Ocean Engineering 251 (2022) 110979

Contents lists available at ScienceDirect

Ocean Engineering

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

# A novel method for estimating missing values in ship principal data

Youngrong Kim [a,*], Sverre Steen [a], Helene Muri [b]

[a] Department of Marine Technology, Norwegian University of Science and Technology, Trondheim, Norway
[b] Industrial Ecology Programme, Department of Energy and Process Engineering, Norwegian University of Science and Technology, Trondheim, Norway

ARTICLE INFO

*Keywords:*
Missing data
Ship principal data
Model-based computation
Regression analysis

ABSTRACT

Missing values in the fleet data set acquired in the marine sector reduce the data available for analysis, which can decrease the statistical power of the model and negatively affects the energy-efficient operation and decision-making. This article presents a method to estimate ship principal data. A model-based computation method using regression analysis was used to handle missing values, and a case study was conducted on principal data from 6,278 container ships in the IHS Sea-Web database. To implement a model for predicting missing values, the entire data set was randomly divided into 80% to 20%, which were used as a training data set and test data set. The prediction performance of models was compared with several regression equations proposed in prior studies, which shows that there is a significant improvement with our method. The goodness of fit of the current method has increased by up to 15.6% over the previous methods. It also showed good applicability for ships with restrictions on certain dimensions, such as the standards for Suez and Panama Canal. The findings presented here may be helpful from the estimation for key parameters of the ship to the computation of missing values in the marine sector.

## 1. Introduction

Data sets acquired from industry are often incomplete, which may be due to various reasons, including sensor failures, measurements outside the range of sensors, malfunctions in data collection systems, power cuts, interruption of transmission lines, and errors in data recording (Imtiaz and Shah, 2008; Khatibisepehr et al., 2013). For instance, in the maritime industry, there may be missing values of 4.4% to 26.0% of the data collected from the machinery system due to various circumstances (Tsitsilonis and Theotokatos, 2018; Lazakis et al., 2019). AIS (Automatic Identification System) mounted on a ship may cause loss of signals registered by the satellite if the time slot is overlapped due to interference with other ships when the ship navigates in congested waters, and in bad weather, such as lightning, the transmission may be lost due to shut off of the receiver (Lloyd's list intelligence, 2017). In addition, entire fleet data, which is widely used for ship operational efficiency, emission prediction from maritime transport, and hull design, is comprehensively collected from various organizations such as ships, owners, shipbuilders, and port authorities (Wang et al., 2016; IHS, 2019). Due to the nature of such data, missing values inevitably exist. If a large fraction of the data is missing, it may lead to inaccurate analysis and prediction, which can negatively affect the energy-efficient operation and decision-making of the fleet (Gutierrez-Torre et al., 2020). Therefore, it is important to process and complete the missing values appropriately before analyzing the acquired data.

Despite the increasing utilization of big data and the use of such in machine learning in the maritime industry, combined with the growing importance of appropriately handling missing values, there are few published studies on missing data. Most of them were to recover missing route information or identify ship behavior patterns through incomplete AIS data analysis (Liu and Chen, 2013; Mao et al., 2018; Dobrkovic et al., 2018; Gutierrez-Torre et al., 2020). There have also been attempts to handle missing data obtained from the machinery system of an operating ship. In Cheliotis et al. (2019)'s study, a hybrid imputation method combining K-nearest neighbor (KNN) and multiple imputation by chained equations algorithms (MICE) has been developed for efficient operation and performance improvement of the main engine systems of ships in operation. Imputation is defined as a method of filling in values of missing data (Little and Rubin, 2019). This method was applied to time-series data collected from a total of eight sensors combined with the main engine. In the process of developing a decision support framework for optimal ship routes based

* Corresponding author.
*E-mail address:* youngrong.kim@ntnu.no (Y. Kim).

https://doi.org/10.1016/j.oceaneng.2022.110979
Received 30 October 2020; Received in revised form 14 January 2022; Accepted 26 February 2022
Available online 17 March 2022

on weather and fuel consumption, Gkerekos and Lazakis (2020) used MICE algorithm applied in Cheliotis et al. (2019) to impute the missing points of weather forecast data. Velasco-Gallego and Lazakis (2020) conducted a comparative study investigating a total of 20 machine learning and time-series prediction algorithms to support a real-time decision-making strategy. In their subsequent study (Velasco-Gallego and Lazakis, 2021), they proposed a new framework by implementing the first-order Markov chain with some multiple imputation methods. In addition to the maritime field, various methods based on machine learning such as multiple regression, random forest, KNN, and support vector regression have been tested and applied to process missing data across the industry (Kim et al., 2017; Andiojaya and Demirhan, 2019; Afrifa-Yamoah et al., 2020; Lin and Tsai, 2020; Jung et al., 2020; Wang et al., 2021). It appears that most research related to missing data handling in the maritime industry is limited to continuous time-series data on a specific ship data, such as the state of the machinery system or the location of the ship. Studies related to stationary data such as ship principal data are rare.

Many studies have been conducted to predict the principal dimensions and particulars. Most of these were intended to be used in the initial design or to optimize design variables for specific vessels, and proposed regression formulas using statistical data of ships. Piko (1980) performed a regression analysis on deadweight tonnage and service speed using the length, breadth, draught, gross tonnage, and power based on Lloyd's shipping database, which became a cornerstone for many subsequent studies. Charchalis and Krefft (2009), Charchalis (2014) attempted to design equations for estimating efficient and optimal main parameters during the initial ship design stage. Parameters were predicted using container capacity and deadweight tonnage, but the range and number of ships used in the research was limited. Kristensen (2012, 2013, 2016) performed extensive statistical analysis on bulk carriers, container ships, and tankers and proposed regression equations for a number of parameters. In particular, they were established by dividing groups according to the size of the vessel, which enabled considering the detailed characteristics of each range. Abramowski et al. (2018) presented regression formulas to estimate key characteristics of container ship based on various combinations of deadweight tonnage, container capacity, in addition to some other variables and proved to be a practical application at the preliminary design stage. Such studies using regression formulas showed generally good accuracy based on several key input variables selected in combination with using domain-knowledge. Conversely, recent work has suggested models based on ANN, which showed better prediction performance than previous ones. Abramowski (2013) applied ANN techniques to optimize the design parameters of cargo ships. In this study, seven parameters were used to implement a model for effective power determination. Moreover, optimization for a single objective of the minimum thrust and multi-objective of the minimum propulsion power and maximum deadweight was performed, while implementing the model. Gurgen et al. (2018) presented a design tool for estimating key details of chemical tankers during the preliminary design phase. In that study, an ANN was used for model implementation and key dimensions such as overall length, length between perpendicular, breadth, draught, and freeboard were predicted using the dead weight and service speed of the vessel as the default input. However, there were concerns about the complexity of the model and the possibility of overfitting to the data set when applying ANN. In this regard, the preceding models implemented based on specific vessel data set were somewhat less applicable in other studies. In addition, all such work including regression analysis and ANN to predict the principal components of vessels always assumed a complete data set and did not address the processing of missing values within the data set. In fact, if some data are missing in the data set, or the composition of the data set is different from the previous studies, the estimation methods mentioned above are difficult to apply.

In fleet-wide studies, such as analysis of ship operational energy efficiency and global greenhouse gas emissions at sea, the principal details of ships are used as important basic data along with time-series data such as AIS and in-service data. Sea-Web is used as the source for ship principle data, and it is found that some parameters are missing for a number of ships. If sufficient amounts of data are obtained or there are few missing values, the problem may be solved by simply removing the missing values, otherwise, it can result in an inappropriate analysis result. Although prior studies using machine learning-based missing data imputation methods (Cheliotis et al., 2019; Velasco-Gallego and Lazakis, 2020) have shown high accuracy, they basically focused on time-series data from specific ships. Furthermore, the interpretation of machine learning-based models itself was difficult, and it was not straightforward to obtain ship principal parameters by applying the same settings to other studies.

In this study, we propose a new method that is designed for estimating missing values in ship principal data. To deal with such values, a model-based computation method using regression analysis is used in this study, which is widely applicable to various data compositions and characteristics. Through the method, the relationship between ship principal data is first identified using correlation analysis and regressional curve fitting functions. Then the missing data is replaced based on regression analysis accompanied by variable selection. This approach complements previous research for estimating ship principal data with respect to handling the missing data. The estimated models and results can be interpreted and, regression expressions for each ship parameter are provided at last, making it easier to apply in other studies. It is believed that the method can be applied also under other circumstances when it is needed to replace erroneous values.

In Section 2, the new method to estimate the missing values for ship principal data is described, and Section 3 shows the results of the model through a case study of container ships. Section 4 compares the performance of the developed model to the regression models of the previous studies and the random forest model. It also verifies the ability of the method to respect particular dimensional restrictions, such as being able to pass through the Suez Canal and Panama Canal. Finally, in Section 5, the conclusions drawn.

## 2. A new method to estimate missing values for ship main particulars

### 2.1. Missing data types

Missing data can cause problems since robust statistical analysis requires values for each variable. Therefore, in situations where missing values are expected, one needs to decide how to handle them. The missing data can be divided into three different types: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) by the cause of missing (Rubin, 1976). MCAR means that the values are lost randomly throughout the data range, regardless of the type and value of the variables. Whilst MAR refers to a case in which the loss of the data is not random across all observations, but only within a subset of the data. If the characteristics of MCAR or MAR are not satisfied, data belongs to the MNAR. MNAR refers to a case in which the values of the missing variables and the reasons for the missing are related.

Missing values can be handled in mainly two ways, either by elimination or imputation of them. Deleting the parameter or variable set that includes missing values from the entire data sets is the easiest and simplest way to handle them. However, it can substantially lower the sample size, leading to a severe lack of statistical power. In particular, it is possible when there are many variables associated in the analysis, and each variable has missing data for several cases, which can lead to biased results, depending on the cause of data missing (Little and Rubin, 2019). If only missing values are removed by applying a pair-wise deletion, the change of subset may lead to distort the analysis results and

make it difficult to interpret. In contrast, imputation can preserve all cases by estimating missing data based on other available information, enabling subsequent statistical analysis of the entire data (Hair et al., 2018).

According to Hair et al. (2018), the method of handling missing values varies depends on the ratio of missing in the data set and its characteristics. If the missing values are less than 10%, they can be removed from the data set or any of the completion methods can be applied. If the missing ratio is between 10 and 20%, hot deck replacement and regression analysis methods are appropriate for MCAR data, and the model-based method is recommended for MAR data. In the case of more than 20%, a regression method is recommended to use for MCAR data, and a model-based method for MAR data.

### 2.2. Missing data handling process

As mentioned in the previous section, there are various methods of processing the missing data depending on the characteristics of the data or the types of missing, and the corresponding results will vary. We propose a model-based computation method using regression analysis that is widely applicable against the ratio and characteristics of missing data, which is able to handle it properly. The main challenge of model-based computation is to establish a model for predicting each target variable that contains missing values in the data set. In fact, many studies have applied regression analysis of statistical data to estimate ship principal parameters and they assumed a complete data set. However, this study aims to complement previous methods from the perspective of missing data handling. In other words, the method proposed in this study is applicable even when the input parameters used in the equations proposed in the previous studies are not in the data set. However, since this algorithm includes several statistical analysis methods, it should be noted that it may not work properly if the size of the data set for filling the missing values is too small. Fig. 1 illustrates the method for completing the ship principal data proposed in this study, and the main steps are composed of the following three steps:

(i) Initial computation: The objective of the first step is to obtain a complete data set by filling in the empty values with plausible values. Multiple regression analysis used in this study requires a complete matrix of variables. However, with the incomplete data set, since the values of some variables are empty, multiple regression analysis cannot be performed directly. Therefore, this step provides a platform for performing a multiple regression analysis in the next step. First, curve fitting is performed between each ship design parameter using a variety of function forms, including linear, quadratic, cube, power, and logarithmic based on the least-squares method. At this point, the overall data sets of each variable are used. Afterward, a single variable and function type that provides the highest $R^2$ value (Coefficient of determination) with the variable to be fitted is identified (Refer to Algorithm 1). Finally, missing values in each ship's case are filled by the curve fitting of the other variable with the highest $R^2$ value (Refer to Algorithm 2). If the corresponding variable is missing in a specific ship case, the next best variable is selected. That is, among the variables that exist in the ship case, the curve fitted value of the variable with the next higher $R^2$ value is used to fill in the missing value. This process repeats until all missing values within the entire data set are filled.

Let the ship data set ($X$) has $N \times M$ matrix containing some missing values:

$$\boldsymbol{X} = (\boldsymbol{X_1}, \boldsymbol{X_2}, \cdots, \boldsymbol{X_M}) = \left.\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1j} \\ x_{21} & x_{22} & \cdots & x_{2j} \\ \vdots & \vdots & \ddots & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} \end{bmatrix}\right\} N \quad \underbrace{\qquad\qquad}_{M}$$

where $x_{ij}$ is $j$th parameter in $i$th ship case, $N$ is the number of ship cases, and $M$ is the number of ship principal parameters.

**Algorithm 1:** Identify the fitted function for each parameter using curve fitting

```
for j ∈ [1, 2, ··· , M] do
    for f ∈ [Curve fitting functions] do
        Calculate R² value between X_j and f(X_−j) using the
        non-missing values.
    end
    Save function f(·) and input parameter X* among X_−j that
    fit best with X_j as f^CV(X*).
end
```

where $f(\cdot)$ is curve fitting function (Refer to Eqs. (1)–(5)), $X_j$ is the $j$th parameter vector in all ship cases, $X_{-j}$ is parameter vector except $j$th parameter in all ship cases, $X^*$ is the selected parameter vector among $X_{-j}$ in all ship cases, and $f^{CV}(\boldsymbol{X}^*)$ is the fitted function that has $X^*$ as an input vector, which shows the highest $R^2$ value between the target parameter vector $X_j$.

**Algorithm 2:** Make an initial guess for all missing values using fitted function

```
for i ∈ [1, 2, ··· , N] do
    for j ∈ [1, 2, ··· , M] do
        Estimate x̂_ij using the fitted function f^CV(·) and
        parameter x_i*, i.e., x̂_ij = f^CV(x_i*).
    end
    Fill in x_ij using the estimated value x̂_ij if x_ij is missing,
    i.e., x_ij = x̂_ij.
end
```

where $x_i{}^*$ is the selected parameter that shows the highest $R^2$ value with the target parameter $x_j$ in $i$th ship case, and $\hat{x}_{ij}$ is the estimated value of the $j$th parameter in $i$th ship case from Algorithm 1 and 2.

(ii) Final imputation: This step is to update the originally missing values with predicted values by performing regression analysis based on the completed data sets obtained from step 1. Before implementing a predictive model for each variable, remaining variables except for a target variable, are converted into the function type with the highest $R^2$ to the target variable, which is to consider the non-linear physical relations between each variable. That is, the curve-fitted values are entered in the terms of the independent variables in subsequent multiple linear regression expressions. The main process in this step is performing multiple regression analysis with the backward elimination method to make a predictive model for each variable. The $p$-value for each variable that makes up the model and the $BIC$ (Bayesian Information Criterion) of the model are evaluated. It starts with all candidate variables and sequentially removes one of which is the least statistically significant for the model, i.e. the variable with the maximum $p$-value. When all models are evaluated according to the number of input variables, the model with the minimum $BIC$ is selected as a final model (Refer to Algorithm 3). Once the predictive model of each variable is set up, the values filled in the previous step are replaced with the newly predicted values from the model (Refer to Algorithm 4).

**Algorithm 3:** Perform multiple regression analysis with backward elimination to make a prediction model for each parameter

**for** $j \in [1, 2, \cdots, M]$ **do**
  Convert parameters $\boldsymbol{X}_{-j}$ to the curve fitted form $\boldsymbol{X}_{-j}{}^{CV}$, i.e., $\boldsymbol{X}_{-j}{}^{CV} = f^{CV}(\boldsymbol{X}_{-j})$.
  **repeat**
    Fit a multiple regression model $f^{MR}(\boldsymbol{X}_{-j}{}^{CV})$ for the target parameter $\boldsymbol{X}_j$.
    Calculate $BIC$ of the model and $p$-value of each input parameter.
    Remove the input parameter that has the highest $p$-value.
  **until** *All input parameters in the model have been removed.*
  Save the multiple regression model that shows the minimum $BIC$ as $f^{MR}(\boldsymbol{X}^{CV*})$.
**end**

where $\boldsymbol{X}_{-j}{}^{CV}$ are converted parameters using the curve fitted form $f^{CV}(\cdot)$ that have $X_{-j}$ as inputs, $\boldsymbol{X}^{CV*}$ are the selected parameters, and $f^{MR}(\cdot)$ is the multiple regression model that has $\boldsymbol{X}^{CV*}$ as inputs, which shows the minimum $BIC$.

**Algorithm 4:** Update the originally missing values using multiple regression model

**for** $i \in [1, 2, \cdots, N]$ **do**
  **for** $j \in [1, 2, \cdots, M]$ **do**
    Estimate $\hat{x}_{ij}$ using the multiple regression model $f^{MR}(\boldsymbol{X_i}^{CV*})$, i.e., $\hat{x}_{ij} = f^{MR}(\boldsymbol{X_i}^{CV*})$.
  **end**
  Replace $x_{ij}$ using the estimated value $\hat{x}_{ij}$ if $x_{ij}$ has been filled in Algorithm 2, i.e., $x_{ij} = \hat{x}_{ij}$.
**end**

where $f^{MR}(\boldsymbol{X_i}^{CV*})$ is the multiple regression model that has $\boldsymbol{X_i}^{CV*}$ as inputs in $i$th ship case, and $\hat{x}_{ij}$ is the estimated value of the $j$th parameter in $i$th ship case from Algorithm 3 and 4.

(iii) Minor adjustment: This step is the process of identifying and correcting implausible values, taking into account the normal range of imputed values. Additional information, known as domain or background knowledge, can be integrated into the modeling process from data processing to model development (Rudin and Wagstaff, 2014; Niknafs and Berry, 2017). If one has domain knowledge about a specific variable, one can consider the practical scope of the obtained data. Some values identified as invalid can be newly estimated based on domain knowledge or replaced with the values estimated from previous steps (Refer to Algorithm 5).

**Algorithm 5:** Correct the originally missing values using domain-knowledge

**for** $i \in [1, 2, \cdots, N]$ **do**
  **for** $j \in [1, 2, \cdots, M]$ **do**
    Estimate $\hat{x}_{ij}$ using the domain-knowledge $f^{DM}(\cdot)$, i.e., $\hat{x}_{ij} = f^{DM}(x_{ij})$.
  **end**
  Replace $x_{ij}$ using the estimated value $\hat{x}_{ij}$ if necessary, i.e., $x_{ij} = \hat{x}_{ij}$.
**end**

where $f^{DM}(\cdot)$ is the estimated function based on domain-knowledge, and $\hat{x}_{ij}$ is the estimated value of the $j$th parameter in $i$th ship case from Algorithm 5.

### 2.3. Regressional curve fitting functions

To perform a curve fitting between each parameter, linear, quadratic, cube, power, and logarithmic functions, which are commonly used for data smoothing, have been applied, and they are fitted on the observed data based on the least-squares method. The intercept term of the expression is excluded from those functions so that the predicted value can start at (0, 0) (Eqs. (1)–(5)). Among the curve fitting functions, the most suitable function for the measured values was applied, which was determined based on the $R^2$ value.

$$Linear function: y = a \cdot x \tag{1}$$

$$Quadratic function: y = a \cdot x^2 \tag{2}$$

$$Cubic function: y = a \cdot x^3 \tag{3}$$

$$Power function: y = a \cdot x^b \tag{4}$$

$$Logarithmic function: y = a \cdot \log x \tag{5}$$

where $a$, $b$ are curvilinear coefficients to be estimated for the model. $y$ is the design parameter, and $x$ is the independent variable.

### 2.4. Multiple regression using backward elimination

According to the number of independent variables, using one independent variable is classified as simple regression analysis, and two or more variables are classified as multiple regression analysis. The basic model of multiple linear regression analysis with $M$ independent variables can be expressed as Eq. (6). The method of minimizing residuals by regression formula is to find a regression coefficient that minimizes the sum of the least-squares errors of the data points, such as Eq. (7). The significance of the estimated regression coefficients in a multiple regression model can be analyzed by performing a $t$-test, which determines whether to reject the null hypothesis that each independent variable has nothing to do with the dependent variable (Mark and Goldberg, 2001).

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} +, \ldots, +\beta_j x_{ij} + \epsilon_i \tag{6}$$

$$\hat{\beta} = argmin_\beta \sum_{i=1}^{N} \left( y_i - \beta_0 - \sum_{j=1}^{L} \beta_j x_{ij} \right)^2 \tag{7}$$

where $y_i$ denotes $i$th observed value of dependent variable, $\beta_j$ signifies regression coefficient, $\beta_0$ is intercept term, $\epsilon_i$ is error term, $x_{ij}$ is $i$th observed value of $j$th independent variable, $N$ is the sample size (ship cases), and $L$ is the total number of independent variables in the regression model.

The multiple regression model has the advantage of being able to include all the candidate variables that can affect the dependent variable. However, if the number of independent variables increases in the model, the complexity of the model increases, which may cause more computational cost and errors. If a statistical model fits too close to a particular data set by including more parameters than can be justified by the data, it may fail to predict additional observations reliably (Anderson and Burnham, 2004). To exclude the redundant explanatory variables, algorithms that add or delete variables based on selected criteria can be introduced (Pituch and Stevens, 2015). They are called variable selection methods, and among them the backward elimination method refers to the process of starting with all candidate variables, sequentially removing a variable of which the most statistically insignificant for the model fit. The criterion for determining the significance of each variable is based on a $p$-value and the variable selection process is repeated until all remaining independent variables satisfy a certain threshold such as $AIC$ (Akaike's Information Criterion), $BIC$, or maximum $p$-value (Konishi and Kitagawa, 2008;

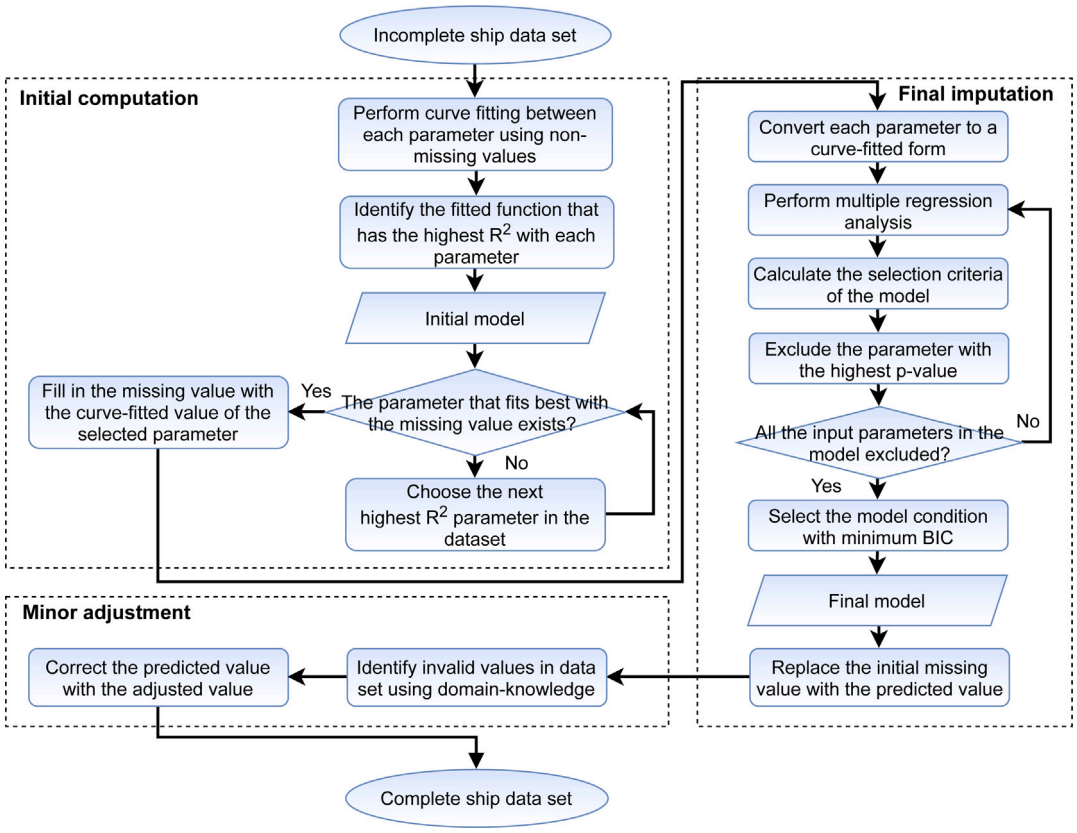


**Fig. 1.** Flowchart of estimating ship principal data considering missing values as proposed in the study.

Montgomery and Runger, 2014). The backward elimination method is most complicated for the initial phase because it contains all candidate variables, but it has the advantage of testing information for all variables. Since the study uses a given data set, backward elimination was applied to be able to test all candidates sequentially as described in the final imputation step.

### 2.5. Experimental evaluation of prediction accuracy

To verify the performance of models for ship principal data, we used error indices such as mean square error ($MSE$), mean absolute error ($MAE$), root mean square error ($RMSE$), coefficient of determination ($R^2$), adjusted coefficient of determination ($Adjusted\ R^2$), Akaike's Information Criterion ($AIC$), and Bayesian Information Criterion ($BIC$) as follows:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{8}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{9}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{10}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{N}\left(y_i - \bar{y}\right)^2} \tag{11}$$

$$Adjusted\ R^2 = 1 - \frac{(1-R^2)\cdot(N-1)}{N-L-1} \tag{12}$$

$$AIC = N\cdot\log(RSS/N) + 2L \tag{13}$$

$$BIC = N\cdot\log(RSS/N) + L\cdot\log(N) \tag{14}$$

where $y_i$ denotes $i$th observed value of dependent variable, $\hat{y}_i$ represents $i$th predicted value of dependent variable, $\bar{y}$ signifies the mean of the observed data, and $RSS$ is the residual sum of squares.

$MSE$ and $MAE$ measure the variance and the average of the residuals, respectively. $RMSE$ is the standard deviation of the prediction errors, which shows a measure of how spread out the residuals are, and the $R^2$ value is based on the proportion of total variation of outcomes explained by the model. In the regression model, as the number of independent variables increases, the $R^2$ value will increase, and as a result, there is a concern that it will be considered the best model (Hair et al., 2018). To compensate for this shortcoming, the $adjusted\ R^2$ value is designed to impose penalties as the number of independent variables increases. Similarly, $AIC$ and $BIC$ serve to select a parsimonious and explainable model by using penalty term for the number of variables and the fitness term of the model. In this study, the $R^2$ value is used as the error index for the curve fitting, and $adjusted\ R^2$ value is used to compare the prediction performance of models in which two or more variables are used. Both $AIC$ and $BIC$ are compared in Section 3.4 for selecting the number of variables when fitting the model.

## 3. Case study

### 3.1. Ship database

IHS Sea-Web database, which contains the following 14 design parameters: auxiliary engine power (AEP), breadth (B), draught (T), deadweight tonnage (DWT), gross tonnage (GT), light displacement

**Table 1**
Descriptive statistics for the principal data of 6,278 container ships from the Sea-Web database.

| Ship principal parameters | Valid data | Missing data | Mean | Median | Std.Dev | Minimum | Maximum | Skewness |
|---|---|---|---|---|---|---|---|---|
| Auxiliary engine power, AEP [kW] | 3892 | 2386 | 1929.6 | 1720.0 | 1177.8 | 50.0 | 5829.0 | 0.66 |
| Breadth, B [m] | 6277 | 1 | 31.6 | 30.2 | 9.8 | 9.5 | 61.5 | 0.63 |
| Draught, T [m] | 6268 | 10 | 11.2 | 11.5 | 3.0 | 1.1 | 16.5 | −0.34 |
| Deadweight tonnage, DWT [t] | 6278 | 0 | 49299.3 | 34577.5 | 44030.0 | 500.0 | 228149.0 | 1.40 |
| Gross tonnage, GT [t] | 6278 | 0 | 43895.2 | 27779.0 | 43506.6 | 355.0 | 232618.0 | 1.68 |
| Light displacement tonnage, LDT [t] | 4559 | 1719 | 15957.6 | 11926.0 | 12202.4 | 358.0 | 66939.0 | 1.30 |
| Length over all, LOA [m] | 6277 | 1 | 221.9 | 208.9 | 80.1 | 48.9 | 400.0 | 0.31 |
| Length between perpendiculars, LBP [m] | 6229 | 49 | 210.8 | 196.6 | 77.0 | 47.5 | 388.1 | 0.34 |
| Main engine cylinder, MEC [-] | 6247 | 31 | 8.1 | 8.0 | 2.0 | 3.0 | 16.0 | 0.85 |
| Main engine power, MEP [kW] | 6273 | 5 | 27620.0 | 21560.0 | 20994.5 | 352.0 | 80905.0 | 0.68 |
| Main engine RPM, MER [-] | 6120 | 158 | 167.9 | 104.0 | 158.0 | 65.0 | 1200.0 | 2.42 |
| Main engine stroke, MES [-] | 6254 | 24 | 2.3 | 2.0 | 0.7 | 2.0 | 4.0 | 1.88 |
| Service speed, V [knot] | 6203 | 75 | 20.7 | 21.0 | 3.5 | 7.5 | 29.2 | −0.60 |
| TEU capacity, TEU [-] | 6254 | 24 | 4073.0 | 2553.5 | 4131.1 | 24.0 | 23756.0 | 1.84 |

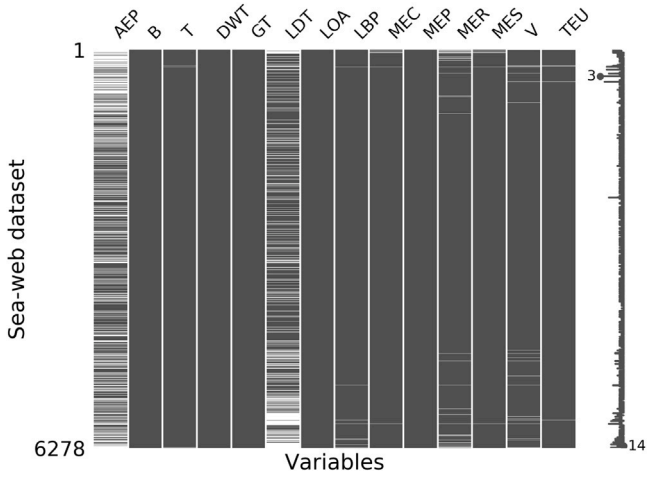


**Fig. 2.** Missing pattern of principal data of container ships collected from the Sea-Web database.

tonnage (LDT), length over all (LOA), length between perpendicular (LBP), main engine cylinder (MEC), main engine power (MEP), main engine RPM (MER), main engine stroke (MES), service speed (V), and container capacity (TEU), was used in this study (IHS, 2019). Here, data analysis and missing data imputation algorithm were mainly performed based on Python programming language. For the case study, the container ship data sets were extracted, consisting of 6,278 vessels from 24 to 23,756 TEU capacity, up until the build year of 2019. The other ship types are not covered in the text, but the results are included in Appendices B and C. Table 1 presents the descriptive statistics of each design parameter, and Fig. 2 visualizes the overall status of missing data. The row of the figure stands for each ship case and the column denotes each parameter. The white cell represents the missing value and the black cell is a non-empty value. The ship data sets are displayed in random order. Among all ship parameters, 38.0%, 27.4%, and 2.5% of the data are missing for the AEP, LDT, and MER. There are also missing data for other parameters, as can be seen from Fig. 2. It should be noted that even if one's data set is different from the data set used in the case study, missing data imputation in ship principal data can be performed according to Algorithms 1–5.

Little's MCAR test is a common method for determining MCAR patterns for missing data in a data set and tests for significant differences between the observed and estimated means for each missing data pattern (Garson, 2015). If the $p$-value of the null hypothesis that the missing data is MCAR is not significant, then the data may be assumed to be MCAR (Little, 1988). Table 2 displays the result of Little's MCAR test against current data sets. It showed that the $p$-value of the null hypothesis is less than 0.05, which means that our data is not missing at random and there may be some sort of a systematic bias included.

**Table 2**
Little's MCAR test for the ship principal data used in the study. Test statistic follows $\chi^2$ distribution asymptotically with degrees of freedom ($df = \sum_{k=1}^{K} M_k - M$) under the null hypothesis that there are no differences between the means of different missing-value patterns. $K$ is the number of missing value patterns among all ship cases, $M_k$ is the number of observed components in pattern $k$, $M$ is the number of ship parameter. $p$-value means the probability that statistics equal to or more extreme than those actually observed in the sample under the assumption that the null hypothesis is correct.

| | $\chi^2$-value | $df$ | $p$-value |
|---|---|---|---|
| Sea-Web database (Container ship) | 4127.053 | 373 | 0.00 |

* Significant at level $p < 0.05$.

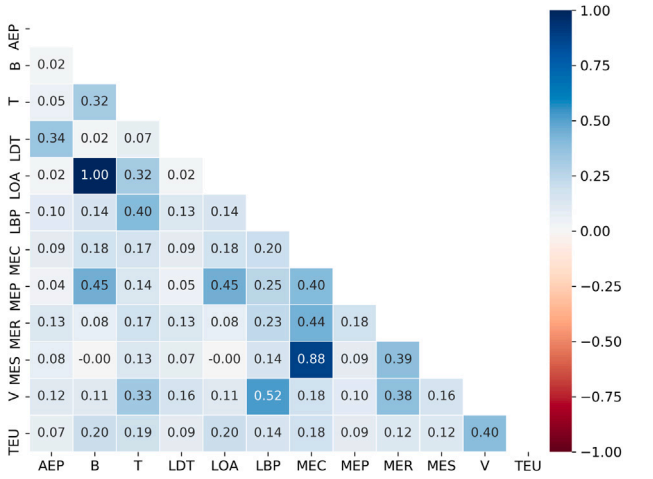


**Fig. 3.** Correlation matrix of missing values.

To achieve additional information about missing characteristics, a correlation analysis between missing and non-missing values for ship variables is performed as depicted in Fig. 3. The notable correlations of missing data are highly correlated variables, such as LOA, B, MES, and MEC. This trend seems to be because the number of missing values is so small that just a few missing values can exaggerate the correlation between the two variables. AEP, LDT, V, and MER, which include relatively higher rate of missing values than other variables, have correlation coefficients in the range of 0.1 to 0.2 with most other variables, indicating that there is almost no correlation between missing values.

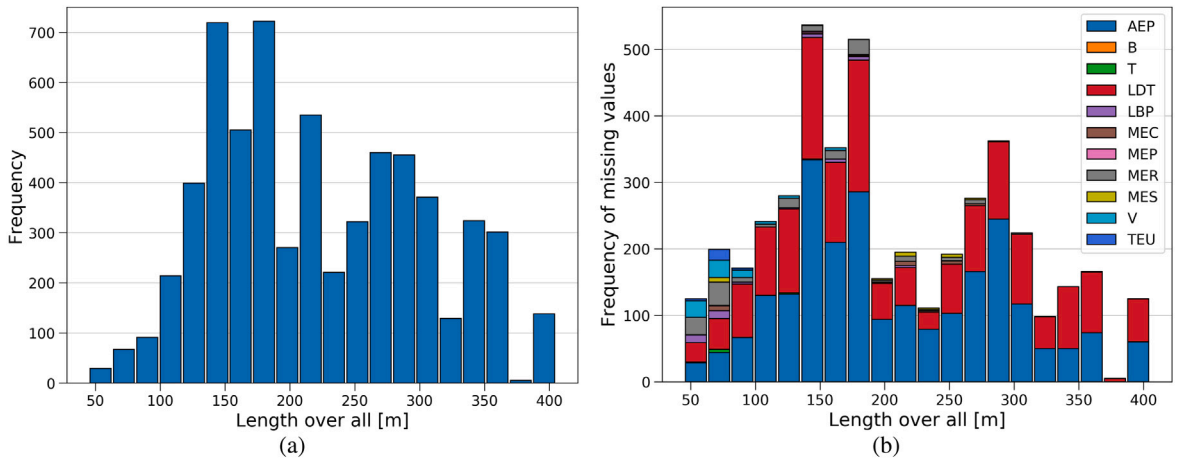


**Fig. 4.** (a)Histogram of the collected ship data by length, (b)Histogram of the cumulative missing values of variables distributed over length.

Fig. 4(a) shows the histogram of container ships by LOA, and Fig. 4(b) reveals the cumulative missing values of variables. Both histograms have similar distributions, which show the shape of the normal distribution centered on 160 meters and 280 m. Fig. 5 denotes the average ratio of missing values against observed data by LOA. While the average missing rates for the most ranges are almost constant at less than 10%, the missing rates for ranges of less than 100 meters are relatively high at 14%–33%. It seems likely that these results are due to the nature of the maritime data, which is generally collected and integrated from various organizations such as ship, owners, shipbuilders, and port authorities.

If a list-wise deletion method that removes missing values in any of the data sets is applied in this case, 46.6% of the total data should be removed, resulting in the inability to use such information and a decrease in statistical power, and one can estimate the biased regression slope. Judging by the nature of the ship principal data, it is inappropriate to apply list-wise deletion or single imputation for the missing values, since the data are not missing completely at random. A model-based computation method using regression analysis intended to be applied in this study does not eliminate missing values but replaces them with plausible values. It also has the advantage of being relatively easy to apply to the ratios and characteristics of various missing data. As discussed above, for ships with length less than 100 meters have a relatively high rate of missing data, the performance of imputation will be checked in a later section.

### 3.2. Correlation between ship principal data

The ship main dimensions and related particulars are determined by various factors such as the cargo volume and weight, and the operational routes required by the ship owner or operator, the strength and stability specified by the rules and regulations of the Classification society, the minimum resistance and friction forces for economic purposes (Papanikolaou, 2014). Moreover, the ship principal parameters such as length, breadth, draught, and height, as well as various other characteristics, are correlated with each other. For instance, for container ships, breadth depends on the row numbers on the deck, thus it is directly related to the number of container capacity on board. An increase in breadth is linked to an increase in cargo capacity and hull resistance, which requires more propulsion power for a ship (Charchalis, 2013). Furthermore, it is important to maintain an appropriate relation between hull length, breadth, draught, and freeboard in terms of securing the ship's stability and integrity (Charchalis and Krefft, 2009). Considering the hull resistance, the wave-making resistance

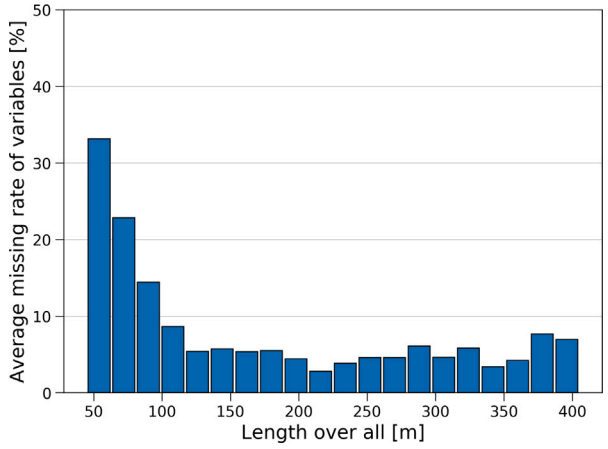


**Fig. 5.** Average missing rate of each variable according to the length over all.

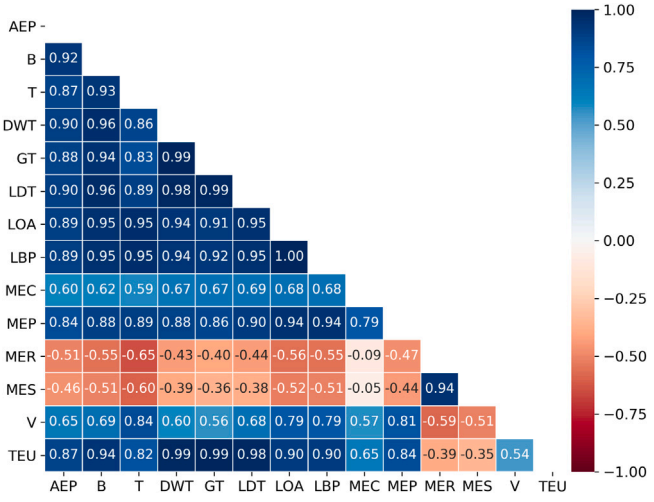


**Fig. 6.** Correlation matrix of non-missing principal particulars of the container ships of the case study.

of the vessel is closely related to the sailing speed and waterline length (Gertler, 1954; Graff, 1964; Tuck, 1987).

Prior to following the procedure of handling the missing data, we have conducted a Pearson correlation analysis between ship main dimensions and related particulars as defined in Eq. (15), and detailed it on the correlation matrix in Fig. 6. The correlation coefficient has been calculated using all data except missing values. As mentioned above, significant correlations are identified between each variable. It can be seen that there is a strong correlation of 0.7 or higher between the volume, weight, cargo quantity, which are composed of L, B, T, and its combination. Variables related to the engine property of the ship such as MEP, MEC, MER, MES, and AEP also have correlation coefficients of 0.3–0.7. Some correlations between other ship principal variables exist.

$$\rho_{X,Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} \tag{15}$$

where $\rho_{X,Y}$ is correlation coefficient between two variables $X$ and $Y$, $E$ is the expected value operator, and $\sigma_X$ and $\sigma_Y$ are standard deviations.

### 3.3. Initial computation

A curve fitting between each variable is performed at the initial computation stage, which is intended to fill in the missing values of the ship data sets and make it possible to implement multiple regression models for variables later. To be specific, it is to take a single variable and function form with the highest goodness of fit for each variable and estimate the missing value using it. Here, it is necessary to define a function that returns the result values of the form shown in Eqs. (1)–(5) for input data and find appropriate unknown coefficients such as 'a' and 'b' for the established function. That is, in order to perform curve fitting on a given data set, optimize.curve_fit of scipy, an open-source Python library, was used in this study (Virtanen et al., 2020). In Fig. 7, the most fitted function among the curve fittings is marked and the degree of fitness ($R^2$ value) is expressed as a heat map. Among the results in Fig. 7, the highest-fitting relationship was extracted for each variable and the final curve fitting results of this step is plotted in Figs. 8(a)–8(n). As can be seen in Fig. 7, the relationship between most parameters is best fitted when applying the power function. This is because the defined power function provides more flexibility than that of the relatively simple functions such as linear, quadratic, cubic, and logarithmic so that the non-linear curves between features can be fitted well. In the given data sets, the relationships of DWT-AEP, GT-B, DWT-T, GT-DWT, TEU-GT, GT-LDT, LBP-LOA, LOA-LBP, MEP-MEC, LOA-MEP, MES-MER, MER-MES, MEP-V, and GT-TEU showed the highest curve fitting results for each other. For GT-DWT, GT-LDT, TEU-GT, and LOA-LBP, the power function was used as the fitted function form, and the exponents of power function were between 0.8 and 1.2, which implies an almost linear relationship.

Regarding the engine factors, it was possible to identify some physical relationships of variables from Figs. 8(a), 8(i), 8(k), 8(l). As the output of the main engine increased, the number of cylinders increased, and as the strokes become from two to four, the rotational speed increased. According to MAN B&W (B&W, 2019), large vessels put a priority on power over speed, so they tend to mount two-stroke engines that have low-speed but good thermal efficiency, low fuel consumption, and high durability. Conversely, the four-stroke engines are mainly installed as a propulsion system for small and medium-sized ships with less than 5,000 kW, or as an auxiliary engine for large ships. It is common to set the rotational speed high to obtain enough power from the auxiliary engine, since the engine is compact and the stroke length can be shortened.

In Fig. 8(m), as the power of the main engine installed on the ship increased, the service speed generally increased logarithmically. In general, main engine power is directly related to maximum speed rather than service speed, but even in the case of the service speed used in this study, it can be seen that $R^2$ value has a prediction accuracy of 0.8387 when it is fitted with a power function. It is noteworthy that some

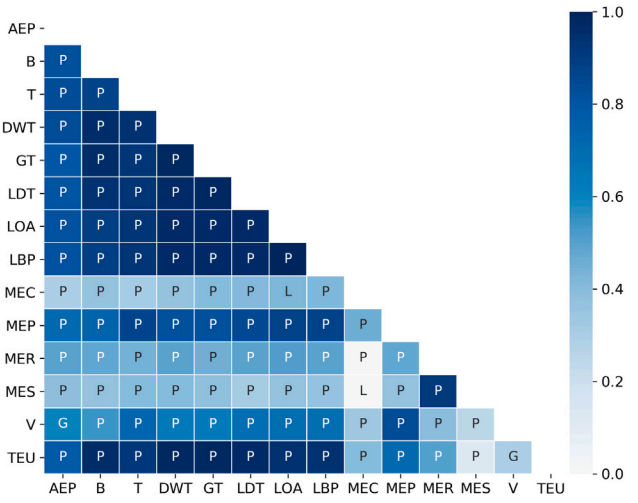


**Fig. 7.** Heat map for curve fitting results of ship principal particulars (L: Linear, Q: Quadratic, C: Cubic, P: Power, G: Logarithmic).

ships with a power of more than 50,000 kW had service speed ranges from 18 to 22 knots, significantly lower than the general trend, and these were found for the case of recently built mega-container ships. Due to rising oil prices in the early 2010s and the adoption of EEDI to new ships to reduce emissions, many shipping companies adopted a slower and more economical voyage speed than previously for their container fleets (Wiesmann, 2010; Meyer et al., 2012). Subsequently, some of the newly built mega container ships were equipped with smaller engines than previous ships of similar size to design slower service speeds (Congress, 2016).

Through this curve fitting process, missing values in the data sets were filled in initially. As explained earlier, the $R^2$ values of most curve fittings were high, but some relationships, such as the main engine cylinder and main engine power showed lower correlations. In addition, there have been instances where the average prediction accuracy of the entire data was good, such as the relationship between the main engine power and service speed, but includes data groups that are out of the general trend. If the $R^2$ value of the resultant model is low or many predicted values deviate from the regression line, it means that a curve-fitted form of a specific single variable is not sufficient to explain the proportion of variance in the dependent variable (Warner, 2020). Therefore, the next section will address the multivariate analysis for explaining more variance.

### 3.4. Final imputation

In this step, multiple regression analysis with backward elimination was performed on the complete data sets obtained from the previous step. According to Mark and Goldberg (2001), the method for testing errors in models produced by stepwise regression is to evaluate the model for data set that are not used to create the model. This method is particularly useful for the case that collects data from different settings or generalizes the model with preventing overfitting. As such, stepwise regression through evaluation criteria such as AIC, BIC, and $p$-value was implemented through statsmodels, a statistical package (Seabold and Perktold, 2010). Therefore, in the entire process of initial computation, final imputation, and minor adjustment, only 80% of the total data set which are randomly selected were used for model implementation and the remaining 20% were only used for the performance evaluation of the final model.

An independent variable with the maximum $p$-value (i.e., the most insignificant variable) was sequentially removed from the model in the backward elimination process, and Fig. 9 shows the maximum $p$-value

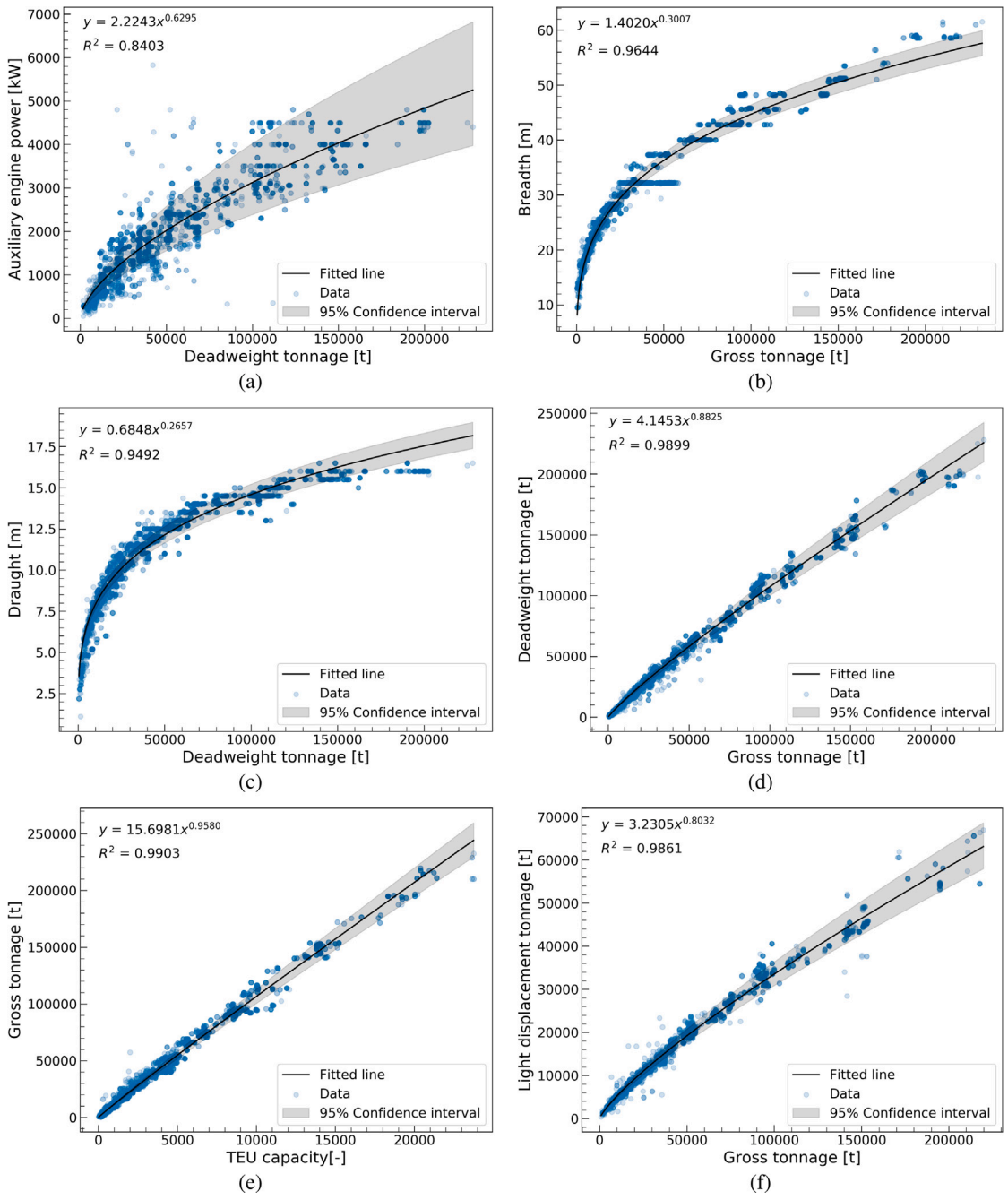


**Fig. 8.** Results of curve fitting for ship main particulars: (a) Auxiliary engine power, (b) Breadth, (c) Draught, (d) Deadweight tonnage, (e) Gross tonnage, (f) Light displacement tonnage, (g) Length over all, (h) Length between perpendicular, (i) Main engine cylinder, (j) Main engine power, (k) Main engine RPM, (l) Main engine stroke, (m) Service speed, (n) TEU capacity.

of independent variables, $AIC$ value, and $BIC$ value of the model according to the number of variables. The minimum values for $AIC$ and $BIC$ are represented by the black edges of the markers. The $AIC$ and $BIC$ include penalty terms for the number of parameters to avoid the possible overfitting problem of the model. Since the $BIC$ puts more penalties for the number of parameters than the $AIC$, fewer variables are selected in the $BIC$ based on the minimum values of $AIC$ and $BIC$, as shown in the figure. Comparing the maximum $p$-values at the minimum points of $AIC$ and $BIC$, some $p$-values for $AIC$ are greater than 0.05, but all $p$-values for $BIC$ are less than 0.05. In general, the significance of independent variables to the dependent variable is based on a $p$-value of 0.05 (Montgomery and Runger, 2014). Thus, the final model was chosen based on a minimum $BIC$ value considering such criterion.

The following Table 3 outlines the results of multiple regression analysis using the backward elimination procedure. Before performing the variable selection process, there is a total of 13 independent variables, and the maximum $p$-value in each regression model is more than 0.05. Through the variable selection process, the $p$-values of all independent variables in the regression model decreased to less than 0.05 by removing the relatively less statistically significant variables. Finally, 7–13 independent variables were selected. The significance of

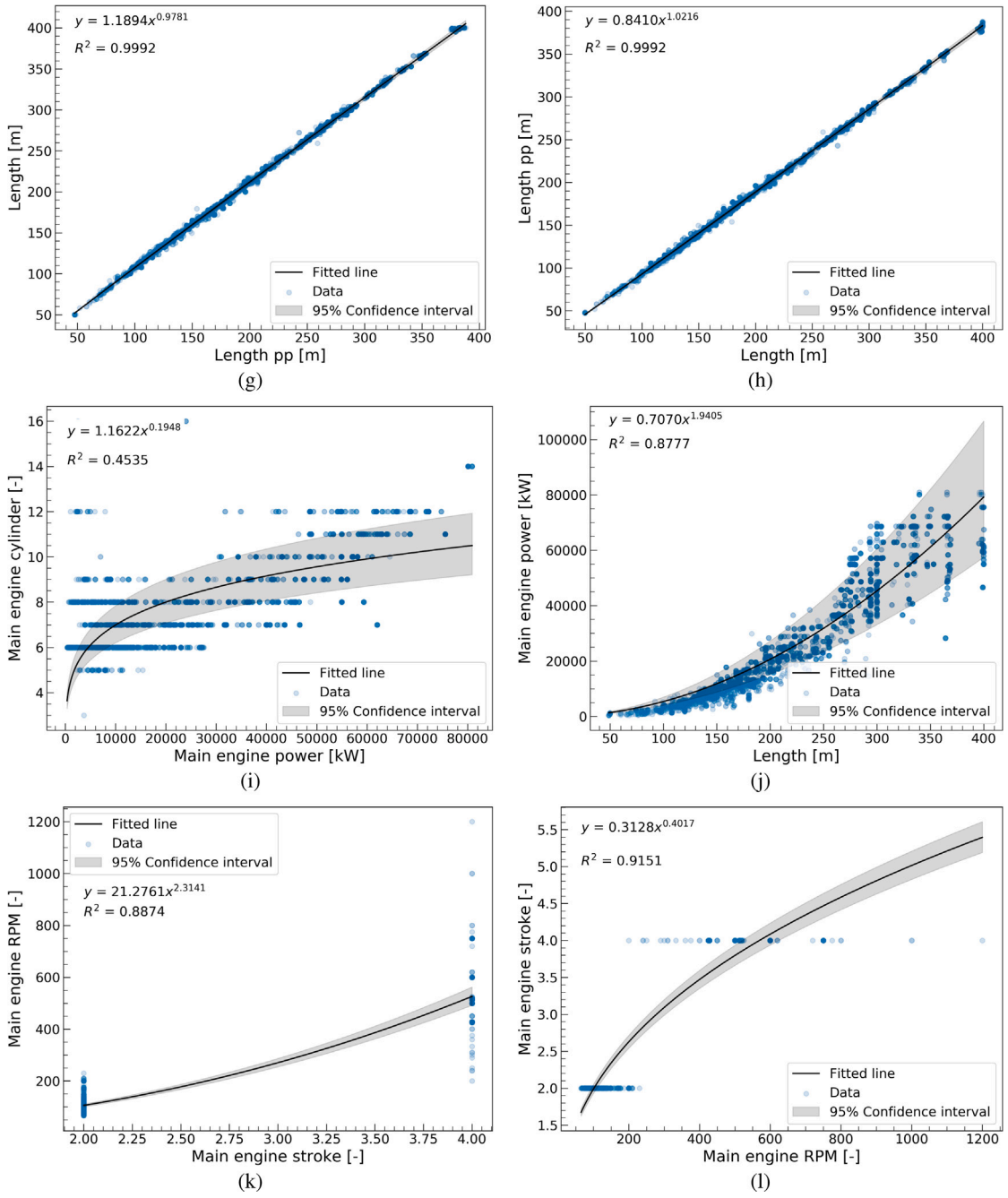


**Fig. 8.** (*continued*).

the final models was evaluated statistically through the $f$-test, and it can be seen that all models were significant at a confidence level of 0.05 as shown in Table 3. Comparing the $adjusted\ R^2$ values of the model for the training data, the $adjusted\ R^2$ was almost maintained even if some variables were removed. This means that the effect of the removed independent variable on the dependent variable is insignificant. Moreover, in the case of independent variables having a similar effect on the dependent variable, unnecessary variables were removed during the backward elimination process, or redundant effects were reduced by adjusting the regression coefficient. For instance, most of the final formulas that require length factor, include only one of LOA or LBP since LOA and LBP have a strong correlation. However, when both LOA and LBP are included in the equations, such as DWT, TEU, V, and T, the redundant effect caused by adding two-length variables at the once is reduced by adjusting the sign and size of the regression coefficients. Another example is LDT, DWT, TEU, and GT related to the overall volume and weight of the ship. The final regression equation of each variable obtained through this process can be found in Appendix C.

### 3.5. Minor adjustment

If one has any prior knowledge of a given variables, we can consider the realistic values based on it. This step corrects the predicted values that are considered inappropriate based on expert judgment. For instance, the main engine stroke is classified into two or four strokes, and the main engine cylinder has to be a positive integer. However,

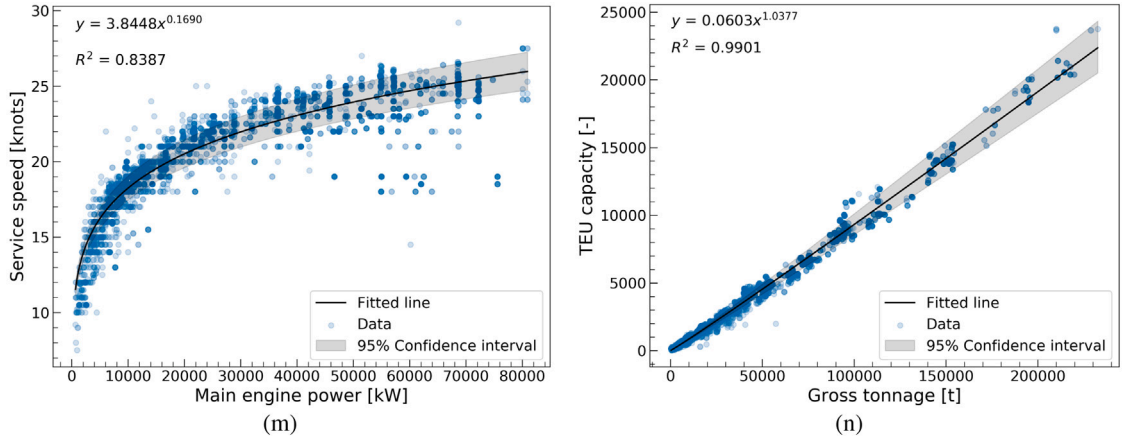


**Fig. 8.** (*continued*).

**Table 3**
The results of multiple regression analysis with backward elimination for training data set of ship principal data. Max $p$-value in the table represents the maximum $p$-value of all independent variables used in the model, and $f$-statistic ($p$-value) means the $f$-test results of the selected model and its $p$-value.

| Ship principal parameters | Full model | | | Selected model | | | |
|---|---|---|---|---|---|---|---|
| | *Adjusted* $R^2$ | Max. $p$-value | No. of inputs | *Adjusted* $R^2$ | Max $p$-value | No. of inputs | $f$-statistic ($p$-value) |
| Auxiliary engine power, AEP [kW] | 0.901 | 0.471 | 13 | 0.901 | 0.000 | 9 | 5064.7 (<0.001) |
| Breadth, B [m] | 0.988 | 0.106 | | 0.988 | 0.002 | 11 | 37,255.1 (<0.001) |
| Draught, T [m] | 0.977 | 0.054 | | 0.977 | 0.000 | 11 | 18,972.8 (<0.001) |
| Deadweight tonnage, DWT [t] | 0.995 | 0.755 | | 0.995 | 0.000 | 12 | 82,946.5 (<0.001) |
| Gross tonnage, GT [t] | 0.996 | 0.236 | | 0.996 | 0.000 | 10 | 153,187.9 (<0.001) |
| Light displacement tonnage, LDT [t] | 0.993 | 0.132 | | 0.993 | 0.000 | 10 | 72,405.7 (<0.001) |
| Length over all, LOA [m] | 0.999 | 0.552 | | 0.999 | 0.001 | 7 | 963,254.6 (<0.001) |
| Length between perpendiculars, LBP [m] | 0.999 | 0.836 | | 0.999 | 0.000 | 8 | 901,284.3 (<0.001) |
| Main engine cylinder, MEC [-] | 0.718 | 0.911 | | 0.718 | 0.000 | 10 | 1,280.3 (<0.001) |
| Main engine power, MEP [kW] | 0.966 | 0.767 | | 0.966 | 0.000 | 11 | 13,099.8 (<0.001) |
| Main engine RPM, MER [-] | 0.921 | 0.728 | | 0.921 | 0.001 | 11 | 5,339.1 (<0.001) |
| Main engine stroke, MES [-] | 0.938 | 0.646 | | 0.938 | 0.000 | 11 | 6,933.3 (<0.001) |
| Service speed, V [knot] | 0.916 | 0.800 | | 0.916 | 0.002 | 11 | 4,984.0 (<0.001) |
| TEU capacity, TEU [-] | 0.993 | 0.001 | | 0.993 | 0.001 | 13 | 63,352.4 (<0.001) |

due to the nature of the regression model, predicted values rarely exist as an integer. Therefore, predicted values of the main engine stroke are adjusted to either 2 or 4 depending on what is closer, and those of the main engine cylinder are rounded off to the nearest positive integer (see Figs. 10(a), 10(b)). In another case, some predicted values may be less than zero due to the intercept term in the predictive model even though the model has outstanding performance generally. Such values are replaced with the curve-fitted value of the initial computation step. The formulas listed in Appendix C are the final imputation results of each variable, and it should be noted that if domain knowledge is applicable (e.g., MEC and MES), the minor adjustment step should be processed for the corresponding result values.

## 4. Results and discussion

### 4.1. Comparison with previous studies

Here, we compare the model proposed in this study with the model developed in the earlier studies and also with the random forest model, which is widely used in the imputation of missing data in machine learning methods. Table 4 summarizes previous studies that established regression equations of main dimensions and related particulars for a container ship. 20% of the total data not used to implement the model was defined as a test data set and the prediction performance of the models listed in Table 3 was evaluated using it. Table 5 shows the results of comparing the prediction performance of the model in this study, the model with the best result among previous studies, and the random forest model against test data set. As an error metrics, $MAE$, $RMSE$, $MSE$, and $adjusted\ R^2$ values defined in Eqs. (8)–(12) were used.

Random forest is an ensemble method that trains a number of decision trees, outperforming in a variety of fields, such as classification and regression of high-dimensional data (Breiman, 2001). Regarding hyperparameters for the random forest model, this study used GridsearchCV of Scikit-Learn library (Kramer, 2016; Bisong, 2019) on the following ranges and took the best subset model among them. (The number of trees = [16-512]; the number of variables in attach split = [3-5]; the other parameters = default of Scikit-Learn library. According to Oshiro et al. (2012), the number of trees at a range between 64 and 128 has shown balanced performance between accuracy, processing time, and memory. In James et al. (2013), the number of variables in each split has been recommended as 1/3 of the number of features. Thus, parameter optimizations have been performed for the range that such values can be included.)

Engine factors such as AEP, MEC, MER, and MES were not covered in the comparison studies. In the case of MEP and V, the $adjusted\ R^2$ has been increased largely from 0.8824 and 0.7578 to 0.9620 and 0.8989, respectively, while $RMSE$ decreased from 7055.58 and 1.70 to 4008.59 and 1.09, respectively. Additionally, since all the $adjusted\ R^2$ and $RMSE$ of other variables have also improved over the previous models, the models proposed in this study are considered to have higher prediction accuracy overall. Since the Sea-Web database used in this study contains a wider range and the number of ships compared to the data sets used in other studies as can be seen from Table 4, the

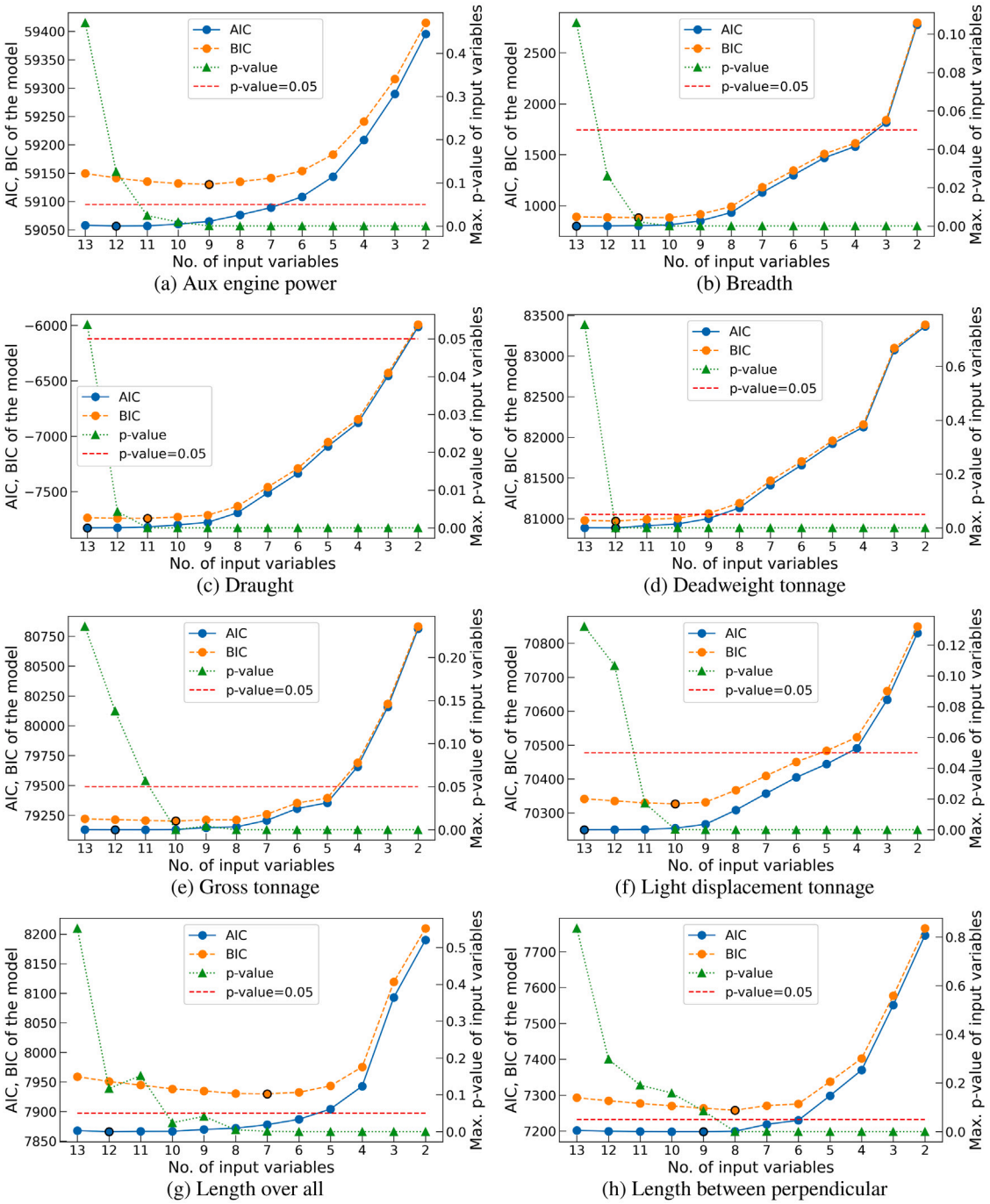


**Fig. 9.** *AIC*, *BIC*, and max. *p*-value according to the decreasing number of independent variables for the model: (a) Auxiliary engine power, (b) Breadth, (c) Draught, (d) Deadweight tonnage, (e) Gross tonnage, (f) Light displacement tonnage, (g) Length over all, (h) Length between perpendicular, (i) Main engine cylinder, (j) Main engine power, (k) Main engine RPM, (l) Main engine stroke, (m) Service speed, (n) TEU capacity.

**Table 4**
Summary of previous studies estimating the main particulars of the container ship using regression analysis.

| Study | Range (TEU) | Build year | No. of ships |
|---|---|---|---|
| Piko (1980) | Abt. 100–3,000 | –1977 | 289 |
| Takahashi et al. (2006) | Abt. 48–8,468 | 1979–2005 | 2,358 |
| Charchalis and Krefft (2009) | Abt. 50–11,000 | – | – |
| Charchalis (2014) | Abt. 1,174–1,388 | – | 17 |
| Kristensen (2016) | Abt. 50–19,500 | 1988–2016 | 2,397 |
| Radfar et al. (2017) | – | 1999–2016 | 985 |
| Abramowski et al. (2018) | Abt. 20–20,000 | 2005–2015 | – |
| Cepowski (2019) | Abt. 90–19,224 | 2000–2018 | 442 |
| This study | Abt. 24–23,756 | 1957–2019 | 6,278 |

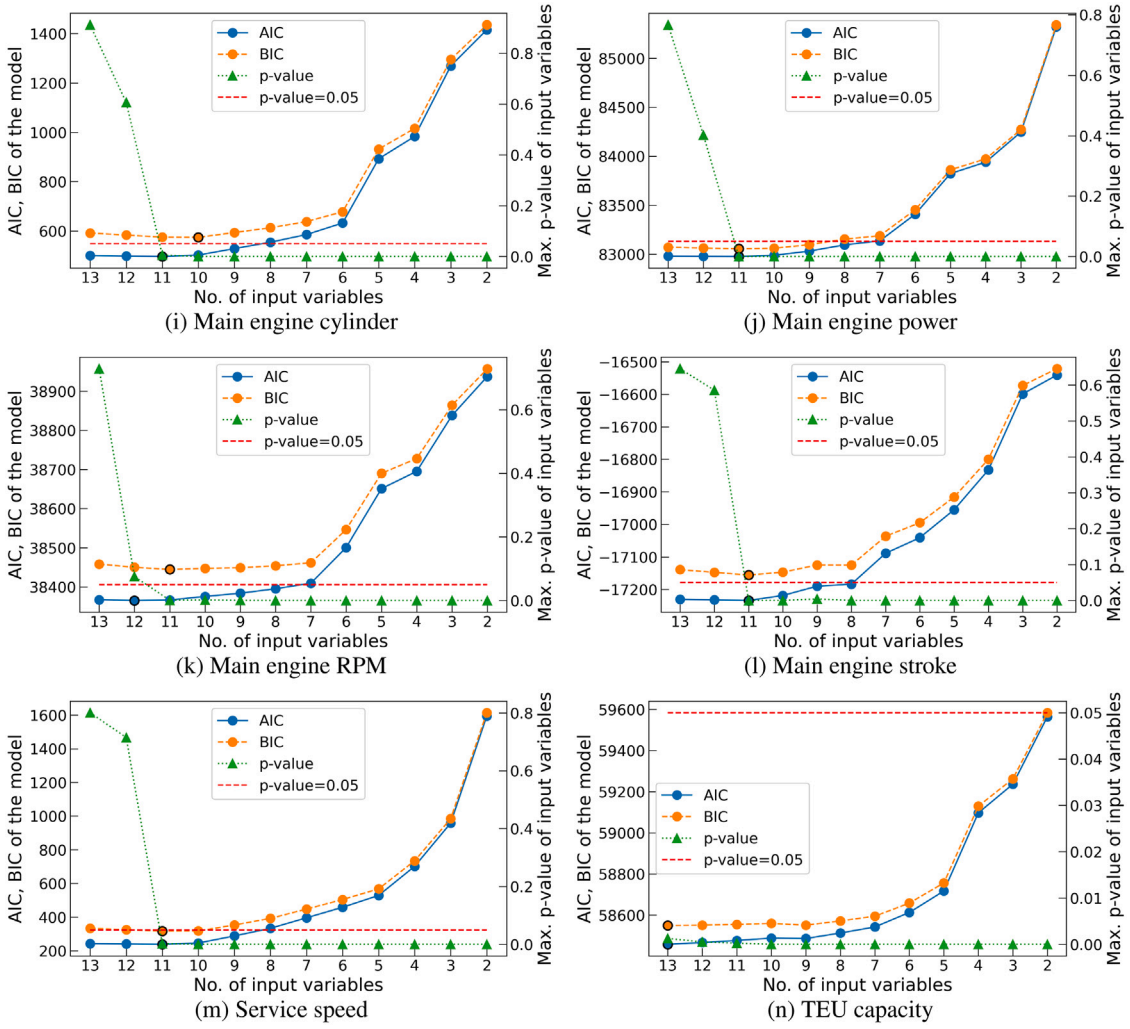


**Fig. 9.** (*continued*).

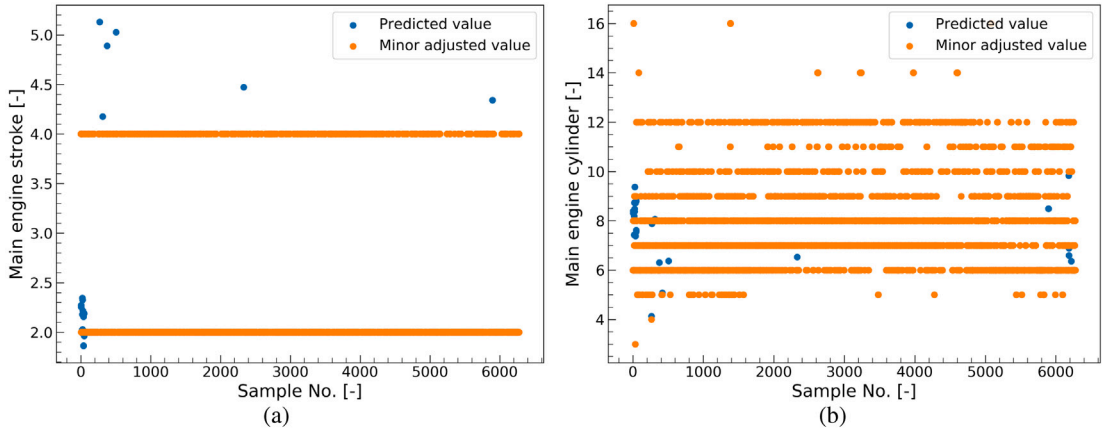


**Fig. 10.** Corrected values through minor adjustment step: (a)Main engine stroke, (b)Main engine cylinder.

scalability of the estimated regression formula is expected to be higher. Some ships in data sets were built before the 1990s, but the influence on the final model is not much because they account for less than 5 percent of the total number of ships. Referring to some examples of the book "Ship design: methodologies of preliminary design" (Papanikolaou, 2014), the predicted results of the proposed algorithm are compared with the method showing the highest accuracy in Table 4 and the simple regression equation from the book in Appendix A. The final

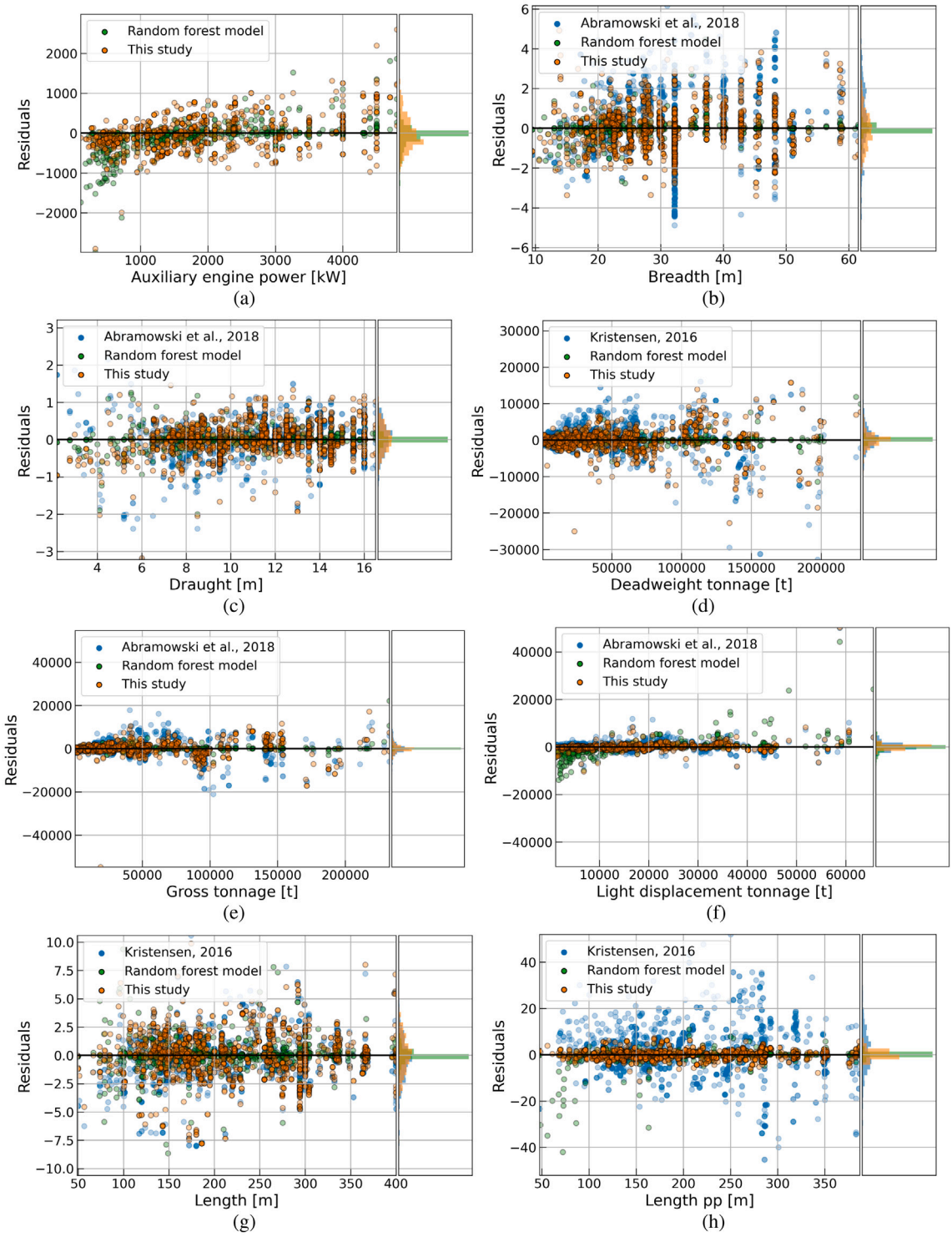


**Fig. 11.** Distribution of residuals for the predicted values between previous studies, random forest model, and this study: (a) Auxiliary engine power, (b) Breadth, (c) Draught, (d) Deadweight tonnage, (e) Gross tonnage, (f) Light displacement tonnage, (g) Length over all, (h) Length between perpendicular, (i) Main engine cylinder, (j) Main engine power, (k) Main engine RPM, (l) Main engine stroke, (m) Service speed, (n) TEU capacity.

equations for different types of ships and their performance are detailed in Appendix C, showing that the model does not only perform well for container ships.

To examine whether the assumptions about the regression model are satisfied, residual analyses were performed (Hair et al., 2018). As can be seen from Figs. 11(a)–11(j), residuals for predicted values are plotted across the range of the variable, and a histogram of residuals is expressed on the right axis. Analyzing the histograms of the current model, they represent a shape close to normal distribution, with no bias to any side around zero, but rather spreading evenly on both sides.

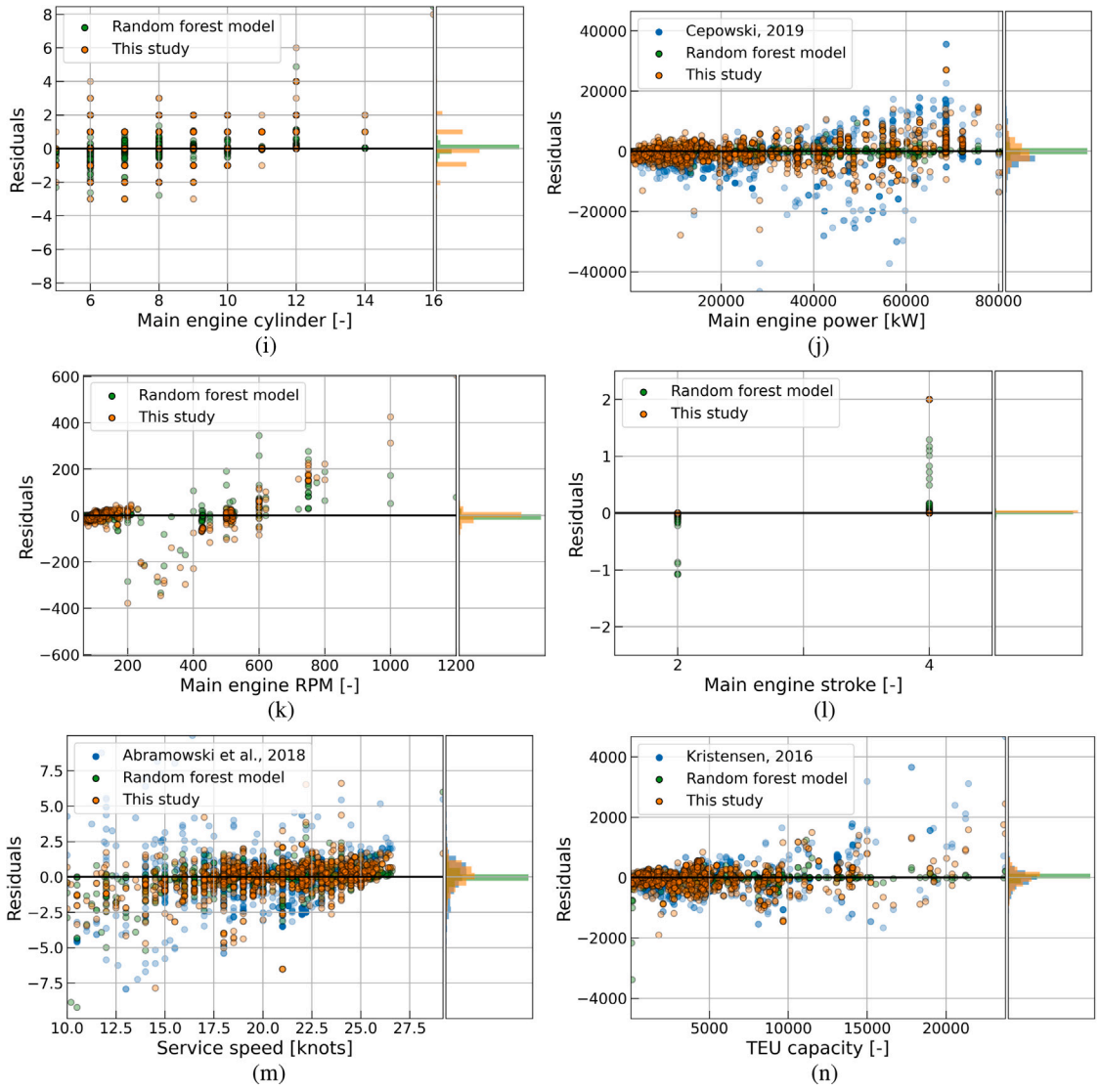


**Fig. 11.** (*continued*).

Moreover, the residuals in the figures do not show a particular pattern and are randomly distributed evenly over the entire range of variables. In particular, the residuals in the range of the ship's length less than 100 m, where the missing points of the collected data was relatively high in this study, are similar to those of other studies. These results provide support for the assumption that the regression models satisfy normality, linearity, and equal variance. Since models of the preceding studies and this study satisfy these assumptions overall and show good accuracy for the test data set, it is expected they will be useful in estimating ship principal data. However, for LBP, B, and V predicted from the previous studies, there are relatively large residuals in some ranges, which one should taken care if using in any analysis.

In the case of LBP, there was a slight discrepancy in the range of 250 to 300 m, according to the residual plot. This is considered to be due to the implementation of the regression equation by dividing the Panamax and Post-Panamax groups by the breadth of 32.2 m. The largest ship that can pass through the Panama Canal is called a Panamax, and is usually designed with 32.2 meters in breadth and 12 meters in loaded draught. However, the maximum width of the old Panama Canal was 32.31 m, and there were some ships having breadth between 32.2 meters and 32.31 meters wide, and 445 vessels were in that range in the current study. Therefore, the ships in the corresponding range were recognized as a Post-Panamax, causing larger residuals. The *adjusted* $R^2$ value of B was 0.9613, with good predictability for the most, while there were relatively large errors at around 32 meters in breadth. This is because the size of the ship was not classified in the previous study and the characteristics of the dimensional constraints were not sufficiently addressed. Moreover, in the range between 10 and 15 knots of service speed, large residuals were observed. This is mainly judged to be a lack of data fitting on feeders of less than 1,000 TEU.

Comparing the random forest model with the developed model from this study, the random forest model shows slightly higher prediction accuracy for most parameters than the current model, as can be seen from Table 5. The random forest model creates as many trees on the subset of the data and combines the output of all the trees, which makes it possible to handle high dimensional data. From these results, a random forest model is also a good method to handle missing data. However, according to the residual distribution of AEP, LDT, LBP, MER, V, and TEU (Figs. 11(a), 11(f), 11(h), 11(k), 11(m), 11(n)), there are some values that have been deviated from the constant residual trends, while the variance of residual distribution is small overall. In this regard, tuning of hyperparameters plays an important role in the performance of random forest models, and sometimes there is a possibility that such problems might occur. Furthermore, the model

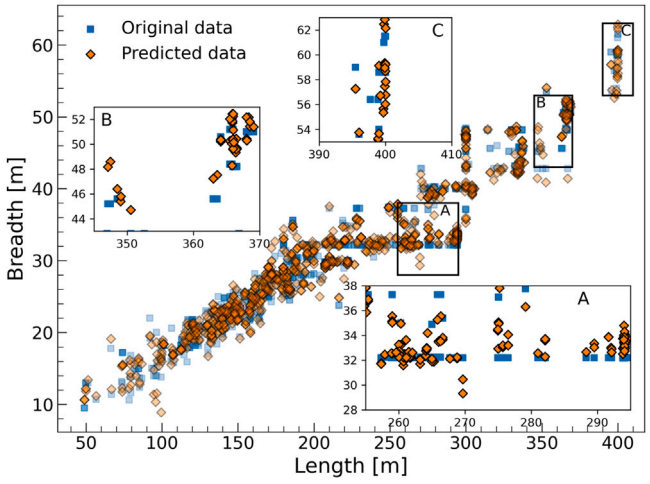


**Fig. 12.** The relation between ship length and breadth regarding dimensional constraints (A: Panama Canal, B: New Panama Canal, C: Suez Canal).

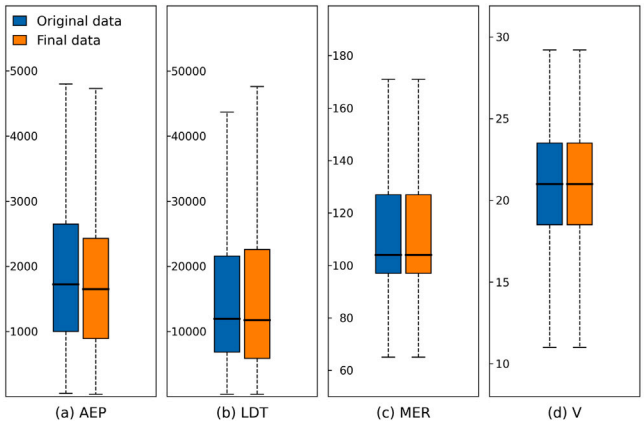


**Fig. 13.** Box plots for descriptive statistics of original data (white boxes) and final data (gray boxes): (a) Auxiliary engine power, (b) Light displacement tonnage, (c) Main engine RPM, (d) Service speed.

training process can be complex and require significant memory storage, as many independent trees are created and different settings of hyperparameters are tuned depending on the characteristics of the data. On the other hand, the model from this study showed consistent residuals across the entire range of data while showing more improved predictive performance over the regression models of previous studies. In addition, the resulting model is intuitive, interpretable, and can be applied easily in other studies. In particular, we believe that this method has sufficient advantages, such as fleet-wide research covered in this paper, that require not only high accuracy of the model but also an overall performance across data ranges.

### 4.2. Validation against dimensional constraints due to operation

Apart from the physical characteristics between ship variables, there are some dimensional restrictions that should be satisfied by the ship not only for navigating certain water areas safely, but entering the terminal and using port facilities (Park and Suh, 2019; Garrido et al., 2020). Table 6 represents the representative dimensions of a container ship for navigating Panama Canal and Suez Canal, which also act as constraints for determining the main dimensions of a ship. If one has domain knowledge for the data set and it is possible to subdivide the samples by clear criteria (e.g., Panamax, Post Panamax vessel) from the initial stage, the implemented model using such a data set may predict the characteristics of corresponding ships more accurately. However, it should be noted that the subdivision of the data set reduces the training sample, which may lead to implementing a model vulnerable to overfitting and outliers.

To confirm the performance of the model against the dimensional constraints for the ship, the breadth and length of container ships are displayed in Fig. 12. According to the 'zoomed in' clusters in boxes A, B, and C in Fig. 12, some clusters are formed around B=32 m, L=366 m, and L=400 m respectively. These clusters seem to represent previous Panamax ships, new Panamax ships, and Suezmax ships. The enlarged plots show that the predicted values of the model satisfy not only the general physical characteristics well, but also the dimensional limitations of canals. Some studies (Takahashi et al., 2006; Kristensen, 2016; Cepowski, 2019) presented in the previous section considered the dimensional constraints by dividing the groups according to the ship size when implementing their models, and it showed higher accuracy than other studies. While this grouping of ships based on domain-knowledge helps improve model accuracy, the detailed grouping reduces the number of data available to implement the model and increases the possibility of overfitting, which may again partly lead to poor performance and low efficiency. In particular, from a missing data processing perspective, the sample size plays an important role in the performance of the resulting model (Heckmann et al., 2014; Hair et al., 2018). Therefore, the estimation method for ship principal data suggested in this study seems to have a novelty in that it shows considerable accuracy without performing grouping and has the benefit of being based on a larger data set than previous studies.

### 4.3. Validation of statistics for the final data

The performance of the missing data imputation was diagnosed by the statistical characteristics of the values superseded. The most missing variables in the database, such as AEP, LDT, MER, and V are shown in Fig. 13 as a box plot. It outlines descriptive statistics of original data and final data to show the statistical characteristics of values filled by this method. In the case of AEP and LDT, the lower quantile (25%) and upper quantile (75%) values vary slightly, but the variation is not large and the mean value is almost maintained. It can be seen that overall statistical values have not changed significantly, even though 38.0% and 27.4% of the data has been replaced. For MER and V, which had relatively fewer missing values than AEP and LDT, the min, max, mean, lower quantile, and upper quantile values are almost maintained.

Fig. 14 compares their imputed values with original values and shows a relationship with the variable having had the highest $R^2$ in the curve fitting to ease the identification of initially missing data. The predicted data are aligned with the trend of the original data in general but it does not exactly lie on the curve fitting line of the initial computation step (Figs. 8(g), 8(k), 8(m), 8(n)). Such variance of replaced values can be interpreted as the effect of the several independent variables affecting the dependent variable is reflected in the model. From these results, it can be seen that the imputed values represent the physical relationship without significantly deviating from the statistical characteristics of the original data.

## 5. Conclusions

This study presented a method of estimating missing values in ship principal data. Data sets of 6,278 container ships from the Sea-Web database were used in a case study, and models for estimating missing values of 14 variables were implemented, including main dimensions such as vessel length, breadth, and draught. In this process, the models were created through curve fitting and variable selection. The selected

**Table 5**
Comparison of prediction performance for the container ship's principal data between previous studies and this study.

| Ship principal parameters | This study | | | | Best result from previous studies | | | | Random forest model | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $MAE$ | $RMSE$ | $MSE$ | $Adj-R^2$ | $MAE$ | $RMSE$ | $MSE$ | $Adj-R^2$ | $MAE$ | $RMSE$ | $MSE$ | $Adj-R^2$ |
| AEP | 341 | 453 | 2.05E+5 | 0.8508 | – | – | – | – | 181 | 352 | 1.24E+5 | 0.9093 |
| B | 0.83 | 1.07 | 1.16 | 0.9875 | 1.49 | 1.90 | 3.61 | 0.9613 | 0.11 | 0.35 | 0.12 | 0.9987 |
| T | 0.32 | 0.43 | 0.19 | 0.9788 | 0.40 | 0.54 | 0.29 | 0.9672 | 0.07 | 0.18 | 0.03 | 0.9963 |
| DWT | 1,877 | 3,141 | 9.87E+6 | 0.9946 | 2,906 | 4,447 | 1.98E+7 | 0.9892 | 355 | 858 | 7.36E+5 | 0.9995 |
| GT | 1,616 | 2,998 | 8.99E+6 | 0.9950 | 2,347 | 3,762 | 1.42E+7 | 0.9921 | 208 | 852 | 3.38E+5 | 0.9995 |
| LDT | 750 | 2,072 | 4.29E+6 | 0.9701 | 1,259 | 2,436 | 5.93E+6 | 0.9591 | 1,051 | 2,855 | 8.15E+6 | 0.9430 |
| LOA | 1.55 | 2.08 | 4.36 | 0.9993 | 1.63 | 2.24 | 5.00 | 0.9992 | 0.47 | 1.66 | 2.76 | 0.9995 |
| LBP | 1.49 | 1.98 | 3.92 | 0.9993 | 8.91 | 12.21 | 149 | 0.9742 | 0.74 | 2.96 | 8.76 | 0.9984 |
| MEC | 0.74 | 1.04 | 1.10 | 0.7062 | – | – | – | – | 0.12 | 0.39 | 0.15 | 0.9594 |
| MEP | 2,760 | 4,009 | 1.61E+7 | 0.9620 | 4,760 | 7,056 | 4.98E+7 | 0.8824 | 378 | 872 | 7.61E+5 | 0.9981 |
| MER | 17.8 | 44.7 | 1,995 | 0.9227 | – | – | – | – | 6.53 | 30.1 | 903 | 0.9649 |
| MES | 0.01 | 0.14 | 0.02 | 0.9609 | – | – | – | – | 0.01 | 0.10 | 0.01 | 0.9823 |
| V | 0.72 | 1.09 | 1.20 | 0.8989 | 1.16 | 1.70 | 2.87 | 0.7578 | 0.30 | 0.73 | 0.53 | 0.9551 |
| TEU | 2,211 | 332 | 1.10E+5 | 0.9931 | 265 | 456 | 2.08E+5 | 0.9872 | 41.0 | 152 | 2.31E+4 | 0.9985 |

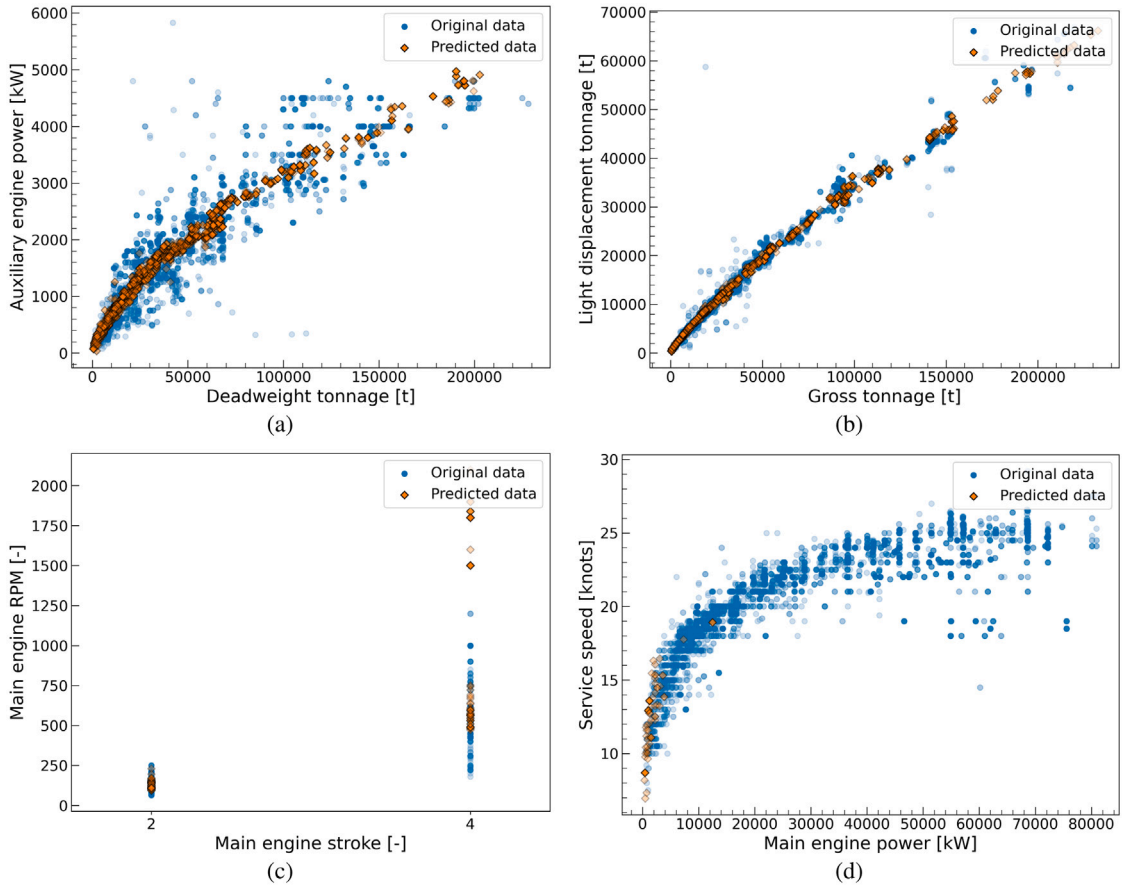


**Fig. 14.** Scatter plots of original data and predicted data: (a)Deadweight tonnage and Auxiliary engine power, (b)Gross tonnage and Light displacement tonnage, (c)Main engine stroke and Main engine RPM, (d)Main engine power and Service speed.

**Table 6**
Dimensional constraints of Panama Canal and Suez Canal for the container ship.

| | Region | B [m] | L [m] | T [m] | DWT [t] | TEU |
|---|---|---|---|---|---|---|
| A | Panama Canal | 32.31 | 294.13 | 12.04 | 52,500 | 5,000 |
| B | Panama Canal (New) | 51.25 | 366.00 | 15.2 | 120,000 | 13,000 |
| C | Suez Canal | 50.00 | 400.00 | 20.1 | | |
| | | 77.50 | 400.00 | 12.2 | | |

variables and the final model were proved to be statistically significant at level 0.05 through $f$-test and $t$-test, respectively. The prediction performance of each model was compared with several regression equations proposed in prior research, and the applicability to the canal passage criteria is verified. Finally, the statistics of complete data were investigated to show consistency with the original data. The main findings of the research are as follows:

- Through correlation analysis and curve fitting, it was found that there are close correlations between many ship principal dimensions and related particulars. Among the fitted results of the variables, pairwise relationships of deadweight tonnage-auxiliary engine power, gross tonnage-breadth, deadweight tonnage-draught, gross tonnage-deadweight tonnage, and cargo quantity-gross tonnage, showed the highest correlations and predictive power with each other.

- As a result of verifying the performance of the model with the test data set, $adjusted\ R^2$ values of the regression equations from earlier works are in the range of 0.7578–0.9992, whilst the ones of this study are 0.8989–0.9993, which shows that there is a significant improvement in the goodness of fit by up to 15.6%. Compared to an ordinary regression model, the presented model illustrates smaller residuals with a constant trend, proving better generality and practicality for estimating ship principal data.
- Comparison of this model with a random forest model, one of the machine learning techniques that is commonly applied for missing data imputation, has been performed. Comparison of this model with a random forest model, one of the machine learning techniques that is commonly applied for missing data imputation, has been performed. The models developed in this study showed slightly lower accuracy than the random forest model but had the advantage of being interpretable, intuitive, and easily applied in other studies.
- Some clusters of ship data were formed around 32 meters in breadth, 366 meters in length, and 400 meters in length, which are the maximum allowable standards of passage through the Suez Canal and Panama Canal. The prediction shows good performance for such dimensional constraints of the ship, even if a detailed classification of data sets is not performed through the model implementation process.
- The statistics for the final values of the auxiliary engine power, light displacement tonnage, main engine RPM, and service speed, which had the most missing values, were identified. The descriptive statistics of completed data sets in this process are almost identical to those of the original data sets and predicted values are aligned with the trend of original values.

Although it is assumed that the proposed algorithm works in different data configurations, in order for this algorithm to function properly, it should be noted that the minimum number of samples is required for statistical analysis methods used in this paper. If there are not enough samples, using the regression equations of previous studies listed in Table 4 or the results of the curve fitting presented in Fig. 8 will probably provide better predictions.

Using the proposed procedure, we were able to properly replace missing values within the ship data sets. The derived regression formulas not only had good predictive power, but reflected physical characteristics and dimensional limitations of ship variables. Therefore, we believe that the methodology suggested in this paper would be applicable from the estimation for the key variables of the ship to the imputation of missing values for data with similar characteristics. In addition, the same principle can be used to replace the erroneous values in the data set with plausible values.

Future research will be to examine the effectiveness of applying data sets of key variables of ships processed in such a manner to the actual marine industry, and it is expected to be able to further improve our current approach. In terms of the accuracy of the model, it is judged that there is still a possibility of improvement as seen through comparison with the random forest model. It will be necessary to further consider advanced machine learning models including the explainable artificial intelligence method from this point of view.

## CRediT authorship contribution statement

**Youngrong Kim:** Investigation, Methodology, Software, Visualization, Writing – original draft. **Sverre Steen:** Methodology, Validation, Writing – review & editing, Supervision. **Helene Muri:** Conceptualization, Resources, Data curation, Writing – review & editing, Supervision.

## Declaration of competing interest

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

## Acknowledgments

This document is the results of the research project “Climate change mitigation in the maritime sector (CLIMMS)”, funded by the Research Council of Norway (grant number 294771), SFI Smart Maritime (RCN grant number 237917), and Norwegian maritime industry. We would like to acknowledge contributions from Anna Ljønes Ringvold, Mario Amin Salgado Delgado and Anders Hammer Strømman on facilitating the work on the database.

## Appendix A. Comparison results of the proposed algorithm and previous studies.

See Fig. A.1.

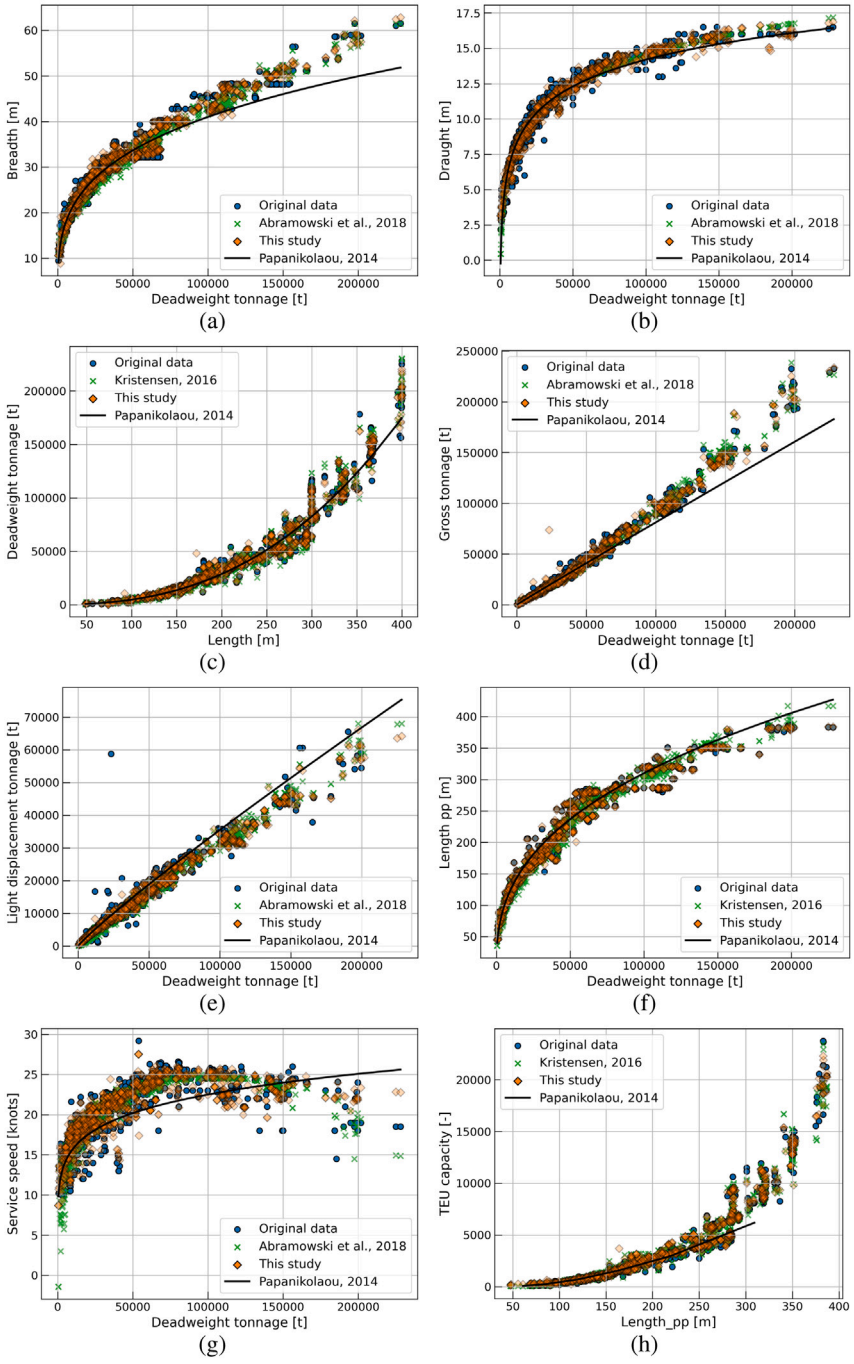


**Fig. A.1.** Comparison plots of the predicted values from the proposed algorithm and previous studies: (a) Breadth, (b) Draught, (c) Deadweight tonnage, (d) Gross tonnage, (e) Light displacement tonnage, (f) Length between perpendicular, (g) Service speed, (h) TEU capacity.

## Appendix B. Prediction of principal data for ship types other than container ships

See Tables B.1–B.3.

**Table B.1**
Sea-Web database used in this study to estimate the principal parameters according to ship type.

| Study | Range | Build year | No. of ships |
|---|---|---|---|
| Bulk carrier | 500–403,880 (DWT) | 1952–2019 | 12,649 |
| Oil tanker | 80–441,585 (DWT) | 1952–2019 | 9,069 |
| Liquefied gas carrier | 140–155,159 (DWT) | 1961–2019 | 2,279 |
| General cargo ship | 20–73,296 (DWT) | 1881–2019 | 16,551 |

**Table B.2**
Comparison of prediction performance for ship principal data of bulk carrier and oil tanker.

| Independent variables | Bulk carrier | | | | Oil tanker | | | |
|---|---|---|---|---|---|---|---|---|
| | $MAE$ | $RMSE$ | $MSE$ | $Adj-R^2$ | $MAE$ | $RMSE$ | $MSE$ | $Adj-R^2$ |
| Auxiliary engine power, AEP [kW] | 80.2 | 113 | 1.27E+4 | 0.6154 | 165 | 263 | 6.90E+4 | 0.6755 |
| Breadth, B [m] | 0.63 | 0.96 | 0.92 | 0.9844 | 0.83 | 1.15 | 1.31 | 0.9954 |
| Draught, T [m] | 0.17 | 0.29 | 0.08 | 0.9909 | 0.32 | 0.44 | 0.19 | 0.9948 |
| Deadweight tonnage, DWT [t] | 1266 | 2057 | 4.2E+6 | 0.9986 | 1301 | 2266 | 5.13E+6 | 0.9994 |
| Gross tonnage, GT [t] | 604 | 1002 | 1.0E+6 | 0.9987 | 814 | 1407 | 1.98E+6 | 0.9991 |
| Light displacement tonnage, LDT [t] | 704 | 1047 | 1.1E+6 | 0.9764 | 750 | 1143 | 1.31E+6 | 0.9929 |
| Length over all, LOA [m] | 1.59 | 2.07 | 4.28 | 0.9980 | 1.31 | 1.87 | 3.49 | 0.9996 |
| Length between perpendiculars, LBP [m] | 1.24 | 1.71 | 2.91 | 0.9986 | 1.25 | 1.80 | 3.26 | 0.9996 |
| Main engine cylinder, MEC [-] | 0.25 | 0.54 | 0.29 | 0.2375 | 0.64 | 1.15 | 1.33 | 0.1633 |
| Main engine power, MEP [kW] | 672 | 994 | 9.89E+5 | 0.9436 | 686 | 1075 | 1.15E+6 | 0.9832 |
| Main engine RPM, MER [-] | 10.9 | 18.4 | 339 | 0.9111 | 162 | 271 | 7.32E+4 | 0.6997 |
| Main engine stroke, MES [-] | 0 | 0.06 | 0 | 0.9766 | 0.07 | 0.36 | 0.13 | 0.8684 |
| Service speed, V [knot] | 0.31 | 0.46 | 0.22 | 0.3896 | 0.62 | 0.87 | 0.76 | 0.8388 |

**Table B.3**
Comparison of prediction performance for ship principal data liquefied gas carrier and general cargo ship.

| Independent variables | Liquefied gas carrier | | | | General cargo ship | | | |
|---|---|---|---|---|---|---|---|---|
| | $MAE$ | $RMSE$ | $MSE$ | $Adj-R^2$ | $MAE$ | $RMSE$ | $MSE$ | $Adj-R^2$ |
| Auxiliary engine power, AEP [kW] | 361 | 496 | 2.46E+5 | 0.8207 | 94.6 | 156 | 2.43E+4 | 0.7335 |
| Breadth, B [m] | 0.67 | 0.88 | 0.78 | 0.9954 | 0.89 | 1.24 | 1.54 | 0.9426 |
| Draught, T [m] | 0.31 | 0.42 | 0.17 | 0.9840 | 0.33 | 0.47 | 0.22 | 0.9560 |
| Deadweight tonnage, DWT [t] | 1640 | 2485 | 6.17E+6 | 0.9949 | 462 | 738 | 5.44E+5 | 0.9892 |
| Gross tonnage, GT [t] | 1683 | 2950 | 8.70E+6 | 0.9955 | 261 | 451 | 2.03E+5 | 0.9917 |
| Light displacement tonnage, LDT [t] | 647 | 1171 | 1.37E+6 | 0.9920 | 314 | 555 | 3.08E+5 | 0.9641 |
| Length over all, LOA [m] | 1.50 | 2.05 | 4.18 | 0.9994 | 1.25 | 1.78 | 3.18 | 0.9976 |
| Length between perpendiculars, LBP [m] | 1.46 | 1.99 | 3.97 | 0.9994 | 1.24 | 1.77 | 3.14 | 0.9973 |
| Main engine cylinder, MEC [-] | 0.64 | 1.03 | 1.06 | 0.6026 | 0.79 | 1.26 | 1.60 | 0.1723 |
| Main engine power, MEP [kW] | 1902 | 2814 | 7.92E+6 | 0.9466 | 341 | 532 | 2.83E+5 | 0.9387 |
| Main engine RPM, MER [-] | 83.3 | 150 | 2.25E+4 | 0.6694 | 202 | 292 | 8.51E+4 | 0.4715 |
| Main engine stroke, MES [-] | 0.07 | 0.37 | 0.14 | 0.8550 | 0.11 | 0.47 | 0.22 | 0.6054 |
| Service speed, V [knot] | 0.76 | 1.09 | 1.19 | 0.8337 | 0.81 | 1.09 | 1.18 | 0.7356 |

## Appendix C. Estimated regression formulas for ship principal parameters

See Tables C.1–C.5

**Table C.1**
Regression coefficients and function forms for ship principal parameters of container ship.

| | Type | Input | | | | | | | | | | | | | | Intercept |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | AEP | B | T | DWT | GT | LDT | LOA | LBP | MEC | MEP | MER | MES | V | TEU | |
| AEP | Form | | P | | P | P | P | P | | P | P | P | P | | | 7.89E+1 |
| | a | | 8.32E−1 | | 1.76E+0 | −1.57E+0 | 9.59E−1 | −7.00E−2 | | −5.69E+0 | 5.11E−1 | 2.38E+6 | −6.84E+2 | | | |
| | b | | 1.77E+0 | | 6.30E−1 | 5.65E−1 | 6.98E−1 | 1.62E+0 | | 1.52E+0 | 6.81E−1 | −2.02E+0 | −1.69E+0 | | | |
| B | Form | | | P | P | P | P | | P | P | P | P | P | P | P | 2.73E+0 |
| | a | | | −4.79E−1 | 4.43E−1 | 1.48E+0 | 1.15E−1 | | −4.11E−1 | 2.56E−1 | 1.31E−1 | 6.86E+1 | −9.88E+0 | −1.34E−2 | 1.04E+0 | |
| | b | | | 1.16E+0 | 3.24E−1 | 3.01E−1 | 3.58E−1 | | 7.99E−1 | 8.03E−1 | 3.19E−1 | −5.23E−1 | −7.43E−1 | 1.30E+0 | 2.98E−1 | |
| T | Form | | P | | P | P | P | P | P | | P | P | P | P | P | −7.34E−2 |
| | a | | −4.10E−1 | | 9.81E−1 | 3.76E−1 | 8.49E−2 | 2.18E−1 | −4.68E−1 | | 1.38E−1 | 2.04E+1 | −3.39E+0 | 2.95E−2 | −2.05E−1 | |
| | b | | 7.80E−1 | | 2.66E−1 | 2.44E−1 | 2.83E−1 | 6.89E−1 | 6.71E−1 | | 2.83E−1 | −4.38E−1 | −7.63E−1 | 1.33E+0 | 2.41E−1 | |
| DWT | Form | P | P | P | | P | P | P | P | P | | P | P | P | P | 6.89E+2 |
| | a | 6.20E−2 | 8.78E−1 | 7.98E−2 | | 1.43E+0 | −2.33E−1 | −4.36E−3 | 1.24E−2 | −1.65E+1 | | −7.48E+8 | 7.13E+3 | −3.23E−2 | 1.10E+1 | |
| | b | 1.30E+0 | 2.62E+0 | 4.61E+0 | | 8.83E−1 | 1.09E+0 | 2.73E+0 | 2.68E+0 | 2.16E+0 | | −2.91E+0 | −2.63E+0 | 3.66E+0 | 8.53E−1 | |

(continued on next page)

**Table C.1** (*continued*).

| | Type | Input | | | | | | | | | | | | | | Intercept |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | AEP | B | T | DWT | GT | LDT | LOA | LBP | MEC | MEP | MER | MES | V | TEU | |
| GT | Form | P | P | | P | | P | P | | P | P | P | P | | P | −1.01E+3 |
| | a | −6.03E−3 | 2.81E−1 | | 5.20E−2 | | 7.53E−2 | 1.53E−4 | | 6.59E+0 | −7.06E−3 | −1.17E+9 | 5.26E+3 | | 2.60E+0 | |
| | b | 1.45E+0 | 2.91E+0 | | 1.13E+0 | | 1.23E+0 | 3.16E+0 | | 2.36E+0 | 1.18E+0 | −3.18E+0 | −2.76E+0 | | 9.58E−1 | |
| LDT | Form | P | P | P | P | P | | | P | | P | P | | P | P | −1.05E+2 |
| | a | 3.18E−2 | 5.15E−1 | 2.46E−2 | −2.78E−1 | 2.17E+0 | | | 9.11E−3 | | 5.21E−2 | 4.21E+7 | | −1.54E−2 | 2.86E+0 | |
| | b | 1.16E+0 | 2.39E+0 | 4.04E+0 | 8.93E−1 | 8.03E−1 | | | 2.39E+0 | | 9.72E−1 | −2.69E+0 | | 4.03E+0 | 7.87E−1 | |
| LOA | Form | P | | P | P | P | | | P | | P | | | P | | 1.53E−1 |
| | a | −2.14E−2 | | 1.33E−1 | −2.49E−1 | 3.45E−1 | | | 1.16E+0 | | 7.07E−2 | | | −5.15E−3 | | |
| | b | 4.73E−1 | | 1.41E+0 | 3.75E−1 | 3.45E−1 | | | 9.78E−1 | | 4.04E−1 | | | 1.84E+0 | | |
| LBP | Form | | P | P | P | | P | P | | | | P | P | P | | 1.19E+0 |
| | a | | −2.81E−1 | −2.40E−1 | 4.48E−1 | | 1.32E−1 | 7.80E−1 | | | | 4.06E+1 | −5.26E+0 | 6.18E−3 | | |
| | b | | 1.09E+0 | 1.45E+0 | 3.83E−1 | | 4.29E−1 | 1.02E+0 | | | | −6.62E−1 | −9.32E−1 | 1.89E+0 | | |
| MEC | Form | P | P | P | P | P | | P | | | P | P | P | | P | −4.80E+0 |
| | a | −3.74E−1 | 1.53E+0 | −1.06E+0 | −1.84E+0 | 1.50E+0 | | 7.00E−1 | | | 1.53E+0 | −7.54E+1 | 5.83E+1 | | −2.00E+0 | |
| | b | 1.97E−1 | 5.00E−1 | 5.23E−1 | 1.61E−1 | 1.56E−1 | | 4.53E−1 | | | 1.95E−1 | −6.59E−2 | −5.64E−2 | | 1.53E−1 | |
| MEP | Form | P | P | P | | P | P | P | | P | | P | P | P | P | −9.27E+3 |
| | a | 1.18E+0 | 1.42E+1 | 3.08E+0 | | −1.52E+1 | 5.40E+0 | 2.64E−1 | | 3.99E+1 | | −1.88E+7 | 4.21E+4 | 4.53E−3 | −2.81E+1 | |
| | b | 9.13E−1 | 1.85E+0 | 3.12E+0 | | 6.25E−1 | 7.89E−1 | 1.94E+0 | | 2.20E+0 | | −1.61E+0 | −2.30E+0 | 4.78E+0 | 6.03E−1 | |
| MER | Form | | P | P | P | P | P | P | P | P | | | P | P | P | 8.80E+1 |
| | a | | 1.44E+4 | 5.18E+2 | −5.12E+3 | −7.24E+3 | 4.72E+3 | 1.21E+5 | −6.62E+4 | −2.48E+2 | | | 1.79E+1 | 1.65E+4 | 5.93E+2 | |
| | b | | −1.89E+0 | −1.26E+0 | −5.58E−1 | −4.79E−1 | −7.95E−1 | −1.69E+0 | −1.65E+0 | −2.53E−1 | | | 2.31E+0 | −2.27E+0 | −5.28E−1 | |
| MES | Form | | P | P | P | P | P | P | | P | P | P | | P | P | −2.15E+0 |
| | a | | −1.13E+1 | −1.90E+0 | 1.02E+1 | 1.11E+1 | −3.81E+0 | −1.59E+1 | | 3.12E+0 | 2.89E+0 | 3.67E−1 | | −3.44E+0 | −3.44E+0 | |
| | b | | −6.25E−1 | −5.46E−1 | −1.91E−1 | −1.74E−1 | −2.24E−1 | −4.92E−1 | | −4.84E−2 | −1.88E−1 | 4.02E−1 | | −8.10E−1 | −1.70E−1 | |
| V | Form | | P | P | P | P | P | P | P | | P | P | P | | G | 4.21E+0 |
| | a | | −1.44E+0 | 3.03E+0 | −5.45E+0 | −1.54E+0 | −1.88E+0 | −2.94E+0 | 4.27E+0 | | 5.59E+0 | 1.36E+1 | −5.02E+0 | | 2.71E+0 | |
| | b | | 3.91E−1 | 5.15E−1 | 1.36E−1 | 1.26E−1 | 1.50E−1 | 3.74E−1 | 3.65E−1 | | 1.69E−1 | −1.96E−1 | −3.80E−1 | | | |
| TEU | Form | P | P | P | P | P | P | P | P | P | P | P | P | P | | 2.40E+2 |
| | a | −2.02E−4 | 1.92E−2 | 7.69E−5 | 2.89E−3 | 1.76E−2 | 1.50E−3 | −6.55E−6 | 2.36E−5 | −1.30E+0 | −1.38E−3 | −1.07E+8 | −7.32E+2 | 1.01E−3 | | |
| | b | 1.49E+0 | 3.02E+0 | 5.64E+0 | 1.17E+0 | 1.04E+0 | 1.27E+0 | 3.31E+0 | 3.25E+0 | 2.32E+0 | 1.18E+0 | −3.31E+0 | −2.69E+0 | 3.73E+0 | | |

* Form: Functional form of each independent variable (L: Linear, Q: Quadratic, C: Cubic, P: Power, G: Logarithmic).
a: Regression coefficient of each independent variable.
b: Exponent number of each independent variable.
Intercept: Intercept term of multiple regression model.

**Table C.2**
Regression coefficients and function forms for ship principal parameters of bulk carrier.

| | Type | Input | | | | | | | | | | | | | Intercept |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | AEP | B | T | DWT | GT | LDT | LOA | LBP | MEC | MEP | MER | MES | V | |
| AEP | Form | | P | P | P | P | P | P | P | P | P | P | P | | −9.96E+1 |
| | a | | 2.69E+1 | 5.93E+1 | −1.29E+2 | 3.83E+1 | 2.03E+1 | −1.51E+1 | 2.19E+1 | 2.54E+2 | 3.13E+0 | 1.18E+3 | −2.31E+2 | | |
| | b | | 8.18E−1 | 7.77E−1 | 2.59E−1 | 2.90E−1 | 3.65E−1 | 8.45E−1 | 8.28E−1 | −1.39E−1 | 4.36E−1 | −5.86E−1 | −7.50E−1 | | |
| B | Form | P | | P | P | P | P | P | P | P | P | P | P | P | −1.61E+3 |
| | a | 5.41E−3 | | −3.06E+0 | 1.40E+0 | 1.01E+0 | 3.55E−2 | 1.46E−1 | −3.17E−1 | 1.61E+3 | 1.09E−2 | 4.06E+1 | −2.19E+0 | −2.01E−3 | |
| | b | 8.52E−1 | | 9.44E−1 | 3.17E−1 | 3.51E−1 | 4.22E−1 | 1.02E+0 | 9.90E−1 | 9.91E−4 | 5.18E−1 | −7.20E−1 | −8.52E−1 | 2.22E+0 | |
| T | Form | P | P | | P | P | P | P | P | P | P | P | P | P | 1.09E+1 |
| | a | 1.48E−3 | −3.23E−1 | | 7.33E−1 | | 3.60E−2 | 2.87E−2 | −7.13E−2 | −1.04E+1 | 7.52E−3 | 1.43E+1 | −6.25E−1 | −4.04E−4 | |
| | b | 8.05E−1 | 9.15E−1 | | 3.18E−1 | 3.46E−1 | 4.13E−1 | 1.02E+0 | 9.97E−1 | −3.46E−2 | 5.16E−1 | −7.46E−1 | −8.77E−1 | 2.38E+0 | |
| DWT | Form | P | P | P | | P | P | P | P | G | | P | P | P | 1.96E+3 |
| | a | −8.92E−4 | 4.07E−1 | 4.34E+0 | | 5.39E−1 | −4.29E−2 | −2.51E−4 | 9.49E−4 | −2.98E+3 | | −1.36E+8 | −3.64E+3 | 2.27E−4 | |
| | b | 2.23E+0 | 2.66E+0 | 3.14E+0 | | 1.07E+0 | 1.27E+0 | 3.23E+0 | 3.19E+0 | | | −2.71E+0 | −2.38E+0 | 5.65E+0 | |
| GT | Form | | P | | P | | P | P | P | | P | P | | P | −3.31E+2 |
| | a | | 7.54E−1 | | 8.54E−1 | | 1.02E−1 | 3.33E−4 | −4.70E−4 | | 2.93E−4 | 4.92E+7 | | −2.21E−4 | |
| | b | | 2.51E+0 | | 9.29E−1 | | 1.19E+0 | 2.99E+0 | 2.95E+0 | | 1.47E+0 | −2.54E+0 | | 5.28E+0 | |
| LDT | Form | P | P | P | P | P | | P | P | P | P | P | P | | 4.71E+3 |
| | a | 4.73E−3 | −6.23E−1 | 1.96E+0 | −3.43E+0 | 4.12E+0 | | −8.42E−3 | 1.66E−2 | −3.15E+3 | 8.94E−3 | −4.69E+6 | −1.43E+3 | | |
| | b | 1.85E+0 | 2.11E+0 | 2.37E+0 | 7.61E−1 | 8.26E−1 | | 2.45E+0 | 2.41E+0 | 1.94E−1 | 1.23E+0 | −2.08E+0 | −1.94E+0 | | |
| LOA | Form | P | P | P | P | P | P | | P | | P | P | P | P | 1.90E−2 |
| | a | −2.54E−2 | 1.31E+0 | 2.44E+0 | −2.18E+0 | −3.83E−1 | −2.47E−1 | | 1.36E+0 | | 8.15E−2 | −9.85E+1 | 3.87E+0 | −7.81E−3 | |
| | b | 7.83E−1 | 8.87E−1 | 9.17E−1 | 3.02E−1 | 3.30E−1 | 3.97E−1 | | 9.74E−1 | | 4.91E−1 | −7.02E−1 | −8.06E−1 | 2.18E+0 | |
| LBP | Form | P | P | P | P | P | P | P | | | P | P | P | P | 2.74E+0 |
| | a | 1.76E−2 | −1.44E+0 | −3.15E+0 | 2.69E+0 | 4.20E−1 | 2.53E−1 | 6.69E−1 | | | −3.12E−2 | 1.25E+2 | −6.25E+0 | 2.56E−3 | |
| | b | 8.07E−1 | 9.08E−1 | 9.44E−1 | 3.10E−1 | 3.39E−1 | 4.08E−1 | 1.02E+0 | | | 5.03E−1 | −7.38E−1 | −8.44E−1 | 2.26E+0 | |
| MEC | Form | P | P | P | P | P | P | P | | | | P | | P | 4.08E+1 |
| | a | 3.54E+0 | −4.83E+1 | −5.37E+1 | 4.73E+1 | 3.98E+1 | 1.87E+1 | −5.85E+1 | | | | 5.83E+0 | | −6.89E+0 | |
| | b | −8.01E−2 | −5.55E−2 | −4.88E−2 | −1.84E−2 | −2.20E−2 | −2.99E−2 | −5.58E−2 | | | | 1.34E−1 | | −5.44E−1 | |
| MEP | Form | P | P | P | | P | P | P | P | P | | P | P | P | −2.72E+2 |
| | a | 5.01E−2 | 4.18E+0 | 2.97E+1 | | 2.32E+0 | 1.70E+0 | 3.47E−1 | −2.73E−1 | −1.56E+3 | | −1.84E+6 | 8.43E+3 | 2.34E−2 | |
| | b | 1.45E+0 | 1.64E+0 | 1.79E+0 | | 6.27E−1 | 7.51E−1 | 1.86E+0 | 1.82E+0 | 1.99E−1 | | −1.33E+0 | −1.73E+0 | 4.19E+0 | |
| MER | Form | P | P | P | | P | | P | P | P | P | | P | P | −3.32E+1 |
| | a | 1.49E+4 | 9.21E+3 | 1.07E+3 | | −8.08E+3 | | −2.72E+5 | 3.35E+5 | 2.46E−2 | −1.76E+4 | | 1.88E+1 | 9.92E+7 | |
| | b | −1.25E+0 | −1.65E+0 | −1.28E+0 | | −5.09E−1 | | −1.56E+0 | −1.52E+0 | 3.58E+0 | −6.63E−1 | | 2.22E+0 | −6.33E+0 | |
| MES | Form | P | P | P | | P | | P | P | P | P | P | | P | 1.06E+0 |
| | a | −1.84E+0 | −6.69E+0 | −4.98E+0 | | 1.30E+1 | | 2.96E+1 | −5.25E+1 | −2.17E−2 | 1.65E+1 | 3.73E−1 | | 7.41E+1 | |
| | b | −3.96E−1 | −4.55E−1 | −3.97E−1 | | −1.53E−1 | | −4.46E−1 | −4.32E−1 | 8.42E−1 | −2.40E−1 | 3.89E−1 | | −2.33E+0 | |
| V | Form | P | P | P | P | | P | P | P | P | P | P | P | | 4.71E+0 |
| | a | −1.02E+0 | −7.76E+0 | −9.98E+0 | 1.79E+1 | | −3.55E+0 | −2.78E+1 | 1.90E+1 | −4.35E+0 | 1.28E+1 | 1.10E+1 | 5.74E+0 | | |
| | b | 8.33E−2 | 9.42E−2 | 9.29E−2 | 3.07E−2 | | 4.07E−2 | 9.76E−2 | 9.56E−2 | −7.77E−2 | 5.55E−2 | −7.94E−2 | −2.18E−1 | | |

* Form: Functional form of each independent variable (L: Linear, Q: Quadratic, C: Cubic, P: Power, G: Logarithmic).
a: Regression coefficient of each independent variable.
b: Exponent number of each independent variable.
Intercept: Intercept term of multiple regression model.

**Table C.3**
Regression coefficients and function forms for ship principal parameters of oil tanker.

| | Type | Input | | | | | | | | | | | | | Intercept |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | AEP | B | T | DWT | GT | LDT | LOA | LBP | MEC | MEP | MER | MES | V | |
| AEP | Form | | | P | P | | P | P | P | | P | P | | P | 1.60E+1 |
| | a | | | 1.47E+1 | −3.83E+1 | | 2.73E+1 | −3.81E+0 | 4.35E+0 | | 7.04E+0 | 4.15E+3 | | −4.74E−3 | |
| | b | | | 9.59E−1 | 3.04E−1 | | 3.86E−1 | 1.03E+0 | 1.01E+0 | | 5.01E−1 | −7.44E−1 | | 3.71E+0 | |
| B | Form | | | P | P | P | P | | P | P | P | P | P | P | 2.63E+1 |
| | a | | | −1.57E+0 | 1.54E+0 | 5.68E−1 | 7.50E−2 | | −8.47E−2 | −2.66E+1 | 2.66E−2 | 9.93E+1 | −8.19E+0 | −3.38E−5 | |
| | b | | | 1.00E+0 | 3.21E−1 | 3.37E−1 | 4.01E−1 | | 1.06E+0 | −2.85E−2 | 5.14E−1 | −7.90E−1 | −1.70E+0 | 3.81E+0 | |
| T | Form | | P | | P | | P | P | P | | P | P | P | P | 3.02E−1 |
| | a | | −2.51E−1 | | 7.43E−1 | | 1.06E−1 | 1.92E−2 | −6.34E−2 | | 1.48E−2 | 4.82E+1 | −3.29E+0 | 3.37E−6 | |
| | b | | 9.74E−1 | | 3.16E−1 | | 3.96E−1 | 1.07E+0 | 1.04E+0 | | 5.10E−1 | −7.69E−1 | −1.71E+0 | 3.84E+0 | |
| DWT | Form | P | P | P | | P | P | P | P | G | | | | | 7.56E+2 |
| | a | −8.46E−2 | 4.37E−1 | 1.51E+1 | | 2.57E−1 | −7.81E−2 | 4.28E−5 | 1.40E−4 | −1.55E+3 | | | | | |
| | b | 1.36E+0 | 2.99E+0 | 2.82E+0 | | 1.06E+0 | 1.22E+0 | 3.49E+0 | 3.41E+0 | | | | | | |
| GT | Form | | P | P | P | | P | P | P | G | | P | P | | −1.32E+3 |
| | a | | 2.60E−1 | 2.99E+0 | 4.71E−1 | | 1.07E−1 | 2.41E−4 | −6.35E−5 | 1.27E+3 | | −6.40E+7 | 1.92E+4 | | |
| | b | | 2.80E+0 | 2.66E+0 | 9.37E−1 | | 1.15E+0 | 3.27E+0 | 3.20E+0 | | | −2.56E+0 | −4.91E+0 | | |
| LDT | Form | P | P | P | P | P | | | P | | P | P | P | P | −1.97E+2 |
| | a | 8.08E−2 | 6.87E−1 | 1.35E+1 | −1.88E+0 | 1.40E+0 | | | 2.38E−3 | | 2.52E−2 | 9.05E+6 | −4.43E+3 | −3.03E−7 | |
| | b | 1.29E+0 | 2.37E+0 | 2.32E+0 | 7.96E−1 | 8.52E−1 | | | 2.72E+0 | | 1.19E+0 | −2.25E+0 | −4.13E+0 | 7.68E+0 | |
| LOA | Form | P | | P | P | P | | | P | | P | P | | P | −2.80E−1 |
| | a | −4.36E−3 | | 6.46E−1 | −6.01E−1 | 3.78E−1 | | | 1.15E+0 | | 4.07E−2 | −3.79E+1 | | 1.03E−4 | |
| | b | 7.66E−1 | | 9.28E−1 | 2.96E−1 | 3.09E−1 | | | 9.78E−1 | | 4.76E−1 | −7.14E−1 | | 3.59E+0 | |
| LBP | Form | P | P | P | P | | P | P | | | P | P | P | P | 5.31E−1 |
| | a | 3.53E−3 | −3.10E−1 | −1.40E+0 | 1.26E+0 | | 1.20E−1 | 7.90E−1 | | | −2.04E−2 | 8.14E+1 | −3.77E+0 | −8.11E−5 | |
| | b | 7.78E−1 | 9.34E−1 | 9.47E−1 | 3.02E−1 | | 3.77E−1 | 1.02E+0 | | | 4.86E−1 | −7.35E−1 | −1.66E+0 | 3.66E+0 | |
| MEC | Form | | P | | | | | P | P | | P | P | P | | 7.57E+1 |
| | a | | −8.59E+1 | | | | | −1.46E+2 | 1.56E+2 | | −2.25E+1 | 1.42E+1 | 5.92E+0 | | |
| | b | | −1.24E−2 | | | | | −1.78E−2 | −1.73E−2 | | −8.31E−3 | 4.43E−2 | 9.10E−2 | | |
| MEP | Form | P | P | P | P | | P | P | P | P | | P | | P | 2.96E+3 |
| | a | 2.77E−1 | 2.94E+0 | 2.59E+1 | −3.65E+0 | | 4.09E+0 | 1.25E−1 | −1.28E−1 | −2.78E+3 | | −5.13E+6 | | 1.06E−5 | |
| | b | 1.18E+0 | 1.82E+0 | 1.83E+0 | 6.08E−1 | | 7.72E−1 | 2.12E+0 | 2.06E+0 | 9.64E−2 | | −1.61E+0 | | 7.01E+0 | |
| MER | Form | P | P | P | | P | | P | | P | P | | P | P | −3.23E+2 |
| | a | 9.42E+3 | −4.63E+3 | 2.88E+3 | | −8.56E+3 | | 7.87E+4 | | 6.76E+0 | −4.34E+3 | | 6.84E+0 | 1.91E+4 | |
| | b | −6.57E−1 | −1.06E+0 | −1.14E+0 | | −3.27E−1 | | −1.05E+0 | | 1.85E+0 | −4.80E−1 | | 2.79E+0 | −2.06E+0 | |
| MES | Form | | P | P | P | P | P | P | | P | | P | | P | −1.21E+0 |
| | a | | 5.24E+0 | −2.97E+0 | 7.29E+0 | 1.05E+1 | −2.64E+0 | −2.40E+1 | | 4.46E−1 | | 4.98E−1 | | 3.78E+0 | |
| | b | | −4.03E−1 | −3.73E−1 | −1.23E−1 | −1.22E−1 | −1.43E−1 | −3.63E−1 | | 3.89E−1 | | 2.40E−1 | | −1.04E+0 | |
| V | Form | P | | P | P | P | P | | G | | G | P | P | | −2.35E−1 |
| | a | −3.35E−1 | | 6.30E+0 | −1.09E+1 | 3.88E+0 | −3.56E+0 | | 2.96E+0 | | 4.16E+0 | 5.13E+0 | 3.41E+0 | | |
| | b | 2.10E−1 | | 2.30E−1 | 7.17E−2 | 7.33E−2 | 8.71E−2 | | | | | −1.54E−1 | −4.26E−1 | | |

* Form: Functional form of each independent variable (L: Linear, Q: Quadratic, C: Cubic, P: Power, G: Logarithmic).
a: Regression coefficient of each independent variable.
b: Exponent number of each independent variable.
Intercept: Intercept term of multiple regression model.

**Table C.4**
Regression coefficients and function forms for ship principal parameters of liquefied gas carrier.

| | Type | Input | | | | | | | | | | | | | Intercept |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | AEP | B | T | DWT | GT | LDT | LOA | LBP | MEC | MEP | MER | MES | V | |
| AEP | Form | | P | P | P | P | P | P | P | P | P | P | P | P | 7.83E+1 |
| | a | | −7.44E−1 | −4.68E−1 | | 2.23E+0 | | | | | 9.73E−1 | 2.12E+5 | −8.80E+2 | | |
| | b | | 1.99E+0 | 3.05E+0 | 7.72E−1 | 6.36E−1 | 8.38E−1 | 2.00E+0 | 1.98E+0 | 1.26E+0 | 7.02E−1 | −1.17E+0 | −7.80E−1 | 4.23E+0 | |
| B | Form | P | | P | P | P | P | P | P | P | P | P | P | P | 1.14E+0 |
| | a | −6.87E−2 | | −3.91E−1 | 4.93E−1 | 1.10E+0 | 1.01E−1 | | −6.15E−2 | | −1.72E−2 | 1.41E+1 | −2.14E+0 | | |
| | b | 4.50E−1 | | 1.34E+0 | 3.43E−1 | 3.10E−1 | 4.10E−1 | 9.71E−1 | 9.49E−1 | 8.10E−1 | 4.08E−1 | −4.27E−1 | −5.55E−1 | 2.47E+0 | |
| T | Form | P | P | | P | P | P | P | P | P | P | P | P | P | 3.59E−1 |
| | a | | −9.49E−1 | | 1.93E+0 | 2.61E−1 | 3.71E−1 | | −3.89E−1 | | | 4.45E+0 | −8.43E−1 | 3.99E−3 | |
| | b | 3.17E−1 | 6.79E−1 | | 2.38E−1 | 2.11E−1 | 2.81E−1 | 6.64E−1 | 6.49E−1 | 5.27E−1 | 2.88E−1 | −3.58E−1 | −5.82E−1 | 1.76E+0 | |
| DWT | Form | P | P | P | | P | P | P | P | P | P | P | P | P | −3.16E+2 |
| | a | | 3.36E+0 | 1.07E−1 | | −5.84E+0 | | | 1.16E−1 | | 1.21E+0 | 6.23E+5 | −4.61E+3 | −2.06E−3 | |
| | b | 9.48E−1 | 2.34E+0 | 4.79E+0 | | 7.49E−1 | 1.00E+0 | 2.41E+0 | 2.38E+0 | 1.55E+0 | 8.14E−1 | −1.29E+0 | −8.24E−1 | 4.88E+0 | |
| GT | Form | P | P | P | P | | P | P | P | P | P | P | P | P | 1.76E+3 |
| | a | 4.20E−1 | 4.76E−1 | 4.03E−3 | −1.51E−2 | | 2.69E−2 | −3.90E−4 | 2.05E−3 | | 1.41E−1 | | −6.39E+3 | 1.46E−4 | |
| | b | 1.11E+0 | 3.05E+0 | 5.29E+0 | 1.24E+0 | | 1.27E+0 | 3.11E+0 | 3.07E+0 | 1.75E+0 | 9.27E−1 | −1.39E+0 | −6.37E−1 | 5.73E+0 | |
| LDT | Form | P | P | P | P | P | | P | P | P | P | P | P | P | 7.33E+2 |
| | a | | 1.23E+0 | 4.57E−2 | | 2.84E+0 | | | 1.39E−2 | −4.39E+1 | 1.38E+0 | | −9.71E+2 | −8.93E−4 | |
| | b | 9.16E−1 | 2.31E+0 | 4.17E+0 | 9.31E−1 | 7.37E−1 | | 2.32E+0 | 2.29E+0 | 1.52E+0 | 8.00E−1 | −1.03E+0 | −7.15E−1 | 4.77E+0 | |
| LOA | Form | P | P | P | P | P | P | | P | P | P | P | P | P | 4.73E−1 |
| | a | | | −7.16E−2 | | 3.27E−1 | | | 1.12E+0 | | 6.00E−2 | | −1.86E+0 | 1.94E−3 | |
| | b | 4.61E−1 | 1.02E+0 | 1.39E+0 | 3.53E−1 | 3.17E−1 | 4.20E−1 | | 9.76E−1 | 8.50E−1 | 4.19E−1 | −4.34E−1 | −5.56E−1 | 2.55E+0 | |
| LBP | Form | P | P | P | P | P | P | P | | P | P | P | P | P | 3.80E−1 |
| | a | | −1.90E−1 | −2.39E−1 | 5.37E−1 | 2.95E−1 | 8.26E−2 | 7.51E−1 | | | −4.20E−2 | 7.99E+0 | | −1.18E−3 | |
| | b | 4.71E−1 | 1.04E+0 | 1.43E+0 | 3.63E−1 | 3.25E−1 | 4.30E−1 | 1.02E+0 | | 8.57E−1 | 4.27E−1 | −4.51E−1 | −5.72E−1 | 2.59E+0 | |
| MEC | Form | P | P | P | P | P | P | P | P | | P | P | P | P | −1.56E+1 |
| | a | | | | | | −4.80E+0 | | | | 6.54E+0 | 5.61E+0 | 2.21E+0 | 5.36E−1 | |
| | b | 7.69E−2 | 1.67E−1 | 1.66E−1 | 5.00E−2 | 5.19E−2 | 6.68E−2 | 1.64E−1 | 1.58E−1 | | 9.57E−2 | 1.13E−1 | 2.52E−1 | 5.87E−1 | |
| MEP | Form | P | P | P | P | P | P | P | P | P | | P | P | P | −4.91E+3 |
| | a | 2.26E+0 | −9.89E−1 | | 9.56E−2 | | 2.94E−1 | 1.30E−2 | −2.25E−2 | 5.84E+1 | | −8.51E+4 | 1.09E+4 | 6.18E−4 | |
| | b | 9.37E−1 | 2.42E+0 | 4.57E+0 | 1.04E+0 | 7.92E−1 | 1.09E+0 | 2.54E+0 | 2.49E+0 | 2.07E+0 | | −5.02E−1 | −3.03E−1 | 5.23E+0 | |
| MER | Form | P | P | P | P | P | P | P | P | P | P | | P | P | −1.13E+2 |
| | a | | 9.01E+3 | 1.97E+3 | −3.78E+3 | −8.50E+3 | | 1.54E+4 | | 1.65E+1 | 2.22E+3 | | 2.22E+1 | | |
| | b | −3.33E−1 | −8.08E−1 | −9.00E−1 | −2.38E−1 | −2.18E−1 | −2.99E−1 | −7.35E−1 | −7.15E−1 | 1.18E+0 | −3.57E−1 | | 2.04E+0 | −1.76E+0 | |
| MES | Form | P | P | P | P | P | P | P | P | P | P | P | | P | −1.24E+0 |
| | a | | −1.16E+1 | −2.69E+0 | 1.59E+1 | 2.61E+1 | −8.53E+0 | −2.93E+1 | | 4.19E−1 | −5.13E+0 | 3.48E−1 | | | |
| | b | −1.67E−1 | −3.58E−1 | −4.13E−1 | −1.15E−1 | −1.06E−1 | −1.40E−1 | −3.31E−1 | −3.24E−1 | 5.27E−1 | −1.44E−1 | 3.53E−1 | | −8.24E−1 | |
| V | Form | P | P | P | P | P | P | P | P | P | P | P | P | | −5.25E+0 |
| | a | | | 2.96E+0 | −5.64E+0 | 5.72E+0 | −2.85E+0 | 7.39E+0 | −7.06E+0 | 1.59E+0 | 3.97E+0 | 6.33E+0 | | | |
| | b | 1.43E−1 | 3.02E−1 | 3.66E−1 | 9.79E−2 | 9.29E−2 | 1.22E−1 | 2.93E−1 | 2.85E−1 | 3.55E−1 | 1.38E−1 | −1.12E−1 | −1.89E−1 | | |

* Form: Functional form of each independent variable (L: Linear, Q: Quadratic, C: Cubic, P: Power, G: Logarithmic).
a: Regression coefficient of each independent variable.
b: Exponent number of each independent variable.
Intercept: Intercept term of multiple regression model.

**Table C.5**
Regression coefficients and function forms for ship principal parameters of general cargo ship.

| | Type | Input | | | | | | | | | | | | | Intercept |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | AEP | B | T | DWT | GT | LDT | LOA | LBP | MEC | MEP | MER | MES | V | |
| AEP | Form | | P | P | P | | P | P | P | | P | P | | | −1.02E+1 |
| | a | | 2.27E−1 | 1.18E+0 | −6.09E−1 | | 3.96E−1 | −6.10E−3 | 1.91E−2 | | 2.46E−1 | 2.11E+3 | | | |
| | b | | 1.94E+0 | 1.82E+0 | 6.05E−1 | | 7.05E−1 | 1.96E+0 | 1.89E+0 | | 8.28E−1 | −9.32E−1 | | | |
| B | Form | P | | P | P | | P | | P | | P | P | | P | −3.79E−1 |
| | a | 1.78E−1 | | −2.08E+0 | 1.54E+0 | | 5.32E−1 | | −1.31E−1 | | 2.61E−1 | 1.79E+1 | | −1.19E−2 | |
| | b | 4.02E−1 | | 7.81E−1 | 2.88E−1 | | 3.07E−1 | | 8.33E−1 | | 3.47E−1 | −3.49E−1 | | 1.43E+0 | |
| T | Form | P | P | | P | P | P | P | P | | P | P | | P | −6.37E−2 |
| | a | 1.34E−2 | −1.38E−1 | | 5.18E−1 | −4.81E−2 | 1.33E−1 | 3.83E−2 | −1.13E−1 | | 4.71E−2 | 9.20E+0 | | 8.35E−3 | |
| | b | 4.57E−1 | 1.04E+0 | | 3.31E−1 | 3.15E−1 | 3.51E−1 | 9.40E−1 | 9.08E−1 | | 4.05E−1 | −4.26E−1 | | 1.76E+0 | |
| DWT | Form | P | P | P | | P | P | P | P | P | P | | P | P | 7.17E+2 |
| | a | −3.48E−1 | 2.11E−1 | 6.61E+0 | | 6.62E−1 | −2.57E−1 | −1.29E−4 | 7.50E−4 | −1.62E+3 | −3.75E−2 | | −6.95E+2 | −1.69E−2 | |
| | b | 1.22E+0 | 3.09E+0 | 3.04E+0 | | 1.04E+0 | 1.12E+0 | 3.27E+0 | 3.16E+0 | −7.31E−1 | 1.17E+0 | | −2.31E+0 | 3.72E+0 | |
| GT | Form | P | P | P | P | | P | P | P | | | | P | P | −2.01E+2 |
| | a | 9.92E−2 | 2.94E−2 | −1.17E+0 | 6.15E−1 | | 3.86E−1 | 4.45E−4 | −7.04E−4 | | | | 1.38E+3 | −1.97E−3 | |
| | b | 1.23E+0 | 2.95E+0 | 2.89E+0 | 9.45E−1 | | 1.10E+0 | 3.14E+0 | 3.02E+0 | | | | −2.32E+0 | 3.91E+0 | |
| LDT | Form | P | P | P | P | P | | | P | | P | | P | | −4.25E+1 |
| | a | 1.34E−1 | 1.30E−1 | 2.97E+0 | −6.96E−1 | 1.42E+0 | | | 1.02E−3 | | 3.54E−2 | | −2.60E+2 | | |
| | b | 1.16E+0 | 2.65E+0 | 2.58E+0 | 8.41E−1 | 8.84E−1 | | | 2.69E+0 | | 1.11E+0 | | −2.09E+0 | | |
| LOA | Form | P | | P | P | P | P | | P | P | P | P | | P | 2.99E+1 |
| | a | −5.44E−2 | | 7.05E−1 | −2.83E−1 | 1.72E−1 | 1.36E−1 | | 1.15E+0 | −3.32E+1 | 8.90E−2 | 1.10E+1 | | 3.04E−2 | |
| | b | 4.35E−1 | | 8.31E−1 | 3.11E−1 | 3.01E−1 | 3.34E−1 | | 9.75E−1 | −1.28E−2 | 3.74E−1 | −3.63E−1 | | 1.58E+0 | |
| LBP | Form | P | P | P | P | | | P | | P | P | | P | P | −2.00E+1 |
| | a | 1.03E−1 | −9.74E−2 | −1.57E+0 | 6.94E−1 | | | 8.24E−1 | | 2.28E+1 | −5.71E−2 | | −1.93E+0 | −2.05E−2 | |
| | b | 4.42E−1 | 1.03E+0 | 8.44E−1 | 3.17E−1 | | | 1.02E+0 | | −1.56E−2 | 3.80E−1 | | −7.16E−1 | 1.59E+0 | |
| MEC | Form | | | P | P | | P | | | | P | P | P | P | −1.26E+2 |
| | a | | | 1.07E+2 | −6.24E+1 | | 1.87E+1 | | | | 3.51E+1 | 4.80E+0 | 1.27E+0 | 1.97E+1 | |
| | b | | | 3.93E−3 | 4.35E−3 | | 8.10E−3 | | | | 1.05E−2 | 1.13E−1 | 1.52E−1 | 2.57E−2 | |
| MEP | Form | P | P | P | P | P | P | P | P | P | | P | | P | 1.28E+3 |
| | a | 2.70E+0 | 1.16E+0 | 1.56E+1 | −1.98E+0 | −9.56E−1 | 1.99E+0 | 7.76E−2 | −1.14E−1 | −1.87E+3 | | −3.26E+4 | | 4.03E−2 | |
| | b | 1.01E+0 | 2.15E+0 | 2.07E+0 | 6.75E−1 | 7.06E−1 | 7.99E−1 | 2.22E+0 | 2.13E+0 | −6.44E−2 | | −1.08E+0 | | 3.86E+0 | |
| MER | Form | P | | P | P | P | P | P | P | P | P | | P | P | −3.45E+2 |
| | a | 1.06E+3 | | 1.70E+3 | −3.86E+3 | −4.34E+3 | 2.15E+3 | 1.82E+4 | −5.04E+3 | 5.51E+1 | −1.24E+3 | | 3.97E+1 | 1.49E+4 | |
| | b | −2.96E−1 | | −6.40E−1 | −1.74E−1 | −1.46E−1 | −1.75E−1 | −5.77E−1 | −5.52E−1 | 1.12E+0 | −2.55E−1 | | 1.60E+0 | −1.65E+0 | |
| MES | Form | | P | | | P | | P | | P | | G | | | −2.09E+0 |
| | a | | 2.73E+0 | | | 8.46E+0 | | −1.23E+1 | | 3.98E−1 | | 1.46E+0 | | | |
| | b | | −2.82E−1 | | | −7.29E−2 | | −2.04E−1 | | 2.28E−1 | | | | | |
| V | Form | | P | P | P | P | P | P | P | P | P | P | | | −8.94E+0 |
| | a | | −6.29E−1 | 3.86E+0 | −5.14E+0 | −1.38E+0 | 1.34E+0 | 3.44E+0 | −2.44E+0 | 8.64E+0 | 3.88E+0 | 3.24E+0 | | | |
| | b | | 3.64E−1 | 3.27E−1 | 1.06E−1 | 1.00E−1 | 1.16E−1 | 3.17E−1 | 3.05E−1 | 2.84E−2 | 1.58E−1 | −1.46E−1 | | | |

* Form: Functional form of each independent variable (L: Linear, Q: Quadratic, C: Cubic, P: Power, G: Logarithmic).
a: Regression coefficient of each independent variable.
b: Exponent number of each independent variable.
Intercept: Intercept term of multiple regression model.

## References


Abramowski, T., 2013. Application of artificial intelligence methods to preliminary design of ships and ship performance optimization. Nav. Eng. J. 125 (3), 101–112.

Abramowski, T., Cepowski, T., Zvolenskỳ, P., 2018. Determination of regression formulas for key design characteristics of container ships at preliminary design stage. New Trends Prod. Eng. 1 (1), 247–257.

Afrifa-Yamoah, E., Mueller, U.A., Taylor, S., Fisher, A., 2020. Missing data imputation of high-resolution temporal climate time series data. Meteorol. Appl. 27 (1), e1873.

Anderson, D., Burnham, K., 2004. Model Selection and Multi-model Inference, Vol. 63, Second ed. (2020), Springer-Verlag, NY, p. 10.

Andiojaya, A., Demirhan, H., 2019. A bagging algorithm for the imputation of missing values in time series. Expert Syst. Appl. 129, 10–26.

Bisong, E., 2019. More supervised machine learning techniques with scikit-learn. In: Building Machine Learning and Deep Learning Models on Google Cloud Platform. Springer, pp. 287–308.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.

B&W, M., 2019. Propulsion trends in container vessels. URL www.man-es.com.

Cepowski, T., 2019. Regression formulas for the estimation of engine total power for tankers, container ships and bulk carriers on the basis of cargo capacity and design speed. Pol. Marit. Res..

Charchalis, A., 2013. Dimensional constraints in ship design. J. KONES 20.

Charchalis, A., 2014. Determination of main dimensions and estimation of propulsion power of a ship. J. KONES 21.

Charchalis, A., Krefft, J., 2009. Main dimensions selection methodology of the container vessels in the preliminary stage. J. KONES 16, 71–78.

Cheliotis, M., Gkerekos, C., Lazakis, I., Theotokatos, G., 2019. A novel data condition and performance hybrid imputation method for energy efficient operations of marine systems. Ocean Eng. 188, 106220.

Congress, G.C., 2016. New mærsk triple-e ships world's largest and most efficient; waste heat recovery and ultra long stroke engines contribute to up to 50% reduction in CO 2/container moved.

Dobrkovic, A., Iacob, M.-E., van Hillegersberg, J., 2018. Maritime pattern extraction and route reconstruction from incomplete AIS data. Int. J. Data Sci. Anal. 5 (2–3), 111–136.

Garrido, J., Saurí, S., Marrero, A., Gül, U., Rúa, C., 2020. Predicting the future capacity and dimensions of container ships. Transp. Res. Rec. 0361198120927395.

Garson, G.D., 2015. Missing values analysis and data imputation. Statistical Associates Publishers, Asheboro, NC.

Gertler, M., 1954. A reanalysis of the original test data for the Taylor Standard Series. Technical Report, DAVID TAYLOR MODEL BASIN WASHINGTON DC.

Gkerekos, C., Lazakis, I., 2020. A novel, data-driven heuristic framework for vessel weather routing. Ocean Eng. 197, 106887.

Graff, W., 1964. Some extensions of DW taylor's standard series. In: Versuch-sanstalt FÜR Binnenschiffbau, Duisburg, Germany, Technische Hochschule Aachen, Forschungsberichte Des Landes Nordrhein Westfalen, Presented At: The Annual Meeting of the Society of Naval Architects and Marine Engineers, SNAME Transactions 1964, New York, USA, Paper: T1964-1 Proceedings.

Gurgen, S., Altin, I., Ozkok, M., 2018. Prediction of main particulars of a chemical tanker at preliminary ship design using artificial neural network. Ships Offshore Struct. 13 (5), 459–465.

Gutierrez-Torre, A., Berral, J.L., Buchaca, D., Guevara, M., Soret, A., Carrera, D., 2020. Improving maritime traffic emission estimations on missing data with CRBMs. Eng. Appl. Artif. Intell. 94, 103793.

Hair, J., Babin, B., Anderson, R., Black, W., 2018. Multivariate data analysis (8. bs.). Cengage Learning, Harlow.

Heckmann, T., Gegg, K., Gegg, A., Becht, M., 2014. Sample size matters: investigating the effect of sample size on a logistic regression susceptibility model for debris flows. Nat. Hazards Earth Syst. Sci. 14 (2), 259.

IHS, 2019. Sea-web ships. URL https://maritime.ihs.com.

Imtiaz, S., Shah, S., 2008. Treatment of missing values in process data analysis. Can. J. Chem. Eng. 86 (5), 838–858.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. Tree-based methods. In: An Introduction to Statistical Learning. Springer, pp. 303–335.

Jung, S., Moon, J., Park, S., Rho, S., Baik, S.W., Hwang, E., 2020. Bagging ensemble of multilayer perceptrons for missing electricity consumption data imputation. Sensors 20 (6), 1772.

Khatibisepehr, S., Huang, B., Khare, S., 2013. Design of inferential sensors in the process industry: A review of Bayesian methods. J. Process Control 23 (10), 1575–1596.

Kim, M., Park, S., Lee, J., Joo, Y., Choi, J.K., 2017. Learning-based adaptive imputation methodwith kNN algorithm for missing power data. Energies 10 (10), 1668.

Konishi, S., Kitagawa, G., 2008. Information Criteria and Statistical Modeling. Springer Science & Business Media.

Kramer, O., 2016. Scikit-learn. In: Machine Learning for Evolution Strategies. Springer, pp. 45–53.

Kristensen, H.O., 2012. Determination of regression formulas for main dimensions of tankers and bulk carriers based on IHS fairplay data. Clean Shipp. Curr. 1 (6).

Kristensen, H.O., 2013. Statistical analysis and determination of regression formulas for main dimensions of container ships based on IHS fairplay data. In: Statistical Analysis and Determination of Regression Formulas for Main Dimensions of Container Ships Based on IHS Fairplay Data. University of Southern Denmark, Denmark.

Kristensen, H.O., 2016. Revision of statistical analysis and determination of regression formulas for main dimensions of container ships based on data from clarkson. Technical Report, Technical University of Denmark & HOK Marineconsult ApS.

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

Lin, W.-C., Tsai, C.-F., 2020. Missing value imputation: a review and analysis of the literature (2006–2017). Artif. Intell. Rev. 53 (2), 1487–1509.

Little, R.J., 1988. A test of missing completely at random for multivariate data with missing values. J. Amer. Statist. Assoc. 83 (404), 1198–1202.

Little, R.J., Rubin, D.B., 2019. Statistical Analysis with Missing Data, Vol. 793. John Wiley and Sons.

Liu, C., Chen, X., 2013. Inference of single vessel behaviour with incomplete satellite-based AIS data. J. Navig. 66 (6), 813.

Lloyd's list intelligence, 2017. Understanding AIS (automatic identification system). URL https://maritimeintelligence.informa.com/.

Mao, S., Tu, E., Zhang, G., Rachmawati, L., Rajabally, E., Huang, G.-B., 2018. An automatic identification system (AIS) database for maritime trajectory prediction and data mining. In: Proceedings of ELM-2016. Springer, pp. 241–257.

Mark, J., Goldberg, M.A., 2001. Multiple regression analysis and mass assessment: A review of the issues. Apprais. J. 56 (1), 89–109.

Meyer, J., Stahlbock, R., Voß, S., 2012. Slow steaming in container shipping. In: 2012 45th Hawaii International Conference on System Sciences. IEEE, pp. 1306–1314.

Montgomery, D.C., Runger, G.C., 2014. Applied Statistics and Probability for Engineers. Wiley.

Niknafs, A., Berry, D., 2017. The impact of domain knowledge on the effectiveness of requirements engineering activities. Empir. Softw. Eng. 22 (1), 80–133.

Oshiro, T.M., Perez, P.S., Baranauskas, J.A., 2012. How many trees in a random forest? In: International Workshop on Machine Learning and Data Mining in Pattern Recognition. Springer, pp. 154–168.

Papanikolaou, A., 2014. Ship Design: Methodologies of Preliminary Design. Springer.

Park, N.K., Suh, S.C., 2019. Tendency toward mega containerships and the constraints of container terminals. J. Mar. Sci. Eng. 7 (5), 131.

Piko, G., 1980. Regression Analysis of Ship Characteristics. Australian Government Publishing Service.

Pituch, K.A., Stevens, J.P., 2015. Applied Multivariate Statistics for the Social Sciences: Analyses with SAS and IBM's SPSS. Routledge.

Radfar, S., Taherkhani, A., Panahi, R., 2017. Standardization of the main dimensions of design container ships in ports—A case study. World J. Eng. Technol. 5 (4), 51–61.

Rubin, D.B., 1976. Inference and missing data. Biometrika 63 (3), 581–592.

Rudin, C., Wagstaff, K.L., 2014. Machine Learning for Science and Society. Springer.

Seabold, S., Perktold, J., 2010. Statsmodels: Econometric and statistical modeling with python. In: Proceedings of the 9th Python in Science Conference, Vol. 57. Austin, TX, p. 61.

Takahashi, H., Goto, A., Abe, M., Kannami, Y., Esaki, T., Mizukami, J., 2006. Study on ship dimensions by statistical analysis: Standard of main dimensions of design ship (Draft). National Inst. for Land and Infrastructure Management, Ministry of Land ....

Tsitsilonis, K.-M., Theotokatos, G., 2018. A novel systematic methodology for ship propulsion engines energy management. J. Cleaner Prod. 204, 212–236.

Tuck, E.O., 1987. Wave resistance of thin ships and catamarans. Appl. Math. Rep. T8701.

Velasco-Gallego, C., Lazakis, I., 2020. Real-time data-driven missing data imputation for short-term sensor data of marine systems. A comparative study. Ocean Eng. 218, 108261.

Velasco-Gallego, C., Lazakis, I., 2021. A novel framework for imputing large gaps of missing values from time series sensor data of marine machinery systems. Ships Offshore Struct. 1–10.

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S.J., Brett, M., Wilson, J., Millman, K.J., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E., Carey, C.J., Polat, I., Feng, Y., Moore, E.W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E.A., Harris, C.R., Archibald, A.M., Ribeiro, A.H., Pedregosa, F., van Mulbregt, P., SciPy 1.0 Contributors, 2020. SciPy 1.0: Fundamental algorithms for scientific computing in python. Nature Methods 17, 261–272. http://dx.doi.org/10.1038/s41592-019-0686-2.

Wang, M.-C., Tsai, C.-F., Lin, W.-C., 2021. Towards missing electric power data imputation for energy management systems. Expert Syst. Appl. 174, 114743.

Wang, H., Zhuge, X., Strazdins, G., Wei, Z., Li, G., Zhang, H., 2016. Data integration and visualisation for demanding marine operations. In: OCEANS 2016-Shanghai. IEEE, pp. 1–7.

Warner, R.M., 2020. Applied Statistics II: Multivariable and Multivariate Techniques. SAGE Publications, Incorporated.

Wiesmann, A., 2010. Slow steaming–a viable long-term option? Wartsila Tech. J. 2, 49–55.