous industrial revolutions due to its massive scale, data richness, AI, ML applications, and the rate of technology changes that it can bring about (Philbeck and Davis, 2018).

Shipping 4.0, an extended version of Industry 4.0, reflects the development and adaptation of digitalization and automation applications for the shipping industry. It can be considered a framework expected to transform the shipping industry with economic growth and new technology innovations. This transformation can affect all aspects of the shipping industry, from port operations (Muhammad et al., 2018) to propulsion monitoring systems under ship energy management digital applications (Aiello et al., 2020b; Ang et al., 2017). The incorporation of various advanced digital technologies under the shipping 4.0 framework such as the IoT (Perera, 2017), cyber-physical systems (Ang et al., 2017),

[*] Corresponding author.
E-mail address: mahmood.taghavi@uit.no (M. Taghavi).

https://doi.org/10.1016/j.oceaneng.2024.117098
Received 2 August 2023; Received in revised form 24 January 2024; Accepted 8 February 2024
Available online 24 February 2024

Nomenclature

E	Expectation Operator
f	Gaussian Distribution
h	Mixture Density Model
J	Number of the Cluster
L	Log-Likelihood Function
P_j	Probability of *j*th Cluster
Q	Expectation Function
V	Singular Vector
x	Data Point (Observed Features)
y	Data Point with all Features (Including Cluster Number)
θ	Parameter Vector of the Gaussian Distribution
Θ	General Parameter Vector
μ	Mean Vector
Σ	Covariance Matrix

ABBREVIATIONS

AI	Artificial Intelligence
AIS	Automatic Identification System
ANN	Artificial Neural Networks
DSS	Decision Support System
EEI	Proposed Energy Efficiency Index
EEDI	Energy Efficiency Design Index
EEOI	Energy Efficiency Operating Indicator
EEXI	Energy Efficiency Existing Ship Index
EM	Expectation Maximization
EP	Engine Power
ES	Engine Speed
FC	Fuel Consumption Rate
GMM	Gaussian Mixture Models
IoT	Internet of Things
IMO	International Maritime Organization
KDE	Kernel Density Estimator
ML	Machine Learning
pdf	Probability Density Function
SOG	Speed Over Ground
SVD	Singular Value Decomposition

cloud computing and blockchain (Lambrou and Ota, 2017; Perera and Czachorowski, 2019), AI, and robotics (Muhammad et al., 2018), advanced big data analytics (Bui and Perera, 2021), and augmented reality (Sepehri et al., 2022) offers a range of significant benefits. These large-scale data sets and data analytics make the path towards digital twin type applications, where the physical asset and its replica of digital asset, i.e. digital model, can be combined through the data flow, to update time-varying features of the digital models. The benefits of such digital twin technology include enhanced safety (Sepehri et al., 2022), increased operational efficiency, less environmental impacts (Aiello et al., 2020a), and the sustainability of shipping operations, along with the potential for developing new and innovative business models within the shipping industry (Lambrou and Ota, 2017; Lambrou et al., 2019).

The shipping 4.0 framework, similar to its counterparts in industry 4.0, is built upon a massive amount of online and offline data generated and stored in ocean-going vessels by onboard IoT. A larger amount of data can facilitate the development of predictive data-driven applications (Rødseth et al., 2016). Conventional mathematical models based empirical ship performances and navigation models cannot process this amount of data/information online or offline. Other challenges in conventional mathematical models can be categorized as: system-model uncertainties, sensor noise and fault conditions, and complex parameter interactions. As a result, such models may not predict actual ship performance and navigation behavior correctly, jeopardizing the validity of navigation and operation strategies and decisions taken afterward (Perera and Mo, 2016). For this purpose, appropriate data analysis approaches should be utilized to extract information from the operational conditions and environment. This is where data-driven methods such as Machine Learning (ML) techniques can be superior to conventional mathematical or empirical models. As an example, a linear constant velocity model is a method being used to predict ship's future trajectories for collision avoidance purposes in many ship navigation systems (Xiao et al., 2019). However, this method is limited to non-complex behaviors. More sophisticated methods, such as kinematic motion models coupled with Extended Kalman Filter (Wang et al., 2022) for online data updates, can be utilized to predict future vessel behavior but in no more than a few minutes, despite the higher accuracy they propose. On the other hand Murray and Perera (2021) developed a deep learning framework that successfully predicts a selected vessel's most likely future trajectories based on its past behavior, even for complex vessel motions, including trajectory outliers. This framework was developed based on Automatic Identification System (Lambrou et al.) data in different clusters as the ship trajectories with similar behavior characteristics in a selected geographic area.

As a result, due to the flexibility, reliability, and accuracy in data-driven methods, the digital twin development has started gaining momentum and attracted significant attention recently, particularly in ship design, construction, maintenance, and operation (Mauro and Kana, 2023). The digital twin approach in system modeling (Pires et al., 2019) can be defined as developing a virtual representation of a physical system or a process, which can demonstrate the same behavior as the actual system or process in simulated conditions. The development of digital twins is an iterative process. In this process, the model evolves by validating and adapting itself to new data from the real system to accurately mirror and predict its behavior over time. This iterative nature of the process necessitates communication and data transmission between the digital twin and the physical system (Assani et al., 2022). The developed digital twin framework can be used to monitor, control, and optimize the system performance with higher efficiency, since the model updates is being done by online data, continuously. As an example, (Lee et al., 2022) propose a digital twin for online simulations of ship operations, decision-making, and navigation control in seaways based on data from a physics-based model of ship motions in different environmental conditions.

1.1. Digital twin applications

In the following, some important applications of data-driven methods and digital twin approaches in the shipping industry are discussed under the respective categories.

Port and Terminal Operations: Digital twins can be a part of port operations, giving the operators more insight into the ongoing activities. In this case, a digital twin can act as a part of a Decision Support System (DSS). Port operators can make more informed decisions through real-time data and analytics on various aspects of port operations provided by the digital twin (Zhou et al., 2021). This can be an effective tool to optimize operations and reduce congestion and delays at terminals, guaranteeing a safer operation. For instance, Pang et al. (2021) outline the development of a digital twin and digital thread framework for a shipyard aimed to improve the efficiency of shipyard operations through real-time data analysis, which enables efficient and effective decision-making processes in different scenarios. The models developed for this application are looking at the operations in a higher level. As a result, they may not be applicable for of energy efficiency enhancement of the shipping industry.

Crew Training and Simulation: Digital twins can have the ability to support crew and port personnel training activities (Major et al., 2021) by simulating real-world scenarios in a safer environment. Seafarers can practice the proper reactions in case of an unforeseen incident and gain confidence, which improves operational efficiency, boosts performance safety, and reduces the risk of accidents such as vessel collisions (Arrichiello and Gualeni, 2020). Models developed for crew training have a different purpose than investigating the vessel performance in detail and cannot be utilized for the purpose of ship energy efficiency.

Ship Design: Digital twins can be used as a framework to simulate and test different designs, materials, and equipment configurations at different stages of the ship design process, from concept designing to prototyping, testing, and manufacturing (Lo et al., 2021). This can help manufacturing companies and engineers to optimize the expected performance of the ship to be built, reduce costs, and improve safety (Arrichiello and Gualeni, 2020) at the design stage. This approach can also be utilized to enhance the operation and maintenance of ship systems, which can result in the efficiency and sustainability of ship operations (Perabo et al., 2020). Models developed for this application are more focused on the ship design phase with simulated configurations with different vessel characteristics, and that may not be applicable for the ship operational phase.

Condition Monitoring and Predictive Maintenance: Digital twins developed by ML algorithms can constantly analyze the online data gathered from sensors installed onboard the vessel and predict potential failures of different machinery and equipment. This can help organizing the maintenance schedule in a more effective way. As a result, methods based on digital twins can offer significant advantages over traditional condition monitoring methods by improving maintenance scheduling and preventing unexpected breakdowns. For example, (Johansen and Nejad, 2019) present a digital twin framework for real-time condition monitoring of drivetrains in marine applications to compensate for the inadequacy of traditional methods. This framework is developed using data from various sensors and system models, allowing for early detection of potential failures. Models used for predictive condition monitoring are trained to classify data points or series of behaviors into classes that lead to different system failures. This framework can predict the failure or give a warning if the probability of the system degradation is significant. In this framework, the model development is performed in a supervised approach, by using the prior knowledge and data about previous failures. However, similar approaches can be adopted towards studying energy efficiency in the shipping industry.

Performance Monitoring, Prediction, and Evaluation: The digital twin has the potential to predict a ship's behavior remotely by tracking sensor data from various systems onboard, such as propulsion and electrical systems (Major et al., 2021). Moreover, since the digital twin can simulate the ship's behavior, it can be used for predicting its performance (Lee et al., 2022), such as fuel oil consumption (Gkerekos et al., 2019). Different configurations or effects of different parameters can be investigated by having a digital twin of a selected vessel. As an example, Taskar and Andersen (2021) investigates the accuracy of various methods for calculating added resistance in ship hydrodynamics using digital twin simulations and full-scale measurements. Coraddu et al. (2019) develop a data-driven digital twin for estimating the speed loss caused by marine fouling using Deep Extreme Learning Machine. The general architecture of the models used for this application reflects some similarities with the model used for energy efficiency evaluation in this research, while used for different purposes. However, most research topics are merely focusing on the model development itself and don't discuss the process of feature selection, i.e., the variables used in model development. Moreover, many research topics in this area, don't consider different localized states and operational modes of the vessel and that can be an important step in a digital development framework.

1.2. Energy efficiency

The shipping industry is the primary mode for global transport, accounting for transportation of around 80% of traded goods globally (Bui and Perera, 2021), which is constantly growing. Due to the rate of energy consumption in this industry and its consequent emissions, International Maritime Organization (Gkerekos et al.) has established strict regulatory requirements to improve the energy efficiency of vessels, such as the Energy Efficiency Design Index (EEDI) for new ships, Energy Efficiency Operating Indicator (EEOI), and the Energy Efficiency Existing Ship Index (EEXI) required to be calculated for every ship (Bazari, 2020). A proper energy efficiency index (EEI) can be used to compare different ships' performances and different behaviors, i.e., operating modes of a selected vessel. For these purposes, the use of onboard sensors and modern data acquisition systems provide an excellent opportunity for researchers. However, the formidable side of ship performance and navigation data is its volume and complexity, which needs some novel approaches for correctly analyzing the structures behind the data, such as ML and Statistical Methods (Rødseth et al., 2016). As a result, the combination of ML techniques and EEIs can form a powerful analysis tool for vessel performance.

1.3. ML algorithms

Data-Driven applications associated with ML algorithms have been extensively used in various transport means, such as Intelligent Transport Systems (ITS) (Zhang et al., 2011), and in the preceding years, it has also made its way into the shipping industry. ML techniques have been extensively used in the categories mentioned in the previous section (Munim et al., 2020) in digital twin development. A significant number of research topics are focused on ship performance predictions, including the fuel consumption of vessels using different ML/AI approaches and techniques (Uyanık et al., 2020). Generally, ML applications can be categorized into classification/clustering, and regression.

In classification, the model is developed in a supervised framework to learn an existing pattern in the data. As an example, Kraus et al. (2018) presented an automated system for classifying ships and determine their types based on their movement patterns and trajectory data using k-nearest neighbors, support vector machines, and random forests.

On the other hand, clustering can be an unsupervised learning approach, and many algorithms have been used to perform it. As an example, Tran (2020) proposed a multicriteria decision making process based on fuzzy clustering method to achieve the optimal loading of the ship and fuel oil consumption of the main diesel engine. Bui and Perera (2021) proposed a big data analytics framework based on a two-step cluster analysis for ship performance monitoring under localized operational conditions to quantify its performance in each of the respective conditions.

From all the methods used for clustering the data, K-Means can be considered as a popular algorithm. For example, Yan et al. (2018) used a distributed parallel K-Means clustering algorithm to perform route divisions for a vessel to enhance its energy efficiency and reduce emissions. However, K-Means impose a severe limitation in cluster shapes, i. e., the resulting clusters in K-Means are assumed to be spherical. On the other hand, the dataset in current research has complex conditions of data distributions, and the clusters may not essentially have a spherical shape. Moreover, in K-means each cluster is modeled only by its centroid's position, and its geometric shape and volume are not measured in terms of its density. The reason for these issues is that the K-means algorithm works based on Euclidean distance between data points. As a result, the captured clusters' boundaries have spherical forms if the cluster centers are far enough from each other, or the feature space is separated into different regions, which is known as *Voronoi tessellation*, if cluster centers are not far enough, and the cluster boundaries are straight lines (Raykov et al., 2016). Furthermore, K-Means is a deterministic approach, which means it assigns each data point to one and only one cluster. To address the weaknesses and issues associated with K-Means, Gaussian Mixture Models (GMM) is selected to perform the clustering in this research study because of the nature of the dataset. On the other hand, since GMM is based on the Bayesian point of view, it gives the probabilities of belonging to each cluster for all data points. In this way, by having a membership function for each data point, the clustering results have more flexibility for function approximation for future prediction purposes and digital twin development.

When the dataset is grouped into appropriate data clusters, the next step would be to understand the relationship, i.e. the correlations, among the respective parameters in the dataset. That can be done by implementing an appropriate regression approach. Regression can also be referred to as a statistical technique that finds a relationship among dependent variables based on the values of one or more independent variables. Various regression techniques are used in research studies for performance prediction purposes, such as Kernel-based support vector regression (Wang et al., 2020), LASSO regression (Wang et al., 2018), multiple linear regression, and Artificial Neural Networks (ANN) (Farag and Ölçer, 2020). In this regard, one of the algorithms that can be used to find the relationship between the variables and the shape of the data is Singular Value Decomposition (SVD). Preserving important information in the dataset, computational efficiency, and approximation with a lower-rank matrix make this method an attractive approach for dimensionality reduction and data compression (Perera and Mo, 2016), image processing (Hameed et al., 2020), fault detection (Li et al., 2019), and information extraction and retrieval applications (He et al., 2015).

1.4. Contribution of the current research

Many research topics in the recent literature have been investigating in relation to ship energy efficiency (Abebe et al., 2020; Gao et al., 2023). These topics have different perspectives ranging from vessel design modifications to operational practices, policies, potential advances in the shipping industry and their feasibility, and alternatives to existing systems, operational procedures, and policies (Barreiro et al., 2022). In this regard, data-driven methods can provide a powerful tool for monitoring, modeling, and predicting ship performance, thus helping to optimize energy efficiency in the shipping industry. However, several studies currently focus on investigating the considerable potential of data-driven approaches with ML algorithms in this field (Jimenez et al., 2022). As an example, data-driven approaches can be adopted to calculate ships' energy efficiency indicators, which can be used to assess the effects of different factors (Yuan et al., 2017) and routine procedures that can have an effect on ship energy efficiency (Shaw and Lin, 2021). These approaches can also be applied to develop data driven models for predicting ships' speed (Abebe et al., 2020) and improving the quality of navigation strategies (Perera and Mo, 2017), potentially enabling route optimization and energy-efficient shipping. The potential of data-driven approaches with ML algorithms can enhance energy efficiency and sustainability of the shipping industry (Huang et al., 2022). Still, more advanced, localized, ship-specific, and reliable methods are needed to provide accurate models of vessel behavior, considering the individual characteristics of various vessels, with online model update capabilities, which can improve energy efficiency in the shipping industry. Most research studies present single models to cover the entire operating range of ocean-going vessels. As an example, Öztürk and Başar (2022) proposed a DSS for energy efficiency in shipping using a multiple linear regression approach and ANNs to predict the fuel consumption and emissions of a selected ship based on RPM, trim, ballast, and weather data gathered from voyage reports of 19 container ships. Although the results show an acceptable fit for the data, using a single model for the entire operating region of 19 vessels raises questions about the validity and probable overgeneralization of the simulation, thus, the decisions made upon them. Therefore, the existing research topics utilizing data-driven and ML approaches are often inadequate for localized operational conditions, leading to inefficiencies and suboptimal performance (Bui and Perera, 2021). Zhang et al. (2019) also proposed a data-driven approach for analyzing and optimizing ship energy efficiency in Arctic waters, in which the proposed model has been developed using a simple ANN. Although, a more sophisticated and detailed model can provide a better basis for optimization purposes. Peng et al. (2020) developed ML-based models aim to predict the energy consumption of ships in ports, and propose reduction strategies, although the statistical deviations in data collection, limits the applicability of this framework. Despite many research topics in the literature, a localized model development approach can fit into different vessel operational conditions, which can improve the accuracy of model predictions.

As mentioned, Digital Twin model development can consist of ML algorithms involving data clustering and regression analysis. Of all these research topics and applications of ML and digital twin development, this research aims to develop a vessel-specific data-driven framework for energy efficiency improvement of the respective vessel with focus on data clustering. This distributed localized operational modes-based framework is intended to be utilized in an onshore operation center, where it can be utilized to facilitate the individual analysis of data transmitted from various vessels. For developing this framework, an unsupervised algorithm, i.e., clustering, has been performed on dataset of a vessel navigational and operational variables to find similar behaviors, or in other words, different operating regions of the vessel. This way, a localized analysis of each operating region, e.g., engine operational modes or trim-draft combined conditions, can be performed, acting as a key component of a digital twin for evaluating ship energy efficiency. The proposed distributed localized operational modes-based model can exhibit an acceptable performance since that can capture more localized vessel operational and navigation conditions. As a result, it is superior to approaches in which the same model is used for the whole operating region of the vessel with different configurations and characteristics. Since localized models are developed in this framework simulating the vessel's behavior, this framework can also serve as a basis for DSSs, while supporting data anomaly detection, and recovery of missing data.

The main contribution of this research study is to develop a digital twin framework, by utilizing machine learning algorithms, to find the most frequent operating regions of a selected vessel. That can also support investigating the Fuel Consumption rate (FC) behavior in the resulting operating regions and find the optimal operation and navigation strategy for the vessel. For this purpose, extensive low-level clustering and subsequent cluster analysis have been performed in the maritime context to serve advanced green ship operations. The final clustering is performed based on the GMM approach in a 3D feature space with the main Engine Power (EP) in kW, Engine Speed (ES) in RPM, and Speed Over Ground (SOG) in kn as the features to find the respective data clusters, i.e., vessel operating regions. The Expectation Maximization (Parzen, 1962) algorithm is applied to calculate the parameters of the cluster distributions (the respective mean and covariance values). The method for clustering presented in this research is built upon a preliminary work (Taghavi and Perera, 2022). Each cluster represents a different navigational and operational conditions. Finding the suitable number of clusters for a given dataset always presents a challenge, as it must be determined and provided as an input to the algorithm before its execution. In this research, a framework for selecting the proper number of clusters for the GMM algorithm is also presented. For this purpose, the separability measures are used to determine the proper number of clusters along with 1D and 2D KDE plots. After finding the number of data clusters, an SVD analysis is performed in each cluster to find the relationship between different variables and the dominant singular directions in each cluster. The presented SVD analysis, can build the basis for future dynamic model development, anomaly detection, and missing data recovery as a part of digital twin applications. In the final step of cluster analysis, two EEIs are defined, and the vessel's performance in different data clusters is compared, and the most efficient behavior and navigational strategy among all operating regions are determined. In the defined EEI, in the absence of cargo amount of the vessel, the draft value is utilized as a proximate measure for cargo amount. This proposed framework for developing digital twin applications in shipping is demonstrated in Fig. 1. The shipping industry can utilize the contributions of this research to meet the IMO energy efficiency requirements enforced by its regulations.

The assumption in this research is that in the development of the digital twin, only datasets from the vessel are used, therefore no conventional mathematical or empirical ship performances and navigation models are extensively used. However, the respective domain knowledge in shipping is used to support data clustering and regression approaches in this study.

In the following, the proposed methodology for digital twin development is discussed in section 2. Section 3 presents computational results and discussion, including the summary of the dataset used, the preprocessing steps taken, cluster number identification methods, feature selection, and data clustering. Finally, the conclusions are presented in section 4.

2. Methodology

This section presents the methodology used for analyzing the respective data sets. As illustrated in Fig. 1, The initial development phase of this framework is data pre-processing. This step includes the removal of anomalies and missing values to ensure a high-quality data for subsequent use in the clustering algorithm. In this research GMM coupled with EM is employed for capturing the clusters. The execution of this clustering algorithm necessitates the pre-determination of the number of data clusters. In this research study the number of data clusters is determined by using 1D and 2D KDE plots, in conjunction with comparing separability measures of the cases with different cluster numbers. Subsequent to capturing the clusters, the cluster analysis step is performed, which encompasses two distinct phases. The first phase of the cluster analysis involves conducting SVD analysis within each cluster to find the relationships between variables. This SVD can serve as a basis for the dynamic digital twin development. The second phase focuses on evaluating the energy efficiency of the main operating regions, i.e. data clusters. For this purpose, two EEIs are proposed, and the performance of the vessel in different clusters is compared based on them, resulting in the identification of the most efficient operating region.

Section 2.1 introduces the proximity measure used for measuring the similarity and distance among data points. Section 2.2 presents the GMM-EM algorithm, used for data clustering. In this section also the steps taken to implement GMM-EM and the respective equations are presented. In 2.3, the KDE plots and separability measure used for identifying the number of clusters in the datasets are discussed. Section 2.4 discusses the SVD technique. Finally, the proposed EEIs used to quantify ship performance are introduced in Section 2.5.

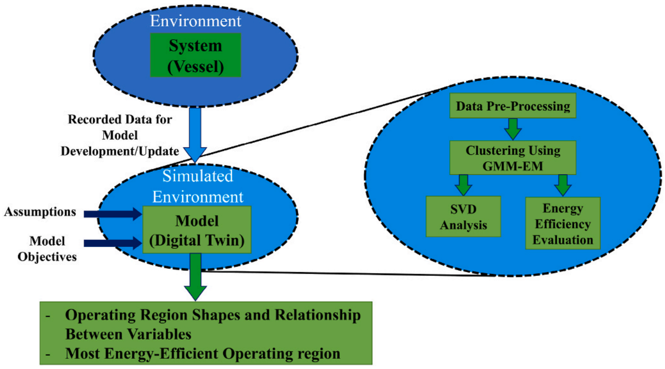

Environment
System (Vessel)
Recorded Data for Model Development/Update
Simulated Environment
Assumptions
Model Objectives
Model (Digital Twin)
Data Pre-Processing
Clustering Using GMM-EM
SVD Analysis
Energy Efficiency Evaluation
- Operating Region Shapes and Relationship Between Variables
- Most Energy-Efficient Operating region

Fig. 1. The proposed framework for digital twin in shipping.

2.1. Proximity measures

A dataset from a marine engine of a selected vessel has been considered for the clustering purpose, since that can create the basis for the proposed digital twin farmwork. As the first step of this study, this dataset is analyzed to identify the respective clusters that can represent the engine operational regions, i.e. the vessel's operating regions.

An unsupervised learning algorithm is used for the same purpose, where the general idea of clustering is based on the density of the data distribution. A dataset can consist of different clusters, it is assumed that data points that are in a near neighborhood of each other can be classified as a data cluster. In other words, there can be proximities among such data points in the same cluster, and the dissimilarity can be maximum between data points in different clusters. In a majority of data analysis applications, proximity and dissimilarity are measured based on distance functions, such as Euclidian or Mahalanobis distances (Theodoridis and Koutroumbas, 2006). The Euclidian distance between two points is simply the length of the direct line connecting them in an N-dimensional space. On the other hand, the Mahalanobis distance, d_M, measures the distance of a point with respect to an estimated mean value (or a point) in an N-dimensional space, such as $x \in R^n$, from a distribution D with mean vector $\mu \in R^n$ and a covariance matrix $\Sigma \in R^{n\times n}$ using Eq. (1):

$$d_M(x, D) = \sqrt{(x-\mu)^T \Sigma^{-1}(x-\mu)} \ (d_M \in R) \tag{1}$$

Based on the concept of proximity in clustering, it is logical to assume each operating region of the selected vessel forms a cluster based on the operational parameters, i.e., the engine follows the same behavior in each operating region. As a result, the term "cluster" is a representation of a specific operating region of the selected vessel in this research study. Based on this, since in this research study, GMM-EM is used to perform the clustering, i.e., in calculating the distance, the covariance matrix of each cluster is also considered, and the Mahalanobis distance is the measure for distance calculations.

2.2. GMM-EM algorithm

Clustering is an unsupervised learning algorithm that can be done by the GMM-EM algorithm, which is one of the most powerful clustering methods. GMM is a probabilistic clustering algorithm based on data distributions that can be updated in the context of the Bayesian approach. In this approach, all the data points are represented as statistically random variables with a related probability density function (pdf) consisting of a mixture of a finite number of Gaussian distributions with unknown parameters (Bishop, 2006) for a large-scale dataset. In other words, a large-scale data distribution can be assumed as a combination of J separate multivariate Gaussian distributions, denoted by $f(x^q;\hat{\theta}(t)|j)$. These distributions, i.e., data dense regions, are considered as clusters in this research study, as the respective localized operation regions for the vessel. Hence, each data point in the dataset belongs to a cluster with a prior probability of P_j. As a result, the pdf of the dataset or mixture density model, h, can be written as Eq. (2).

$$h(x^q; \widehat{\Theta}(t)) = \sum_{j=1}^{J} f(x^q; \widehat{\theta}(t)|j) P_j$$

$$f(x^q; \widehat{\theta}(t)|j) = \frac{1}{\sqrt{(2\pi)^n |\Sigma_j|}} exp\left(-\frac{1}{2}(x^q-\mu_j)^T \Sigma_j^{-1}(x^q-\mu_j)\right)$$

$$\Theta = \begin{pmatrix} \theta \\ P \end{pmatrix}, P = \begin{pmatrix} P_1 \\ P_2 \\ \vdots \\ P_J \end{pmatrix}, \theta = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_J \end{pmatrix}, \theta_j = \begin{pmatrix} \mu_j \\ \Sigma_j \end{pmatrix}$$

$$\sum_{j=1}^{J} P_j = 1 \tag{2}$$

The parameter vector Θ contains two sets of parameters, θ, and P, which are going to be estimated using the EM algorithm. θ_j consists of the mean vector and covariance matrix of the *j*th cluster, and the value of P_j for each data point is the probability that this data point belongs to cluster *j*. As a result, the summation of all P_j values should be equal to one (Eq. (2)). Ultimately, each point is assigned to a cluster whose P_j is the maximum for that data point. One should be noted that the value of *J,* i.e., the number of data clusters, should be selected before implementing the GMM-EM algorithm. In section 2.3, the proposed approaches in this research study to select a proper number of data clusters in a selected dataset are discussed.

The nature of the GMM-EM algorithm is defining a joint distribution over the observed and latent variables. As a result, estimating the GMM parameters can be achieved via the EM algorithm. In the following formulations, vector $\boldsymbol{x}$ is the data point that is obtained from the observations, i.e., sensor measurements. Additionally, a new set of variables, $\boldsymbol{y}$, is defined, with an observed part, $\boldsymbol{x}$, and an unobserved part, $\boldsymbol{j}$, which corresponds to the cluster number of the data point $\boldsymbol{x}$.

In this algorithm, the likelihood function describes the joint probability of the observed data as a function of the parameters of the Gaussian distribution. The log-likelihood is the logarithm of the same likelihood function. In order to find the parameters, the likelihood or log-likelihood function is maximized. Based on this, the next step is the maximization of the likelihood function. Since the logarithm function is monotonically increasing, the log-likelihood function, defined as Eq. (3), is maximized using the EM algorithm to estimate the parameters.

$$L(\theta) = \sum_{q=1}^{Q} [lnf(y^q; \theta | x^q)] \tag{3}$$

The EM algorithm is an iterative scheme consisting of two steps. The first step is the E-Step that calculates the expectation of the log-likelihood function. In the M-Step, as the second step, the derivative of the expectation of the log-likelihood function with respect to the parameters is calculated and set to zero. This way, the values for parameters that maximize the log-likelihood function are found. In the EM algorithm iterative scheme, only in the distribution of θ in the E-step the last calculated values of $\theta(t)$ are used. This means that the parameter values in this section of the E-step are considered constant values. In the following, the equations for these steps are presented.

2.2.1. E-step

For this step, a new function, *Q*, is defined as Eq. (4), which calculates the expectation of the log-likelihood function. As mentioned, in the distribution of this function, $\theta(t)$ is considered a constant, and the values calculated in the previous iteration are used.

$$Q(\theta; \widehat{\theta}(t)) = E\{L(\theta)|\widehat{\theta}(t)\} = \sum_{q=1}^{Q} [E\{lnf(y^q; \theta | x^q)|\widehat{\theta}(t)\}]$$

$$= \sum_{q=1}^{Q} \sum_{j=1}^{J} [ln[f(x^q; \theta | j).P_j].P(j; \widehat{\Theta}(t)|x^q)]$$

$$= \sum_{q=1}^{Q} \sum_{j=1}^{J} \Bigg[\bigg[-\frac{n}{2} ln(2\pi) - ln(|\Sigma_j|)$$

$$-\frac{1}{2}(x^q - \mu_j)^T \Sigma_j^{-1} (x^q - \mu_j) + ln(P_j)\bigg].P(j; \widehat{\Theta}(t)|x^q)\Bigg] \tag{4}$$

2.2.2. M-step

In the M-step, the derivatives of function Q with respect to Σ_j, μ_j, and P_j are calculated and set to zero. The solution to the resulting equations is the values of Σ_j, μ_j, and P_j for the next iteration, which is presented in Eq. (5). As it can be seen in these equations, the resulting values are calculated based on the values for the previous iteration. This process is repeated until convergence.

$$P(j; \widehat{\Theta}(t)|x^q) = \frac{f(x^q; \widehat{\theta}(t)|j)\widehat{P}_j(t)}{\sum_{i=1}^{J} f(x^q; \widehat{\theta}(t)|i)\widehat{P}_i(t)}$$

$$\widehat{\mu}_i(t+1) = \frac{\sum_{q=1}^{Q} P(i; \widehat{\Theta}(t)|x^q)x^q}{\sum_{q=1}^{Q} P(i; \widehat{\Theta}(t)|x^q)}$$

$$\widehat{\Sigma}_i(t+1) = \frac{\sum_{q=1}^{Q} P(i; \widehat{\Theta}(t)|x^q)(x^q - \widehat{\mu}_i(t+1))(x^q - \widehat{\mu}_i(t+1))^T}{\sum_{q=1}^{Q} P(i; \widehat{\Theta}(t)|x^q)}$$

$$\widehat{P}_i(t+1) = \frac{1}{Q} \sum_{q=1}^{Q} P(i; \widehat{\Theta}(t)|x^q) \tag{5}$$

After deriving the equations for the parameters, the iterative scheme for the EM algorithm can be implemented. For this purpose, the algorithm starts from initial estimated values of Σ_j, μ_j, and P_j. Then, parameter values are calculated based on Eq. (5) for subsequent iterations. This iterative process is repeated until the convergence of all parameters. In other words, when the change in parameters in two successive iterations are negligible it can be assumed that the algorithm is converged. Since any change in the parameter values will be mirrored in the likelihood function, the absolute difference in the log-likelihood function between two successive iterations is less than a specified threshold. Typically, if it is less than 1%, then it is assumed that the convergence is achieved, and the iterative algorithm is terminated.

2.3. Data cluster identification

Since in GMM algorithm the cluster labels for the data points are not given in advance, there is no information on the number of clusters in the dataset. The number of clusters as the first step is determined in the clustering step, where several techniques can be used. This research study proposes two approaches to ensure the proper number of clusters is selected and those methods can complement to each other to verify the selected number of clusters, appropriately.

2.3.1. KDE plots

As the first method to find a proper number for the clusters, 1D and 2D Kernel Density Estimator (KDE) diagrams of different operating parameters are used. Using this approach, an initial idea about the number of different clusters, i.e., the operating regions of the vessel, can be achieved. The respective KDE diagrams are plotted based on the Parzen Density Estimation approach (Parzen, 1962), i.e., a statistical approach that considers the variables as stochastic variables and is used to find the distribution of a given dataset. In these diagrams, the joint pdf of the respective variables is plotted. The regions with higher densities are represented as peaks in 1D KDE diagrams with the denser regions in 2D KDE diagrams. These regions with higher data densities can be assumed to be clusters due to their data concentrated regions, and that can be a measure of the number of clusters in the whole dataset.

2.3.2. Separability measure

Apart from the KDE plot, which provides an overview of the data distribution and gives an initial estimation on the respective number of clusters from the observations, another approach is needed to calculate a quantitative index for comparing the result of clustering with different cluster numbers and finding a proper number of clusters. For this purpose, to compare different cluster configurations, a separability measure based on cluster scatter matrices is used (Theodoridis and Koutroumbas, 2006). This measure evaluates the clustering result based on how the clusters are compact in their structure as well as separated from other clusters. In other words, this measure is based on the concept that the data points in a cluster should be closer to each other and data points in different clusters should be as far as possible from the same data points. This separability measure is generally a ratio of the distance between points in a selected cluster vs. the distance between other clusters.

To define the separability measure, the following terms should be introduced.

The within-class scatter matrix, S_w, is defined as:

$$S_w = \sum_{i=1}^{M} P_i \Sigma_i \tag{6}$$

where Σ_i is the covariance matrix of cluster ω_i.

$$\Sigma_i = E\left[(x-\mu_i)(x-\mu_i)^T\right] \tag{7}$$

And P_i is the ratio of the number of data points in cluster ω_i, n_i, and the total number of data points, N, and is calculated as:

$$P_i = {}^{n_i}\!/_{N} \tag{8}$$

Based on this definition, S_w can measure the average covariance and variance of all the features over all classes.

The between-class scatter matrix, S_b, is defined as:

$$S_b = \sum_{i=1}^{M} P_i(\mu_i-\mu_o)(\mu_i-\mu_o)^T \tag{9}$$

where P_i can be derived from the previous equation, and μ_o is the global mean vector calculated using Eq. (10).

$$\mu_o = \sum_{i=1}^{M} P_i \mu_i \tag{10}$$

In this formulation, *trace{Sb}* can be interpreted as a measure of the average distance of the mean of all classes from the global mean value.

The mixture scatter matrix, S_m, is defined as:

$$S_m = E\left[(x-\mu_o)(x-\mu_o)^T\right] \tag{11}$$

In other words, S_m is the covariance matrix with respect to the global mean. It can be shown that:

$$S_m = S_w + S_b \tag{12}$$

Since comparing two matrices is impossible, an operator or a function that maps each matrix to a number should be selected. For this purpose, different operators, such as trace and determinant, can be used to form the final criterion. In this research study, the following criterion, *SM*, is used since it is invariant under linear transformations:

$$SM = trace\left\{S_w^{-1} S_m\right\} \tag{13}$$

This criterion takes larger values when samples in the feature space are well separated as data clusters and the respective data points are clustered around their mean values.

In order to use this separability measure, *SM*, to find the proper number of data clusters, a clustering algorithm is implemented with different cluster numbers, then a separability measure using scatter matrices is utilized to understand cluster distributions. The proper number of data clusters is decided by comparing the separability measures of all scenarios.

2.4. Singular Value Decomposition (SVD)

SVD analysis on a matrix is one of the most powerful algorithms in linear algebra used for analysis of multivariate data. This technique is based on decomposing a matrix into three separate matrices and finding the singular values and singular vectors of the matrix. Given an $l \times n$ matrix X of rank r, using SVD it can be represented as the product of three matrices: U, Y, and V of the dimensions $l\times l$, $l\times n$, and $n \times n$ respectively so that:

$$X = U\begin{bmatrix} \Lambda^{\frac{1}{2}} & 0 \\ 0 & 0 \end{bmatrix} V^H \text{ or } Y \equiv \begin{bmatrix} \Lambda^{\frac{1}{2}} & 0 \\ 0 & 0 \end{bmatrix} = U^H X V \tag{14}$$

where subscript H denotes the Hermitian operation, that is, complex conjugation and transposition. $\Lambda^{\frac{1}{2}}$ is the $r\times r$ diagonal matrix with elements $\sqrt{\lambda_i}$, and each λ_i is a nonzero eigenvalue of the associated matrix $X^H X$. U and V are unitary matrices that transform X into the special diagonal structure of Y. This equation can be rewritten as Eq. (15) by assuming u_i and v_i, the eigenvectors corresponding to the nonzero eigenvalues of the matrices XX^H and $X^H X$, as the column vectors of matrices U and V.

$$\begin{aligned} X &= \begin{bmatrix} u_0 & u_1 & \cdots & u_{r-1} \end{bmatrix} \begin{bmatrix} \sqrt{\lambda_0} & & & \\ & \sqrt{\lambda_1} & & 0 \\ & & \ddots & \\ & 0 & & \sqrt{\lambda_{r-1}} \end{bmatrix} \begin{bmatrix} v_0^H \\ v_1^H \\ \vdots \\ v_{r-1}^H \end{bmatrix} \\ &= U_r \Lambda^{\frac{1}{2}} V_r^H \end{aligned} \tag{15}$$

where U_r, called the left singular vectors, denotes the $l\times r$ matrix that consists of the first r columns of U, and V_r, called right singular vectors, denotes the $r\times n$ matrix formed by the first r columns of V. U_r represent the directions in the input space with the highest variances. V_r represent the directions in the projected space associated with the left singular vectors.

For a symmetric matrix X, the left singular vectors and right singular vectors are identical because the transpose of an orthogonal matrix is itself. As a result, the columns of the matrix U, which represent the left singular vectors, can be interpreted as the principal components of the data, and can provide valuable insights into the structure, shape, and properties of the data. As mentioned, for each singular vector, there is also a singular value. The magnitude of the singular value represents the amount of information in the respective covariance direction of a selected data cluster.

2.5. Energy efficiency index (EEI)

Two *EEIs* are defined in this research study based on *EEXI*, with focusing on FC of the selected vessel considering the availability of the performance and navigation variables in the dataset. The proposed *EEIs* calculate the FC rate per distance. For this purpose, from the vessel's FC, which is tons per day, the average rate of FC in tons per minute is calculated by dividing by 1440. On the other hand, it is assumed that the vessel's speed is constant during each minute since the data sampling rate is 1 min. As a result, from SOG values, the distance the vessel moved every minute can be calculated. Based on these values, the first proposed *EEI*, namely *EEI₁*, can be defined as the amount of fuel consumed per nautical mile, in tons, and is calculated by dividing an average rate of FC in tons per minute by the distance traveled per minute in nautical miles as described in Eq. (16).

$$EEI_1 = \frac{FC \div (1440)}{SOG \div 60} = \frac{FC}{SOG \times 24} \tag{16}$$

The loading condition can significantly affect ship resistance, thus the FC of a vessel at a certain speed can be increased due increased ship resistance. However, the dataset used in this study does not consist of the loading condition of the selected vessel for different data points. In this research study, to account for this effect, the vessel's draft values are considered as a measure for the loading condition of the vessel, i.e., the ship cargo amount. As a result, to cancel the effect of draft value, i.e., the loading condition of the vessel, on FC, the proposed index in Eq. (16) has been revised by dividing it by the draft value, which is defined as EEI_2 in Eq. (17). This way, the effects of cargo conditions on the vessel are approximately taken into account by considering its draft values. The vessel energy efficiency in different operating regions can be compared under this proposed index. In other words, EEI_2 can compare the energy consumption per unit distance for a similar loading condition.

$$EEI_2 = \frac{FC}{SOG \times draft \times 24} \tag{17}$$

The proposed *EEIs* are calculated for all the main clusters, i.e., the main operating regions, to determine the most energy efficient configurations of navigational and operational parameters of the selected vessel.

3. Computational results and discussion

In this section, the outcomes of data analysis are presented. In section 3.1, the dataset used in this research study is introduced. Section 3.2 discusses the preprocessing steps. In section 3.3, the results of approaches taken for determining the number of data clusters are presented. In section 3.4, the feature selection step is discussed. In section 3.5, the final data clustering results are presented. In section 3.6, the result of SVD analysis for different operating regions are presented. Finally, in section 3.7, the energy efficiency evaluation of the vessel in different operating regions is represented.

3.1. Dataset summary

The proposed framework is developed using engine datasets from a selected ocean-going vessel described in Table 1. The data points were recorded almost every minute, so the total number of data points is 499,920. The respective ship performance and navigation variables are represented by their statistical distributions to identify the operational modes using GMM.

Table 1
Ship specifications.

Parameter	Particulars
Ship Type	Chemical Tanker
Ship Length	135 (m)
Ship Beam	25 (m)
Deadweight (at Designed Draft)	9500 (tons)
Main Engine Type	Dual Fuel Engine with MCR 4500 (kW) at 720 (RPM)
Gearbox Reduction Ratio	7:1
Propeller Type	A Controllable Pitch Propeller with a Diameter of 5.5 (m) and 4 Blades

3.2. Data preprocessing

As mentioned, a ship navigation and operation dataset consisting of 12 months of the selected vessel is used. However, not all these data points can be used in the clustering process because some time intervals do not consist of engine operational data due to data erroneous conditions or not utilization of the marine engine. Since the scope of this research study is to investigate the operating regions of the vessel, only high-quality data points associated with vessel operational conditions are considered. Fig. 2 shows the marine engine operational data, however the time periods, when the engine is not operating are also noted in the same figure. In this figure, the EP time series for about 70 days, i.e., 100,000 min or data points of the ship's operation, is plotted along with vessel positions, i.e., longitude and latitude. The figure reveals that when the engine is not generating any power, the vessel position remains constant, indicating that the ship is stationary probably in a port.

The data points that have approximately zero values of EP, SOG, and EP are removed from the dataset. As a result, 277,160 data points remain for analysis and the cluster analysis is performed on the resulting dataset.

The dataset is normalized so that the differences in values of different variables do not cause a bias in the analysis and put more importance on the variables with higher magnitudes. In this way, all the features contribute similarly to the log-likelihood function, and the convergence is achieved faster.

3.3. Appropriate data cluster number identification

The methods used for selecting the appropriate number of clusters and their outcomes are discussed in this section.

3.3.1. KDE plots

As the first step, the KDE diagrams are used to get the initial information to determine the respective number of data clusters. For this purpose, 1D and 2D KDE diagrams of the current dataset with EP and ES values are plotted and are shown in Fig. 3. Based on the 1D and 2D KDE plots in this figure, at least 7 regions with denser regions are observed. The centers of these dense regions are connected with red lines to their associated peaks in 1D KDE plots. Hence, it can be concluded that the entire feature space of the pdf of the dataset is a combination of at least 7 dense data distributions, each can be approximated to a cluster. One should note that KDE plots for combinations of other variables are used to find the number of existing data clusters, but here, just one of them is presented.

3.3.2. Separability measure

The second method for determining the number of data clusters and confirm the findings from the previous section involves utilizing the

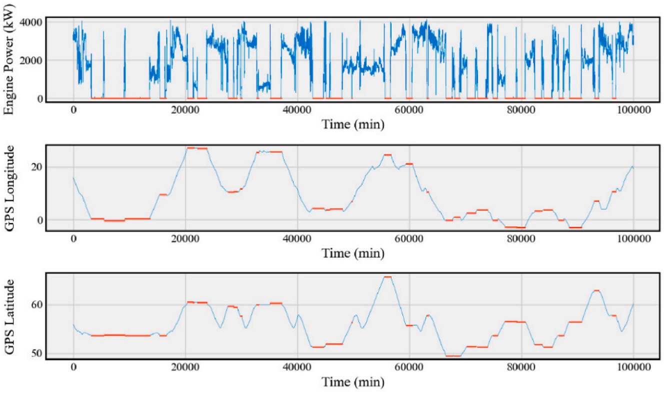


Fig. 2. Time Series Plots of EP and Vessel position in Longitudes and Latitudes for 70 Days.

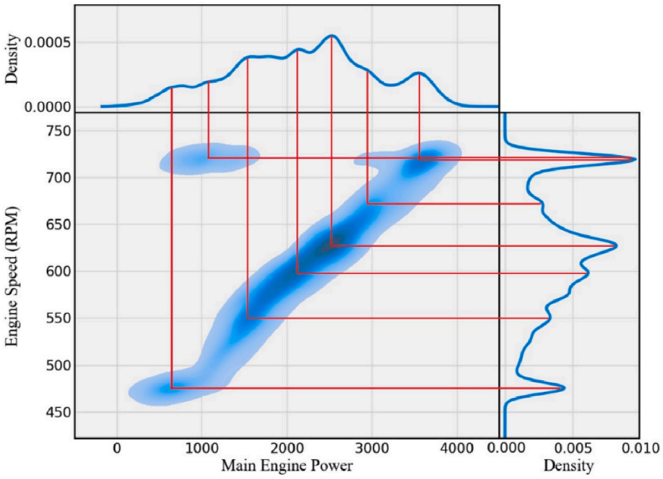


Fig. 3. 2D KDE plot of EP and ES

separability measure outlined in section 2.3.2. For this purpose, the GMM-EM algorithm is implemented for different cluster numbers, and the introduced separability measure is calculated for each case. In the clustering and separability measure calculations, a 7D feature space has been considered to make sure all the aspects of the vessel's operation have been considered and maximum separability is achieved. The following ship performance and navigation parameters are considered for this analysis: EP, SOG, ES, Average Draft, Trim, Relative Wind Speed and Direction.

The separability measure, SM, values for different cluster numbers are plotted in Fig. 4. As seen from this figure, the highest value for the separability measure is achieved for 7 clusters after passing the initial maximum at 3 clusters. One should note that the 3-cluster situation represents a situation, where the respective dataset can be approximated into 3 operational regions. That can be a high-level overview of the vessel operations. However, this study is interested in low-level vessel operational conditions, therefore the 7-cluster situation is considered as a more suitable situation, where more localized ship performance and navigation information can be preserved. Furthermore, this result confirms the number of data clusters observed in KDE plots (see Fig. 3). As mentioned in 2.3.1, the KDE plots already show 7 peaks in 1D KDE and 7 dense regions associated with them in 2D KDE, which represent existing clusters. The same results have been observed using a 3D feature space with EP, ES, and SOG as the variables, as shown in Fig. 5. In this figure also, 7 clusters have the highest separability measure among all cluster numbers after passing the first peak.

One should note that the clustering results performed in this research will serve as a basis for digital twin development in the future study. This digital twin development can consist of both clustering/classification and regression approaches. Hence, having a more detailed and low-level clustering as the main focus of this study increases the model's accuracy. In clustering with higher cluster numbers, the vessel's behavior is more similar in data points of a selected cluster, and it is easier to fit a model to each cluster with higher accuracy. In contrast, in a more general cluster, there may be difficulties in fitting one model into the diverse behaviors of the vessel in subclusters, which can lead to lower accuracy. To avoid this, a two-step clustering may be necessary. For example, in case of selecting 3 as the cluster number, another clustering step should be performed in each resulting cluster to investigate the vessel operations in more detail, classify the data points in more similar operational conditions, and develop a more accurate localized model. This way, their subclusters, representing various vessel operation modes, can be found. For example, (Bui and Perera, 2021) performed a second clustering step in the primary clusters to find the subclusters. However, two-step clustering imposes more computational costs on the algorithm. As a result, 7 is selected as the best option for cluster number to avoid a second clustering step.

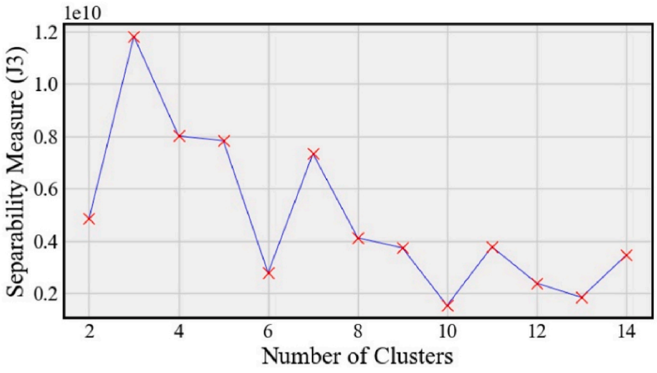


Fig. 4. Separability measure J_3 **values for different numbers of clusters for 7D feature space.**

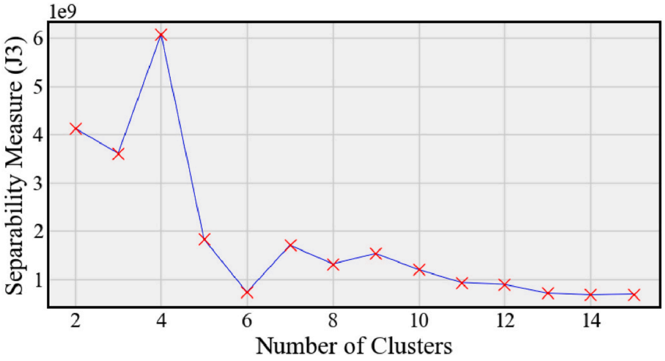


Fig. 5. Separability measure J_3 **values for different numbers of clusters for 3D feature space.**

Based on different criteria discussed previously, it is concluded to perform the clustering algorithm to find 7 clusters or operating regions of the vessel for this dataset.

3.4. Feature selection

In previous sections, it was mentioned that for increasing the separability of the clusters, a higher dimension had been considered for identifying the number of clusters and concluded that the proper number of clusters is 7. However, the computational cost is higher for clustering in a 7D data space, while the clustering result doesn't change proportionally to the computational cost imposed on the algorithm. As a result, the simpler model with the 3D feature space is selected to have a faster approach for online applications in both model development and necessary model updates. To ensure that choosing 3 features does not affect the results considerably, the resulting cluster centers and similar elements in covariance matrices for 3D and 7D clustering are compared, and no considerable difference is observed. This means adding more features to the model will not necessarily result in a more accurate model. Based on this, the 3D cluster with the following operational features of the vessel is selected for further analysis, and the results for this feature space are presented: EP in kW, ES in RPM, and SOG in Kn.

Main Engine FC values are also in the dataset but are not selected as the primary variable in the clustering process. The main reason is that EP and FC have a high correlation in the whole operating range, as shown in Fig. 6. Therefore, selecting a low-dimensional data space that has the most important information can improve the convergence of data clustering. The scatter plots of EP and FC with SOG and ES are presented in Figs. 7 and 8 respectively, representing a data space with complex shapes of data clusters. On the other hand, it can be seen that EP can

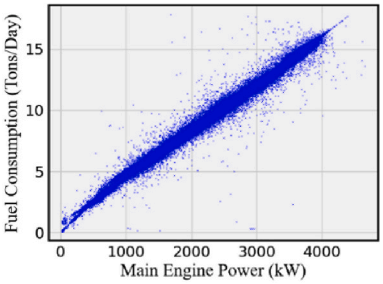


Fig. 6. EP-FC scatter plot.

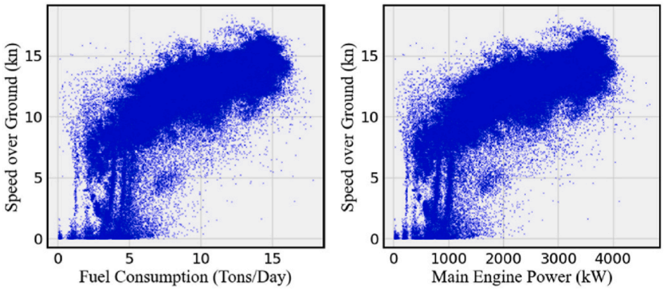


Fig. 7. FC-SOG & EP-SOG scatter plots.

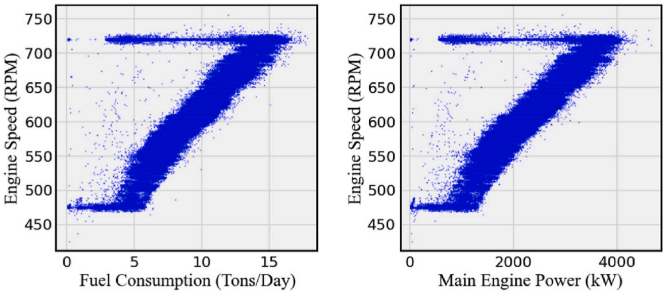


Fig. 8. FC-ES & EP-ES scatter plots.

provide almost all the information FC can provide, since scatter plots with FC are almost the same as scatter plots with EP. As a result, using FC as a feature does not increase the separability between clusters significantly. In other words, adding FC as the fourth feature does not add much information compared to the computational costs it adds to the algorithm, since that can introduce additional complexity into the clustering algorithm. Another reason for not considering FC as the fourth feature is that more than 20% of the data points have missing values for FC, and removing this amount of data points can reduce the accuracy and generalization property of the clustering process significantly.

After finding the operating regions using the clustering algorithm, the FC in tons per day will be introduced to the dataset as the fourth feature along the same data clusters to investigate the energy efficiency and FC of each cluster. It should be mentioned that the data points where the FC is not recorded are separated from the dataset for energy-efficiency evaluation to improve the accuracy of the respective calculations. In the resulting 4D dataset, an SVD analysis is also performed to find the correlation between features and the dominant singular directions in each cluster. The results of the 3D clustering using the GMM-EM approach are presented in the next section.

3.5. *Data clustering*

In this section, the results of the clustering algorithm are presented. As mentioned earlier, 7 clusters are considered for this analysis, representing the vessel's operating regions. Since estimating the parameters of the GMM is achieved by maximizing the log-likelihood function, the EM can be considered as an optimization algorithm whose purpose is to maximize the log-likelihood function. Thus, the change in the log-likelihood function's value at each iteration is usually taken as a measure of the convergence of the EM algorithm. The log-likelihood function for each iteration of the EM algorithm is plotted and presented in Fig. 9. It can be seen from this figure that the algorithm has converged before 150 iterations.

The result of the cluster analysis and characteristics of the resulting clusters, i.e., the operating regions are presented in Table 2. As shown in this table, there are 3 major clusters or dominant operating regions of the vessel, in which the vessel was operating for more than 78% of the time. Other clusters correspond to transient regions or when the vessel is starting its journey from a stationary condition and leaving the port.

All the resulting clusters are presented in two 3D scatter plots from different angles in Fig. 10, in which their size and orientation can be observed and compared. In this figure, different clusters are determined by different colors.

Each cluster is also plotted separately in Fig. 11. The main clusters, which correspond to the dominant operating regions of the vessel, can be observed individually in Fig. 11 (a), (b), and (c), respectively. Scatter plots in Fig. 11 (d), (e), and (f) correspond to the regions where the vessel started from a stationary point, i.e., zero SOG, moving towards entering one of the main operating regions. For further reference, these clusters are called acceleration regions. Because at some points in the acceleration regions, the SOG is zero and GPS coordinates are constant, these regions are the time when the vessel is at the port, and the engine is used for feeding other power requirements, e.g., using for auxiliary systems such as generators. As a result, the engine generates power, but the ship's SOG is zero. When the vessel leaves the port, the SOG won't increase suddenly to the cruise speed, and there should be a period with positive acceleration from the stationary point to the cruising speed. Because the acceleration is not that high, the increase in speed is not very sharp. One should note that the two acceleration regions have two different RPMs and EPs. The ES in cluster 7 is higher than the ES in cluster 4. This difference indicates that a different path has been taken for increasing the speed in each accelerating region, which can be due to different gear configurations or loading conditions of the ship propulsion system.

Cluster 6 corresponds to the transient region of the vessel operations. The transient region can be defined as the points that the vessel transits between data clusters, i.e., localized operational modes. In order to

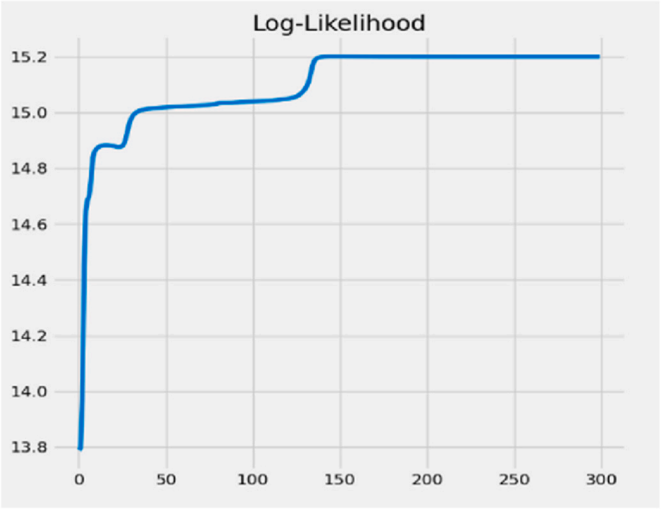


Fig. 9. Log-likelihood function for the 3D feature space.

Table 2
Characteristics of the operating regions of the vessel.

Cluster Number	Operating Hours	Average EP (kW)	Average SOG (kn)	Average ES (RPM)
1	922.40	1555.92	11.20	549.40
2	1272.12	2447.19	13.05	624.52
3	1446.82	1397.22	6.21	719.61
4	252.10	2642.22	12.63	631.30
5	327.60	3624.99	14.85	718.72
6	109.88	635.00	6.34	475.04
7	289.65	2058.89	7.10	608.47

move from one cluster to the other, the vessel should pass through a continuous range, therefore the transient regions have emerged. For this purpose, it must go through points that cannot fit in any other cluster. It is the least used operating region, in which the ship was only operated less than 110 h in the whole year. This cluster is shown in Fig. 11 (g). As it is obvious in this figure, data points of this cluster are distributed over a wide range of EP, ES, and SOG values.

3.6. Singular Value Decomposition (SVD) analysis results

Another observation from the acceleration region scatter plots is that data points are closely located on planes parallel to the SOG-EP plane. As can be seen in Fig. 11 (d), (e), and (f), the difference between the highest ES from the ES mean value is less than 3% for cluster 4 and less than 1% for clusters 7 and 5, which is no more than about 5 RPM. As a result, the vessel's behavior in these regions can be considered as constant speed. This difference can be due to the sensor measurement noise. The intensity of the data points around the ES mean value is higher in these plots. Moreover, the distribution of these points is similar to a Gaussian distribution in the ES direction, which is a logical and common assumption for the measurement noise.

Another observation from this figure is that the distribution of the data points in the ES direction in the acceleration regions shown in Fig. 11 seems to be discrete, although ES is, in nature, a continuous-time signal. The reason for these gaps in the ES values is the discretization error for the measurement and recording of the data, i.e., zero order hold (Ogata, 1995).

In Fig. 12 (a), ES as a time series is plotted for 2 acceleration regions, clusters 7 and 4, to demonstrate the constant speed in these clusters. As shown in this figure, the acceleration regions have a constant ES, indicating that they are located on a 2D surface rather than a 3D surface (see Fig. 12(b)), which means they can be represented in a lower dimension space.

To investigate the shape of the clusters, the FC values are first added to the data points; then an SVD analysis is performed in each cluster's new 4D feature space. The resulting singular vectors, i.e., principal components found by the SVD analysis, singular values, and percentage of information are presented in Table 3. As mentioned earlier, the magnitude of the singular value represents the amount of information in the respective direction. Based on the values presented in this table, more than 95% of the information is stored in the first two singular vectors, which means 95% of the information is located on a plane made by the first two singular vectors. As a result, this data can be compressed using singular vectors without losing any notable information, which means lower memory is needed for storing them.

3.7. Energy efficiency analysis

This section presents the results of energy efficiency assessment of the main clusters, i.e., the main operating regions. For this purpose, both EEIs presented in section 2.5 are calculated for the operating regions of the vessel, and the most energy-efficient operating region is observed.

The results of the EEI_1 calculation for all the data points in all the main operating regions are plotted in Fig. 13. It is evident from this figure that operating region 1 has the best EEI_1 among the main operating regions and operating region 2 has the higher FC per distance.

As mentioned in section 2.5, to consider the effects of vessel cargo, draft value has been selected as an approximate indicator of the vessel load. In order to verify the effect of draft on FC, 2D scatter plots of SOG and the calculated EEI_1 with the draft value color bar for the main operating regions are plotted. These plots are shown in Fig. 14. In these plots, darker regions are associated with higher draft values. As shown in this figure, at any speed, the points with higher draft values have higher EEI_1, which shows the draft has a monotonic relationship with the calculated EEI_1. As a result, EEI_2 can be considered as a more comprehensive index for the energy efficiency evaluation of the vessels.

EEI_2 values for the main operating regions are plotted in Fig. 15. As shown in this figure, the mean value for EEI_2 for operating region 1 is the smallest, which means this operating region has a better energy efficiency among all other main operating regions. Among all the main operating regions, number 2 has the highest EEI_2 value, which makes it the least efficient operating region among the main operating regions.

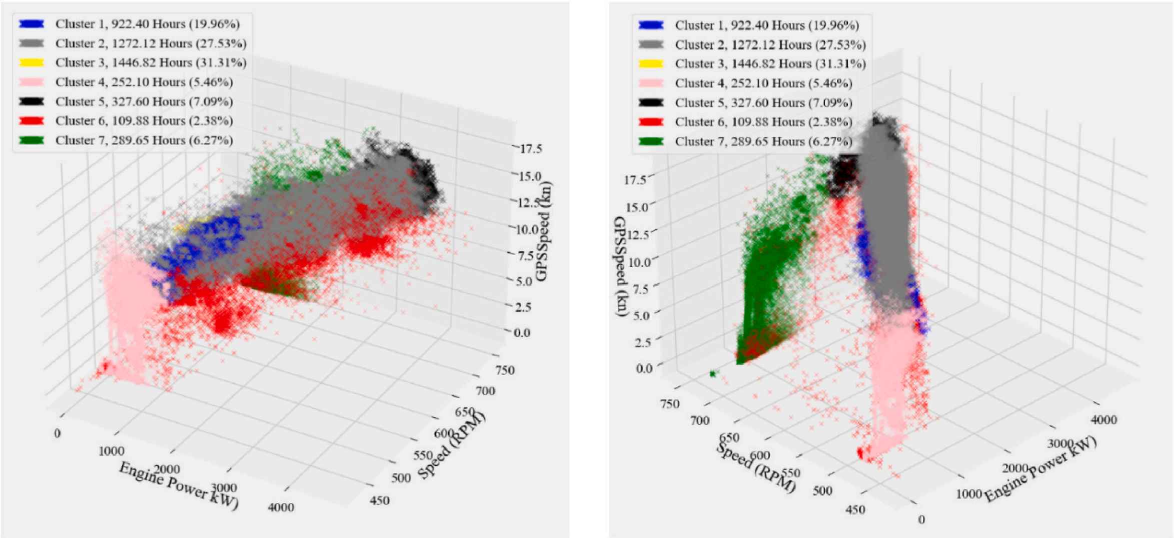
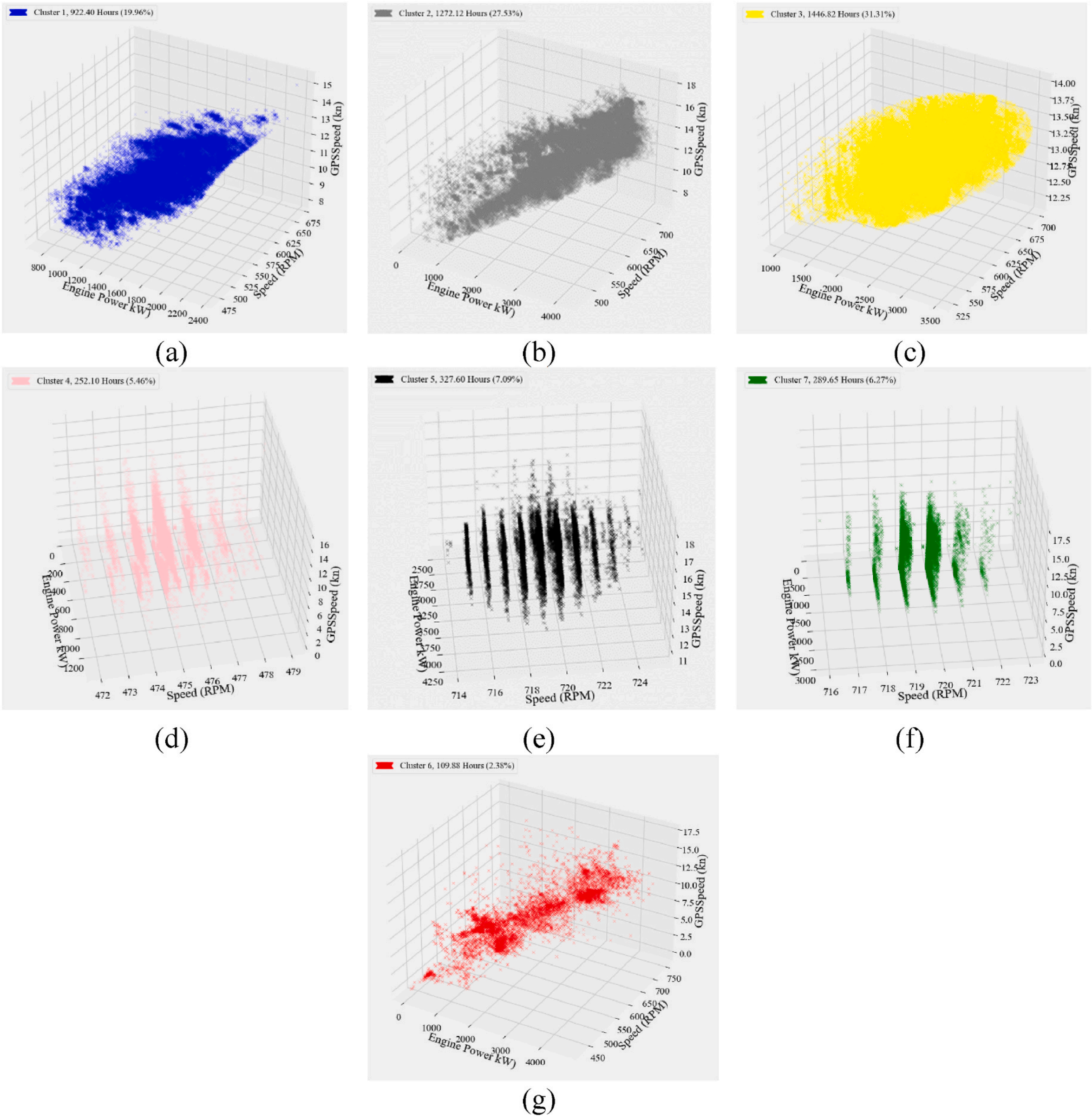


Fig. 10. Scatter plots of all the clusters.

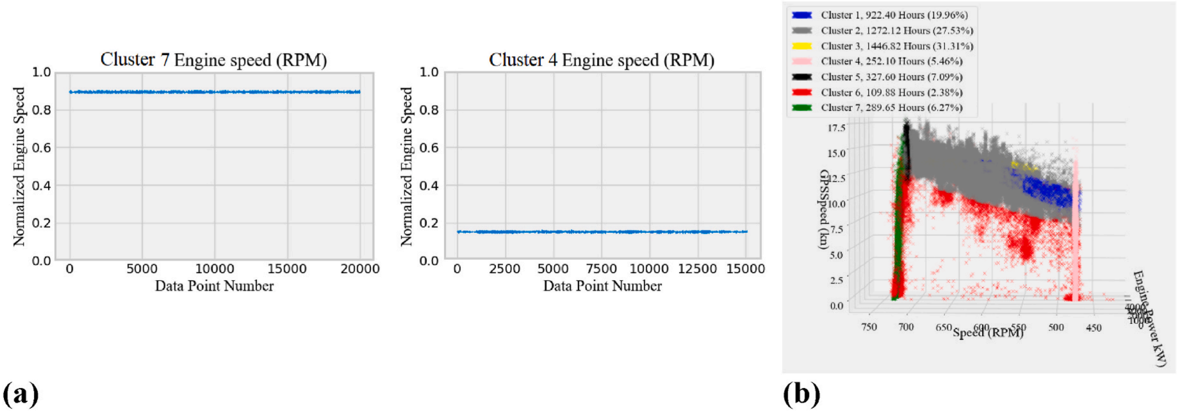


(a) (b) (c)

(d) (e) (f)

(g)

Fig. 11. Scatter plots of different clusters.



(a) **(b)**

Fig. 12. (a) Clusters 4 and 7 ES time series (b) clusters scatter plots.

Table 3
Singular vectors found by SVD for each cluster.

Cluster Number	Singular Vector		Singular Value	Percentage of Information
Cluster 1	V1	[0.996758, 0.080344, 0.004182]	1.401341e-02	89.93
	V2	[0.003702, 0.006120, −0.999974]	1.293317e-03	8.30
	V3	[0.080367, −0.996748, −0.005803]	2.723988e-04	1.75
	V4	[0.080367, −0.996748, −0.005803]	3.095186e-06	0.02
Cluster 2	V1	[9.999934e-01, -3.565913e-03, -5.683130e-04]	4.799843e-02	88.72
	V2	[3.565749e-03, 9.999936e-01, -2.897690e-04]	5.212245e-03	9.63
	V3	[-5.693427e-04, -2.877406e-04, -9.999998e-01]	8.829149e-04	1.63
	V4	[0.080367, −0.996748, −0.005803]	7.768488e-06	0.02
Cluster 3	V1	[9.973184e-01, 7.309133e-02, 3.699815e-03]	1.285398e-02	95.28
	V2	[3.630880e-03, 1.075987e-03, -9.999928e-01]	4.288604e-04	3.18
	V3	[7.309479e-02, -9.973246e-01, -8.077163e-04]	2.026303e-04	1.50
	V4	[0.080367, −0.996748, −0.005803]	5.459539e-06	0.04
Cluster 4	V1	[0.998191, 0.060013, 0.003577]	2.991752e-02	93.43
	V2	[0.003830, −0.004106, −0.999984]	2.094001e-03	6.54
	V3	[0.059997, −0.998189, 0.004328]	2.196780e-06	0.02
	V4	[0.080367, −0.996748, −0.005803]	7.794754e-06	0.01
Cluster 5	V1	[0.996758, 0.080344, 0.004182]	3.922104e-03	65.21
	V2	[0.003702, 0.006120, −0.999974]	2.064051e-03	34.32
	V3	[0.080367, −0.996748, −0.005803]	3.962610e-06	0.41
	V4	[0.080367, −0.996748, −0.005803]	2.483428e-05	0.06
Cluster 6	V1	[0.996758, 0.080344, 0.004182]	9.426789e-02	69.27
	V2	[0.003702, 0.006120, −0.999974]	3.389410e-02	24.91
	V3	[0.080367, −0.996748, −0.005803]	7.866624e-03	5.78
	V4	[0.080367, −0.996748, −0.005803]	5.564750e-05	0.04
Cluster 7	V1	[0.996758, 0.080344, 0.004182]	4.734389e-02	92.93
	V2	[0.003702, 0.006120, −0.999974]	3.589138e-03	7.05
	V3	[0.080367, −0.996748, −0.005803]	4.697735e-06	0.01
	V4	[0.080367, −0.996748, −0.005803]	5.363718e-06	0.01

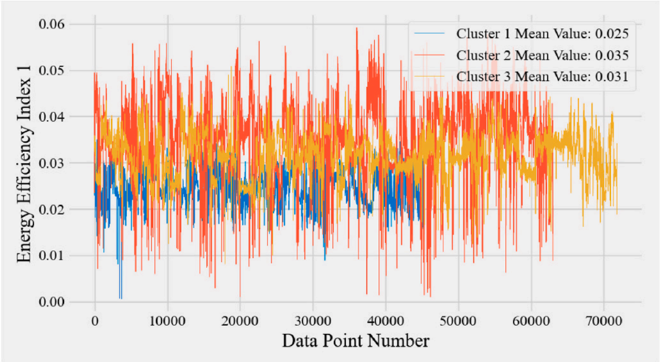


Fig. 13. Results of the EEI_1 based on Eq. (16). for Main Operating Regions.

4. Conclusions

This research study proposes a methodology for developing a framework designed to assess the energy efficiency of a selected vessel under various operating conditions. The developed framework can be generalized and adapted to different vessels, allowing the creation of a digital twin framework for each, given sufficient data recorded from them. However, due to limitations in data availability, this framework is used only for a single vessel in this research to demonstrate the feasibility of developing digital twins capable of incorporating detailed information of the respective vessel.

The main contribution of this research is finding the operating regions of the selected vessel, investigating the FC behavior in different operating regions, which can support the optimal operation and navigation strategy for a selected vessel based on its previous behavior. The findings of this research can help ship navigators to plan their voyages in a more energy-efficient way.

In this research, a cluster analysis is performed using the GMM-EM approach in a 3D feature space with EP in kW, ES in RPM, and SOG in kn to find the operating regions of the vessel. It is also observed that adding more features to the model will not necessarily result in a more accurate model. This is due to the reason that the important parameter correlations are preserved within the limited feature space. This approach captures 7 clusters, i.e., operation regions of the vessel. The engine's most frequent and dominant operating regions and, more importantly, their shapes are detected. Each cluster represents a different navigational and operational condition. Some clusters correspond to main operational regions, and others are associated with transient regions.

Two approaches, namely KDE plots and a separability measure, have been presented in this research for determining the proper number of clusters for the GMM-EM algorithm. These methods can be used for other applications and create a concrete and quantitative basis for selecting the proper number of clusters in a dataset.

There are 3 major clusters, i.e., dominant operating regions of the vessel, in which the vessel was operated for more than 78% of its operating time for the selected vessel. The cluster analysis proposed in this research also builds the initial basis for the digital twin framework, where the knowledge can be used to develop model evaluation conditions. In this framework, the variables are represented as statistical distributions, which are later used in the structure identification step. The novelty of this investigation is that this study considers dataset has complex conditions of data distributions with various clusters or operating regions of a selected vessel.

An SVD analysis is performed to find the dominant singular directions in each cluster and investigate the behavior of each cluster and the correlation between ship performance and navigation variables. The result of the SVD analysis can be utilized in model development and finding the relationship between different variables in each cluster. Based on the SVD analysis, it is observed that each data cluster is located on a lower dimensional feature space. This observation gives the opportunity for data compression without losing any considerable information.

In order to find the most efficient operating region of a selected vessel, an *EEI* called EEI_2 is defined based on the vessel's FC, SOG, and draft values. The most efficient cluster is selected based on the calculated EEI_2 for each cluster. Ship owners can use the approach presented here to find the most optimal configuration for the navigational and

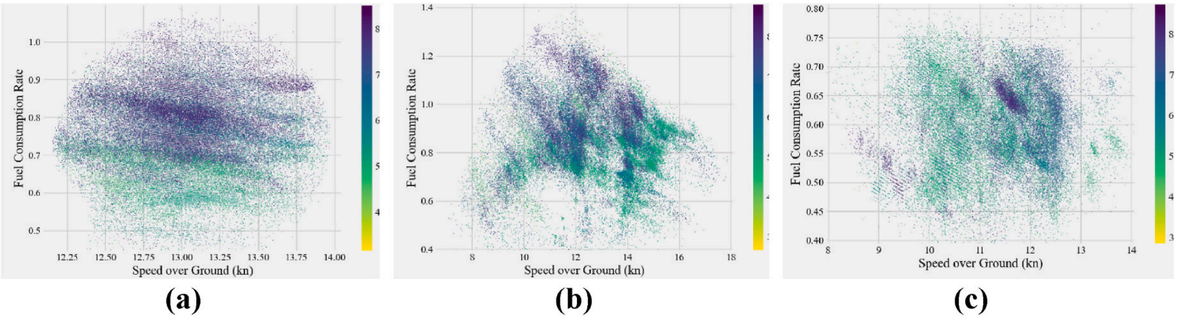


Fig. 14. Scatter Plots of SOG and calculated EEI_1 based on Eq. (16) with draft color bar for (a) Operating Region 1, (b) Operating Region 2, and (c) Operating Region 3.

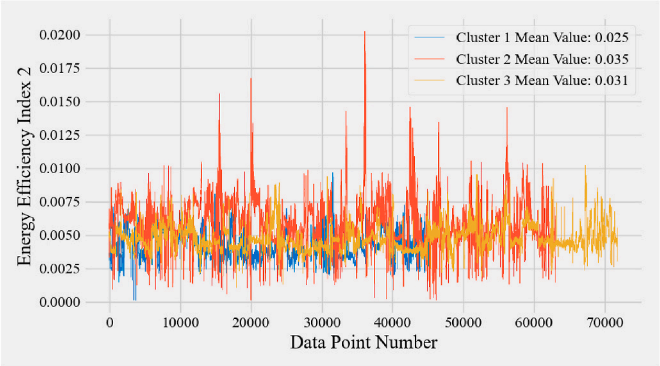


Fig. 15. Results of the EEI_2 based on Eq. (17). for Main clusters.

operational variables of the vessel and make their voyages more energy efficient.

A more accurate analysis of the energy efficiency can be conducted if data on the vessel's cargo load is available. In this research study, however, the draft value is utilized as a proximate measure for cargo amount. Future research could benefit from actual cargo data instead of draft value to enhance the presented EEI_2.

Since the GMM-EM algorithm is an iterative algorithm based on data, an increase in the size of the dataset necessitates more computational time. Nevertheless, given a moderate-quality dataset, an acceptable framework can be developed employing the methodology presented in this research study.

As a subsequent phase of this research study, life cycle cost analysis can be integrated into the developed framework. Moreover, including a predictive dynamic model of the vessel within the framework can form a decision support system for the onshore operation center.

CRediT authorship contribution statement

Mahmood Taghavi: Conceptualization, Formal analysis, Methodology, Software, Validation, Visualization, Writing – original draft. **Lokukaluge P. Perera:** Supervision, Validation, Writing – review & editing.

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.The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:Mahmood Taghavi.

Data availability

The data that has been used is confidential.

Acknowledgements

The project has received funding from the European Union's Horizon 2020 research and innovation program under the grant agreement No 857840 for the SeaTech H2020 project (seatech2020.eu). The opinions expressed in this document reflect only the author's view and in no way reflect the European Commission's opinions. The European Commission is not responsible for any use that may be made of the information it contains.

References

Abebe, M., Shin, Y., Noh, Y., Lee, S., Lee, I., 2020. Machine learning approaches for ship speed prediction towards energy efficient shipping. Appl. Sci. 10 (7), 2325.
Aiello, G., Giallanza, A., Mascarella, G., 2020a. Towards Shipping 4.0. A preliminary gap analysis. Procedia Manuf. 42, 24–29.
Aiello, G., Giallanza, A., Vacante, S., Fasoli, S., Mascarella, G., 2020b. Propulsion monitoring system for digitized ship management: preliminary results from a case study. Procedia Manuf. 42, 16–23.
Ang, J.H., Goh, C., Saldivar, A.A.F., Li, Y., 2017. Energy-efficient through-life smart design, manufacturing and operation of ships in an industry 4.0 environment. Energies 10 (5), 610.
Arrichiello, V., Gualeni, P., 2020. Systems engineering and digital twin: a vision for the future of cruise ships design, production and operations. Int. J. Interact. Des. Manuf. 14, 115–122.
Assani, N., Matić, P., Katalinić, M., 2022. Ship's digital twin—a review of modelling challenges and applications. Appl. Sci. 12 (12), 6039.
Bai, C., Dallasega, P., Orzes, G., Sarkis, J., 2020. Industry 4.0 technologies assessment: a sustainability perspective. Int. J. Prod. Econ. 229, 107776.
Barreiro, J., Zaragoza, S., Diaz-Casas, V., 2022. Review of ship energy efficiency. Ocean Engineering 257, 111594.
Bazari, Z., 2020. MARPOL Annex VI Chapter 4–Energy Efficiency Regulations, National Workshop on Ratification and Implementation of MARPOL Annex VI for Egypt, pp. 3–19.
Bishop, C., 2006. Pattern Recognition and Machine Learning, vol. 2. Springer google schola, pp. 35–42.
Bui, K.Q., Perera, L.P., 2021. Advanced data analytics for ship performance monitoring under localized operational conditions. Ocean Engineering 235, 109392.
Coraddu, A., Oneto, L., Baldi, F., Cipollini, F., Atlar, M., Savio, S., 2019. Data-driven ship digital twin for estimating the speed loss caused by the marine fouling. Ocean Engineering 186, 106063.
Farag, Y.B., Ölçer, A.I., 2020. The development of a ship performance model in varying operating conditions based on ANN and regression techniques. Ocean Engineering 198, 106972.
Gao, Y., Chang, D., Chen, C.-H., 2023. A digital twin-based approach for optimizing operation energy consumption at automated container terminals. J. Clean. Prod. 385, 135782.
Gkerekos, C., Lazakis, I., Theotokatos, G., 2019. Machine learning models for predicting ship main engine Fuel Oil Consumption: a comparative study. Ocean Engineering 188, 106282.
Hameed, Z., Zahia, S., Garcia-Zapirain, B., Javier Aguirre, J., Maria Vanegas, A., 2020. Breast cancer histopathology image classification using an ensemble of deep learning models. Sensors 20 (16), 4373.
He, T., Ye, W., Pan, Q., Liu, X., 2015. An automatic abrupt information extraction method based on singular value decomposition and higher-order statistics. Meas. Sci. Technol. 27 (2), 025007.

Huang, L., Pena, B., Liu, Y., Anderlini, E., 2022. Machine learning in sustainable ship design and operation: a review. Ocean Engineering 266, 112907.

Jamwal, A., Agrawal, R., Sharma, M., Giallanza, A., 2021. Industry 4.0 technologies for manufacturing sustainability: a systematic review and future research directions. Appl. Sci. 11 (12), 5725.

Jimenez, V.J., Kim, H., Munim, Z.H., 2022. A review of ship energy efficiency research and directions towards emission reduction in the maritime industry. J. Clean. Prod. 366, 132888.

Johansen, S.S., Nejad, A.R., 2019. On Digital Twin Condition Monitoring Approach for Drivetrains in Marine Applications, International Conference on Offshore Mechanics and Arctic Engineering. American Society of Mechanical Engineers. V010T009A013.

Kraus, P., Mohrdieck, C., Schwenker, F., 2018. Ship Classification Based on Trajectory Data with Machine-Learning Methods, 2018 19th International Radar Symposium (IRS). IEEE, pp. 1–10.

Lambrou, M., Ota, M., 2017. Shipping 4.0: Technology Stack and Digital Innovation Challenges. IAME 2017 Conference, pp. 1–20.

Lambrou, M., Watanabe, D., Iida, J., 2019. Shipping digitalization management: conceptualization, typology and antecedents. Journal of Shipping and Trade 4 (1), 11.

Lee, J.-H., Nam, Y.-S., Kim, Y., Liu, Y., Lee, J., Yang, H., 2022. Real-time digital twin for ship operation in waves. Ocean Engineering 266, 112867.

Li, G., Hou, Y., Wu, A., 2017. Fourth Industrial Revolution: technological drivers, impacts and coping methods. Chin. Geogr. Sci. 27, 626–637.

Li, H., Liu, T., Wu, X., Chen, Q., 2019. Research on bearing fault feature extraction based on singular value decomposition and optimized frequency band entropy. Mech. Syst. Signal Process. 118, 477–502.

Lo, C., Chen, C., Zhong, R.Y., 2021. A review of digital twin in product design and development. Adv. Eng. Inf. 48, 101297.

Major, P.Y., Li, G., Zhang, H., Hildre, H.P., 2021. Real-time Digital Twin of Research Vessel for Remote Monitoring. Proceedings of 35th European Council for Modelling and Simulation.

Mauro, F., Kana, A., 2023. Digital twin for ship life-cycle: a critical systematic review. Ocean Engineering 269, 113479.

Mohamed, M., 2018. Challenges and benefits of industry 4.0: an overview. Int. J. Supply Oper. Manag. 5 (3), 256–265.

Muhammad, B., Kumar, A., Cianca, E., Lindgren, P., 2018. Improving Port Operations through the Application of Robotics and Automation within the Framework of Shipping 4.0, 2018 21st International Symposium on Wireless Personal Multimedia Communications (WPMC). IEEE, pp. 387–392.

Munim, Z.H., Dushenko, M., Jimenez, V.J., Shakil, M.H., Imset, M., 2020. Big data and artificial intelligence in the maritime industry: a bibliometric review and future research directions. Marit. Pol. Manag. 47 (5), 577–597.

Murray, B., Perera, L.P., 2021. An AIS-based deep learning framework for regional ship behavior prediction. Reliab. Eng. Syst. Saf. 215, 107819.

Namazi, H., Taghavipour, A., 2021. Traffic flow and emissions improvement via vehicle-to-vehicle and vehicle-to-infrastructure communication for an intelligent intersection. Asian J. Control 23 (5), 2328–2342.

Ogata, K., 1995. Discrete-time Control Systems. Prentice-Hall, Inc.

Öztürk, O.B., Başar, E., 2022. Multiple linear regression analysis and artificial neural networks based decision support system for energy efficiency in shipping. Ocean Engineering 243, 110209.

Pang, T.Y., Pelaez Restrepo, J.D., Cheng, C.-T., Yasin, A., Lim, H., Miletic, M., 2021. Developing a digital twin and digital thread framework for an 'Industry 4.0'Shipyard. Appl. Sci. 11 (3), 1097.

Parzen, E., 1962. On estimation of a probability density function and mode. Ann. Math. Stat. 33 (3), 1065–1076.

Peng, Y., Liu, H., Li, X., Huang, J., Wang, W., 2020. Machine learning method for energy consumption prediction of ships in port considering green ports. J. Clean. Prod. 264, 121564.

Perabo, F., Park, D., Zadeh, M.K., Smogeli, Ø., Jamt, L., 2020. Digital Twin Modelling of Ship Power and Propulsion Systems: Application of the Open Simulation Platform (Osp), 2020 IEEE 29th International Symposium on Industrial Electronics (ISIE). IEEE, pp. 1265–1270.

Perera, L.P., 2017. Handling big data in ship performance and navigation monitoring. Smart Ship Technology 89–97.

Perera, L.P., Czachorowski, K., 2019. Decentralized System Intelligence in Data Driven Networks for Shipping Industrial Applications: Digital Models to Blockchain Technologies, OCEANS 2019-Marseille. IEEE, pp. 1–6.

Perera, L.P., Mo, B., 2016. Data Compression of Ship Performance and Navigation Information under Deep Learning, International Conference on Offshore Mechanics and Arctic Engineering. American Society of Mechanical Engineers. V007T006A086.

Perera, L.P., Mo, B., 2017. Machine intelligence based data handling framework for ship energy efficiency. IEEE Trans. Veh. Technol. 66 (10), 8659–8666.

Philbeck, T., Davis, N., 2018. The fourth industrial revolution. J. Int. Aff. 72 (1), 17–22.

Pires, F., Cachada, A., Barbosa, J., Moreira, A.P., Leitão, P., 2019. Digital Twin in Industry 4.0: Technologies, Applications and Challenges, 2019 IEEE 17th International Conference on Industrial Informatics (INDIN). IEEE, pp. 721–726.

Raykov, Y.P., Boukouvalas, A., Baig, F., Little, M.A., 2016. What to do when K-means clustering fails: a simple yet principled alternative algorithm. PLoS One 11 (9), e0162259.

Rødseth, Ø.J., Perera, L.P., Mo, B., 2016. Big Data in Shipping-Challenges and Opportunities.

Sepehri, A., Vandchali, H.R., Siddiqui, A.W., Montewka, J., 2022. The impact of shipping 4.0 on controlling shipping accidents: a systematic literature review. Ocean Engineering 243, 110162.

Shaw, H.-J., Lin, C.-K., 2021. Marine big data analysis of ships for the energy efficiency changes of the hull and maintenance evaluation based on the ISO 19030 standard. Ocean Engineering 232, 108953.

Taghavi, M., Perera, L.P., 2022. Data driven digital twin applications towards green ship operations. In: International Conference on Offshore Mechanics and Arctic Engineering. American Society of Mechanical Engineers. V05AT06A028.

Taskar, B., Andersen, P., 2021. Comparison of added resistance methods using digital twin and full-scale data. Ocean Engineering 229, 108710.

Theodoridis, S., Koutroumbas, K., 2006. Pattern Recognition. Elsevier.

Tran, T.A., 2020. Effect of ship loading on marine diesel engine fuel consumption for bulk carriers based on the fuzzy clustering method. Ocean Engineering 207, 107383.

Uyanık, T., Karatuğ, Ç., Arslanoğlu, Y., 2020. Machine learning approach to ship fuel consumption: a case of container vessel. Transport. Res. Transport Environ. 84, 102389.

Wang, S., Ji, B., Zhao, J., Liu, W., Xu, T., 2018. Predicting ship fuel consumption based on LASSO regression. Transport. Res. Transport Environ. 65, 817–824.

Wang, Y., Perera, L.P., Batalden, B.-M., 2022. The comparison of two kinematic motion models for autonomous shipping maneuvers. In: International Conference on Offshore Mechanics and Arctic Engineering. American Society of Mechanical Engineers. V05AT06A031.

Wang, Z., Xu, H., Xia, L., Zou, Z., Soares, C.G., 2020. Kernel-based support vector regression for nonparametric modeling of ship maneuvering motion. Ocean Engineering 216, 107994.

Xiao, Z., Fu, X., Zhang, L., Goh, R.S.M., 2019. Traffic pattern mining and forecasting technologies in maritime traffic service networks: a comprehensive survey. IEEE Trans. Intell. Transport. Syst. 21 (5), 1796–1825.

Xu, M., David, J.M., Kim, S.H., 2018. The fourth industrial revolution: opportunities and challenges. Int. J. Financ. Res. 9 (2), 90–95.

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 Engineering 169, 457–468.

Yuan, Y., Li, Z., Malekian, R., Yan, X., 2017. Analysis of the operational ship energy efficiency considering navigation environmental impacts. Journal of Marine Engineering & Technology 16 (3), 150–159.

Zhang, C., Zhang, D., Zhang, M., Mao, W., 2019. Data-driven ship energy efficiency analysis and optimization model for route planning in ice-covered Arctic waters. Ocean Engineering 186, 106071.

Zhang, J., Wang, F.-Y., Wang, K., Lin, W.-H., Xu, X., Chen, C., 2011. Data-driven intelligent transportation systems: a survey. IEEE Trans. Intell. Transport. Syst. 12 (4), 1624–1639.

Zhou, C., Xu, J., Miller-Hooks, E., Zhou, W., Chen, C.-H., Lee, L.H., Chew, E.P., Li, H., 2021. Analytics with digital-twinning: a decision support system for maintaining a resilient port. Decis. Support Syst. 143, 113496.