Ocean Engineering 315 (2025) 119722

Contents lists available at ScienceDirect

Ocean Engineering

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

Research paper

Enhancing data quality in maritime transportation: A practical method for imputing missing ship static data

Ruikai Sun [a,*], Wessam Abouarghoub [a,b], Emrah Demir [a]

[a] Logistics and Operations Management, Cardiff Business School, Cardiff University, Cardiff, United Kingdom
[b] Department of Operations and Project Management, College of Business, Alfaisal University, Riyadh, Kingdom of Saudi Arabia

ARTICLE INFO

Keywords:
Data science
Maritime transport
Multiple nonlinear regression
Stepwise regression
Missing data
Data imputation

ABSTRACT

Maritime transport research, including emissions estimation, shipping network design, and bunker management relies on complete high-quality data. Incomplete ship static data often lead to bias and misleading conclusions. Existing imputation methods either depend on small data samples with poor imputation accuracy or are overly complex with practical limitations. This study addresses these issues by evaluating existing methods and proposing a new imputation approach to enhance the quality of ship static data. First, we introduce a stepwise multiple nonlinear regression method to simplify the imputation process and improve accuracy. Second, we propose a novel evaluation metric, the coverage rate, to assess the model performance. Finally, from a total of 14 models, we use a decision matrix to select the optimal model for imputing missing ship static data. These models are applied to a real dataset with missing values and cross-validated using multiple databases to ensure robustness. The proposed method maximizes the coverage rate, approaching nearly 100 percent for missing data. The most significant improvement was observed in main engine RPM imputation, where the average adjusted R-squared increased by at least 20.74%. Based on a large training dataset of 38,018 ships, this method can be directly applied to other maritime transport studies.

1. Introduction

Missing data presents a significant challenge in quantitative research, and maritime research is no exception. Similar to other industries, the use of big data and machine learning (ML) offers diverse solutions for data-driven decision-making in the maritime industry (Jeon et al., 2021; Yu et al., 2022b). Big Data Analytics (BDA) and Artificial Intelligence (AI) have gradually shaped decision-making processes in maritime operations (Yang et al., 2019; Munim et al., 2020; Yu et al., 2022a). Alongside these emerging technologies, Operations Research (OR) in maritime also relies on accurate and high-quality data. Optimization problems such as port berth allocation, liner network design, bunker management require clean and sufficient data to validate solutions (Ksciuk et al., 2023). The integration of BDA and OR can effectively address data uncertainty and further enhance data quality (Raeesi et al., 2023). While managing smaller datasets allows for more straightforward cleaning and procession, large databases introduce greater complexity. The increased prevalence of missing data in such large databases further complicates the process, making it more time-consuming and reducing the accuracy of data processing outcomes (Cheliotis et al., 2019; Cammin et al., 2020; Peng et al., 2020).

Quantitative research in maritime transport, covering areas such as emissions estimation, shipping network design and bunker management depends on the analysis of comprehensive high quality datasets. Which can be broadly classified into two categories: ship activity data and ship static data. The distinction between these categories is based on whether the data change over time (Yan et al., 2021). Ship activity data are primarily collected through Automatic Identification Systems (AIS), which automatically transmit messages via very high frequency (VHF) signals. Each AIS entity actively monitors the VHF medium to find available time slots for data transmission (Liang et al., 2024; Kelly, 2022). However, in regions with a dense concentration of AIS entities, network overloading can occur, leading to AIS data loss. Consequently, areas with a high volume of ships are particularly susceptible to data loss. VHF signals are also susceptible to environmental factors such as rain or fog, which can cause further data loss. Additionally, physical obstructions such as land or other vessels can diminish signal reception rates (Shepperson et al., 2018). Apart from environmental conditions and physical obstructions, AIS message loss can also arise from network overloads and timeouts within the AIS system (Last et al., 2014). These factors lead to data gaps and affect the completeness of ship activity data used in maritime transport research. Ship static data are

[*] Corresponding author.
E-mail address: sunr10@cardiff.ac.uk (R. Sun).

https://doi.org/10.1016/j.oceaneng.2024.119722
Received 3 July 2024; Received in revised form 30 October 2024; Accepted 3 November 2024
Available online 20 November 2024

similar prone to data loss. Ship static data include information such as ship dimensions (e.g. overall length, beam), engine specifications (e.g. main engine power), and other parameters (e.g., service speed) that remain constant over time and operating conditions (Kim et al., 2020). Detailed information on each vessel within a fleet can be collected from various sources, including ship owners, shipbuilders, and port authorities (Wang et al., 2016). However, the collection standards of these databases sources may not be fully consistent, and the uploaded static data for ships may not be entirely compatible with one another. This issue can be exacerbated by factors such as increasing processing costs and concerns over data confidentiality, leading to missing static data from ships (Jeon et al., 2021; Gao et al., 2023; Skarlatos et al., 2024).

Both types of data are essential to shipping research, yet most existing studies on missing data imputation have focused solely on ship activity data (Nguyen et al., 2015, 2018; He et al., 2021; Duan et al., 2022). The research in ship static data remains limited. For instance, in studies estimating ship or port emissions, missing ship static data can lead to a significant underestimation of ship emissions, by as much as 49% in some cases (Zhang et al., 2019; Gutierrez-Torre et al., 2020; Huang et al., 2020; Peng et al., 2020). Low-quality real data in maritime operations research can only provide a rough estimate for parameter estimation (Umang et al., 2013; Wawrzyniak et al., 2020). In berth allocation problems, ship network design and speed optimization, high-quality real data is critical. In ship or port emission inventory studies, key ship static data include main engine power, main engine revolutions per minute (RPM), and service speed (Tichavska and Tovar, 2015; Huang et al., 2018; Xu and Yang, 2020). Similarly, for ship and port operations optimization, static data such as deadweight tonnage (DWT), length, breadth, draught and service speed are essential (Christiansen et al., 2020; Yan et al., 2020; Martin-Iradi et al., 2024). As optimization studies increasingly incorporate environmental impacts into their objective functions, the need for accurate ship static data in emission studies grows (Du et al., 2011; Reinhardt et al., 2020; Ksciuk et al., 2023). Consequently, this research focuses on seven key static ship data parameters: main engine power, main engine RPM, service speed, overall length, beam, draft, dead weight tonnage (DWT).

Our own observations from shipping databases (i.e. Refinitiv, Clarkson) and evidence from the literature based on data from other shipping databases (e.g., IHS Maritime & Trade, China Classification Society and Lloyd’s Register) support the notion that a substantial amount of ship static data is missing key variables. For example, one study shows that main engine power, main engine RPM, and ship service speed have missing rates ranging from 11% to 44% (Merien-Paul et al., 2018; Zhang et al., 2019; Huang et al., 2020). When sufficient data are available or only a few values are missing, these gaps can be solved by estimating the missing values. However, failure to address these data gaps lead to inaccurate analyses and poor decision-making (Sun et al., 2025).

Shipping studies typically employ three main approaches to address the challenge of missing ship static data: approximation methods, statistical models, and machine learning (ML) algorithms. Approximation involves using static data from ships of similar size, age, manufacturer and operator when specific data is unavailable. Alternatively, average static data from ships of the same type is often used (Chen et al., 2021; Nguyen et al., 2022; Yang et al., 2023). However, these methods require access to extensive ship databases and are limited by their inability to estimate data for ships outside the database’s scope. Additionally, using average values fails to preserve the relationships between variables, leading to poor imputation accuracy. In contrast, statistical methods allow for modeling the relationships between different variables, which is beneficial for maritime studies (Mi et al., 2020). The most commonly applied statistical models in the literature, including those used in IMO reports, are multiple linear regression and non-linear regression models (Abramowski et al., 2018; Cepowski, 2019b; IMO, 2020, 2021; Schwarzkopf et al., 2021; Kim et al., 2022). These models provide concise imputation formulas, making the data substitution process more practical. However, they are limited by significant errors due to small training samples and outdated data. Furthermore, some models require numerous input parameters, restricting their applicability under certain conditions. Recently, several studies have used ML algorithms to impute missing ship static data (Gao et al., 2023; Skarlatos et al., 2024). While ML models generally outperform traditional methods in terms of accuracy, their complexity often renders them as ’black boxes’ making it difficult to understand how predictions or decisions are derived. This lack of transparency hinders the interpretation of underlying relationships within the data (Cheong et al., 2023). Additionally, ML models are prone to overfitting, which reduces their effectiveness in real-world applications. Therefore, a ship static data imputation method is needed that balances accuracy, interpretability and practicality for effective use in maritime operations.

This paper proposes a Stepwise Multiple Nonlinear Regression (SMNLR) method to estimate missing values in ship static data. The approach begins by categorizing ship static parameters into two groups: independent and dependent variables, based on the extent of missing data. First, ship static data are grouped by ship type and deadweight tonnage (DWT), as the literature supports the existence of a nonlinear relationship among these static parameters (Cepowski, 2019a; Cepowski and Chorab, 2021; Papanikolaou, 2014; Piko, 1980), and correlation analysis and nonlinear regression techniques are employed to identify the best relationship among these parameters. In this study these relationships form the basis for multiple nonlinear regression models (Cepowski and Chorab, 2021). Independent parameters are initially imputed based on the best nonlinear relationships identified. Nonlinear functions of the independent variables are incorporated as parameters in a multiple linear regression model to estimate the coefficients of multiple nonlinear regressions. Simultaneously, the stepwise regression is used within the multiple regression models to select independent parameters that correlate with the dependent variables. This study compares the performance of six models from the literature (Abramowski et al., 2018; Cepowski, 2019a; Schwarzkopf et al., 2021; IMO, 2020, 2021; Kim et al., 2022) as well as two ML models, against the proposed method for imputing missing ship static data. The two ML models are used as benchmarks. Additionally, six new models are proposed under the SMNLR method, based on the proposed grouping rules and regression parameters. In total, 14 models are tested, and based on the results, a decision matrix is generated. All models are applied to a real dataset with missing values and cross-validated using different databases.

This paper makes four significant contributions. Firstly, it introduces an alternative approach to improve data completeness for maritime transport research, which also enhances the model’s imputation accuracy. Additionally, using complete training data and basic formulas, this method can be directly applied to impute ship static data. Secondly, the paper introduces the coverage rate indicator, an innovation addition to the evaluation of model performance for imputing ship static data. The coverage rate provides a quantifiable measure of the imputation model’s success, offering a comprehensive assessment of its effectiveness. Thirdly, the paper presents the development of a decision matrix specifically designed for imputation across various missing data scenarios. This matrix facilitates precise, case-specific imputations, further enhancing data completion accuracy. The application of this decision matrix allows the coverage rate to reach 100%, delivering a high-quality ship static dataset to use for maritime research. Finally, the proposed approach can be applied to impute other types of cross-sectional data.

The rest of the paper is structured as follows. Section 2 provides a brief literature review. Section 3 outlines previously proposed methods for imputing missing data in the shipping industry. Section 4 introduces the SMNLR method for imputing missing ship static data, while Section 5 presents the results of the proposed model, including a validation case study on container ships from the Clarkson database. This section also compares the performance of the proposed model with that of previously studied models. Finally, Section 6 concludes the paper and outlines direction for future research.

2. Literature review

Missing data refers to the absence or incompleteness of data points within a dataset. It is a common issue encountered in various fields, including maritime transportation, where data are collected from various sources and may be subject to various factors that result in missing values. In maritime transportation research, missing data poses a significant challenge as it can hinder accurate analysis and decision-making processes. Understanding the extent and nature of missing data is crucial for conducting comprehensive studies on energy efficiency, emissions estimation, and other aspects of maritime operations. To address the issue of missing data, researchers in maritime transport have explored different methods and techniques for data imputation. Therefore, this section offers a brief overview of the missing data concept discussed in the literature and its relevance in maritime research.

2.1. The concept of missing data

According to Rubin (1976) and Rubin and Little (2019), three types of missing data situation can occur during data collection: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). In the case of MCAR, data loss is equally likely in all cases, indicating that there is no discernible relationship or pattern with missing values. Analyzing only the available data without estimation does not introduce bias into the results. However, the assumption of MCAR is often regarded as strict and unlikely to hold in practical scenarios (Muthén et al., 1987; Raghunathan, 2004). MAR refers to data loss that occurs in a systematic and predictable manner based on other information that is also missing. The missingness in MAR is partially or completely predictable. For instance, the likelihood of data being missing may depend on certain observed variables (Rubin and Little, 2019). MNAR, on the other hand, is a common type of lost data. It is assumed that the probability of losing a value is due to unknown reasons or for reasons for which there is no relevant information (Santos et al., 2019). MNAR implies that the missingness cannot be attributed to randomness or predictability.

These causes of data loss are neither systematic nor predictable based on other information, yet they are not entirely random either. Since assumptions about the causes of data loss can significantly influence the assumptions underlying statistical modeling techniques, it is essential to distinguish between different types of missing data. When the missing data deviate from the assumption of MCAR and indicate a potential pattern of missingness, it is feasible to develop models for estimating missing data (Stead and Wheat, 2020).

Little (1988) introduced a multivariate test to determine whether the data are MCAR. This test examines the variation in means among subgroups with the same missing data patterns for each variable in the dataset. By contrasting the mean of the observed variable for each missing data pattern with the overall expected mean (estimated using the Expectation Maximization Estimation (EM) algorithm), one can assess whether the data are MCAR. The test statistic is derived from the sum of squares of the standardized differences between the means of the subsample and the expected overall mean. This is further weighted by the estimated variance–covariance matrix and the count of observations in each subgroup (Enders, 2010). Under the null hypothesis that the data is MCAR, this test statistic asymptotically follows a chi-square distribution. A statistically significant outcome of this test suggests that the data may not be of the MCAR type. Methods for handling missing data can be classified into two broad categories based on their approach. The first category comprises methods that rely on discarding sample portions with incomplete information. The second category encompasses methods that replace missing data with values imputed based on the estimates of the models. A detailed decision-making process implemented in this study to select appropriate missing data handling methods is illustrated in Fig. 1.

The proportion of missing data is a critical metric used to assess the extent of information loss within a dataset. Generally, missing data rates below 5% or 10% are considered insignificant (Lee and Huber, 2011). However, when the missing data rate surpasses 50%, the study's findings should be interpreted as hypothesis-generating rather than conclusive (Dong and Peng, 2013; Jakobsen et al., 2017). High rates of missing data can compromise the validity and reliability of a study's results.

2.2. Missing data research in maritime transport

Technological advancements have led to a substantial increase in the diversity and volume of collectable data in the maritime sector. However, the expansion of database sizes does not necessarily translate to improved data quality, and new challenges related to data omissions continue to emerge (Batini et al., 2009). Maritime research has primarily concentrated on two types of missing data: ship activity data and ship static data. Most studies on missing ship data have focused on utilizing AIS data to reconstruct missing ship trajectories or identify behavioral patterns of ships (Sang et al., 2015; Nguyen et al., 2015; Dobrkovic et al., 2018; Nguyen et al., 2018; Guo et al., 2021).

In contrast to ship activity data, research on inferring missing ship static data is relatively limited. There are three commonly preferred methodologies for handling missing static data, multiple liner regression, nonliner regression, and ML algorithms. As early as 1980, Piko (1980) conducted a regression analysis using the Lloyd's Register ship service database to explore the correlation between deadweight tonnage (DWT) and service speed with parameters such as length, width, draft, gross tonnage, and power. This seminal work laid the foundation for subsequent studies. Numerous research efforts have since employed linear regression to identify relationships among key ship design indicators, providing general guidance for establishing basic hull dimensions and total engine power during the preliminary stages of container ship design. McArthur and Osland (2013) identified a linear relationship between a ship's main engine power and gross tonnage to impute missing data, using essential design variables like DWT and TEU as input variables. While these models are broadly applicable, their accuracy is often limited (Charchalis and Krefft, 2009; Charchalis, 2013, 2014). For example, Huang et al. (2020) use a polynomial regression to develop models for estimating missing main engine power data across different types of ships. Similarly, Peng et al. (2020) adopts a stratified random sampling method to impute missing ship static data by categorizing ships based on size, type, main engine power, and other factors, thereby reducing the uncertainty in emission estimation caused by missing data.

Beyond linear regression, there has been a shift towards exploring nonlinear regression. Schwarzkopf et al. (2021) developed various curve fits to estimate absent ship attributes crucial for modeling ship emissions, such as gross tonnage, main or auxiliary engine power, engine rating, and service speed. These attributes are frequently missing in current datasets. Cepowski (2019b) used nonlinear regression to estimate total engine power based on the deadweight or speed for tankers, bulk carriers and container ships from 2000 to 2018. Although the model aligns well with core indicators such as GT and TEU, its precision in estimating the total engine power is compromised. To address this, Cepowski (2019a) refined the approach, categorizing ships by size and conducting distinct nonlinear regressions for each group. The results showed that the method had the highest accuracy compared to the ungrouped regression. Building on this, a 2021 study by Cepowski and Chorab (2021) integrated nonliner regression with a stochastic search function, marking a 44% surge in model accuracy compared to using solely nonlinear regression (Cepowski and Chorab, 2021). Notably, the primary focus of these methodologies has been on

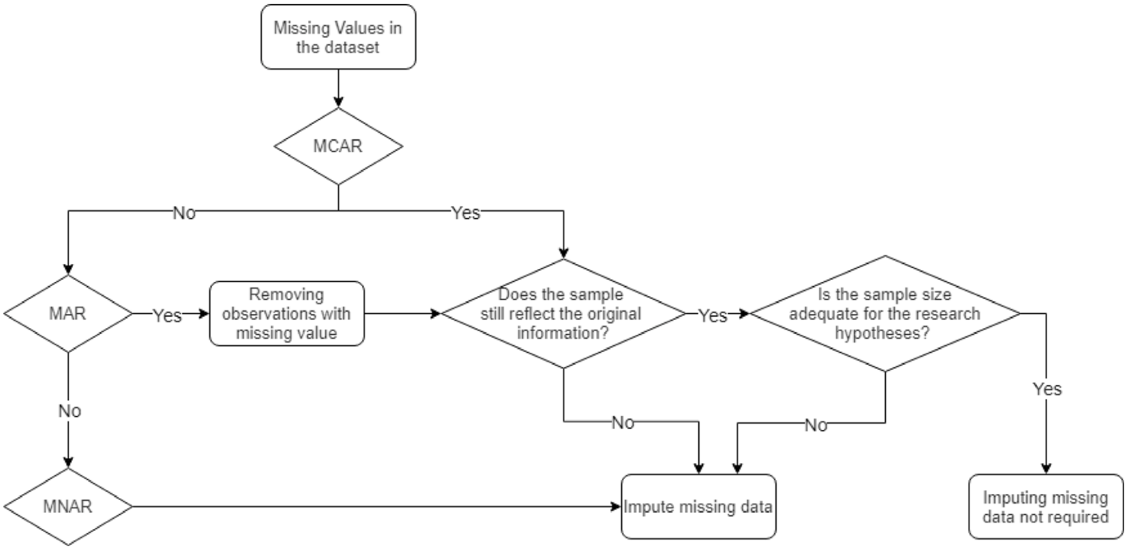

Missing Values in the dataset
MCAR
No
Yes
MAR
Yes
Removing observations with missing value
Does the sample still reflect the original information?
Yes
Is the sample size adequate for the research hypotheses?
No
No
No
Yes
MNAR
Impute missing data
Imputing missing data not required

Fig. 1. Decision-making process for selecting missing data handling methods.

parameters such as main engine power and other fundamental indicators (i.e., LOA, LBP). They do not encompass ship speed, main engine RPM or other variables pivotal for maritime transport research. More recently, Kim et al. (2022) applied a model-based computation method using regression analysis for estimating missing ship data, enhancing the goodness-of-fit by 15.6% over several regression models alone. However, this method requires large datasets for effective training and depends on a large number of input variables. As a result, we need a model that balances its complexity and accuracy.

ML algorithms have been adopted alongside traditional regression techniques to impute missing data. ML algorithms often outperform conventional models in terms of accuracy. Gurgen et al. (2018) implemented an artificial neural network, and the findings revealed that the primary details of the chemical tankers were deduced with greater precision than the data from the sample vessels. To the best of our knowledge, many of these studies have employed ML methods to impute missing AIS data. The rationale lies in AIS data being time-series in nature with ample data available for model training. Conversely, when addressing ship static data, which are static and interconnected during ship design, regression models have been the go-to for missing data imputation. While they may not be as accurate as ML models, their interpretability is commendable, elucidating the relationship between parameters (Gilpin et al., 2018). It is noteworthy that research on missing AIS data far outpaces that on missing ship static data. This can be attributed to the versatility of AIS data, which finds applications ranging from maritime rescue and accident prediction to route optimization (Yang et al., 2019). Historically, the datasets of sampled ships exhibited high completeness with modest volume, thereby sidelining concerns of missing parameters. Yet, with the contemporary focus on carbon neutrality and BDA in shipping, there is an increased demand for extensive ship datasets. After all, a model, regardless of its accuracy, loses its utility if it addresses only a minuscule fraction of missing data instances. A comprehensive comparison of the models is shown in Table 1, with Table 2 describing the notation used in the literature and Table 1.

Table 1 offers a comparative analysis between the methodologies highlighted in the literature review discussion and the method proposed in this paper. In particular, this study uses the largest data sample for training. The proposed method leverages the ship's fundamental parameters to infer its attributes. These parameters are categorized as “independen” and “dependent” according to the technical characteristics of the ship. Independent parameters cover mainly the static dimensional data of the ship, such as the DWT and its total length, while dependent parameters considers the ship's capabilities, including attributes like main engine power and ship service speed.

Estimating dependent parameters is more challenging due to their involvement of multiple design parameters, resulting in higher rates of missing data. Although employing fewer parameters, the SMNLR method presented in this paper maintains or surpasses the accuracy of other methods. Using only the low missing rate parameters also implies a higher model coverage rate. In addition, the number of model comparison benchmarks used in this study is greater than in previous studies. More indicators were also used to evaluate the models, giving a more complete understanding of the performance of the model.

3. Methodological approaches for imputing missing data

This paper utilizes known ship static parameters to deduce the missing data. In the following sections we explore and discuss the eight methods proposed in the literature.

3.1. Imputation methods from the literature

In this paper, eight representative missing data imputation methods are considered from the literature, which is shown in Table 3. Method 1 is the Random Forest, which is a ML algorithm that consists of decision trees. The final output category or value is determined by the majority score of individual decision tree prediction categories (Breiman, 2001). Without dimensionality reduction, it can handle high dimensional data (Pantanowitz and Marwala, 2009; Tang and Ishwaran, 2017). Method 2 is a Generalized Additive Model (GAM). It is a type of Generalized Linear Model (GLM) where the linear response variables are assumed to have a linear relationship with an unknown smoothing function of the predictor variables. GAMs were first formulated by Hastie (2017) with the intention of integrating the characteristics of GLMs with additive models. However, it is important to note that this increased flexibility may come at the cost of reduced interpretability (Ravindra et al., 2019). Method 3 is based on nonlinear regression method. The method presents regression equations to estimate the latest data on ship static data based on various parameters, including the container ship load weight, the number of containers and their combinations at the preliminary design stage. These regression formulas are generated based on an evolutionary algorithm to find an optimal combination. Method 4 is a multivariate nonlinear regression method. A set of regression equations for ships according to their ship types such as tankers, bulk carriers and container ships is created. Method 5 use

Table 1
A summary of different imputation models in literature review.

Reference	Segment focus	Sample size	The number of methods compared	Input parameter	Output parameter	Model performance evaluation indicator
Charchalis (2013)	Tankers, Bulk, Containers	3200 Ships	None	LWL, BEAM, DRA, CB, CWP, SSP, H	MEP	Absolute error, Relative error, Correlation coefficient
Charchalis (2014)	Container	17 Containers	None	TEU	DWT, LBP, LBP/DRA, BEAM/DRA, DRA/BEAM	SE
Abramowski et al. (2018)	Container	3573 Containers	None	DWT, TEU	LBP, BEAM, DRA, D, GT, SP, CB, CWL, MEP, LV, Disp, SSP, LBD, LBT, TEU, DWT	SE, R-squared
Gurgen et al. (2018)	Chemical Tanker	100 Tankers	None	DWT, SSP	LOA, LBP, BEAM, D	MAPE, Correlation coefficient
Cepowski (2019a)	Tanker	1723 Tankers	3	DWT, SSP	LBP, BEAM, DRA	SE
Cepowski (2019b)	Tankers, Bulk, Containers	1710 Tankers, 1248 Bulkers, 442 Containers	2	DWT, SSP	MEP	SE, R-squared
Cepowski and Chorab (2021)	Container	215 Containers	2	DWT, SSP	LBP, BEAM, DRA, D	RMSE, Correlation coefficient
Kim et al. (2022)	Container	6278 Containers	2	AEP, BEAM, DRA, DWT, GT, LDT, LOA, LBP, MEC, MEP, MER, MES, SSP, TEU	AEP, BEAM, DRA, DWT, GT, LDT, LOA, LBP, MEC, MEP, MER, MES, SSP, TEU	MSE, MAE, RMSE, Adjusted-R-squared
This paper	Tankers, Bulk, Containers	17 980 Tankers, 12 374 Bulkers, 7664 Containers	8	DWT, LOA, DRA, BEAM	MEP, MER, SSP, DWT, LOA, DRA, BEAM	MAE, RMSE, Adjusted-R-squared, Friedman-Nemenyi test, Model coverage rate

Table 2
Technical parameter notations for ships.

Notation	Description	Notation	Description
AEP	Auxiliary engine power [kW]	LDT	Light displacement tonnage, LDT [t]
BEAM	Breadth [m]	LOA	Length overall, LOA [m]
CB	Block coefficient [–]	LV	Light vessel mass [t]
CWL	Waterplane area coefficient [–]	LWL	Length at waterline [m]
CWP	Waterplane coefficient [–]	MEC	Main engine cylinder, [–]
D	Depth [m]	MEP	Main engine power [kW]
Disp	Displacement mass [t]	MER	Main engine RPM, [–]
DWT	Deadweight capacity [t]	MES	Main engine stroke, [–]
GT	Gross tonnage [–]	SP	Final price [$ million]
H	Height [m]	DRA	Draft [m]
LBD	the result of L·B·D [m^3]	TEU	Number of containers [–]
LBP	Length between perpendiculars [m]	SSP	Ship service speed [knot]
LBT	the result of L·B·T [m^3]		

a nonlinear regression model with asymptotic behavior to explain the correlation between predicted and input parameters. The main difference between this model and the previous model is that the exponential model can be used to simulate the asymptotic patterns observed at SSP and MER. Larger ships install larger marine diesel engines with faster ship speed and lower MER. This is because of the assumption that larger ship marine diesel engines typically drive larger diameter, lower pitch propellers. Therefore, asymptotic behavior is implemented as the MER cannot be reduced below zero. Although ultralow-speed cylinder diesel engines can run at approximately 60 RPM, lower ratings are very rare. The allowed MER value is therefore capped at this point and the smaller estimate is corrected to 60 RPM. Method 6 is presented in the fourth IMO GHG study (2020), a new algorithm implemented to impute missing ship static data. The algorithm is based on a multiple linear regression created for each ship type, considering the known design parameters of each ship (IMO, 2020). Method 7 is taken from another IMO report. In this method, there is a power function relationship between the ship static data and DWT (IMO, 2021). This method fits the relationship between the known static data and the DWT using a power function to obtain the corresponding coefficients. Method 8 is a model-based approach with regression analysis calculations and is applicable to estimate missing values (Kim et al., 2022). First, it identifies the fitted function for each parameter using curve fitting and chooses the best function according to the R-square value. After getting a complete data set, this method performs a multiple regression analysis with backward elimination to make a prediction model for each parameter.

Table 3 shows the difference between the selected eight methods and the gap that exists among them. These methods can be divided into two categories, ML model and statistical model. The application can be categorized into two types that need to train or provide formulas in the literature that can be used directly, which reflects the practicality of the method. The size of the training data set is also a parameter worth comparing and determines the performance of the method. The type and number of input parameters determine whether the model is limited in its use. This is because the input parameters can also be missing, making the model unusable. Methods 1 and 2 are ML models, and both need to be trained using a large amount of data before they can be used. They are also prone to overfitting and perform worse on new datasets. In addition, the models lack transparency and are not easy to interpret. Since these two ML models have high accuracy, they are used in this paper as upper bound performance benchmarks

Table 3
Comparative table for literature method.

Methods name	Reference	Methods classification	Methods application	Training dataset size	Input parameter	Advantages	Gap
Method 1	Breiman (2001)	Machine Learning	Need to train	\	DWT, LOA, Draft, Beam	-High accuracy -Robustness to noise point	-Lack of Interpretability -Computational complexity -Propensity to overfit -High data requirements
Method 2	Hastie (2017)	Machine Learning	Need to train	\	DWT, LOA, Draft, Beam	-Moderate accuracy -Flexibility -Moderate interpretability	-Propensity to overfit -High data requirements -Sensitivity to smoothing parameters
Method 3	Abramowski et al. (2018)	Statistical model	Input data directly	3573	DWT	-Easy to use -High interpretability -Only need DWT information	-Small training sample -Low accuracy -Sensitive to noise point
Method 4	Piko (1980), Cepowski (2019b)	Statistical model	Input data directly	3400	DWT, SSP	-Easy to use -Simple input parameter	-Small training sample -Low accuracy -Sensitive to noise point
Method 5	Schwarzkopf et al. (2021)	Statistical model	Input data directly	Not mentioned	GT	-Easy to use -Consider the asymptotic character of the data -Only need GT information	-Low accuracy -Outdated coefficient
Method 6	IMO (2020)	Statistical model	Need to train	\	LOA, MEP, DWT, SSP	-Moderate accuracy -High interpretability	-High data requirements -Not fit nonlinear data well -Limited application scenarios
Method 7	IMO (2021)	Statistical model	Input data directly	Not mentioned	DWT	-Easy to use -Only need DWT information	-Low accuracy -Outdated coefficient
Method 8	Kim et al. (2022)	Statistical model	Need to train	6278	DWT, LOA, MEP, MER, SSP, DRA, Beam, GT	-Moderate accuracy -High interpretability	-Computational complexity -High data requirements -Reliability of input parameter not considered
Proposed method	\	Statistical model	Input data directly	38,018	DWT, LOA, Draft, Beam	-High accuracy -High interpretability -Easy to use -Simple input parameter -Considers the impact of different sized ships	

for other statistical models (Thiyagalingam et al., 2022). Statistical methods also have some gaps. Methods 3, 4,5 and 7 provide formulas for direct use. However, these formulas are based on smaller training samples and older data, which affects the accuracy of the model. Methods 3 and 4 also do not address how to impute the MER. Method 6 provides regression parameters but not regression coefficient and therefore needs to be fitted based on the case study data. Method 6 requires the use of SSP to impute MEP and MEP to impute SSP, so when the ship is missing both SSP and MEP the data cannot be imputed. In addition, MER requires the imputed MEP and SSP parameters, but these two parameters have a high missing rate on their own, leading to limitations in the scenarios in which the method can be used. Method 8 provides formulas and parameters, but the model requires a large number of input parameters. SSP, MEP, and MER require 11 parameters to be imputed. These parameters include AEP, MEP, MER, and LDT which have very high missing rates of their own. Although Method 8 can avoid missing input parameters by two-round imputation. However, if the missing rate of input parameters is high, then it may result in the model being trained based on a large number of estimated values instead of real values. This can introduce bias in model training. Therefore, the method proposed in this paper is improved based on the above gaps, and the specific steps are shown in detail in Section 4.

3.2. Measuring the output accuracy

This study employs various metrics, such as root mean square error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and adjusted R^2, to thoroughly assess the effectiveness of the predictions. RMSE is particularly effective for evaluating a model's sensitivity to outliers and large errors, as it emphasizes these by squaring the residuals. In contrast, MAE and MAPE offer insight into the overall error performance of the model. While MAE is scale-dependent, allowing for the observation of error magnitudes across individual variables, MAPE is scale-independent, making it useful for assessing model performance across different variables. R-squared and adjusted R-squared measure the model's ability to explain the variance in the data. The adjusted R-squared account for the number of predictor variables, providing a more accurate and reliable evaluation, especially in models with multiple predictors. Together, these metrics offer a comprehensive assessment of model performance, balancing error magnitudes, robustness to outliers, and the model's explanatory power. Their formulas are shown in Eq. (1)–(5). In these formulas, y_i and $\hat{y}_i$ are the actual value and predicted value, respectively. $\bar{y}_i$ is the mean of actual value. N is number of records in the data set, and p is the number of independent variables. These metrics are used to gauge the numerical precision of models. In the case of RMSE, MAE, and MAPE, lower values indicate more accurate model estimations. For R-squared and Adjusted R-squared, the values usually range from zero to one. When these values are close to one, it means better performance.

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

$$MAE = \sum_{i=1}^{N} |\hat{y}_i - y_i| / N \tag{2}$$

$$MAPE = 100 \times \sum_{i=1}^{N} |\hat{y}_i - y_i| / |(y_i \times N)| \tag{3}$$

$$R^2 = 1 - \sum_{i=1}^{N} (\hat{y}_i - y_i)^2 / \sum_{i=1}^{N} (\bar{y}_i - y_i)^2 \tag{4}$$

$$Adjusted\, R^2 = [(1 - R^2)(N - 1)] / (N - p - 1) \tag{5}$$

3.3. Measuring the output coverage rate

The model coverage rate refers to how much percentage of data can be imputed by the model's output variables based on the missing rate of model's input variables. It is the ratio of the size of complete data after being imputed to the total sample size and can be calculated as shown in Eq. (6). In this formula, C_{rate} is coverage rate, $N_{imputed}$ is the sample size of dataset can be imputed by models. N_{total} is the total sample size of dataset. The model coverage rate is a very important metric for imputing missing data and possesses similar concepts in other areas of missing data research (Blankers et al., 2010; Enders, 2022). This is because if the model introduces too many parameters or if the input parameters themselves have a high missing rate, this can lead to a limited number of scenarios in which the model can be used, such as Methods 6 and 8. Introducing more parameters can improve the model's accuracy performance, and the coverage rate can provide a reference for which parameters to introduce to balance the practicality and accuracy of the model.

$$C_{rate} = N_{imputed} / N_{total} \tag{6}$$

4. Stepwise Multiple Nonlinear Regression (SMNLR) method

Different approaches can be employed to address missing data, which can vary depending on the nature of the data or the type of missingness. The outcomes obtained from these methods may also differ accordingly. This research introduces an alternative method for data imputation, namely SMNLR. The primary objective in imputing ship static parameters is to develop a model that possesses broad applicability, strong interpretability, and high accuracy. Most previous research solely relied on regression analysis to estimate ship static parameters. However, the limited size of their training samples compromises the accuracy of their models. Conversely, hybrid methods and ML models proposed in other studies necessitate a sufficiently large sample size to be effective. The SMNLR method is based on the latest commercial database of global ships and uses a combination of grouped regression, multiple nonlinear regression and step-forward linear regression to obtain a series of imputation formulas. With SMNLR method, it is possible to impute ship static data for small sample sizes. Fig. 2 illustrates the flowchart of SMNLR method.

4.1. Data preparation

As discussed earlier, there are seven target ship static parameters to be imputed in this paper. These include MEP, MER, SSP, LOA, BEAM, DRA, DWT. Firstly, we obtain the raw static data of global merchant fleet in service from the Refinitiv database. From this dataset, parameters with a missing rate less than 10% missing rate were identified and considered as independent parameters. These independent parameters include the target parameters (LOA, BEAM, DRA, DWT). The reason for categorizing based on missing rate is that using these low missing rate parameters as input parameters to the model expands the applicability of the model and increases model stability. Existing models with high accuracy require a large number of input parameters. However, these input parameters can also have missing rates. Increasing the number of input parameters also increases the probability of missing input parameters, thus limiting the scenarios in which the model can be used. Therefore in this paper, choosing parameters with low missing rates can effectively improve this situation. In addition existing models also have the situation of avoiding the missing input parameters of the model by two rounds of imputation. However, if input parameters have a high missing rate then it may lead to the input to the model are estimated values, which will bias the results (Lin and Tsai, 2020). Subsequently, the remaining target parameters were used as dependent parameters, including MEP, MER, SSP.

4.2. Grouping data

Both ship type and DWT significantly influence the distribution of other ship static data (Piko, 1980; Barrass, 2004; Cepowski, 2019b). Ships of different sizes and functions are designed according to distinct principles, which result in notable variations in their static parameters. For example, the main engine power of a large ship increases with size, but its service speed may increase only slightly or remain constant. Additionally, LNG tankers typically have higher service speeds compared to oil tankers of similar size. By grouping the data and performing regressions individually for each group, the model is better able to capture the specific characteristics of different ship types. This grouping approach is crucial to improving the accuracy of the model (Prais and Aitchison, 1954; Cepowski, 2019b). We first classified the data into three categories based on ship type: bulk carriers, container ships, and tankers. Then, we further subdivided the data within each category based on DWT (Kanamoto et al., 2021). The specific DWT classification rules are provided in Table 7. Each grouped dataset is then used to train the models individually, followed by the identification of the best nonlinear relationship between the parameters within each group.

4.3. Choose best nonlinear relationship

There is a well-established nonlinear relationship between static parameters (Papanikolaou, 2014; Cepowski, 2019b; Cepowski and Chorab, 2021; Rinauro et al., 2024). Therefore, in this paper, multiple nonlinear regressions are applied to fit the static parameters of the ships. To efficiently derive the multiple nonlinear regression equation, it is crucial to determine the relationship between the parameters. Nonlinear regression estimates the functional association between a continuous curve and a discrete set of data points representing the coordinates on a surface. In simpler terms, an analytic function $y = f(x)$ is derived to approximate or pass through a sequence of data points (x, y) (O'Hagan, 1978).

We fit the observed data for each parameter by least squares using linear, quadratic, cubic, power, logarithmic and exponential functions. These six functions are the basic nonlinear function (Arlinghaus, 2023). Assuming the data has m independent and n dependent parameters, this step generates $(m+n-1)\times(m+n)\times 6$ nonlinear functions. The best function to describe the nonlinear relationship between the independent and dependent variables is selected based on the R-squared values of the

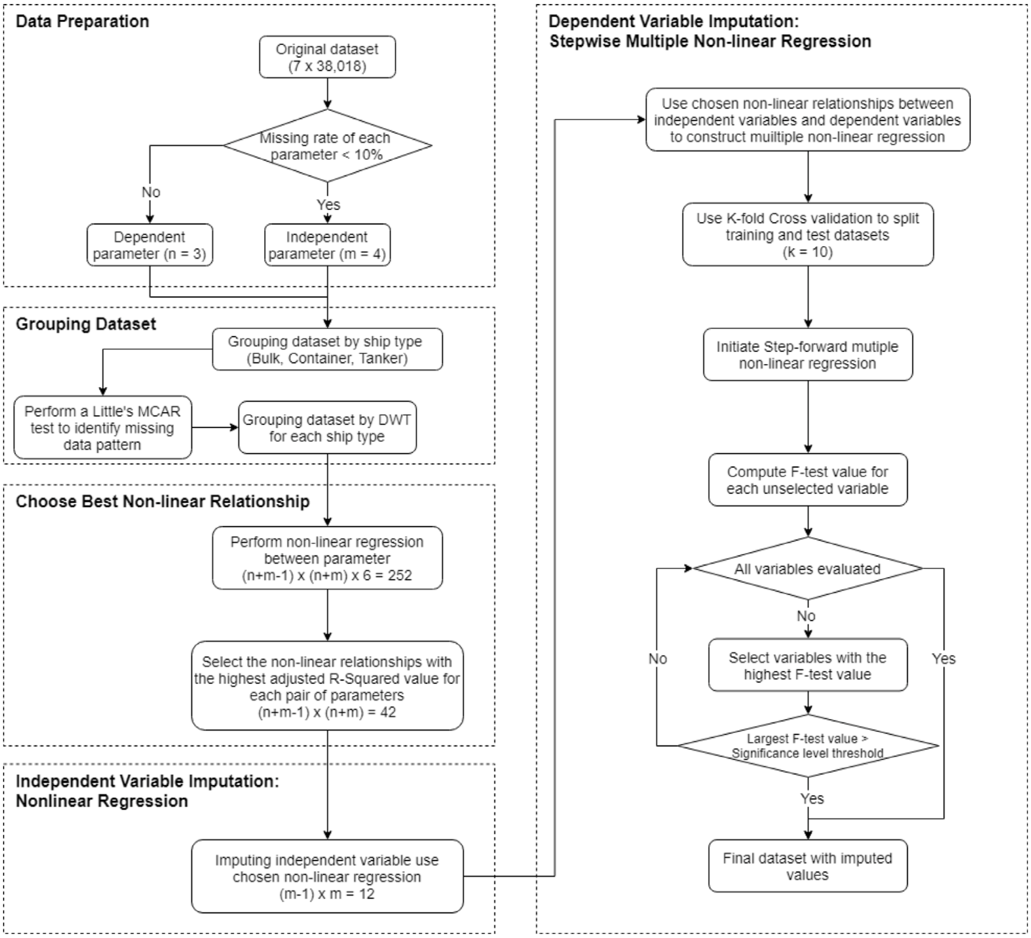

Data Preparation
Original dataset (7 x 38,018)
Missing rate of each parameter < 10%
No
Yes
Dependent parameter (n = 3)
Independent parameter (m = 4)
Grouping Dataset
Grouping dataset by ship type (Bulk, Container, Tanker)
Perform a Little's MCAR test to identify missing data pattern
Grouping dataset by DWT for each ship type
Choose Best Non-linear Relationship
Perform non-linear regression between parameter (n+m-1) x (n+m) x 6 = 252
Select the non-linear relationships with the highest adjusted R-Squared value for each pair of parameters (n+m-1) x (n+m) = 42
Independent Variable Imputation: Nonlinear Regression
Imputing independent variable use chosen non-linear regression (m-1) x m = 12
Dependent Variable Imputation: Stepwise Multiple Non-linear Regression
Use chosen non-linear relationships between independent variables and dependent variables to construct muiltiple non-linear regression
Use K-fold Cross validation to split training and test datasets (k = 10)
Initiate Step-forward mutiple non-linear regression
Compute F-test value for each unselected variable
All variables evaluated
No
Select variables with the highest F-test value
No
Yes
Largest F-test value > Significance level threshold
Yes
Final dataset with imputed values

Fig. 2. A flowchart of the Stepwise Multiple Nonlinear Regression (SMNLR) method.

functions. The number of nonlinear functions can be filtered to $(m+n-1)\times(m+n)$. Based on these nonlinear functions, we get a dataset inside which are nonlinear regression estimates $\hat{y}_{ij}$ of independent parameters y_j on dependent parameter x_i. Using these estimates, we can further perform a step-forward multiple nonlinear regression on them to obtain the final imputation value.

4.4. *Nonlinear regression*

Independent variables also have missing values. For the independent parameters with low missing rate, this paper first imputes these parameters using the best nonlinear relationship obtained from previous section. If both parameters which had the best nonlinear relationship were missing then the second-best parameter is used to impute, and so on. Since the missing rate of these variables is less than 10%, replacing them with estimates from the nonlinear regression cannot affect the performance of the model, when the stepwise multiple nonlinear regression is trained. This makes SMNLR more stable and avoids missing input variables.

4.5. *Stepwise multiple nonlinear regression*

Stepwise regression serves as a method for constructing a regression model, where the selection of predictor variables is done automatically (Shen and Ren, 2014). Stepwise regression serves as a method for constructing a regression model, where the selection of predictor variables is done automatically. The primary objective is to autonomously choose the most essential variables from a broad range of options, enabling the creation of a regression model used for prediction or explanation (Hocking, 1976). Its essence is to establish the optimal multiple linear regression equation (Jenelius, 2019).

To conduct multiple nonlinear regression, this paper substitutes the nonlinear function into a multiple linear regression. The forward selection method is used to build multiple nonlinear regression model. This method involves adding one regression variable at a time until no more variables can be introduced in the process as follows. In the previous step, a one-variable nonlinear regression model is built for each of the independent variables x_i and the dependent variable y_j. It could get the estimates $\hat{y}_{ij}$ for each x_i. Now we create a one variable linear regression model for each $\hat{y}_{ij}$ (Eq. (7)) and calculate the value of the F-test statistic for the regression coefficient of each $\hat{y}_{ij}$. F_{max_1} is the maximum of these values. F_1 is a corresponding critical value, based on given significant level. If $F_{max\_1} > F_1$, then $\hat{y}_{1j}$ is introduced into the regression model, and this is the first independent variable selected.

$$\hat{y}_j = \beta_0 + \beta_i \hat{y}_{ij} + \epsilon, i = 1, \dots, n \tag{7}$$

In the second step, the unselected independent variables are introduced into the regression equation separately to create a two variable regression model (Eq. (8)) of the dependent variable y_j against the selected and unselected independent variables. In total, there are a total number of n-1 models. The value of the statistic for the F-test of the regression coefficient of the variable is again calculated and its maximum value, F_{max_2}, is chosen. For a given significant level, the corresponding critical value is noted as F_2. If $F_{max\_2} > F_2$, then $\hat{y}_{2j}$ is introduced into the regression model, which is the second selected independent variable. If there is no selected variable, then the variable introduction process is terminated.

$$\hat{y}_j = \beta_0 + \beta_1 \hat{y}_{1j} + \beta_i \hat{y}_{ij} + \epsilon, i = 1, \dots, n-1 \tag{8}$$

Table 4
Description of parameters.

Code	Name	Description
Name	Ship name	A proper noun chosen at the shipowner's discretion
ID	IMO number	The seven-digit number of the IMO ship identification number assigned to all ships when constructed.
TYP	Ship Type	The classification of ship
STYP	Ship Sub-Type	The sub classification of ship
SSTYP	Ship Sub-Sub-Type	The sub-sub classification of ship
DWT	Dead weight tonnage	Deadweight tonnage is a measure of how much weight a ship is carrying or can safely carry.
GT	Gross Tonnage	Gross tonnage is a nonlinear measure of a ship's overall internal volume.
LOA	Overall Length	Overall Length refers to the maximum length of a vessel from the two points on the hull measured perpendicular to the waterline.
BEAM	Beam	The beam of a ship is its width at the widest point.
DRA	Draught	The draft or draught of a ship's hull is the vertical distance between the waterline and the bottom of the hull
Cubic	Cubic Capacity	Cubic capacity in cubic meters, the total capacity of goods a vessel can handle in its holds or tanks.
SSP	Service Speed	The average speed maintained by a ship under normal load and weather conditions
MEP1	Main Engine Power 1	The total power supplied by the main engine installed on a ship (Data source 1)
MEP2	Main Engine Power 2	The total power supplied by the main engine installed on a ship (Data source 2)
MEP	Main Engine Power mixed	The total power supplied by the main engine installed on a ship (All data source)
MER	Main Engine RPM	The revolutions per minute of main engine
MEY	Main Engine Built Year	The year in which main engine was constructed
AEP	Auxiliary Engine Power	The total power supplied by the auxiliary engine installed on a ship
AER	Auxiliary Engine RPM	The revolutions per minute of auxiliary engine
AEY	Auxiliary Engine Built Year	The year in which auxiliary engine was constructed
SHY	Ship Built Year	The year in which ship was constructed

The process is then repeated until all variables are selected, or the F-test value is greater than the critical value at a certain step, then the process is terminated. We obtain the multiple regression model for y_j as shown in Eq. (9). $\hat{y}_j$ is the final imputed value of the entire model for y_j and m is the serial number of the selected variable.

$$\hat{y}_j = \beta_0 + \beta_1 \hat{y}_{1j} + \beta_2 \hat{y}_{2j} + \cdots + \beta_m \hat{y}_{mj} + \epsilon \tag{9}$$

The SMNLR method proposed in this paper effectively addresses the limitations of existing approaches. First, by selecting input parameters based on the rate of missing data, the model's stability is significantly improved. Second, by grouping the data according to ship types and sizes, the model can capture the unique characteristics of different ships, leading to a better overall fit. Third, selecting the most suitable nonlinear function for each parameter, rather than applying a uniform function, enhances the model's flexibility, allowing it to better capture variations in the data. Fourth, the use of multiple nonlinear regression enables the model to account for more ship-specific features, resulting in a closer alignment with real-world data. Finally, due to the availability of ample training samples, the equations and regression parameters generated by the statistical methods can be directly applied, eliminating the need for additional training, and thus enhancing the model's practicality.

5. Computational experiments

This study uses data from Refinitiv database, which includes a comprehensive set of 330 parameters per vessel. We have extracted data on bulk carriers, container ships and tanker from January 1982 to October 2023, encompassing 38,018 ships with capacities ranging from 152 to 400,000 DWT. This dataset includes most merchant ships currently in service as of October 2023, ensuring that the model derived from it could be directly used in most scenarios. Table 4 reports details of the codes and descriptions of all the parameters used in this case study. 'Main Engine Power mixed' represents a combined dataset of MEP from two distinct sources. Due to space constraints, we only detail the imputation process for container ship. The results of other types of ships are available in supplementary material.

5.1. Case study analysis

This section outlines the selection of independent and dependent parameters for imputation. Table 5 reports basic descriptive statistics for container ship parameters, including the number of valid observations, number of missing observations and average value. Skewness and kurtosis describe the state of data distribution. A skewness value of zero and kurtosis of three indicates a normal distribution. Our results suggest that most parameters do not follow a normal distribution.

Fig. 3 classifies missing data rates into three groups: less than 10%, between 10% and 50%, and more than 50%. Initially, variables with a missing rate exceeding 50% and non-technical variables are excluded. For variables with a missing rate under 10%, we consider them as independent variables, as model based on these yields a higher coverage rate. This approach is explained in more detail in the results section. The independent variables selected for our study are DWT, GT, LOA, Beam, and Draft. The dependent variables selected are MEP, MER, and SSP. Fig. 4 illustrates missing rates in different DWT groups. Both large and small ships are more prone to missing data. The dependent variables display similar missing rates in each group, making them suitable for the next imputation step.

The quality of the data has a significant impact on the results of the study. In the Refinitiv's database, we found out that the quality of most of parameters' known value is relatively good. However, there were some values in the data that defied common sense, such as ships produced in 1917, 2000 m ship length. This paper based on the largest merchant ship data and other literature (IMO, 2020) to derive the extreme value deletion rules, shown in Table 6.

A random selection of extreme values are also cross-checked with Clarkson's database to ensure that the rules are reasonable and practical. As mentioned in the previous section, the data need to pass Little's MCAR test to indicate that there is a potential missing pattern in the missing data. The p-value is 0, which is less than 0.05. Therefore, it can be deduced that the missing data is not MCAR and can proceed to the next step.

Before processing the missing data, we performed a Pearson correlation analysis between the parameters. Figure S1 – Figure S3 (in

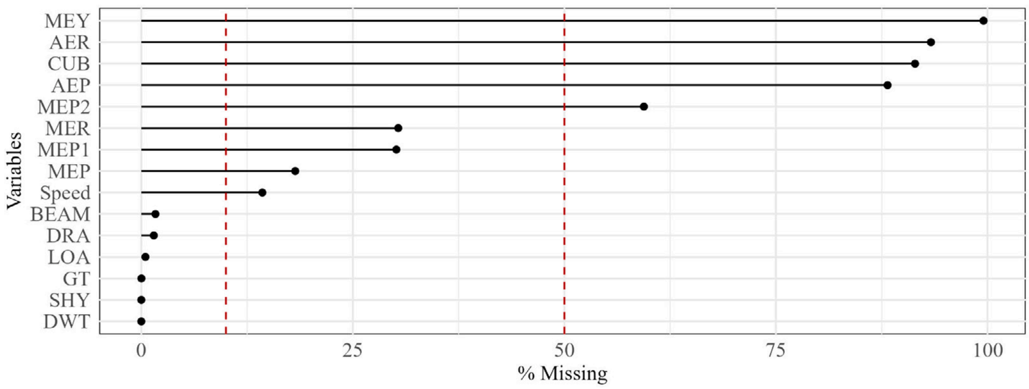


Fig. 3. Missing rates for all parameters.

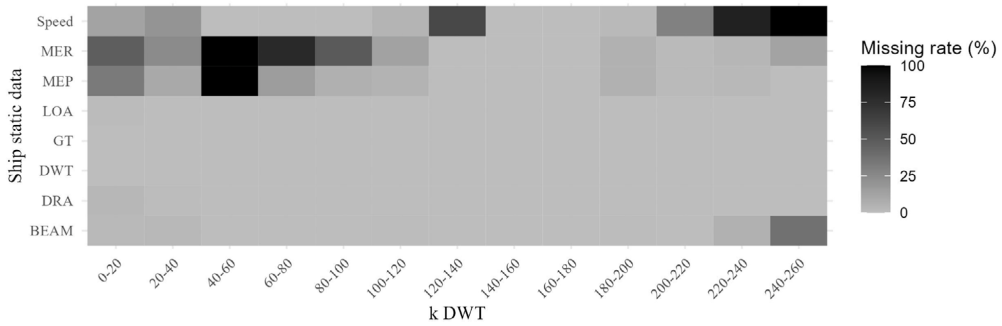


Fig. 4. Missing data distribution of ship static data.

Table 5
Descriptive statistics for container ships.

Parameters	Valid data	NA data	Mean	Std. dev	Skewness	Kurtosis
DWT	7664	0	41,488	46,517	1.66	2.36
GT	7663	1	41,221	44,704	1.74	3.17
LOA	7627	37	202	87	0.4	−0.6
BEAM	7536	128	30	11	0.58	1.25
DRA	7551	113	10	3	−0.11	−0.7
Cubic	656	7008	20,105	29,866	3.74	24.07
SSP	6568	1096	20	4	−0.49	−0.3
MEP1	5354	2310	28,564	21,939	0.67	−0.93
MEP2	3112	4552	25,673	21,653	0.88	−0.46
MER	6270	1394	26,370	21,546	0.81	−0.66
MEY	5337	2327	171	175	3.2	15.48
AEP	36	7628	1887	447	−4.05	15.25
AER	905	6759	1584	3181	14.66	260.85
AEY	512	7152	962	382	0.76	0.19
SHY	7664	0	2004	12	−1.1	1.24
MEP	7664	0	41,488	46,517	1.66	2.36

Table 6
Maximum container ship parameters.

Name	Extreme value point
DWT	400,000 DWT
GT	410,000 GT
LOA	460 m
BEAM	75 m
DRA	25 m
SSP	30 knots
MEP	81,000 kW
MER	4000 round per minute
SHY	Ship built year = 1982

the supplementary files) show three paired correlation matrices of independent and dependent parameters. These matrices illustrate the relationship among parameters (i.e., SSP, MEP, MER). The bottom left part of the figure shows the scatter plots between the parameters, showing the nonlinear relationships between the independent and dependent parameters. However, the exact relationship needs to be determined by performing a nonlinear regression. The diagonal plot illustrates the frequency distribution of each parameter. The correlation between the parameters is shown at the top right of the figure. The absolute correlation value between all parameters is greater than 0.4. The correlation between the MEP and the dependent parameters is around 0.9, which is the best. The correlation between the MER and the dependent parameters is around 0.5, which is the worst, compared with SSP and MEP. In addition, we observe that GT is highly linearly correlated with DWT, which combined with the significance of these two parameters, can be treated as one parameter. Hence, we exclude the GT.

To allow the model to be applied for different ship characteristics, the data was categorized and subjected to separate regression analyses for enhancing the precision of the outcomes. We grouped the data according to different DWT group and ship types (Kanamoto et al., 2021), the detailed grouping results are shown in Tables 7 and 8. At the same time, we test the effect of the number and type of parameters on the results. One type of model introduces only four independent parameters, while another type of model introduces all parameters into the model except the target parameter. Based on our proposed model, there are six test models as shown in Table 9.

5.2. Independent parameter imputation

This section finds the best nonlinear relationship between each parameter and the independent parameter is imputed. The results in Fig. 5 were obtained after conducting a nonlinear regression between each parameter, showing that the relationship between most variables is the power function or the logarithmic function. The adjusted R-square for the main engine power and the independent parameters were

Table 7
Regression classification DWT group.

Code	Tanker classification	DWT	Code	Bulk/container classification	DWT
T000	Small	0 – 25,000	B000/C000	Handysize + Minibulk	0 – 20,000
T025	Medium	25,000 – 50,000	B020/C020	Handymax	20,000 – 40,000
T050	Panamax	50,000 – 75,000	B040/C040	Panamax	40,000 – 65,000
T075	Aframax	75,000 – 120,000	B065/C065	Neo-Panamax	65,000 – 85,000
T120	Suezmax	120,000 – 200,000	B085/C085	Capesize	85,000 – 120,000
T200	VLCC + ULCC	200,000 +	B120/C120	VLOC	120,000 +

Table 8
Regression classification ship type group.

Tanker classification	Container classification	Bulker classification
LPG Tankers	Reefers	Minibulk
Oil Tankers	Ro-Ros	Handysize
Other Tankers	Containers	Handymax
Chemical Tankers		Panamax
LNG		OverPanamax
Other Dry		Small Capesize
		Capesize
		VLOC

Table 9
Proposed test models.

Model name	Grouping method	Input parameters
Method 9a	No group	Independent
Method 9b	No group	All
Method 9c	DWT	Independent
Method 9d	DWT	All
Method 9e	Ship type	Independent
Method 9f	Ship type	All

both above 0.8, while the adjusted R-square for the speed and the independent parameters were between 0.5 and 0.75. The adjusted R-square values for the RPM of the main engine were generally low, around 0.5, which is consistent with the results of the previous correlation analysis. This chosen nonlinear relationship is used in stepwise multiple nonlinear regression. Furthermore, the R-square of the nonlinear relationships between the independent parameters is between 0.85 and 0.95. So, we directly use these nonlinear regressions to impute missing independent parameters.

5.3. Dependent parameter imputation

In this subsection, the best nonlinear relationship obtained in the previous step is analyzed by multiple regression, using the step-forward method. The K-Fold cross-validation method allows optimal utilization of the limited training data and enables an evaluation process that assesses the model's performance on the test dataset. This method is widely accepted due to its simplicity and ability to provide more objective and conservative estimates of model performance compared to alternative approaches, such as a basic train/test split. The procedure entails the random partitioning of the set of observations into k groups or folds, each of which has an approximately equal number of elements. The initial fold is designated as the validation set, while the approach is trained using the remaining k-1 folds (James et al., 2013). In this study, the value of k is determined as 10 due to empirical evidence indicating that this particular value produces test error rate estimates that remain unaffected by both excessive bias and excessive variance (Kuhn and Johnson, 2013). After determining the training and test sets, the data is substituted into step-forward linear regression model, which select the optimal model based on the principle of minimum RMSE.

5.4. Results and discussion

After a series of calculations, the final model's performance results were obtained as shown in Appendix B. Three tables representing the model's performance on Bulk, Container and Tanker data. Each table contains representations of 14 models, Method 1–8 are models from the literature and Method 9a–9f are test models from the SMNLR method proposed in this paper. Method 1 undoubtedly performs best during training. Method 2 is GAM model, which is somewhere between ML and traditional regression models in terms of interpretation. We use it as a second benchmark. Therefore, we take these two methods as the benchmark for the performance of all training models. The metric for evaluating the model performance consists of two main components, one is the model accuracy, and the other is the model coverage rate. The model accuracy metric is the RMSE, adjusted R-squared and MAE. The '/' in the table indicates that the method is not applicable to this parameter or to this ship type.

The performance of the model varies for different ship type and imputation parameters. From an overall perspective, excluding the Random Forest model, the Method 9d has the highest accuracy in the most situation. Turning to the performance between six sub-models in Method 9. We can see that the accuracy of the grouped model is significantly higher than that of the ungrouped model. DWT group methods give the better results than ship type group methods. This can be explained that the ship static data of the different ship subtypes do not differ much for dry bulks and containers. However, for tankers we find that the ship static data for LNG and LPG vessels are significantly different from the other vessel types. They are faster, wider, and longer for the same tonnage. Therefore, for tanker's SSP have an advantage in being grouped by ship subtype. At the same time, the model with all parameters included has a higher accuracy than the model with independent parameters. The accuracy of Method 9c is similar to as Method 9d. We use the high coverage rate Method 9c as a benchmark for accuracy in Method 9 to compare with other methods. Method 6 is an IMO method, this paper uses its coverage rate as a benchmark to see how much the model's coverage rate improves.

The performance of imputing the dependent parameters is different. For SSP, Method 9c improves the imputation average accuracy by at least 2.93% depending on different ship type and compared methods. The coverage rate is improved from 0.95% to 1.99% depending on different ship type. For MER, Method 9c improves the imputation average accuracy by at least 20.74%. The coverage rate is improved 14.30% to 27.34%. For MEP Method 9c improves the imputation average accuracy by at least 2.52%, depending on different ship type. The coverage rate is improved 14.30% to 27.39%. By applying SMNLR method, MER has the greatest improvement in imputation accuracy compared to the previous method. But the SMNLR method also has a large gap compared with the ML method. SSP also has a good improvement and the SMNLR method has very little gap compared with the ML method. MEP has the smallest improvement among three ship static data in accuracy. However, the SMNLR method performs closest to the ML model. Ship type also affects model performance. The SMNLR method has the best performance for imputing the ship static data of container ships and has also achieved impressive results on bulk carriers and tankers. Especially considering the improvement in coverage rate of at average 13.84% over the Method 6. In the case of choosing a single model the method 9c is the optimal option, considering both accuracy and coverage rate. In addition to methods' overall accuracy and coverage rate, this paper also analyzes the effect of ship size and type on the accuracy of the methods. Figs. 6–8 are the bar charts of the results for each model's

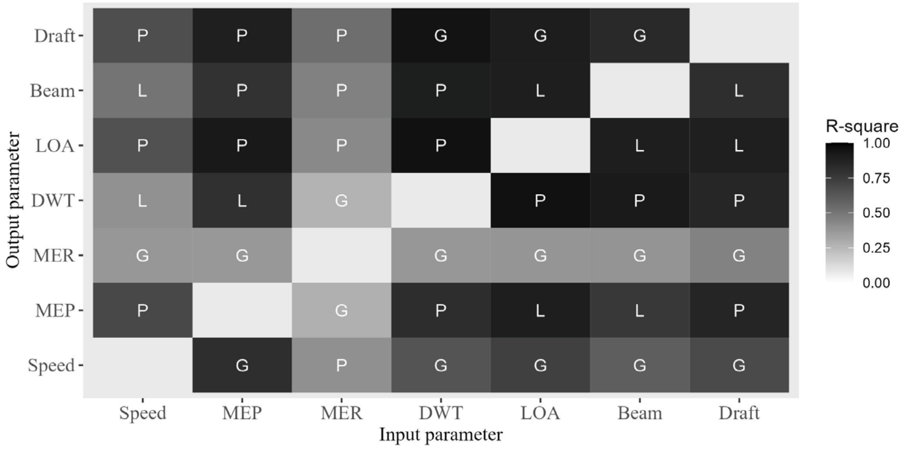


Fig. 5. Heat map for curve fitting results of dependent and independent variables (L: Linear, Q: Quadratic, C: Cubic, P: Power, G: Logarithmic, E: Exponential).

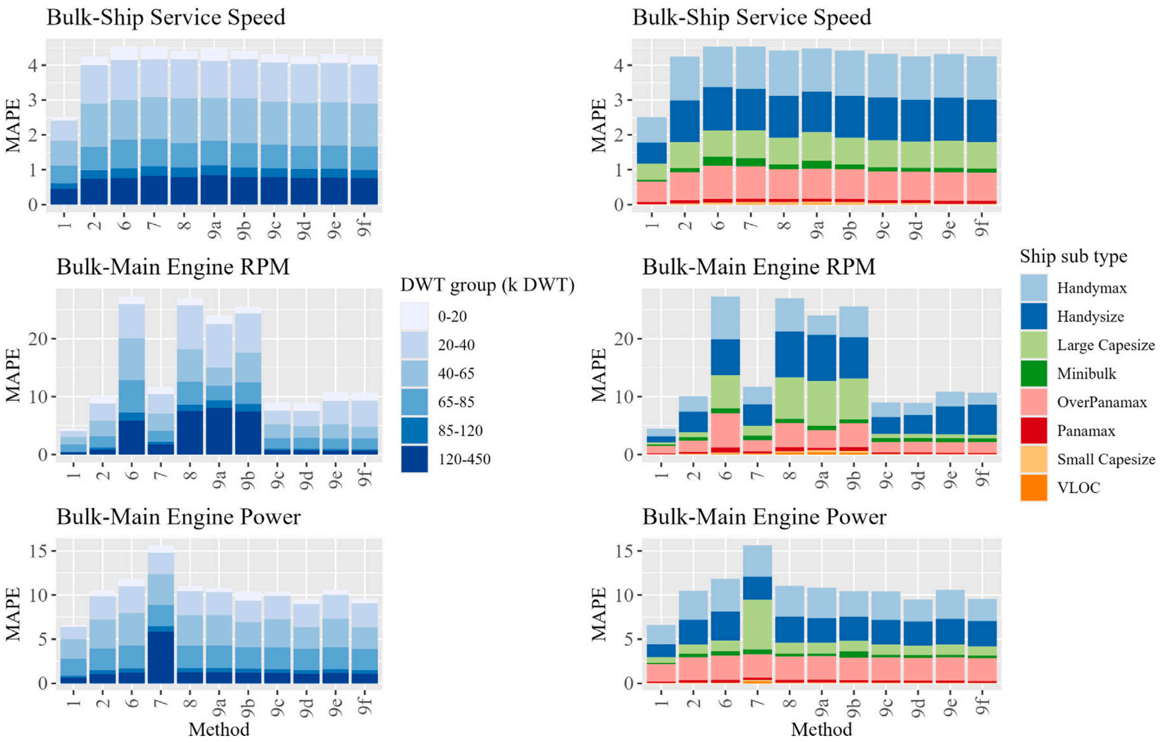


Fig. 6. Method MAPE performance of bulk.

accuracy. The *x*-axis represents the method used, and the *y*-axis is the MAPE value. The proportion of different ship sizes and ship types' MAPE has been calculated, shown in different colors in the graph. The methods that performed well in the table are selected for this analysis. Small ships contribute the largest proportion of the imputation model's MAPE. Models grouped according to ship size effectively improve the model's accuracy. Effectively reduced imputation errors in main engine RPM for larger ships. Models grouped according to ship type are only valid for specific ship types like reefer and other tanker.

In order to test the stability of the methods, this paper measures the performance of 14 methods with different missing rates of the data. The complete data previously used to train the methods was divided into two parts, 80% for training dataset and 20% for test dataset. Simulated datasets with missing rates ranging from 10% to 90% were generated through the MNAR mechanism using the training dataset. These simulated datasets will be applied to train the method and then the method will be applied on the test data. This process will be repeated 10 times and the MAE will be recorded for each time. Finally, the mean value of MAE is then used to measure the performance of different methods. The results are shown in Fig. 9. Only method 1, method 2, method 8, method 9c and method 9d are included in the figure. This is because methods with too large MAE will be excluded from the picture, showing the trend of the remaining methods more clearly. Notable among the excluded methods are method 9e and method 9f, which are modeled based on ship sub type. In most cases their errors increase dramatically when the missing data rate is higher than 40%, which is due to the fact that the sample size becomes smaller for many ship sub type leading to a decrease in model accuracy. It can be found in the figure that with the increase of missing rate, the MAE error of each method has a tendency to grow upward. Although the MAE error of Method 1 is maintained at a low level at the beginning, the growth rate of MAE error is obviously higher than that of the other methods. The performance of Method 2, Method 8, Method 9c and Method 9d is not affected much by the missing rate and basically maintains the

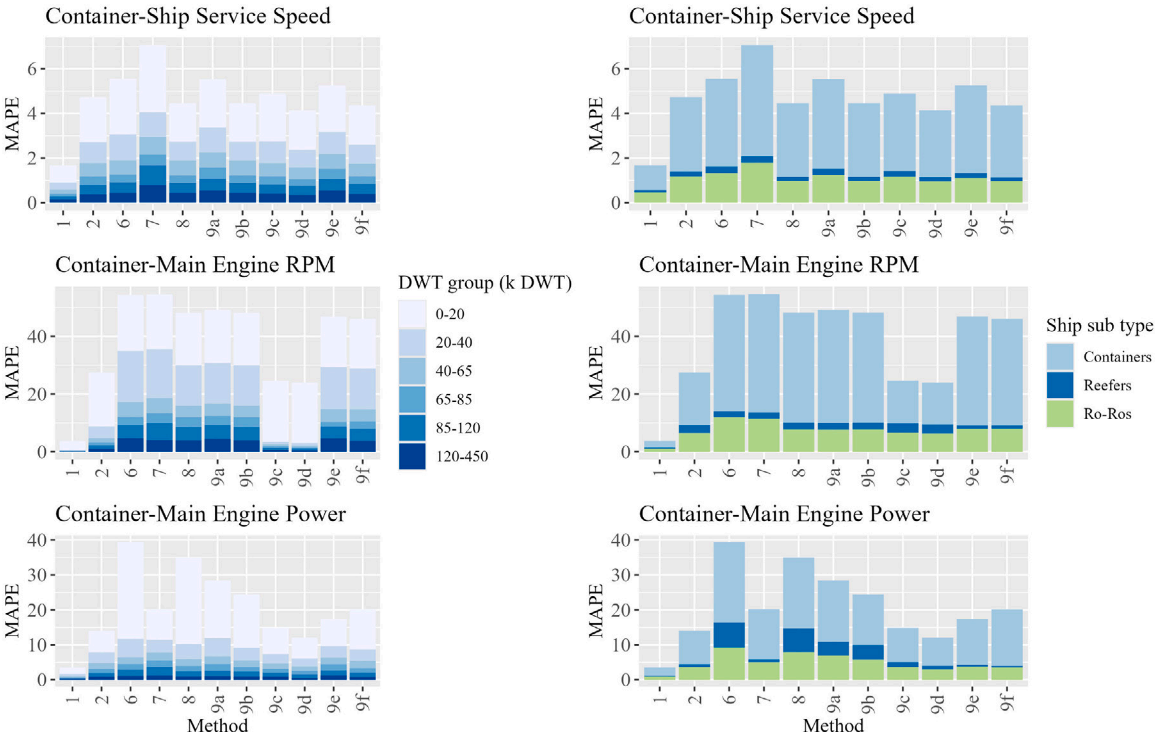


Fig. 7. Method MAPE performance of container.

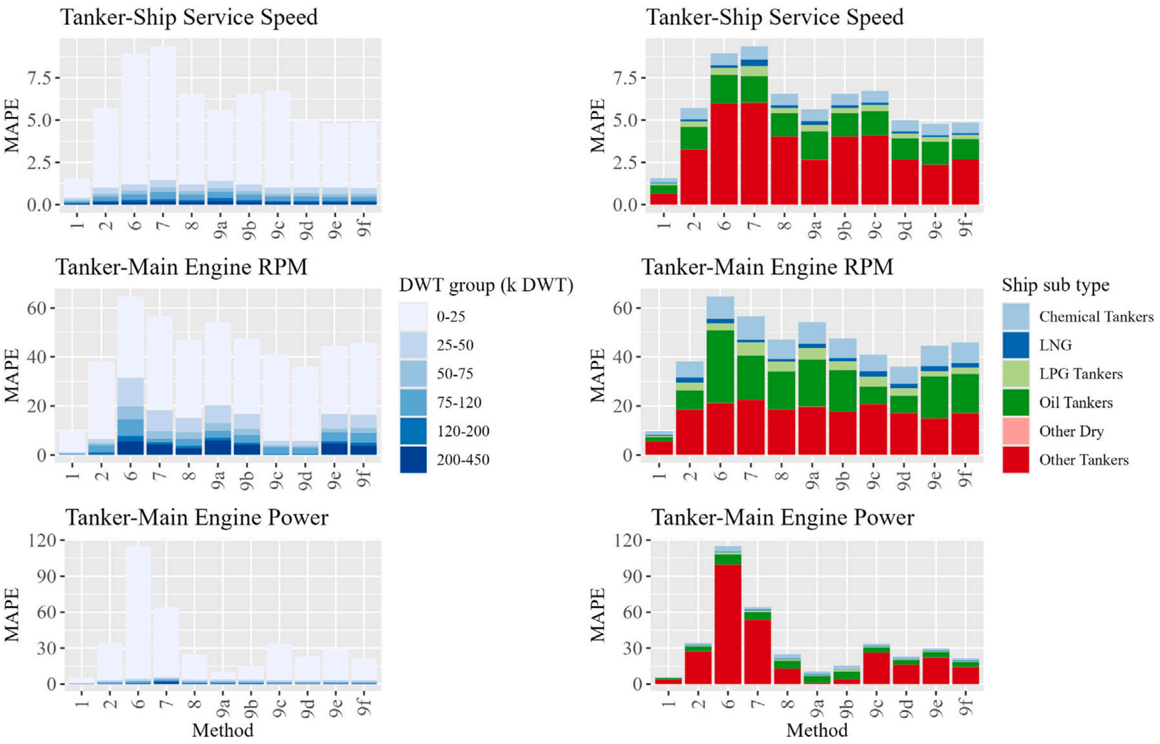


Fig. 8. Method MAPE performance of tanker.

similar level. Method 9d has the best performance among these four methods, consistent with the results of previous experiments. Method 8 has a higher error in imputing the MEP and MER, and therefore does not appear in these figures. The test results also demonstrate the impact of dataset size on the performance of various imputation methods. As the simulated data includes 10% to 90% missing values, the size of the training dataset consequently ranges from 10% to 90%. As illustrated in the figure, the method proposed in this paper maintains its performance even with a smaller training dataset size. Statistical methods are less sensitive to data quality than machine learning methods, and it is also verified that the method proposed in this paper is also effective for high missing rate data.

Furthermore, we can generate a decision matrix based on the performance of the model and choose which models to use for missing

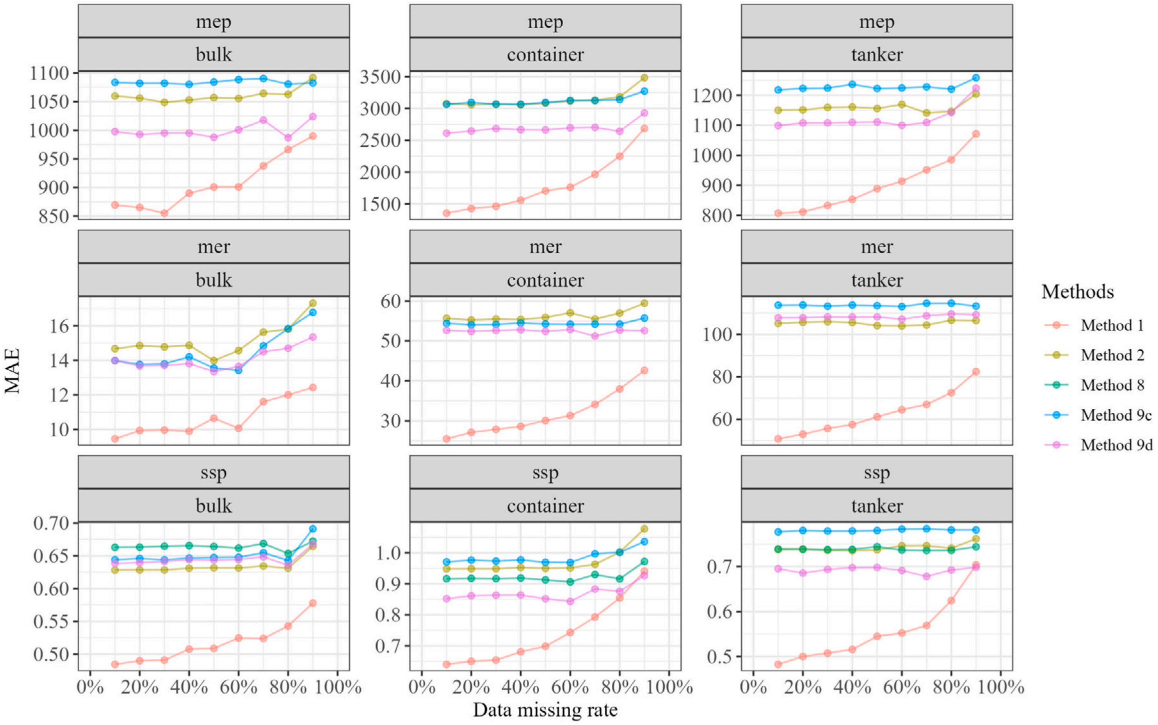


Fig. 9. MAE performance based on different data missing rate.

data. As shown in Table 10, we divided the missing data into two cases, one is when the ship is missing only the target-dependent parameters, then Method 9d and Method 9f can be used to impute the data. When the ship has only independent parameters (missing rate under 10%), Method 9c and Method 9e can be used. The selection of the imputing model based on the decision matrix allows a maximum accuracy and a 100% coverage rate. The decision matrix is applied to a missing dataset of real ships to verify its performance. The dataset consists of 3914 bulk carriers, 3353 container ships and 9021 tankers. These ships are missing at least one of SSP, MEP, MER and partially missing DWT, LOA, BEAM and DRA. Since the imputation results cannot be verified using real values, this paper utilizes fitted images to observe the performance of the decision matrix. The results are shown in Figs. 10–12. The blue points in the figure represent the complete real data, and the red points are the missing data imputed using the decision matrix. After the missing data are imputed by the decision matrix, they perfectly reproduce the characteristics of the different sizes of ships. Observing the fit figure of Aframax Tanker (T075), it can be seen that the model perfectly distinguishes the characteristics of two different types of vessels: LNG tanker and oil tanker. The model is very stable and there are no extreme values. All the imputed points are within the range of the original data. The decision matrix has some error in estimating the MER for the smaller vessels (B000, C000, T000), which is consistent with the training results. The main reason is that there is less complete data for small ships, which leads to underfitting of the model after training. However, the decision matrix has good performance for other ships.

Table 10
Imputing method decision matrix.

Ship type	Target parameters	Only missing target dependent parameter	Not missing independent parameters
Bulk	Speed	Method 9d	Method 9c
	MER	Method 9d	Method 9c
	MEP	Method 9d	Method 9c
Container	Speed	Method 9d	Method 9c
	MER	Method 9d	Method 9c
	MEP	Method 9d	Method 9c
Tanker	Speed	Method 9f	Method 9e
	MER	Method 9d	Method 9c
	MEP	Method 9d	Method 9c

5.5. Validation

The efficacy of the missing data estimates in this study is evaluated based on the statistical characteristics of the replaced values. We collected data on 1107 container ships with missing data from the Refinitiv database and used the Clarkson database for validation. Of these, 810 had missing SSP data, 196 lacked MEP data, and 564 were missing MER data. We applied 14 methods to the Refinitiv data for imputation and compared the results with Clarkson database. Method 3, 4, and 5 were excluded from validation, which did not involve MER and underperformed in imputing SSP and MEP. For quantitative assessment of the significance of differences between methods, we employed the Friedman-Nemenyi test. This test compares each algorithm's average ranking against other methods' critical difference (CD) (Demšar, 2006). If the CD ranges of the two methods do not overlap, it signifies a statistically significant difference. We input the absolute errors of all methods and actual values to obtain model rankings, with the results illustrated in Fig. 13. On the *x*-axis, the rankings reflect model accuracy, where a lower score indicates higher accuracy and fewer errors. The middle point of each line, marked in red, denotes the average ranking value, while the top and bottom points represent the CD region boundaries.

The outcomes aligned with previous training performance, yet some useful observations emerged. Method 9d surprisingly outperformed Method 1 in the SSP validation dataset. For the MER, Method 9c and 9d significantly surpassed other approaches. Regarding MEP, the performance of each method was comparatively similar, with Method 9d matching the effectiveness of Method 1, showing no significant differences between the two. The SMNLR method demonstrated practical accuracy comparable to that of ML methods. In certain instances, it even exceeded the performance of ML approaches. Additionally, these methods offer greater interpretability, allowing for a clearer understanding of the relationships between different ship static parameters

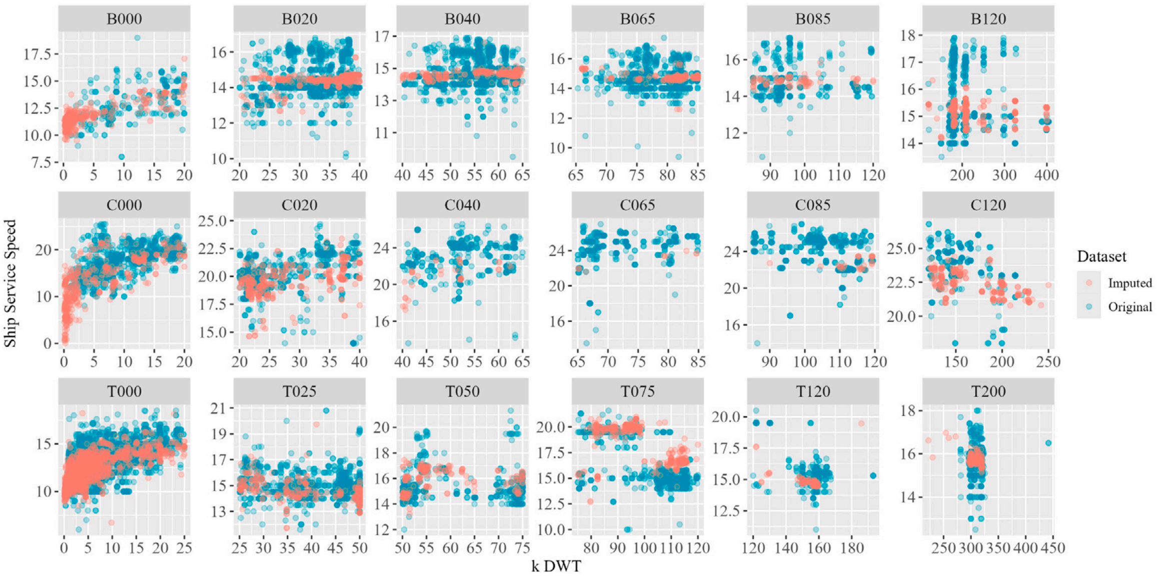


Fig. 10. Decision matrix performance on ship service speed.

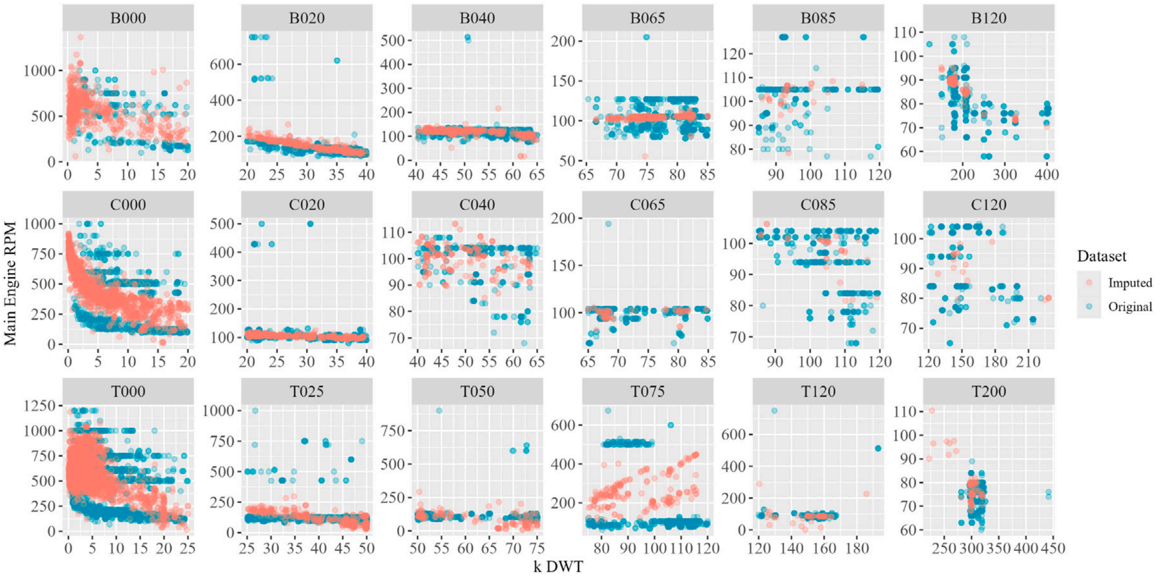


Fig. 11. Decision matrix performance on main engine RPM.

through their equations. These equations can also be used to impute missing data directly without additional training.

6. Conclusion

This study proposes the SMNLR method, using a sample of 38,018 vessels to estimate seven essential ship static parameters. For validation, k-fold cross-validation and Friedman-Nemenyi tests have been applied to the final imputing model, derived from nonlinear regression with forward stepwise selection. Unlike other methods, the SMNLR method employs only four input parameters with a low missing rates to effectively impute ship static data. Based on a large amount of complete training data, the SMNLR method generates a series of equations and regression coefficient matrices. Details and instructions for usage are provided in the Supplemental Materials. These equations and coefficients can be directly applied to impute missing static ship data in other maritime studies. For example, in port emission estimation, ship static data such as SSP, MEP and MER are required to calculate emissions generated during ship activities. In maritime safety studies, ship size and SSP data are used to calculate ship domains for safe navigation and collision avoidance. Imputing missing data improves data completeness and accuracy, enhancing the reliability of model results. In studies with small samples of ship static data, the imputation can be done directly using the training results provided in this paper, without the need to collect large amounts of additional data to train the model. The SMNLR method is also applicable for imputing missing data in other cross-sectional datasets, such as vehicle characteristic data, across different domains.

The SMNLR method demonstrated higher accuracy compared to other methods introduced in previous research. The most notable improvement was in the imputation of the main engine RPM, with adjusted R-squared values increasing by at least 20.74% compared to six other methods documented in the literature. Improvements were also observed in the imputation of ship service speed and main engine power, with increases of at least 2.93% for ship service speed and 2.52% for the main engine power, reaching precision levels comparable to machine learning (ML) algorithms However, imputing data

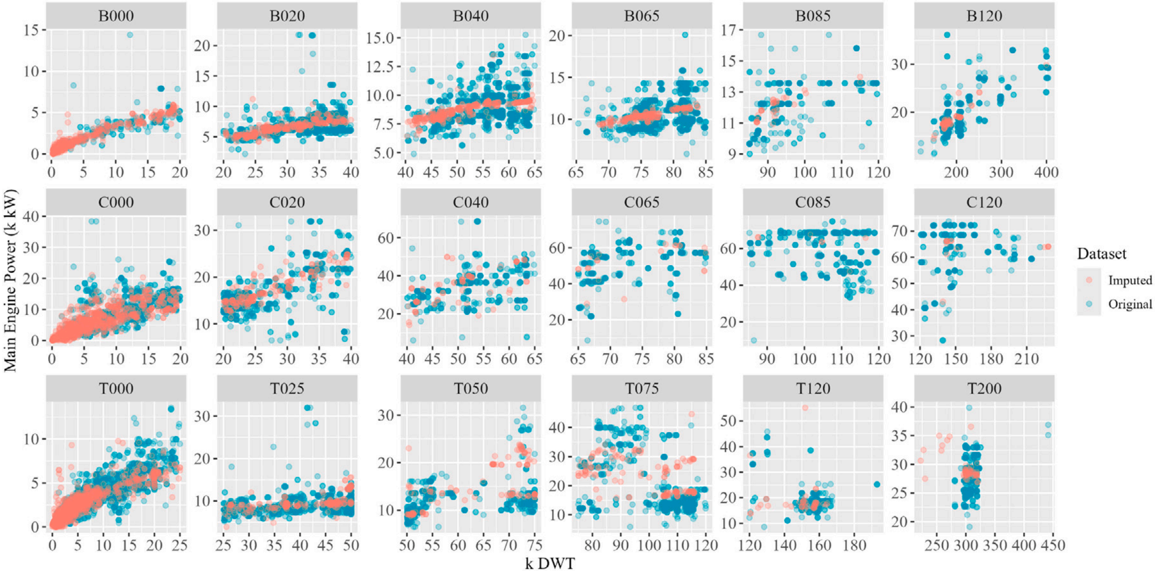
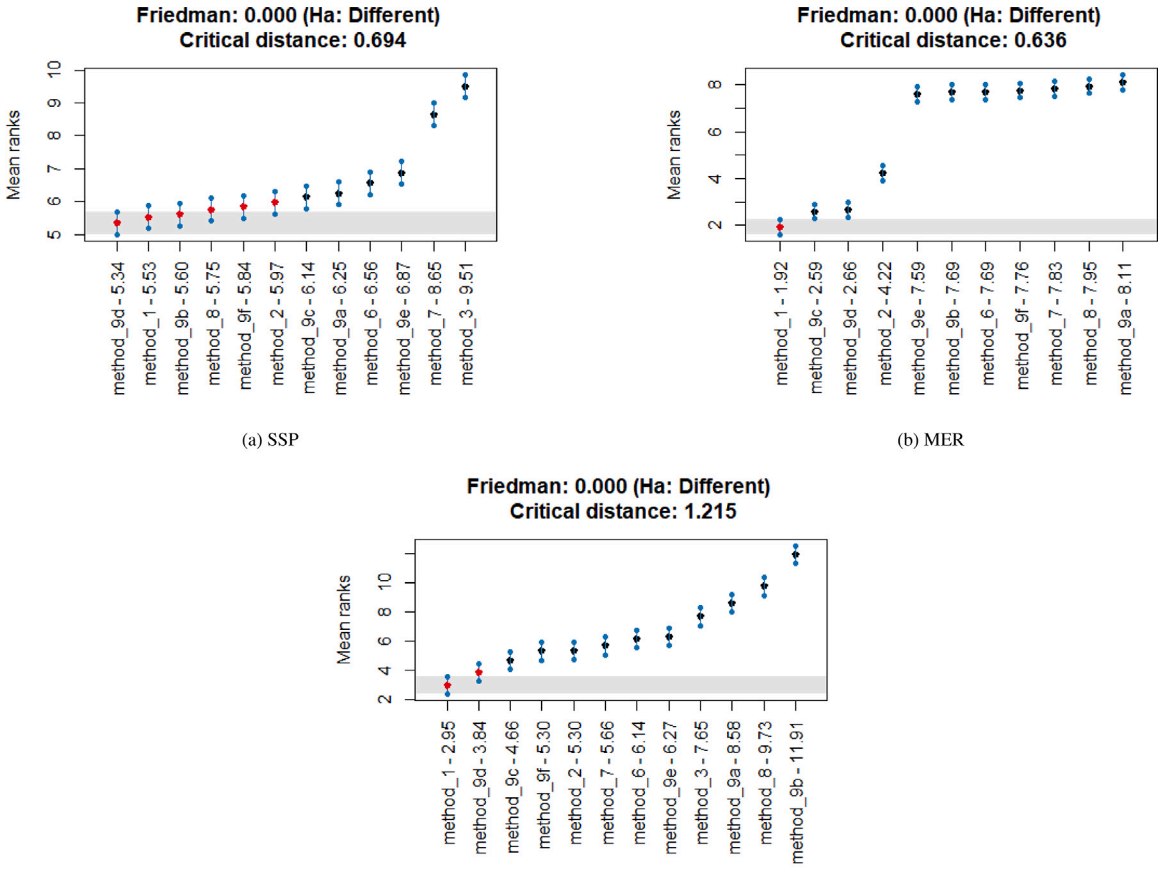


Fig. 12. Decision matrix performance on main engine power.



(a) SSP

(b) MER

(c) MEP

Fig. 13. Friedman-Nemenyi test for container validation data.

for vessels under 25k DWT showed higher error rates, indicating the need for specialized imputation techniques for small-sized vessels. The coverage rate for three dependent parameters across three types of ships improved by a minimum of 0.48% and a maximum of 27.39% compared to six methods from the literature. The proposed model demonstrated high accuracy and coverage rate across various types and sizes of ships, especially for container ships, and was validated against different databases. The SMNLR method sometimes outperformed the Random forest model in accuracy, highlighting the effectiveness of size-based group regressions. A decision matrix derived from model performance helps in selecting the most suitable method for achieving a coverage rate of 100%. Correlation analysis revealed strong relationships (around 0.9) between independent parameters and between main engine power and other independent parameters, while basic parameters showed moderate correlations (around 0.75). Nonlinear regression indicate that ship static parameters primarily follow power and logarithmic relationships.

The SMNLR model has several limitations. Firstly, the SMNLR method is more suitable for ship static data, and this paper does not test its performance on other types of maritime data. Secondly, while the SMNLR method is applied in this paper to cross-sectional data, and is theoretically valid for cross-sectional data outside the maritime domain, it is not suitable for time series data. Thirdly, this paper tests the method using container ship data from different sources, demonstrating good performance across datasets, however, it does not validate the method with bulk carriers or tankers, which requires further research. Lastly, for small ships, a specialized imputation method is needed. Future research should focus on validating the datasets imputed by the SMNLR method against actual maritime research data and consider additional target parameters in the regression analysis. The SMNLR method should also be compared across other imputation methods (e.g., multiple imputation, hot-deck, expectation–maximization-based imputation) across different data sizes and missing rates.

CRediT authorship contribution statement

Ruikai Sun: Writing – original draft, Visualization, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Wessam Abouarghoub:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Emrah Demir:** Writing – original draft, Supervision, Methodology.

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

The support provided by China Scholarship Council (No. 202108060143) and Cardiff University is acknowledged.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.oceaneng.2024.119722.

References

Abramowski, T., Cepowski, T., Zvolenský, 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.

Arlinghaus, S., 2023. Practical Handbook of Curve Fitting. CRC Press, Florida, USA.

Barrass, B., 2004. Ship Design and Performance for Masters and Mates. Elsevier, Amsterdam, The Netherlands.

Batini, C., Cappiello, C., Francalanci, C., Maurino, A., 2009. Methodologies for data quality assessment and improvement. ACM Comput. Surv. 41 (3), 1–52.

Blankers, M., Koeter, M.W., Schippers, G.M., et al., 2010. Missing data approaches in ehealth research: simulation study and a tutorial for nonmathematically inclined researchers. J. Med. Internet Res. 12 (5), e1448.

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

Cammin, P., Yu, J., Heilig, L., Voß, S., 2020. Monitoring of air emissions in maritime ports. Transp. Res. D 87, 102479.

Cepowski, T., 2019a. Determination of regression formulas for main tanker dimensions at the preliminary design stage. Ships Offshore Struct. 14 (3), 320–330.

Cepowski, T., 2019b. 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. Polish Marit. Res..

Cepowski, T., Chorab, P., 2021. Determination of design formulas for container ships at the preliminary design stage using artificial neural network and multiple nonlinear regression. Ocean Eng. 238, 109727.

Charchalis, A., 2013. Dimensional constraints in ship design. J. KONES 20, 29–34.

Charchalis, A., 2014. Determination of main dimensions and estimation of propulsion power of a ship. J. KONES 21, 39–44.

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.

Chen, S., Meng, Q., Jia, P., Kuang, H., 2021. An operational-mode-based method for estimating ship emissions in port waters. Transp. Res. D 101, 103080.

Cheong, R.C.K., Lim, J.M.-Y., Parthiban, R., 2023. Missing traffic data imputation for artificial intelligence in intelligent transportation systems: review of methods, limitations, and challenges. IEEE Access 11, 34080–34093.

Christiansen, M., Hellsten, E., Pisinger, D., Sacramento, D., Vilhelmsen, C., 2020. Liner shipping network design. European J. Oper. Res. 286 (1), 1–20.

Demšar, J., 2006. Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. 7, 1–30.

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), 111–136.

Dong, Y., Peng, C.-Y.J., 2013. Principled missing data methods for researchers. SpringerPlus 2, 1–17.

Du, Y., Chen, Q., Quan, X., Long, L., Fung, R.Y., 2011. Berth allocation considering fuel consumption and vessel emissions. Transp. Res. E 47 (6), 1021–1037.

Duan, H., Ma, F., Miao, L., Zhang, C., 2022. A semi-supervised deep learning approach for vessel trajectory classification based on AIS data. Ocean & Coastal Management 218, 106015.

Enders, C.K., 2010. Applied Missing Data Analysis. Guilford Press, Buckinghamshire, United Kingdom.

Enders, C.K., 2022. Applied Missing Data Analysis. Guilford Publications.

Gao, J., Cai, Z., Sun, W., Jiao, Y., 2023. A novel method for imputing missing values in ship static data based on generative adversarial networks. J. Mar. Sci. Eng. 11 (4), 806.

Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M., Kagal, L., 2018. Explaining explanations: An overview of interpretability of machine learning. In: 2018 IEEE 5th International Conference on Data Science and Advanced Analytics. DSAA, IEEE, pp. 80–89.

Guo, S., Mou, J., Chen, L., Chen, P., 2021. Improved kinematic interpolation for AIS trajectory reconstruction. Ocean Eng. 234, 109256.

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.

Hastie, T.J., 2017. Generalized additive models. In: Statistical Models in S. Routledge, Oxfordshire, United Kingdom, pp. 249–307.

He, W., Lei, J., Chu, X., Xie, S., Zhong, C., Li, Z., 2021. A visual analysis approach to understand and explore quality problems of AIS data. J. Mar. Sci. Eng. 9 (2), 198.

Hocking, R.R., 1976. A biometrics invited paper. The analysis and selection of variables in linear regression. Biometrics 1–49.

Huang, L., Wen, Y., Geng, X., Zhou, C., Xiao, C., 2018. Integrating multi-source maritime information to estimate ship exhaust emissions under wind, wave and current conditions. Transp. Res. D 59, 148–159.

Huang, L., Wen, Y., Zhang, Y., Zhou, C., Zhang, F., Yang, T., 2020. Dynamic calculation of ship exhaust emissions based on real-time AIS data. Transp. Res. D 80, 102277.

IMO, 2020. Fourth IMO Greenhouse Gas Study 2020. Report, International Maritime Organization, URL: https://greenvoyage2050.imo.org/wp-content/uploads/2021/07/Fourth-IMO-GHG-Study-2020-Full-report-and-annexes_compressed.pdf. (Accessed 17 April 2024).

IMO, 2021. Calculation of the Attained Energy Efficiency Existing Ship Index (EEXI) Resolution MEPC.333(76). Report, International Maritime Organization, URL: https://wwwcdn.imo.org/localresources/en/OurWork/Environment/Documents/Air%20pollution/MEPC.333(76).pdf. (Accessed 17 April 2024).

Jakobsen, J.C., Gluud, C., Wetterslev, J., Winkel, P., 2017. When and how should multiple imputation be used for handling missing data in randomised clinical trials–a practical guide with flowcharts. BMC Med. Res. Methodol. 17 (1), 1–10.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An Introduction to Statistical Learning. Vol. 112, Springer, New York, USA.

Jenelius, E., 2019. Data-driven metro train crowding prediction based on real-time load data. IEEE Trans. Intell. Transp. Syst. 21 (6), 2254–2265.

Jeon, M., Noh, Y., Jeon, K., Lee, S., Lee, I., 2021. Data gap analysis of ship and maritime data using meta learning. Appl. Soft Comput. 101, 107048.

Kanamoto, K., Murong, L., Nakashima, M., Shibasaki, R., 2021. Can maritime big data be applied to shipping industry analysis? focussing on commodities and vessel sizes of dry bulk carriers. Marit. Econ. Logist. 23 (2), 211–236.

Kelly, P., 2022. A novel technique to identify AIS transmissions from vessels which attempt to obscure their position by switching their AIS transponder from normal transmit power mode to low transmit power mode. Expert Syst. Appl. 202, 117205.

Kim, S.-H., Roh, M.-I., Oh, M.-J., Park, S.-W., Kim, I.-I., 2020. Estimation of ship operational efficiency from AIS data using big data technology. Int. J. Nav. Archit. Ocean Eng. 12, 440–454.

Kim, Y., Steen, S., Muri, H., 2022. A novel method for estimating missing values in ship principal data. Ocean Eng. 251, 110979.

Ksciuk, J., Kuhlemann, S., Tierney, K., Koberstein, A., 2023. Uncertainty in maritime ship routing and scheduling: A literature review. European J. Oper. Res. 308 (2), 499–524.

Kuhn, M., Johnson, K., 2013. Applied Predictive Modeling. Vol. 26, Springer, New York, USA.

Last, P., Bahlke, C., Hering-Bertram, M., Linsen, L., 2014. Comprehensive analysis of automatic identification system (AIS) data in regard to vessel movement prediction. J. Navig. 67 (5), 791–809.

Lee, J.H., Huber, Jr., J., 2011. Multiple imputation with large proportions of missing data: How much is too much? In: United Kingdom Stata Users' Group Meetings 2011. Stata Users Group.

Liang, M., Su, J., Liu, R.W., Lam, J.S.L., 2024. Aisclean: AIS data-driven vessel trajectory reconstruction under uncertain conditions. Ocean Eng. 306, 117987.

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.

Martin-Iradi, B., Pacino, D., Ropke, S., 2024. An adaptive large neighborhood search heuristic for the multi-port continuous berth allocation problem. European J. Oper. Res. 152–167.

McArthur, D.P., Osland, L., 2013. Ships in a city harbour: An economic valuation of atmospheric emissions. Transp. Res. D 21, 47–52.

Merien-Paul, R.H., Enshaei, H., Jayasinghe, S.G., 2018. In-situ data vs. bottom-up approaches in estimations of marine fuel consumptions and emissions. Transp. Res. D 62, 619–632.

Mi, J.-X., Li, A.-D., Zhou, L.-F., 2020. Review study of interpretation methods for future interpretable machine learning. IEEE Access 8, 191969–191985.

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

Muthén, B., Kaplan, D., Hollis, M., 1987. On structural equation modeling with data that are not missing completely at random. Psychometrika 52 (3), 431–462.

Nguyen, V.-S., Im, N.-k., Lee, S.-m., 2015. The interpolation method for the missing AIS data of ship. J. Navig. Port Res. 39 (5), 377–384.

Nguyen, D., Vadaine, R., Hajduch, G., Garello, R., Fablet, R., 2018. A multi-task deep learning architecture for maritime surveillance using AIS data streams. In: 2018 IEEE 5th International Conference on Data Science and Advanced Analytics. DSAA, IEEE, pp. 331–340.

Nguyen, P.-N., Woo, S.-H., Kim, H., 2022. Ship emissions in hotelling phase and loading/unloading in southeast Asia ports. Transp. Res. D 105, 103223.

O'Hagan, A., 1978. Curve fitting and optimal design for prediction. J. R. Stat. Soc. Ser. B Stat. Methodol. 40 (1), 1–24.

Pantanowitz, A., Marwala, T., 2009. Missing data imputation through the use of the random forest algorithm. In: Advances in Computational Intelligence. Springer, pp. 53–62.

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

Peng, X., Wen, Y., Wu, L., Xiao, C., Zhou, C., Han, D., 2020. A sampling method for calculating regional ship emission inventories. Transp. Res. D 89, 102617.

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

Prais, S.J., Aitchison, J., 1954. The grouping of observations in regression analysis. Revue Inst. Int. Stat. 1–22.

Raeesi, R., Sahebjamnia, N., Mansouri, S.A., 2023. The synergistic effect of operational research and big data analytics in greening container terminal operations: A review and future directions. European J. Oper. Res. 310 (3), 943–973.

Raghunathan, T.E., 2004. What do we do with missing data? Some options for analysis of incomplete data. Annu. Rev. Public. Health 25, 99–117.

Ravindra, K., Rattan, P., Mor, S., Aggarwal, A.N., 2019. Generalized additive models: Building evidence of air pollution, climate change and human health. Environ. Int. 132, 104987.

Reinhardt, L.B., Pisinger, D., Sigurd, M.M., Ahmt, J., 2020. Speed optimizations for liner networks with business constraints. European J. Oper. Res. 285 (3), 1127–1140.

Rinauro, B., Begovic, E., Mauro, F., Rosano, G., 2024. Regression analysis for container ships in the early design stage. Ocean Eng. 292, 116499.

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

Rubin, D.B., Little, R.J., 2019. Statistical Analysis with Missing Data. John Wiley & Sons, New Jersey, USA.

Sang, L.-z., Wall, A., Mao, Z., Yan, X.-p., Wang, J., 2015. A novel method for restoring the trajectory of the inland waterway ship by using AIS data. Ocean Eng. 110, 183–194.

Santos, M.S., Pereira, R.C., Costa, A.F., Soares, J.P., Santos, J., Abreu, P.H., 2019. Generating synthetic missing data: A review by missing mechanism. IEEE Access 7, 11651–11667.

Schwarzkopf, D.A., Petrik, R., Matthias, V., Quante, M., Majamäki, E., Jalkanen, J.-P., 2021. A ship emission modeling system with scenario capabilities. Atmos. Environ. X 12, 100132.

Shen, W.W., Ren, J.M., 2014. Multiple stepwise regression analysis crack open degree data in gravity dam. In: Applied Mechanics and Materials. Vol. 477, Trans Tech Publ, pp. 888–891.

Shepperson, J.L., Hintzen, N.T., Szostek, C.L., Bell, E., Murray, L.G., Kaiser, M.J., 2018. A comparison of VMS and AIS data: The effect of data coverage and vessel position recording frequency on estimates of fishing footprints. ICES J. Mar. Sci. 75 (3), 988–998.

Skarlatos, K., Papageorgiou, G., Biris, P., Skamnia, E., Economou, P., Bersimis, S., 2024. Ship engine model selection by applying machine learning classification techniques using imputation and dimensionality reduction. J. Mar. Sci. Eng. 12 (1), 97.

Stead, A.D., Wheat, P., 2020. The case for the use of multiple imputation missing data methods in stochastic frontier analysis with illustration using English local highway data. European J. Oper. Res. 280 (1), 59–77.

Sun, R., Abouarghoub, W., Demir, E., Potter, A., 2025. A comprehensive analysis of strategies for reducing GHG emissions in maritime ports. Mar. Policy 171, 106455. http://dx.doi.org/10.1016/j.marpol.2024.106455, URL: https://www.sciencedirect.com/science/article/pii/S0308597X2400455X.

Tang, F., Ishwaran, H., 2017. Random forest missing data algorithms. Stat. Anal. Data Min.: ASA Data Sci. J. 10 (6), 363–377.

Thiyagalingam, J., Shankar, M., Fox, G., Hey, T., 2022. Scientific machine learning benchmarks. Nat. Rev. Phys. 4 (6), 413–420.

Tichavska, M., Tovar, B., 2015. Environmental cost and eco-efficiency from vessel emissions in Las Palmas Port. Transp. Res. E 83, 126–140.

Umang, N., Bierlaire, M., Vacca, I., 2013. Exact and heuristic methods to solve the berth allocation problem in bulk ports. Transp. Res. E 54, 14–31.

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.

Wawrzyniak, J., Drozdowski, M., Sanlaville, E., 2020. Selecting algorithms for large berth allocation problems. European J. Oper. Res. 283 (3), 844–862.

Xu, H., Yang, D., 2020. LNG-fuelled container ship sailing on the Arctic Sea: Economic and emission assessment. Transp. Res. D 87, 102556.

Yan, R., Wang, S., Du, Y., 2020. Development of a two-stage ship fuel consumption prediction and reduction model for a dry bulk ship. Transp. Res. E 138, 101930.

Yan, R., Wang, S., Psaraftis, H.N., 2021. Data analytics for fuel consumption management in maritime transportation: Status and perspectives. Transp. Res. E 155, 102489.

Yang, D., Liao, S., Lun, Y.V., Bai, X., 2023. Towards sustainable port management: Data-driven global container ports turnover rate assessment. Transp. Res. E 175, 103169.

Yang, D., Wu, L., Wang, S., Jia, H., Li, K.X., 2019. How big data enriches maritime research–a critical review of Automatic Identification System (AIS) data applications. Transp. Rev. 39 (6), 755–773.

Yu, Y., Sun, R., Sun, Y., Shu, Y., 2022a. Integrated carbon emission estimation method and energy conservation analysis: the Port of Los Angles case study. J. Mar. Sci. Eng. 10 (6), 717.

Yu, Y., Sun, R., Sun, Y., Wu, J., Zhu, W., 2022b. China's port carbon emission reduction: A study of emission-driven factors. Atmosphere 13 (4), 550.

Zhang, Y., Fung, J.C., Chan, J.W., Lau, A.K., 2019. The significance of incorporating unidentified vessels into AIS-based ship emission inventory. Atmos. Environ. 203, 102–113.