MULTIPLE MODEL ADAPTIVE ESTIMATION COUPLED WITH NONLINEAR FUNCTION APPROXIMATION AND GAUSSIAN MIXTURE MODELS FOR PREDICTING FUEL CONSUMPTION IN MARINE ENGINES

Mahmood Taghavi
UiT The Arctic University of Norway
Tromsø, Norway

Lokukaluge Prasad Perera
UiT The Arctic University of Norway
Tromsø, Norway

ABSTRACT

Digital twin type models can be developed for physical systems that are complex nonlinear a system of systems (SoS). However, such models are usually difficult to represent by linear equations. Therefore, an adequate linearization technique should be introduced. Therefore, linear models as digital twins can be interpreted easily and need much less computational power when applied to various industrial applications. On the other hand, a linearization approach can increase the respective system-model errors and impose significant constraints on the models of SoS, i.e., since linear models can be applicable only in limited operating regions. This research study aims to combine positive characteristics of both linear and nonlinear modelling into a digital twin development framework by having the properties of linear digital twin models locally while the model framework is covering the whole operating region of the SoS. An industrial application of marine engines as an SoS is considered for this study, where the respective models have been used to predict engine fuel consumption.

For this purpose, firstly, a dataset is selected from a marine engine of a selected ocean-going vessel. Then, several localized linear operational regions of the respective data set are identified using an unsupervised data-driven technique, i.e., on the engine propeller combinator diagram. For developing the localized models: firstly, the Gaussian Mixture Models method is used to cluster the data points into different operational regions of the engine propeller combinator diagram. Then, a nonlinear model of the relationship between features is developed in each cluster using the polynomial regression approach. Then, these models are combined using the Multiple Model Adaptive Estimation (MMAE) method to create an overall model for the marine engine as an SoS. The same model is utilized to predict the respective fuel consumption based on engine operational conditions.

Keywords: Digital Twin, Clustering, Multiple Model Adaptive Estimation, Marine Engine

1. INTRODUCTION

The commercial and economic importance of the shipping industry is indispensable. This industry accounts for the transportation of around 90% of traded goods globally [1]. In 2019, the international maritime trade volume was equal to 11,076 million tons of loaded goods [2], which is constantly growing. International shipping accounts for 2.9% of the world's CO2 emissions. Due to this rate of energy consumption in this industry and its consequent emissions, the International Maritime Organization (IMO) has devised strict rules for extensive reduction of greenhouse gas emissions [3], such as the Energy Efficiency Design Index (EEDI) for new ships, the mandatory Ship Energy Efficiency Management Plan (SEEMP) for all ships, and the attained Energy Efficiency Existing Ship Index (EEXI) required to be calculated for every ship [4]. Furthermore, the SEEMP also provides an approach for shipping companies to manage the ship and fleet-level efficiency over extended periods, namely the voluntary use of the Energy Efficiency Operational Indicator (EEOI) for new and existing ships [5].

As a result, in devising strategies for the future of this industry, these imposed regulations, emission reduction, and improving its efficiency should be considered. Adapting to these regulations is not easily accessible with conventional methods, and technological innovations should be considered to meet these requirements, such as digital twin-type applications, i.e., digitalization and machine learning (ML)-based approaches [6].

Also, shipping requires advanced databases, analytics, decision support systems [6], and data handling frameworks, which makes digital twin-based approaches and techniques more attractive and desirable. The effectiveness and flexibility of these methods attracted a great deal of attention from researchers and have been applied to many industrial applications. However, the initial research studies done on various ML and AI approaches have made the pathway toward digital twin type applications. As an example, Petersen and Jacobsen [7] present a statistical model of fuel efficiency in ship propulsion systems through high-quality sensory data sets, using a 10-fold ANN-based approach as a nonlinear supervised learning method for regression analysis.

Since fuel cost is a significant portion of the vessel's operational cost, improving fuel efficiency can be attractive for all ship owners. However, an accurate estimation of the fuel consumption (FC) in ocean-going vessels can play an essential role in such a task. Several different ways of estimating the FC of vessels are presented in the literature. For instance, Beşikçi et al. [8] developed an ANN-based estimator for the FC of a selected vessel for various operating conditions to support energy-efficient ship operations. Uyanık et al. [9] compared different prediction models for FC estimation, i.e., Multiple Linear Regression, Ridge and LASSO Regression, Support Vector Regression, Tree-Based Algorithms, and Boosting Algorithms, and concluded that Multiple Linear Regression has the best performance among all. Also, Anan et al. [10] utilized AI and high-dimensional statistical analysis to visualize ship performance and improve fuel efficiency using weather data, along with vessel navigation and ship system operation data in onboard and onshore cloud/edge platforms.

Furthermore, the International Maritime Organization (IMO) adopted a mandatory Fuel Oil Data Collection System (DCS) for international shipping. Under such amendments to MARPOL Annex VI on the DCS for fuel oil consumption of ships, ships of 5,000 gross tonnages and above are required to collect FC data, as well as other additional, specified data, including proxies for transport work [11].

The problem related to the mandatory DCS is that there are some operating points that the FC is not recorded for them, which can be an effect of faulty data acquisition systems or sensors, saturations or noise conditions of the sensors, or even human errors. As a result, in case of a missing FC value or anomaly values for FC, a methodology that can estimate the respective FC value for any given operating point can be valuable. This process can be approximated to a parameter interpolation or extrapolation process due to missing values. In other words, the missing data points of FC values should be identified. Then the FC can be estimated for them based on other operating parameters of a marine engine in such situations.

In order to estimate the FC of a selected engine, one should have a model of the respective engine operations. There are different models, such as control-oriented models, in the literature developed to estimate different operating states of marine engines, but most of them require in-cylinder measurements, such as pressure and temperature values [12, 13, 14]. Even though modern engines can produce such data sets, ship owners are hardly collecting and utilizing such data sets with higher sampling rates for energy efficiency applications in shipping.

On the other hand, the aim of this research is to develop a model using an available dataset of a selected vessel, i.e., without using the in-cylinder properties values. In this research, the FC is estimated by a combination of algebraic equations using the operating conditions of a marine engine as variables, i.e., engine operation modes.

These localized models, i.e., in engine operation modes, are combined and selected using the Multiple Model Adaptive Estimation (MMAE) algorithm. The MMAE algorithm is a powerful tool, using a bank of N parallel linear or nonlinear models. At any given operating point or region, it selects a single model or combination of models to generate the desired state. MMAE has been widely used for prediction purposes in recent literature. Barrios et al. used MMAE coupled with Adaptive Extended Kalman Filters (EKF) for predicting vehicle position using Global Positioning System (GPS) data [15]. Song et al. [16] used an improved MMAE approach coupled with Sage–Husa adaptive unscented Kalman filter for integrated navigation with time-varying noise levels. This approach improved the system's robustness to various noise levels, which enhanced the filter's performance in time-varying noisy measurements. Zhang et al. [17] presented a new scheme of weighted MMAE in which the conventional weighting algorithm is replaced by a dynamic weighting signal generator algorithm that relaxes the convergence conditions. The results confirmed that the proposed MMAE scheme is effective for different parameter estimations under various uncertainties.

The main contribution of this research study is to develop a data-driven framework for estimating the FC of a selected vessel using operational variables, i.e., available data sets, of a marine engine using ML techniques to develop the proposed digital twin type applications. For this purpose, firstly, a cluster analysis is performed, where 4 clusters, along with their mean and covariance values, are captured. In each cluster, a regression analysis is performed to find the respective functions of engine speed (ES) in RPM and FC in tons per day with main engine power (EP) in kW as the independent variables. Two polynomial regressions are performed in each data cluster, i.e., based on the data points in that cluster, one for finding the relationship between ES and EP and another for finding the relationship between the respective FC and EP. Then, based on the measured ES and EP, one of the developed models is utilized to estimate the future operational conditions of the engine. The model selection is based on the posterior probability calculated in the MMAE approach. At this stage, the MMAE has the role of selecting the best model among the developed cluster models based on their respective residual vector and covariance matrix.

2. METHODOLOGY

The proposed framework is developed using data from a selected vessel for two months. 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.

One should note that data sets of two months are considered for this study. The data set from the first month of the vessel's operation is selected for clustering and developing the model, including polynomials and the MMAE framework. The data set from the second month is used as new data values or test data set for evaluating the developed MMAE framework.

The time series for EP and FC for the second month are plotted in Fig. 1. As it is obvious in this figure, there are many missing data points in which the engine was running, but the FC is not recorded. Some anomalies are also noted in the same data sets, where the FC has abnormally high values while the engine runs normally. In the current research, all missing data points are removed from the data set in the preprocessing step, but in future work, the values for FC for these data points will be recovered using the proposed framework. Since this research is to prove the proposed method works, only data sets with clean values, i.e., without anomalies and missing data, are used for the performance evaluation of the proposed digital twin. One should note that that could help improve the initial models developed to support MMAE since those models are based on high-quality data sets.

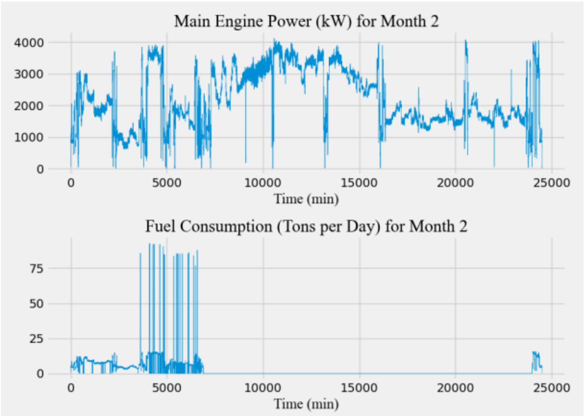

FIGURE 1: TIME SERIES FOR EP AND FC FOR THE SECOND MONTH BEFORE PREPROCESSING

In the following, the methodology for the clustering step is presented, then the general idea of the MMAE approach is briefly discussed.

2.1 GMM-EM Basic Concepts and Ideas

GMM can be considered as a probabilistic clustering algorithm [18]. In this method, it is assumed that the distribution of the data points can be estimated by combining J separate multivariate Gaussian distributions, f. Based on this, the total distribution of a random variable x in the dataset or Mixture Density Model (MDM), h, can be written as Eq. 1. The general parameter vector is Θ, which contains two sets of parameters, θ, and P. θ_j contains the mean vector and covariance matrix of the j^{th} cluster, μ_j and Σ_j, and P_j is the posterior probability of cluster j.

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

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

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

For each data point, x^q, the probability of belonging to one of the Gaussian distributions is higher, which is assumed to be the resulting cluster of that data point.

The remaining step is to estimate the values for the model parameters. Estimating the model parameters can be achieved using the EM algorithm. For formulating the EM algorithm, a new set of variables, y, is defined, with an observed part, x, and an unobserved part, j. Vector x is the vector of parameters obtained from the measurements. In order to find the parameters, a likelihood function is formed, and the estimation is done based on the distribution of this likelihood function. Since the logarithm function is monotonically increasing, to make the calculations more straightforward, the log-likelihood function, $L(\theta)$, is defined as Eq. 2 and can be maximized using the EM algorithm to estimate unknown parameters. In this equation, M is the number of data points in the data set.

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

The EM algorithm is an iterative scheme that consists of two steps. In the first step, or the E-Step, the expectation of the log-likelihood function is calculated. One should note that in the calculation of the expectation, and only in the distribution of θ, the last values of θ are used. In the second step, which is the maximization step, the derivative of the expectation of the log-likelihood function with respect to the parameters is calculated and set to zero to find the optimal values for the model parameters. The second step of the EM algorithm is called the M-Step. For the E-step, a new function, Q, is defined as Eq. 3, which calculates the expectation of the log-likelihood function with the assumption mentioned earlier.

$$
\begin{aligned}
Q\left(\theta;\hat{\theta}(t)\right) &= E\{L(\theta)|\hat{\theta}(t)\} = \sum_{q=1}^{M}\left[E\{lnf(y^q;\theta|x^q)|\hat{\theta}(t)\}\right] \\
&= \sum_{q=1}^{M}\sum_{j=1}^{J}\left[ln\left[f(x^q;\theta|j).P_j\right].P\left(j;\widehat{\Theta}(t)|x^q\right)\right] \\
&= \sum_{q=1}^{M}\sum_{j=1}^{J}\left[\left[-\frac{n}{2}ln(2\pi) - ln\left(|\Sigma_j|\right)\right.\right. \\
&\qquad -\frac{1}{2}\left(x^q-\mu_j\right)^T\Sigma_j^{-1}\left(x^q-\mu_j\right) \\
&\qquad \left.\left.+\,ln\left(P_j\right)\right].P\left(j;\widehat{\Theta}(t)|x^q\right)\right]
\end{aligned}
\tag{3}
$$

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 results, which are presented in Eq. 4, are the values of Σ_j, μ_j, and P_j for the next iteration [18].

$$
\begin{gathered}
P\left(j;\widehat{\Theta}(t)|x^q\right) = \frac{f\left(x^q;\hat{\theta}(t)|j\right)\hat{P}_j(t)}{\sum_{i=1}^{J} f\left(x^q;\hat{\theta}(t)|i\right)\hat{P}_i(t)} \\
\hat{\mu}_i(t+1) = \frac{\sum_{q=1}^{M} P\left(i;\widehat{\Theta}(t)|x^q\right)x^q}{\sum_{q=1}^{M} P\left(i;\widehat{\Theta}(t)|x^q\right)} \\
\hat{\Sigma}_i(t+1) = \frac{\sum_{q=1}^{M}\left[P\left(i;\widehat{\Theta}(t)|x^q\right)\left(x^q-\hat{\mu}_i(t+1)\right)\left(x^q-\hat{\mu}_i(t+1)\right)^T\right]}{\sum_{q=1}^{M} P\left(i;\widehat{\Theta}(t)|x^q\right)} \\
\hat{P}_i(t+1) = \frac{1}{M}\sum_{q=1}^{M} P\left(i;\widehat{\Theta}(t)|x^q\right)
\end{gathered}
\tag{4}
$$

Eq. 4 shows the iterative nature of the EM algorithm, in which the values of Σ_j, μ_j, and P_j in each iteration are calculated based on the previous iteration values. The initial values for the parameters of this algorithm are selected randomly. This iterative process is terminated after the convergence of parameters to stable values in GMMs is achieved. Additional criteria can be selected for terminating the proposed iterative scheme in the EM algorithm. In this research study, when the parameters' values do not change more than 2% in two successive steps, it is assumed the algorithm is converged.

In this research, the following parameters are used for the cluster analysis:

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

In the following section, the steps to fit a polynomial to each cluster in the dataset are presented, then the concepts of MMAE and its implementation based on the developed polynomials are presented.

2.2 Localized Model Development

As mentioned earlier, in each data cluster, a polynomial is considered for estimating FC and ES based on EP. In order to fit the polynomials, their coefficients, b_js should be estimated using the respective data points. For this purpose, the steepest descent algorithm is used to estimate the constants of the polynomial. $h(x)$ is the function or hypothesis assumed as the model, i.e., the polynomial. $J(a)$ is the cost function for this hypothesis. $h(x)$ and $J(a)$ are defined in Eq. 5.

$$
\begin{aligned}
h(x) &= b_0 + b_1x + b_2x^2 = \hat{y} \\
J(a) &= \frac{1}{2}\sum_{i=1}^{m} {e_i}^2 = \frac{1}{2}\sum_{i=1}^{m}(y_i-\hat{y}_i)^2 = \frac{1}{2}\sum_{i=1}^{m}\left(y_i-h(x_i)\right)^2 \\
&= \frac{1}{2}\sum_{i=1}^{m}(y_i-b_0-b_1x_i-b_2{x_i}^2)^2
\end{aligned}
\tag{5}
$$

In order to find the coefficients of the hypothesis, the steepest descent algorithm is used. This algorithm is an iterative scheme based on the rule in Eq. 6.

$$
b_j \coloneqq b_j - \alpha\frac{\partial J}{\partial b_j} \tag{6}
$$

Where ≔ sign means the old value will be overwritten by the new value on the right side. In this equation, α is the step size. As a result, the update rule for the coefficients can be written as:

$$
b_j \coloneqq b_j + \sum_{i=1}^{m}(y_i-b_0-b_1x_i-b_2{x_i}^2)\,{x_i}^j \tag{7}
$$

2.3 MMAE Framework

This section presents the framework for selecting and combining the models from a model bank, i.e., the models developed in each cluster. Firstly, the general concepts and ideas behind the MMAE approach are presented, then the implementation of MMAE for the current research study is discussed.

2.3.1 MMAE Basic Concepts and General Idea

A physical system that generates data can be approximately modeled as one of the N possible models, which can also be expended towards the digital twin model development steps. The MMAE algorithm uses a bank of N parallel linear or nonlinear models, and at any given situation, it selects a model or combination of models to generate the output or desired state of a digital twin. The selection of a model or a combination of models to generate the final estimate, as a part of the general digital twin model, is decided based on the posterior probabilities of each model with respect to the measured variable. At each step, one or more variables/states are measured, which is denoted by $y(k)$, and based on the resulting estimation error with respect to the measurement, the posterior probabilities are calculated for this digital twin application. The posterior probabilities determine which model has the best estimate among all models in a digital twin application. In other words, the posterior probabilities are defined as follows:

$$p_i(k) = \text{Probability}\left[i^{th}\ model\ is\ true|y(k)\right] \tag{8}$$

As a result, the posterior probabilities are used to weigh local estimates to generate the final estimated values for the respective model parameters. The posterior probability of the *i*

As a result, the posterior probabilities are used to weigh local estimates to generate the final estimated values for the respective model parameters. The posterior probability of the i^{th} model at time k is denoted by $p_i(k)$, which is the probability that the parameter estimated by the i^{th} model, $\hat{x}(k)$, is the estimated value of the respective parameter, $x(k)$. As a result, $p_i(k)$ should satisfy the following constraints:

$$p_i(k) \geq 0, \qquad \sum_{i=1}^{N} p_i(k) = 1 \tag{9}$$

The posterior probabilities are calculated by the Posterior Probability Evaluator (PPE). Each linear or nonlinear model has its own estimate, residual vector, and error covariance matrix. The PPE evaluates the posterior probability for each model based on the residual vector and its covariance matrix.

To calculate the posterior probability, the first step is to calculate the residual vector. The residual vector for each model is defined as the difference between the measured vector, $y(k)$, and the estimated vector, $\hat{y}_i(k)$, of the digital twin application from that model for the same variable(s). Thus, the residual vector is calculated based on the following equation.

$$r_i(k) = y(k) - \hat{y}_i(k) \tag{10}$$

The covariance matrix of the same variable(s) is, by definition, the expected value of the squared residual value.

$$\begin{aligned} S_i(k) = cov[r_i(k)] &= E[r_i(k).r_i(k)^T|H_r] \\ &= E\Big[\big(y(k) \\ &- \hat{y}_i(k)\big)\big(y(k) - \hat{y}_i(k)\big)^T|H_r\Big] \\ &= \int_{-\infty}^{+\infty} \big(y(k) \\ &- \hat{y}_i(k)\big)\big(y(k) - \hat{y}_i(k)\big)^T p(y(k)|H_i)\,dy \end{aligned} \tag{11}$$

Where H_i is the hypothesis or model i, and $p(y(k)|H_i)$ is the probability density function of the i^{th} model for vector $y(k)$. The last equation is the same as the definition of the covariance in a multivariate probability density function. As a result, $S_i(k)$ is the same as the covariance matrix of the i^{th} model.

As mentioned in the clustering section, it is assumed that all clusters have Gaussian distributions, therefore the dynamic probability evaluation $p_i(k)$ for the i^{th} model at iteration, k is calculated based on Eq. 12.

$$p_i(k) = \left(\frac{\beta_i(k)e^{-\frac{1}{2}\omega_i(k)}}{\sum_{j=1}^{N}\beta_j(k)e^{-\frac{1}{2}\omega_j(k)}p_j(k-1)}\right)p_i(k-1) \tag{12}$$

Where,

$$\begin{aligned} \beta_i(k) &= \frac{1}{(2\pi)^{\frac{m}{2}}\sqrt{\det\big(S_i(k)\big)}} \\ \omega_i(k) &= r_i(k)^T S_i(k)^{-1} r_i(k) \end{aligned} \tag{13}$$

After calculating the posterior probabilities, $p_i(k)$, the final variable estimates, $\hat{x}(k)$, can be calculated as the sum of each model variable estimates, $\hat{x}_i(k)$, multiplied by the associated posterior probabilities.

$$\hat{x}(k) = \sum_{i=1}^{N} p_i(k)\hat{x}_i(k) \tag{14}$$

In the case of a scalar system, the residual vector is replaced by the residual value, and the covariance matrix is replaced by the variance value of the respective parameter.

2-3-2- MMAE Framework Development for FC Estimation

As mentioned in this research, the MMAE approach is used for combining the models found in each data cluster. Two polynomial regressions are performed in each cluster based on the respective data points, one step for finding the relationship between ES and EP values, and another for finding the relationship between the respective FC and EP values. MMAE is used to find and combine the best models among these polynomials for estimating the FC values based on the posterior probabilities, where that approach has been categorized as the development steps for digital twin. Adapting the MMAE approach for the current research goals is presented in this section, and the general steps of this process are shown in Fig. 2.

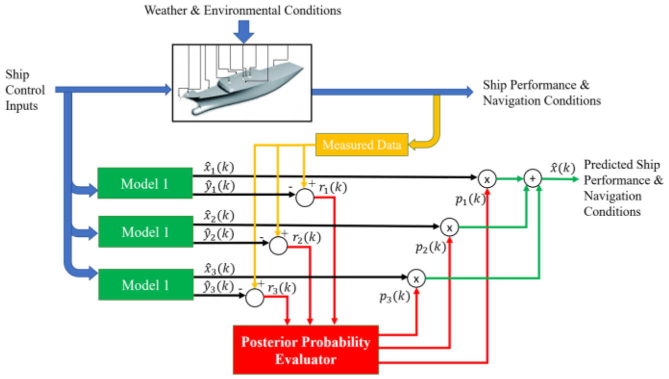

FIGURE 2: MMAE DIAGRAM

As mentioned earlier, the idea behind this framework is that, firstly, a cluster analysis is performed to find the operating regions of a marine engine using an existing dataset of the selected vessel. Then, the offline models are developed for ES-EP and FC-EP in each cluster using polynomial regression. In estimating the future states for a new dataset, only the ES and EP values are assumed to be recorded, and the FC will be estimated based on the selected models. For this purpose, the recorded ES values at the respective EP values are compared to the estimated ES values by the respective ES-EP polynomials. This comparison is made in each cluster, and their associated residual values will be calculated. The PPE calculates the posterior probability for each model based on the residual value and the variance for EP in each cluster. The final value for FC is the weighted sum of the estimated values for FC calculated using each model of this proposed digital twin framework.

In the proposed framework, the following points should be mentioned:

- Since in the selected approach, the models for each cluster are developed in advance, and there is no model update at each time step, the covariance matrix of each model is the same throughout time steps and $S_i(k) = S_i = cov_i$.
- Moreover, in the proposed method, only one feature is recorded, and one feature is to be estimated. Based on this, all the residual and covariance equations are written in a scalar form. As a result, the covariance matrix and residual vector are replaced by the variance and residual value for the measured variable, respectively.
- The normalized dataset formed by mapping all the features into [-1, 1] intervals is used for the clustering to prevent bias in favor of features with higher values. As a result, all the calculations and estimations in the MMAE framework are also performed in the normalized feature space.
- The data set for one month is used to perform the clustering and polynomial regression, and the FC estimation is performed for the data set for the following month, where the estimated FC is compared with the measured FC. The data set for the second month, for which the FC is known, is selected as the test data to evaluate the proposed model. For future reference, the data set for the second month will be called the query points.

As mentioned, the clusters are formed using the GMM-EM approach, and all the clusters are assumed to have Gaussian distributions. In the covariance matrix of each data cluster, the diagonal elements are the variance values of each feature.

3. RESULTS AND DISCUSSION

In this section, the results of the clustering algorithm and then the estimation step are presented.

3.1 Clustering Results

The detailed implementation and results of the proposed clustering algorithm for a dataset of one month of the selected vessel are presented in [19]. Based on these results, 4 clusters or operating regions are identified for this vessel using the selected operating parameters of the marine engine. It is worth mentioning that data points with missing values and anomalies are removed from the dataset before performing the clustering during the preprocessing step. Also, the data points associated with the times when the engine was not functional, and no power was being generated are also removed from the dataset since this research aims to analyze and develop a model for the operating region of the engine. Therefore, some selected data anomalies are removed from this data set during the preprocessing step.

In Fig. 3, all the data points are plotted in a 3-D space with different colors for each cluster. The percentage of data points belonging to each cluster is calculated and presented in Table 2. As shown in this table, the dominant cluster or operating region is the third one, and nearly 56% of the data points are from this cluster. In other words, the vessel operated for nearly 56% of the time in this data cluster or operating region.

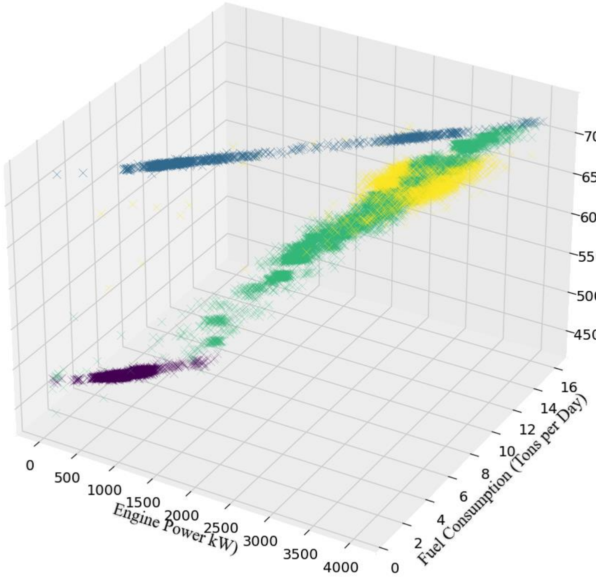

FIGURE 3: FINAL CLUSTER CONFIGURATION FOR THE 3D SPACE

TABLE 2: PERCENTAGES OF DATA POINTS BELONGING TO CLUSTERS IN THE 3D FEATURE SPACE

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

The resulting cluster centers or cluster mean values for the selected feature space are presented as follows. The first row in all the mean value vectors is the EP mean value, the second is the FC mean value, and the third is the ES mean value.

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

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

For the clustering algorithm, all the data points are normalized, so the algorithm is not biased in favor of the features with the higher values. The covariance matrices of the captured clusters for the normalized data can be written as follows:

$$cov_1 = \begin{bmatrix} 0.00120847 & 0.00127509 & -1.5828e-05 \\ 0.00127509 & 0.0014179 & -1.87791e-05 \\ -1.5828e-05 & -1.87791e-05 & 1.46093e-05 \end{bmatrix}$$

$$cov_2 = \begin{bmatrix} 0.0571969 & 0.053608 & -2.57862e-05 \\ 0.053608 & 0.0502977 & -2.68347e-05 \\ -2.57862e-05 & -2.68347e-05 & 1.15797e-05 \end{bmatrix}$$

$$cov_3 = \begin{bmatrix} 0.0150404 & 0.0151152 & 0.015841 \\ 0.0151152 & 0.0152681 & 0.0159561 \\ 0.015841 & 0.0159561 & 0.0184682 \end{bmatrix}$$

$$cov_4 = \begin{bmatrix} 0.00542729 & 0.0049016 & 0.00105455 \\ 0.0049016 & 0.00540206 & 0.00112822 \\ 0.00105455 & 0.00112822 & 0.00268443 \end{bmatrix}$$

The diagonal elements of these covariance matrices are the variance values for the respective operational parameters.

Now that the clusters have been captured by the GMMs and EM algorithm, the next step is to develop a model for each cluster and use MMAE to combine these localized models to generate a general model to cover the whole engine operating region.

3.2 Estimation Step Results

In the following, the results of the polynomial fitting to the dataset in each cluster are presented, then the implementation of MMAE based on the developed polynomials is discussed.

3.2.1 Polynomial Regression Results and Discussion

Using the NumPy toolbox in Python, a 2nd order polynomial is fitted to data points in each cluster. The resulting polynomials for each data cluster using the mentioned algorithm are presented in Tables 3 & 4. In each cluster, one polynomial is fitted for calculating ES from EP, and another one is found for estimating FC from EP.

TABLE 3: RESULTING 2ND ORDER POLYNOMIALS FOR CALCULATING ES FROM EP FOR EACH CLUSTER

Cluster 1	$ES = 9.510 \times 10^{-2} \times EP^2 - 4.460 \times 10^{-2} \times EP + 0.135$
Cluster 2	$ES = 3.966 \times 10^{-4} \times EP^2 - 9.058 \times 10^{-4} \times EP + 0.972$
Cluster 3	$ES = -0.187 \times EP^2 - 1.308 \times EP - 8.311 \times 10^{-2}$
Cluster 4	$ES = -8.263 \times 10^{-2} \times EP^2 + 0.225 \times EP + 0.661$

TABLE 4: RESULTING 2ND ORDER POLYNOMIALS FOR CALCULATING FC FROM EP FOR EACH CLUSTER

Cluster 1	$FC = 0.728 \times EP^2 + 0.812 \times EP + 8.903 \times 10^{-3}$
Cluster 2	$FC = 2.313 \times 10^{-2} \times EP^2 + 0.915 \times EP + 5.529 \times 10^{-2}$
Cluster 3	$FC = 0.105 \times EP^2 + 0.875 \times EP + 2.717 \times 10^{-2}$
Cluster 4	$FC = -1.193 \times 10^{-2} \times EP^2 + 0.928 \times EP + 5.427 \times 10^{-2}$

3.2.2 MMAE Implementation Results

The first step in implementing MMAE is to calculate the residual values between the measured and the estimated values of the respective variables of each cluster. Fig. 4 shows the absolute residual values for different cluster models for all the query points, i.e., data points from the second month. Also, the minimum residual values among these four models are presented in Fig. 5. This figure shows that the minimum residual value among all the clusters is below 0.05 for the majority of the data points, which shows a good fit for the polynomial regression. The average value for the minimum residual value is 0.026.

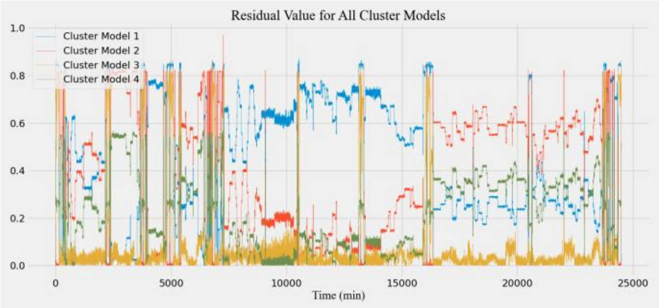

FIGURE 4: CLUSTER MODELS RESIDUALS FOR DIFFERENT CLUSTER MODELS FOR ALL THE QUERY POINTS

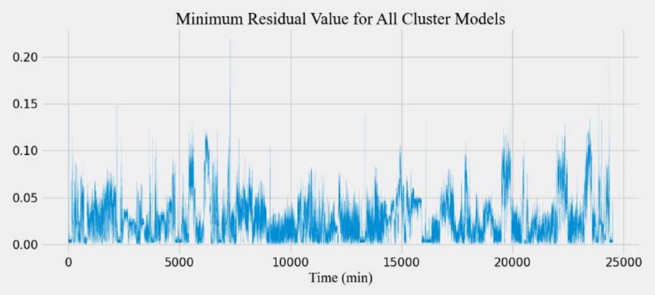

FIGURE 5: MINIMUM CLUSTER MODELS RESIDUALS FOR ALL THE QUERY POINTS

Fig. 6 plots the estimated FC values from all cluster models in the same diagram. One should note that at each point, one model or a combination of them can be selected to generate the final value for FC. As mentioned before, this selection is performed based on the posterior probability of each cluster model. The posterior probabilities of each model for all the query points are plotted in Fig. 7. As seen from this figure, the model from cluster 3, the dominant cluster in the first month, is responsible for estimating the FC values more than any other model. This means that in the second month, the 3rd cluster also defines the dominant operating region of the selected vessel.

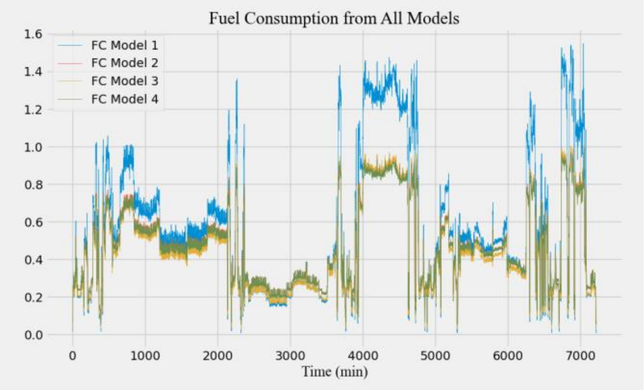

FIGURE 6: ESTIMATED FC VALUES FROM ALL CLUSTER MODELS

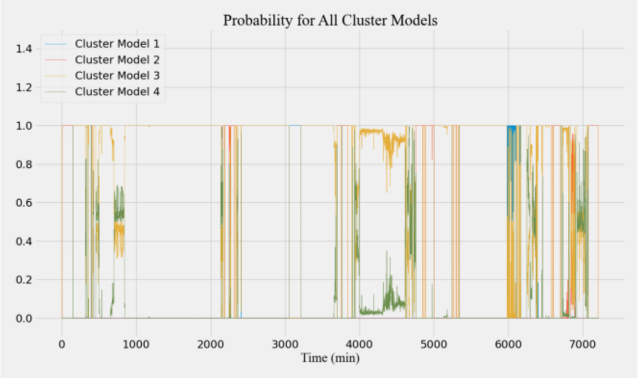

FIGURE 7: POSTERIOR PROBABILITIES OF EACH MODEL FOR ALL THE QUERY POINTS

The result of using MMAE for combining the models for estimating the FC for the query points is presented in Fig. 8. In this figure, the estimated FC and the measured FC from the dataset are plotted in the same diagram. As demonstrated in this figure, a reasonable estimation of the FC has been performed using the proposed approach. The estimation error is also calculated and plotted in Fig. 9. The average value of the absolute error of the estimated FC for the query points is equal to 0.011, which is 2.5% of the mean value for the measured FC for the same query points.

4. CONCLUSION AND FUTURE WORKS

This section presents the conclusions of the proposed framework for estimating the FC of the selected vessel based on digital twin type applications. This framework consists of data clustering using the GMM-EM algorithm coupled with the MMAE approach for digital twin type model development. Based on the results presented in the previous section, the following points can be concluded:

- In this research, EP is used as the independent variable, ES as the measured variable/state, and FC is estimated for any given query point with known EP and ES. The results show this approach has a small error in estimating FC for a given EP and ES.
- In each cluster, two polynomials are fitted for approximating ES and FC as a function of EP. These polynomials are the models developed in each cluster. Based on the measured ES, MMAE is used to determine the model from which cluster should be used to estimate and generate the FC for the query point.
- MMAE is a powerful and effective approach for function approximation and model development of marine engines, and that can further be implemented under digital twin type frameworks.
- The proposed framework can be used for estimating the missing values not only for FC but also for any other operating variable of the engine. This framework can also be very effective for preparing the dataset in the preprocessing step.
- In this research, only EP, ES, and FC are considered as the features for analyzing the engine performance. If more features, such as speed over ground, trim, draft, loading condition, and weather conditions, are included, more accurate results can be achieved, and a more comprehensive digital twin framework can be developed.

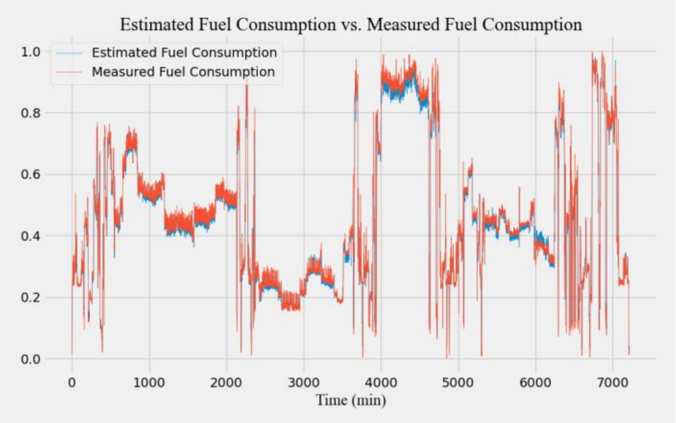

FIGURE 8: ESTIMATED FC AND MEASURED FC FOR THE QUERY POINTS

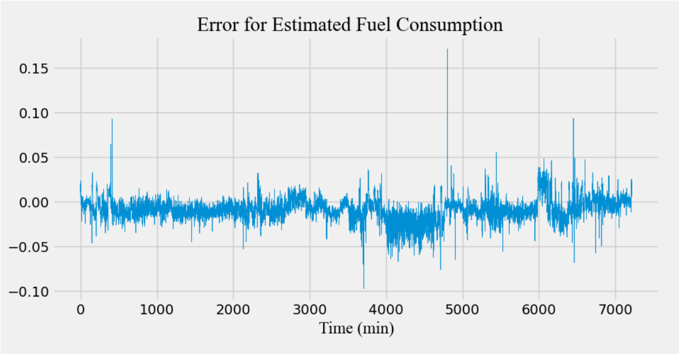

FIGURE 9: ESTIMATION ERROR FOR FC OF THE QUERY POINTS

ACKNOWLEDGEMENTS

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

[1] Organisation for Economic Co-operation and Development, Ocean shipping and shipbuilding
URL:
https://www.oecd.org/ocean/topics/ocean-shipping/

[2] United Nations Conference on Trade and Development UNCTAD 2020, Review of Marine Transport, Table 1.1- Page 4

[3] IMO, 2021. 'IMO-Norway project is supporting States to implement energy efficiency measures and explore opportunities for low carbon shipping.'
URL:
https://www.imo.org/en/MediaCentre/PressBriefings/Pages/06 GHGinitialstrategy

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

[5] Marine Environment Protection Committee, 2009. Guidelines for voluntary use of the ship energy efficiency operational indicator (EEOI). International Maritime Organization, Report.\

[6] Norwegian Shipowners' Association., 2021, March. Maritime Outlook.
URL:
https://rederi.no/en

[7] Petersen, J.P., Jacobsen, D.J. and Winther, O., 2012. Statistical modelling for ship propulsion efficiency. Journal of marine science and technology, 17(1), pp.30-39.

[8] Beşikçi, E.B., Arslan, O., Turan, O. and Ölçer, A.I., 2016. An artificial neural network based decision support system for energy efficient ship operations. Computers & Operations Research, 66, pp.393-401.

[9] Uyanık, T., Karatuğ, Ç. and Arslanoğlu, Y., 2020. Machine learning approach to ship fuel consumption: A case of container vessel. Transportation Research Part D: Transport and Environment, 84, p.102389.

[10] Anan, T., Higuchi, H. and Hamada, N., 2017. New artificial intelligence technology improving fuel efficiency and reducing CO2 emissions of ships through use of operational big data. Fujitsu Sci. Tech. J, 53, pp.23-28.

[11] IMO, 'Data collection system for fuel oil consumption of ships.'
URL:
https://www.imo.org/en/OurWork/Environment/Pages/Data-Collection-System.aspx

[12] Llamas, X. and Eriksson, L., 2019. Control-oriented modeling of two-stroke diesel engines with exhaust gas recirculation for marine applications. Proceedings of the Institution of Mechanical Engineers, Part M: Journal of Engineering for the Maritime Environment, 233(2), pp.551-574.

[13] Nikzadfar, K. and Shamekhi, A.H., 2011. Developing a state space model for a turbocharged diesel engine using least square method. Khaje Nasir University of Technology, SAE International.

[14] Nikzadfar, K. and Shamekhi, A.H., 2015. An extended mean value model (EMVM) for control-oriented modeling of diesel engines transient performance and emissions. Fuel, 154, pp.275-292.

[15] Barrios, C., Himberg, H., Motai, Y. and Sadek, A., 2006, September. Multiple model framework of adaptive extended Kalman filtering for predicting vehicle location. In 2006 IEEE Intelligent Transportation Systems Conference (pp. 1053-1059). IEEE.

[16] Song, J., Li, J., Wei, X., Hu, C., Zhang, Z., Zhao, L. and Jiao, Y., 2022. Improved Multiple-Model Adaptive Estimation Method for Integrated Navigation with Time-Varying Noise. Sensors, 22(16), p.5976.

[17] Zhang, W., Wang, S. and Zhang, Y., 2018. Multiple-model adaptive estimation with a new weighting algorithm. Complexity, 2018.

[18] Theodoridis, S. and Koutroumbas, K., 1999, July. Pattern recognition and neural networks. In Advanced Course on Artificial Intelligence (pp. 169-195). Springer, Berlin, Heidelberg.

[19] Taghavi, M. and Perera, L.P., 2022, June. Data Driven Digital Twin Applications Towards Green Ship Operations. In International Conference on Offshore Mechanics and Arctic Engineering (Vol. 85895, p. V05AT06A028). American Society of Mechanical Engineers.