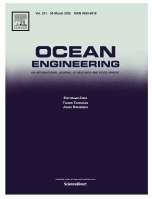
Ocean Engineering 333 (2025) 121407

Contents lists available at ScienceDirect

Ocean Engineering

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



Research paper

Development of regression models for estimating main particulars of RoPax vessels in the conceptual design stage

Francesco Mauro [a,b,*], Ahmed Salem [b]

[a] University of Trieste, 34100, Trieste, Italy
[b] Sharjah Maritime Academy, 180018, Sharjah, Khorfakkan, UAE

ARTICLE INFO

Keywords:
Ship design
Regression models
RoPax vessels
Multiple linear regression
Forest tree algorithms.

ABSTRACT

The design of new ships is a process that requires knowledge of several aspects of naval architecture and marine engineering. During the early design stage, one of the first issues that designers should face is the preliminary estimation of the vessel's main dimensions, respecting the desiderata of the ship owner. Therefore, it is relevant to provide designers with suitable tools that may help estimate the principal dimensions, consider conventional methods and investigate the applicability of modern techniques based on machine learning. The present work focuses on applying different regression techniques to a database of RoPax vessels, finding mathematical instruments to evaluate the ship's main dimensions. Conventional regression techniques are first investigated here to compare with the existing formulae provided by other databases. The study is then extended by applying multiple linear regression and forest tree algorithms, seeking an improvement of conventional formulations available in the literature. The results highlight how the most modern regression techniques allow for better coverage of the design space, allowing the use of more than one input to obtain the final dimensions.

1. Introduction

Ship design is a complex process that requires multiple phases with a constantly increasing level of detail, ranging from the conceptual design of the vessel up to the detailed design of the subsystems (Andrews, 1998; Caprace and Rigo, 2011). The preliminary stages of the design process include the most crucial and relevant decisions, influencing the direction of the whole design (Watson, 1998; Papanikolaou et al., 2022). In this phase, the designer should choose the vessel's dimensions, satisfying the global requirements of the shipowner (Schneekluth and Bertram, 1998; Ljulj et al., 2020).

The praxis for the design of ships is using simple regression or diagrams that allow for determining the dimensions as a function of a design-driven parameter (Žanić et al., 1992; Kalokairinos et al., 2005; Grubišić and Begović, 2001), varying according to the vessel type (Papanikolaou, 2014; Abramowski et al., 2018). For the specific case of RoPax vessels, the driving parameters are the displacement, the deadweight or the lane metres (Trincas et al., 1994). For this purpose, the literature provides examples of simple regression models for various parameters, with some regressions based on displacement, deadweight, and others as a function of lane meters or length (Piko, 1980; Putra et al., 2022; Kristensen, 2016; Friis et al., 2002; Novak et al., 2020). However, in the preliminary design stage, the shipowner may also require the satisfaction of additional parameters, such as ship speed, installed power, or passenger capacity.

Simple regression analyses in the literature do not account for these additional parameters when estimating a ship's main dimensions. Furthermore, the database used for such regressions, in the case of RoPax vessels, refers to old ships, not accounting for recent designs. Therefore, besides the inclusion of modern ships in the starting database, an alternative approach is needed to enhance the accuracy of the regression models used for this evaluation. Other methods, ranging from multiple linear regressions to advanced machine learning techniques (Cepowski and Chorab, 2021; Majanarić et al., 2022), allow for the inclusion of more parameters in the regression process. These approaches require an initial, comprehensive database of ships containing all relevant dimensions and parameters. To ensure reliability, the database must be homogeneous, with outliers and incomplete datasets removed. The quality of the database also influences the type of regression that can be applied, particularly when using machine learning techniques (Asrol et al., 2021).

In the present study, the database consists of 87 RoPax vessels, derived from an initial population of 127 ships after the removal of incomplete data and outliers (Clarksons, 2024; Ferry-site, 2024). Given the limited sample size of 87 vessels, it is not recommended to apply machine learning techniques such as neural networks, which re-

* Corresponding author.
E-mail addresses: fmauro@units.it (F. Mauro), Ahmed.Salem@sma.ac.ae (A. Salem).

https://doi.org/10.1016/j.oceaneng.2025.121407
Received 20 January 2025; Received in revised form 13 April 2025; Accepted 27 April 2025
Available online 18 May 2025

quire a larger dataset to perform effectively (Clausen et al. 2001; Gurgen et al. 2018). However, other methods, such as random forest, can handle smaller datasets and are thus suitable for this study (Ekinci et al., 2011; Rinauro et al., 2024). For the execution and testing of the different regression models, the final dataset has been divided into a training set, used to fit the regressions, and a test set for evaluating and comparing the different models.

Therefore, this work compares the simple conventional regression formulae applied to the present database with the multiple linear regressions and the random forest technique. To achieve this aim, the paper follows the outline presented below:

- Section 2 provides a summary of the regression models available in the literature for predicting the main dimensions of RoPax vessels.
- Section 3 outlines the initial database and details the process used to exclude outliers and divide the training and test sets.
- Section 4 introduces the regression strategy for simple and multiple linear regressions, as well as the random forest approach, along with the methodology used to assess the quality of the regressions.
- Section 5 presents the results of the regression analysis on the training set.
- Section 6 presents the application of the regressions on the test set, along with a discussion of the obtained results.

The study provides a step forward in the main dimensions estimation of RoPax vessels, employing a modern database as a source for the regression process. The comparison between the formulations available in the literature and the ones derived from the study highlights a significant difference in the regression formulation, due to the different population of the database, that includes modern vessels. Therefore, even the new simple regression formulae, derived from the new database, are a significant improvement for designers in the initial estimate of RoPax vessels' main dimensions, compared to old formulations available in the literature. Furthermore, the study demonstrates how advanced regression strategies outperform simpler methods by achieving higher regression accuracy, enabling designers to incorporate more parameters when estimating a ship's main dimensions. The differences, advantages, and disadvantages of these advanced techniques are thoroughly explained and discussed in the paper.

2. Available regression models for predicting the main dimensions of RoPax vessels

The selection of a ship's dimensions is a key focus in the preliminary design phase of a vessel. Consequently, the literature offers various methods and formulations specifically for RoPax vessels. These formulations differ in terms of the equation structure and the independent variable used for regression. However, all the equations rely on a single independent variable.

This section presents the equations based on the different independent variables used to determine the ship's dimensions, namely:

- Deadweight.
- Lane metres.
- Length.
- Others.

Each of the above mentioned variables depends on the availability of data at the beginning of the process and the strategy adopted for determining the main dimensions.

2.1. Regression models based on deadweight

In ship design, one of the most straightforward ways to estimate the vessels' dimensions is by using deadweight DWT as the independent variable. Analysing a database of 107 Ropax ships, Piko (1980) proposes a set of power regressions having the following form:

$$x = \alpha(DWT)^{\beta} \cdot \epsilon \tag{1}$$

From the regression analysis the following parameters α and β result for the main dimensions:

$$L = 61.7(DWT)^{0.423} \tag{2}$$

$$B = 11.3(DWT)^{0.303} \tag{3}$$

$$T = 3.94(DWT)^{0.297} \tag{4}$$

The same model according to Eq. (1) is also used to define the horsepower (in Hp) installed onboard:

$$HP = 1640(DWT)^{0.905} \tag{5}$$

All these equations underline a strong correlation between the deadweight and all the regressed parameters; therefore, the study of Piko (1980) includes also an alternative formulation for the main dimensions, according to a quadratic model:

$$x = \alpha + \beta(DWT) + \gamma(DWT)^2 + \epsilon \tag{6}$$

From the regression analysis the following parameters α, β and γ result for the main dimensions:

$$L = 62.0 + 13.6(DWT) - 0.314(DWT)^2 \tag{7}$$

$$B = 13.5 + 1.07(DWT) - 0.0151(DWT)^2 \tag{8}$$

$$T = 4.42 + 0.423(DWT) - 0.00844(DWT)^2 \tag{9}$$

For the horsepower, instead of a quadratic model, an alternative linear model is proposed according to the following formulation:

$$HP = 1010 + 1340(DWT) \tag{10}$$

The study reports the quality of the regression in the form of the determination coefficient R^2, highlighting higher values for the power regressions compared to the quadratic and linear models. It is then advisable to use the power model as a reference for the determination of main dimension as a function of the deadweight.

2.2. Regression models based on lane metres

Another possibility for estimating the main dimensions of a RoPax vessel in the preliminary design stage is by using the lane metres LM as the independent variable. Even though the lane metres are an important parameter for the design of RoPax ships, the literature does not report specific regression using LM as an independent variable. The only studies refer to Ro-ro ships.

Kristensen (2016) proposes a set of regressions as a function of LM. Part of the regressions are based on a power model as follows:

$$x = \alpha(LM)^{\beta} \cdot \epsilon \tag{11}$$

others are based on linear regressions in the following form:

$$x = \alpha + \beta(LM) + \epsilon \tag{12}$$

According to the study, the following set of equations for the main dimensions is proposed:

$$L_{PP} = \begin{cases} 20.4(LM)^{0.259} & \text{if } LM < 1,402 \\ 11.18(LM)^{0.342} & \text{if } LM \geq 1,402 \end{cases} \tag{13}$$

$$B = 5.49(LM)^{0.198} \tag{14}$$

$$T = \begin{cases} 1.9(LM)^{0.16} & \text{if } LM < 2,000 \\ 5.81 + 0.0003(LM) & \text{if } LM \geq 2,000 \end{cases} \tag{15}$$

$$D = 11.42 + 0.00172(LM) \tag{16}$$

Another study provided by Putra et al. (2022) gives other formulations for the main dimensions, analysing a database of vessels below 2500 GT. The given relationships are linear, thus following the model of Eq. (12). Regressions have the following form:

$$L_{OA} = 22.632 + 0.223(LM) \tag{17}$$

$$L_{PP} = 20.039 + 0.197(LM) \tag{18}$$

$$B = 7.698 + 0.038(LM) \tag{19}$$

No regressions are given for T and D. In addition to the main dimensions, the study provides a formulation for the gross tonnage GT:

$$GT = -31.382 + 6.034(LM) \tag{20}$$

The quality of the regression is assessed according to the determination coefficient R^2, finding values above 0.7 for the given relations. In any case, the study highlights a strong correlation in the database between the lane metres and the regressed variables.

2.3. Regression models based on length

The third kind of regression available for the main dimensions of the RoPax vessels employs the length between perpendiculars L_{PP} as the independent variable. The study of Kristensen (2016) reports a set of regressions, according to a linear model:

$$x = \alpha + \beta(L_{PP}) + \epsilon \tag{21}$$

From the regression analysis, the following parameters α and β result for the main dimensions:

$$L_{OA} = 1.93 + 1.078(L_{PP}) \tag{22}$$
$$B = 7.5 + 0.116(L_{PP}) \tag{23}$$
$$T_{min} = 0.95 + 0.028(L_{PP}) \tag{24}$$
$$T_{max} = 2.45 + 0.028(L_{PP}) \tag{25}$$
$$D = 6.94 + 0.05(L_{PP}) \tag{26}$$

where $T_{\min}$ and $T_{\max}$ are the minimum and maximum draught of the unit, respectively. Also in this case, the study indicates a strong correlation between the main dimensions and L_{PP}. The study of Kristensen (2016), as reported also by Friis et al. (2002) and Papanikolaou (2014), is not using the same amount of data for all the proposed regression models in the case of RoPax vessels.

However, besides the main dimensions, the study also reports some statistics for other characteristics relevant to the RoPax ships, such as the number of passengers N_p and the number of vehicles N_v. Analysing the data with the same model as in Eq. (21), the following regression can be derived from the reported data:

$$N_p = 24.784 + 9.1746(L_{PP}) \tag{27}$$
$$N_v = 180.16 + 4.0638(L_{PP}) \tag{28}$$

Even though a correlation exist for such variables, the determination coefficient R^2 for the two regressions is low.

In the recent years, a study from Novak et al. (2020) analyses a database of 128 vessels built until 2019, thus referring to a set of ships which includes also more modern vessels compared to previously mentioned studies. The authors propose two linear regressions for the breadth and draught according to the model presented in Eq. (21). The regressions have the following forms:

$$B = 0.1026(L_{PP}) + 8.8904 \tag{29}$$
$$T = 0.0271(L_{PP}) + 1.6105 \tag{30}$$

The formulations have a relatively high correlation coefficient of 0.679 and 0.726, respectively.

2.4. Other regressions

Besides length, deadweight and lane metres, other characteristics of the ship can be used as an independent variable for the estimation of main dimensions. As an example, the study of Kristensen (2016) proposes a regression for the length by employing the number of passengers N_p as the independent variable. The resulting power model is:

$$L_{PP} = 22.5(N_p)^{0.255} \tag{31}$$

Another parameter that could be of interest for the design of a RoPax vessel is the number of vehicles N_v. Therefore it could be possible that this parameter is available in the early stages of design. According to the study of Kristensen (2016), there is a strong correlation between the number of vehicles and the lane metres LM. This correlation results in the following linear model:

$$LM = 26.392 + 2.4461(N_v) \tag{32}$$

The selection of these kinds of equations depends on the availability of these data in the initial stage of design.

2.5. Strategies for selecting the main dimensions of RoPax ships

As highlighted in the previous sections, there are multiple methodologies available for the determination of the main dimensions of a RoPax vessel. There could be different inputs from shipowners, and depending on them, it may be possible to employ different equations.

If the deadweight DWT is given, it is possible to directly derive the dimensions applying the equations from (2) to (4). If the shipowner imposes a determinate amount of lane metres LM, then the equations from (13) to (16) could be used for the first estimate. Alternatively, if the L_{PP} is known, the equations to apply are from (22) to (26).

All these simple cases require the application of a unique set of equations. However, if other parameters are given, a combination of equation sets is required. If the number of passenger N_p is provided, then Eq. (31) estimates the L_{PP} and, consequently, equations from (22) to (26) determine the remaining dimensions. As another example, if the number of vehicles N_v is given, the lane metres LM can be obtained from Eq. (32) and then all the other dimensions are derived from the application of equations from (13) to (16).

In any case, all these kinds of solutions imply the availability of only one initial parameter for the estimation of the main dimensions. However, it is possible to have initial requirements that are not covered by the existing set of regressions, such as the speed V_s or the required power P_B. Furthermore, the sets of equations that could be used do not refer to homogeneous databases; therefore, the results may be affected by many uncertainties. It is necessary to identify alternative methods to derive a set of equations or models that allow for deriving the main dimensions from homogeneous databases, including the possibility of using more than one independent variable.

3. The RoPax database

The starting point for executing a regression analysis is the availability of a suitable database. This study utilises an initial set of 127 ships from an online database, containing information on the main dimensions, deadweight, displacement, lane meters, installed power, number of passengers, and vessel speed (Clarksons, 2024; Ferry-site, 2024).

3.1. Outliers and incomplete data elimination

For a statistical regression, it is crucial to eliminate outliers and incomplete datasets from the initial population, which means removing data points that lie outside the overall pattern in a distribution. Several methods in the literature exist for removing outliers from a population; this work adopts the interquartile range (IQR) rule. This common rule states that a data point is considered an outlier if it is more than 1.5 times the IQR above the third quartile (Q_3) or below the first quartile (Q_1), which means:

$$\begin{cases} x & \leq 1.5IQR - Q_1 \\ x & \geq 1.5IQR + Q_3 \end{cases} \tag{33}$$

where IQR is the distance between the first and the third quartiles:

$$IQR = Q_3 - Q_1 \tag{34}$$

Prior to outliers detection, the incomplete sets of data have been removed from the database, thus removing all ships lacking complete

parameter data. From such analysis, 22 ships have been removed, keeping only the ship having a complete set of parameters. Outlier detection has been conducted for all parameters of the database, resulting in the exclusion of additional 18 data points. Therefore, the final population is composed of 87 RoPax vessels. The vessels have a complete set of data for all the parameters. Because they all derive from the same data source, the data are homogeneous, especially for dimensions like the length where multiple definitions could be present (i.e. length overall, length between perpendiculars, or length at the waterline). In this specific case, data refer to the length between perpendiculars L_{PP}. Unfortunately, the data at disposal did not provide some additional information regarding the design of the vessel, thus not allowing to know which are the specific rules they have to satisfy and that could be a reason for becoming an outlier for the database.

3.2. Database analysis

The database is composed of a set of reasonably recent ships. Fig. 1 presents the histogram of the vessels' construction year. The ten oldest ships in the database were built before 2000 and, in any case, not before 1996. Most RoPax vessels (about 43) were built between 2000 and 2004. The most recent ships (about 4) are between 2020 and 2024. This composition represents an improvement over previous studies, as it presents more modern ships and does not include excessively old vessels.

Fig. 2 shows the distribution of L_{PP} values in the database. The data distribution highlights that most data (57 ships) lie between 160 and 200 metres. The data have a minimum length of 104 metres, a maximum of 226 and a mean of 177.21 with a standard deviation of 28.36 metres. As a consequence, the data are not uniformly distributed between the maximum and the minimum.

Fig. 3 shows the distribution of the breadth B among the database. The data have a minimum of 19 metres, a maximum of 35 and a mean of 26.51 with a standard deviation of 3.36 metres. As with the case of L_{PP}, the data are not uniformly distributed between the minimum and maximum.

Fig. 4 shows the distribution of the depth D among the database. The data have a minimum of 7 metres, a maximum of 22 and a mean of 12.69 with a standard deviation of 4.37 metres. The distribution shows a primary peak around 9 metres and two smaller peaks at 15 and 21 metres, respectively. This characteristic affects the homogeneity of the database as the data are clustered around the three peaks. Such behaviour may suggest that the origin of the data does not refer to the same conventions for all the vessels. Therefore, a statistical regression

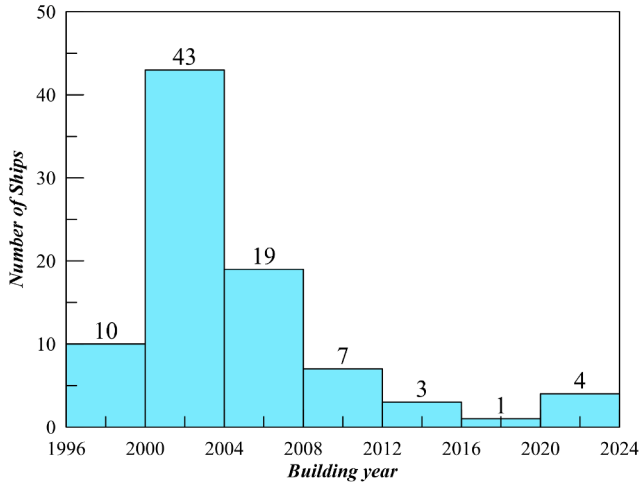


Fig. 1. RoPax construction years.

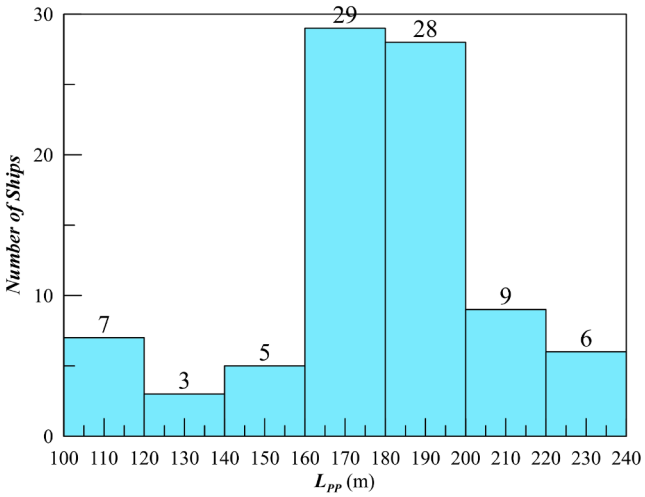


Fig. 2. L_{PP} distribution.

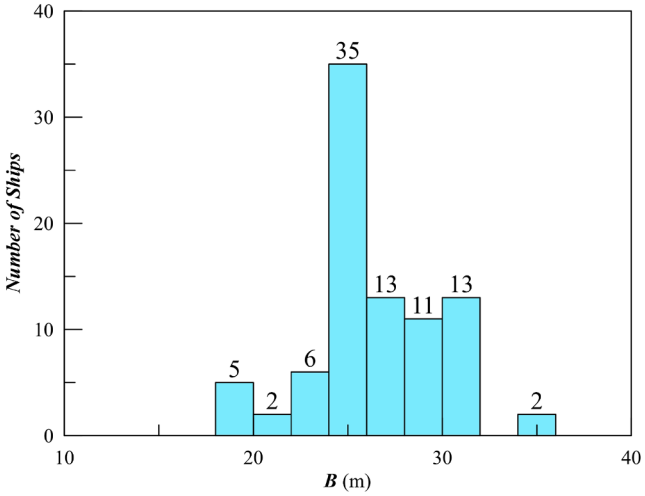


Fig. 3. B distribution.

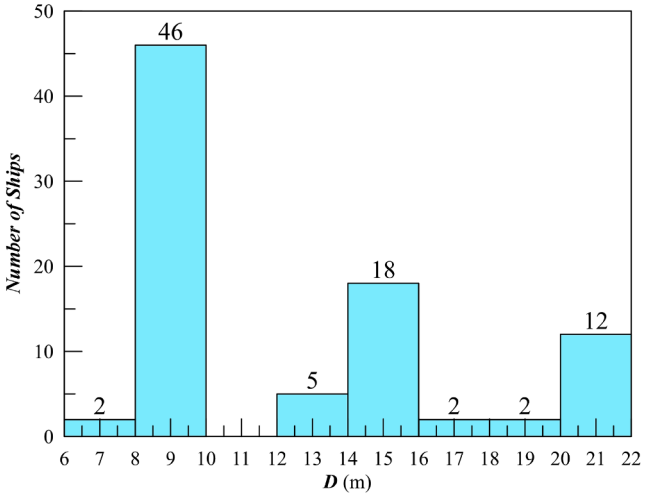


Fig. 4. D distribution.

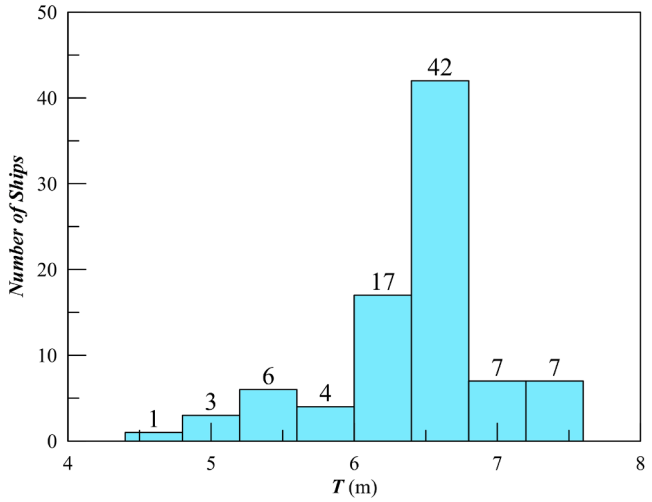


Fig. 5. T distribution.

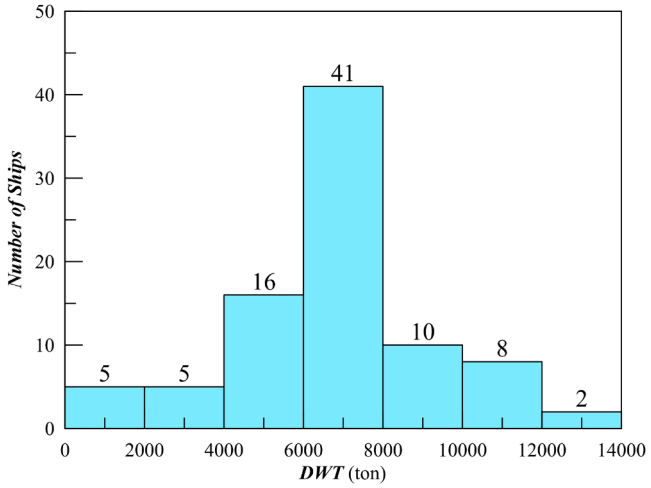


Fig. 7. DWT distribution.

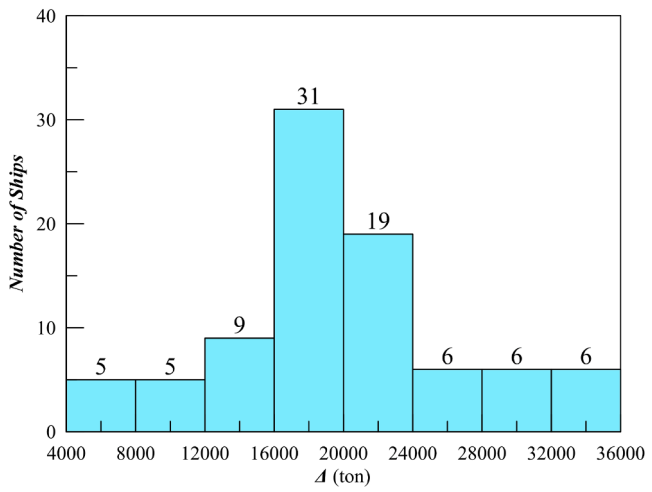


Fig. 6. Δ distribution.

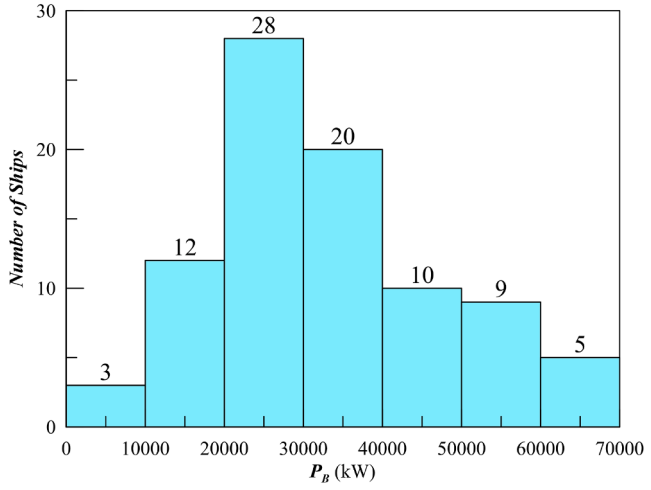


Fig. 8. P_B distribution.

on this data could generate an unsatisfactory reproduction of the initial population.

Fig. 5 shows the distribution of the draught T in the database. The data have a minimum of 5 metres, a maximum of 7 and a mean of 6.46 with a standard deviation of 0.56 metres. The data present a peak at 6.6 metres. The resulting distribution is skewed towards the maximum region; thus, there should be no issue in performing regression for this main dimension.

Fig. 6 shows the distribution of the displacement Δ among the database. The data have a minimum of 5200 tons, a maximum of 33,700 and a mean of 19,679.67 with a standard deviation of 6,748.31 tons. Analysing the figure, it is possible to conclude that the data are normally distributed between the minimum and the maximum.

Fig. 7 shows the distribution of the deadweight DWT among the database. The data have a minimum of 1200 tons, a maximum of 12,785 and a mean of 6,837.68 with a standard deviation of 2,524.07 tons. Displacement and deadweight are strongly correlated. Therefore, the distribution of DWT is similar to that of Δ; hence, the same considerations of DWT are valid for Δ.

Fig. 8 shows the distribution of the installed power P_B in the database. The data have a minimum of 8,640 kW, a maximum of 67,200 and a mean of 32,688.71 with a standard deviation of 14,484.26 kW. The peak of the distribution is at 25,000 kW, resulting in a moderate skew towards the minimum value.

Fig. 9 shows the distribution of the vessel speed V_s in the database. The data have a minimum of 18 knots, a maximum of 31 and a mean of 24.05 with a standard deviation of 3.09 knots. The highest density in the population is between 20 and 24 knots, with a consequent skew of the distribution towards the minimum of the data.

Fig. 10 shows the distribution of the number of passengers N_p among the database. The data have a minimum of 264 passengers, a maximum of 3000 and a mean of 1307.14, with a standard deviation of 689.56 passengers. The peak of the distribution is between 800 and 1,200, highlighting more density in the tale going to the maximum with respect to the minimum.

Fig. 11 shows the distribution of the lane metres LM among the database. The data have a minimum of 360, a maximum of 5500 and a mean of 2,228.78, with a standard deviation of 1,080.43 lane metres. The peak of the distribution is around 1600 lane metres, resulting in a skewness toward lower values.

However, for properly analysing the initial database, it is necessary to evaluate not only individual variables but also possible correlations among them. To this end, it is advisable to calculate the correlation matrix **C** between all the parameters. Following the same order as in

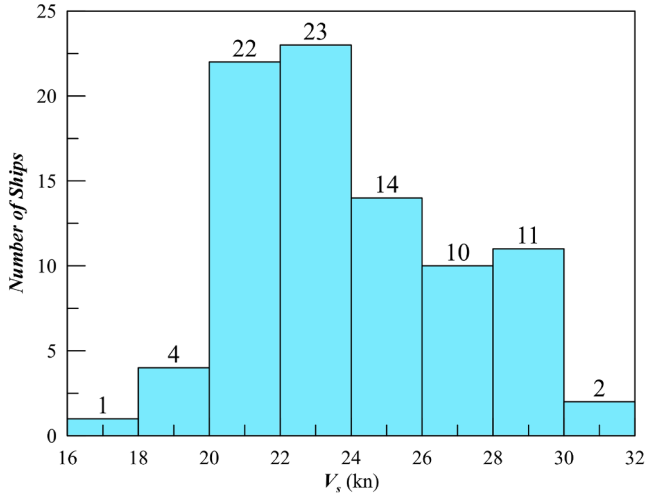


Fig. 9. V_s distribution.

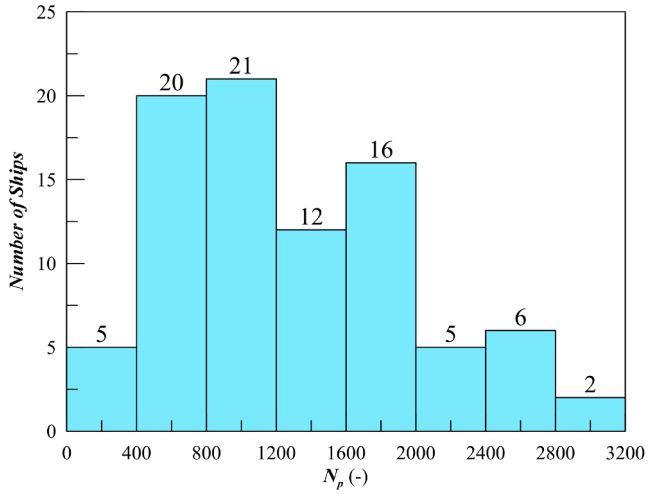


Fig. 10. N_p distribution.

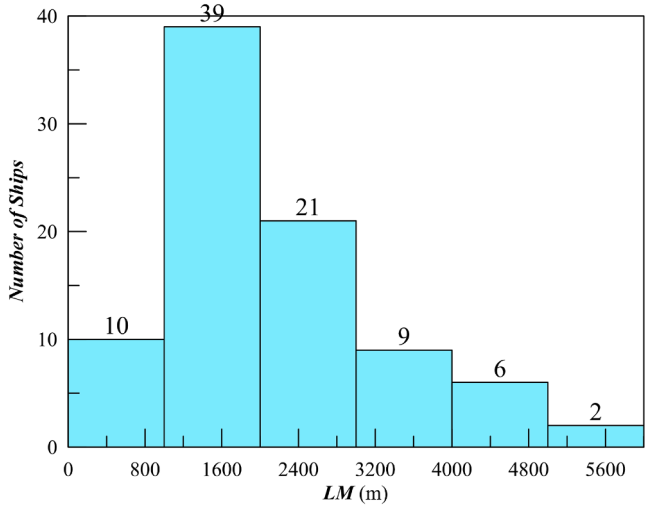


Fig. 11. LM distribution.

the above description, the correlation matrix has the following form:

$$\mathbf{C} = \begin{bmatrix} 1 & 0.76 & 0.20 & 0.68 & 0.86 & 0.89 & 0.40 & 0.17 & 0.21 & 0.75 \\ 0.76 & 1 & 0.38 & 0.57 & 0.73 & 0.93 & 0.21 & -0.15 & 0.28 & 0.66 \\ 0.20 & 0.38 & 1 & 0.14 & 0.12 & 0.35 & 0.29 & 0.03 & 0.34 & 0.16 \\ 0.68 & 0.57 & 0.14 & 1 & 0.54 & 0.61 & 0.64 & 0.49 & 0.43 & 0.32 \\ 0.86 & 0.73 & 0.12 & 0.54 & 1 & 0.83 & 0.14 & -0.09 & -0.06 & 0.86 \\ 0.89 & 0.93 & 0.35 & 0.61 & 0.83 & 1 & 0.34 & 0^* & 0.29 & 0.73 \\ 0.40 & 0.21 & 0.29 & 0.64 & 0.14 & 0.34 & 1 & 0.81 & 0.57 & 0.04 \\ 0.17 & -0.15 & 0.03 & 0.49 & -0.09 & 0^* & 0.81 & 1 & 0.48 & -0.18 \\ 0.21 & 0.28 & 0.34 & 0.43 & -0.06 & 0.29 & 0.57 & 0.48 & 1 & -0.16 \\ 0.75 & 0.66 & 0.16 & 0.32 & 0.86 & 0.73 & 0.04 & -0.18 & -0.16 & 1 \end{bmatrix} \tag{35}$$

In the matrix, the term 0^* indicates a value lower than 10^{-2}. Generally the more a term in **C** is close to 1 the most possibility is to have a correlation between two variables. However, there is no criterion to establish the appropriate threshold for detecting a correlation. On the other hand, it is possible to evaluate the matrix of the p-values **P** for testing the hypothesis that there is no relationship between the observed phenomena. In case a diagonal value of **P** is below a given threshold (generally 0.05), then the corresponding correlation in **C** is considered significant. The matrix **P** has the following form:

$$\mathbf{P} = \begin{bmatrix} 1 & 0^* & 0.07 & 0^* & 0^* & 0^* & 0^* & 0.12 & 0.05 & 0^* \\ 0^* & 1 & 0^* & 0^* & 0^* & 0^* & 0.05 & 0.17 & 0.01 & 0^* \\ 0.07 & 0^* & 1 & 0.20 & 0.29 & 0^* & 0.01 & 0.77 & 0^* & 0.14 \\ 0^* & 0^* & 0.20 & 1 & 0^* & 0^* & 0^* & 0^* & 0^* & 0^* \\ 0^* & 0^* & 0.29 & 0^* & 1 & 0^* & 0.21 & 0.42 & 0.59 & 0^* \\ 0^* & 0^* & 0^* & 0^* & 0^* & 1 & 0^* & 0.98 & 0.01 & 0^* \\ 0^* & 0.05 & 0.01 & 0^* & 0.21 & 0^* & 1 & 0^* & 0^* & 0.74 \\ 0.12 & 0.17 & 0.77 & 0^* & 0.42 & 0.98 & 0^* & 1 & 0^* & 0.09 \\ 0.05 & 0.01 & 0^* & 0^* & 0.59 & 0.01 & 0^* & 0^* & 1 & 0.14 \\ 0^* & 0^* & 0.14 & 0^* & 0^* & 0^* & 0.74 & 0.09 & 0.14 & 1 \end{bmatrix} \tag{36}$$

Also in this case, the term 0^* indicates a value lower than 10^{-2}. According to the criterion above, a strong correlation is identified when the p-value is less than 0.01, medium if between 0.01 and 0.05, and low otherwise.

From the analysis of the p-values matrix **P**, the following considerations can be drawn on the correlation between variables:

- *Length* L: the length (for brevity L_{PP} is indicated as L in the rest of the paper) presents a strong correlation with B, T, Δ, DWT,P_B and LM. The correlation with D and N_p is medium, while the correlation with V_s is low. Therefore, L is a suitable candidate to be an independent variable for the estimation of the main dimension of a RoPax with this database.
- *Breadth* B: the breadth presents a strong correlation with L, D, T, Δ, DWT and LM. The correlation with P_B and N_p is medium, while the correlation with V_s is low. Even though the literature studies are not considering B as an independent variable, the correlations suggest that also B can be considered as an independent variable with the actual database.
- *Depth* D: the depth presents a strong correlation with B, DWT and N_p. The correlation with L and P_B is medium, while the correlation with T, Δ, V_s and $LM$is low. Therefore, for the present database, D is not a suitable candidate to be an independent variable.
- *Draught* T: the draught presents a strong correlation with all the other variables except for D, where the correlation is low. Usually, T is not used as an independent variable for the regressions (as highlighted in the literature); however, it could be considered as a good candidate for the present database.
- *Displacement* Δ: the displacement presents a strong correlation with L, B, T, DWT and LM. The correlation with D, P_B, V_s and N_p is low. Then, Δ can be considered for the present database as a good candidate to be an independent variable for the estimation of the main dimensions.

Table 1
Qualitative correlation between the variables of the RoPax database.

	L	B	D	T	Δ	DWT	P_B	V_s	N_p	LM
L	–	strong	medium	strong	strong	strong	strong	low	medium	strong
B	strong	–	strong	strong	strong	strong	medium	low	medium	strong
D	medium	strong	–	low	low	strong	medium	low	strong	low
T	strong	strong	low	–	strong	strong	strong	strong	strong	strong
Δ	strong	strong	low	strong	–	strong	low	low	low	strong
DWT	strong	strong	strong	strong	strong	–	strong	low	medium	strong
P_B	strong	medium	medium	strong	low	strong	–	strong	strong	low
V_s	low	low	low	strong	low	low	strong	–	strong	medium
N_p	medium	medium	strong	strong	low	medium	strong	strong	–	low
LM	strong	strong	low	strong	strong	strong	low	medium	low	–

- *Deadweight* DWT: the deadweight presents a strong correlation with all the other variables except for V_s and N_p. With V_s the correlation is low, while with N_p the correlation is medium. Then, DWT can be considered a good candidate to be an independent variable for the estimation of RoPax main dimensions.
- *Power* P_B: the installed power presents a strong correlation with L, T, DWT, V_s and N_p. The correlation with B and D is medium, while the correlation with Δ and LM is low. Therefore, P_B can be considered as a possible candidate to be an independent variable for the estimation of the main dimensions.
- *Speed* V_s: the speed presents a strong correlation with T, P_B and N_p. The correlation with LM is medium, while the correlation with L, B, D, Δ and DWT is low. Therefore, V_s is not a suitable candidate to be the principal independent variable for the main dimensions estimation.
- *Number of passengers* N_p: the number of passengers presents a strong correlation with D, T, P_B and V_S. The correlation with L, B and DWT is medium, while the correlation with Δ and LM is low. Therefore, N_p is not the best candidate to be the principal independent variable for the main dimensions estimation of RoPax.
- *Lane metres* LM: the lane metres presents a strong correlation with L, B, T, Δ and DWT. The correlation with V_s is medium, while the correlation with D, P_B and N_p is low. Therefore, LM can be a suitable independent variable for the estimation of RoPax main dimensions.

All the qualitative correlations listed above are summarised in Table 1, while an overview of the pairwise distributions of the variables is given in Fig. 12.

Analysing the figure it is possible to observe how the data are distributed among the variables, highlighting how simple regressions, like the ones described in Section 2, well represent the main dimensions and the other parameters for the cases of L, Δ, DWT and LM. Such a consideration suggests to try the simple regression analysis using those variables as the independent one. However, the necessities of satisfy all the requirement of a ship owner may request the availability of regressions that includes more than a parameter as independent variable, thus including characteristics like P_B, V_s or N_p that are not traditionally used in simple regressions. The correlation study, thus the matrix $\mathbf{P}$ is helpful in advising which variables are not correlated with each other, avoiding, for instance, autocorrelation problems in multiple linear regression analysis. However, such points will be further discussed in the coming sections.

3.3. Training and test datasets

The present work employs both traditional regression techniques, like simple and multiple linear regressions, and machine learning techniques as the random tree forest. Conventional studies in the field of regressions consider all the data at disposal in a database to fit the regression and usually the quality of fit is assessed on the data used for determining the models (Fisher, 1922; Freedman, 2009).

However, with the rising applicability of machine learning techniques for regression problems, it is necessary to change conventional approaches in such a way to have a fair comparison between machine learning-based regressions and traditional ones. In fact, once a machine learning method is applied, the original database should be divided into two parts, one used for the training of the models and one for the evaluation of the models, in such a way to provide an unbiased estimation of the final models.

In this study, the same data division approach has also been adopted for simple and multiple linear regressions. The initial dataset of 87 RoPax vessels has been divided into a training set (80 % of the ships) and a test set (the remaining 20 %). The test set, often referred to as the holdout set, is obtained by removing a random sample from the full dataset.

As a result of this holdout methodology, the training set contains 70 ships, while the test set consists of 17 ships. Although this division is generally performed randomly, the relatively small size of the dataset in this case may lead to the test set not fully covering the variability of the dataset. Therefore, a stratified random sampling has been performed trying to ensure good coverage of the represented data. However, the relatively low number of data, does not allow to well represent the tales of the distribution for all the variables as this may require the availability of thousands of data. This is not a problem for model verification, as, in any case, the dataset is well covered by the provided samples.

Fig. 13 shows the distributions of the training and test sets for the length. It can be observed that the test partition is covering the whole range of the database. Fig. 14 shows the distributions of the training and test sets for the breadth. From the figure it results that only the upper tail of the database is not represented by the test set for this variable. Fig. 15 shows the distribution of the training and test sets for the depth. Also in this case, the distribution of the test data is well covering the database. Fig. 16 shows the distribution of the training and test sets for the draught. The test partition for this variable is lacking in coverage only for the upper tail of the distribution. Figs. 17 and 18 show the training and test sets for the DWT and Δ, respectively. In both cases, the test set covers properly the whole database. Figs. 19 and 20 show the training and test sets for the installed power and vessel speed, respectively. For both variables, the test set is not covering just the upper tail of the distributions. Fig. 21 shows the training and test sets for the number of passengers. For this variable, the test set is not covering the upper tail of the distribution. Finally, Fig. 22 shows the training and test sets for the lane metres. In this case the test set is covering the whole extension of the database. Thanks to the stratified approach adopted for selecting the holdout test set, it can be concluded that the resulting test data form a representative partition of the initial dataset. This makes them suitable for verifying the generalisation capabilities of the regression models.

Therefore, all regression analyses were conducted using the training set, with training performance assessed for each model. Afterwards, the models were evaluated on the test set to verify their performance and validate the generalisation of the obtained regression results.

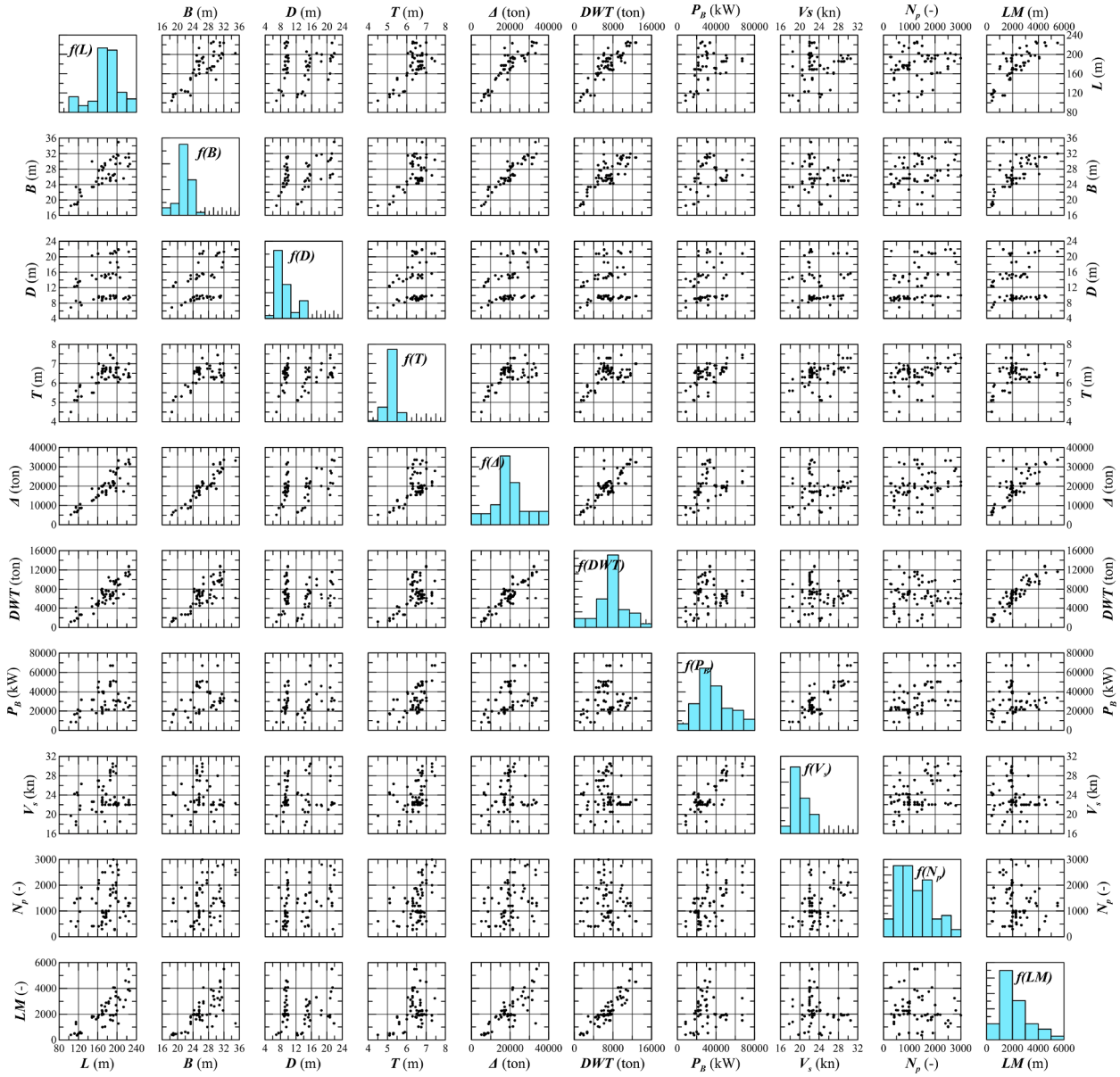


Fig. 12. Pairwise distributions of database variables.

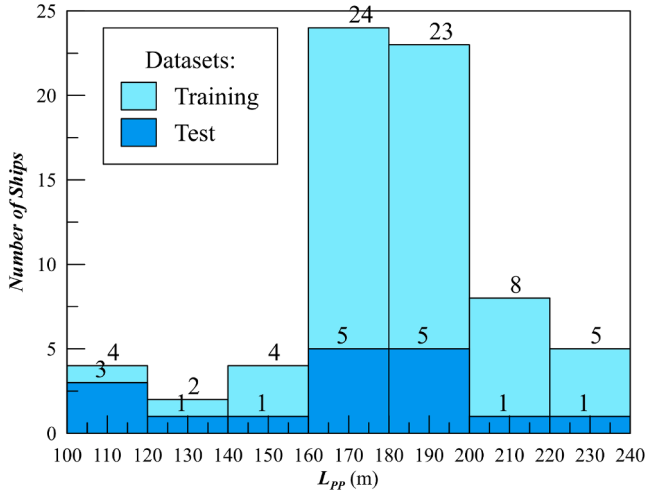


Fig. 13. L distribution for training and test sets.

4. Regression methodology

In the present study, different types of regression models are investigated to predict the main dimensions of a RoPax vessel, using the database presented in Section 3. The present section describes the methodology adopted to perform the regression analysis by means of the following models:

- Simple models (linear, power or logarithmic).
- Multiple linear regressions.
- Forest trees regressions.

In addition to describing the aforementioned methods, this section also presents the parameters used to evaluate the quality of the resulting regressions.

4.1. Simple regressions

According to the literature, the most common approach for performing regression analysis on a ship's main dimensions is through simple

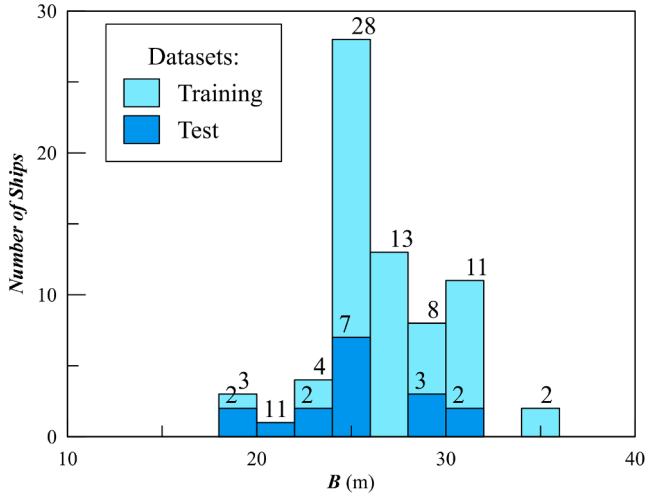


Fig. 14. B distribution for training and test sets.

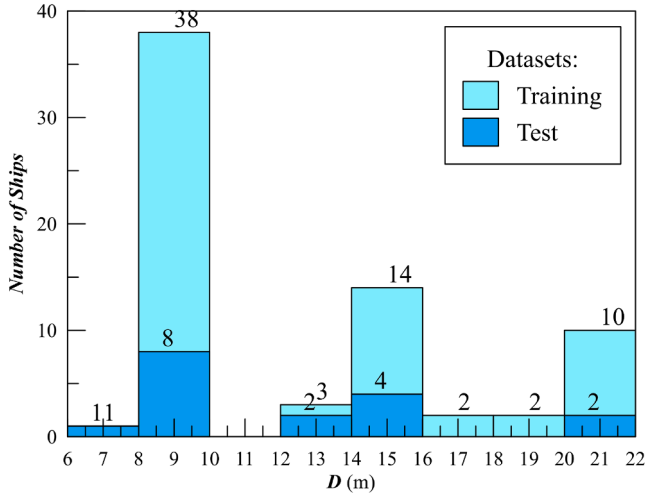


Fig. 15. D distribution for training and test sets.

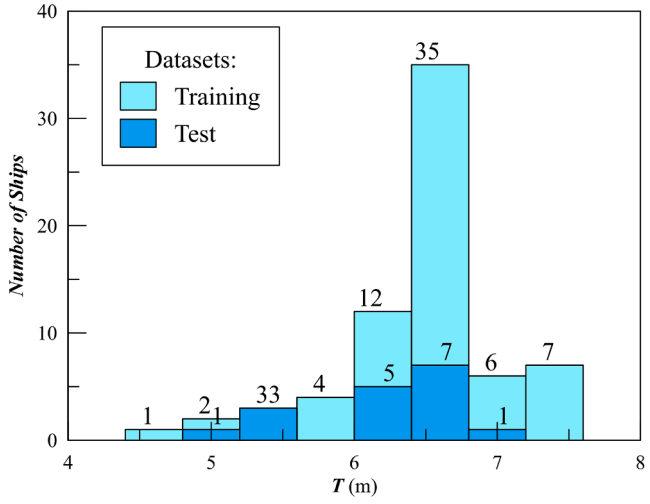


Fig. 16. T distribution for training and test sets.

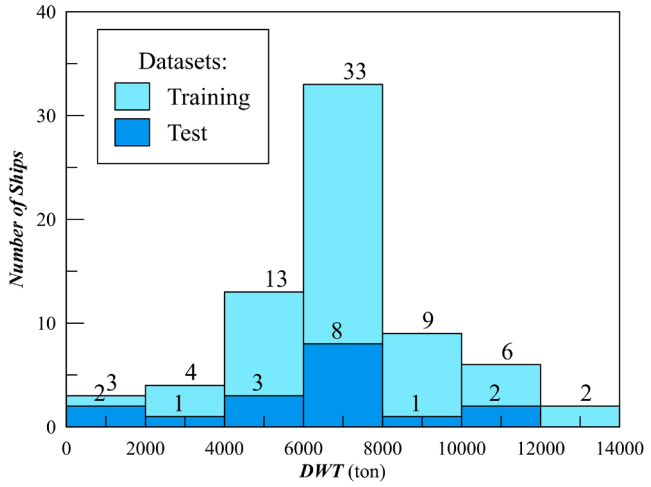


Fig. 17. DWT distribution for training and test sets.

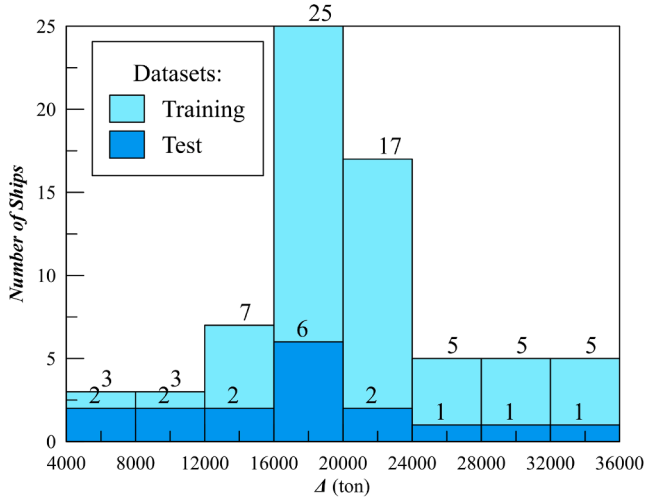


Fig. 18. Δ distribution for training and test sets.

regression models. These models are typically based on linear, power, or logarithmic regressions. A linear model takes the following general form:

$$y = \alpha + \beta x + \epsilon \tag{37}$$

where α and β are the regression coefficients and ϵ is the residual error. A common way to evaluate α and β is by means of the least squares method. According to this methodology, the following estimation is valid:

$$\beta = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{(x_i - \bar{x})^2} \tag{38}$$

$$\alpha = \bar{y} - \beta \bar{x} \tag{39}$$

where (x_i, y_i) are the data points and $\bar{x}$ and $\bar{y}$ are the mean values of the vectors.

When a power or a logarithmic model are selected, the following equations became representative:

$$y = \alpha x^{\beta} \cdot \epsilon \tag{40}$$

$$y = \alpha + \beta \ln x + \epsilon \tag{41}$$

Also for Eqs. (40) and (41) the least square method can be used to identify the regression parameters α and β. In fact, by means of proper transformations, both the equations can be rewritten in the form of Eq. (37),

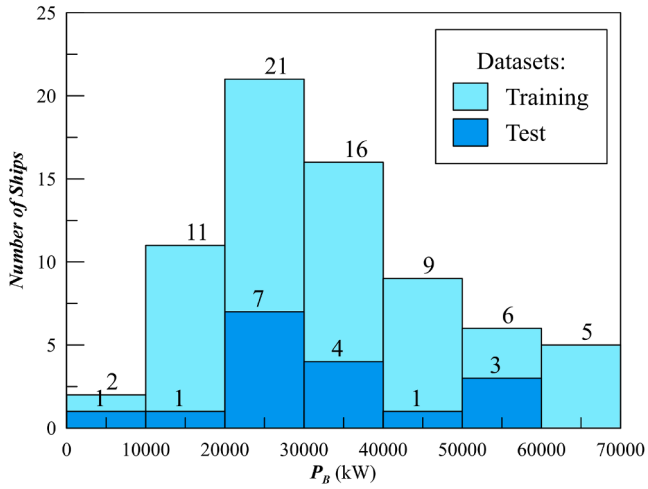


Fig. 19. P_B distribution for training and test sets.

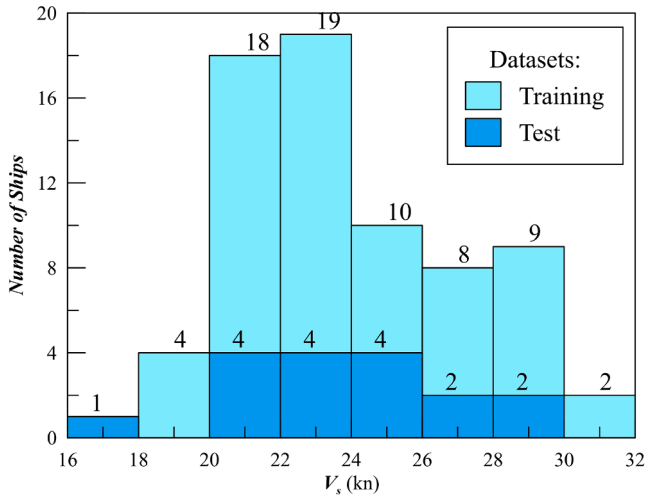


Fig. 20. V_s distribution for training and test sets.

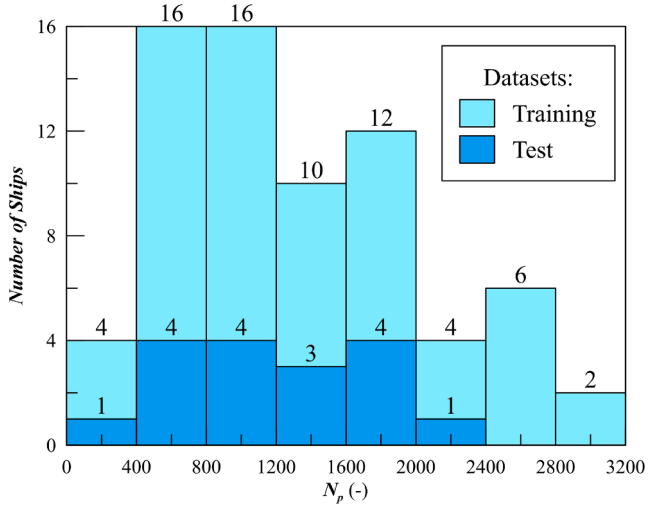


Fig. 21. N_p distribution for training and test sets.

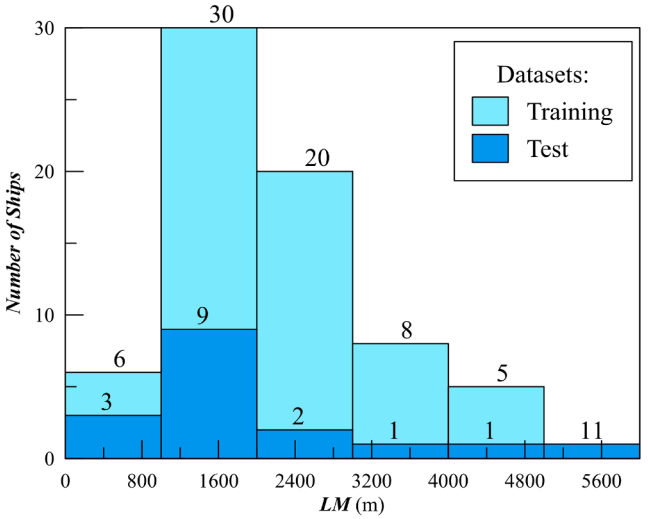


Fig. 22. LM distribution for training and test sets.

thus applying Eqs. (38) and (39) for the determination of regression parameters.

In this case, simple regression models use deadweight (DWT) as the independent variable, while displacement (Δ), lane meters (LM), and length (L) are the dependent variables. For DWT, LM, and L, it is possible to compare the regression results with those from studies available in the literature and presented in Section 2.

4.2. Multiple linear regressions

As highlighted in previous sections, ship designers may consider multiple parameters when selecting the main dimensions of a vessel, based on the requirements of the shipowner. While simple regressions are limited to a single independent variable, multiple linear regression (MLR) enables the use of several input variables simultaneously.

The general model for a multiple linear regression is given by the following matrix formulation:

$$\mathbf{y} = \boldsymbol{\alpha}\mathbf{x} + \boldsymbol{\epsilon} \tag{42}$$

where $\mathbf{y}$ is the matrix of the measured variable, $\mathbf{x}$ is the matrix of the independent variables, $\boldsymbol{\alpha}$ is the matrix of the regression coefficients and $\boldsymbol{\epsilon}$ are the errors.

According to the matrix formulation of Eq. (42), the matrix of the regression coefficients $\boldsymbol{\alpha}$ is unknown. The problem can be solved as follows:

$$\boldsymbol{\alpha} = \left(\mathbf{x}'\mathbf{x}\right)^{-1} \cdot \mathbf{x}'\mathbf{y} \tag{43}$$

where $\mathbf{x}'$ is the transpose of matrix $\mathbf{x}$ and $\left(\mathbf{x}'\mathbf{x}\right)^{-1}$ is the inverse of matrix $\mathbf{x}'\mathbf{x}$.

In the present study, the multiple linear regression analysis is applied to obtain regressions that includes more than one independent variable. The following cases are tested:

- Vessel speed V_s and deadweight DWT.
- Vessel speed V_s and lane metres LM.
- Number of passengers N_p and deadweight DWT.
- Number of passengers N_p and LM.
- Number of passengers N_p, vessel speed V_s and deadweight DWT.
- Number of passengers N_p, vessel speed V_s and lane metres LM.

These combinations were selected to avoid autocorrelation among independent variables, as identified in Section 3. Moreover, vessel speed and number of passengers are parameters often specified by shipowners.

4.3. Forest tree algorithms

Multiple linear regression is not the only technique to obtain a model that consider more than one independent variable. In this sense, machine learning techniques could be a possible alternative solution. There are multiple methodologies in machine learning that could be used for regression purposes, starting from neural networks up to forest tree algorithms. Having to deal with a relatively low number of data, the neural networks are not advisable to be used. On the other hand, forest tree works well with a low amount of data and, as shown by Rinauro et al. (2024) they can provide an enhancement in the accuracy of the regression compared to standard methods.

Forest tree algorithm is a Supervised Machine Learning Algorithm widely used in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression. Random forest is a versatile machine learning algorithm that leverages an ensemble of multiple decision trees to generate predictions or classifications. The random forest algorithm delivers a consolidated and more accurate result by combining the outputs of these trees. Its widespread popularity stems from its user-friendly nature and adaptability, which enables the effective tackling of classification and regression problems. The algorithm's strength lies in its ability to handle complex datasets and mitigate overfitting, making it a valuable tool for various predictive tasks in machine learning. One of the most relevant features of the Random Forest Algorithm is that it can handle the data set containing continuous variables, as in the case of regression, and categorical variables, as in the case of classification. It performs better for classification and regression tasks. Despite its strengths, one drawback of the forest tree algorithm is its black-box nature–the model does not provide explicit regression coefficients, making it less interpretable than traditional regression models.

In this study, forest tree algorithms are applied to the same variable combinations tested in the multiple linear regression models, enabling a direct comparison of performance between the two approaches.

4.4. Regression quality assessment

Beyond the methodology used to derive regression models, it is crucial to adopt appropriate metrics to assess their quality. A consistent evaluation framework allows for the comparison of models based on different techniques. For this purpose, in the present study use has been made of the following fit coefficients to asses the quality of the regressions and compare the different formulations: coefficient of determination R^2, Pearson coefficient Prs, $MAPE$ (Mean Absolute Percentage Error), $RMSE$ (Root Mean Square Error) and $RRMSE$ (Relative Root Mean Square Error). In particular, for multiple linear regression analyses, additional parameters have been evaluated: the adjusted coefficient of determination R^2_{adj} for the regression itself and the SE (Standard Error), t-stud and p-value for all the regression coefficients.

Fit coefficients have the following formulations:

$$R^2 = 1 - \frac{SS_E}{SS_{tot}} = 1 - \frac{\sum_{i=1}^{n}\left(y_i - y_i^*\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2} \tag{44}$$

$$R^2_{adj} = 1 - \left(1 - R^2\right)\frac{n-1}{n-n_p-1} \tag{45}$$

$$MAPE = \frac{\sum_{i=1}^{n}\left|y_i - y_i^*\right|}{n} \tag{46}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - y_i^*\right)^2}{n}} \tag{47}$$

$$RRMSE = \sqrt{\frac{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - y_i^*\right)^2}{\sum_{i=1}^{n}\left(y_i^*\right)^2}} \tag{48}$$

$$Prs = \frac{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)\left(y_i - \bar{y^*}\right)}{\sqrt{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}\sqrt{\sum_{i=1}^{n}\left(y_i - \bar{y^*}\right)^2}} \tag{49}$$

where y_i are the n data, y_i^* are the predicted values, $\bar{y}$ is the mean of the data and $\bar{y^*}$ is the mean of the predicted values. n_p is the number of parameters used in the regression model.

To judge the quality of the regressions by means of the above mentioned parameters it is worthy to consider that R^2, R^2_{adj} and Prs are attributes to maximise, while $MAPE$, $RMSE$ and $RRMSE$ are attributes to minimise.

The regression quality assessment is here performed in two steps. Firstly, the quality is assessed on the training set used to determine the regression structures. This assessment allows to judge the performances of the conventional regression techniques such as simple and multiple linear regressions, but may be not suitable for machine learning approaches like the forest trees as it may be affected by overfitting issues (even though the algorithm itself is minimising the issue compared to traditional trees regression models). For this purpose, a second step is needed to judge the regressions by using the test set (not used to determine the regression models), thus providing a proof of the general applicability of the obtained regressions.

5. Regression results

This section presents the results of the regression analysis performed according to the methods described in Section 4, thus with simple, multivariate regressions and forest tree. Results are explained by mean of graphs and tables, comparing, where possible, the results with the studies available in the literature.

5.1. Simple regressions

Simple regression analysis has been carried out considering three typologies of regression: linear regression, power regression and logarithmic regression. Besides, different independent variables have been considered to carry out the regression analysis, the deadweight DWT, the displacement Δ, the lane metres LM and the length L. This section reports the results of the simple regression analysis, comparing the different typologies of regressions tested among them and with the literature studies previously described in Section 2.

5.1.1. Regressions as a function of DWT

As mentioned above, the regressions as a function of DWT have been performed according to a linear, a power and a logarithmic model. The following formulae have been derived for all the variables of the database (except, of course, for DWT) according to the linear model expressed by Eq. (37):

$$L = 0.0091(DWT) + 115.9789 \tag{50}$$

$$B = 0.0009(DWT) + 20.5499 \tag{51}$$

$$D = 0.0002(DWT) + 11.5837 \tag{52}$$

$$T = 0.0001(DWT) + 5.7960 \tag{53}$$

$$\Delta = 2.0746(DWT) + 5724.4001 \tag{54}$$

$$P_B = 0.6073(DWT) + 28796.9300 \tag{55}$$

$$V_s = -0.0001(DWT) + 24.8490 \tag{56}$$

$$N_p = -0.0257(DWT) + 1513.5019 \tag{57}$$

$$LM = 0.3675(DWT) - 289.7683 \tag{58}$$

The application of a power model according to Eq. (40) leads to the following formulae:

$$L = 10.6444(DWT)^{0.3207} \tag{59}$$

$$B = 4.3893(DWT)^{0.2054} \tag{60}$$

$$D = 5.1065(DWT)^{0.0984} \tag{61}$$

$$T = 2.0488(DWT)^{0.1316} \tag{62}$$

$$\Delta = 35.6960(DWT)^{0.7159} \tag{63}$$

$$P_B = 1534.6988(DWT)^{0.3386} \tag{64}$$

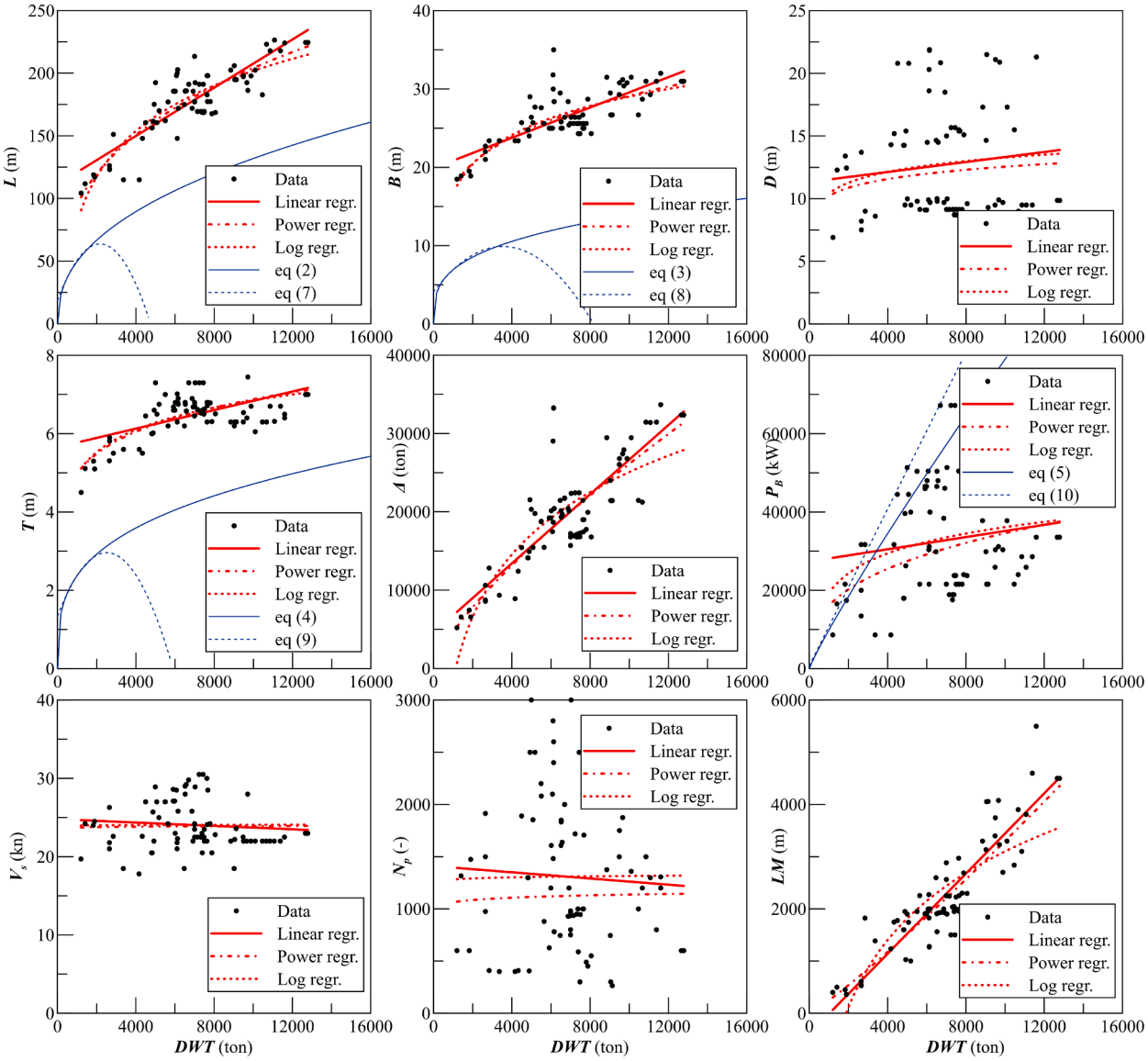


Fig. 23. Simple regressions on the present database as a function of DWT.

$$V_s = 21.9767(DWT)^{0.0094} \quad (65)$$

$$N_p = 1253.5970(DWT)^{-0.0112} \quad (66)$$

$$LM = 0.1819(DWT)^{1.0624} \quad (67)$$

Finally, by applying the logarithmic model of Eq. (41) the following formulae can be derived from the database:

$$L = -266.3449 + 50.8307 \ln DWT \quad (68)$$

$$B = -17.6005 + 5.0609 \ln DWT \quad (69)$$

$$D = 1.5114 + 1.2860 \ln DWT \quad (70)$$

$$T = -0.1928 + 0.7652 \ln DWT \quad (71)$$

$$\Delta = -76644.9219 + 11041.8915 \ln DWT \quad (72)$$

$$P_B = -27743.8010 + 6932.8717 \ln DWT \quad (73)$$

$$V_s = 22.5528 + 0.1722 \ln DWT \quad (74)$$

$$N_p = 1509.4259 - 19.9158 \ln DWT \quad (75)$$

$$LM = -13469.0507 + 1794.8726 \ln DWT \quad (76)$$

Fig. 23 shows all the derived equations together with the comparison with available literature studies, while Table 2 reports the quality of the regressions obtained for the formulae as a function of DWT.

From a preliminary analysis of the results it is possible to compare the obtained regressions with the studies available in the literature. Considering the study of Piko (1980), it is evident that the database used for that study is composed of small old ship and extrapolating the results to larger ships like in the present database underestimate all the results. The quadratic model adopted by Piko (1980) highlights a validity only for ships with less than 2500 ton of DWT. Therefore the simple regressions as a function of DWT proposed in this work are for sure an improvement compared to the available studies. However several considerations and ranking between the regression types can be made.

The simple regressions present three different models, the linear, the power and the logarithmic one. Analysing in detail the results of the quality of fit presented in Table 2, it is possible to better understand which one is the best model for each variable. Furthermore, it is also possible to assess whether a regression model is significant or not for the given variable. A detailed analysis variable by variable of the results is reported in Appendix A.

5.1.2. *Regressions as a function of* Δ

As for the case of DWT, the regressions as a function of Δ have been also performed according to a linear, a power and a logarithmic model.

Table 2
Quality of fit for the simple regressions as a function of DWT.

	Linear					Power					Logarithmic				
	R^2	$MAPE$	$RMSE$	$RRMSE$	Prs	R^2	$MAPE$	$RMSE$	$RRMSE$	Prs	R^2	$MAPE$	$RMSE$	$RRMSE$	Prs
L	0.700	0.068	14.478	0.129	0.836	0.748	0.060	13.262	0.118	0.865	0.745	0.063	13.346	0.119	0.863
B	0.456	0.064	2.371	0.054	0.675	0.496	0.061	2.282	0.052	0.705	0.499	0.062	2.276	0.052	0.706
D	0.009	0.316	4.407	0.147	0.095	0.006	0.294	4.442	0.152	0.127	0.017	0.317	4.389	0.146	0.130
T	0.221	0.060	0.471	0.022	0.470	0.383	0.053	0.420	0.019	0.622	0.412	0.051	0.409	0.019	0.642
Δ	0.611	0.158	4.0E3	3.387	0.782	0.615	0.143	4.0E3	3.400	0.786	0.590	0.164	4.1E3	3.478	0.768
P_B	0.009	0.458	1.4E4	9.686	0.099	0.018	0.385	1.5E4	10.28	0.172	0.044	0.440	1.4E4	9.517	0.210
V_s	0.008	0.103	3.061	0.074	0.089	0.003	0.103	3.079	0.075	0.024	0.000	0.105	3.073	0.074	0.025
N_p	0.007	0.680	7.1E2	2.330	0.087	0.077	0.577	7.4E2	2.632	0.011	0.000	0.681	7.1E2	2.339	0.012
LM	0.786	0.185	4.6E2	1.169	0.887	0.780	0.178	4.7E2	1.202	0.889	0.639	0.282	6.0E2	1.520	0.799

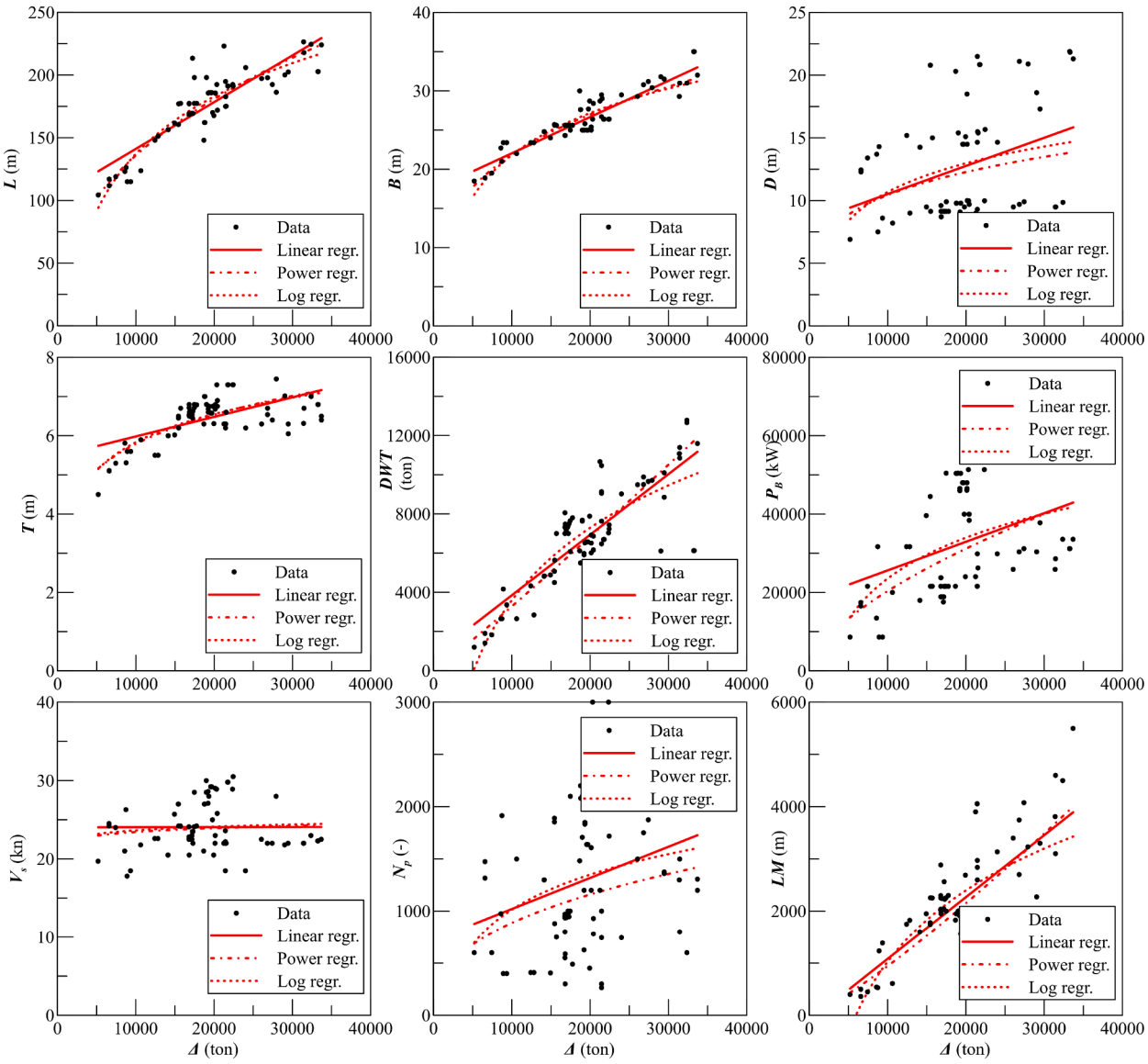


Fig. 24. Simple regressions on the present database as a function of Δ.

By employing the linear model, according to Eq. (37), the following formulae can be derived:

$$L = 0.0036(\Delta) + 107.0725 \tag{77}$$

$$B = 0.0005(\Delta) + 17.4034 \tag{78}$$

$$D = 0.0002(\Delta) + 7.6581 \tag{79}$$

$$T = 4.6856E-5(\Delta) + 5.5705 \tag{80}$$

$$DWT = 0.2948(\Delta) + 1006.3352 \tag{81}$$

$$P_B = 0.7774(\Delta) + 17370.90044 \tag{82}$$

$$V_s = 7.1971E-6(\Delta) + 23.9167 \tag{83}$$

$$N_p = .0360(\Delta) + 610.7479 \tag{84}$$

$$LM = 0.1044(\Delta) - 158.5059 \tag{85}$$

Table 3
Quality of fit for the simple regressions as a function of Δ.

	Linear					Power					Logarithmic				
	R^2	$MAPE$	$RMSE$	$RRMSE$	Prs	R^2	$MAPE$	$RMSE$	$RRMSE$	Prs	R^2	$MAPE$	$RMSE$	$RRMSE$	Prs
L	0.762	0.055	12.87	0.114	0.873	0.817	0.043	11.28	0.100	0.905	0.831	0.045	10.85	0.096	0.911
B	0.867	0.037	1.168	0.027	0.931	0.853	0.036	1.232	0.028	0.924	0.826	0.040	1.339	0.030	0.909
D	0.137	0.298	4.111	0.137	0.371	0.095	0.285	4.211	0.144	0.345	0.109	0.305	4.178	0.139	0.330
T	0.319	0.057	0.441	0.020	0.565	0.459	0.051	0.393	0.018	0.680	0.491	0.049	0.381	0.017	0.701
DWT	0.611	0.213	1.5E3	2.174	0.782	0.579	0.177	1.5E3	2.281	0.779	0.643	0.191	1.4E3	2.085	0.801
P_B	0.114	0.414	1.3E4	9.159	0.338	0.099	0.347	1.4E4	9.580	0.365	0.167	0.385	1.3E4	8.883	0.409
V_s	0.000	0.105	3.073	0.074	0.015	0.007	0.104	3.063	0.074	0.105	0.011	0.106	3.056	0.074	0.108
N_p	0.105	0.648	6.7E2	2.213	0.324	0.029	0.548	7.0E2	2.483	0.319	0.093	0.659	6.8E2	2.227	0.306
LM	0.447	0.313	7.4E2	1.883	0.668	0.428	0.292	7.6E2	1.960	0.666	0.449	0.294	7.4E2	1.880	0.670

The application of a power model according to Eq. (40) leads to the following formulae:

$$L = 3.2581(\Delta)^{0.4056} \quad (86)$$
$$B = 1.2892(\Delta)^{0.3072} \quad (87)$$
$$D = 0.7203(\Delta)^{0.2865} \quad (88)$$
$$T = 1.2094(\Delta)^{0.1706} \quad (89)$$
$$DWT = 0.2051(\Delta)^{1.0506} \quad (90)$$
$$P_B = 40.1151(\Delta)^{0.6713} \quad (91)$$
$$V_s = 16.3955(\Delta)^{0.0381} \quad (92)$$
$$N_p = 15.1221(\Delta)^{0.4386} \quad (93)$$
$$LM = 0.0457(\Delta)^{1.0855} \quad (94)$$

Finally, by applying the logarithmic model of Eq. (41) the following formulae can be derived from the database:

$$L = -461.4721 + 65.0390 \ln \Delta \quad (95)$$
$$B = -50.9617 + 7.8903 \ln \Delta \quad (96)$$
$$D = -26.0660 + 3.9443 \ln \Delta \quad (97)$$
$$T = -3.4476 + 1.0113 \ln \Delta \quad (98)$$
$$DWT = -44878.4991 + 5261.0606 \ln \Delta \quad (99)$$
$$P_B = -127561.0944 + 16333.9041 \ln \Delta \quad (100)$$
$$V_s = 15.2096 + 0.8988 \ln \Delta \quad (101)$$
$$N_p = -4481.8894 + 590.6043 \ln \Delta \quad (102)$$
$$LM = -15681.3318 + 1821.6192 \ln \Delta \quad (103)$$

Fig. 24 shows the obtained regressions for the present database. For the case of Δ no regressions are available from the literature; therefore, no comparison is possible with previous studies. Table 3 reports the quality of fit for the regression obtained as a function of Δ.

Also in this case, the simple regressions have been carried out according to the linear, the power and the logarithmic model. Analysing in detail the results reported in Table 3 it is possible to understand which one is the best model for each variable. A detailed variable by variable analysis is reported in Appendix A

5.1.3. Regressions as a function of LM

Also for the regressions as a function of the lane metres LM the analysis has been carried out with either liner, power and logarithmic model. The application of the linear model of Eq. (37) leads to the following formulae:

$$L = 0.0191(LM) + 135.8495 \quad (104)$$
$$B = 0.0019(LM) + 22.4928 \quad (105)$$
$$D = 0.0005(LM) + 11.6186 \quad (106)$$
$$T = 0.0001(LM) + 6.2063 \quad (107)$$
$$DWT = 2.1410(LM) + 2099.3561 \quad (108)$$
$$P_B = -0.1125(LM) + 33263.7538 \quad (109)$$
$$V_s = -0.0006(LM) + 25.4287 \quad (110)$$
$$N_p = -0.1558(LM) + 1687.0549 \quad (111)$$
$$\Delta = 4.2818(LM) + 10441.0113 \quad (112)$$

The application of a power model according to Eq. (40) leads to the following formulae:

$$L = 27.1790(LM)^{0.2463} \quad (113)$$
$$B = 8.1104(LM)^{0.1559} \quad (114)$$
$$D = 7.4847(LM)^{0.0631} \quad (115)$$
$$T = 3.4874(LM)^{0.0816} \quad (116)$$
$$DWT = 18.6103(LM)^{0.7677} \quad (117)$$
$$P_B = 5794.5780(LM)^{0.2454} \quad (118)$$
$$V_s = 26.8194(LM)^{-0.0153} \quad (119)$$
$$N_p = 5419.8755(LM)^{-0.2054} \quad (120)$$
$$\Delta = 324.7679(LM)^{0.5345} \quad (121)$$

Finally, by applying the logarithmic model of Eq. (41) the following formulae can be derived from the database:

$$L = -119.6957 + 39.2823 \ln LM \quad (122)$$
$$B = -2.2645 + 3.8145 \ln LM \quad (123)$$
$$D = 6.1339 + 0.8739 \ln LM \quad (124)$$
$$T = 2.9711 + 0.4657 \ln LM \quad (125)$$
$$DWT = -23692.1931 + 4026.9431 \ln LM \quad (126)$$
$$P_B = 6491.7647 + 3486.3961 \ln LM \quad (127)$$
$$V_s = 27.2772 - 0.4228 \ln LM \quad (128)$$
$$N_p = 3076.7567 - 229.0082 \ln LM \quad (129)$$
$$\Delta = -42225.0027 + 8196.2275 \ln LM \quad (130)$$

Fig. 25 shows the obtained regressions for the present database together with a comparison with the regression available in the literature, in particular with the study of Kristensen (2016) and Putra et al. (2022) already presented in this work. Taking a look to the Figure allows for comparing the literature regression with the present database. It is evident for the regressions of Putra et al. (2022), that the validity of those formulae can be considered only for small vessels with less than 1000 metres of LM. In fact, by increasing the LM the regressions are overestimating the variables compared to the database values.

Different is the case of the regressions proposed by Kristensen (2016). For the length L, there is an underestimation compared to the values of the present database, for the breadth B the underestimation is less strong then L. Considering the depth D, the equation of Kristensen (2016) fit well a subpopulation of the presented data but overestimate the mean value of the whole database. Finally, for the draught T, there is a underestimation for small vessels and an overestimation for the larger ones.

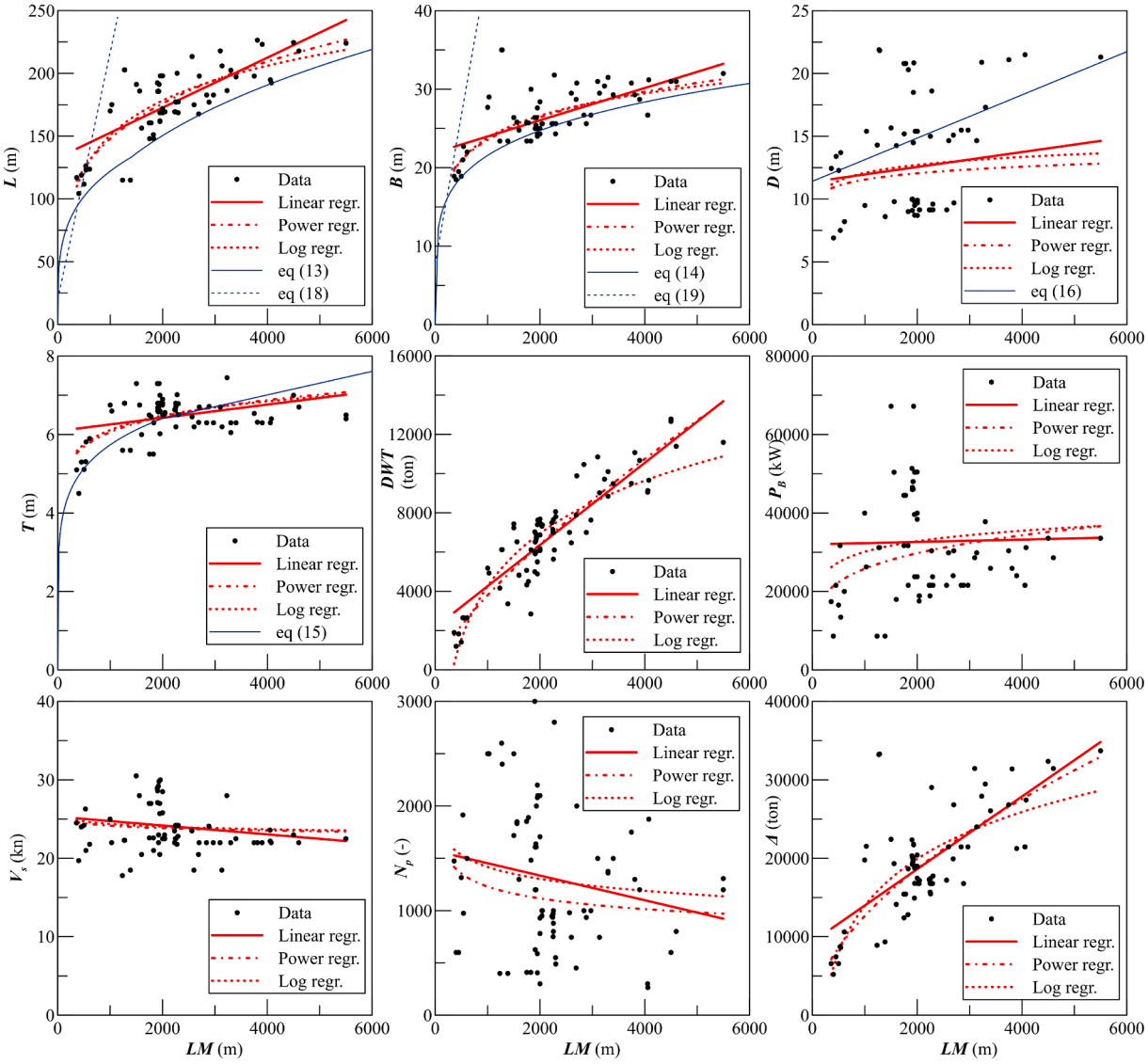


Fig. 25. Simple regressions on the present database as a function of LM.

Table 4
Quality of fit for the simple regressions as a function of LM.

	Linear					Power					Logarithmic				
	R^2	$MAPE$	$RMSE$	$RRMSE$	Prs	R^2	$MAPE$	$RMSE$	$RRMSE$	Prs	R^2	$MAPE$	$RMSE$	$RRMSE$	Prs
L	0.531	0.089	18.09	0.161	0.729	0.606	0.079	16.59	0.148	0.780	0.615	0.080	16.38	0.146	0.784
B	0.348	0.068	2.597	0.060	0.589	0.388	0.065	2.516	0.058	0.624	0.392	0.067	2.507	0.057	0.626
D	0.013	0.316	4.396	0.147	0.117	0.012	0.294	4.454	0.153	0.104	0.010	0.317	4.403	0.147	0.104
T	0.065	0.063	0.516	0.024	0.255	0.190	0.059	0.481	0.022	0.443	0.211	0.058	0.474	0.022	0.460
DWT	0.786	0.183	1.1E3	1.611	0.887	0.795	0.151	1.1E3	1.586	0.896	0.765	0.172	1.1E3	1.691	0.874
Δ	0.447	0.212	4.7E3	4.042	0.668	0.448	0.179	4.7E3	4.086	0.678	0.450	0.187	4.7E3	4.030	0.671
P_B	0.000	0.465	1.4E4	9.734	0.007	0.048	0.391	1.5E4	10.45	0.091	0.015	0.451	1.4E4	9.659	0.124
V_s	0.039	0.101	3.013	0.073	0.198	0.001	0.101	3.072	0.075	0.070	0.005	0.104	3.066	0.074	0.072
N_p	0.048	0.665	6.9E2	2.282	0.219	0.050	0.561	7.3E2	2.592	0.149	0.028	0.669	7.0E2	2.306	0.169

For the current database, simple regressions were performed using the linear, power, and logarithmic models. By analysing the results in Table 4, it is possible to determine the best model for each variable, with LM as the independent variable. A detailed analysis of all the results obtained for each variable is reported in Appendix A.

5.1.4. Regressions as a function of L

Finally, the same analysis of the previous variables has been carried out on the length L. The application of the linear model of Eq. (37) allows for obtaining the following formulae:

$$\Delta = 213.1201(L) - 18051.2622 \quad (131)$$

$$B = 0.0924(L) + 10.1930 \quad (132)$$

$$D = 0.0355(L) + 6.4223 \quad (133)$$

$$T = 0.0128(L) + 4.2212 \quad (134)$$

$$DWT = 76.9636(L) - 6846.2519 \quad (135)$$

$$P_B = 200.5077(L) - 2899.1137 \quad (136)$$

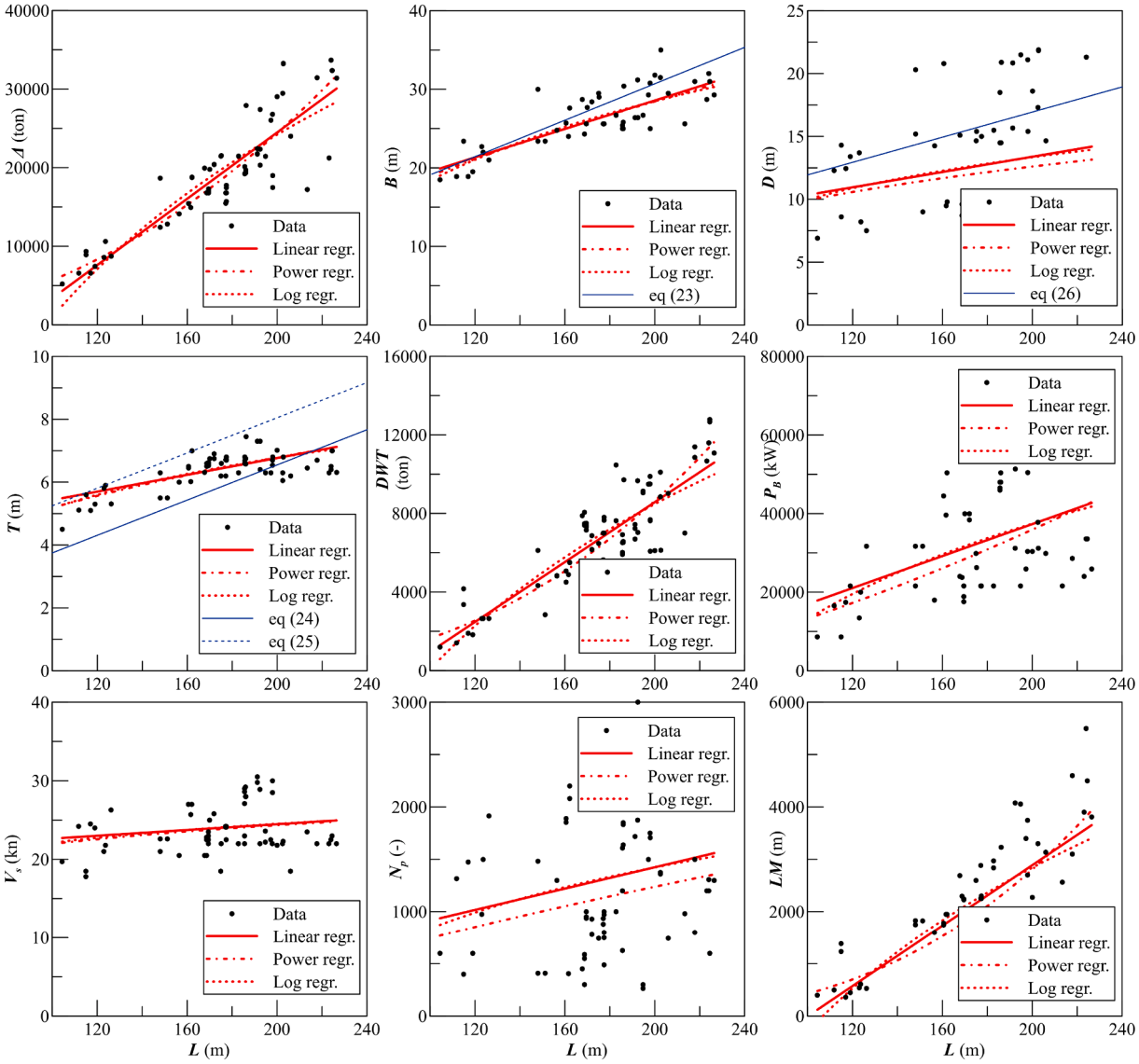


Fig. 26. Simple regressions on the present database as a function of L.

$$V_s = 0.0176(L) + 20.9076 \quad (137)$$

$$N_p = 5.3590(L) + 375.1592 \quad (138)$$

$$LM = 27.7879(L) - 2716.2716 \quad (139)$$

The application of a power model according to Eq. (40) leads to the following formulae:

$$\Delta = 0.3414(L)^{2.1106} \quad (140)$$

$$B = 1.2204(L)^{0.5951} \quad (141)$$

$$D = 1.2705(L)^{0.4355} \quad (142)$$

$$T = 0.9113(L)^{0.3793} \quad (143)$$

$$DWT = 0.0200(L)^{2.4493} \quad (144)$$

$$P_B = 14.5642(L)^{1.2732} \quad (145)$$

$$V_s = 11.0828(L)^{0.1482} \quad (146)$$

$$N_p = 28.5136(L)^{0.7120} \quad (147)$$

$$LM = 0.0028(L)^{2.6026} \quad (148)$$

Finally, by applying the logarithmic model of Eq. (41) the following formulae can be derived from the database:

$$\Delta = -156096.4172 + 34046.0794 \ln L \quad (149)$$

$$B = -50.9541 + 15.0134 \ln L \quad (150)$$

$$D = -17.9308 + 5.9340 \ln L \quad (151)$$

$$T = -5.1158 + 2.2469 \ln L \quad (152)$$

$$DWT = -56988.5017 + 12353.3481 \ln L \quad (153)$$

$$P_B = -149126.5784 + 35191.5620 \ln L \quad (154)$$

$$V_s = 5.9155 + 3.5061 \ln L \quad (155)$$

$$N_p = -3331.9941 + 901.7176 \ln L \quad (156)$$

$$LM = -20448.7100 + 4387.5977 \ln L \quad (157)$$

Fig. 26 shows the obtained regressions for the present database, together with a comparison of the already presented regressions of Kristensen (2016) obtained as a function of L. Analysing the Figure it is possible to observe that for the breadth B the regression of Kristensen (2016) is starting from the same level of the newly fitted regressions but the slope of the line is higher, resulting in an overestimation of the data for longer ships. Concerning the depth D, the regression of Kristensen (2016) is fitting well one of the subpopulation but is overestimating the global mean of the present database. Finally, for the draught T the study of Kristensen (2016) reports two curves, indicating the maximum and minimum draughts with two parallel lines. The new regressions stay

Table 5
Quality of fit for the simple regressions as a function of L.

	Linear					Power					Logarithmic				
	R^2	$MAPE$	$RMSE$	$RRMSE$	Prs	R^2	$MAPE$	$RMSE$	$RRMSE$	Prs	R^2	$MAPE$	$RMSE$	$RRMSE$	Prs
B	0.577	0.062	2.091	0.048	0.759	0.577	0.061	2.090	0.048	0.760	0.575	0.063	2.096	0.048	0.758
D	0.044	0.315	4.326	0.144	0.212	0.023	0.293	4.373	0.150	0.215	0.047	0.315	4.321	0.144	0.217
T	0.400	0.051	0.414	0.019	0.632	0.438	0.050	0.400	0.018	0.665	0.466	0.048	0.390	0.018	0.682
DWT	0.700	0.161	1.3E3	1.910	0.836	0.687	0.169	1.3E3	1.965	0.834	0.681	0.176	1.3E3	1.970	0.825
Δ	0.762	0.120	3.1E3	2.646	0.873	0.766	0.105	3.1E3	2.640	0.875	0.735	0.141	3.3E3	2.796	0.857
P_B	0.128	0.399	1.3E4	9.089	0.358	0.082	0.350	1.4E4	9.684	0.342	0.149	0.385	1.3E4	8.979	0.386
V_s	0.022	0.106	3.038	0.074	0.151	0.029	0.104	3.029	0.074	0.180	0.034	0.106	3.021	0.073	0.185
N_p	0.039	0.672	7.0E2	2.293	0.198	0.032	0.563	7.2E2	2.568	0.200	0.041	0.672	7.0E2	2.290	0.204
LM	0.531	0.265	6.8E2	1.733	0.729	0.548	0.277	6.7E2	1.738	0.746	0.500	0.275	7.1E2	1.790	0.707

between the two lines indicating that maximum and minimum draughts are containing all the draughts of the present database.

Analysing in detail the results reported in Table 5 it is possible to understand which one is the best model for each variable, for the case of L as the independent variable. A detailed analysis of the results obtained variable by variable is performed in Appendix A.

5.1.5. Best simple regressions

To summarise the results obtained for simple regression analysis, it is worth reporting which one are the best regressions obtained for each dependent variable. This allows a designer to clearly identify which equation is the most suitable to use in the prediction of main vessel general parameters. In the following, the best models identified for each variable are reported:

- *Length* L: the variable is described by regressions as a function of DWT, Δ and LM. For the regressions as a function of DWT, the best model is provided by Eq. (59), which is a power model. The logarithmic model of Eq. (95) is the best option for the regressions as a function of Δ. For the regressions as a function of LM, also the logarithmic model Eq. (122) identifies the best solution.
- *Breadth* B: the breadth is described by all the employed independent variables. For the regressions as a function of DWT, the best option is the logarithmic model of Eq. (69). For the regressions as a function of Δ, Eq. (78), the linear model, gives the best quality of fit. The logarithmic model of Eq. (123) is the best option for the regressions as a function of LM, while Eq. (135) (linear model) is the best for regressions as a function of L.
- *Depth* D: the depth D is represented by all the four independent variables employed in the analysis. For the regressions as a function of DWT, the best model is represented by Eq. (70), which means the logarithmic model. The linear model represents the best solution for the regressions as a function of Δ Eq. (79) and for the regressions as a function of LM Eq. (106). The logarithmic model, represented by Eq. (151),is the best option for the regressions as a function of LM.
- *Draught* T: for the draught T, models are avilable for all the considered independent variables. For the regressions as a function of DWT, the best model is the logarithmic one, represented by Eq. (71). The logarithmic model is the best solution also for all the remaining independent variables, which means it is advisable to use Eq. (98) for the regression as a function of Δ, Eq. (125) for the regressions as a function of LM, and Eq. (152) for the regressions as a function of L.
- *Displacement* Δ: the displacement Δ is represented by the models obtained as a function of DWT, LM and L. For the regressions as a function of DWT, the best solution is provided by the power model, represented by Eq. (63). For the regressions as a function of LM, the best option is the logarithmic model of Eq. (126), while the power model of Eq. (140) is the best solution for the regressions as a function of L.
- *Deadweigth* DWT: models for DWT are obtained as a function of Δ, LM and L. The logarithmic model of Eq. (96) is the best solution for the regressions as a function of Δ. For the regressions as a function of LM, the best option is provided by the power model of Eq. (117). The linear model of Eq. (135) is the best solution for the regressions as a function of L.
- *Installed power* P_B: the installed power P_B is modelled for all the four independent variables considered in the study. The logarithmic model is the best solution for the regressions as a function of DWT, Δ and L, represented by Eqs. (73), (100), and (154), respectively. The power model of Eq. (118), is the best solution for the regressions as a function of LM.
- *Vessel speed* V_s: the vessel speed V_s is described by models functions of all the four considered independent variables. The linear model is the best solution for the regressions as a function of DWT and LM, represented by Eqs. (56) and (110), respectively. The logarithmic model is the best option for the remaining two variables, resulting in Eq. (101) for the regressions as a function of Δ and Eq. (155) for the regressions as function of L.
- *Number of passengers* N_p: the number of passengers N_p is described by the models obtained as a function of all the four considered independent variables. The power model is the best option for regressions as a function of DWT and LM, represented by Eqs. (66) and (120), respectively. The logarithmic model represents the best solution for the regressions as a function of L, which means Eq. (156). For the regressions as a function of Δ, the best solution is the linear model, represented by Eq. (84).
- *Lane metres* LM: the lane metres LM is represented by the regressions as a function of DWT, Δ and L. For the regressions as a function of DWT, the best solution is the linear model of Eq. (58). For the regressions as a function of Δ, the best option is the logarithmic model of Eq. (103). Finally, for the regressions as a function of L, the best solution is provided by the power model of Eq. (148).

These regression formulae will be afterwards used for the verification process and comparison with other regression models in Section 6.

5.2. Multiple linear regressions

The present section presents the more relevant results of the multiple linear regression analysis on the RoPax database. Additional results, like the complete set of regression coefficient and tests on the heteroskedasticity of the regressions, are reported in Appendix B. All the regression have been performed by starting from a complete 4th order model, eliminating automatically the terms not significant for the global results in term of quality of fit of the regression. Such an approach allows for obtaining the regression with the minimum number of relevant terms to fit the selected variable.

The results are presented grouped by the regression types, analysing separately the following cases:

- Regressions as a function of V_s and DWT.
- Regressions as a function of V_s and LM.
- Regressions as a function of N_p and DWT.

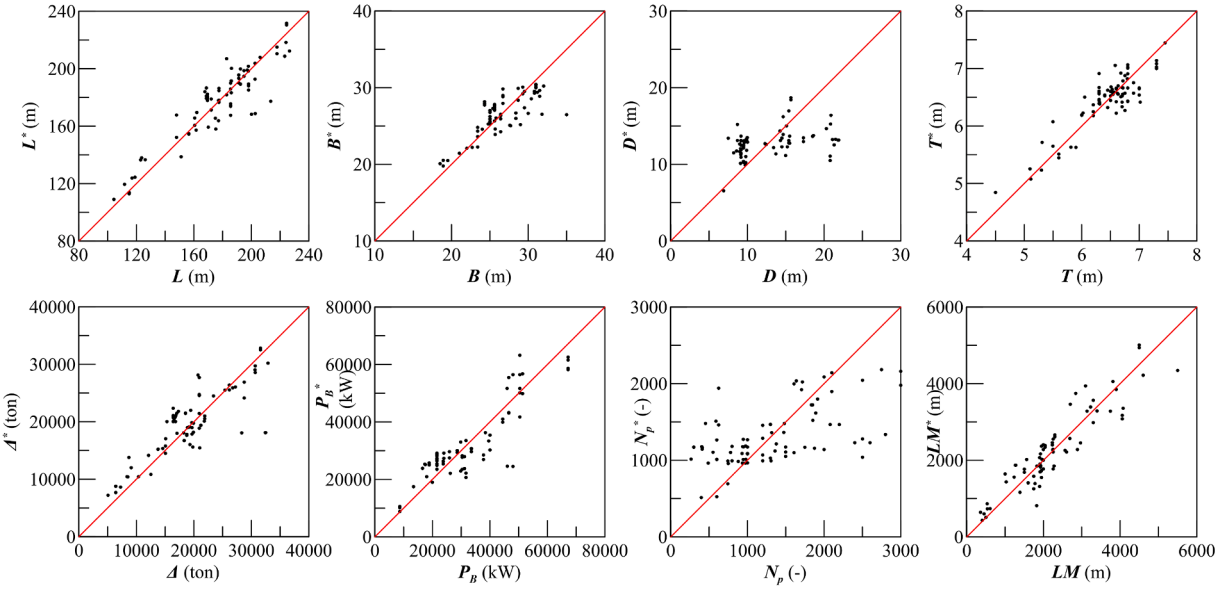


Fig. 27. Database values vs predicted values in multiple linear regressions as a function of V_s and DWT.

Table 6
Quality of fit of the regressions as a function of V_s and DWT.

	R^2	R^2_{adj}	$MAPE$	$RMSE$	$RRMSE$	Prs
L	0.805	0.779	0.047	11.66	0.104	0.897
B	0.456	0.440	0.064	2.372	0.054	0.675
D	0.209	0.106	0.254	3.936	0.131	0.457
T	0.845	0.728	0.032	0.266	0.012	0.867
Δ	0.611	0.600	0.158	4.0E3	3.388	0.782
P_B	0.843	0.806	0.178	5.8E3	3.853	0.918
N_p	0.406	0.268	0.477	5.5E2	1.802	0.637
LM	0.827	0.810	0.163	4.1E2	1.054	0.909

Table 7
Quality of fit of the regressions as a function of V_s and LM.

	R^2	R^2_{adj}	$MAPE$	$RMSE$	$RRMSE$	Prs
L	0.703	0.669	0.061	14.41	0.128	0.838
B	0.348	0.328	0.068	2.597	0.060	0.589
D	0.237	0.138	0.242	3.865	0.129	0.487
T	0.681	0.627	0.034	0.301	0.014	0.825
DWT	0.845	0.830	0.127	9.5E2	1.373	0.919
Δ	0.535	0.482	0.176	4.3E3	3.706	0.731
P_B	0.796	0.777	0.198	6.6E3	4.387	0.892
N_p	0.338	0.281	0.490	5.8E2	1.903	0.581

- Regressions as a function of N_p and LM.
- Regressions as a function of N_p, V_s and DWT.
- Regressions as a function of N_p, V_s and LM.

The following subsections give a complete overview of the obtained regressions and the obtained quality of fit indicators.

5.2.1. Regressions as a function of V_s and DWT

The first set of regressions has been performed taking into account V_s and DWT as the independent variable; therefore, regressions have been provided for L, B, D, T, Δ, P_B, N_p and LM. The regression results in terms of coefficients and associated t-stud and p-values is reported in Table B.3.

Fig. 27 shows the predicted values of the regressions against the original data for all the fitted regressions. In the picture, the spreading along the bisector plotted in red highlights the deviation of the predicted data from the database one; therefore, the more scattered is the diagram the worst is the quality of fit of the regression. As for the simple regressions in the previous section, the quality of fit has been evaluated according to the R^2, $MAPE$, $RMSE$, $RRMSE$ and Prs indicator. In addition, as it is usual for multiple linear regression analysis, the R^2_{adj} has been also evaluated for all the regressions.

Table 6 reports the quality of fit indicators for all the dependent variable regressions. A detailed analysis of the results is reported in Appendix A.

5.2.2. Regressions as a function of V_s and LM

An alternative to the previous set of regression is given by considering LM as an independent variable in place of DWT. By adopting such kind of initial set, regressions can be obtained for L, B, D, T, Δ, P_B, N_p and DWT. The regression results in terms of coefficients and associated t-stud and p-values is reported in Table B.4.

Fig. 28 shows the predicted values of the regressions against the original data for all the fitted regressions. As per the previous example, the picture is useful to roughly understand the quality of the regression from the spreading of the data set. The detailed analysis of the quality of fit is performed according to the R^2, R^2_{adj}, $MAPE$, $RMSE$, $RRMSE$ and Prs indicators.

Table 7 reports the obtained quality of fit indicators for all the considered dependent variables. A detailed analysis of the results, with considerations provided for each of the analised variables is reported in Appendix A.

5.2.3. Regressions as a function of N_p and DWT

Alternative set of independent variable is composed of N_p and DWT. By adopting this set of independent variable, regressions can be obtained for L, B, D, T, Δ, P_B, V_s and LM. The regression results in terms of coefficients and associated t-stud and p-values are reported in Table B.5.

Fig. 29 shows the predicted values of the regressions against the original data for the fitted regressions. Such a picture allows for a graphical recognition of the quality of fit of the regressions, by considering the scattering of the plotted datasets. The detailed analysis of the quality of fit is performed with the R^2, R^2_{adj}, $MAPE$, $RMSE$, $RRMSE$ and Prs indicators.

Table 8 reports the obtained quality of fit indicators for all the selected dependent variables. A detailed analysis of the variables obtained for each of the regressions is reported in Appendix A.

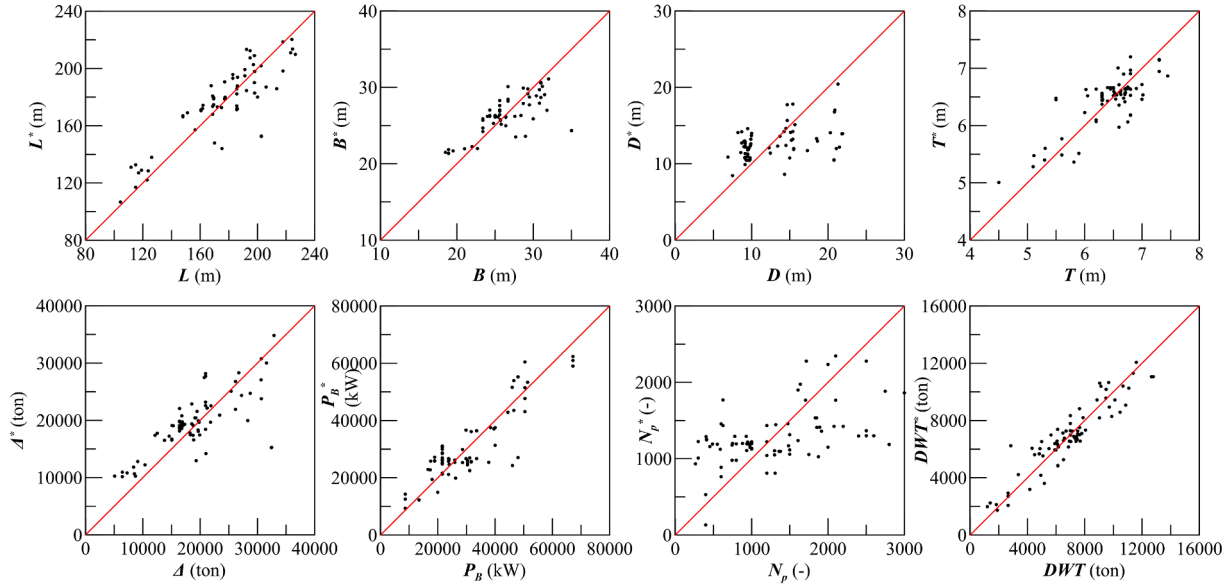


Fig. 28. Database values vs predicted values in multiple linear regressions as a function of V_s and LM.

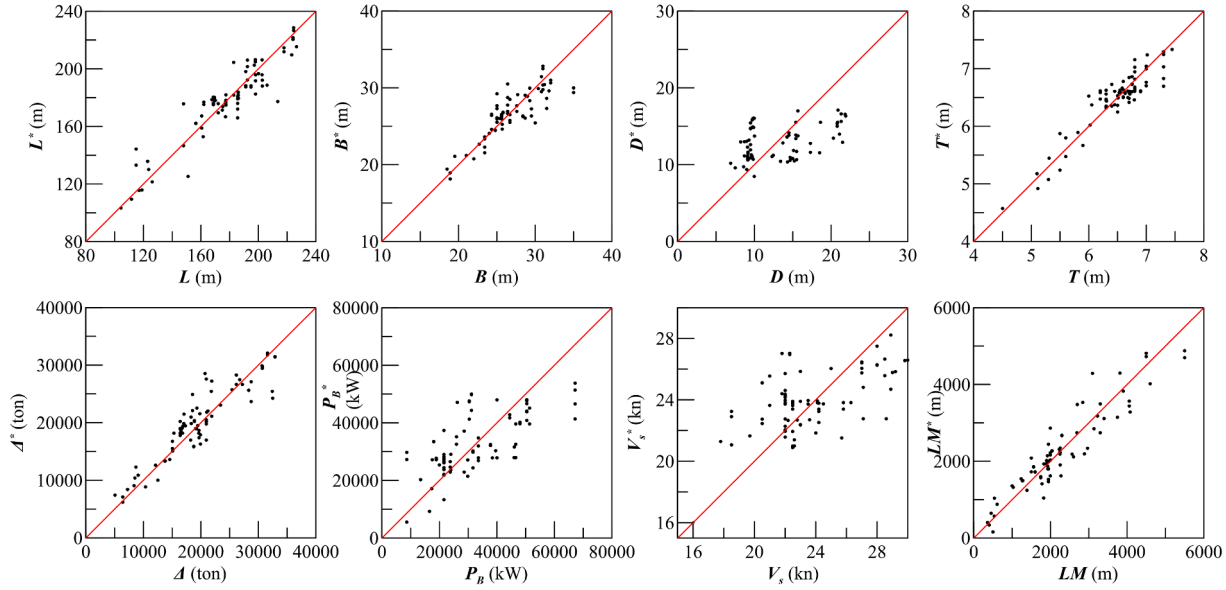


Fig. 29. Database values vs predicted values in multiple linear regressions as a function of N_p and DWT.

Table 8
Quality of fit of the regressions as a function of N_p and DWT.

	R^2	R^2_{adj}	$MAPE$	$RMSE$	$RRMSE$	Prs
L	0.849	0.818	0.045	10.24	0.091	0.921
B	0.731	0.680	0.045	1.668	0.038	0.855
D	0.187	0.137	0.275	3.991	0.133	0.432
T	0.829	0.796	0.026	0.221	0.010	0.910
Δ	0.815	0.780	0.112	2.7E3	2.338	0.902
P_B	0.548	0.462	0.294	9.9E3	6.541	0.740
V_s	0.271	0.226	0.090	2.623	0.063	0.521
LM	0.817	0.806	0.177	4.3E2	1.082	0.904

5.2.4. Regressions as a function of N_p and LM

As performed for the regression including V_s, also for N_p it is possible to switch the second independent variable from DWT to LM, obtaining this new set of multiple linear regressions. By adopting this new set of independent variables it is possible obtaining regressions for L, B, D, T, Δ, P_B, V_s and DWT. The regressions results in terms of coefficients and associated t-stud and p-values is reported in Table B.6.

Fig. 30 shows the predicted values of the regressions against the original data for all the fitted regressions. The figure is useful for a fast qualitative understanding of the quality of fit of the regression, by looking at the spreading of the fitted dataset. The deteiled analysis of the quality of fit is performed according to the R^2, R^2_{adj}, $MAPE$, $RMSE$, $RRMSE$ and Prs indicators.

Table 9 reports the obtained quality of fit indicators for all the considered variables. A detailed analysis of the results is reported in Appendix A.

5.2.5. Regressions as a function of N_p, V_s and DWT

The previous set of multiple linear regressions was a function of two independent variables not strictly correlated between each other. However, it is possible to increase also the number of not correlated independent variables, selecting besides V_s and N_p either DWT or LM. In this section the regressions as a function of V_s, N_p and DWT are presented. By adopting this set of independent variables it is possible obtaining regressions for L, B, D, T, Δ, P_B and LM. The regression results in

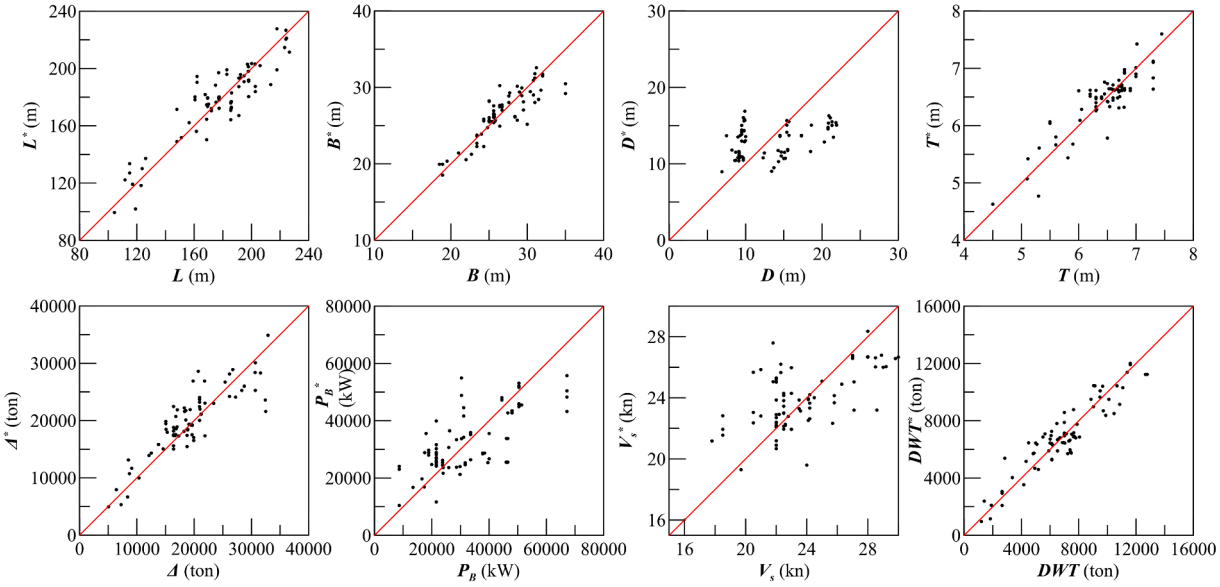


Fig. 30. Database values vs predicted values in multiple linear regressions as a function of N_p and LM.

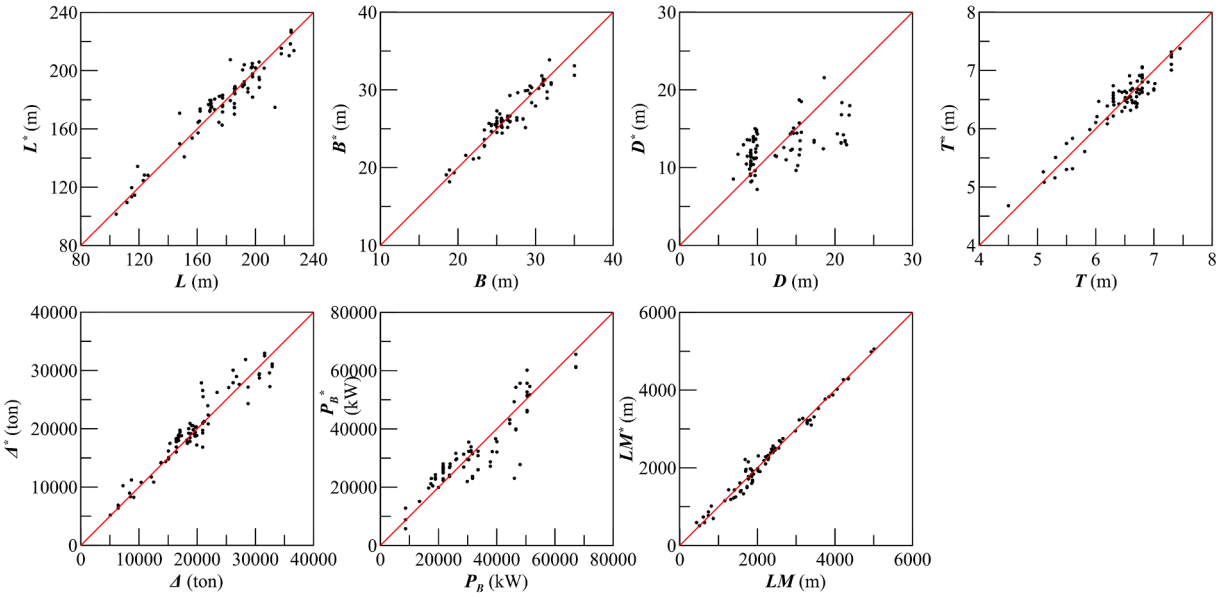


Fig. 31. Database values vs predicted values in multiple linear regressions as a function ofN_p, V_s and DWT.

Table 9
Quality of fit of the regressions as a function of N_p and LM.

	R^2	R^2_{adj}	$MAPE$	$RMSE$	$RRMSE$	Prs
L	0.807	0.785	0.054	11.61	0.103	0.898
B	0.732	0.687	0.046	1.662	0.038	0.856
D	0.229	0.169	0.272	3.885	0.129	0.479
T	0.764	0.718	0.030	0.259	0.012	0.874
DWT	0.838	0.823	0.136	9.7E2	1.400	0.915
Δ	0.725	0.699	0.137	3.3E3	2.846	0.852
P_B	0.525	0.453	0.291	1.0E4	6.709	0.724
V_s	0.431	0.356	0.075	2.318	0.056	0.656

Table 10
Quality of fit of the regressions as a function of N_p, V_s and DWT.

	R^2	R^2_{adj}	$MAPE$	$RMSE$	$RRMSE$	Prs
L	0.908	0.865	0.032	8.005	0.071	0.953
B	0.836	0.802	0.037	1.301	0.030	0.914
D	0.473	0.314	0.205	3.213	0.107	0.688
T	0.877	0.840	0.023	0.187	0.009	0.936
Δ	0.883	0.861	0.079	2.2E3	1.857	0.939
P_B	0.842	0.818	0.167	5.8E3	3.866	0.917
LM	0.856	0.826	0.130	3.8E2	0.960	0.925

terms of coefficients and associated t-stud and p-values are reported in Tables B.7 and B.8.

Fig. 31 shows the predicted values of the regressions against the original data for all the fitted regression in order to have a global overview of the quality of fit of the obtained models. The detailed analysis of the quality of fit is performed according to the R^2, R^2_{adj}, $MAPE$, $RMSE$, $RRMSE$ and Prs indicators.

Table 10 reports the obtained quality of fit indicators for all the considered variables. A detailed variable by variable analysis of the obtained regressions is provided in Appendix A.

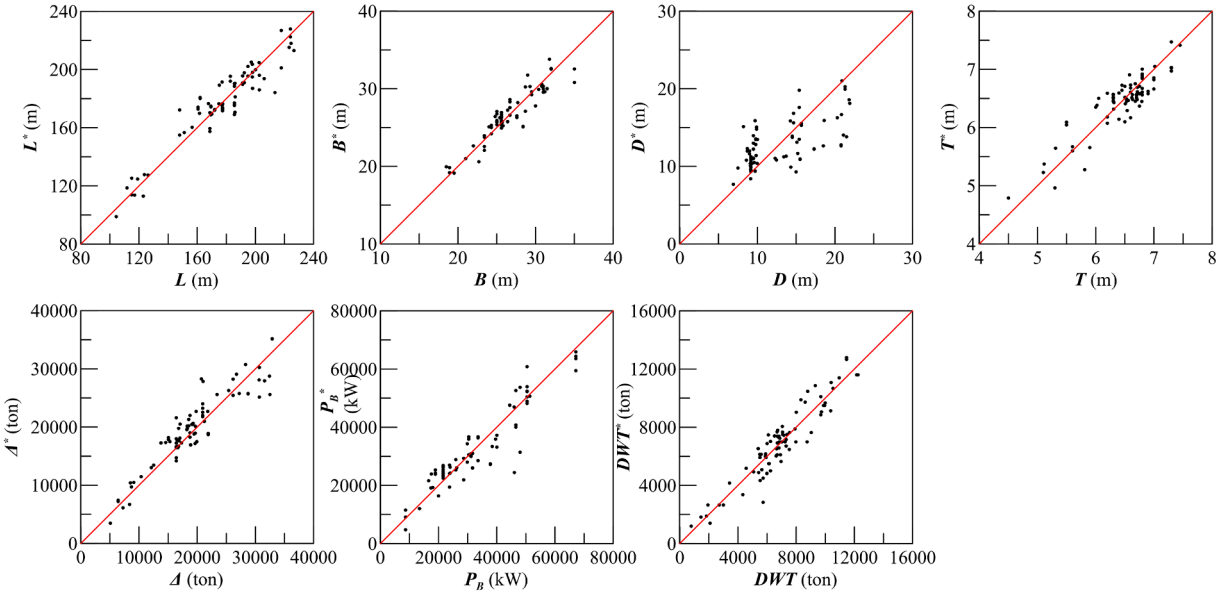


Fig. 32. Database values vs predicted values in multiple linear regressions as a function of N_p, V_s and LM.

Table 11
Quality of fit of the regressions as a function of N_p, V_s and LM.

	R^2	R^2_{adj}	$MAPE$	$RMSE$	$RRMSE$	Prs
L	0.900	0.875	0.035	8.343	0.074	0.948
B	0.874	0.823	0.028	1.139	0.026	0.935
D	0.410	0.285	0.219	3.400	0.113	0.640
T	0.867	0.809	0.022	0.194	0.009	0.931
DWT	0.870	0.843	0.123	8.7E2	1.254	0.933
Δ	0.847	0.808	0.106	2.5E3	2.126	0.920
P_B	0.910	0.865	0.104	4.4E3	2.920	0.953

5.2.6. Regressions as a function of N_p, V_s and LM

As for the case of two independent variables, it is possible to switch between LM and DWT as additional variable to V_s and N_p. Here the case of LM is analysed, resulting in the reproduction of regression models for L, B, D, T, Δ, P_B and DWT. The regression results in term of coefficients, and associated t-stud and p-values are reported in Tables B.9 and B.10.

Fig. 32 shows the predicted values of the regressions against the original data for all the fitted regressions. As for the previous cases, the plot allows for checking the quality of the regression according to the spreading of the dataset. The detailed analysis of the quality of fit is performed according to the R^2, R^2_{adj}, $MAPE$, $RMSE$, $RRMSE$ and Prs indicators.

Table 11 reports the obtained quality of fit indicators for all the considered variables. Appendix A reports a detailed variable by variable analysis of the results.

5.3. Forest tree

Apart from multiple linear regressions, forest tree regressions are an advanced technique well-suited to investigating how the main dimensions of RoPax ships depend on one or more parameters. The forest tree algorithm allows the classification of the output through the averaged prediction of more individual trees (Ho, 1998), thus reducing the overfitting problem of individual trees. Here, the MATLAB application for the determination of forest tree is applied to the database, providing regression for the quantities of interest.

The ensemble aggregation method employed in the calculation uses a least squares boosting, with a maximum number of 100 learning cycles. The hyperparameters are automatically optimised by the algorithm, by searching for the best tree objects increasing the fitting quality at each learning cycle.

The analysis has been performed for the same conditions of the multiple linear regression analysis, thus employing the same combinations of independent variables described in the previous sections. In the following, the detailed vision of the obtained forest tree is given, keeping in mind that the method is a black-box method, therefore no explicit formula is available for the estimation of the design parameters.

5.3.1. Forest trees as a function of V_s and DWT

Taking into consideration V_s and DWT as independent variables for the generation of the forest tree, it is possible to generate distinct trees for L, B, D, T, Δ, P_B, N_p and LM. The quality of fit for the tree models is evaluated according to the R^2, $MAPE$, $RMSE$, $RRMSE$ and Prs indicators, remembering that Prs and R^2 are quantities to maximise and the remaining indexes to minimise to have a good quality of fit.

Fig. 33 shows the predicted variables against the original data for all the fitted trees, plotting also the respective predictions according to the multiple linear regressions (MLR in the picture) described in the previous section. The picture allows for having a direct comparison between the spreading of the data obtained with the forest tree and the MLR. It is immediately evident that the data plotted by the forest tree have less scattering than the predictions using MLR.

However for the effective evaluation of the quality of fit all the indices have been calculated and the results are reported in Table 12. It has to be observed that for all the obtained models, all the indicators state that the forest tree have a better quality of fit than the respective MLR models. A detailed variable by variable analysis is performed in Appendix A.

5.3.2. Forest trees as a function of V_s and LM

Taking into consideration V_s and LM as independent variables for the generation of the forest tree, it is possible to generate distinct trees for L, B, D, T, Δ, P_B, N_p and DWT. The quality of fit for the tree models is evaluated according to the R^2, $MAPE$, $RMSE$, $RRMSE$ and Prs indicators, remembering that Prs and R^2 are quantities to maximise and the remaining indexes to minimise to have a good quality of fit.

Fig. 34 shows the predicted variables against the original data for all the fitted trees, plotting also the respective predictions according to the multiple linear regressions (MLR in the picture) described in the previous section. The picture allows for having a direct comparison between the spreading of the data obtained with the forest tree and the MLR. It

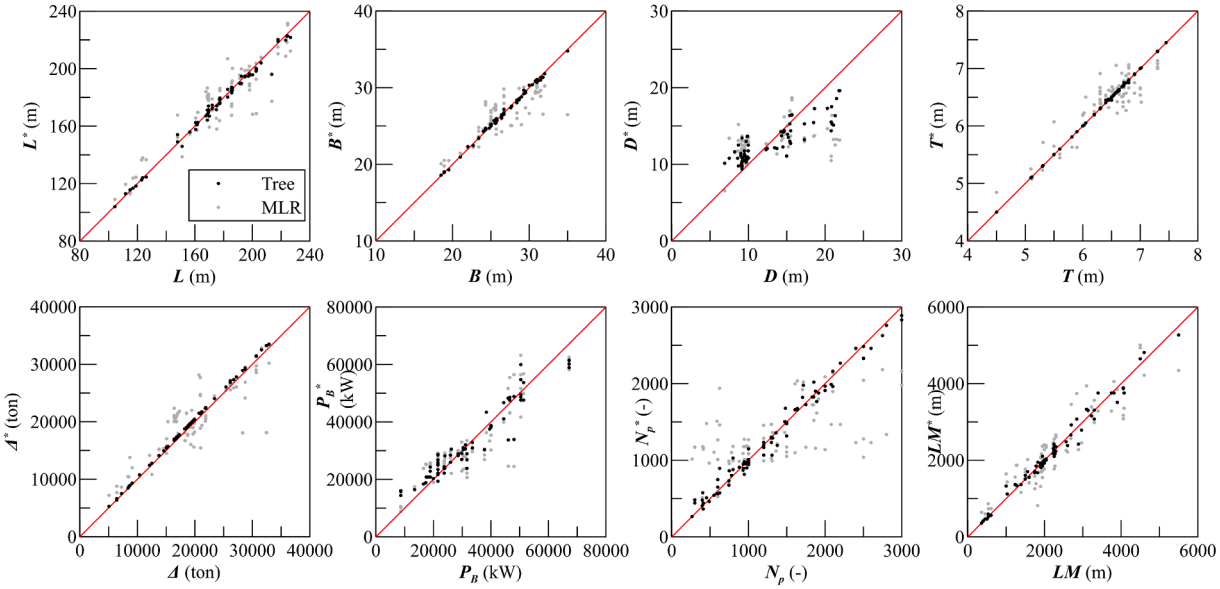


Fig. 33. Database values vs predicted values in forest trees as a function of V_s and DWT.

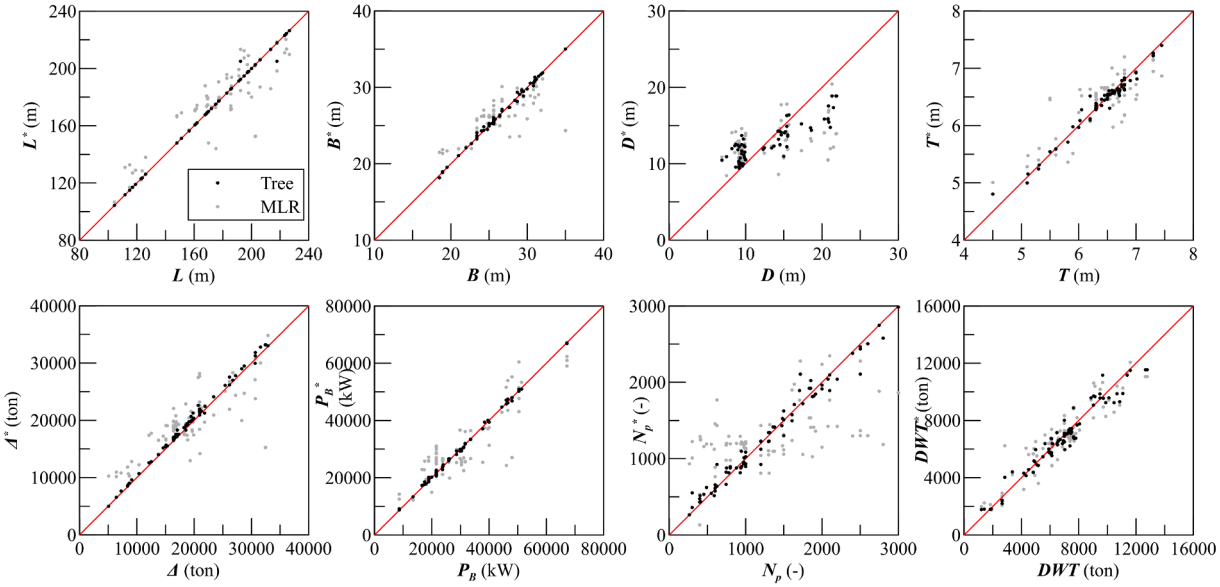


Fig. 34. Database values vs predicted values in forest trees as a function of V_s and LM.

Table 12
Quality of fit of the forest trees as a function of V_s and DWT.

	R^2	$MAPE$	$RMSE$	$RRMSE$	Prs
L	0.993	0.005	2.578	0.023	0.996
B	0.978	0.009	0.518	0.012	0.989
D	0.616	0.161	2.615	0.088	0.816
T	0.938	0.008	0.147	0.007	0.970
Δ	0.988	0.016	7.8E2	0.665	0.994
P_B	0.959	0.044	3.1E3	2.047	0.980
N_p	0.857	0.081	2.4E2	0.813	0.932
LM	0.970	0.033	1.9E2	0.503	0.985

Table 13
Quality of fit of the forest trees as a function of V_s and LM.

	R^2	$MAPE$	$RMSE$	$RRMSE$	Prs
L	0.963	0.023	5.814	0.053	0.987
B	0.941	0.008	0.847	0.019	0.971
D	0.623	0.163	2.592	0.087	0.818
T	0.936	0.010	0.149	0.007	0.968
DWT	0.978	0.038	3.9E2	0.581	0.989
Δ	0.919	0.053	2.0E3	1.743	0.959
P_B	0.963	0.050	2.9E3	1.938	0.982
N_p	0.818	0.116	2.7E2	0.924	0.909

is immediately evident that, as for the previous case, the data plotted by the forest tree have less scattering than the predictions using MLR.

The quality of fit analysis is reported in Table 13, where all the indicators are reported for each obtained forest tree model. As a final remark, it can be observed that the obtained indicators for the quality of fit are all indicating that the forest tree fit better than MLR all the analysed variables. Appendix A reports a variable by variable analysis of the obtained results.

5.3.3. Forest trees as a function of N_p and DWT

Taking into consideration N_p and DWT as independent variables for the generation of the forest tree, it is possible to generate distinct

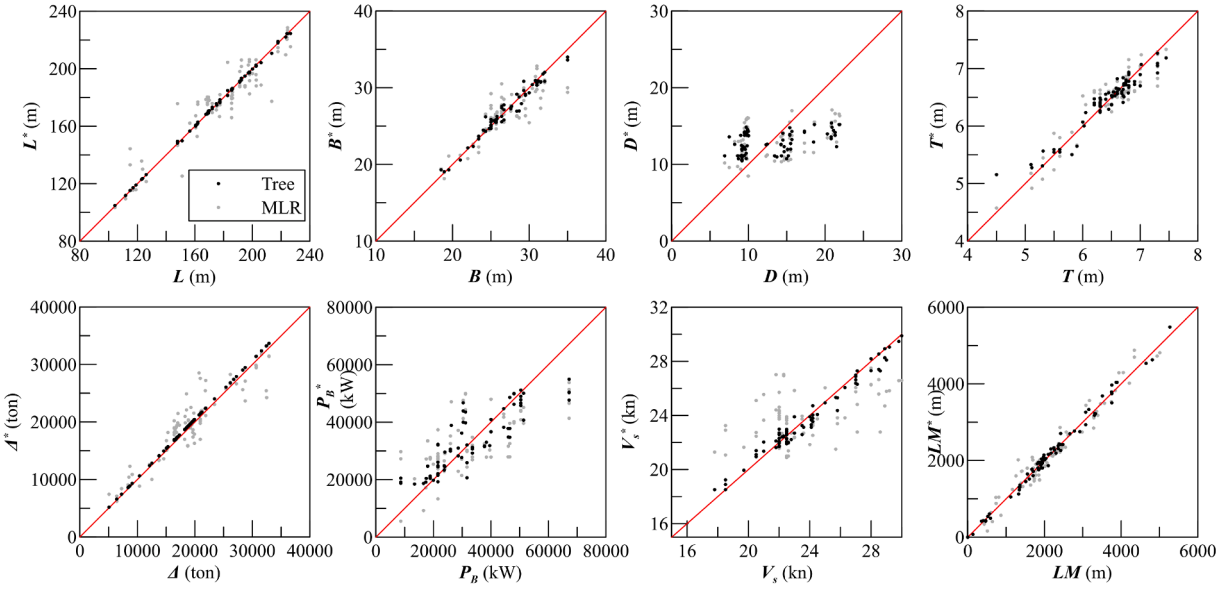


Fig. 35. Database values vs predicted values in forest trees as a function of N_p and DWT.

Table 14
Quality of fit of the forest trees as a function of N_p and DWT.

	R^2	$MAPE$	$RMSE$	$RRMSE$	Prs
L	0.992	0.005	2.768	0.025	0.996
B	0.982	0.011	0.472	0.011	0.991
D	0.278	0.257	3.585	0.120	0.553
T	0.953	0.006	0.127	0.006	0.977
Δ	0.988	0.027	7.9E2	0.677	0.994
P_B	0.854	0.178	5.8E3	3.849	0.937
V_s	0.721	1.052	1.633	0.040	0.862
LM	0.990	0.018	1.1E2	0.295	0.995

Table 15
Quality of fit of the forest trees as a function of N_p and LM.

	R^2	$MAPE$	$RMSE$	$RRMSE$	Prs
L	0.976	0.009	4.724	0.043	0.988
B	0.916	0.019	1.007	0.023	0.958
D	0.340	0.230	3.427	0.115	0.597
T	0.930	0.008	0.156	0.007	0.967
DWT	0.966	0.058	4.9E2	0.723	0.986
Δ	0.974	0.028	1.1E3	0.988	0.987
P_B	0.867	0.157	5.5E3	3.659	0.942
V_s	0.751	0.049	1.542	0.037	0.875

trees for L, B, D, T, Δ, P_B, V_s and LM. The quality of fit for the tree models is evaluated according to the R^2, $MAPE$, $RMSE$, $RRMSE$ and Prs indicators, remembering that Prs and R^2 are quantities to maximise and the remaining indexes to minimise to have a good quality of fit.

Fig. 35 shows the predicted variables against the original data for all the fitted trees, plotting also the respective predictions according to the multiple linear regressions (MLR in the picture) described in the previous section. The picture allows for having a direct comparison between the spreading of the data obtained with the forest tree and the MLR. It is immediately evident that, as for the previous case, the data plotted by the forest tree have less scattering than the predictions using MLR.

The quality of fit analysis is reported in Table 14, where all the indicators are reported for each obtained forest tree model. As a final remark, comparing the quality of fit indicators, the forest tree fits better than the MLR all the variables, except for the depth D where the quality of fit is similar between the different models. A detailed analysis of the obtained results is presented in Appendix A.

5.3.4. Forest trees as a function of N_p and LM

Taking into consideration N_p and LM as independent variables for the generation of the forest tree, it is possible to generate distinct trees for L, B, D, T, Δ, P_B, V_s and DWT. The quality of fit for the tree models is evaluated according to the R^2, $MAPE$, $RMSE$, $RRMSE$ and Prs indicators, remembering that Prs and R^2 are quantities to maximise and the remaining indexes to minimise to have a good quality of fit.

Fig. 36 shows the predicted variables against the original data for all the fitted trees, plotting also the respective predictions according to the multiple linear regressions (MLR in the picture) described in the previous section. The picture allows for having a direct comparison between the spreading of the data obtained with the forest tree and the MLR. It is immediately evident that, as for the previous case, the data plotted by the forest tree have less scattering than the predictions using MLR.

Table 15 reports the quality of fit indicators for all the developed forest tree models. As a final remark, comparing the quality of fit indicators, the forest tree fits better than the MLR all the variables, except for the depth D where the quality of fit is similar between the different models. Appendix A reports a detailed variable by variable analysis of the results.

5.3.5. Forest trees as a function of N_p, V_s and DWT

Taking into consideration V_s, N_p and DWT as independent variables for the generation of the forest tree, it is possible to generate distinct trees for L, B, D, T, Δ, P_B and LM. The quality of fit for the tree models is evaluated according to the R^2, $MAPE$, $RMSE$, $RRMSE$ and Prs indicators, remembering that Prs and R^2 are quantities to maximise and the remaining indexes to minimise to have a good quality of fit.

Fig. 37 shows the predicted variables against the original data for all the fitted trees, plotting also the respective predictions according to the multiple linear regressions (MLR in the picture) described in the previous section. The picture allows for having a direct comparison between the spreading of the data obtained with the forest tree and the MLR. It is immediately evident that, as for the previous case, the data plotted by the forest tree have less scattering than the predictions using MLR.

Table 16 reports the quality of fit indicators for all the developed forest tree models. Analysing the quality of fit indicators and comparing

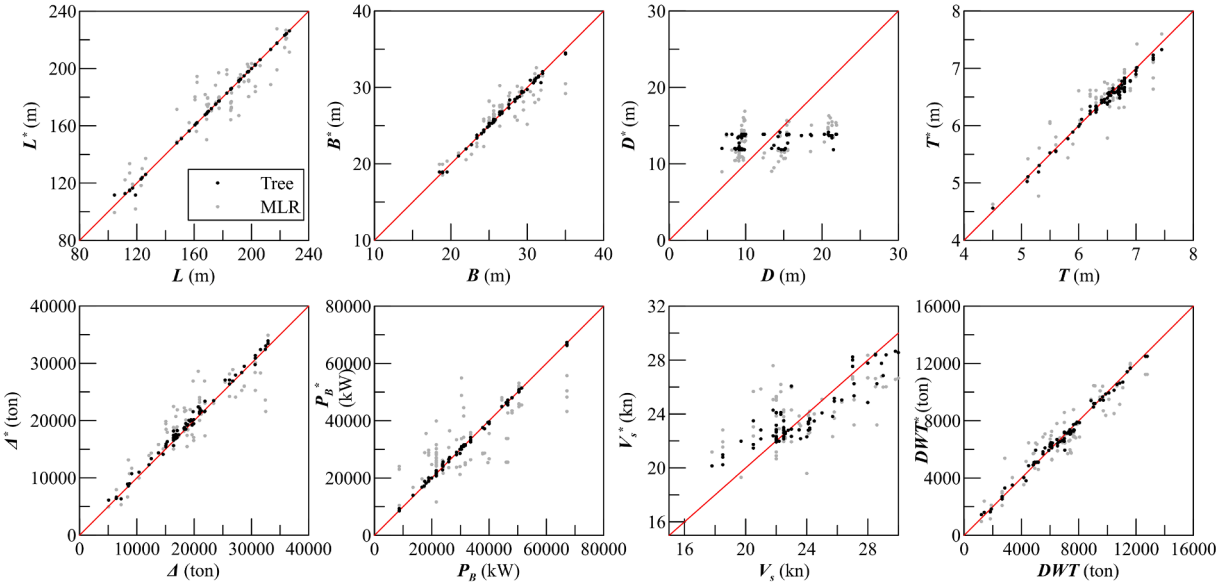


Fig. 36. Database values vs predicted values in forest trees as a function of N_p and LM.

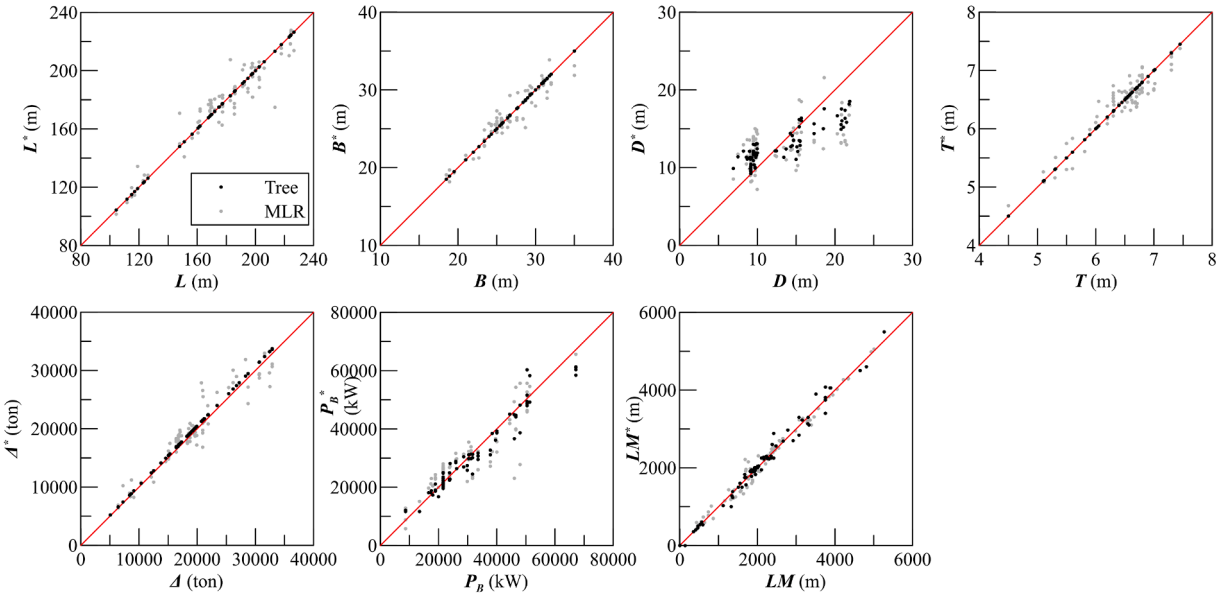


Fig. 37. Database values vs predicted values in forest trees as a function of N_p, V_s and DWT.

Table 16
Quality of fit of the forest trees as a function of N_p, V_s and DWT.

	R^2	$MAPE$	$RMSE$	$RRMSE$	Prs
L	0.991	0.008	2.805	0.025	0.996
B	0.969	0.008	0.610	0.014	0.985
D	0.563	0.181	2.789	0.094	0.789
T	0.971	0.007	0.099	0.005	0.986
Δ	0.989	0.016	7.4E2	0.640	0.995
P_B	0.951	0.035	3.4E3	2.241	0.976
LM	0.990	0.025	1.1E2	0.298	0.995

it with the corresponding cases of the MLR, it can be stated that for all the variables the forest tree models fit the data better than the MLR ones. A detailed analysis of the results is reported in Appendix A.

5.3.6. Forest trees as a function of N_p, V_s and LM

Taking into consideration V_s, N_p and LM as independent variables for the generation of the forest tree, it is possible to generate distinct trees for L, B, D, T, Δ, P_B and DWT. The quality of fit for the tree models is evaluated according to the R^2, $MAPE$, $RMSE$, $RRMSE$ and Prs indicators, remembering that Prs and R^2 are quantities to maximise and the remaining indexes to minimise to have a good quality of fit.

Fig. 38 shows the predicted variables against the original data for all the fitted trees, plotting also the respective predictions according to the multiple linear regressions (MLR in the picture) described in the previous section. The picture allows for having a direct comparison between the spreading of the data obtained with the forest tree and the MLR. It is immediately evident that, as for the previous case, the data plotted by the forest tree have less scattering than the predictions using MLR.

Table 17 reports the quality of fit indicators for all the developed forest tree models. From the analysis of the quality of fit indicators, it can be concluded that the forest tree models have a better fitting of the data compared to the corresponding MLR except for the depth D. For D MLR and forest tree present similar values and in both cases the regressions are not significant. Appendix A provides a detailed variable by variable analysis of the obtained results.

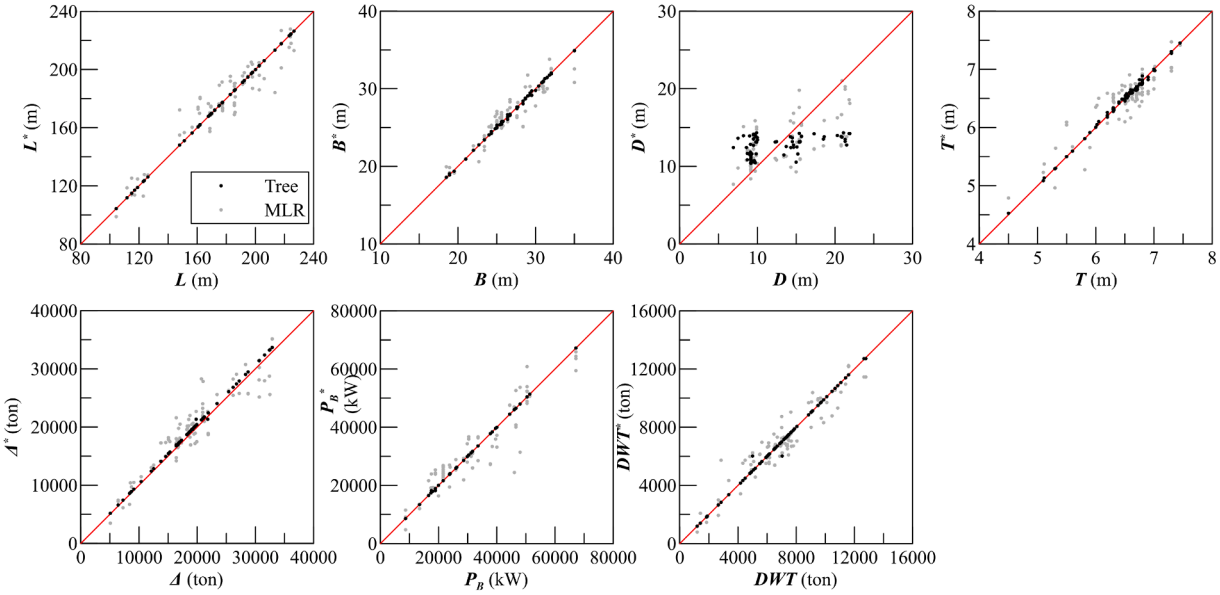


Fig. 38. Database values vs predicted values in multiple linear regressions as a function of N_p, V_s and LM.

Table 17
Quality of fit of the forest trees as a function of N_p, V_s and LM.

	R^2	$MAPE$	$RMSE$	$RRMSE$	Prs
L	0.994	0.004	2.292	0.021	0.997
B	0.954	0.013	0.743	0.017	0.977
D	0.557	0.179	2.808	0.094	0.775
T	0.834	0.022	0.240	0.011	0.915
DWT	0.973	0.051	4.4E2	0.641	0.988
Δ	0.959	0.048	1.4E3	1.243	0.980
P_B	0.958	0.034	3.1E3	2.074	0.980

6. Models verification and remarks

In the previous section, the simple regression, the multiple linear regression and the forest tree analysis have been presented with reference to the training set employed for their development. Such an analysis allows for establishing the intrinsic quality of fit of the models but is not giving an indication of the possible generalisation of the model to different ships. To this end, the obtained models will be here applied to the test set held out from the initial database, which means considering ships not used in the development of regression models. Evaluating the quality of fit of the models on this dataset may allow for a wider comprehension of the significance of the obtained models and the establishment of a ranking among the different formulations. Afterwards, some concluding remarks will be drawn, concerning the obtained results on the test set and the general behaviour of the prediction models.

6.1. Model verification

To verify the possible application of the models derived in Section 5 to a general set of new vessels, it is handy to verify the quality of fit of the regressions on a dataset external from the training set. Therefore, this section presents the application of the described models to the test dataset introduced in Section 4.

To judge the quality of fit on the test set, the same strategy explained for the training set is applied, making use of the performance indicators R^2, $MAPE$, $RMSE$, $RRMSE$ and Prs. The results are presented per independent variable, analysing which is the best model among the different proposed solutions. To facilitate the representation of the results, the models obtained for the multiple linear regressions and forest tree models have been renamed as follows:

- Models as a function of V_s and DWT: the MLR model is identified by *mlr-0* and the forest tree by *ft-0*.
- Models as function of V_s and LM: the MLR model is identified by *mlr-1* and the forest tree by *ft-1*.
- Models as a function of N_p and DWT: the MLR model is identified by *mlr-2* and the forest tree by *ft-2*.
- Models as a function of N_p and LM: the MLR model is identified by *mlr-3* and the forest tree by *ft-3*.
- Models as a function of N_p, V_s and DWT: the MLR model is identified by *mlr-4* and the forest tree by *ft-4*.
- Models as a function of N_p, V_s and LM: the MLR model is identified by *mlr-5* and the forest tree by *ft-5*.

For the simple regressions, only the best model for each of the four categories (given DWT, given Δ, given LM and given L) is analysed, according to the ranking provided by the quality of fit study on the training set.

6.2. Length L

All multiple linear regressions and forest tree models describe the length. For the simple regressions, only the models as a function of DWT, Δ and LM are present. In the specific, according to the study on the training set, the power model is selected for the function of DWT, the logarithmic for the function of Δ and LM. The results of quality of fit obtained for the test set is reported in Table 18.

From a first observation of the obtained quality of fit indicators, it can be stated that the regression models are significant for the independent variable L, except for the forest tree models *ft-0*,*ft-2* and *ft-4*. For the three cited models, the quality of fit values highlight a drastic decrease compared to the ones obtained using the training set, suggesting that the three provided models are not predicting well L for a general set of ships. The quality of fit values obtained for the other models differ from the ones obtained with the training set but remains significant for the selected variable.

In conclusion, for the estimation of L the models giving the higher quality of fit are the forest tree *ft-1*, *ft-3* and *ft-5*. However, for estimating L, also the conventional simple regressions could provide a significant estimate as well as the multiple linear regressions. Therefore, according to the dependent variables available by designers in the early design stage, the corresponding provided simple and multiple regression models can be employed for a new Ro-Pax vessel.

Table 18
Quality of fit on the test set for the L prediction models.

Regression		R^2	$MAPE$	$RMSE$	$RRMSE$	Prs
Simple	Given DWT	0.813	0.070	14.03	0.267	0.912
	Given Δ	0.878	0.053	12.07	0.217	0.938
	Given LM	0.756	0.079	16.02	0.304	0.874
MLR	mlr-0	0.889	0.059	11.15	0.206	0.947
	mlr-1	0.889	0.058	11.17	0.204	0.956
	mlr-2	0.864	0.058	12.31	0.228	0.932
	mlr-3	0.866	0.064	12.27	0.226	0.932
	mlr-4	0.765	0.082	16.23	0.297	0.894
	mlr-5	0.829	0.072	13.85	0.256	0.925
Forest tree	ft-0	0.631	0.018	9.911	0.177	0.812
	ft-1	0.912	0.021	4.843	0.087	0.963
	ft-2	0.505	0.022	11.48	0.205	0.746
	ft-3	0.914	0.009	4.779	0.086	0.960
	ft-4	0.538	0.027	11.08	0.198	0.760
	ft-5	0.916	0.013	4.722	0.085	0.958

Table 19
Quality of fit on the test set for the B prediction models.

Regression		R^2	$MAPE$	$RMSE$	$RRMSE$	Prs
Simple	Given DWT	0.728	0.067	1.896	0.089	0.883
	Given Δ	0.838	0.047	1.461	0.070	0.917
	Given L	0.568	0.077	3.944	0.114	0.764
	Given LM	0.771	0.062	2.404	0.083	0.904
MLR	mlr-0	0.699	0.068	1.990	0.094	0.888
	mlr-1	0.651	0.075	2.145	0.101	0.868
	mlr-2	0.670	0.058	2.086	0.099	0.821
	mlr-3	0.738	0.050	1.858	0.089	0.862
	mlr-4	0.788	0.058	1.671	0.080	0.898
	mlr-5	0.707	0.062	1.966	0.095	0.853
Forest tree	ft-0	0.767	0.023	1.167	0.053	0.885
	ft-1	0.541	0.025	1.637	0.075	0.754
	ft-2	0.826	0.018	1.008	0.047	0.917
	ft-3	0.699	0.018	1.327	0.062	0.856
	ft-4	0.909	0.010	0.731	0.034	0.954
	ft-5	0.640	0.025	1.452	0.068	0.814

6.3. Breadth B

All the simple, multiple linear regressions and forest tree models describe the breadth. In the specific, according to the study on the training set, the logarithmic model is selected for the function of DWT and LM, the linear for the function of Δ and L. The results of quality of fit obtained for the test set is reported in Table 19.

Taking a look to the quality of fit indicators, it can be stated that all the models proposed are moderate significant or significant. For the simple regressions, the models as a function of DWT, Δ and LM are significant, while the one as a function of LM is moderate significant. The multiple linear regression models are not giving a particular improvement compared to simple regressions; however, *mlr-0*, *mlr-1* and *mlr-2* are moderate significant, while the other are significant.

Different is the case of forest tree models. Here, the quality of fit values indicate that, except for model *ft-1*, all the models are significant. However, comparing the quality of fit values on the training set and the test set, also in this case there is a drastic decrease of the quality of fit for the forest tree models.

In conclusion, also for the prediction of B, a designer can use even the simple models proposed to achieve an improvement compared to available literature data. However, the best quality of fit is provided by the forest tree models *ft-2* and *ft-4*, which are giving a consistent improvement in applicability compared to simple and multiple linear regression models.

6.4. Depth D

All the simple, multiple linear regressions and forest tree models describe the depth. For the simple regression models, according to the study on the training set, the logarithmic model is selected for the function of DWT and L, the linear for the function of Δ and LM. The results of quality of fit obtained for the test set is reported in Table 20.

Table 20
Quality of fit on the test set for the D prediction models.

Regression		R^2	$MAPE$	$RMSE$	$RRMSE$	Prs
Simple	Given DWT	0.012	0.305	3.785	0.268	0.139
	Given Δ	0.008	0.306	3.716	0.273	0.247
	Given L	0.008	0.303	3.854	0.277	0.099
	Given LM	0.002	0.300	3.621	0.263	0.251
MLR	mlr-0	0.003	0.287	4.174	0.281	0.158
	mlr-1	0.002	0.263	4.387	0.304	0.319
	mlr-2	0.040	0.265	4.040	0.275	0.186
	mlr-3	0.074	0.267	3.810	0.261	0.330
	mlr-4	0.001	0.469	6.372	0.431	0.133
	mlr-5	0.068	0.264	3.822	0.266	0.380
Forest tree	ft-0	0.001	0.325	4.867	0.330	0.021
	ft-1	0.137	0.215	4.362	0.308	0.542
	ft-2	0.001	0.334	5.089	0.343	0.022
	ft-3	0.001	0.319	4.819	0.326	0.103
	ft-4	0.001	0.318	4.782	0.324	0.001
	ft-5	0.001	0.309	4.859	0.326	0.143

Looking at the values reported in the table, none of the models is significant for this variable. R^2 and Prs indicators are extremely low, while the remaining indicators are quite high. It has to be observed that, for all models, the quality of fit indicators are worst than the values obtained for the training set and presented in the previous section.

In conclusion, in case a designer should choose an option to estimate D, the best model is given by *ft-1*. However, also for this model, the regression cannot be considered as significant for the given variable.

6.5. Draught T

All the simple, multiple linear regressions and forest tree models describe the draugth. For the simple regression models, according to the study on the training set, the logarithmic model is selected for all the dependent variables. The results of quality of fit obtained for the test set is reported in Table 21.

Considering the value present in the table, it can be stated that all the regression methods give options that are at least moderate significant for the selected variable. The simple models provide moderate significant results, with the regression according to given L providing the best fitting results. For the multiple linear regression models, the best solutions are provided by *mlr-2* and *mlr-3*, which provide significant regressions for the given variable. In the case of forest tree, also in this case there is a decrease in the fitting performances by using the test set instead of the training set. As a consequence, the models are only moderate signif-

Table 21
Quality of fit on the test set for the T prediction models.

Regression		R^2	$MAPE$	$RMSE$	$RRMSE$	Prs
Simple	Given DWT	0.652	0.047	0.340	0.032	0.849
	Given Δ	0.677	0.047	0.327	0.031	0.839
	Given L	0.682	0.045	0.324	0.031	0.846
	Given LM	0.494	0.062	0.410	0.039	0.754
MLR	mlr-0	0.679	0.045	0.327	0.031	0.869
	mlr-1	0.576	0.048	0.376	0.036	0.825
	mlr-2	0.841	0.032	0.230	0.022	0.919
	mlr-3	0.814	0.029	0.249	0.024	0.921
	mlr-4	0.708	0.043	0.312	0.030	0.892
	mlr-5	0.390	0.055	0.451	0.044	0.817
Forest tree	ft-0	0.306	0.032	0.279	0.026	0.557
	ft-1	0.261	0.038	0.288	0.027	0.514
	ft-2	0.671	0.020	0.192	0.018	0.827
	ft-3	0.691	0.021	0.186	0.018	0.836
	ft-4	0.310	0.034	0.278	0.026	0.618
	ft-5	0.654	0.021	0.197	0.019	0.823

Table 22
Quality of fit on the test set for the Δ prediction models.

Regression		R^2	$MAPE$	$RMSE$	$RRMSE$	Prs
Simple	Given DWT	0.877	0.134	2.6E3	4.607	0.952
	Given L	0.865	0.124	2.7E3	4.908	0.930
	Given LM	0.892	0.135	2.4E3	4.343	0.952
MLR	mlr-0	0.865	0.172	2.7E3	4.785	0.962
	mlr-1	0.802	0.233	3.3E3	5.849	0.903
	mlr-2	0.877	0.140	2.6E3	4.625	0.942
	mlr-3	0.884	0.147	2.5E3	4.461	0.949
	mlr-4	0.912	0.118	2.2E3	3.944	0.970
	mlr-5	0.884	0.166	2.5E3	4.442	0.951
Forest tree	ft-0	0.666	0.073	2.4E3	3.945	0.839
	ft-1	0.001	0.154	4.2E3	7.090	0.337
	ft-2	0.956	0.029	865.9	1.457	0.978
	ft-3	0.662	0.104	2.4E3	4.062	0.816
	ft-4	0.801	0.038	1.8E3	3.069	0.904
	ft-5	0.880	0.053	1.4E3	2.377	0.954

icant, performing worst than the multiple linear regression models. In several cases, like for *ft-0*, *ft-1* and *ft-4* the goodness of fit indicators are less good than the simple regressions.

In conclusion, if a designer want to estimate the draught T, it is suggested to use either model *mlr-2* or *mlr-3*, resulting in a significant improvement compared to regressions available in the literature.

6.6. Displacement Δ

All the multiple linear regressions and forest tree models describe the displacement. For the simple regression models, according to the study on the training set, the power model is selected for the function of DWT and L, the logarithmic model for the function of LM. Of course, no model is present as a function of Δ. The results of quality of fit obtained for the test set is reported in Table 22.

The values in the Table highlights that all the regression methodologies employed provides at least one model which is significant for the selected variable. Considering the simple regressions, the best solution is given by the model as a function of LM; however, also the other two models have similar performance indicators and, therefore, could be used as a suitable alternative. The multiple linear regression models are all significant for the selected variable. The best solution is given by *mlr-4*; however, also other models like *mlr-3* and *mlr-5* give similar performance indicators. Different is the case of forest tree models. Here there is a scattering between the models. There are models like *ft-1* that are not significant for the given variable, and models like *ft-2* which have extremely good performance indicators. In any case, also for the displacement, there is a huge difference between the quality of fit indicators obtained for the forest tree models by using the test and the training set, observing a drastic reduction of the quality of fit for most of the tested models.

In conclusion, for the estimation of Δ a designer could choose *ft-2* as the best option. However, model*mlr-4* is a suitable alternative. The best models of each category provide a suitable estimation of the displacement.

6.7. Deadweight DWT

The deadweight is not described by all the multiple linear regression and forest tree models. Only models *mlr-1*, *mlr-3*, *mlr-5* and, consequently, *ft-1*, *ft-3* and *ft-5* describe the considered variable. For the simple regression models, according to the study on the training set, the logarithmic model is selected for the function of Δ, the power model for the function of LM and the linear model for the function of L. Of course, no model is present as a function of DWT. The results of quality of fit obtained for the test set is reported in Table 23.

The values reported in the table indicates that the models obtained with simple regressions and multiple linear regressions are all signif-

Table 23
Quality of fit on the test set for the DWT prediction models.

Regression		R^2	$MAPE$	$RMSE$	$RRMSE$	Prs
Simple	Given Δ	0.904	0.115	856.5	2.671	0.960
	Given L	0.841	0.169	1.1E3	3.408	0.924
	Given LM	0.873	0.122	986.7	2.992	0.948
MLR	mlr-1	0.902	0.134	864.1	2.655	0.960
	mlr-3	0.928	0.129	742.6	2.263	0.967
	mlr-5	0.918	0.118	791.9	2.417	0.964
Forest tree	ft-1	0.584	0.134	1.1E3	3.054	0.772
	ft-3	0.764	0.098	795.2	2.293	0.877
	ft-5	0.609	0.133	1.0E3	2.953	0.785

icant for the selected value, while the forest tree ones are moderate significant. The best simple model is the one as a function of Δ but also the other two options give similar results for the quality of fit. For the multiple linear regressions all the models are extremely good, especially *mlr-3*. Different is the case of the forest tree models, where all the provided regressions underperform the goodness of fit compared to the values registered on the training set.

In conclusion, the best model that could be selected for the estimation of DWT is *mlr-3*; however, also the other option given by simple and multiple linear regressions can be applied with good results. For this variable it is not advisable to use the forest tree models.

6.8. Installed power P_B

All the simple, multiple linear regressions and forest tree models describe the installed power. For the simple regression models, according to the study on the training set, the logarithmic model is selected for the function of DWT, Δ and L and the power model for LM. The results of quality of fit obtained for the test set is reported in Table 24.

The values of the quality of ft indicators reported in the Table highlights that the models obtained with the simple regressions are not significant for this variable. However, the models obtained with the multiple regressions and forest tree provides better values for the quality of fit. For the multiple linear regressions, the best solution is given by model *mlr-1*, which, according to the quality of fit indicators, is moderate significant for the present variable. In the case of forest tree, the best solution is *ft-5*, providing a model which is extremely significant for the estimation of P_B.

In conclusion, the best model that can be used for the estimation of P_B is provided by *ft-5*. For the case of installed power, all the provided models are an unicum in the literature, as no other regression is available for its estimation in the early design stage. Therefore, also the simpler models could be a help for RoPax designers in estimating P_B.

Table 24
Quality of fit on the test set for the P_B prediction models.

Regression		R^2	$MAPE$	$RMSE$	$RRMSE$	Prs
Simple	Given DWT	0.114	0.447	1.2E5	15.95	0.349
	Given Δ	0.163	0.385	1.1E5	15.94	0.439
	Given L	0.345	0.336	1.0E5	14.04	0.589
	Given LM	0.054	0.466	1.2E5	16.43	0.254
MLR	mlr-0	0.513	0.235	8.7E3	11.51	0.797
	mlr-1	0.629	0.182	7.6E3	10.35	0.813
	mlr-2	0.332	0.410	1.0E4	13.94	0.641
	mlr-3	0.466	0.360	9.2E3	12.60	0.706
	mlr-4	0.618	0.221	7.8E3	10.52	0.802
	mlr-5	0.398	0.259	9.7E3	13.57	0.781
Forest tree	ft-0	0.619	0.146	6.2E3	7.968	0.803
	ft-1	0.342	0.170	8.2E3	10.94	0.714
	ft-2	0.458	0.138	7.4E3	9.690	0.728
	ft-3	0.816	0.088	4.3E3	5.507	0.911
	ft-4	0.766	0.092	4.9E3	6.430	0.888
	ft-5	0.952	0.042	2.2E3	2.854	0.977

Table 25
Quality of fit on the test set for the V_s prediction models.

Regression		R^2	$MAPE$	$RMSE$	$RRMSE$	Prs
Simple	Given DWT	0.005	0.101	3.031	0.149	0.084
	Given Δ	0.002	0.111	3.074	0.153	0.016
	Given L	0.058	0.110	2.950	0.146	0.249
	Given LM	0.036	0.098	2.984	0.147	0.204
MLR	mlr-2	0.269	0.097	2.599	0.129	0.543
	mlr-3	0.187	0.100	2.741	0.137	0.506
Forest tree	ft-2	0.001	0.112	2.973	0.147	0.001
	ft-3	0.001	0.112	2.979	0.147	0.087

6.9. Ship speed V_s

The ship speed is not described by all the multiple linear regression and forest tree models. Only models *mlr-2*, *mlr-3* and, consequently, *ft-2* and *ft-3* describe the considered variable. For the simple regression models, according to the study on the training set, the logarithmic model is selected for the function of Δ and L, the linear model for the function of LM and DWT. The results of quality of fit obtained for the test set is reported in Table 25.

The values of the quality of fit indicators in the Table show extremely poor performances for all the presented models. Especially for the forest tree model there is a huge difference for the quality of fit indicators obtained on the training and the test set. Therefore, none of the models is significant for the present variable.

In conclusion, for the estimation of V_s the best option is given by *mlr-2*. However, the model is not significant for the present variable but is still an unicum in the literature for the estimation of V_s in the early design stage of RoPax ships.

6.10. Number of passengers N_p

The number of passenger is not described by all the multiple linear regression and forest tree models. Only models *mlr-0*, *mlr-1* and, consequently, *ft-0* and *ft-1* describe the considered variable. For the simple regression models, according to the study on the training set, the power model is selected for the function of DWT and LM, the linear model for the function of Δ and the logarithmic model for the function L. The results of quality of fit obtained for the test set is reported in Table 26.

The values of the quality of fit indicators in the Table show extremely poor performances for all the presented models. Especially for the forest tree model there is a huge difference for the quality of fit indicators obtained on the training and the test set. Therefore, none of the models is significant for the present variable.

In conclusion, the best solution for the estimation of N_p is given by *ft-1*. However, the poor values of the quality of fit indicators does not suggest to use with confidence this o one of the other models, even though are the sole available for the estimation of V_s in the early design stage of a RoPax.

Table 26
Quality of fit on the test set for the N_p prediction models.

Regression		R^2	$MAPE$	$RMSE$	$RRMSE$	Prs
Simple	Given DWT	0.015	0.545	536.3	3.856	0.072
	Given Δ	0.095	0.479	557.1	4.105	0.133
	Given L	0.023	0.570	525.9	3.569	0.240
	Given LM	0.062	0.579	548.5	3.855	0.085
MLR	mlr-0	0.001	0.603	633.0	3.997	0.545
	mlr-1	0.170	0.504	484.7	3.199	0.613
Forest tree	ft-0	0.001	0.731	791.8	5.257	0.001
	ft-1	0.216	0.567	698.9	4.646	0.488

Table 27
Quality of fit on the test set for the LM prediction models.

Regression		R^2	$MAPE$	$RMSE$	$RRMSE$	Prs
Simple	Given DWT	0.818	0.205	556.8	2.993	0.934
	Given Δ	0.789	0.261	598.8	3.328	0.948
	Given L	0.721	0.332	688.7	3.791	0.878
MLR	mlr-0	0.866	0.263	476.8	2.499	0.957
	mlr-2	0.879	0.238	452.7	2.383	0.961
	mlr-4	0.847	0.332	510.0	2.682	0.955
Forest tree	ft-0	0.835	0.077	286.6	1.425	0.917
	ft-2	0.846	0.088	277.4	1.413	0.931
	ft-4	0.981	0.024	96.76	0.485	0.992

6.11. Lane metres LM

The lane metres are not described by all the multiple linear regression and forest tree models. Only models *mlr-0*, *mlr-2*, *mlr-4* and, consequently, *ft-0*, *ft-2* and *ft-4* describe the considered variable. For the simple regression models, according to the study on the training set, the power model is selected for the function of L, the linear model for the function of DWT and the logarithmic model for the function Δ. Of course, no model as a function of LM is present. The results of quality of fit obtained for the test set is reported in Table 27.

The results reported in the Table shows that all the provided models are significant for the present variable. The best option between the simple regression methodologies is the one as a function of DWT but also the other options have similar quality of fit indicators. For the multiple linear regressions the best solution is provided by *mlr-2*. Also in this case, the other multiple regression models have similar performances. In the case of forest tree models, all the regressions are significant, with a preference for *ft-4*.

In conclusion, the best option for the estimation of LM is *ft-4* but also the other models can be employed with some confidence. The availability of robust methods for the estimation of LM is a plus, as they represent an unicum for the literature on RoPax ships.

6.12. Concluding remarks

This study presents and evaluates three distinct types of regression models for predicting the main dimensions and general particulars of RoPax vessels. From the analysis of the initial database presented in Section 3 it was immediately evident that some variables where more critical to analyse than others. This is the specific case of the depth D, where the initial distribution of data presents more than one subpopulation, and other quantities like the installed power P_B or the number of passengers N_p where the spreading of data was high for large ships.

Such a behaviour has been confirmed by the execution of the regression analysis presented in Section 5. In fact, the worst regressions remains the ones for the critical variables mentioned above. In any case, the execution of three different level of models highlights an increasing level of fidelity. Taking into consideration the quality of fit indicators (R^2, $MAPE$, $RMSE$, $RRMSE$ and Prs) calculated in Section 5, for all the variables the best models are represented by regression trees. The second best solution is given by the multiple linear regression models, while the simple regressions provide the worst quality of fit. The increase of quality of MLR compared to simple regressions is given by the fact that, the use of more than one independent variable, allows for considering the correlation between one of the independent variables with the dependent one. This is the case, for example of P_B, where the regressions as a function of V_s capture the strong correlation between V_s and P_B.

While testing the obtained regressions on a test-set, the results in terms of quality of fit were quite different. The models that all around performed better on the training set (i.e. the forest trees) were not confirm such a high level of fitting quality on the test set. A lot of the models having really high values for the training set results not significant for

the test set. Such a behaviour indicates that the forest tree models are good for reproducing exactly a population of data but, in this case, are not suitable to be used in a general way on a different set of ships. In any case, for some of the variables, forest tree models provide the best solution.

However, it should be highlighted that the selection of the prediction model that could be adopted for the estimation of main dimensions and general particulars depends on the initial variables at disposal by the designer. Here, the simple models cover the classical methods used in the literature for the estimation of main dimensions. More complex models require the availability of more than one variable to start the prediction process. In the present work, the selection of the multi-variables elements has been performed by considering design issues together with avoiding problems of auto-correlation between variables, resulting in the 6 models for multiple linear regressions and other 6 for the forest trees.

Of utmost importance is also the reproducibility of the data. Not all the methods are reproducible. The simple regression provided in the paper are directly applicable using equations from (50) to (157). Multiple linear regression models can be reproduced by implementing the equations using the coefficients presented in Appendix B. Different is the case of forest tree. The method is a so-called black-box method as other machine learning techniques, thus the resulting model is intrinsically embedded in the calculation machine. Therefore, a designer should implement the model directly from his own database. Such a consideration is limiting the reproducibility of the present work. However, the potential shown by the method in terms of quality of fit is undoubted.

The methods and equations presented in this paper can be used by designers according to different strategies for the determination of main dimensions. A potential user can estimate main dimensions according to the multiple linear regression method if all the required independent variables can be provided. Otherwise, use can be made of simple regression models, as per classical literature. In such a case, a user can estimate the main dimensions starting from the deadweight DWT, the displacement Δ or the lane metres LM. Alternatively, could also estimate the length L with one of the method above and then using the regression as a function of L for the estimate of remaining parameters.

The application example on the test set shows that there is not a directly preferable way to proceed with; however, the application of a higher fidelity method gives more confidence in the obtained results.

7. Conclusions

This paper presents a comprehensive analysis of a database comprising 87 RoPax vessels, and introduces various regression techniques to estimate the ships' main dimensions and other relevant general particulars during the early design phase. The study is limited to the case of large RoPax vessels, having a length longer than 100 metres, and it is not advisable to adopt the obtained regressions outside their limiting range. The dataset was split into training and test subsets to enable model development and validation, respectively. Starting from the implementation of simple regression models as a function of deadweight, displacement, lane meters and length, multiple linear regression models have been provided to enhance the capability of the fitting models, by considering more than one independent variable. Besides, machine learning techniques have been also implemented in the form of forest trees to further enhance the quality of the obtained regression models.

As first result, the comparison between the available models in the literature for RoPax and the simple regressions provided in this work highlights how the literature models doesn't fit well the current database. This is mainly due to the age of the vessels considered in the analysis as the present database uses modern ships, while the previous studies employ quite old vessels. Therefore, the provided simple regression models as a function of the deadweight, displacement, lane metres and length are a consistent improvement compared to literature data.

As a second result, the paper provides 46 multiple linear regression models for the main dimensions and general particulars of RoPax vessels, employing couple or triplets of independent variables for the analysis. The detailed analysis of the regressions, reported in the paper and in the appendix, highlight the quality of the provided models, which represent an improvement compared to the simple regression models for the quality of fit. The multiple linear regressions are capable of providing a higher quality of fit compared to the simple regression models, allowing the designers to take into consideration more than one single independent variable as initial parameter for the main dimensions estimation.

Finally, the work provides the results and analysis of 46 forest tree models for the prediction of the main dimension and general particulars of RoPax vessel. The forest tree models are generated employing the same independent variables of the multiple linear regression models. Forest trees have a better quality of fit of all the previous model, as indicated by all the performance indicators employed in the study. Such a machine learning technique is powerful for the prediction of main dimension of the vessel; however, the technique is a black-box model and is not easy to reproduce from this study. Furthermore, the testing on the test set highlights that not all the models grant the same quality of fit obtained on the training set.

In any case, the application of all the presented methods to the test set highlights that suitable prediction can be performed also with simple regression models or multiple linear regressions. The provided models have a high level of reliability except for the depth of the ship, because of a strange initial population in the database. However, between the main dimensions, the depth is the less relevant, as it could be derived from regulations while the draught is known.

This study underscores the necessity of examining more modern databases to develop models that aid designers in estimating the main particulars of ships. It emphasizes that modern machine learning techniques, alongside conventional methods, can also be employed in predictions, even though a larger initial population sample is advisable. To further this purpose, additional research will be conducted on different vessel types to determine if the findings of this work on RoPax vessels can be extended to other ships.

CRediT authorship contribution statement

Francesco Mauro: Writing – review & editing, Writing – original draft, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization; **Ahmed Salem:** Writing – review & editing, Investigation, Conceptualization.

Declaration of competing interest

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

Appendix A. Detailed analysis of the obtained regressions

The present appendix reports a detailed analysis of the regressions obtained by all the three proposed regression methods, providing to the reader a more in-depth overview of the results presented by the present research. Therefore, the following subsections reports the results obtained for simple, multiple regressions and forest three results.

A.1. Simple regressions

Here the results of the simple regression analysis are reported distinguishing between the cases obtained by using the four independent variables DWT, Δ, LM and L.

A.1.1. Regressions as a function of DWT

Hereafter a detailed analysis is performed variable by variable for the regressions presented in Section 5.1.1:

- *Length L*: the length can be modelled according to formulae (50), (59), and (68). All the quality of fit indicators in Table 2 suggest that the best fit option for L is the power model expressed by formula (59). The level of indicators like R^2 or Prs state that the regression is significant for the selected variable.
- *Breadth B*: the breadth can be modelled according to formulae (51), (60), and (69). Also in this case, all the quality of fit indicators in Table 2 suggest that the best fit option is given by the power model of formula (60). The level of indicators like R^2 or Prs state that the regression is moderate significant for the selected variable.
- *Depth D*: the depth can be modelled according to formulae (52), (61), and (70). In this case, there is a discrepancy between the indications given by the fit quality indicators. According to R^2, $RMSE$, $RRMSE$ and Prs the best model is the logarithmic one of formula (70); but according to $MAPE$ the best model is the power one of formula (61). In any case, the differences between the three models are minimal and, therefore, the formulae can be considered equivalent. However, considering the extremely low level of the indicators like R^2 and Prs, it can be stated that the regressions are not significant for the selected variable.
- *Draught T*: the draught can be modelled according to formulae (53), (62), and (71). There is a discrepancy between the quality of fit indicators, as for the R^2 the best model is the power one formula (62) while for the others the best model is the logarithmic one formula (71). Differences between the two formulae are not so large and, therefore, the models can be considered equivalent. Considering the level of the indicators like R^2 and Prs, it can be stated that the regression is moderate significant for the selected variable.
- *Displacement* Δ: the displacement can be modelled according to formulae (54), (63), and (72). According to all the quality of fit indicators of Table 2, the best model is the power one, expressed by formula (63). According to the level of the indicators like R^2 and Prs, the regression is significant for the selected variable.
- *Installed power* P_B: the installed power can be modelled according to formulae (55), (64), and (73). in this case, there is a discrepancy between the quality of fit indicators, as the best model for R^2 is the power one of formula (64), while for the others the best model is the logarithmic of formula (73). In any case, the quality of fit is so low that both models can be considered not significant for the selected variable.
- *Speed* V_s: the speed can be modelled according to formulae (56), (65), and (74). According to the quality of fit indicators of Table 2, the best model is the linear one of formula (56). However, the level of indicators like R^2 and Prs indicate that the regression is not significant for the selected variable.
- *Number of passengers* N_p: the number of passengers can be modelled according to formulae (57), (66), and (75). According to the quality of fit indicators of Table 2, the best model is the linear one of formula (57). However, the level of indicators like R^2 and Prs indicate that the regression is not significant for the selected variable.
- *Lane metres LM*: the lane meters can be modelled according to formulae (58), (67), and (76). According to the quality of fit indicators of Table 2, the best model is the power one of formula (67). The level of indicators like R^2 and Prs allows to state that the regression is significant for the selected variable.

A.1.2. Models as a function of Δ

Hereafter, a detailed analysis is presented variable by variable of the regressions presented in Section 5.1.2:

- *Length L*: the length can be modelled according to formulae (77), (86), and (95). Considering the quality of fit indicators of Table 3, the R^2 values highlights the power regression of formula (86) as the best option, while the others indices indicate the logarithmic one formula (95). In any case, the values between the two formulae are similar and, therefore, they could be considered equivalent. Taking a look to the values of R^2 and Prs it can be stated that the regression is significant for the selected variable.
- *Breadth B*: the breadth can be modelled according to formulae (78), (87), and (96). There is a discrepancy between the quality of fit indicators in finding the best regression among the proposed models. $RMSE$ and $RRMSE$ indicates the linear model of formula (78) as the best, while the other indicators point to the power model of formula (87). As the values of $RMSE$ and $RRMSE$ are quite similar, it can be stated that formula (87) is the best option. The values of R^2 and Prs state that the regression is significant for the selected variable.
- *Depth D*: the depth can be modelled according to formulae (79), (88), and (97). Considering the quality of fit indicators of Table 3, the best regression is the linear one of formula (79). However, considering the level of indicators R^2 and Prs it can be stated that the regression is not significant for the selected variable.
- *Draught T*: the draught can be modelled according to formulae (80), (89), and (98). Taking a look to the quality of fit indicators, the R^2 suggests the power regression formula (89) as the best but the other indicators are in favour of the logarithmic one formula (98). Therefore, the best regression for T is formula (98) and, considering the level of the indicators R^2 and Prs, it can be stated that the regression is moderate significant for the variable.
- *Deadweight DWT*: the deadweight can be modelled according to formulae (81), (90), and (99). The quality of fit indicators of Table 3 show more variability. According to R^2 and $MAPE$, the best regression is formula (90), the others are in favour of formula (99). Therefore, the logarithmic regression formula (99) can be considered as the best option. The level of indicators R^2 and Prs state that the regression is significant for the selected variable.
- *Installed Power* P_B: the installed power can be modelled according to formulae (82), (91), and (100). The R^2 and $MAPE$ indicators suggest the power model formula (91) as the best model, the others are in favour of the logarithmic one formula (100). Therefore formula (100) can be considered the best regression. However, the level of indicators R^2 and Prs state that the regression is not significant for the selected variable.
- *Speed* V_s: the speed can be modelled according to formulae (83), (92), and (101). R^2 value suggests the power model formula (92) as the best regression, the other parameters the logarithmic one formula (101). However, the discrepancy between the two regressions is relatively low and they can be considered equivalent. The level of indicators R^2 and Prs state that the regression is not significant for the selected variable.
- *Number of passengers* N_p: the number of passengers can be modelled according to formulae (84), (93), and (102). All the indicators, except for the Prs state that the best model is the logarithmic one formula (102), while Prs indicates the linear model formula (84). Therefore, formula (102) can be considered as the best model. The level of indicators R^2 and Prs state that the regression is not significant for the selected variable.
- *Lane metres LM*: the lane metres can be modelled according to formulae (85), (94), and (103). Except for R^2, the indicators of Table 3 suggest the linear model formula (85) as the best model, the R^2 indicates the power one formula (94). Therefore, the linear model can be considered as the best regression among the three available options. The level of indicators R^2 and Prs state that the regression is moderate significant for the selected variable.

A.1.3. Models as a function of LM

A detailed analysis of each variable of the models presented in Section 5.1.3 is presented below:

- *Length L*: the length can be modelled according to formulae (104), (113), and (122). Considering the quality of fit indicators of Table 4, the best model for R^2 and $MAPE$ is the power one formula (113) while for the others is the logarithmic model formula (122). Therefore, the best model can be the logarithmic regression. According to the level of R^2 and Prs it can be stated that the regression model is moderate significant for the present variable.
- *Breadth B*: the breadth can be modelled according to formulae (105), (114), and (123). Considering the quality of fit indicators of Table 4, the best model for R^2 and $MAPE$ is the power one formula (114) while for the others is the logarithmic model formula (123). The difference between the two models is relatively small; therefore, they can be considered equivalent. According to the level of R^2 and Prs it can be stated that the regression model is moderate significant for the present variable.
- *Depth D*: the depth can be modelled according to formulae (106), (115), and (124). Considering the quality of fit indicators of Table 4, the best model is the linear one of formula (106). According to the level of R^2 and Prs it can be stated that the regression is not significant for the present variable.
- *Draught T*: the draught can be modelled according to formulae (107), (116), and (125). Considering the quality of fit indicators of Table 4, the best model for R^2 and $MAPE$ is the power one formula (116) while for the others is the logarithmic model formula (125). The difference between the two models is relatively small; therefore, they can be considered equivalent. According to the level of R^2 and Prs it can be stated that the regression model is moderate significant for the present variable.
- *Deadweight DWT*: the deadweight can be modelled according to formulae (108), (117), and (126). Considering the quality of fit indicators of Table 4, the best model is the power one of formula (117). According to the level of R^2 and Prs it can be stated that the regression is significant for the present variable.
- *Installed power* P_B: the installed power can be modelled according to formulae (109), (118), and (127). Considering the quality of fit indicators of Table 4, the best model for R^2 and $MAPE$ is the power one formula (118) while for the others is the logarithmic model formula (127). The difference between the two models is relatively small; therefore, they can be considered equivalent. According to the level of R^2 and Prs it can be stated that the regression model is not significant for the present variable.
- *Speed* V_s: the vessel speed can be modelled according to formulae (110), (119), and (128). Considering the quality of fit indicators of Table 4, the best model is the linear one of formula (110). According to the level of R^2 and Prs it can be stated that the regression model is not significant for the present variable.
- *Number of passengers* N_p: the number of passengers can be modelled according to formulae (111), (120), and (129). Considering the quality of fit indicators of Table 4, the best model is the linear one of formula (111). According to the level of R^2 and Prs it can be stated that the regression model is not significant for the present variable.
- *Displacement* Δ: the displacement can be modelled according to formulae (112), (121), and (130). Considering the quality of fit indicators of Table 4, the best model is the power one of formula (121). According to the level of R^2 and Prs it can be stated that the regression model is significant for the present variable.

A.1.4. Models as a function of L

Hereafter, a detailed analysis is presented variable by variable for the regression models presented in Section 5.1.4:

- *Displacement* Δ: the displacement can be modelled according to formulae (131), (140), and (149). According to the quality of fit indicators of Table 5, the best regression is the power model of formula (140). The level of indicators like R^2 and Prs state that the regression is significant for the selected variable.
- *Breadth B*: the breadth can be modelled according to formulae (132), (141), and (150). Taking into consideration the R^2 and $MAPE$, the best model is the power one of formula (141). The other indicators are equivalent for the power and the linear model. Therefore, the power model of formula (141 can be considered the best regression for B. The level of R^2 and Prs state that the regression is moderate significant for the selected variable.
- *Depth D*: the depth can be modelled according to formulae (133), (142), and (151). For this case, the indicators shown in Table 5 are discordant. The R^2, Prs and $RMSE$ indicate the linear model as the best solution, $MAPE$ the power model and $RRMSE$ the logarithmic and the linear. However, the differences are relatively small between the three regressions. In any case, according to the level of R^2 and Prs, the regressions are not significant for the selected variable.
- *Draught T*: the draught can be modelled according to formulae (134), (143), and (152). Taking into consideration the R^2 values, the best regression is given by the power model formula (143). The other indicators are in favour of the logarithmic model of formula (152). Therefore, the best model could be the logarithmic one. The level of R^2 and Prs state that the regression is moderate significant for the selected variable.
- *Deadweight DWT*: the deadweight can be modelled according to formulae (135), (144), and (153). According to the $MAPE$, the best model is the linear one formula (135). For the other indices the best regression is the power model formula (144). Therefore the best regression can be given by the power model. The level of R^2 and Prs state that the regression is significant for the selected variable.
- *Installed power* P_B: the installed power can be modelled according to formulae (136), (145), and (154). The R^2 value is in favor of the power regression of formula (136). However, the other indices in Table 5 indicate the logarithmic model formula (154) as the best solution. In any case, the level of the R^2 and Prs state thar the regression is moderate significant for the selected variable.
- *Speed* V_s: the speed can be modelled according to formulae (137), (146), and (155). According to the quality of fit indicators, all the models are similar, with a preference for the power model formula (146) according to the R^2 and $MAPE$ values and for the logarithmic model formula (155) according to the other indicators. However, the level of R^2 and Prs state that the regression is not significant for the selected variable.
- *Number of passengers* N_p: the number of passengers can be modelled according to formulae (138), (147), and (156). The quality of fit indicators of Table 5 are in favour of the power model of formula (147), except for the $RRMSE$, which indicates the logarithmic model formula (156) as the best option. In any case, the level of R^2 and Prs state that the regression is not significant for the selected variable.
- *Lane metres LM*: the lane metres can be modelled according to formulae (139), (148), and (157). The values of the quality of fit indicators of Table 5 suggest that the best model is the power one of formula (148). The level of R^2 and Prs state that the regression is significant for the selected variable.

A.2. Multiple linear regressions models

Here a detailed variable by variable analysis is presented for all the regression models obtained by means of multiple linear regression analysis. Each Subsection reports the cases obtained for each of the tuples of independent variables considered in the analysis.

A.2.1. Regressions as a function of V_s and DWT

In the following, a detailed analysis of the results presented in Section 5.2.1 for each variable is performed:

- *Length L*: the length is well captured by the obtained multiple linear regression. Indicators like R^2, R^2_{adj} and Prs are relatively high, being all above 0.8. The other indicators are small, confirming the good

quality of the obtained regression. Therefore, it can be concluded that the regression of L as a function of V_s and DWT is significant.

- *Breadth B*: the breadth is moderate well captured by the obtained multiple linear regression. Considering the indicators visible in Table 6, the R^2, R^2_{adj} and the Prs are between 0.5 and 0.8, showing an average quality of the regression. Therefore, it can be concluded that the regression of B as a function of V_s and DWT is moderate significant.
- *Depth D*: the multiple linear regression does not provide a particularly good representation of the depth. The indicators like R^2, R^2_{adj} and Prs are between 0.1 and 0.5, showing a bad quality of the regression. Therefore, it can be con concluded that the regression of D as a function of V_s and DWT is not significant.
- *Draught T*: the draught is well captured by the obtained multiple linear regression. The indicators of Table 6 highlights high value for R^2, R^2_{adj} and Prs (all above 0.7) and small values for the remaining ones. Therefore, it can be concluded that the regression of T as a function of V_s and DWT is significant.
- *Displacement* Δ: the displacement is well captured by the obtained multiple linear regression. The indicators of Table 6 show high values above 0.65 for the R^2, R^2_{adj} and Prs and all the other indicators are relatively small. Therefore, the regression of Δ as a function of V_s and DWT is significant.
- *Installed power* P_B: the installed power is well captured by the obtained multiple linear regression. According to the quality of fit indicators in Table 6, the regression is quite good. In fact, R^2, R^2_{adj} and Prs are all above 0.75 and the other indicators are relatively low. Therefore, the regression of P_B as a function of V_s and DWT is significant.
- *Number of passengers* N_p: the number of passenger is not really well captured by the obtained multiple linear regression. The R^2, R^2_{adj} and Prs are between 0.25 and 0.6 and the other indicators have quite high relative values. Therefore, the regression of N_p as a function of V_s and DWT is slightly significant.
- *Lane metres LM*: the lane metres are well captured by the obtained multiple linear regression. The qulity of fit indicators of Table 6 show values of R^2, R^2_{adj} and Prs above 0.8 and the remaining coefficients are significantly low. Therefore, the regression of LM as a function of V_s and DWT is significant.

A.2.2. Regressions as a function of V_s and LM

A detailed analysis for each variable is hereafter reported for the regression analysis presented in Section 5.2.2:

- *Length L*: the length is well represented by the obtained multiple linear regression. The quality of fit indicators of Table 7 highlight values of R^2, R^2_{adj} and Prs all above or close to 0.75. The other indicators are relatively low. Therefore, it can be concluded that the regression of L as a function of V_s and LM is significant.
- *Breadth B*: B is moderately well represented by the obtained multiple linear regression. The indicators in Table 7 show values between 0.45 and 0.7 for the R^2, R^2_{adj} and Prs, while the other indicators remain reasonably low. Therefore, it can be concluded that the regression of B as a function of V_s and LM is moderate significant.
- *Depth D*: the multiple linear regression does not provide a particularly good representation of the depth. The values of indicators like R^2, R^2_{adj} and Prs remain low, ranging between 0.15 and 0.5. The remaining coefficients have relatively high values compared to other regressions. Therefore it can be concluded that the regression of D as a function of V_s and LM is not significant.
- *Draught T*: the draugth is well captured by the obtained multiple linear regression. Considering the values of indicators like R^2, R^2_{adj} and Prs (all above or close to 0.65), the quality of fit is good. Therefore, it can be concluded that the regression of T as a function of V_s and LM is significant.
- *Displacement* Δ: the displacement is quite well captured by the obtained multiple linear regression. The indicators in Table 7 shows values for R^2, R^2_{adj} and Prs all above or close to 0.58. Other indicators can be considered relatively low compared to other regressions. Therefore, it can be concluded that the regression of Δ as a function of V_s and LM is moderate significant.
- *Installed power* P_B: the installed power is well captured by the obtained multiple linear regression. The values reported for R^2, R^2_{adj} and Prs are all above 0.75, indicating a good quality of fit. Other indicators present small values, which is also positive in this sense. Therefore, it can be concluded that the regression of P_B as a function of V_s and LM is significant.
- *Number of passengers* N_p: the number of passenger are not properly captured by the obtained multiple linear regressions. The quality of fit indicators of Table 7 show values between 0.25 and 0.5 for R^2, R^2_{adj} and Prs, while other indicators are relatively high. Therefore, it can be concluded that the regression of N_p as a function of V_s and LM is slightly significant.
- *Deadweight DWT*: the deadweight is well represented by the obtained multiple linear regression. The R^2, R^2_{adj} and Prs indicators are all above 0.85, while the others are low, indicating a good quality of fit. Therefore, it can be concluded that the regression of DWT as a function of V_s and N_p is significant.

A.2.3. Regressions as a function of N_p and DWT

The following items reports a detailed analysis of the quality of fit for each variable , resulting from the regressions reported in Section 5.2.2:

- *Length L*: the length is well represented by the obtained multiple linear regression. The quality of fit indicators of Table 8 present high values for the $R62$, R^2_{adj} and Prs, all above 0.8, while the remaining indicators are low, highlighting the good quality of the regression. Therefore, it can be stated that the regression of L as a function of N_p and DWT is significant.
- *Breadth B*: the breadth is quite well represented by the obtained multiple linear regression. The R^2, R^2_{adj} and Prs are all above 0.8, while the remaining indicators are relatively low. Therefore, it can be concluded that the regression of B as a function of N_p and DWT is significant.
- *Depth D*: the depth is not particularly well represented by the obtained multiple linear regression. The indicators reported in Table 8 shows low values for the R^2, R^2_{adj} and Prs, all ranging between 0.14 and 0.5, while the remaining indicators are relatively high. Therefore, it can be stated that the regression of D as a function of N_p and DWT is not significant.
- *Draught T*: the draugth is well represented by the obtained multiple linear regression. The quality of fit indicators R^2, R^2_{adj} and Prs are all above 0.8, while the other indicators are relatively low, highlighting the good quality of the regression. Therefore, it can be concluded that the regression of T as a function of N_p and DWT is significant.
- *Displacement* Δ: the displacement is well represented by the obtained multiple linear regression. Table 8 shows how R^2, R^2_{adj} and Prs got values higher than 0.8, while the other indices are low. Therefore, it can be concluded that the regression of Δ as a function of N_p and DWT is significant.
- *Installed Power* P_B: the installed power is moderately well represented by the obtained multiple linear regression. The quality of fit indicators of Table 8 show that the value of R^2, R^2_{adj} and Prs are ranging between 0.4 and 0.7, while the other indicators remain moderately high. Therefore, it can be stated that the regression of P_B as a function of N_p and DWT is moderate significant.
- *Speed* V_s: the speed is moderately well represented by the obtained multiple linear regressions. R^2, R^2_{adj} and Prs have values ranging between 0.25 and 0.6, while the remaining indicators are moderately high. Therefore, it can be stated that the regression of V_s as a function of N_p and DWT is moderate significant.
- *Lane metres LM*: the lane metres are well represented by the obtained multiple linear regression. Table 8 reports values above 0.85 for the R^2, R^2_{adj} and Prs, while the remaining indicators are low,

highlighting the good quality of the regression. Therefore, it can be stated that the regression of LM as a function of N_p and DWT is significant.

A.2.4. Regressions as a function of N_p and LM

Hereafter, a detailed analysis of the quality of fit for the regressions reported in Section 5.2.3, is reported for each variable:

- *Length* L: the length is well represented by the obtained multiple linear regression. The level of the quality of fit indicators R^2, R^2_{adj} and Prs, all above 0.8, highlights a good quality of the regressions. At the same time, the low level of the remaining indicators confirm this conclusion. Therefore, it can be stated that the regression of L as a function of N_p and LM is significant.
- *Breadth* B: the breadth is quite well represented by the obtained multiple linear regression. The indicators R^2, R^2_{adj} and Prs are all above 0.7, while the remaining indicators are relatively low. Therefore, it can be concluded that the regression of B as a function of N_p and LM is significant.
- *Depth* D: the depth is not well represented by the obtained multiple linear regression. The level of R^2, R^2_{adj} and Prs in Table 9 reveals values between 0.15 and 0.5, while the remaining indicators remains relatively high, implying a bad quality of fit. Therefore, it can be concluded that the regression of D as a function of N_p and LM is not significant.
- *Draugth* T: the draught is well represented by the obtained multiple linear regression. The indicators of Table 9 present value above 0.75 for the R^2, R^2_{adj} and Prs, while the remaining coefficients remains relatively low. Therefore, it can be stated that the regression of T as a function of N_p and LM is significant.
- *Displacement* Δ: the displacement is well represented by the obtained multiple linear regression. R^2, R^2_{adj} and Prs indicators in Table 9 have values above 0.75, while the other indicators remains relatively low. Therefore, it can be concluded that the regression of Δ as a function of N_p and LM is significant.
- *Installed power* P_B: the installed power P_B is moderate well represented by the obtained multiple linear regression. The quality of fit indicators of Table 9 have values ranging from 0.45 to 0.75 for the R^2, R^2_{adj} and Prs, while the remaining coefficients have moderately high values. Therefore, it can be stated that the regression of P_B as a function of N_p and LM is moderate significant.
- *Speed* V_s: the speed is moderately well represented by the obtained multiple linear regressions. R^2, R^2_{adj} and Prs have values ranging between 0.35 and 0.65, while the remaining indicators are moderately high. Therefore, it can be stated that the regression of V_s as a function of N_p and LM is moderate significant.
- *Deadweight* DWT: the deadweight is well represented by the obtained multiple linear regression. The values of the indicators of Table 9 show levels above 0.85 for the R^2, R^2_{adj} and Prs, while the others remain low, highlighting the quality of the obtained fit. Therefore, it can be stated that the regression of DWT as a function of N_p and LM is significant.

A.2.5. Regressions as a function of N_p, V_s and DWT

In the following a detailed analysis of the obtained data for the regressions reported in Section 5.2.5 is performed for each variable:

- *Length* L: the length is well represented by the obtained multiple linear regression. The level of indicators R^2, R^2_{adj} and Prs is always above 0.85 while the remaining indicators are low, highlighting the good quality of the regression. Therefore, it can be concluded that the regression of L as a function of V_s, N_p and DWT is significant.
- *Breadth* B: the breadth is well represented by the obtained multiple linear regression. Table 10 shows values of R^2, R^2_{adj} and Prs above 0.8 and the remaining indicators are relatively low. Therefore, it can be stated that the regression of B as a function of V_s, N_p and DWT is significant.
- *Depth* D: the depth is not particularly well represented by the obtained multiple linear regression. The level of R^2, R^2_{adj} and Prs is ranging from 0.2 to 0.6, while the remaining coefficients in Table 10 remains high. Therefore, it can be concluded that the regression of D as a function of V_s, N_p and DWT is slightly significant.
- *Draught* T: the draught is well captured by the obtained multiple linear regression. The indicators in Table 10 have values of R^2, R^2_{adj} and Prs is always above 0.85, while the remaining indicators are low. Therefore, it can be concluded that the regression of T as a function of $V-s$, N_p and DWT is significant.
- *Displacement* Δ: the displacement is well represented by the obtained multiple linear regression. The indicators R^2, R^2_{adj} and Prs are always above 0.8, while the remaining indicators remains relatively low. Therefore, it can be concluded that the regression of Δ as a function of V_s, N_p and DWT is significant.
- *Installed power* P_B: the installed power is well represented by the obtained multiple linear regression. Table 10 shows values of R^2, R^2_{adj} and Prs always above 0.8, while the remaining coefficients are relatively low. Therefore, it can be stated that the regression of P_B as a function of V_s, N_p and DWT is significant.
- *Lane metres* LM: the lane metres are well represented by the obtained multiple linear regression. The level of R^2, R^2_{adj} and Prs is always above 0.8, while the level of the remaining indicators is relatively low. Therefore, it can be concluded that the regression of LM as a function of V_s, N_p and DWT is significant.

A.2.6. Regressions as a function of N_p, V_s and LM

In the following, a detailed analysis of the obtained regressions (presented in Section 5.2.6) is performed:

- *Length* L: the length is well represented by the obtained multiple linear regression. The indicators of Table 11 highlights the good quality of the regression. In fact, R^2, R^2_{adj} and Prs have values above 0.85, while the remaining indicators have low values. Therefore, it can be concluded that the regression of L as a function of V_s, N_p and LM is significant.
- *Breadth* B: the breadth is well represented by the obtained multiple linear regression. R^2, R^2_{adj} and Prs are all above 0.8, while the remaining indicators are relatively low. Therefore, it can be concluded that the regression of B as a function of V_s, N_p and LM is significant.
- *Depth* D: the depth is not properly well represented by the obtained multiple linear regression. According to the data in Table 11, the R^2, R^2_{adj} and Prs are ranging between 0.35 and 0.7, while the remaining indicators remain relatively high. Therefore, it can be stated that the regression of D as a function of V_s, N_p and LM is moderate significant.
- *Draught* T: the draught is well represented by the obtained multiple linear regression. The R^2, R^2_{adj} and Prs values, visible in Table 11, are all above 0.75 and the remaining indicators are relatively low. Therefore, it can be stated that the regression of T as a function of V_s, N_p and LM is significant.
- *Displacement* Δ: the displacement is well represented by the obtained multiple linear regression. The data in Table 11 highlight that the values of R^2, R^2_{adj} and Prs are all above 0.8 while remaining indices are low. Therefore, it can be concluded that the regression of Δ as a function of V_s, N_p and LM is significant.
- *Installed power* P_B: the installed power is well represented by the obtained multiple linear regression. The R^2, R^2_{adj} and Prs have values above 0.8, while other indicators remain low highlighting the good quality of the regression. Therefore, it can be concluded that the regression of P_B as a function of V_s, N_p and LM is significant.
- *Deadweight* DWT: the deadweight is well represented by the obtained multiple linear regression. The R^2, R^2_{adj} and Prs in Table 11 are all above 0.85, while the remaining indices are all relatively low. Therefore, it can be concluded that the regression of DWT as a function of V_s, N_p and LM is significant.

A.3. Forest tree models

Here a detailed variable by variable analysis is presented for all the regression models obtained by means of forest tree regression analysis. Each Subsection reports the cases obtained for each of the tuples of independent variables considered in the analysis.

A.3.1. Regressions as a function of V_s and DWT

In the following a detailed description of the quality of fit for the regressions presented in Section 5.3.1 is presented variable by variable:

- *Length* L: the length is well represented by the obtained forest tree. Taking a look to the quality of fit indicators of Table 12, it is possible to observe that R^2 and Prs are above 0.99 and the remaining ones are relatively low. This is an indication of an excellent quality of fit. Therefore, it can be stated that the forest tree for L as a function of V_s and DWT is significant.
- *Breadth* B: the breadth is well represented by the obtained forest tree. The R^2 and Prs are both above 0.97 and the remaining indicators remain low, highlighting the excellent quality of the fit. Therefore it can be stated that the forest tree model for B as a function of V_s and DWT is significant.
- *Depth* D: the depth is moderately well represented by the obtained forest tree. The R^2 and Prs in Table 12 have values ranging from 0.61 to 0.81 while the other coefficients remain slightly low. Therefore, it can be concluded that the forest tree model for D as a function of V_s an DWT is moderate significant.
- *Draugth* T:the draugth is really well represented by the obtained forest tree model. R^2 and Prs values are both above 0.9 and the other coefficients are significantly low. This is an indicator of an excellent quality of fit of the model. Therefore, it can be concluded that the forest tree model for T as a function of V_s and DWT is significant.
- *Displacement* Δ: the displacement is really well represented by the obtained forest tree model. The R^2 and Prs values in Table 12 are both above 0.9 and the remaining indicators remain low. Therefore, it can be concluded that the forest tree model for Δ as a function of V_s and DWT is significant.
- *Installed power* P_B: the installed power is well represented by the obtained forest tree model. The R^2 and Prs values are above 0.9 while the other indices are relatively low. Therefore, it can be concluded that the forest tree model of P_B as a function of V_s and DWT is significant.
- *Number of passengers* N_p: the number of passengers are well captured by the obtained forest tree model. The level of R^2 and Prs visible in Table 12 is above 0.85 while the remaining indices are relatively low. Therefore, it can be stated that the forest tree model of N_p as a function of V_s and DWT is significant.
- *Lane metres* LM:the lane metres are well represented by the obtained forest tree model. The R^2 and Prs indices are both above 0.95 while the remaining ones are all relatively low. Therefore, it can be stated that the forest tree model of LM as a function of V_s and DWT is significant.

A.3.2. Regressions as a function of V_s and LM

In the following a detailed analysis of the obtained data, for the regressions obtained in Section 5.3.2 is performed variable by variable:

- *Length* L: the length is well represented by the obtained forest tree model. The values of R^2 and Prs in Table 13 are both above 0.95, while the remaining indices are all relatively low. Therefore it can be concluded that the forest tree model of L as a function of V_s and LM is significant.
- *Breadth* B: the breadth is well represented by the obtained forest tree model. The values of R^2 and Prs in Table 13 are both above 0.95, while the remaining indices are all relatively low. Therefore it can be concluded that the forest tree model of B as a function of V_s and LM is significant.
- *Depth* D: the depth is moderately well represented by the obtained forest tree model. The quality of fit indicators R^2 and Prs are ranging between 0.7 and 0.8 while the other indices in Table 13 remains slightly low. Therefore, it can be stated that the forest tree model of D as a function of V_s and LM is moderate significant.
- *Draugth* T: the draught is well represented by the obtained forest tree model. The quality of fit indicators in Table 13 shows values for R^2 and Prs above 0.95 while the remaining coefficients are low. Therefore, it can be concluded that the forest tree model of T as a function of V_s and LM is significant.
- *Displacement* Δ: the displacement is well represented by the obtained forest tree model. The values of R^2 and Prs in Table 13 are both above 0.95, while the remaining indices are all relatively low. Therefore it can be concluded that the forest tree model of Δ as a function of V_s and LM is significant.
- *Installed power* P_B: the installed power is really well represented by the obtained forest tree model. R^2 and Prs indicators are both above 0.95, while the remaining indicators are low, highlighting the extremely good quality of the regression. Therefore, it can be concluded that the model of P_B as a function of V_s and LM is significant.
- *Number of passengers* N_p: the number of passengers is well represented by the obtained forest tree model. The indicators like R^2 and Prs in Table 13 have values above 0.90 while the remaining coefficients are relatively low. Therefore, it can be concluded that the forest tree model of N_p as a function of V_s and LM is significant.
- *Deadweight* DWT: the deadweight is well captured by the obtained forest tree model. The quality of fit indicators of Table 13 have values above 0.9 for the R^2 and Prs while the remaining coefficients are low. Therefore, it can be concluded that the forest tree model of DWT as a function of V_s and LM is significant.

A.3.3. Regressions as a function of N_p and DWT

In the following a detailed analysis of the obtained data in Section 5.3.3 is performed variable by variable:

- *Length* L: the length is really well represented by the obtained forest tree model. The indicators R_2 and Prs are both above 0.95, while the others are low, highlighting the excellent quality of the fitted model. Therefore, it can be concluded that the forest tree model of L as a function of N_p and DWT is significant.
- *Breadth* B: the breadth is well represented by the obtained forest tree model. The R^2 and Prs coefficients in Table 14 are both above 0.95, while the remaining ones are relatively low. Therefore, it can be stated that the forest tree model of B as a function of N_p and DWT is significant.
- *Depth* D: the depth is not well represented by the obtained forest tree model. The R^2 and Prs indicators are ranging from 0.2 to 0.5 while the other coefficients are relatively high. Therefore, it can be stated that the forest tree model of D as a function of N_p and DWT is not particularly significant.
- *Draugth* T: the draught is well represented by the obtained forest tree model. The level of R^2 and Prs is above 0.9 while the remaining indices remain relatively low. Therefore, it can be concluded that the forest tree model of T as a function of N_p and DWT is significant.
- *Displacement* Δ: the displacement is really well represented by the obtained forest tree model. The indicators R_2 and Prs are both above 0.95, while the others are low, highlighting the excellent quality of the fitted model. Therefore, it can be concluded that the forest tree model of Δ as a function of N_p and DWT is significant.
- *Installed power* P_B: the installed power is quite well represented by the obtained forest tree model. The level of the R^2 and Prs reported in Table 14 is ranging between 0.70 and 0.88, while the remaining indices are moderately low. Therefore, it can be stated that the forest tree model of P_B as a function of N_p and DWT is moderate significant.
- *Speed* V_s: the speed is well represented by the obtained forest tree model. The R^2 and Prs indices are above 0.70 while the remain-

ing ones are relatively low. Therefore, it can be concluded that the forest tree model of V_s as a function of N_p and DWT is moderate significant.
- *Lane metres* LM: the lane metres is well represented by the obtained forest tree model. The indicators R_2 and Prs are both 0.95, while the others are low, highlighting the good quality of the fitted model. Therefore, it can be concluded that the forest tree model of LM as a function of N_p and DWT is significant.

A.3.4. Regressions as a function of N_p and LM

Hereafter, a detailed analysis variable by variable for the results presented in Section 5.3.4 is presented:

- *Length* L: the length is well represented by the obtained forest tree model. The R^2 and the Prs in Table 15 are above 0.95, while the remaining indicators are low. Therefore, it can be stated that the forest tree model of L as a function of N_p and LM is significant.
- *Breadth* B: the breadth is well represented by the obtained forest tree model. The R^2 and the Prs in Table 15 are above 0.95, while the remaining indicators are low. Therefore, it can be stated that the forest tree model of L as a function of N_p and LM is significant.
- *Depth* D: the depth is not well represented by the obtained forest tree model. The R^2 and Prs are ranging from 0.1 and 0.45, while the remaining indices are relatively high. Therefore it can be concluded that the forest tree model of D as a function of N_p and LM is not significant.
- *Draught* T: the draught is well represented by the obtained forest tree model. The R^2 and Prs in Table 15 are above 0.90, while the remaining coefficients are low. Therefore, it can be concluded that the forest tree model for T as a function of N_p and LM is significant.
- *Displacement* Δ: the displacement is well represented by the obtained forest tree model. The R^2 and the Prs in Table 15 are above 0.95, while the remaining indicators are low. Therefore, it can be stated that the forest tree model of Δ as a function of N_p and LM is significant.
- *Installed power* P_B: the installed power is well represented by the obtained forest tree model. The R^2 and the Prs in Table 15 are above 0.95, while the remaining indicators are low. Therefore, it can be stated that the forest tree model of P_B as a function of N_p and LM is significant.
- *Speed* V_s: the speed is well represented by the obtained forest tree model. The R^2 and Prs indices are both above 0.8 and the remaining coefficients are relatively low. Therefore, it can be concluded that the forest tree model of V_s as a function of N_p and LM is significant.
- *Deadweight* DWT: the deadweight is well represented by the obtained forest tree model. The indicators of Table 15 show values above 0.95 for the Prs and R^2, while the remaining indices are relatively low. Therefore, it can be stated that the forest tree model of DWT as a function of N_p and LM is significant.

A.3.5. Regressions as a function of N_p, V_s and DWT

Hereafter, a detailed analysis variable by variable for the regressions presented in Section 5.3.5 is presented:

- *Length* L: the length is really well represented by the obtained forest tree model. The indicators in Table 16 highlight values above 0.95 for R^2 and Prs and extremely low values for the remaining indicators. Therefore, it can be concluded that the forest tree model of L as a function of N_p, V_s and DWT is significant.
- *Breadth* B: the breadth is really well represented by the obtained forest tree model. The indicators in Table 16 highlight values above 0.95 for R^2 and Prs and extremely low values for the remaining indicators. Therefore, it can be concluded that the forest tree model of B as a function of N_p, V_s and DWT is significant.
- *Depth* D: the depth is quite well captured by the obtained forest tree model. The R^2 and Prs values are ranging from 0.65 to 0.9 while the other indices are moderately low. Therefore, it can be concluded that the forest tree model of D as a function of N_p, V_s and DWT is moderate significant.
- *Draught* T: the draught is really well represented by the obtained forest tree model. The indicators in Table 16 highlight values above 0.95 for R^2 and Prs and extremely low values for the remaining indicators. Therefore, it can be concluded that the forest tree model of T as a function of N_p, V_s and DWT is significant.
- *Displacement* Δ: the displacement is really well represented by the obtained forest tree model. The indicators in Table 16 highlight values above 0.95 for R^2 and Prs and extremely low values for the remaining indicators. Therefore, it can be concluded that the forest tree model of Δ as a function of N_p, V_s and DWT is significant.
- *Installed power* P_B: the installed power is well represented by the obtained forest tree model. The R^2 and Prs coefficients are above 0.9 while the remaining indices are relatively low. Therefore, it can be stated that the forest tree model of P_B as a function of N_p, V_s and DWT is significant.
- *Lane metres* LM: the lane metres is really well represented by the obtained forest tree model. The indicators in Table 16 highlight values above 0.95 for R^2 and Prs and extremely low values for the remaining indicators. Therefore, it can be concluded that the forest tree model of LM as a function of N_p, V_s and DWT is significant.

A.3.6. Regressions as a function of N_p, V_s and LM

In the following, a detailed analysis variable by variable for the regression models presented in Section 5.3.6 is presented:

- *Length* L: the length is really well represented by the obtained forest tree model. The R^2 and Prs shown in Table 17 are both above 0.95 and the other indices are extremely low, highlighting the excellent quality of the fitting model. Therefore, it can be concluded that the forest tree model of L as a function of N_p, V_s and LM is significant.
- *Breadth* B: the breadth is really well represented by the obtained forest tree model. The R^2 and Prs shown in Table 17 are both above 0.95 and the other indices are extremely low, highlighting the excellent quality of the fitting model. Therefore, it can be concluded that the forest tree model of B as a function of N_p, V_s and LM is significant.
- *Depth* D: the depth is not well represented by the obtained forest tree model. The level of R^2 and Prs is ranging from 0.2 and 0.5, while the remaining indices are relatively high. Therefore, it can be concluded that the forest tree model of D as a function of N_p, V_s and LM is not significant.
- *Draught* T: the draught is well represented by the obtained forest tree model. R^2 and Prs indicators are both above 0.90, while the remaining coefficients remain low. Therefore, it can be concluded that the forest tree model of T as a function of N_p, V_s and LM is significant.
- *Displacement* Δ: the displacement is really well represented by the obtained forest tree model. The R^2 and Prs shown in Table 17 are both above 0.95 and the other indices are extremely low, highlighting the excellent quality of the fitting model. Therefore, it can be concluded that the forest tree model of Δ as a function of N_p, V_s and LM is significant.
- *Installed power* P_B: the installed power is really well represented by the obtained forest tree model. The R^2 and Prs shown in Table 17 are both above 0.95 and the other indices are extremely low, highlighting the excellent quality of the fitting model. Therefore, it can be concluded that the forest tree model of P_B as a function of N_p, V_s and LM is significant.
- *Deadweight* DWT: the deadweight is well represented by the obtained forest tree model. R^2 and Prs indicators are both above 0.90, while the remaining coefficients remain low. Therefore, it can be concluded that the forest tree model of DWT as a function of N_p, V_s and LM is significant.

Appendix B. Multiple linear regression coefficients and additional analyses

For each of the multiple linear regressions reported in Section 5 a detailed and extended analysis has been conducted. First of all, the multicollinearity has been checked according to the Variance Inflation Factor (VIF), highlighting no collinearities for all the analysed models. Such a result is due to the fact that the independent variables have been chosen taking into consideration the correlation between variables. In any case, Table B.1 reports the obtained values for the VIF factor. In the Table, the models refer to the nomenclature introduced in Section 6.

Table B.1
VIF values for the different multiple linear regression models.

model	V_s	N_p	DWT	LM	multicol.
mdl-0	1.0188	–	1.10188	–	NO
mdl-1	1.0005	–	–	1.0005	NO
mdl-2	–	1.0030	1.0030	–	NO
mdl-3	–	1.0353	–	1.0353	NO
mdl-4	1.5360	1.5121	1.0459	–	NO
mdl-5	1.5100	1.5626	–	1.0614	NO

For all the regressions, the normality of the residuals has been checked with the Kolmogorov-Smirnov test, giving positive results for all tested cases. Furthermore, the heteroscedasticity has been evaluated according to the Breush-Pagan test, reporting the results in Table B.2. According to the Breush-Pagan test, 19 models out of 46 are affected by moderate heteroscedasticity. In fact, the low values of the T-values of the test suggest that the heteroscedasticity is low and essentially due to the nature of the database data, as no collinearity is detected between the independent variables. In any case, being the objective of the regressions the estimation of the independent variable and not the influence of each parameter on the final regression, the detection of non-constant variance does not require the manipulation of input data to eliminate the problem.

Finally, to ensure the reproducibility of the study, this appendix reports all the coefficients obtained during the regression analysis. The data reports the estimated coefficients together with the associated SE, t-Stud and p-value. The data associated to the model *mlr-0* are reported in Table B.3. Table B.4 reports the data of model *mlr-1*, while Tables B.5 and B.6 report the data of models *mlr-2* and *mlr-3*, respectively. Due to the high number of regression coefficients, the data for model *mlr-4* are split into Tables B.7 and B.8. The same for *mlr-5*, where the data are reported in Tables B.9 and B.10.

Table B.2
Breush-Pagan test results for the detection of heteroscedasticity.

model	T	df	p-val	Heterosced.	model	T	df	p-val	Heterosced.
mlr-0-L	5.3828	2	0.0678	NO	*mlr-1-L*	12.3467	2	0.0021	YES
mlr-0-B	10.1615	2	0.0062	YES	*mlr-1-B*	14.0700	2	0.0009	YES
mlr-0-D	5.8500	2	0.0537	NO	*mlr-1-D*	6.9571	3	0.0733	NO
mlr-0-T	1.9761	2	0.3723	NO	*mlr-1-T*	9.1288	3	0.0276	YES
*mlr-0-*Δ	3.3457	3	0.3413	NO	*mlr-1-*Δ	2.7254	2	0.2560	NO
mlr-0-P_B	2.3386	3	0.5052	NO	*mlr-1-*P_B	1.1446	2	0.5642	NO
mlr-0-N_p	6.5104	2	0.0386	YES	*mlr-1-*N_p	6.0687	2	0.0481	YES
mlr-0-LM	10.6443	1	0.0011	YES	*mlr-1-DWT*	7.9892	2	0.0184	YES
mlr-2-L	7.9495	3	0.0471	YES	*mlr-3-L*	2.4105	2	0.2996	NO
mlr-2-B	4.6506	3	0.1992	NO	*mlr-3-B*	0.7457	2	0.6888	NO
mlr-2-D	0.4600	2	0.7945	NO	*mlr-3-D*	3.7538	3	0.2893	NO
mlr-2-T	2.4983	2	0.2868	NO	*mlr-3-T*	0.0012	1	0.9729	NO
*mlr-2-*Δ	1.6753	2	0.4327	NO	*mlr-3-*Δ	15.8208	2	0.0004	YES
mlr-2-P_B	8.1633	2	0.0169	YES	*mlr-3-*P_B	8.3643	2	0.0153	YES
mlr-2-V_s	3.3598	3	0.3394	NO	*mlr-3-*V_s	19.6624	2	0.0001	YES
mlr-2-LM	5.1130	3	0.1637	NO	*mlr-3-DWT*	25.2297	3	0.0000	YES
mlr-4-L	11.6644	3	0.0086	YES	*mlr-5-L*	5.8250	2	0.0543	NO
mlr-4-B	0.6842	2	0.7103	NO	*mlr-5-B*	7.0013	2	0.0302	YES
mlr-4-D	0.3254	2	0.8499	NO	*mlr-5-D*	0.6558	2	0.7205	NO
mlr-4-T	5.9039	2	0.0522	NO	*mlr-5-T*	0.9564	2	0.6199	NO
*mlr-4-*Δ	2.8179	2	0.2444	NO	*mlr-5-*Δ	15.4236	2	0.0004	YES
mlr-4-P_B	0.4309	3	0.9338	NO	*mlr-5-*P_B	12.7546	2	0.0017	YES
mlr-4-LM	0.1100	3	0.9906	NO	*mlr-5-DWT*	15.0734	2	0.0005	YES

Table B.3
Multiple linear regression coefficients for the regressions as a function of V_s and DWT.

Length L					Breadth B				
Variable	Coefficient	SE	t-Stud	p-value	Variable	Coefficient	SE	t-Stud	p-value
–	−8187.3022	2070.5570	−3.9541	0.00016	–	20.8351	2.1508	9.6868	2.76706e-15
DWT	1.1680	0.3076	3.7966	0.00028	DWT	0.0018	0.0003	4.7626	8.01129e-06
V_s	1074.8750	270.7609	3.9698	0.00015	V_s	−0.1468	0.0823	−1.7825	0.07831
$DWT \cdot V_s$	−0.1498	0.0400	−3.7461	0.00033	DWT^2	−6.6138e-08	2.7762e-08	−2.3822	0.01949
V_s^2	−46.1891	11.7038	−3.9464	0.00017					
$DWT \cdot V_s^2$	0.0064	0.0017	3.7337	0.00035					
V_s^3	0.6583	0.1672	3.9366	0.00017					
$DWT \cdot V_s^3$	−9.1310e-05	2.4481e-05	−3.7297	0.00035					
Depth D					Draught T				
Variable	Coefficient	SE	t-Stud	p-value	Variable	Coefficient	SE	t-Stud	p-value
–	−366.2691	181.3983	−2.0191	0.04686	–	−162.1063	51.8850	−3.1243	0.00251
DWT	0.0270	0.0100	2.7028	0.00841	DWT	0.0249	0.0078	3.1895	0.00206
V_s	45.2096	22.7064	1.9910	0.04993	V_s	21.9180	6.7816	3.2319	0.00181
DWT^2	−2.0861e-06	7.9773e-07	−2.6151	0.01068	DWT^2	−1.6589e-07	6.2569e-08	−2.6513	0.00972
$DWT \cdot V_s$	−0.0011	0.0004	−2.6741	0.00910	$DWT \cdot V_s$	−0.0031	0.0010	−3.1416	0.00238
V_s^2	−1.9025	0.9467	−2.0096	0.04788	V_s^2	−0.9598	0.2940	−3.2640	0.00163
$DWT^2 \cdot V_s$	9.0448e-08	3.4825e-08	2.5971	0.01120	$DWT^2 \cdot V_s$	6.2478e-09	2.7360e-09	2.2835	0.02510
V_s^3	0.0282	0.0131	2.1569	0.03401	$DWT \cdot V_s^2$	0.0001	4.3038e-05	3.1769	0.00214
					V_s^3	0.0139	0.0042	3.3044	0.00144
					$DWT \cdot V_s^3$	−1.9921e-06	6.1579e-07	−3.2351	0.00179
Displacement Δ					Installed power P_B				
Variable	Coefficient	SE	t-Stud	p-value	Variable	Coefficient	SE	t-Stud	p-value
–	4575.8903	1189.9755	3.8453	0.00023	–	−4609516.8483	2250123.2920	−2.0485	0.04382
DWT	2.2089	0.1633	13.5203	6.83835e-23	DWT	61.0966	25.9324	2.3559	0.02095
					V_s	773528.1942	379709.4560	2.0371	0.04498
					$DWT \cdot V_s$	−5.0310	2.2264	−2.2596	0.02659
					V_s^2	−48679.5602	23836.7177	−2.0422	0.04446
					$DWT \cdot V_s^2$	0.1049	0.0477	2.1993	0.03077
					V_s^3	1362.3990	659.8316	2.0647	0.04222
					V_s^4	−14.2491	6.7995	−2.0955	0.03932
Number of passengers N_p					Lane metres LM				
Variable	Coefficient	SE	t-Stud	p-value	Variable	Coefficient	SE	t-Stud	p-value
–	−568697.7025	191361.7437	−2.9718	0.00393	–	38176.9107	18769.3119	2.0340	0.04522
DWT	3.2185	1.4448	2.2275	0.02879	DWT	0.1980	0.0766	2.5831	0.01158
V_s	93973.9812	32296.6494	2.9097	0.00471	V_s	−4924.7857	2380.1122	−2.0691	0.04172
DWT^2	−0.0002	0.0001	−2.1017	0.03879	DWT^2	1.4229e-05	5.4699e-06	2.6013	0.01103
$DWT \cdot V_s$	−0.1402	0.0629	−2.2279	0.02876	V_s^2	210.7398	99.4720	2.1185	0.03718
V_s^2	−5791.2841	2028.6099	−2.8548	0.00551	V_s^3	−2.9784	1.3712	−2.1720	0.03277
$DWT^2 \cdot V_s$	1.0553e-05	5.0523e-06	2.0887	0.03999					
V_s^3	157.4439	56.1648	2.8032	0.00638					
V_s^4	−1.5851	0.5778	−2.7430	0.00754					

Table B.4
Multiple linear regression coefficients for the regressions as a function of V_s and LM.

Length L					Breadth B				
Variable	Coefficient	SE	t-Stud	p-value	Variable	Coefficient	SE	t-Stud	p-value
–	−1775.2835	591.2607	−3.0025	0.00357	–	20.0413	1.0843	18.4825	2.43531e-31
V_s	229.9712	75.1271	3.0610	0.00300	LM	0.0037	0.0008	4.4103	3.03870e-05
LM	0.0799	0.0227	3.5190	0.00071	LM^2	−3.2205e-07	1.5312e-07	−2.1031	0.03844
V_s^2	−9.4565	3.1472	−3.0047	0.00354					
$V_s \cdot LM$	−0.0017	0.0009	−1.8744	0.06452					
LM^2	−3.5418e-06	9.0372e-07	−3.9191	0.00018					
V_s^3	0.1313	0.0433	3.0276	0.00331					
Depth D					Draugth T				
Variable	Coefficient	SE	t-Stud	p-value	Variable	Coefficient	SE	t-Stud	p-value
–	−638.7960	202.4972	−3.1545	0.00228	–	−26.1857	13.7411	−1.9056	0.06028
V_s	75.9464	24.1109	3.1498	0.00231	V_s	3.7748	1.7385	2.1712	0.03287
LM	0.1301	0.0410	3.1739	0.00215	LM	0.0018	0.0002	6.6242	3.70326e-09
V_s^2	−2.8764	0.9508	−3.0250	0.00336	V_s^2	−0.1552	0.0726	−2.1377	0.03559
$V_s \cdot LM$	−0.0109	0.0034	−3.1329	0.00243	LM^2	−5.7143e-07	1.1940e-07	−4.7855	7.67725e-06
LM^2	−2.3350e-06	1.4451e-06	−1.6157	0.11018	V_s^3	0.0021	0.0010	2.1446	0.03501
V_s^3	0.0350	0.0124	2.8096	0.00626	LM^3	5.3853e-11	1.4361e-11	3.7497	0.00033
$V_s^2 \cdot LM$	0.0002	7.5160e-05	3.1452	0.00234					
LM^3	3.6553e-10	1.7478e-10	2.0913	0.03975					
Displacement Δ					Installed power P_B				
Variable	Coefficient	SE	t-Stud	p-value	Variable	Coefficient	SE	t-Stud	p-value
–	−413093.9495	179638.5352	−2.2995	0.02401	–	383892.1292	356943.3338	1.0754	0.28550
V_s	54087.1509	22763.5942	2.3760	0.01983	V_s	−36682.5534	31156.2655	−1.1773	0.24267
LM	4.6392	0.4570	10.1506	3.76407e-16	LM	−443.5873	363.2971	−1.2210	0.22580
V_s^2	−2292.3352	951.3230	−2.4096	0.01820	V_s^2	886.3373	676.1951	1.3107	0.19383
V_s^3	32.1251	13.1157	2.4493	0.01644	$V_s \cdot LM$	40.6619	31.4130	1.2944	0.19938
					LM^2	0.1486	0.0859	1.7300	0.08763
					$V_s^2 \cdot LM$	−0.8910	0.6746	−1.3207	0.19048
					$V_s \cdot LM^2$	−0.0134	0.0074	−1.8094	0.07429
					LM^3	7.9426e-07	3.3737e-07	2.3542	0.02111
					$V_s^2 \cdot LM^2$	0.0002	0.0001	1.8245	0.07194
Number of passengers N_p					Deadweight DWT				
Variable	Coefficient	SE	t-Stud	p-value	Variable	Coefficient	SE	t-Stud	p-value
–	−69616.1503	29945.0969	−2.3247	0.02261	–	−139176.4782	39713.2823	−3.5045	0.00074
V_s	8174.6295	3585.9891	2.2796	0.02529	V_s	17899.8479	5029.4182	3.5590	0.00062
LM	10.5791	5.7847	1.8287	0.07115	LM	3.3838	0.3337	10.1374	4.56379e-16
V_s^2	−310.0921	142.2733	−2.1795	0.03223	V_s^2	−755.4373	210.1498	−3.5947	0.00055
$V_s \cdot LM$	−0.9051	0.4957	−1.8255	0.07164	LM^2	−0.0002	6.0235e-05	−4.0923	0.00010
V_s^3	3.8743	1.8759	2.0652	0.04213	V_s^3	10.5151	2.8967	3.6299	0.00049
$V_s^2 \cdot LM$	0.0190	0.0106	1.7974	0.07603					

Table B.5
Multiple linear regression coefficients for the regressions as a function of N_p and DWT.

Length L					Breadth B				
Variable	Coefficient	SE	t-Stud	p-value	Variable	Coefficient	SE	t-Stud	p-value
–	0.0000	0.0000	NaN	NaN	–	0.0000	0.0000	NaN	NaN
DWT	0.0424	0.0062	6.8097	2.106419e-09	DWT	0.0051	0.0009	5.3408	9.30877e-07
N_p	0.1990	0.0362	5.4888	5.26883e-07	N_p	0.0521	0.0066	7.7975	2.68494e-11
DWT^2	−3.4211e-06	1.0728e-06	−3.1887	0.00208	DWT^2	−5.3362e-07	1.7201e-07	−3.1022	0.00269
$DWT \cdot N_p$	−4.0627e-05	9.2934e-06	−4.3716	3.91306e-05	$DWT \cdot N_p$	−3.2475e-06	9.2674e-07	−3.5042	0.00077
N_p^2	−0.0001	3.7074e-05	−3.7605	0.00033	N_p^2	−5.7167e-05	8.5640e-06	−6.6752	3.56893e-09
DWT^3	1.1761e-10	4.8908e-11	2.4047	0.01865	DWT^3	2.5160e-11	8.3141e-12	3.0261	0.00337
$DWT^2 \cdot N_p$	1.8098e-09	1.0272e-09	1.7617	0.08218	$DWT \cdot N_p^2$	2.98745e-09	7.7290e-10	3.8651	0.00023
$DWT \cdot N_p^2$	2.7581e-08	5.4869e-09	5.0267	3.29609e-06	N_p^3	2.2834e-08	4.2005e-09	5.4360	6.34874e-07
N_p^3	2.3525e-08	1.0623e-08	2.2144	0.02983	$DWT \cdot N_p^3$	−7.2679e-13	1.8456e-13	−3.9379	0.00018
$DWT^2 \cdot N_p^2$	−9.8614e-13	4.4465e-13	−2.2177	0.02959	N_p^4	−2.8715e-12	6.9591e-13	−4.1262	9.34870e-05
$DWT \cdot N_p^3$	−3.4986e-12	1.5295e-12	−2.2874	0.02499					
Depth D					Draught T				
Variable	Coefficient	SE	t-Stud	p-value	Variable	Coefficient	SE	t-Stud	p-value
–	0.0000	0.0000	NaN	NaN	–	0.0000	0.0000	NaN	NaN
DWT	0.0013	0.0028	0.4564	0.64936	DWT	0.0014	0.0001	10.7230	8.53826e-17
N_p	0.0240	0.0106	2.2576	0.02687	N_p	0.0096	0.0009	10.2542	6.37546e-16
DWT^2	4.2867e-07	6.24681e-07	0.6862	0.49468	DWT^2	−1.4349e-07	2.3292e-08	−6.1606	3.31321e-08
$DWT \cdot N_p$	−4.7399e-06	3.8997e-06	−1.2154	0.22800	$DWT \cdot N_p$	−7.8412e-07	1.1635e-07	−6.7388	2.8544e-09
N_p^2	−1.2198e-05	8.6696e-06	−1.4069	0.16356	N_p^2	−8.2708e-06	1.1657e-06	−7.0946	6.16840e-10
DWT^3	−4.1042e-11	3.4456e-11	−1.1911	0.23734	DWT^3	6.1491e-12	1.0511e-12	5.8500	1.20612e-07
$DWT^2 \cdot N_p$	−6.1674e-10	7.4076e-10	−0.8325	0.40773	$DWT^2 \cdot N_p$	−1.0964e-11	6.1662e-12	−1.7781	0.07943
$DWT \cdot N_p^2$	5.8693e-09	2.0313e-09	2.8894	0.00504	$DWT \cdot N_p^2$	6.2342e-10	1.1040e-10	5.6465	2.7791e-07
N_p^3	−2.2370e-09	1.1756e-09	−1.9029	0.06088	N_p^3	2.5596e-09	5.3450e-10	4.7887	8.24918e-06
$DWT^3 \cdot N_p$	7.2299e-14	4.2100e-14	1.7173	0.09004	$DWT \cdot N_p^3$	−1.1684e-13	2.5210e-14	−4.6346	1.4778e-05
$DWT^2 \cdot N_p^2$	−3.5422e-13	1.3807e-13	−2.56536	0.01230	N_p^4	−2.5755e-13	8.7335e-14	−2.9490	0.00424
Displacement Δ					Installed power P_B				
Variable	Coefficient	SE	t-Stud	p-value	Variable	Coefficient	SE	t-Stud	p-value
–	0.0000	0.0000	NaN	NaN	–	1935.3044	14327.4967	0.1350	0.89289
DWT	2.5607	1.1650	2.1979	0.03099	DWT	19.6098	6.1037	3.2127	0.00190
N_p	22.3941	7.9726	2.8088	0.00631	N_p	−46.7003	21.4432	−2.1778	0.03240
DWT^2	−9.5414e-05	0.0001	−0.9409	0.34972	DWT^2	−0.0030	0.0010	−2.9217	0.00453
$DWT \cdot N_p$	−0.0029	0.0024	−1.2100	0.23002	$DWT \cdot N_p$	0.0017	0.0009	1.9285	0.05737
N_p^2	−0.0290	0.0123	−2.3577	0.02095	N_p^2	0.0349	0.01469	2.3808	0.01967
$DWT^2 \cdot N_p$	3.5180e-07	1.9729e-07	1.7831	0.07855	DWT^3	1.32212e-07	5.0178e-08	2.6348	0.01012
$DWT \cdot N_p^2$	1.8048e-06	1.2324e-06	1.4645	0.14717	N_p^3	−7.4734e-06	3.0567e-06	−2.4448	0.01671
N_p^3	1.3768e-05	6.60487e-06	2.0845	0.04043					
$DWT^2 \cdot N_p^2$	−1.7979e-10	9.0843e-11	−1.9792	0.05141					
N_p^4	−2.3335e-09	1.0450e-09	−2.2330	0.0284					
Speed V_s					Lane metres LM				
Variable	Coefficient	SE	t-Stud	p-value	Variable	Coefficient	SE	t-Stud	p-value
–	16.9947	4.1067	4.1382	8.67985e-05	–	0.0000	0.0000	NaN	NaN
DWT	0.0001	0.0011	0.1338	0.89385	DWT	0.3309	0.3725	0.8882	0.37728
N_p	0.0098	0.0052	1.8681	0.06544	N_p	−0.7783	1.8601	−0.4184	0.67682
DWT^2	5.2775e-08	7.8738e-08	0.6702	0.50464	DWT^2	4.3521e-05	7.2359e-05	0.6014	0.54936
$DWT \cdot N_p$	1.3210e-08	1.1041e-06	0.0119	0.9904	$DWT \cdot N_p$	0.0002	0.0003	0.7714	0.44291
N_p^2	−4.6552e-06	1.8985e-06	−2.4519	0.01641	N_p^2	0.0009	0.0023	0.4019	0.68890
$DWT^2 \cdot N_p$	−1.2528e-10	7.1655e-11	−1.7483	0.08428	DWT^3	−2.7498e-09	3.5299e-09	−0.7790	0.43846
$DWT \cdot N_p^2$	6.0345e-10	2.6765e-10	2.2545	0.02693	$DWT^2 \cdot N_p$	−1.7182e-07	8.3713e-08	−2.0525	0.04365
					$DWT \cdot N_p^2$	4.4389e-07	2.2515e-07	1.9715	0.05239
					N_p^3	−1.2255e-06	9.8973e-07	−1.2383	0.21951
					$DWT^3 \cdot N_p$	9.9231e-12	4.23459e-12	2.3433	0.02179
					$DWT \cdot N_p^3$	−9.2669e-11	4.7878e-11	−1.9354	0.05674
					N_p^4	2.9524e-10	1.5725e-10	1.8774	0.06439

Table B.6
Multiple linear regression coefficients for the regressions as a function of N_p and LM.

Length L					Breadth B				
Variable	Coefficient	SE	t-Stud	p-value	Variable	Coefficient	SE	t-Stud	p-value
–	0.0000	0.0000	NaN	NaN	–	0.0000	0.0000	NaN	NaN
N_p	0.2251	0.0358	6.28822	1.86056e-08	N_p	0.0542	0.0063	8.5936	8.8362e-13
LM	0.0760	0.0133	5.6887	2.26940e-07	LM	0.0143	0.0038	3.7183	0.00038
N_p^2	−0.0002	5.4344e-05	−3.7428	0.00035	N_p^2	−6.1150e-05	7.8581e-06	−7.7817	3.1017e-11
$N_p \cdot LM$	−4.8990e-05	2.2307e-05	−2.1961	0.03112	$N_p \cdot LM$	−4.9095e-06	2.6198e-06	−1.8740	0.06482
LM^2	−9.7964e-06	2.9638e-06	−3.3053	0.00144	LM^2	−6.6978e-06	3.1589e-06	−2.1203	0.03728
N_p^3	8.6450e-08	2.6856e-08	3.2190	0.00189	N_p^3	2.6565e-08	3.7495e-09	7.0851	6.42731e-10
$N_p^2 \cdot LM$	1.9584e-08	9.8592e-09	1.9863	0.05059	$N_p^2 \cdot LM$	4.7603e-09	2.1993e-09	2.1644	0.03361
$N_p \cdot LM^2$	1.1503e-08	5.0394e-09	2.2826	0.02524	LM^3	1.6518e-09	8.4213e-10	1.9615	0.05352
N_p^4	−1.3499e-11	4.2937e-12	−3.1439	0.00237	N_p^4	−3.7341e-12	6.1862e-13	−6.0362	5.5733e-08
$N_p^2 \cdot LM^2$	−5.5845e-12	2.2400e-12	−2.4929	0.01484	$N_p^3 \cdot LM$	−1.2779e-12	5.3256e-13	−2.3996	0.01889
					LM^4	−1.4013e-13	7.3902e-14	−1.8961	0.06178
Depth D					Draught T				
Variable	Coefficient	SE	t-Stud	p-value	Variable	Coefficient	SE	t-Stud	p-value
–	12.2166	3.1744	3.8484	0.00023	–	0.0000	0.0000	NaN	NaN
N_p	−0.0117	0.0076	−1.5462	0.12590	N_p	0.0091	0.0008	10.8096	4.9888e-17
LM	0.0009	0.0004	2.3481	0.02127	LM	0.0049	0.0005	8.6883	5.2792e-13
N_p^2	1.0968e-05	5.3886e-06	2.0354	0.04503	N_p^2	−8.2293e-06	1.1345e-06	−7.2532	2.91741e-10
N_p^3	−2.3866e-09	1.1233e-09	−2.1246	0.0366	$N_p \cdot LM$	−9.3302e-07	1.5549e-07	−6.0001	6.2484e-08
					LM^2	−2.1754e-06	4.4594e-07	−4.8783	5.7463e-06
					N_p^2	3.2886e-09	5.7113e-10	5.7580	1.70660e-07
					$N_p^2 \cdot LM$	3.0364e-10	5.7819e-11	5.2515	1.3299e-06
					LM^3	4.4360e-10	1.1964e-10	3.7077	0.00039
					N_p^4	−4.8149e-13	9.3713e-14	−5.1379	2.0868e-06
					LM^4	−3.1300e-14	1.0547e-14	−2.9675	0.00401
Displacement Δ					Installed power P_B				
Variable	Coefficient	SE	t-Stud	p-value	Variable	Coefficient	SE	t-Stud	p-value
–	4091.2998	3607.8021	1.1340	0.26017	–	3557.8698	20515.3812	0.1734	0.86277
N_p	−6.8466	5.1563	−1.3278	0.18801	N_p	−20.9432	31.0536	−0.6744	0.50206
LM	6.0016	1.6956	3.5393	0.00067	LM	33.0620	15.4903	2.1343	0.03599
N_p^2	0.0051	0.0017	2.8715	0.00522	N_p^2	0.0252	0.0164	1.5363	0.12855
$N_p \cdot LM$	0.0046	0.0020	2.2405	0.0278	$N_p \cdot LM$	0.0016	0.0128	0.1264	0.89973
LM^2	−0.0004	0.0002	−2.2643	0.02625	LM^2	−0.0130	0.0042	−3.0711	0.00294
$N_p^2 \cdot LM$	−2.2974e-06	7.6880e-07	−2.9882	0.00372	N_p^3	−7.8614e-06	3.0120e-06	−2.6099	0.01087
					$N_p^2 \cdot LM$	4.9180e-06	2.8294e-06	1.7381	0.08618
					$N_p \cdot LM^2$	−2.9521e-06	1.7009e-06	−1.7355	0.08664
					LM^3	1.8380e-06	4.7334e-07	3.8832	0.00021
Speed V_s					Deadweight DWT				
Variable	Coefficient	SE	t-Stud	p-value	Variable	Coefficient	SE	t-Stud	p-value
–	8.2886	3.3555	2.4701	0.01566	–	−1859.7805	840.9318	−2.2115	0.02981
N_p	0.0164	0.0042	3.8905	0.0002	N_p	2.4291	0.7818	3.1069	0.00260
LM	0.0113	0.0026	4.2190	6.49104e-05	LM	4.1475	0.4265	9.7229	2.97319e-15
N_p^2	−4.9160e-06	1.4856e-06	−3.3089	0.00141	N_p^2	−0.0003	0.0002	−1.9568	0.05381
$N_p \cdot LM$	−6.0296e-06	1.7473e-06	−3.4508	0.00090	$N_p \cdot LM$	−0.0003	0.0001	−2.2050	0.03028
LM^2	−3.3602e-06	9.2730e-07	−3.6236	0.00051	LM^2	−0.0002	6.0612e-05	−4.7259	9.51622e-06
$N_p^2 \cdot LM$	2.0769e-09	6.4429e-10	3.2235	0.00184					
LM^3	3.8757e-10	1.1381e-10	3.4052	0.00104					

Table B.7
Multiple linear regression coefficients for the regressions as a function of N_p, V_s and DWT. (Part I).

Length L					Breadth B				
Variable	Coefficient	SE	t-Stud	p-value	Variable	Coefficient	SE	t-Stud	p-value
–	−1248.0528	583.9765	−2.1371	0.03603	–	0.0000	0.0000	NaN	NaN
DWT	0.2317	0.0998	2.3201	0.02321	DWT	0.1335	0.0372	3.5876	0.00064
V_s	112.7001	53.4170	2.1098	0.03840	N_p	−0.4889	0.1818	−2.6887	0.00913
N_p	1.0508	0.5128	2.0489	0.04416	DWT^2	−1.0009e-05	3.0700e-06	−3.2602	0.00178
DWT^2	5.6750e-07	3.2921e-07	1.7238	0.08909	$DWT \cdot V_s$	−0.0136	0.0037	−3.6242	0.00057
$DWT \cdot V_s$	−0.0195	0.0090	2.1629	0.03391	V_s^2	−0.2429	0.1827	−1.3295	0.18840
V_s^2	−2.2346	1.2230	−1.8270	0.07188	$DWT \cdot N_p$	4.1664e-05	1.1454e-05	3.6374	0.00055
$DWT \cdot N_p$	−0.0001	8.3226e-05	−1.8243	0.07230	$V_s \cdot N_p$	0.0628	0.0226	2.7776	0.00717
$V_s \cdot N_p$	−0.0909	0.0415	−2.1891	0.03187	N_p^2	−0.0001	4.5336e-05	−2.2734	0.02636
N_p^2	2.4602e-05	1.0616e-05	2.3173	0.02337	DWT^3	1.7460e-1	8.6782e-12	2.0120	0.04842
$DWT \cdot V_s^2$	0.0004	0.0002	1.9811	0.05144	$DWT^2 \cdot V_s$	1.0217e-06	3.1237e-07	3.2708	0.00172
$DWT^2 \cdot N_p$	−6.5618e-10	3.2186e-10	−2.0387	0.04520	$DWT \cdot V_s^2$	0.0003	9.3392e-05	3.6989	0.00045
$DWT \cdot N_P \cdot V_s$	1.3879e-05	6.7397e-06	2.0592	0.04313	V_s^3	0.0371	0.0162	2.2929	0.02514
$V_s^2 \cdot N_p$	0.0018	0.0008	2.1137	0.03804	$DWT^2 \cdot N_p$	−3.1157e-09	8.8509e-10	−3.5201	0.00080
$DWT \cdot N_p^2$	−3.0026e-09	1.6170e-09	−1.8569	0.06746	$DWT \cdot V_s \cdot N_p$	−1.7488e-06	4.8972e-07	−3.5710	0.00068
$DWT \cdot V_s^2 \cdot N_p$	−2.8373e-07	1.3870e-07	−2.0455	0.04450	$V_s^2 \cdot N_p$	−0.0028	0.0010	−2.8616	0.00568
					$V_s \cdot N_p^2$	8.4363e-06	3.6426e-06	2.3159	0.02377
					$DWT^2 \cdot V_s^2$	−2.6081e-08	7.6868e-09	−3.3930	0.00119
					V_s^4	−0.0010	0.0003	−2.9208	0.00481
					$DWT^2 \cdot V_s \cdot N_p$	1.3187e-10	3.7910e-11	3.4786	0.00091
					$V_s^3 \cdot N_p$	4.5440e-05	1.5377e-05	2.9549	0.00437
					$V_s^2 \cdot N_p^2$	−1.6896e-07	7.22182e-08	−2.3396	0.02243

Depth D					Draugth T				
Variable	Coefficient	SE	t-Stud	p-value	Variable	Coefficient	SE	t-Stud	p-value
–	177.5708	145.7283	1.2185	0.22690	–	57.3342	23.0144	2.4912	0.01503
DWT	−0.0315	0.0244	−1.2902	0.20099	DWT	0.0081	0.0026	3.1262	0.00255
V_s	−19.0405	14.3180	−1.3298	0.18765	V_s	−8.7213	3.0248	−2.8832	0.00518
N_p	0.1592	0.0718	2.2168	0.02970	N_p	−0.0344	0.0211	−1.6327	0.10688
$DWT \cdot V_s$	0.0037	0.0023	1.5513	0.12507	DWT^2	−1.9770e-08	3.0964e-09	−6.3850	1.47247e-08
V_s^2	0.5121	0.3475	1.4733	0.14488	$DWT \cdot V_s$	−0.0006	0.0002	−2.7449	0.00763
$DWT \cdot N_p$	−1.4833e-05	5.6984e-06	−2.6030	0.01115	V_s^2	0.4407	0.1345	3.2755	0.00162
$V_s \cdot N_p$	−0.0118	0.0052	−2.2558	0.02703	$DWT \cdot N_p$	−1.1122e-05	2.4857e-06	−4.4744	2.80507e-05
N_p^2	2.2784e-05	8.1423e-06	2.7982	0.00654	$V_s \cdot N_p$	0.0070	0.0027	2.5773	0.01200
$DWT \cdot V_s^2$	−0.0001	5.8356e-05	−1.7508	0.08411	$DWT \cdot V_s^2$	1.2681e-05	4.9782e-06	2.5473	0.01299
$DWT \cdot V_s \cdot N_p$	6.5332e-07	2.4660e-07	2.6493	0.00985	V_s^3	−0.0070	0.0020	−3.4776	0.00086
$V_s^2 \cdot N_p$	0.0002	0.0001	2.0959	0.03950	$DWT \cdot V_s \cdot N_p$	8.8337e-07	2.0135e-07	4.3871	3.85525e-05
$V_s \cdot N_p^2$	−9.5588e-07	3.3296e-07	−2.8708	0.00533	$V_s^2 \cdot N_p$	−0.0003	0.0001	−3.2533	0.00173
					$DWT \cdot V_s^2 \cdot N_p$	−1.7427e-08	4.0468e-09	−4.3063	5.15886e-05
					$V_s^3 \cdot N_p$	6.6089e-06	1.78400e-06	3.7044	0.00041

Table B.8
Multiple linear regression coefficients for the regressions as a function of N_p, V_s and DWT. (Part II).

Displacement Δ					Installed power P_B				
Variable	Coefficient	SE	t-Stud	p-value	Variable	Coefficient	SE	t-Stud	p-value
–	0.0000	0.0000	NaN	NaN	–	−510999.3849	348439.1089	−1.4665	0.14679
DWT	107.5351	39.5673	2.7177	0.00835	DWT	125.7389	58.5647	2.1470	0.03511
V_s	−17646.7134	7681.4392	−2.2973	0.02473	V_s	36970.7252	31854.1045	1.1606	0.24957
N_p	−267.5379	211.6793	−1.2638	0.21065	N_p	807.7937	317.5934	2.5434	0.01309
DWT^2	0.0002	8.2464e-05	3.2638	0.00173	$DWT \cdot V_s$	−9.5374	5.2753	−1.8079	0.0747
$DWT \cdot V_s$	−9.7659	3.6770	−2.6559	0.00987	V_s^2	−567.5217	726.5743	−0.7810	0.43727
V_s^2	1652.0369	710.8934	2.3238	0.02317	$DWT \cdot N_p$	−0.1495	0.0511	−2.9206	0.00464
$DWT \cdot N_p$	−0.0902	0.0338	−2.6688	0.00953	$V_s \cdot N_p$	−66.5726	25.9318	−2.5672	0.01230
$V_s \cdot N_p$	46.2768	26.6628	1.7356	0.08722	N_p^2	0.0269	0.0093	2.8756	0.00528
N_p^2	0.0599	0.0300	1.9974	0.04983	$DWT \cdot V_s^2$	0.1756	0.1192	1.4733	0.14495
$DWT \cdot V_s^2$	0.2172	0.0852	2.5473	0.01315	$DWT \cdot V_s \cdot N_p$	0.0118	0.0041	2.8645	0.00544
V_s^3	−37.0994	16.5859	−2.2368	0.02863	$V_s^2 \cdot N_p$	1.2828	0.5209	2.4624	0.01615
$DWT^2 \cdot N_p$	−1.9174e-07	8.1025e-08	−2.3664	0.02085	N_p^3	−5.4266e-06	1.9894e-06	−2.7276	0.00798
$DWT \cdot V_s \cdot N_p$	0.0087	0.0030	2.8550	0.00572	$DWT \cdot V_s^2 \cdot N_p$	−0.0002	8.3388e-05	−2.7536	0.00743
$V_s^2 \cdot N_p$	−2.7212	1.2374	−2.1990	0.03133					
$DWT \cdot N_p^2$	−9.2105e-06	4.7822e-06	−1.9259	0.05835					
$V_s \cdot N_p^2$	−0.0025	0.0012	−1.9391	0.05669					
$DWT \cdot V_s^2 \cdot N_p$	−0.0002	7.1995e-05	−2.8367	0.00602					
$V_s^3 \cdot N_p$	0.0517	0.0208	2.4843	0.01548					
$DWT \cdot V_s \cdot N_p^2$	3.9997e-07	2.0622e-07	1.9395	0.05664					
Lane metres LM									
Variable	Coefficient	SE	t-Stud	p-value					
–	467.4116	259.9962	1.7977	0.07585					
DWT	0.1748	0.0748	2.3343	0.02199					
N_p	−0.1781	0.0701	−2.5391	0.01297					
DWT^2	1.5065e-05	5.2970e-06	2.8441	0.00560					

Table B.9
Multiple linear regression coefficients for the regressions as a function of N_p, V_s and LM. (Part I).

Length L					Breadth B				
Variable	Coefficient	SE	t-Stud	p-value	Variable	Coefficient	SE	t-Stud	p-value
–	−999.4391	427.7201	−2.3366	0.02220	–	0.0000	0.0000	NaN	NaN
V_s	129.2679	53.5539	2.4137	0.01829	V_s	2.6609	1.1470	2.3198	0.02335
N_p	0.1163	0.0409	2.8409	0.00582	N_p	0.0004	0.0184	0.0258	0.97941
LM	0.0965	0.0310	3.1045	0.0027	LM	0.0300	0.0265	1.1348	0.26042
V_s^2	−5.2761	2.2230	−2.3734	0.02025	V_s^2	−0.0786	0.0505	−1.5548	0.12462
$V_s \cdot N_p$	−0.0083	0.0019	−4.3303	4.66899e-05	$V_s \cdot N_p$	−0.0003	0.0007	−0.4038	0.68757
N_p^2	3.21996e-05	6.2111e-06	5.1841	1.86343e-06	N_p^2	4.0984e-06	1.1323e-06	3.6194	0.00056
$V_s \cdot LM$	−0.0038	0.0015	−2.5423	0.01313	$V_s \cdot LM$	−0.0054	0.0022	−2.4546	0.01666
$N_p \cdot LM$	−1.0997e-05	2.0364e-05	−0.5400	0.5908	$N_p \cdot LM$	2.1301e-05	1.6236e-05	1.3119	0.19393
LM^2	6.0210e-07	2.1617e-06	0.2785	0.78139	LM^2	8.9754e-06	1.4309e-05	0.6272	0.53260
V_s^3	0.0785	0.0305	2.5701	0.01220	$V_s^2 \cdot LM$	0.0001	6.4145e-05	2.9085	0.00490
$V_s \cdot N_p \cdot LM$	2.6275e-06	9.2797e-07	2.8314	0.00598	$V_s \cdot N_p \cdot LM$	−7.6179e-07	6.9242e-07	−1.1001	0.27512
$N_p^2 \cdot LM$	−1.4983e-08	2.8829e-09	−5.1973	1.76993e-06	$N_p^2 \cdot LM$	−1.6162e-09	5.4390e-10	−2.9714	0.00409
$N_p \cdot LM^2$	−3.9113e-09	1.7848e-09	−2.1914	0.03160	$V_s \cdot LM^2$	8.3866e-07	1.0342e-06	0.8108	0.42026
					$N_P \cdot LM^2$	−7.0312e-09	3.6379e-09	−1.9327	0.05743
					LM^3	−5.2223e-09	2.1581e-09	−2.4197	0.01820
					$V_s^2 \cdot LM^2$	−5.5853e-08	2.5863e-08	−2.1595	0.03433
					$V_s \cdot N_p \cdot LM^2$	3.0513e-10	1.5745e-10	1.9379	0.05678
					$V_s \cdot LM^3$	2.3332e-10	9.5224e-11	2.4502	0.01685
Depth D					Draugth T				
Variable	Coefficient	SE	t-Stud	p-value	Variable	Coefficient	SE	t-Stud	p-value
–	−74.0202	56.5452	−1.3090	0.19462	–	17.6184	7.7752	2.2659	0.02641
V_s	4.5163	4.6181	0.9779	0.33132	V_s	−1.4558	0.7637	−1.9061	0.06056
N_p	0.1323	0.0471	2.8090	0.00637	N_p	0.0059	0.0020	2.8693	0.00537
LM	0.0401	0.0123	3.2590	0.00169	LM	−0.0052	0.0040	−1.2914	0.20060
V_s^2	−0.0299	0.0952	−0.3140	0.75438	V_s^2	0.0362	0.0186	1.9465	0.05543
$V_s \cdot N_p$	−0.0098	0.0037	−2.6353	0.01025	$V_s \cdot N_p$	−0.0001	8.6028e-05	−2.1628	0.03382
N_p^2	1.4399e-05	5.2151e-06	2.7610	0.00728	N_p^2	−7.2915e-07	3.8381e-07	−1.8997	0.06141
$V_s \cdot LM$	−0.0015	0.0005	−3.0372	0.00331	$V_s \cdot LM$	0.0007	0.0003	1.7932	0.07707
$N_p \cdot LM$	−2.6425e-05	7.0337e-06	−3.7569	0.00034	$N_p \cdot LM$	−2.1139e-06	9.9060e-07	−2.1340	0.03619
LM^2	−2.5089e-06	1.3142e-06	−1.9090	0.06018	LM^2	−5.3399e-07	9.8554e-08	−5.4182	7.38737e-07
$V_s^2 \cdot N_p$	0.0001	7.2835e-05	2.0088	0.04825	N_p^3	1.6564e-10	8.0143e-11	2.0668	0.04229
N_p^3	−3.1535e-09	1.0929e-09	−2.8853	0.00513	$V_s^2 \cdot LM$	−1.6956e-05	9.6655e-06	−1.7543	0.08357
$V_s \cdot N_p \cdot LM$	1.1468e-06	3.0142e-07	3.8049	0.00029	$V_s \cdot N_p \cdot LM$	8.2974e-08	4.3110e-08	1.9246	0.05816
LM^3	3.6621e-10	1.5793e-10	2.3186	0.02321	LM^3	4.7570e-11	1.1864e-11	4.0093	0.00014

Table B.10
Multiple linear regression coefficients for the regressions as a function of N_p, V_s and LM. (Part II).

Displacement Δ					Installed power P_B				
Variable	Coefficient	SE	t-Stud	p-value	Variable	Coefficient	SE	t-Stud	p-value
–	35954.3946	24737.9753	1.4534	0.1502	–	0.0000	0.0000	NaN	NaN
V_s	−5111.7692	1874.8323	−2.7265	0.00796	V_s	5551.8294	4522.5810	1.2275	0.22377
N_p	43.6079	11.3252	3.8504	0.00024	N_p	−75.2072	74.4374	−1.0103	0.31586
LM	20.7648	8.6139	2.4106	0.01837	LM	−228.7228	87.0084	−2.6287	0.01055
V_s^2	164.1776	39.1412	4.1944	7.41500e-05	V_s^2	−184.6708	197.8458	−0.9334	0.35386
$V_s \cdot N_p$	−2.3184	0.5165	−4.4882	2.55006e-05	$V_s \cdot N_p$	1.3980	3.1813	0.4394	0.66172
N_p^2	0.0068	0.0016	4.1092	0.00010	N_p^2	0.0270	0.0085	3.1823	0.00219
$V_s \cdot LM$	−0.6782	0.3921	−1.7296	0.0878	$V_s \cdot LM$	6.3486	3.6855	1.7225	0.08944
$N_p \cdot LM$	−0.0085	0.0052	−1.6101	0.1115	$N_p \cdot LM$	0.0682	0.0628	1.0858	0.28130
LM^2	−0.0003	0.0001	−1.8839	0.06344	LM^2	0.1834	0.0524	3.5011	0.00081
$V_s \cdot N_p \cdot LM$	0.0005	0.0002	2.3610	0.02082	N_p^3	−5.5655e-06	1.7758e-06	−3.1339	0.00253
$N_p^2 \cdot LM$	−2.5469e-06	7.5462e-07	−3.3750	0.00117	$V_s^2 \cdot LM$	0.1867	0.1053	1.7732	0.08059
					$V_s \cdot N_p \cdot LM$	−0.0026	0.0026	−0.9815	0.32973
					$V_s \cdot LM^2$	−0.0083	0.0023	−3.5980	0.00059
					$N_p \cdot LM^2$	−2.1248e-05	1.2374e-05	−1.7171	0.09043
					LM^3	−2.5623e-05	8.1597e-06	−3.1402	0.00248
					$V_s \cdot N_p \cdot LM^2$	8.7889e-07	5.2465e-07	1.6751	0.09842
					$V_s \cdot LM^3$	1.1742e-06	3.6185e-07	3.2451	0.00181

Deadweight DWT				
Variable	Coefficient	SE	t-Stud	p-value
–	−118296.1464	39443.1036	−2.9991	0.00367
V_s	14773.2463	4973.0715	2.9706	0.00399
N_p	−0.5592	1.6751	−0.3338	0.73943
LM	14.6333	5.7554	2.5425	0.01306
V_s^2	−622.3221	207.3525	−3.0012	0.00364
N_p^2	0.0006	0.0005	1.1252	0.26408
$V_s \cdot LM$	−0.4904	0.2580	−1.9008	0.06116
$N_p \cdot LM$	0.0009	0.0007	1.3673	0.17559
LM^2	−0.0034	0.0015	−2.3305	0.02246
V_s^3	8.8478	2.8612	3.0922	0.00278
$N_p^2 \cdot LM$	−5.1310e-07	2.6285e-07	−1.9520	0.05466
$V_s \cdot LM^2$	0.0001	6.5781e-05	2.1417	0.03545

References

Abramowski, T., Cepowski, T., Zvolensky, 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.

Andrews, D. J., 1998. A comprehensive methodology for the design of ships (and other complex systems). Proc. Math. Phys. Eng. Sci. 454 (1968), 187–211.

Asrol, M., Papilo, P., Gunawan, F. E., 2021. Support vector machine with k-fold validation to improve the industry's sustainability performance classification. Procedia Comput. Sci. 179, 854–862.

Caprace, J. D., Rigo, P., 2011. Ship complexity assessment at the concept design stage. J. Mar Sci Technol 16, 68–75.

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.

Clarksons, 2024. Shipping and trade data analysis and research services. on-line.

Clausen, H. B., Lutzen, M., Hansen, A. F., Bjorneboe, N., 2001. Bayesian and neural networks for preliminary ship design. Mar. Technol. 38 (4), 268–277.

Ekinci, S., Celabi, U. B., Bal, M., Amasyali, M. F., Boyaci, U. K., 2011. Predictions of oil/chemical tanker main design parameters using computational intelligence techniques. Appl. Soft Comput. 11 (2), 2356–2366.

Ferry-site, 2024. "Official website", available at: on-line. http://www.ferry-site.dk.

Fisher, R. A., 1922. The goodness of fit of regression formulae, and the distribution of regression coefficients. J. R. Stat. Soc. 85 (4) (4), 597–612.

Freedman, D. A., 2009. Statistical Models: Theory and Practice. Cambridge University Press.

Friis, A. M., Andersen, P., Jensen, J. J., 2002. Ship Design (Part I and II). Technical University of Denmark.

Grubišić, I., Begović, E., 2001. Multi-attribute concept design model of the adriatic type of fishing vessel. Brodogradnja 49 (1), 39–54.

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

Ho, H. T., 1998. The random subspace method for constructing decision forests. IEEE Trans. Pattern Anal. Mach. Intell. 20 (8), 832–844.

Kalokairinos, E., Mavroeidis, T., Radou, G., Zachariou, Z., 2005. Regression Analysis of Basic Ship Design Values for Merchant Ships. Master's thesis. National University of Athens.

Kristensen, H. O., 2016. Analysis of Technical Data of Ro-Ro Ships. Technical Report. Konges Lyngby: HOK Marineconsult Aps.

Ljulj, A., Slapničar, V., Grubišić, I., 2020. Multi-attribute concept design procedure of a generic naval vessel. Alex. Eng. J. 59 (3), 1725–1734.

Majanarić, D., Šegota, S. B., Lorencin, I., Car, Z., 2022. Prediction of main particulars of container ships using artificial intelligence algorithms. Ocean Eng. 265, 112571.

Novak, L. J., Majanarić, D., Deihalla, R., Zamarin, A., 2020. An analysis of basic parameters of Ro-Pax ships and double-ended ferries as basis for new hybrid ferries design. Pomorski zbornik 3, 33–48.

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

Papanikolaou, A., Harries, S., Hooijmans, P., Marzi, J., Nena, R., Torben, S., Yrjan, A., Boden, B., 2022. A holistic approach to ship design: tools and applications. Journal of Ship Research 66, 25–63.

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

Putra, B., Aryawan, W., Suisetyono, A., 2022. Lane meters correlation analysis towards the main dimensions of Ro-ro ships under 2000 GT. In: Proceedings of the 4th International Conference on Marine Technology, pp. 141–147.

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

Schneekluth, E., Bertram, V., 1998. Ship Design for Efficiency and Economy, 2nd edition. Butterworth-Heinemann.

Trincas, G., Žanić, V., Grubišić, I., 1994. Comprehensive concept of fast Ro-ro ships by multiattribute decision making. In: IMDC 94, Proceeding of the 5th International Marine Design Conference, Delft.

Watson, D., 1998. Practical Ship Design, vol. I. Elsevier.

Žanić, V., Grubišić, I., Trincas, G., 1992. Multiattribute decision-making system based on random generation of non dominated solutions: an application to fishing vessel design. In: Proceedings of the PRAD, 5th International symposium on the practical design of ships and mobile units, pp. 17–22.