Journal of Marine Science and Engineering

MDPI

Article

Estimation of Tanker Ships' Lightship Displacement Using Multiple Linear Regression and XGBoost Machine Learning

Vlado Frančić [1,*], Nermin Hasanspahić [2,*], Mario Mandušić [3], and Marko Strabić [1]

[1] Faculty of Maritime Studies, University of Rijeka, 51000 Rijeka, Croatia
[2] Maritime Department, University of Dubrovnik, 20000 Dubrovnik, Croatia
[3] Independent researcher, 20000 Dubrovnik, Croatia
* Correspondence: vlado.francic@uniri.hr (V.F.); nermin.hasanspahic@unidu.hr (N.H.)

Abstract: It is of the utmost importance to accurately estimate different ships' weights during their design stages. Additionally, lightship displacement (LD) data are not always easily accessible to shipping stakeholders, while other ships' dimensions are within hand's reach (for example, through data from the online Automatic Identification System (AIS)). Therefore, determining lightship displacement might be a difficult task, and it is traditionally performed with the help of mathematical equations developed by shipbuilders. Distinct from the traditional approach, this study offers the possibility of employing machine learning methods to estimate lightship displacement weight as accurately as possible. This paper estimates oil tankers' lightship displacement using two ships' dimensions, length overall, and breadth. The dimensions of oil tanker ships were collected from the INTERTANKO Chartering Questionnaire Q88, available online, and, because of similar block coefficients, all tanker sizes were used for estimation. Furthermore, multiple linear regression and extreme gradient boosting (XGBoost) machine learning methods were utilised to estimate lightship displacement. Results show that XGBoost and multiple linear regression machine learning methods provide similar results, and both could be powerful tools for estimating the lightship displacement of all types of ships.

Keywords: oil tanker; lightship displacement; length overall; breadth; machine learning; XGBoost

Citation: Frančić, V.; Hasanspahić, N.; Mandušić, M.; Strabić, M. Estimation of Tanker Ships' Lightship Displacement Using Multiple Linear Regression and XGBoost Machine Learning. *J. Mar. Sci. Eng.* **2023**, *11*, 961. https://doi.org/10.3390/jmse11050961

Academic Editor: Carlos Guedes Soares

Received: 22 March 2023
Revised: 27 April 2023
Accepted: 28 April 2023
Published: 30 April 2023

1. Introduction

During the preliminary and final stages of ship design, estimating the various weight groups and positions of the ship's centroids is fundamental. Any miscalculations and omissions might significantly influence the ship's cargo transport capacity, speed, stability, and overall safety during a voyage [1]. Another important reason for accurate estimation is the relationship between a ship's construction cost and its weight, especially its structural steel weight. Therefore, it is necessary to estimate the weights of different groups of ships as accurately as possible during the preliminary design of the ship. Furthermore, a shipbuilder's initial tender conditions to a shipowner depend on the accuracy of the estimation of various steel weight groups [1].

According to the International Convention for the Safety of Life at Sea (SOLAS II-1/2.21 and SOLAS II-2/3.28), "lightweight is the displacement of a ship in tonnes without cargo, fuel, lubricating oil, ballast water, fresh water and feedwater in tanks, consumable stores, and passengers and crew and their effects" [2]. "Lightship condition is a ship complete in all respects, but without consumables, stores, cargo, crew and effects, and without any liquids on board except that machinery and piping fluids, such as lubricants and hydraulics, are at operating levels" [3].

Lightship condition is not important just for ship-stability reasons. For example, in the ship recycling industry, ships are valued primarily based on their lightship displace-

ment (LD), that is, the weight of the ship when ready for service [4]. It is the most important unit of measure in the shipbreaking process [5,6]. The price of a ship sold for scrapping is usually quoted in USD ($) per ton of LD. Furthermore, based on lightweight displacement, the recycling shipyard calculates the estimated recycling costs and plans recycling activities [7]. However, lightship weight data are often not easily accessible (for example, if a buyer of a ship to be recycled wants to check its weight before price negotiations between brokers). Therefore, this paper aims to develop a lightship displacement (LD) estimation model based on two easily accessible ship dimensions: length overall (LOA) and breadth (B). Length overall (LOA) is a ship's length measured from the foremost point of its stem to the aftermost part of its stern. Breadth (B) is a ship's width at its widest point. It must be noted that a ship's length and breadth are correlated dimensions, and length to breadth (L/B) is a dimensionless ratio that is different for different ship types. Typical values for cargo ships are 6.0–7.0, those of tankers and bulk carriers are 5.5–6.5, and those of passenger ships are 6.0–8.0 [8]. Larger values of this ratio are favourable for a ship's speed but unfavourable for its manoeuvrability. LOA and B are easily accessible data that can be accessed from various sources, the most popular being the various Automatic Identification System (AIS) platforms available online. In this way, lightship weight for a specific ship might be obtained using those two ship dimensions.

Čudina, in [9], presented design procedures and mathematical modelling that could be applicable in the initial design stages of merchant ships. Mathematical models presented in the paper include the optimisation of the main characteristics of bulk carriers and tankers and the optimisation of the commercial effects of new buildings. In [10], Duru developed a method for estimating the lightship weight and deadweight of a projected fishing-vessel design. This newly developed method includes several existing methods, such as the cubic number method, the rate per meter method, and the rational equation method, and is recommended by various classification societies. Lin and Shaw [11] presented a feature-based estimation (FSE) method to estimate a ship's steel weight and centre of gravity in the preliminary designing phase. Their approach utilises principal component analysis (PCA) to identify principal parameters from the ship's parameters and its main structural components as well as to develop equations for estimating steel weight. Furthermore, a regression that is based on each structural section's characteristics is used to adjust the estimated weight. In [12], the authors analysed 58 ships of different types and capacities and systematically investigated their lightship weight distribution. The authors determined the limiting lightship longitudinal and vertical centre of gravity range for bulk carriers, crude oil tankers, liquefied gas carriers (LGC), container ships, and pure car carriers (PCC). Obreja and Chiroşcă, in [13], developed a PHP-Ship Weight computer code, aiming to estimate the components of the lightship and the deadweight of ships. The estimation of the ship's weight components in their software is based on parametric models. In [14], the author determined regression formulas for the main dimensions of bulk carriers and tankers. Tankers and bulk carriers were categorised into groups according to size, and regression formulas were developed for each group separately. In [15], the authors developed an artificial neural network (ANN) to predict the main particulars of a chemical tanker at the preliminary design stage. The obtained results are in good agreement with actual data and corroborate that the model can be used to predict the main dimensions of chemical tankers; however, its applications for chemical tankers with innovative designs might be inadequate. An empirical model to predict the lightship displacement of jack-up rigs was developed in [16]. The authors compiled a lightship displacement dataset for jack-up rigs and utilised linear regression to develop a model to estimate lightship displacement in jack-up rigs. Their model explains 91% of the variation in lightship displacement. A model to calculate lightship weight using empirical methods was presented in [17]. The authors presented and analysed existing frameworks for estimating lightship weight based on empirical data from existing ships. The study integrated and calibrated a number of empirical methods and established the use of the least squares method for determining the method that identifies lightship values closest to the ones observed in

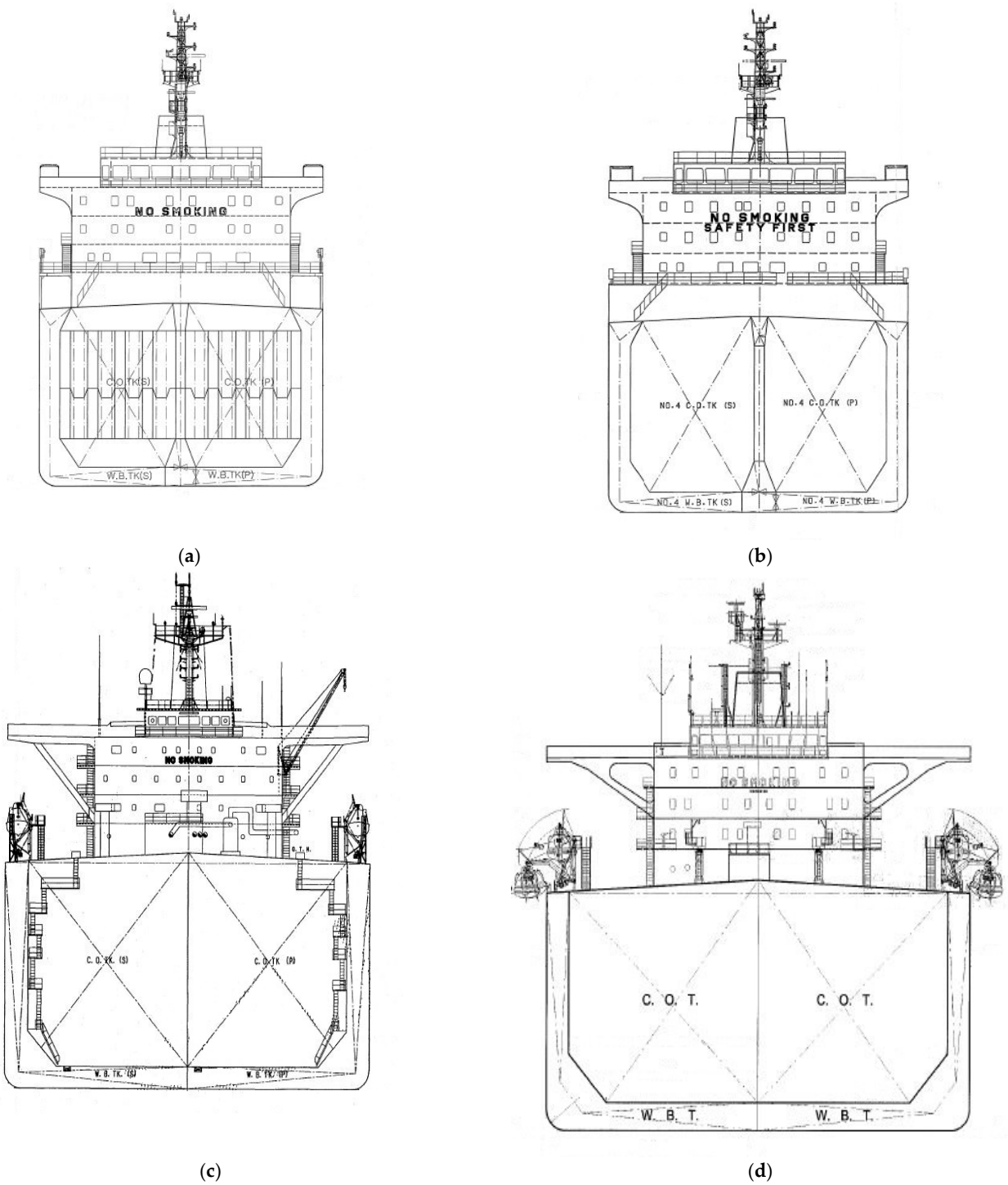
databases. Lightship weight was estimated for three types of cargo ships, tankers, bulk carriers, and container ships. Furthermore, it is necessary to mention a study by Cepowski [18]. The author developed regression formulas for main tanker dimensions (length between perpendiculars—LBP, breadth—B, and draught moulded—T). Regression formulas were developed separately for Handysize, Medium Range, Panamax, Post Panamax, Aframax, Suezmax, and VLCC tankers, and the variables used in the study were deadweight and velocity. In [19], the authors used an artificial neural network (ANN) and multiple nonlinear regression to estimate the length between perpendiculars for container ships. The variables used in the study were the twenty-foot equivalent unit (TEU) and velocity of the ships.

However, to the best of the authors' knowledge, there is no research in the literature aimed at estimating the lightship weight of merchant ships that employs only two characteristics (dimensions) of the ship, namely length overall and breadth. Therefore, this paper focuses on estimating the lightship displacement (LD) of existing merchant ships, specifically oil tankers, utilising linear regression and extreme gradient boosting machine learning (XGBoost ML) methodologies to analyse the known dimensions (LOA and B) of a tanker ship.

The structure of tanker ships is specific because of the nature of the dangerous cargo they transport—primarily crude oil and its products. Because tanker ship accidents can cause catastrophic consequences, including fatalities, injuries, loss of ships and cargo, and substantial environmental pollution, the International Maritime Organization (IMO) has introduced additional safety measures in the form of structural improvements (for example, double hull). In addition, the International Association of Classification Societies (IACS) adopted the Common Structural Rules for Bulk Carriers and Oil Tankers to ensure maximum structural safety for oil tankers; the latest rule entered into force in July 2019. Although the weight of marine steel is optimised to reduce the weight of an empty ship so that it can carry as much cargo as possible, introducing new rules to increase safety has resulted in a 2% to 8% increase in lightship weight [20,21]. This research focuses on oil tankers because they exhibit specific structures as a result of the dangerous nature of the cargo they transport and of the improved construction standards implemented by the International Maritime Organization (IMO) and the IACS. This has resulted in oil tankers having a higher steel content compared to other types of ships, making them desirable in ship-recycling industries such as the Bangladesh shipbreaking industry [22].

Tanker ships can be categorised into five major groups according to their size (Figure 1) [9]:

- Handy-sized oil tankers of up to 45,000–50,000 deadweight (DWT) with a breadth (B) limited by the ability to pass through the Panama Canal;
- Panamax-sized oil tankers that are built to pass through the Panama Canal and, in most cases, with an LOA limited to 228.6 m;
- Aframax-sized oil tankers with an approximate DWT between 80,000 and 110,000 tons at the maximum draught;
- Suezmax-sized oil tankers with an approximate DWT between 150,000 and 170,000 tons;
- Very large crude carriers (VLCC), or a group of oil tankers of approximately 300,000 DWT.



(**a**) (**b**)

(**c**) (**d**)

Figure 1. Midship sections of (**a**) handy-sized, (**b**) Panamax, (**c**) Aframax, and (**d**) Suezmax tankers.

However, lightship displacement was not estimated separately for each group of oil tankers in this paper because of their similar hull designs and block coefficients (see Table 1), but tankers of all sizes were included in the research.

Table 1. Block coefficients of tanker groups [23].

Tanker Group	Handy	Panamax	Aframax	Suezmax	VLCC
Block coeff. (C_B)	0.747–0.763	0.817	0.797	0.799	0.788

The authors developed a framework (using multiple linear regression and extreme gradient boosting machine learning, as shown in Sections 3 and 4) through which a tanker's lightship displacement can be estimated using known variables (length overall and breadth) and a dummy variable (tanker size group). In this way, existing tankers' lightship weight estimation might be facilitated, enabling easier and more accurate estimations for interested stakeholders, such as ship designers and shipowners.

The rest of the paper is organised as follows: Section 2 presents formulas for the estimation of the lightship weight of merchant ships. Section 3 is Methodology, where the logic of the research procedure and methods is described. Finally, research results and a discussion of them are given in Section 4, while concluding thoughts and future research directions in this area are given in Section 5.

2. Lightship Weight Calculations

As mentioned in the introduction section, an estimation that is as accurate as possible of the weight components of a ship is fundamental during ship design because any inaccuracies might have a major effect on the ship's safety [1]. Therefore, it is of crucial importance to estimate various ship components' weights as accurately as possible. In this section, the authors present equations adopted from [1] to calculate ship components' weights.

A ship's displacement can be calculated using Equation (1):

$$\Delta = DWT + W_{LS} \quad (1)$$

where Δ is the weight of displaced water, DWT is deadweight (transport capacity), and W_{LS} is lightship weight (weight of an empty ship). Deadweight can be calculated using Equation (2):

$$DWT = W_{LO} + W_F + W_{PR} + W_P + W_{CR} + B \quad (2)$$

where W_{LO} is payload weight, W_F is fuel weight including lubricating oil, W_{PR} is the weight of provisions and freshwater supplies, W_P is the weight of passengers (if any) and their effects (luggage), W_{CR} is the weight of crew and their effects, and B is the weight of non-permanent ballast.

For the purposes of the estimate, lightship weight can be considered as the sum of three main components and the margin of uncertainty:

$$W_{LS} = W_{ST} + W_{OT} + W_M + R \quad (3)$$

where W_{ST} is the weight of the steel structure, W_{OT} is the weight of the outfitting, W_M is the weight of machinery, and R is a reserve.

The weight of the steel structure includes the weight of all components of a ship's steel structure and corresponds approximately to a shipyard's steelwork. Thus, in addition to all of a ship's plates and stiffeners, the mounting base of the engine, the superstructure and deckhouses (even if they are made of different materials, such as aluminium, for example), the masts, the rudder, the rudder shaft, the hatch coamings, and the bulwark are also included in this component group as well [1].

The outfitting weight includes the weight of all fittings of a "naked" ship and all the ship's separable outfittings, except for the machinery outfitting. Specific components of the W_{ST} can be taken as components of W_{OT}, such as the masts and the rudder, noting that it depends on the shipyard or ship designer [1].

W_M can be calculated as:

$$W_M = W_{MM} + W_{MS} + W_{MR} \quad (4)$$

where W_{MM} is the weight of the main machinery, W_{MS} is propeller shaft and propeller weight, and W_{MR} is the remaining machinery weight.

Main machinery weight includes the main engine weight and gearbox weight (if any), the turbine weight for turbine-driven ships, and the gearbox and boilers.

The remaining machinery weight includes the weight of pumps of any kind, any piping inside the engine room, funnels, main electric generators, transformers and switchboards, any supporting mechanical components of the main engine, and alike.

Percentages of tanker weight groups relative to lightship weight are presented in Table 2 [24].

Table 2. Percentages of tanker weight groups relative to lightship weight.

Tanker Size (*DWT*)	DWT/Δ (%)	W_{ST}/W_{LS} (%)	W_{OT}/W_{LS} (%)	W_M/W_{LS} (%)
25,000–120,000	65–83	73–83	5–12	11–16
≥200,000	83–88	75–83	9–13	9–16

The reserve (or margin of uncertainty) R is laid down in the preliminary design to cover possible inexact initial approximations of the various weight groups. For example, the typical R-value in the preliminary design stage in lightship weight is 1–2% for tankers [1].

The lightship structure and equipment load compose static and dynamic components. The static load results from gravity, and the dynamic load can be divided into quasi-static and inertial elements. The quasi-static load results from gravity, considering a ship's instantaneous roll and pitch inclinations. The inertial load results from instantaneous local accelerations on the lightship structure and equipment caused by a ship's motions in six degrees of freedom (DoF) [25].

Furthermore, a regression analysis can be used to estimate the lightship weight, and some examples of regression formulas (Equations (5) and (6)) for the oil tankers presented in [1] are shown below:

$$LS = 2.9186 * DWT^{0.75548} \quad (5)$$

$$LS/\Delta = 1.62433 * DWT^{-0.207684} \quad (6)$$

where LS is lightship weight, DWT is deadweight, and Δ is tanker displacement.

Lightship weight calculation is presented in this section, whereas an empty ship's weight is divided into components. An accurate estimate of each component's weight during the preliminary and final stages of ship design is of the utmost importance for a ship's safety and for economic reasons (overall cargo capacity). Apart from the "traditional" approach to lightship weight estimation, other methods might be beneficial for ship designers, shipyards, and shipowners (and buyers of second-hand and scrap ships). The following section presents a methodology using multiple linear regression and machine learning to estimate the lightship weight of tanker ships.

3. Methodology

The research process is illustrated in Figure 2.

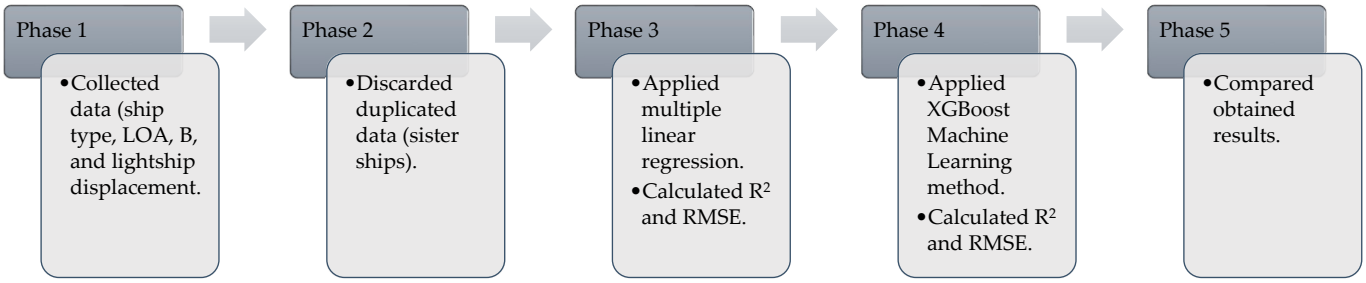


Figure 2. Research phases in this study.

In the first research phase, the authors compiled ship data. Data were collected from the International Association of Independent Tanker Owners (INTERTANKO) Chartering Questionnaire Q88, available online. Questionnaire 88 includes updated information for the assessment of a ship's suitability and risk when chartering tankers. It is the periodically revised (5th version currently) accepted tanker industry standard for information on all types of tankers and tanker terminals worldwide for vetting purposes. Only oil tanker ships were included in this study, irrespective of their size. Three tanker dimensions were used for this research: length overall (LOA), breadth (B), and lightship displacement (LD). Additionally, each tanker's group, according to its size, was added as a dummy variable to improve prediction results (the block coefficient was indirectly included through the tanker size group). Duplicated data (sister ships) were excluded from the study and discarded in the second phase of the research. In total, 80 oil tankers were included in the study (Table A1). Figure 3 presents the sample size according to tanker group.

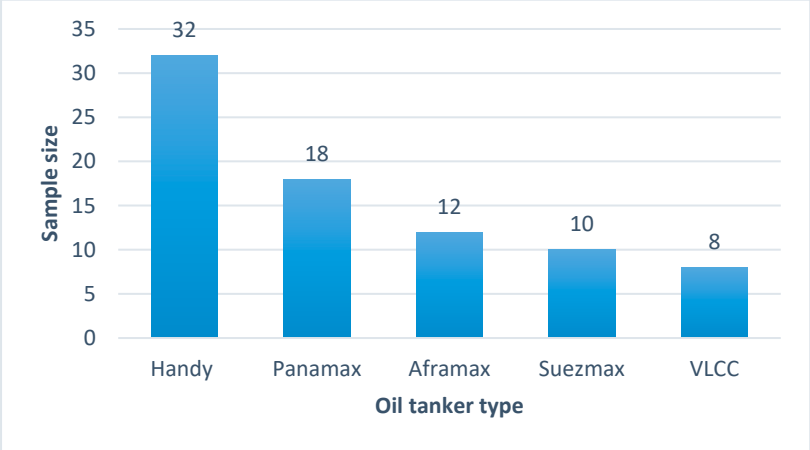


Figure 3. Sample size according to oil tanker type.

A multiple linear regression (MLR) analysis was performed using dimensions obtained from the data collected in the third research phase (80 tankers). This analysis used the ordinary least squares (OLS) method performed in Python using the "statsmodels" package. It describes the relationship between dependent and explanatory (independent) variables [26]. In this paper, the dependent variable was oil tanker lightship displacement (LD), the independent variables were length overall (LOA) and ship breadth (B), and the dummy variable was the tanker group. It is a very efficient method that can be used for

prediction, forecasting, or error reduction [27]. The general multiple regression equation is:

$$Y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + e_i \text{ for } i = 1,\ldots n \quad (7)$$

where Y_i is the response or dependent variable, α is an intercept parameter, $x_{i1}\ldots x_{ip}$ are explanatory variables, $\beta_1\ldots\beta_p$ are regression coefficients, and e_i is the error variable [28]. Multiple linear regression was performed using the software Python's "statsmodels" library to develop a model for estimating lightship displacement based on length overall (LOA), breadth (B), and groups of oil tanker ships.

In phase four, the authors applied the extreme gradient boosting machine learning method (XGBoost ML) using the Python module XGBoost Regressor to check how successfully ML can predict lightship displacement and whether there any differences between the two models. Machine learning is considered a segment of artificial intelligence (AI). ML methodologies can be used to create models based on "training data" (sample data) to make predictions or decisions [29]. It is a branch of computational algorithms developed to mimic human intelligence by learning from the environment. ML-based techniques have been successfully applied in various fields, such as computer vision, spacecraft engineering, finance, entertainment, and medical and are considered a workhorse in the big-data era [30]. The extreme gradient boosting (XGBoost) method, which is "an optimised distributed gradient boosting library designed to be highly efficient, flexible and portable" [31], was used to predict the same data. It applies machine learning algorithms under a gradient boosting methodology. XGBoost provides a parallel tree boosting that is able to rapidly and precisely resolve various data problems [31]. "It can be used to solve regression, classification, ranking, and user-defined prediction problems" [32]. Finally, acquired predictions were compared based on R^2 and root mean square error (*RMSE*), and conclusions were drawn. R^2 is the coefficient of determination and it gives information about the model's goodness of fit [33]:

$$R^2 = 1 - \frac{\sum\left(LD_i - \widehat{LD}_i\right)^2}{\sum(LD_i - \overline{LD})^2} \quad (8)$$

where $\sum\left(LD_i - \widehat{LD}_i\right)$ is the sum of residuals and $\sum(LD_i - \overline{LD})$ is the sum of the distances at which the data occur away from the mean. $\widehat{LD}_i$ is the predicted *LD* value while $\overline{LD}$ is the mean of *LD* values. *RMSE* measures distance between predicted and actual values and indicates the absolute fit of the model to the data [34]:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\widehat{LD}_i - LD_i)^2} \quad (9)$$

where *N* is the number of values.

4. Results and Discussion

Multiple linear regression (MLR) was performed using the Python "statsmodels" package. This study aimed to predict oil tanker lightship displacement using two dimensions—LOA and B. However, since all oil tanker groups were included in the regression, the tanker group was introduced as an additional (dummy) variable to improve LD prediction results. Therefore, the Python *get_dummies()* function was used to convert categorical variables (oil tanker groups) into indicator (dummy) variables. This coding enables the usage of categorical variables in different machine learning prediction models. Dummy coding in Python uses only two values (one and zero) to deliver all needed information on group membership [35]. It is important to emphasise that, in order to reduce the correlations between the dummy variables, the function *drop_first = True* was used. This function assists in reducing the extra column created during dummy variable creation. In this particular case, the Aframax group of oil tankers was removed since it was,

according to its dimensions, just between other tanker groups. In simple terms, if the tanker group is not coded as Handy, Panamax, Suezmax, or VLCC, then it is an Aframax tanker. Multiple linear regression results are presented in Table 3.

Table 3. Results of multiple linear regression with six independent variables.

	Coef. (β)	Std. Err.	t Value	$p > \|t\|$	
const	−4364.13	2442.18	−1.79	0.078	R^2: 0.991
LOA	47.85	12.07	3.96	0.000	Adjusted R^2: 0.990
B	253.55	76.84	3.30	0.001	F-statistic: 1318.0
Handy	−2585.83	1411.42	−1.83	0.071	
Panamax	−2224.17	806.05	−2.76	0.007	
Suezmax	3331.72	606.19	5.49	0.000	
VLCC	17310	1071.89	16.15	0.000	

As shown in Table 3, the adjusted R^2 is 0.99, and the p-value for the coefficients LOA and B is <0.05. Thus, a multiple regression model was adopted, including one dependent variable (LD) and two independent variables (LOA and B) in the case of the Aframax oil tanker group (6). It can be said that the model successfully explains 99.0% of the variance in the data and has statistical significance. From the results of multiple linear regression (Table 3), Equation (10) can be used for estimating the lightship displacement of Aframax oil tanker ships:

$$LD\ (Aframax) = 47.85\ LOA + \ 253.55\ B - 4364.13 \quad (10)$$

In order to estimate other tanker groups' lightship displacement, Equations (11)–(14) apply:

$$LD\ (Handy) = 47.85\ LOA + \ 253.55\ B - 6949.96 \quad (11)$$

$$LD\ (Panamax) = 47.85\ LOA + \ 253.55\ B - 6588.3 \quad (12)$$

$$LD\ (Suezmax) = 47.85\ LOA + \ 253.55\ B - 1032.41 \quad (13)$$

$$LD\ (VLCC) = 47.85\ LOA + \ 253.55\ B + 12945.87 \quad (14)$$

Additionally, it must be noted that the constant's standard deviation is 2442.18 (Table 3). The average differences in tons and percentages (and absolute values) for each tanker group, identified via MLR, are presented in Table 4.

Table 4. Average differences between MLR-predicted and actual LD values.

Tanker Group	Handy-Sized	Panamax	Aframax	Suezmax	VLCC
RMSE	621.00	989.97	1070.11	1925.65	2035.16
Average diff. (%)	3.5	−1.5	−0.4	−0.6	−0.2

As presented in Table 4, the most considerable average difference in percentage is for the Handy-sized group of tankers. However, as VLCC tankers are the group with the largest size, for the tanker with an LD of 47,818 t, the difference of 0.2% is 95.6 t, and for the Handy-sized tanker with an LD of 3008.13 t, the difference of 3.5% is 105.3 t. From these examples, it can be concluded that the difference is relatively small when converted in tons as a result of large differences in lightship displacement.

A Studentized Breusch–Pagan test (bptest) was performed in the statistical software Python to test the model for heteroskedasticity. If a p-value is less significant than 0.05, it indicates that the null hypothesis can be rejected (the variance does not change in the residual—homoskedasticity), and, therefore, heteroskedasticity exists. Another important

assumption in multiple linear regression is autocorrelation, meaning there is no correlation between residuals (they are independent of each other). The Durbin–Watson (DW) statistical test was performed to check for autocorrelation. If there is no autocorrelation, the DW statistic value is between 1.5 and 2.5 (rule of thumb), and the *p*-value is above 0.05. The results of both tests are shown in Table 5.

Table 5. Results of heteroskedasticity and autocorrelation of the model.

Studentized Breusch–Pagan Test	
BP = 3.619	*p*-value = 0.005
Durbin–Watson test	
DW = 1.347	

From the results of the tests, as presented in Table 5, it can be concluded that the obtained model shows features of heteroskedasticity and autocorrelation. However, as stated in [36], "heteroskedasticity has never been a reason to throw out an otherwise good model." Therefore, the authors believe the model could be a valuable predictive tool and usable in practice. Furthermore, their research aimed to use another tool that could be used for the prediction of LD using two known dimensions and comparing obtained values.

One such tool might be the extreme gradient boosting machine learning method focusing on predictive accuracy. A graphical representation of the relation between the variables is given in Figure 4.

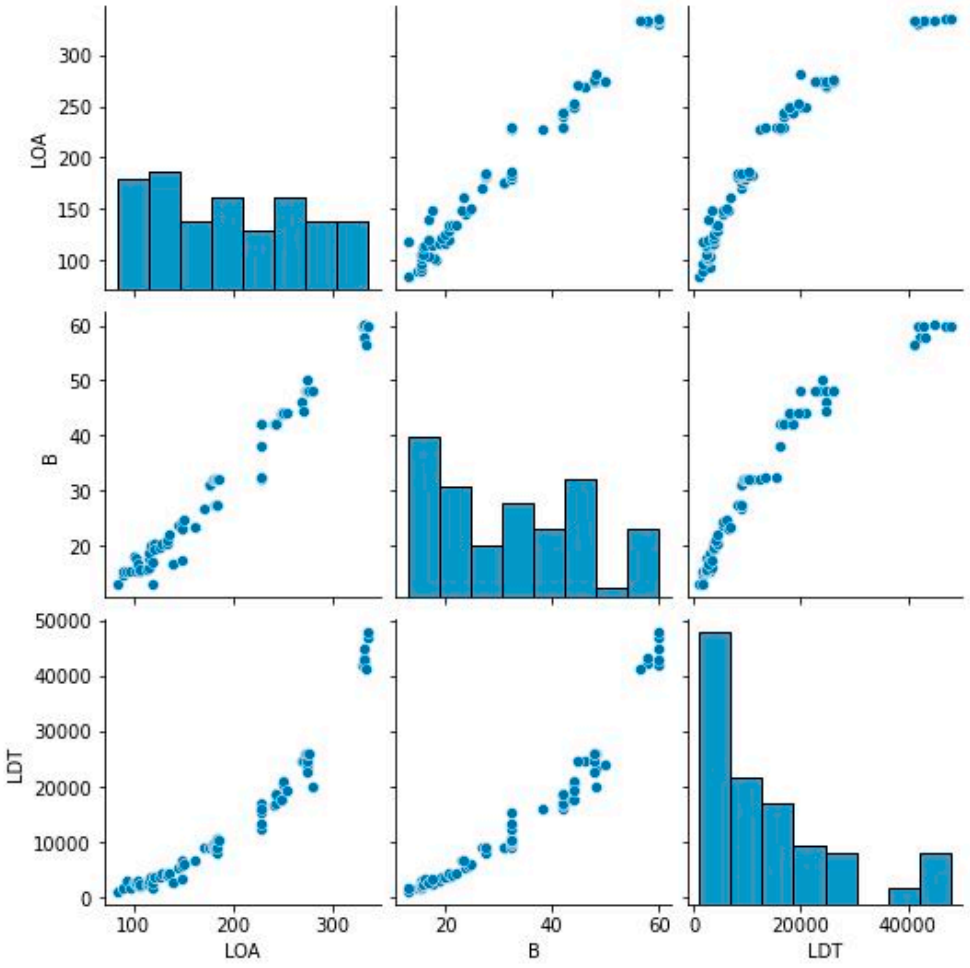


Figure 4. Relations between variables.

The correlation matrix of the variables is presented in Figure 5.

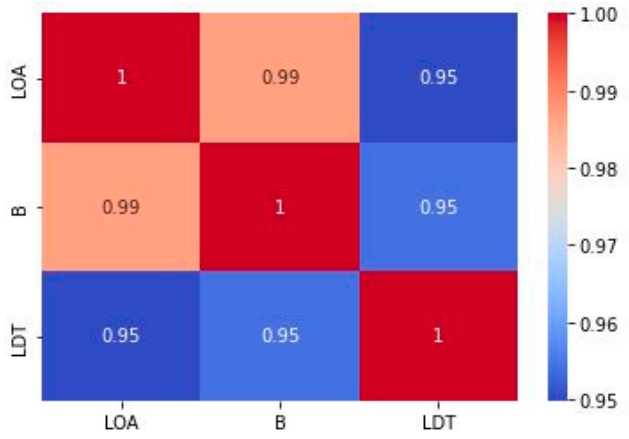


Figure 5. Variable correlation matrix.

The correlation matrix (Figure 4) shows that the correlation between LOA and B is 0.99, between LOA and LD is 0.95, and between B and LD is 0.95. Thus, it can be concluded that both LOA and B will have a high predictive power on the LD. Additionally, as presented in Figure 4, the ratio of LD and variables LOA and B seems more polynomial than linear. This feature might create minor problems for linear models such as regression. However, XGBoost copes well with nonlinearity, even without adjustments, preventing overfitting [37]. Data were applied to the XGBoost ML method, and descriptive statistical results are presented in Table 6.

Table 6. Descriptive statistics of the variables.

	Count	Mean	Std	Min	25%	50%	75%	Max
LOA	80	195.03	75.33	83.4	124.49	182.55	249.87	336.17
B	80	32.54	14.34	13.0	19.65	31.6	44.0	60.04
LD	80	13833.28	12557.47	1133.0	3722.72	9038.0	19564.3	47818.0

XGBoost machine learning resulted in $R^2 = 0.990$ and $RMSE = 1211.99$. Like in MLR, LOA and B have high predictive power; however, B has greater significance according to XGBoost (Figure 6).

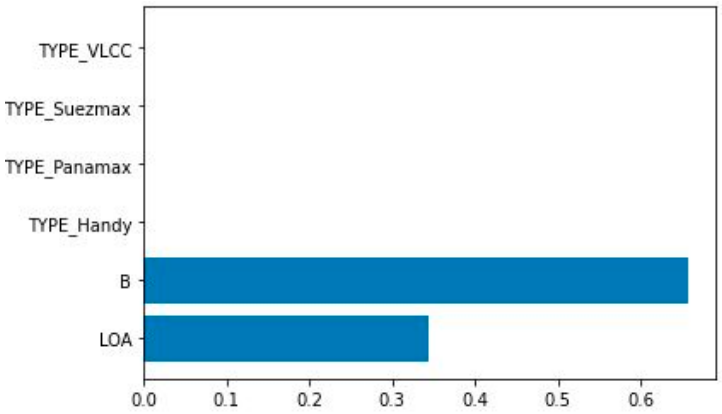


Figure 6. Importance of independent variables (LOA and B).

The data set was divided into a training set (85%) and a test set (15%) to evaluate the XGBoost performance. The algorithm was trained on the first set, and the second set was used to compare predicted values with actual ones. The disadvantage of this evaluation

tool is that it can have a high variance, meaning that training and test set data differences might result in significant differences when estimating the model's accuracy. Nevertheless, as shown in Figure 7, the predicted LD fits well with the original LD. Even though the differences between predicted and original values are minor (Figure 7), the model needs further evaluation.

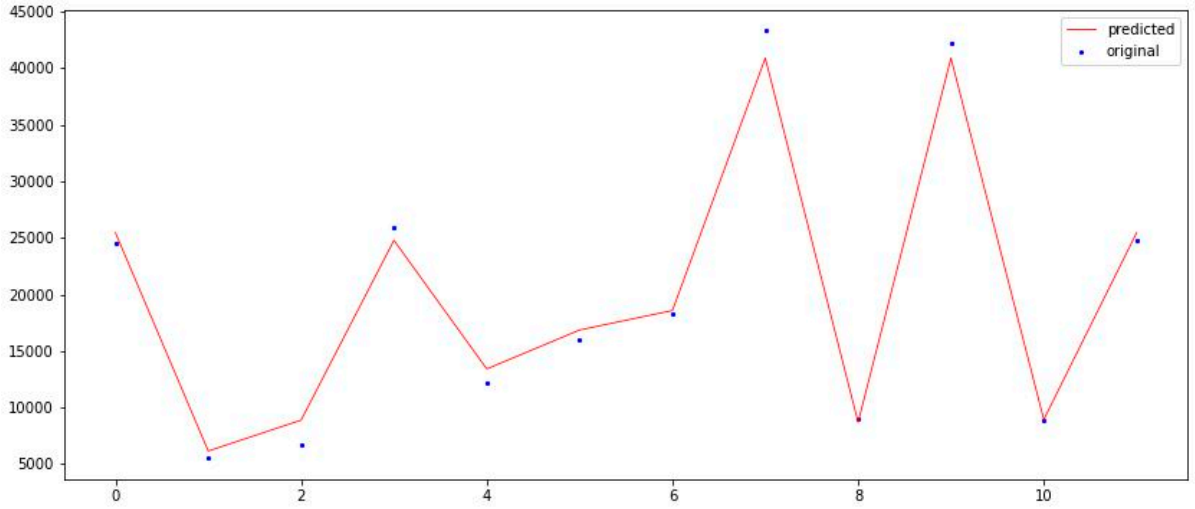


Figure 7. Differences between predicted and original LD values.

Another tool for estimating the accuracy of the model is cross validation (CV). It is a tool that can be used to estimate the performance of an ML algorithm with less variance than can be achieved by dividing the dataset into training and test sets. It works by dividing the dataset into k parts (for example, k = 3, k = 5, or k = 10). Each data split is called a fold. The ML algorithm is trained on k-1 folds with one held back and tested on the held back fold. This procedure is repeated so that each fold of the dataset is given a chance to act as the held-back test set. Cross validation results in k different performance scores that can be summarised using a mean and a standard deviation. In this research, the dataset was divided into 10 parts (k = 10). The mean k-fold cross validation R^2 score was 0.97, meaning that the obtained model can explain 97% of the variance. It is important to mention that the standard deviation obtained for the CV R^2 score was 0.04. The mean CV *RMSE* score was 1018.77, with a standard deviation of 257.35. Results of the CV test on a random sample are presented in Table 7.

Table 7. Predicted and original LD differences (random sample).

No.	Predicted LD	Original LD	Difference (t)	Difference (%)
1	2753.8	2752.0	1.8	0.07
2	6139.7	6135.0	4.7	0.08
3	3414.8	3415.0	−0.2	0.00
4	41261.1	41261.0	0.1	0.00
5	40904.2	43308.0	2403.8	5.55
6	17648.5	17642.6	5.9	0.03
7	10813.9	10821.4	−7.5	0.07
8	19996.8	19993.0	3.8	0.02
9	4158.9	4154.8	4.1	0.10
10	3126.3	3125.8	0.5	0.02

As shown in Table 7, nine rows have minimal differences between predicted and original LD (1–4 and 6–10), but the fifth row has a significant difference, notwithstanding

the generally good results that were obtained. This probably occurred as a result of occasions wherein the data set contained two or more ships with very similar LOA and B values, but there was a difference between LD values. The differences could be due to various reasons, such as ice-class ships, ships with a higher superstructure, various additional equipment installed (e.g., a water ballast treatment system), or other additional structures, machinery, or equipment added for specific trading areas. In addition, a reason for the LD difference between tankers of the same dimensions could be that one was built with a certain percentage more steel than the minimum class requirements in order to strengthen the ship structure, extending the ship's life. Average cross-validation differences in tons and percentages for each tanker group are presented in Table 8.

Table 8. Average CV differences for tanker groups.

Tanker Group	Handy-Sized	Panamax	Aframax	Suezmax	VLCC
RMSE	405.00	348.94	82.56	521.34	961.00
Average diff. (%)	−1.38	−0.72	−0.13	−0.23	1.07

As presented in Table 8, the Handy-sized group exhibited the most considerable differences, followed by the VLCC group. The most negligible differences between original and predicted LD values were obtained for the Aframax and Suezmax groups.

Results obtained using the XGBoost ML method showed that the model explains 99.0% of the variance, but, when evaluation using k-fold cross validation was performed, the obtained R^2 mean score was 0.97. Thus, it is shown that ML, especially the XGBoost method, can be applied in the shipping industry to this particular example for predicting the lightship displacement weight of oil tanker ships, using only two known variables, LOA and B, and by introducing tanker group (by size).

The results of the regression performance measurement show that, in this case, multiple linear regression together with the XGBoost regressor have high predictive power. For example, in this research, the linear regression value of R^2 is 0.990 compared to the XGBoost R^2 value of 0.990 (CV R^2 = 0.97), and the *RMSE* is 1193.41 compared to 1211.99 (CV *RMSE* = 1018.77), respectively, as presented in Table 9.

Table 9. Comparison of R^2 and RMSE obtained.

	MLR	XGBoost ML (CV)
RMSE	1193.41	1018.77
R^2	0.990	0.970

In addition, a comparison of the results for different tanker size groups are presented in Table 10. As can be concluded from Table 10, the worst fitting was obtained for Handy –sized and VLCC tankers. Furthermore, it needs to be stressed that results utilising XGBoost were better when compared to MLR.

Table 10. Comparison of results obtained utilising MLR and XGBoost by tanker size groups.

Tanker Size Group	Handy	Panamax	Aframax	Suezmax	VLCC
MLR *RMSE*	621.00	989.97	1070.11	1925.65	2035.16
MLR av. diff. (%)	3.5	−1.5	−0.4	−0.6	−0.2
XGBoost *RMSE*	405.00	348.94	82.56	521.34	961.00
XGBoost av. diff. (%)	−1.38	−0.72	−0.13	−0.23	1.07

The XGBoost ML method has shown its usability for predicting the lightship displacement of oil tankers and, together with the MLR, confirmed another possible application in the maritime industry. Lightship displacement tonnage is valuable information for numerous stakeholders, and, because of its seldom being readily available, a method for its estimation has been developed. Furthermore, there might be considerable differences

in LD as a result of different equipment or machinery weights or even more superstructure decks on ships with the same LOA and B. However, the model obtained in this research estimates LD rather accurately, which was confirmed by R^2 and *RMSE*. Moreover, cross validation confirmed its high predictive power, and the model can be used for LD estimation in practice.

It is also worth mentioning that about 85% of a ship's weight (LD) is reusable steel, making oil tankers and bulk carriers highly desirable ships for recycling as a result of their high steel content [22]. Furthermore, from an economic point of view, according to data from 2021, the average price offered for scrap steel was around USD 450 per lightship displacement ton (Indian subcontinent) [38]. Therefore, if the LD estimation error were transmuted to US dollars, an estimation error of 100 tons is worth USD 45,000, which is a relatively expensive error. However, in terms of LD percentage, an estimation error of 100 tons for an average Aframax crude oil tanker is 0.6% of the total LD.

5. Conclusions

This paper shows that the lightship displacement of tanker ships can be estimated using two ship dimensions, length overall (LOA) and breadth (B). Furthermore, multiple linear regression is a valuable tool that can accurately predict dependent variables based on independent variables. In this research, the authors introduced additional dummy variables (tanker group according to size) to predict lightship displacement using multiple linear regression. The XGBoost machine learning method was introduced to compare the lightship displacement tonnage prediction. The novelty of this research is introducing the XGBoost machine learning method to estimate lightship displacement data, which showed its advantages and confirmed its usability in the shipping industry.

However, it should be mentioned that the study also takes into consideration some limitations. Firstly, it can be argued that our collected data sample is relatively small. This is connected to the fact that some data regarding the same-sized group of tankers were duplicated (sister ships). After discarding duplicate data, 80 oil tankers were used to build a model for predicting existing tankers' lightship displacement. Nevertheless, in addition to two independent variables (LOA and B), dummy variables (tanker groups) were introduced in this study, enabling more precise lightship displacement predictions. Therefore, it is recommended that more data should be collected to predict required oil tanker dimensions more accurately in future research. Secondly, it must be emphasised that machine learning, in general, is poor at generalising to examples outside the scope of the training set. Tree-based algorithms (XGBoost) are inferior for large extrapolations. This model predicts well LD values between a minimum LOA of 83.4 m and a maximum LOA of 336.17 m and between a minimum B of 13.0 m and a maximum B of 60 m. Having said that, the developed model is applicable to the shipping industry, covering the majority of tanker size groups, but could be inaccurate for the design of new tanker ships and can be used to estimate the lightship displacement of existing tankers within the abovementioned LOA and B ranges.

The results of this research might be used by maritime shipping stakeholders seeking to estimate the lightship displacement of a tanker ship, and the results will be satisfactory. The authors' opinion is that further research on the use of tanker ships' length overall and breadth for estimating lightship displacement that employs the multiple linear regression and XGBoost ML methods, has large sample sizes, and classifies tankers according to their size will generate more accurate predictions and facilitate the estimation of tankers' lightship displacement. Therefore, in future research, only one oil tanker category size will be sampled (for example, Aframax-sized oil tankers). Moreover, introducing additional variables, such as the draft of a tanker and its year and country of building, might provide additional insights into estimating lightship displacement. It should be emphasised that the authors focused on oil tankers in this paper, but the same procedure can be utilised to predict lightship displacement for other ship types.

Author Contributions: Conceptualization, N.H. and V.F.; methodology, N.H and V.F.; software, M.M.; validation, N.H., V.F., and M.S.; formal analysis, N.H. and V.F.; investigation, N.H. and V.F.; resources, N.H. and V.F.; data curation, N.H.; writing—original draft preparation, N.H and V.F.; writing—review and editing, N.H., V.F., and M.S.; visualization, N.H. and M.M; supervision, V.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are available from corresponding authors on request.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. Dimensions and size classes of tankers used in the study.

No.	Tanker Group	LOA (m)	B (m)	LD (t)	No.	Tanker Group	LOA (m)	B (m)	LD (t)
1	Handy	83.40	13.00	1133.00	41	Panamax	182.55	27.40	8699.00
2	Handy	88.80	14.80	1748.70	42	Panamax	183.00	32.20	10,821.40
3	Handy	90.00	15.20	1840.57	43	Panamax	183.17	32.24	10,200.00
4	Handy	92.35	15.20	3008.13	44	Panamax	184.32	27.40	8123.00
5	Handy	95.80	15.20	1810.90	45	Panamax	184.33	27.44	8966.00
6	Handy	99.60	18.00	2817.00	46	Panamax	185.93	32.23	10,249.00
7	Handy	102.21	15.50	2373.00	47	Panamax	228.17	32.20	12,198.00
8	Handy	102.70	17.80	2592.87	48	Panamax	228.40	38.03	16,045.60
9	Handy	103.60	16.60	2940.00	49	Panamax	228.60	32.26	13,421.35
10	Handy	105.29	15.20	2217.00	50	Panamax	228.60	32.35	15,238.90
11	Handy	106.20	15.60	2462.00	51	Aframax	228.60	42.00	16,871.20
12	Handy	113.08	15.70	2500.00	52	Aframax	228.60	42.04	16,075.00
13	Handy	114.87	16.00	3230.00	53	Aframax	240.99	42.00	16,639.00
14	Handy	115.00	17.60	2464.80	54	Aframax	243.00	42.00	18,295.00
15	Handy	115.50	18.70	3125.83	55	Aframax	243.80	42.04	18,637.00
16	Handy	116.50	20.00	3704.53	56	Aframax	243.96	42.00	16,915.00
17	Handy	118.87	13.00	1607.00	57	Aframax	244.15	42.00	18,498.00
18	Handy	119.10	16.90	2826.70	58	Aframax	248.96	43.80	17,642.60
19	Handy	120.00	20.43	3748.20	59	Aframax	249.00	44.00	17,771.90
20	Handy	121.40	19.20	3728.79	60	Aframax	249.85	44.06	20,200.00
21	Handy	125.53	19.80	3802.50	61	Aframax	249.96	44.00	20,924.00
22	Handy	128.60	20.40	4403.00	62	Aframax	253.59	44.03	19,421.40
23	Handy	134.16	20.52	4272.00	63	Suezmax	269.17	46.04	24,742.90
24	Handy	134.30	21.20	4154.80	64	Suezmax	270.45	44.60	24,531.00
25	Handy	134.85	22.00	4248.60	65	Suezmax	273.70	48.04	25,939.60
26	Handy	139.90	16.70	2752.00	66	Suezmax	274.00	48.04	23,414.00
27	Handy	145.53	23.73	5375.01	67	Suezmax	274.18	50.00	23,978.00
28	Handy	147.83	24.23	5562.00	68	Suezmax	274.20	48.00	22,616.00
29	Handy	148.43	23.00	6580.00	69	Suezmax	274.22	48.00	24,788.00
30	Handy	149.35	17.30	3415.00	70	Suezmax	274.50	48.00	25,941.40
31	Handy	149.93	24.60	6135.00	71	Suezmax	277.08	48.00	25,900.00
32	Handy	161.12	23.25	6671.00	72	Suezmax	281.20	48.20	19,993.00
33	Panamax	170.17	26.63	8881.20	73	VLCC	329.88	60.00	41,789.30
34	Panamax	175.90	31.00	8937.00	74	VLCC	332.00	58.00	42,173.00

35	Panamax	175.93	31.00	8871.00	75	VLCC	332.95	58.00	43,308.00
36	Panamax	179.88	32.23	9110.00	76	VLCC	333.00	60.00	42,749.80
37	Panamax	179.96	32.20	9710.00	77	VLCC	333.00	60.05	44,900.00
38	Panamax	181.78	27.42	8311.00	78	VLCC	333.40	56.50	41,261.00
39	Panamax	182.50	32.20	9981.00	79	VLCC	336.00	60.00	46,974.00
40	Panamax	182.55	27.34	8942.00	80	VLCC	336.17	60.00	47,818.00

References

1. Papanikolau, A. Ship *Design; Methodologies of Preliminary Design*, 1st ed.; Springer: Berlin/Heidelberg, Germany, 2014. ISBN 978-94-017-8750-5.
2. International Maritime Organization. *SOLAS Consolidated Edition*; IMO: London, UK, 2020.
3. International Maritime Organization. *Resolution MSC.267(85). Adoption of the International Code on Intact Stability (2008 IS Code);* IMO: London, UK, 2008.
4. Creese, R.C.; Nandeshwar, A.; Sibal, P. Ship Deconstruction Cost Models. In Proceedings of the 2002 AACE International Transactions, 46th Annual Meeting of AACE International, Portland, OR, USA, 23–26 June 2002; AACE International: Morgantown, WV, USA, 2002.
5. Pour, B.; Noshadi, E; Fard, M. Analysis of Ships Supply and Demand Principles in the World Sea Trade. *Int. J. Account. Financ. Manag. IJAFM* **2012**, *4*, 161–169.
6. Karlis, T.; Polemis, D. Ship demolition activity: A monetary flow process approach. *Sci. J. Marit. Res.* **2016**, *30*, 128–132.
7. Jain, K.P.; Pruyn, J.F.J.; Hopman, J.J. Quantitative assessment of material composition of end-of-life ships using onboard documentation. *Resour. Conserv. Recycl.* **2016**, *107*, 1–9. https://doi.org/10.1016/j.resconrec.2015.11.017.
8. Molland, A.F.; Turnock, S.R.; Hudson, D.A. *Ship Resistance and Propulsion. Practical Estimation of Ship Propulsive Power*; Cambridge University Press: New York, NY, USA, 2011. ISBN 978-0-521-76052-2.
9. Čudina, P. Design Procedure and Mathematical Models in the Concept Design of Tankers and Bulk Carriers. *Brodogradnja* **2008**, *59*, 323–339.
10. Duru, S.C. Lightship Component Masses in Preliminary Design Examplified for Fishing Vessel. *Int. J. Sci. Eng. Res.* **2016**, *7,* 49–53.
11. Lin, C.-K.; Shaw, H.-J. Feature based estimation of steel weight in shipbuilding. *Ocean. Eng.* **2015**, *107*, 193–203.
12. Prabu, C.S.K.; Nagarajan, V.; Sha, O.P. Study on the Lightship Characteristics of Merchant Ships. *Brodogradnja* **2020**, *71*, 37–70.
13. Obreja, D.; Chiroşcă, A.-M. Preliminary estimation of ship weight. *Ann. "Dunarea de Jos" Univ. Galati Fascicle XI –Shipbuild.* **2015**.
14. Kristensen, H.O. *Determination of Regression Formulas for Main Dimensions of Tankers and Bulk Carriers Based on HIS Fairplay Data. Project no. 2010-56, Emissionsbeslutningsstøttesystem, Work Package 2*; Report no. 2; Technical University of Denmark: Kongens, Denmark, 2012.
15. Gurgen, S.; Altin, I.; Ozkok, M. Prediction of main particulars of a chemical tanker at preliminary ship design using artificial neural network. *Ships Offshore Struct.* **2018**, *13*, 459–465.
16. Kaiser, M.J.; Snyder, B.F. Empirical models of jackup rig lightship displacement. *Ships Offshore Struct.* **2013**, *8*, 468–476.
17. Slapničar, V.; Zadro, K.; Ložar, V.; Ćatipović, I. The lightship mass calculation model of a merchant ship by empirical methods. *Pedagog. Pedagog.* **2021**, *93*, 73–87. https://doi.org/10.53656/ped21-6s.06lig.
18. Cepowski, T. Determination of regression formulas for main tanker dimensions at the preliminary design stage. *Ships Offshore Struct.* **2018**, *14*, 320–330. https://doi.org/10.1080/17445302.2018.1498570.
19. Cepowski, T.; Chorab, P.; Łozowicka, D. Application of an artificial neural network and multiple nonlinear regression to estimate container ship length between perpendiculars. *Pol. Marit. Res.* **2021**, *28*, 36–45. https://doi.org/10.2478/pomr-2021-0019.
20. Horn, G.; Cronin, D. The Common Structural Rules-Initial Designs and Future Developments. *Tanker Struct. Coop. Forum Shipbuild. Meet.* **2010**.
21. Herbert Engineering Corp. *Design and Construction of Oil Tankers. Prepared for: Enbridge Northern Gateway Project, 2012;* Report No. 2012-020-1; Herbert Engineering Corp: Annapolis, MD, USA, 2012.
22. Sujauddin, M.; Koide, R.; Komatsu, T.; Hossain, M.; Tokoro, C.; Murakami, S**.** Characterisation of ship breaking industry in Bangladesh. *J. Mater. Cycles Waste Manag.* **2014**, *17*, 72–83.
23. Newcastle University. Typical Ship Principal Dimensions. Revision 1. 2011. Available online: https://eprints.ncl.ac.uk/file_store/production/179602/76BCF714-08E2-4C75-BC08-DD4CBEE63EC3.pdf (accessed on 7 April 2023).
24. Schneekluth, H.; Bertram, V. *Ship Design for Efficiency and Economy*, 2nd ed.; Butterworth-Heinemann: Oxford, UK, 1998. ISBN 0 7506 4133 9.
25. American Bureau of Shipping (ABS). Guide for 'Safehull-Dynamic Loading Approach' For Vessels. 2018. Available online: https://ww2.eagle.org/content/dam/eagle/rules-and-guides/current/design_and_analysis/140_safehulldlaforvessels/DLA-Vessels_Guide_e-May18.pdf (accessed on 7 April 2023).
26. Ćorović, B.; Djurović, P. Research of Marine Accidents through the Prism of Human Factors. *Promet. Traffic Transp.* **2013**, *25,* 369–377.

27. Uyanik, G.K.; Güler, N. A study on multiple linear regression analysis. *Procedia –Soc. Behav. Sci.* **2013**, *106*, 234–240.
28. Kloke, J.; McKean, J.W. *Nonparametric Statistical Methods Using R*; CRC Press: Boca Raton, FL, USA; Taylor and Francis Group: Abingdon-on-Thames, UK, 2015. ISBN 13:978-1-4398-7344-1.
29. Koehrsen, W. Towards Data Science. Learning Algorithm to Deliver Business Value. How to Train, Tune, and Validate a Machine Learning Model 2018. Available online: https://towardsdatascience.com/modeling-teaching-a-machine-learning-algorithm-to-deliver-business-value-ad0205ca4c86 (accessed on 7 June 2021).
30. El Naqa, I. What Is Machine Learning? In *Machine Learning in Radiation Oncology*; El Naqa, I., Li, R., Murphy, M.J., Eds.; Springer Nature: Berlin/Heidelberg, Germany, 2015; pp. 3–11. https://doi.org/10.1007/978-3-319-18305-3.
31. Scalable and Flexible Gradient Boosting. About XGBoost. Available online: https://xgboost.ai/about (accessed on 5 June 2021).
32. Morde, V.; Setty, V.A. Towards Data Science. XGBoost Algorithm: Long May She Reign! 2019. Available online: https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d (accessed on 7 June 2021).
33. Newcastle University. Coefficient of Determination, R-Squared. Available online: https://www.ncl.ac.uk/webtemplate/ask-assets/external/maths-resources/statistics/regression-and-correlation/coefficient-of-determination-r-squared.html (accessed on 7 April 2023).
34. Kaggle. What Is Root Mean Square Error (RMSE)? Available online: https://www.kaggle.com/general/215997 (accessed on 7 April 2023.
35. Statology. How to Use Pandas Get Dummies–pd.get_dummies. 2021. Available online: https://www.statology.org/pandas-get-dummies/ (accessed on 15 June 2021).
36. Mankiw, N.G. A Quick Refresher Course in Macroeconomics. *J. Econ. Lit.* **1990**, *XXVIII*, 1645–1660.
37. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the KDD '16 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; Association for Computing Machinery, New York, NY, USA, 2016; pp. 785–794. https://doi.org/10.1145/2939672.2939785.
38. Bimco.org. Crude Oil Tanker Demolition Stalls as Second-Hand Prices Win. 2021. Available online: https://www.bimco.org/news/market_analysis/2021/20210311_2021_crude_oil_tanker_demolition_stalls_as_second-hand_prices_win (accessed on 25 September 2021).