Singular Value Decomposition-based Multiple Model Approach towards Developing Digital Twin Applications in Ship Performance Prediction

Mahmood Taghavi
UiT The Arctic University of Norway
Tromsø, Norway

Lokukaluge Prasad Perera
UiT The Arctic University of Norway, Tromsø, Norway
SINTEF Digital, SINTEF AS, Oslo, Norway

ABSTRACT

Representing the entire operational range of an ocean-going vessel with linear equations is often a formidable task. In this research study, a data-driven localized model is presented for ship performance prediction as a part of the digital twin development. For this purpose, different operational conditions of the vessel, i.e., data clusters, are identified using the Gaussian Mixture Models (GMM) coupled with the Expectation Maximization (EM) algorithm. Subsequently, Singular Value Decomposition (SVD) as a part of the Eigensystem Realization Algorithm (ERA) is applied to each cluster to establish the relationships between different operational and navigational variables and capture the system dynamics in localized operational conditions in each cluster.

KEY WORDS: Vessel Performance Prediction, Control-Oriented Model, Data-Driven Model, Digital Twin

INTRODUCTION

In the preceding years, the application of numerical modeling and computational simulations has become more prevalent, revolutionizing various industries towards more data-driven predictive approaches. This paradigm shift, is gaining momentum increasingly in diverse fields of research, ranging from engineering (Taghavi et al., 2023; Bhuvela et al., 2023; Namazi and Taghavipour, 2021; Alvandifar et al., 2021) to healthcare (Shoaib and Ramamohan, 2022; Bagherian et al., 2020) and environmental sciences (Xiu et al., 2020; Hardesty, 2017; Chen, 2020).

The maritime industry, with complex operational dynamics in its various sectors and its ever-increasing reliance on efficiency and precision, is no exception to this trend. Simulations and numerical modeling, when used for detailed analysis of the behavior of the various systems of ocean-going vessels, can play a crucial role in the design and optimization phase (Barone et al., 2023), safety and risk analysis (Chang et al., 2021), operational efficiency improvement (Barone et al., 2023), predictive maintenance scheduling (Liu et al., 2022; Makridis et al., 2020), autonomous vessels development (Wang et l., 2022; Hasan et al., 2023), and economic analysis.

Moreover, since shipping is the fundamental mode of international trade, lower operational costs in this industry can lead to reduced freight rates imposed on the overall supply chain. As a result, reducing the operational costs associated with this industry can have a profound effect on the global economy. As a result, many research topics have been investigating the economic aspect of the shipping industry (Akbar et al., 2021). Of all the operational costs of shipping, fuel consumption accounts for approximately 45-50% (Rodrigue 2020), a reduction of which has the potential to yield substantial economic benefits. Hence, improving fuel efficiency can be attractive for all ship owners, attracting more interest among the researchers. As an example, Taghavifar and Perera (2023) assess the lifecycle emissions and costs associated with using Liquefied Natural Gas (LNG) as an alternative fuel for ocean-going diesel-operated ships. A reliable model that simulates a ship's behavior is crucial for enhancing fuel efficiency. Such a model facilitates optimization and predictive analysis, enabling the testing of various potential scenarios to assess their effectiveness. By identifying key factors contributing to fuel consumption using the developed model, fuel consumption across diverse operational scenarios can be optimized.

Furthermore, due to the high rate of energy consumption within the maritime sector and its significant contribution to global emissions, the International Maritime Organization (IMO) has devised strict regulations to substantially reduce pollutants caused by vessels. These regulations include the Energy Efficiency Design Index (EEDI), the Ship Energy Efficiency Management Plan (SEEMP), the calculation of the Energy Efficiency Existing Ship Index (EEXI) for all ships, Carbon Intensity Index (CII), and the monitoring of the Energy Efficiency Operational Indicator (EEOI) (Bazari, 2020). Complying with these regulations necessitates the adoption of advanced technological innovations, such as machine learning (ML)-based approaches and

Digital Twin (DT)-type applications (Norwegian Shipowners' Association, 2021). As a result, the development and application of DT technology can be regarded as a pivotal element within this industry. While it is essential to investigate the trustworthiness characteristics of DT applications in the maritime industry, such as explainability, fairness, and accountability (Namazi and Perera, 2023), DT can offer a framework for predicting vessel behavior (Taghavi and Perera, 2023) and enhancing decision-making processes (West et al. 2021).

The entire operating range of an ocean-going vessel exhibits nonlinear behavior, presenting the challenging task of finding a linear model capable of capturing the vessel's dynamics across its entire operating range. Conversely, a set of highly nonlinear equations that accounts for all aspects of the vessel's performance may result in an unsatisfactory generalization of the resulting model. Furthermore, the resultant nonlinear model may not be suitable for optimization and model predictive control applications. To address these challenges, the use of localized models based on frequent operational conditions can be considered a viable solution. This approach enables the approximation of the vessel's nonlinear behaviors while achieving improved generalization and reduced required computational resources. The simplicity inherent in each localized model can facilitate a more straightforward interpretation and analysis of the System of Systems (SoS) model's outcomes.

In this research study, a localized control-oriented model is presented for ship performance prediction using a dataset of a selected vessel. This model can simulate the dynamic behavior and operation of a vessel, which serves as a part of the development of DT models for the shipping industry and that can be utilized to reduce the respective emissions. The presented framework is based on the idea that, within frequently encountered operating conditions, the selected vessel's behavior can be approximated with linear system dynamics. Analysis of the Singular Values (SV) within each operating region, as presented in detail in (Taghavi and Perera, 2022), reveals a dominant singular direction in each operating region, confirming that a linear equation can effectively approximate the relationship between different variables while containing most of the information. For this purpose, in the first step, operating conditions of the vessel are captured by employing cluster analysis. In this framework, each cluster is considered as an operating region of the vessel. The Gaussian Mixture Models (GMM) approach is implemented for capturing the data clusters, followed by the Expectation Maximization (EM) algorithm to calculate the parameters of the clusters, namely, the respective mean and covariance values. Utilizing this approach, the most frequent operating regions of the vessel and their shapes are detected. In this framework the Engine Speed (ES) in RPM and Main Engine Power (EP) in kW are considered as the inputs of the model and the Fuel Consumption (FC) in tons per day is considered as the output. In the next step, the Eigensystem Realization Algorithm (ERA) is used coupled with Observer Kalman Filter Identification (OKID) for capturing the dynamics of the vessel's behavior in the main operating region. During this phase, a Singular Value Decomposition (SVD) analysis is also performed to approximate the linear dynamics describing the system. The final output of this framework is a discrete linear state-space model tailored to serve as a foundation of DT development. Moreover, the proposed reduced-order model can be utilized for control purposes, where rapid and reliable predictions are paramount.

MATERIALS AND METHODS

In this section methods and steps taken for developing the proposed framework is discussed. In 2-1, a summary of the dataset is presented. In 2-2, the concept of GMM-EMM algorithm is briefly discussed. In 2-3, the ERA-OKID algorithm is presented. It's important to recognize that in the clustering step, the main engine-related variables are represented as statistical distributions to identify the respective operational modes of the main engine. However, in the ERA algorithm variables are considered as time series data to capture the dynamical nature and behavior of the system.

Dataset Summary

The proposed framework is developed using data from a selected vessel for one month. The data sampling time is 1 minute, and the specifications of the selected vessel are presented in Table 1.

Table 1. Ship Specifications

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 5.5 (m) Diameter and 4 Blades.

GMM-EM Basic Concepts and Ideas

This research is built upon a prior study (Taghavi and Perera, 2022), where the clustering step with GMM-EM was thoroughly explained. Consequently, in this current study, the discussion on the GMM-EM concept is kept concise.

GMM can be considered as a probabilistic clustering algorithm (Theodoridis and Koutroumbas, 1999). In this method, it is assumed that the distribution of the data points can be formed by combining J distinct multivariate Gaussian distributions, $f\left(x^q;\hat{\theta}(t)\right)$. Consequently, the overall distribution of a random variable x in the dataset or Mixture Density Model (MDM) denoted as h, can be written as Eq. 1. The general parameter vector is represented by Θ, which contains two sets of parameters, θ, and P. θ_j comprises the mean vector and covariance matrix of the j^{th} cluster, μ_j and Σ_j, and P_j represents the posterior probability of cluster j. Each data point, denoted as x^q, is assumed to belong to the Gaussian distribution with the highest probability, representing the cluster to which that data point is assigned.

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

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

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

The next step involves the estimation of model parameters. This parameter estimation can be accomplished utilizing the EM algorithm. To formulate the EM algorithm, a new set of variables denoted as y is defined, which comprises an observed part, *x*, and an unobserved part, *j*. The vector *x* represents the parameters obtained from the measurements. In order to estimate the unknown parameters, the log-likelihood function, $L(\theta)$, is defined as Eq. 2, which will be maximized using the EM algorithm. In this equation, *M* is the total number of data points within the data set.

$$L(\theta) = \sum_{q=1}^{M} [lnf(y^q; \theta | x^q)] \tag{2}$$

The EM algorithm is an iterative process consisting of two main steps. In the initial step known as the E-Step, the expectation of the log-likelihood function is computed. In the second step, referred to as the M-Step or the maximization step, the derivative of the expectation of the log-likelihood function with respect to Σ_j, μ_j, and P_j are calculated and set to zero. Eq. 3 illustrates the iterative nature of the EM algorithm. In this process, the values of Σ_j, μ_j, and P_j in each iteration are calculated based on the previous iteration values (Theodoridis and Koutroumbas, 1999).

$$\begin{aligned}
P\left(j; \widehat{\Theta}(t) | x^q\right) &= \frac{f\left(x^q; \hat{\theta}(t) | j\right) \hat{P}_j(t)}{\sum_{i=1}^{J} f\left(x^q; \hat{\theta}(t) | i\right) \hat{P}_i(t)} \\
\hat{\mu}_i(t+1) &= \frac{\sum_{q=1}^{M} P\left(i; \widehat{\Theta}(t) | x^q\right) x^q}{\sum_{q=1}^{M} P\left(i; \widehat{\Theta}(t) | x^q\right)} \\
\hat{\Sigma}_i(t+1) &= \frac{\sum_{q=1}^{M} \left[P\left(i; \widehat{\Theta}(t) | x^q\right)\left(x^q - \hat{\mu}_i(t+1)\right)\left(x^q - \hat{\mu}_i(t+1)\right)^T\right]}{\sum_{q=1}^{M} P\left(i; \widehat{\Theta}(t) | x^q\right)} \\
\hat{P}_i(t+1) &= \frac{1}{M} \sum_{q=1}^{M} P\left(i; \widehat{\Theta}(t) | x^q\right)
\end{aligned} \tag{3}$$

The initial values for the parameters of this algorithm are randomly chosen. This iterative process continues until the GMM parameters converge to stable values. In this particular research study, the algorithm is considered to be converged when the parameter values exhibit less than a 2% change over two consecutive steps. In this research, the cluster analysis is performed using the following parameters:

- Main Engine Power (kW) (EP)
- Fuel Consumption Rate (Tons per Day) (FC)
- Engine Speed (RPM) (ES)

In the following section, steps taken to implement the ERA algorithm are presented.

Eigensystem Realization Algorithm (ERA)

ERA is a system identification technique used to determine a low-dimensional linear state-space representation of a dynamical system (Brunton and Kutz, 2022). This algorithm estimates the state-space matrices (*A*, *B*, *C*, and *D*) of a linear time-invariant system from sensor measurements of an impulse response experiment, based on the "minimal realization" theory of Ho and Kalman (Ho and Kálmán, 1966). ERA is based entirely on impulse response measurements and does not require prior knowledge of a model.

In the context of this study, the system dynamics are characterized by the discrete linear time-invariant state-space representation, as expressed in Eq. 4. Generally, this method can be utilized for a Multi Input Multi Output (MIMO) system, and u_k, and output, y_k, can be both vectors.

$$\begin{aligned}
x_{k+1} &= Ax_k + Bu_k \\
y_k &= Cx_k + Du_k
\end{aligned} \tag{4}$$

Where $x_k \epsilon \mathcal{R}^n$, $u_k \epsilon \mathcal{R}^p$, and $y_k \epsilon \mathcal{R}^q$.

In ERA, the input is assumed to be discrete-time delta function as described in Eq. 5.

$$u_k^\delta = \begin{cases} I, & k = 0 \\ o, & k = 1, 2, 3, \ldots \end{cases} \tag{5}$$

As a result, the response y_k will be a discrete-time impulse response as described in Eq. 6.

$$y_k^\delta = \begin{cases} D, & k = 0 \\ CA^{k-1}B, & k = 1, 2, 3, \ldots \end{cases} \tag{6}$$

In general, it is assumed that *q* impulse responses are performed, one for each of the separate input channels. The output responses are collected for each impulsive input, and at a given time-step *k*, the output vector in response to the *j*th impulsive input will form the *j*th column of y_k^δ. As a result, at each time step *k*, y_k^δ is a $p \times q$ matrix equal to $CA^{k-1}B$ (Brunton and Kutz, 2022). The next step is to form the Hankel matrix *H*, presented in Eq. 7, by stacking the shifted time-series of impulse-responses into a single matrix. It is worth mentioning that the matrix *H* is constructed purely from measurements.

$$H = \begin{bmatrix} y_1^\delta & y_2^\delta & y_3^\delta & \cdots \\ y_2^\delta & y_3^\delta & y_4^\delta & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} = \begin{bmatrix} CB & CAB & CA^2B & \cdots \\ CAB & CA^2B & CA^3B & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} = \mathcal{O}\mathcal{C} \tag{7}$$

Where:

$$\mathcal{C} = \begin{bmatrix} C \\ CA \\ CA^2 \\ \vdots \end{bmatrix}, \qquad \mathcal{O} = \begin{bmatrix} B & AB & A^2B & \ldots \end{bmatrix} \tag{8}$$

Taking the SVD of the Hankel matrix, Eq. 9, and computing the singular values and singular vectors, yields the dominant temporal patterns in the time-series data.

$$H = \mathcal{O}\mathcal{C} = U\Sigma V^* = \begin{bmatrix} \widetilde{U} & U_t \end{bmatrix} \begin{bmatrix} \widetilde{\Sigma} & 0 \\ 0 & \Sigma_t \end{bmatrix} \begin{bmatrix} \widetilde{V}^* \\ V_t^* \end{bmatrix} \approx \widetilde{U}\widetilde{\Sigma}\widetilde{U}^* = \left(\widetilde{U}\widetilde{\Sigma}^{\frac{1}{2}}\right)\left(\widetilde{\Sigma}^{\frac{1}{2}}\widetilde{V}^*\right) \tag{9}$$

Where * denotes the complex conjugate transpose.

The singular values are ordered based on their value in Σ from top to buttom. The smaller singular values are stored in Σ_t. To get a reduced order rank *r* model that captures most of the system dynamics Σ_t, U_t and

V_t^* are truncated, and the Henkel matrix is approximated by $\tilde{U}\tilde{\Sigma}\tilde{U}^*$. The value of *r* is selected by evaluating the of the amount singular values in a way that smaller singular values and their corresponding rows and columns in U and V^* are truncated. This decision can be made by plotting and observing the rate of decay of singular values (Brunton and Kutz, 2022). Based on the Henkel matrix, another matrix, H', can be formed by shifting each element of the Henkel matrix one time step in the future. Based on the definitions in Eq. 8, this matrix can be denoted as $\mathcal{O}A\mathcal{C}$.

$$H' = \begin{bmatrix} y_2^\delta & y_3^\delta & y_4^\delta & \cdots \\ y_3^\delta & y_4^\delta & y_5^\delta & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} = \begin{bmatrix} CAB & CA^2B & CA^3B & \cdots \\ CA^2B & CA^3B & CA^4B & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} = \mathcal{O}A\mathcal{C} \tag{10}$$

By comparing the matrices H and H', the reduced-order model can be constructed as presented in Eq. 11.

$$\begin{aligned} \tilde{A} &= \tilde{\Sigma}^{-\frac{1}{2}}\tilde{U}^* H' \tilde{V}\tilde{\Sigma}^{-\frac{1}{2}} \\ \tilde{B} &= \tilde{\Sigma}^{-\frac{1}{2}}\tilde{V}^* \begin{bmatrix} I_p & 0 \\ 0 & 0 \end{bmatrix} \\ \tilde{C} &= \begin{bmatrix} I_q & 0 \\ 0 & 0 \end{bmatrix} \tilde{U}\tilde{\Sigma}^{\frac{1}{2}} \end{aligned} \tag{11}$$

Where I_p is the $p \times p$ identity matrix, which extracts the first p columns of $\tilde{U}\tilde{\Sigma}^{\frac{1}{2}}$, and I_q is the $q \times q$ identity matrix, which extracts the first q rows of $\tilde{U}\tilde{\Sigma}^{\frac{1}{2}}$. Thus, the input–output dynamics can be expressed in terms of a reduced system with a low-dimensional state $\tilde{x} \epsilon \mathcal{R}^r$.

$$\begin{aligned} \tilde{x}_{k+1} &= \tilde{A}\tilde{x}_k + \tilde{B}u_k \\ y_k &= \tilde{C}\tilde{x}_k \end{aligned} \tag{12}$$

It's important to recognize that matrices *A*, *B*, and *C* are not known in prior, and the goal of this algorithm is to approximate these matrices in a way that best describe the model dynamics. This approach is purely data driven and there is no information about the system dynamics, the number of system states (*n* is unknown), and the states themselves. As a result, there is no measurement of these states in different time steps in advance, and the only measurements present are the inputs and outputs of the system. Consequently, the final model will be a reduced order model of rank *r*. In the resulted model there will be some latent variables that are necessary to describe the input-output dynamics. In order to have a physical interpretation of these states, a full-state measurement is needed to be performed.

ERA is fundamentally based on the assumption that it is possible to measure an impulse response from the system. However, in practice, applying an isolated impulse response is a challenging task for many complex systems. Moreover, the effect of measurement noise can degrade the measurement results. To address this challenges, Observer Kalman Filter Identification (OKID) technique has been developed. Utilizing the OKID technique, it becomes feasible to estimate the optimal impulse response that is the most consistent with the non-impulse input and output data originating from an existing system of interest. In this case, the input *u* is not an impulse, and at each discrete time step, *u* has a distinct value, representing the input signal's dynamic nature. Based on this, state and output values at each time step can be written as Table 2. Based on the calculated values for each time step, the respective impulse response can be calculated as Eq. 13.

Table 2. Input, state, and output values at each time step

Time Step	*u*	*x*
0	u_0	0
1	u_1	Bu_0
2	u_2	$ABu_0 + Bu_1$
3	u_3	$A^2Bu_0 + ABu_1 + Bu_2$
⋮	⋮	⋮

$$\begin{aligned} [y_0 \quad y_1 \quad y_2 \quad \cdots] &= [D \quad CB \quad CAB \quad \cdots] \begin{bmatrix} u_0 & u_1 & u_2 & \cdots \\ 0 & u_0 & u_1 & \cdots \\ 0 & 0 & u_0 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \\ &= [y_0^\delta \quad y_1^\delta \quad y_2^\delta \quad \cdots] \begin{bmatrix} u_0 & u_1 & u_2 & \cdots \\ 0 & u_0 & u_1 & \cdots \\ 0 & 0 & u_0 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \end{aligned} \tag{13}$$

$$\begin{aligned} Impulse\ Response &= [y_0^\delta \quad y_1^\delta \quad y_2^\delta \quad \cdots] \\ &= [y_0 \quad y_1 \quad y_2 \quad \cdots] \left(\begin{bmatrix} u_0 & u_1 & u_2 & \cdots \\ 0 & u_0 & u_1 & \cdots \\ 0 & 0 & u_0 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \right)^{-1} \end{aligned}$$

Now that the respective impulse response of the system has been determined, ERA algorithm can be employed to estimate the *A*, *B*, and *C* matrices that best capture the dynamics of the system.

A time series of inputs and outputs data should be utilized for the implementation of the ERA algorithm.

For this purpose, a time series consisting of 500 instances from the main operating region of the vessel has been selected as the training set. Additionally, another time series containing 400 instances, also from the same operating region, has been selected as the test set. In this research study, the EP has been selected as the measured input, while the FC is chosen as the measured output of the system. Under the framework of ERA, it is assumed that there is no information regarding other system states, and a reduced order model is solely developed based on the available input-output data.

DATA ANALYSIS RESULTS AND DISCUSSION

In this section, the results at each phase of the data analysis are discussed. In the first section, the results of the preprocessing step and their effect on the quality of data is presented. After that, the resulting clusters, which represent the distinct operating regions of the vessel, are presented. Finally, the outcomes derived from the implementing of the ERA-OKID algorithm are comprehensively discussed.

Preprocessing

The preprocessing phase is a crucial step prior to any ML-based data-driven approach, and the performance of the developed model is highly dependent on the quality of the data used. In this section, results for the preprocessing step are presented.

Time series for EP, ES, and FC values for the entire month are plotted in

Fig. 1. This figure clearly shows numerous instances of missing data in which the engine was running, but the FC is missing. Additionally, certain data sets exhibit irregularities, such as abnormally high FC values during normal engine operation. In the current research, all data points with missing instances are excluded from the data set in the preprocessing phase. However, in the future, the values for FC for these data points can be recovered using the proposed framework. For the purpose of validating the effectiveness of the proposed framework, this research exclusively utilizes data sets cleaned from anomalies and missing values.

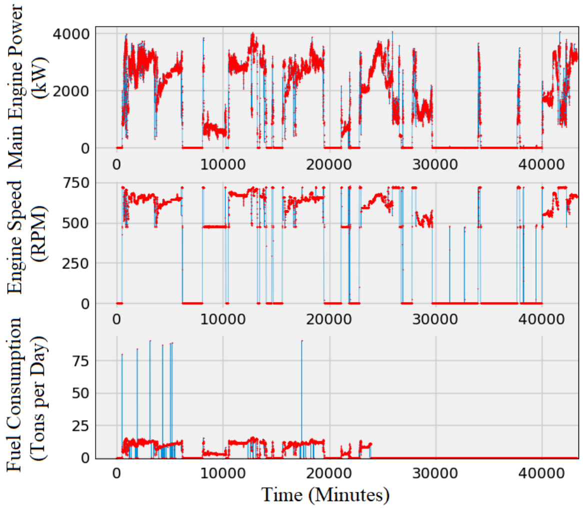


Fig. 1. Time Series Plots for EP (Kw), ES (RPM), and FC (Tons Per Day)

Fig. 2 shows the histograms for the main engine power after removing the missing data based on the approach explained. The data vitalization has improved by removing the missing data and data outliers in this figure. It should be mentioned that the regions where the engine is not running should be distinguished from the outliers. In line with the research objective of identifying the engine's operating regions, these regions where the engine was inactive are removed alongside with the outliers.

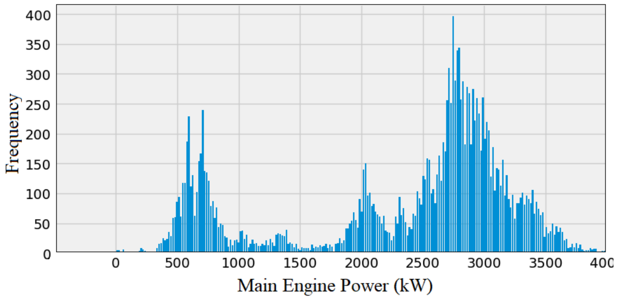


Fig. 2. Histograms For the EP (Kw) After Removing the Missing Data

Clustering Results

The comprehensive implementation and results of the GMM-EM algorithm on the dataset from the selected vessel are presented in detail in a previous work (Taghavi and Perera, 2022). It's crucial to note that in the clustering algorithm, all data points are normalized. This step ensures that the algorithm remains unbiased and does not disproportionately favor features with higher values. From these findings, four distinct clusters or operational regions were identified for the vessel, using the selected operational variables of the marine engine. The resulting mean values for each cluster for the selected feature space are presented as follows. Within all the mean value vectors, the first row corresponds to the mean value of EP, the second row indicates the mean value of FC, and the third row represents the mean value of ES.

$$\mu_1 = \begin{bmatrix} 1760.14 \\ 7.65 \\ 719.62 \end{bmatrix}, \quad \mu_2 = \begin{bmatrix} 642.52 \\ 3.03 \\ 475.01 \end{bmatrix},$$

$$\mu_3 = \begin{bmatrix} 2702.72 \\ 10.73 \\ 614.11 \end{bmatrix}, \quad \mu_4 = \begin{bmatrix} 2993.81 \\ 11.87 \\ 662.63 \end{bmatrix}$$

Fig. 3 displays all data points in a 3D space, with each cluster distinguished by a unique color. The percentage of data points corresponding to each cluster is calculated and presented in Table 3. This table indicates that the dominant cluster or operating region is the third one, encompassing approximately 56% of the data points. This implies that the vessel was operating for nearly 56% of the observed period in this data cluster or operating region.

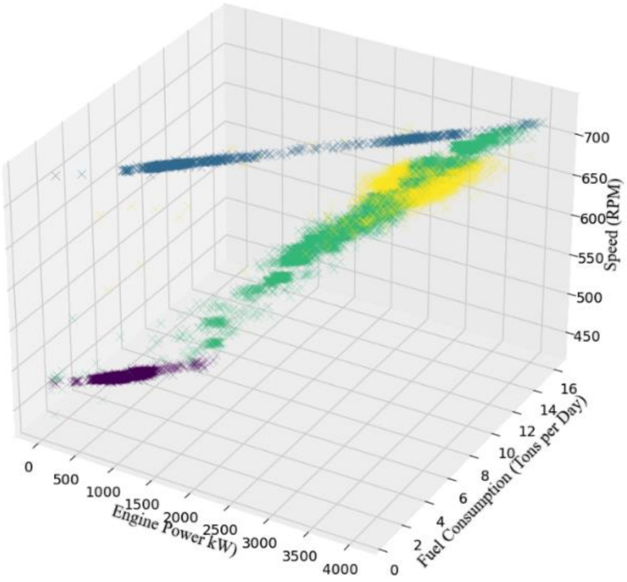


Fig. 3. Final Cluster Configuration (Taghavi and Perera, 2022)

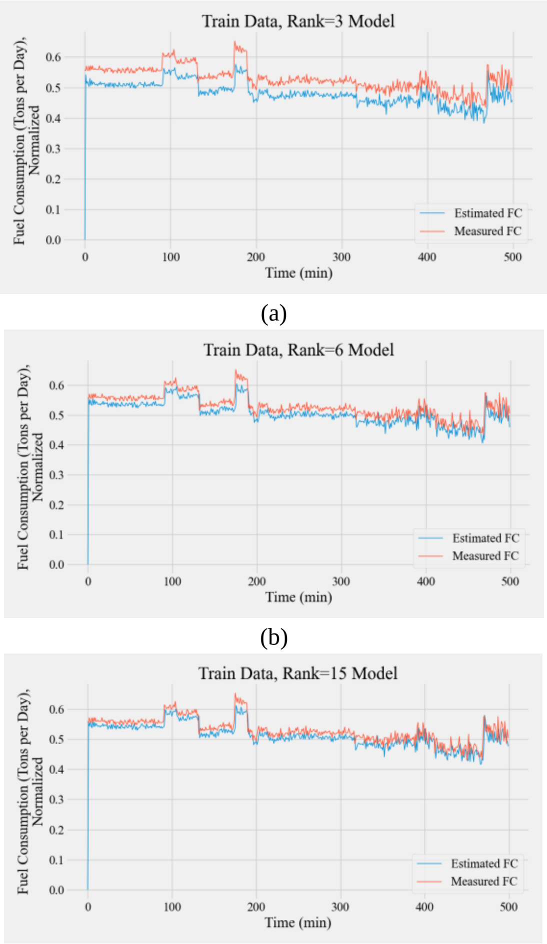
Table 3. Percentages of Data Points Belong to Each Cluster

Number of Clusters	Data point density (%)
1	6.45
2	17.03
3	55.63
4	20.89

ERA-OKID Results

In this section, the results of employing the OKID-ERA algorithm are presented. In this research study, to implement ERA, as previously mentioned, one input, namely the EP, and one output, the FC, has been selected. The algorithm is applied to a training set comprising 500 discrete time steps, resulting in the derivation of reduced-order approximations for matrices A, B, and C. The resulted matrices are of rank *r*, which is selected based on the singular values of the Henkel matrix. The selected rank represents the number of hidden or latent states considered for the model. The ERA algorithm has been executed for different values for the rank of the model in order to evaluate its impact on the results.

The outcome of the steps executed in the OKID-ERA algorithm for different rank values are illustrated in Fig. 4. and Fig. 5 for train and test data respectively. As depicted in this figure, the developed model exhibits a better performance in estimating the FC consumption as the rank selected for the model increases. Nevertheless, increasing the rank of the system to higher values doesn’t seem to enhance the performance of the model proportionally. This suggests that a reduced-order model with a rank even less than 10 can effectively capture most of the system dynamics with an acceptable accuracy.

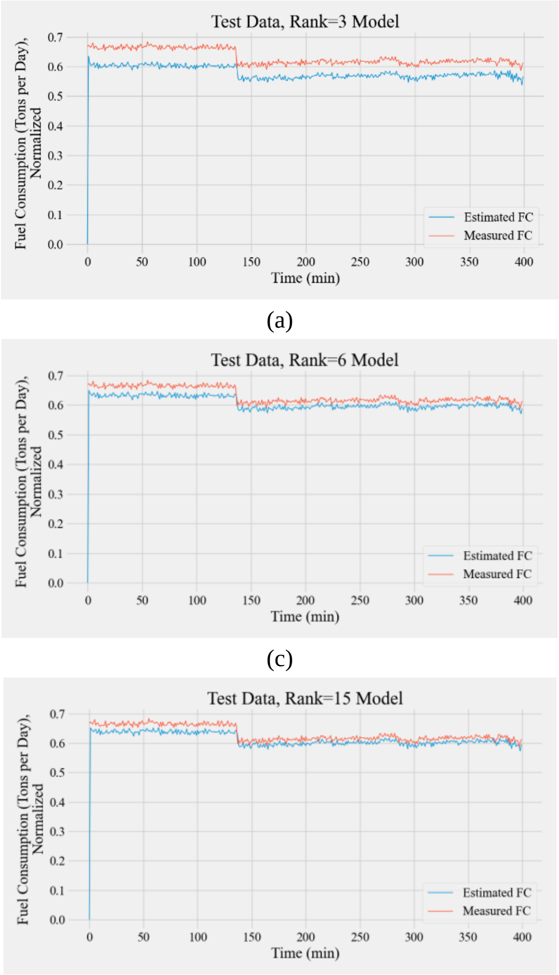


(a)

(b)

(c)

Fig. 4. Measured and Estimated Train FC for a Model of Rank Equal to 3 (a), 6(b), and 15(c)



(a)

(c)

(e)

Fig. 5. Measured and Estimated Test FC for a Model of Rank Equal to 3 (a), 6(b), and 15(c)

It is worth mentioning that the computational cost of ERA depends on several factors, including the size of the Hankel matrix, the SVD operation, and state-space model construction. The size of the Hankel matrix, and thus the computational cost, grows with the amount of data. The cost of SVD is proportional to the dimensions of the Hankel matrix. The selection of significant singular values and vectors, which corresponds to the rank r, impact the cost. For a fixed data size, increasing the rank r -by keeping more singular values and vectors in memory- can slightly increase the computation time, but not as significantly as the costs associated with the matrix operations of the Hankel matrix itself. The rank r is mostly effective in the computational cost related to the state-space model construction and further controller development. In the state-space model, the computational cost is related to the matrix operations involved in constructing the A, B, and C matrices of the state space model. This cost scales with the square or cube of the rank r, depending on the specifics of the matrix operations. The major influence of the rank of the system is on the controller design. Control algorithms, especially those relying on online computation, become less efficient if the model is unnecessarily complex. A system with a higher rank r can lead to increased computational demands for the controller, both in terms of memory usage and processing power, which can be a limiting factor in the implementation of control algorithms in online applications.

The percentage of average errors with respect to the mean value of the FC in the test data for different ranks of the system are also presented in Table 4. Error values in this table indicate as the rank of the system increases, the accuracy of the estimation improves. However, improving the rank of the model to much higher values will not improve the estimation error proportionally.

Table 4- Percentage of Average Error for Different Ranks of the system

Rank	**r = 3**	**r = 6**	**r = 15**	**r = 30**
Average Error (%)	9.333	4.351	3.371	2.44

CONCLUSIONS

This section presents the conclusions and contributions of the proposed framework for data-driven model development of the vessel. Based on the results presented in the previous section, the following points can be concluded:

- ERA coupled with OKID algorithm has been employed to construct a state-space representation of the main operating region of a selected vessel. This predictive model can serve as a part of the development of DT models for the shipping industry and that can be utilized to reduce the respective emissions.
- In the proposed model, EP is designated as the input, and FC as the output of the system.
- To develop this model, in the initial phase, GMM-EM algorithm has been executed to capture the dominant operating region of the vessel. Subsequently, a time series of data is selected from this main operating region for implementing the ERA-OKID algorithm.
- The state-space representation is developed using three different values for the rank of the model, with each rank representing the number of hidden or latent states considered for the model.
- It is observed that by increasing the rank, the performance of the model improves. However, it is important to note that the estimation error improvement is not linearly proportional to the rank increase. Thus, increasing the rank value to considerably higher values does not result in substantial improvements in estimation accuracy.
- Increasing the rank of the system does not significantly affect the computational cost in the state-space model development stage. However, in simpler models, i.e., models with smaller ranks, the controller will demand less computational resources.
- The rank of the model is selected based on the singular values of the Henkel matrix.
- The proposed model has been developed solely based on the recorded data onboard the vessel, without incorporating any assumption or knowledge about the physical behavior or dynamics of the vessel.
- The proposed model has only one input and one output. For a more comprehensive representation, additional inputs, outputs, and states of the system can be introduced to have a more holistic model of the vessel performance.
- In this research, only the main operating region of the engine is considered for model development. The same approach can be applied to other operating regions of the vessel as the next step of this study. Subsequently, the localized models for different operating regions can be combined to construct an overall localized model for ship performance prediction. Despite the overall nonlinearity of the vessel's operating range, linear models can approximate the dynamics of each specific operating region.
- ERA-OKID is proved to be a powerful and effective algorithm for development of state-space representation that can capture most of the input-output dynamics of complex systems based solely on recorded data. This approach can further be implemented under DT type frameworks for other applications, too.
- The presented research here is part of a broader study on vessel performance quantification. For this purpose, different Key Performance Indicators (KPI) have been defined to evaluate the vessel performance, with FC being a critical factor. However, the FC data contains many missing values and anomalies. This paper develops a method to estimate these missing and anomalous FC values, thereby enhancing the reliability of KPIs for performance quantification.

ACKNOWLEDGEMENTS

The project has received funding from the European Union's Horizon 2020 research and innovation program under the grant agreement No 857840 (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.

This work is also supported by the MARKOM II project under project title "digital twin for green ship operations (D2G 2023)" under the contract number PMK-2022-10022.

REFERENCES

Akbar, A, Aasen, A.K, Msakni, M.K, Fagerholt, K, Lindstad, E, and Meisel, F (2021). "An economic analysis of introducing autonomous ships in a short-sea liner shipping network," *International Transactions in Operational Research*, 28(4), 1740-1764.

Alvandifar, N, Saffar-Avval, M, Amani, E, Mehdizadeh, A.R, Ebrahimipour, M.R, Entezari, S, Namazi, H, Esfandiari-Mehni, M, and Ahmadibeni, G (2021). "Experimental study of partially metal foam wrapped tube bundles," *International Journal of Thermal Sciences*, 162, 106798.

Bagherian, A, Baghani, M, George, D, Rémond, Y, Chappard, C, Patlazhan, S, and Baniassadi, M (2020). "A novel numerical model for the prediction of patient-dependent bone density loss in microgravity based on micro-CT images," *Continuum Mechanics and Thermodynamics*, 32, 927-943.

Barone, G, Buonomano, A, Del Papa, G, Forzano, C, Giuzio, G.F, Maka, R, Russo, G, and Vanoli, R (2023). "Improving the Energy Efficiency of Ships: Modelling, Simulation, and Optimization of Cost-effective Technologies," *Modelling and Optimisation of Ship Energy Systems 2023*.

Barone, G, Buonomano, A, Forzano, C, and Palombo, A (2021). "Implementing the dynamic simulation approach for the design and optimization of ships energy systems: Methodology and applicability to modern cruise ships," *Renewable and Sustainable Energy Reviews*, 150, 111488.

Bazari, Z (2020). "MARPOL Annex VI Chapter 4–Energy efficiency regulations," *National Workshop on Ratification and Implementation of MARPOL Annex VI for Egypt*, 25, 3-19.

Bhuvela, P, Taghavi, H, and Nasiri, A (2023). "Design Methodology for a Medium Voltage Single Stage LLC Resonant Solar PV Inverter," *12th International Conference on Renewable Energy Research and Applications (ICRERA),* IEEE, 556-562.

Brunton, S.L, and Kutz, J.N (2022). "*Data-driven science and engineering: Machine learning, dynamical systems, and control*," Cambridge University Press.

Chang, C.H, Kontovas, C, Yu, Q, and Yang, Z (2021). "Risk assessment of the operations of maritime autonomous surface ships," *Reliability Engineering & System Safety*, 207, 107324.

Chen, C, He, W, Zhou, H, Xue, Y, and Zhu, M (2020). "A comparative study among machine learning and numerical models for simulating groundwater dynamics in the Heihe River Basin, northwestern China," *Scientific reports*, 10(1), 3904.

Hardesty, B.D, Harari, J, Isobe, A, Lebreton, L, Maximenko, N, Potemra, J, Van Sebille, E, Vethaak, A.D, and Wilcox, C (2017). "Using numerical model simulations to improve the understanding of micro-plastic distribution and pathways in the marine environment," *Frontiers in marine science*, 4, 30.

Hasan, A, Widyotriatmo, A, Fagerhaug, E, and Osen, O (2023). "Predictive digital twins for autonomous ships," *IEEE Conference on Control Technology and Applications (CCTA),* IEEE, 1128-1133.

Ho, B.L, and Kálmán, R.E (1966). "Effective construction of linear state-variable models from input/output functions: Die Konstruktion von linearen Modeilen in der Darstellung durch Zustandsvariable aus den Beziehungen für Ein-und Ausgangsgrößen," *at-Automatisierungstechnik*, 14(1-12), 545-548.

Liu, S, Chen, H, Shang, B, and Papanikolaou, A (2022). "Supporting predictive maintenance of a ship by analysis of onboard measurements," *Journal of Marine Science and Engineering*, 10(2), 215.

Makridis, G, Kyriazis, D, and Plitsos, S (2020). "Predictive maintenance leveraging machine learning for time-series forecasting in the maritime industry," *23rd international conference on intelligent transportation systems (ITSC),* IEEE, 1-8.

Namazi, H, and Perera, L.P (2023). "Trustworthiness Evaluation Framework for Digital Ship Navigators in Bridge Simulator Environments," *International Conference on Offshore Mechanics and Arctic Engineering*, OMAE, American Society of Mechanical Engineers, 86878, V005T06A037.

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

Norwegian Shipowners' Association, (2021). "*Maritime Outlook*".

Rodrigue, J.P, (2020). "*The geography of transport systems. Routledge*".

Shoaib, M, and Ramamohan, V (2022). "Simulation modeling and analysis of primary health center operations," *Simulation*, 98(3), 183-208.

Taghavi, H, El Shafei, A, and Nasiri A (2023). "Liquid Cooling System for a High Power, Medium Frequency, and Medium Voltage Isolated Power Converter," *12th International Conference on Renewable Energy Research and Applications (ICRERA),* IEEE, 405-413.

Taghavi, M, and Perera, L.P (2022). "Data Driven Digital Twin Applications Towards Green Ship Operations," *International Conference on Offshore Mechanics and Arctic Engineering*, OMAE, American Society of Mechanical Engineers, 85895, V05AT06A028.

Taghavi, M, and Perera, L.P (2023). "Multiple Model Adaptive Estimation Coupled with Nonlinear Function Approximation and Gaussian Mixture Models for Predicting Fuel Consumption in Marine Engines," *International Conference on Offshore Mechanics and Arctic Engineering*, OMAE, American Society of Mechanical Engineers, 86878, V005T06A034.

Taghavifar, H, and Perera, L.P (2023). "Life cycle emission and cost assessment for LNG-retrofitted vessels: the risk and sensitivity analyses under fuel property and load variations," *Ocean Engineering*, 282, 114940.

Theodoridis, S, and Koutroumbas, K (1999), "*Pattern recognition and neural networks*," Springer Berlin Heidelberg.

Wang, Y, Perera, L. P, and Batalden, B.M (2022). "The Comparison of Two Kinematic Motion Models for Autonomous Shipping Maneuvers," *International Conference on Offshore Mechanics and Arctic Engineering*, OMAE, American Society of Mechanical Engineers, 85895, V05AT06A031.

West, S, Stoll, O, Meierhofer, J, and Züst, S (2021). "Digital twin providing new opportunities for value co-creation through supporting decision-making," *Applied Sciences*, 11(9), 3750.

Xiu, Z., Nie, W., Yan, J., Chen, D., Cai, P., Liu, Q, Du, T, and Yang, B (2020). "Numerical simulation study on dust pollution characteristics and optimal dust control air flow rates during coal mine production," *Journal of Cleaner Production*, 248, 119197.