

Regression Models for Estimating Cruise Ship Main Dimensions in the Early Stages of Design

Francesco MAURO^a Ahmed SALEM^b Enrico DEL PIERO^{a,1}, Luca BRAIDOTTI^a
Serena BERTAGNA^a and Vittorio BUCCI^a

^a*Department of Engineering and Architecture, University of Trieste, Trieste, Italy*
^b*Sharjah Maritime Academy, Khorfakkan, UAE*

ORCID ID: Francesco Mauro <https://orcid.org/0000-0003-3471-9411>, Ahmed Salem
<https://orcid.org/0000-0002-0929-6173>, Enrico Del Piero
<https://orcid.org/0009-0001-7719-6166>, Luca Braidotti
<https://orcid.org/0000-0003-0255-6822>, Serena Bertagna
<https://orcid.org/0000-0002-1100-5048>, Vittorio Bucci
<https://orcid.org/0000-0001-5929-8655>

Abstract. Cruise ships are complex object to design, requiring the knowledge of multiple aspects of naval architecture and marine engineering. During the concept phase of design, the first step consists in estimating the ship main dimensions as a function of the desiderata of the ship owners. Therefore, it is necessary to provide designers with suitable tools for a fast estimation of the principal ship parameters. The literature reports the availability of simple regression methods and the application of modern machine learning techniques for assessing this issue. The present paper focuses on the traditional regression methods, providing for the first time on cruise ships multiple linear regressions that could be directly applied by designers in the early stages of design. Starting from a database of modern cruise vessels, models have been developed not only for the main dimensions but also for other relevant preliminary attributes like the deadweight, displacement and installed power. The results highlights how the multiple linear regressions provide a good estimate for all the considered parameters, giving to designers an instrument easy to apply.

Keywords. Ship design, Regression models, Cruise ships, Multiple linear regressions

1. Introduction

The design of a modern cruise ship (or of a ship in general) is a complex process, requiring multiple phases with a constantly increasing level of details [1]. The process ranges from the concept design phase to the detail design of subsystems [2]. The first decisions that a designer should take are the most relevant, as they influence the whole design process [3,4].

¹Corresponding Author: Enrico Del Piero, Department of Engineering and Architecture, University of Trieste, Trieste 34127, Italy, enrico.delpiero@phd.units.it.

Targeting the early design stage, the standard process implies the use of simple regression or diagrams to estimate dimensions as a function of design driven parameters [5,6,7]. Of course, due to evident diversities, the key driven parameters vary per each ship type [8,9]. More specifically, for cruise ships the driven parameters are the deadweight, the displacement or the number of passenger onboard [10,11]. The open literature does not address a wide range of examples of regression analysis on cruise ships. Examples of regressions are provided concerning the powering performance of the vessel [11], the installed power [12] and the baricentric vertical motions of the ship [13]. There are no studies aiming at determining the main dimensions as a function of the key parameters. Furthermore, during preliminary design, the shipowner may request the satisfaction of additional parameters, such as ship speed or passenger capacity.

Usually, simple regressions available in literature for other kinds of ships are not considering these additional parameters, except for the work of [14] on RoPax vessels. It is also important to state that, except for some recent publications [12,14], the available regressions for different kinds of ships are based on outdated databases which are not including recent ships [15,16,17]. It is, therefore, essential to provide regressions on a set of recent vessels in order to give to designers a useful design tool.

The present work aims at developing regression models for cruise ship main particulars employing multiple linear regression techniques, in such a way to determine formulations of immediate rapid applicability for designers. As highlighted by [14], the quality of fit and the generalisation of the model for multiple linear regression method is not that much lower compared to more advanced techniques.

In the present study, the database consists of 126 Cruise vessels, derived from an initial population of 147 ships after the removal of incomplete data and outliers [18]. Given the limited sample size, it is not advisable to use advanced machine learning techniques, which requires higher sample sizes to work proficiently [19,20]. However, the employment of multiple linear regressions does not affect their reliability with such a database size [21,22,14]. Therefore, the decision to limit this work on cruise ship to simple and multiple linear regressions is appropriate. 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.

The paper provides a step forward for the estimation of main dimensions and general particulars of cruise ships, giving the first example of regression sets applicable by designers. The methods, the results and their interpretations are explained and discussed in the paper.

2. The cruise ship database

The first step to start with the regression analysis is the definition of the initial cruise ships database. Interrogating the Clarkson database [18], it has been possible to obtain data for the length between perpendiculars L , the breadth B , the draught T , the depth D , the displacement Δ , the deadweight DWT , the installed power P_B , the vessel speed V_s and the number of passengers onboard N_p . A total of 147 ships was found in the initial screening of cruise ships; however, it is necessary to analyse this first set to eliminate incomplete data and outliers.

The literature provides several methodologies to detect outliers in a population; here, the interquartile range (IQR) rule is used. 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), meaning:

$$\begin{cases} x \leq 1.5IQR - Q_1 \\ x \geq 1.5IQR + Q_3 \end{cases} \quad (1)$$

where the *IQR* is the distance between the first and the third quartile:

$$IQR = Q_3 - Q_1 \quad (2)$$

In any case, prior to start with outliers detection, the incomplete dataset should be removed, to ensure the homogeneity of the database. This process implies the elimination of all the ships that are lacking at list one of the parameters that should be used in the regression analysis. With this elimination process, 9 of the initial 147 ships have been discarded, keeping only ships with a complete series of parameters. Afterwards, the outliers detection process has been applied to the remaining ships, discarding additional 12 ships. This results in a final number of 126 complete datasets. Figure 1 shows the pairwise distributions of the database variables for the final set of 126 ships.

Analysing the p-value matrix **P** of the database variables, the following considerations on correlations can be drawn:

- *Length L*: the length L presents a strong correlation with B , T , DWT , Δ , P_B , V_s and N_p . The correlation with D is low. Therefore, L can be considered a suitable candidate to be an independent variable for regression model development.
- *Breadth B*: the breadth B presents a strong correlation with L , T , DWT , Δ , P_B , V_s and N_p . The correlation with D is low. Therefore, B can be considered a suitable independent variable for the regression analysis.
- *Depth D*: the depth D presents a low correlation with all the other variables in the database. No strong or medium correlation is present. Then, D is not a suitable candidate to be an independent variable for regression model development.
- *Draught T*: the draught T presents a strong correlation level with L , B , DWT , Δ , P_B , V_s and N_p . The correlation with D is low. Therefore, T is a suitable candidate to be an independent variable for the regression model development.
- *Deadweight DWT*: the deadweight DWT presents a strong correlation level with L , B , T , P_B , V_s and N_p . The correlation with D is low. Then, DWT is a suitable candidate to be an independent variable for the regression model development.
- *Displacement Δ*: the displacement Δ presents a strong correlation with L , B , T , DWT , Δ , P_B , V_s and N_p . The correlation with D is low. Therefore, Δ is a suitable candidate to be an independent variable for the regression model development.
- *Installed Power P_B*: the installed power P_B presents a strong correlation with L , B , T , DWT , Δ , P_B , V_s and N_p . The correlation with D is low. Then, P_B is a suitable candidate to be an independent variable for the regression model development.
- *Vessel Speed V_s*: the vessel speed V_s presents a strong correlation with L , B , T , DWT , Δ , P_B , V_s and N_p . The correlation with D is low. Then, V_s is a suitable candidate to be an independent variable for the regression model development.
- *Number of passengers N_p*: the number of passengers N_p presents a strong correlation with L , B , T , DWT , Δ , P_B and V_s . The correlation with D is low. Therefore, N_p can be considered as a suitable candidate to be an independent variable for the regression model development.

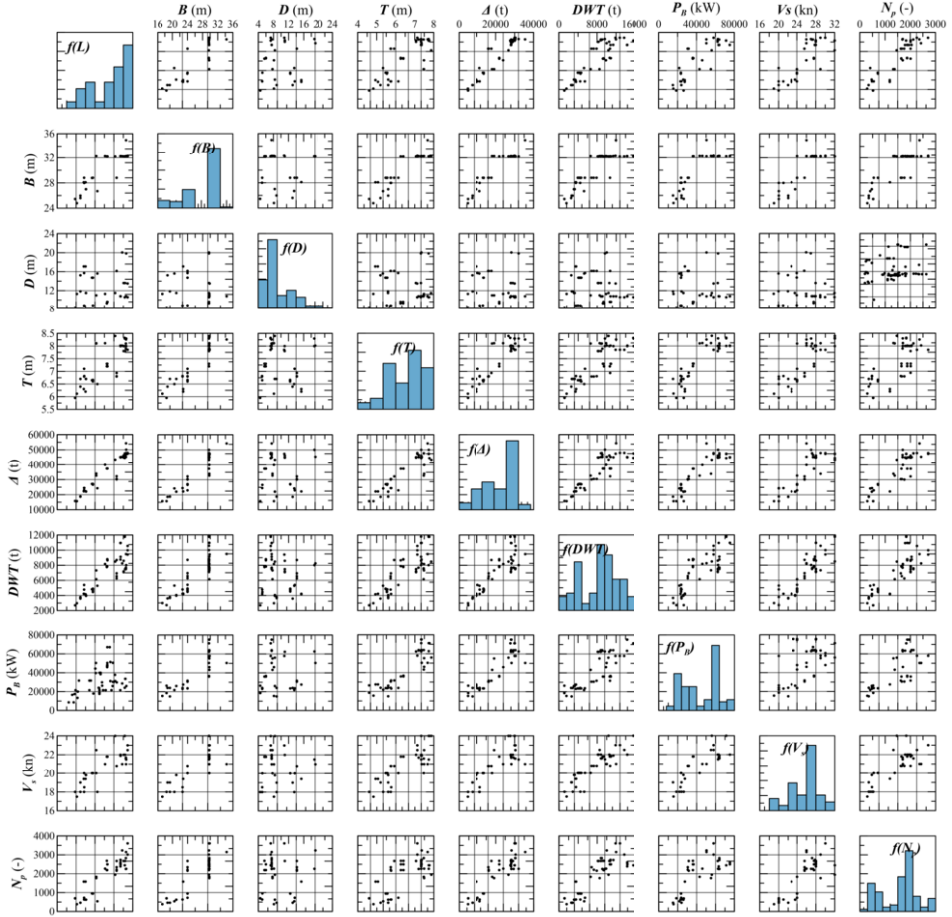


Figure 1. Pairwise distribution of database variables.

The above considerations are summarised in Table 1, reporting the level of correlation (low, medium or high) between the variables. In general, all variables are correlated with each others, except for the depth D .

As a principal consideration, it can be stated that variables like V_s , N_p or P_B (generally asked by owners as principal requirements) can be employed besides others like DWT and Δ as independent variables.

The initial dataset of 126 cruise 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 101 ships, while the test set consists of

Table 1. Correlation level between the variables in the database.

	<i>L</i>	<i>B</i>	<i>D</i>	<i>T</i>	<i>DWT</i>	Δ	P_B	V_s	N_p
<i>L</i>	-	high	low	high	high	high	high	high	high
<i>B</i>	high	-	low	high	high	high	high	high	high
<i>D</i>	low	low	-	low	low	low	low	low	low
<i>T</i>	high	high	low	-	high	high	high	high	high
<i>DWT</i>	high	high	low	high	-	high	high	high	high
Δ	high	high	low	high	high	-	high	high	high
P_B	high	high	low	high	high	high	-	high	high
V_s	high	high	low	high	high	high	high	-	high
N_p	high	high	low	high	high	high	high	high	-

25 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 [14]. This is not a problem for model verification, as, in any case, the dataset is well covered by the provided samples.

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

3. Regression methodology

The following matrix formulation provides the general model for a multiple linear regression:

$$\mathbf{y} = \alpha \mathbf{x} + \varepsilon \quad (3)$$

where $\mathbf{y}$ is the matrix of the observed variables, $\mathbf{x}$ is the matrix of the considered independent variables, α is the matrix of regression coefficients and ε is the matrix of residual errors.

According to the matrix formulation of equation (3), the matrix of the regression coefficients α represents the unknowns of the problem, which can be solved as follows:

$$\alpha = (\mathbf{x}'\mathbf{x})^{-1} \cdot \mathbf{x}'\mathbf{y} \quad (4)$$

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

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:

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

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

These combinations were selected as vessel speed and number of passengers are parameters often specified by shipowners. However, as highlighted in the previous Section, the variables are correlated within each other, and this could be source of autocorrelation as it will be discussed during the analysis of regressions results.

The present study uses the following fit coefficients to assess the quality of the regressions and compare the different formulations: determination coefficient R^2 , Pearson coefficient Prs , mean absolute percentage error $MAPE$, root mean squared error $RMSE$ and relative root mean squared error $RRMSE$. Furthermore, for the case of multiple linear regressions, additional parameters have been evaluated throughout the analysis: the adjusted determination coefficient R^2_{adj} , the standard error SE and the t-Student and p-value for each regression coefficients [23,24].

The quality of fit indicators have the following form:

$$R^2 = 1 - \frac{SSE}{SStot} = 1 - \frac{\sum_{i=1}^n (y_i - y_i^*)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (5)$$

$$R^2_{adj} = 1 - (1 - R^2) \frac{n-1}{n-n_p-1} \quad (6)$$

$$MAPE = \frac{\sum_{i=1}^n |y_i - y_i^*|}{n} \quad (7)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_i - y_i^*)^2}{n}} \quad (8)$$

$$RRMSE = \sqrt{\frac{\frac{1}{n} \sum_{i=1}^n (y_i - y_i^*)^2}{\sum_{i=1}^n (y_i^*)^2}} \quad (9)$$

$$Prs = \frac{\sum_{i=1}^n (y_i - \bar{y})(y_i - \bar{y}^*)}{\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y}^*)^2}} \quad (10)$$

were y_i are the n initial 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 of fit indicators are evaluated for all the proposed regressions on the training set employed to develop the regressions. A second step is performed 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.

4. Regression results

The present section reports the main and most relevant results of the multiple linear regression analysis on the training set of the cruise ships database. For all the regressions, the starting model was a complete 4th order polynomial. From the complete form, a

Table 2. Regression coefficients for model *mlr-0*.

Length L		Breadth B		Depth D		Draught T	
Variable	Coefficient	Variable	Coefficient	Variable	Coefficient	Variable	Coefficient
-	7.490E+01	-	0.000E+00	-	-1.263E+02	-	2.169E+03
DWT	3.154E-02	DWT	3.497E-01	DWT	3.543E-02	DWT	6.179E-02
DWT^2	-1.279E-06	V_s	-1.291E+02	V_s	7.119E+00	V_s	-4.364E+02
		DWT^2	1.206E-05	DWT^2	-1.944E-06	$DWTV_s$	-8.811E-03
		$DWTV_s$	-6.010E-02	$DWTV_s$	-1.795E-03	V_s^2	3.285E+01
		V_s^2	2.097E+01	DWT^2V_s	9.809E-08	$DWTV_s^2$	4.207E-04
		DWT^2V_s	-1.198E-06			V_s^3	-1.094E+00
		$DWTV_s^2$	3.400E-03			$DWTV_s^3$	-6.705E-06
		V_s^3	-1.118E+00			V_s^4	1.361E-02
		$DWT^2V_s^2$	2.958E-08				
		$DWTV_s^3$	-6.313E-05				
		V_s^4	1.974E-02				
Displacement Δ		Installed Power P_B		Number of passengers N_p			
Variable	Coefficient	Variable	Coefficient	Variable	Coefficient		
-	3.995E+07	-	2.156E+05	-	5.131E+04		
DWT	1.314E+03	DWT	2.375E+01	DWT	-8.404E+00		
V_s	-8.110E+06	V_s	-2.987E+04	V_s	-5.259E+03		
$DWTV_s$	-1.902E+02	$DWTV_s$	-9.584E-01	DWT^2	-2.265E-05		
V_s^2	6.147E+05	V_s^2	9.761E+02	$DWTV_s$	8.982E-01		
$DWTV_s^2$	9.211E+00			V_s^2	1.297E+02		
V_s^3	-2.062E+04			$DWTV_s^2$	-2.199E-02		
$DWTV_s^3$	-1.487E-01						
V_s^4	2.586E+02						

stepwise process has been implemented, removing the terms not significant to the global quality of fit of the regression [22,14,11]. Such an approach allows for obtaining the regression with the minimum number of relevant terms to fit the selected variable. For the sake of brevity, the paper provide only the regression coefficients, the associated statistics of p-value and t-Stud is not reported but could be made available on request.

4.0.1. Model *mlr-0*: regressions as a function of V_s and DWT

The first set of regression has been performed by considering V_s and DWT as independent variables. Then, regressions have been performed for L , B , D , T , Δ , P_B and N_p . Table 2 shows the regression coefficient of the resulting model. Figure 2 shows the predicted values of the regressions against the original data of the training set for all the developed regression models. In the figure, the spread along the red bisector line highlights the deviation of the predicted data from the database values. Therefore, the greater the scatter in the diagram, the poorer the quality of the regression fit. Table 3 shows the quality of fit indicators for all the dependent variable regressions. It can be observed that, except for D , all the regressions have a good fit quality.

4.0.2. Model *mlr-1*: regressions as a function of V_s and Δ

This set of regression has been performed by considering V_s and Δ as independent variables. Then, regressions have been performed for L , B , D , T , DWT , P_B and N_p . Table 4

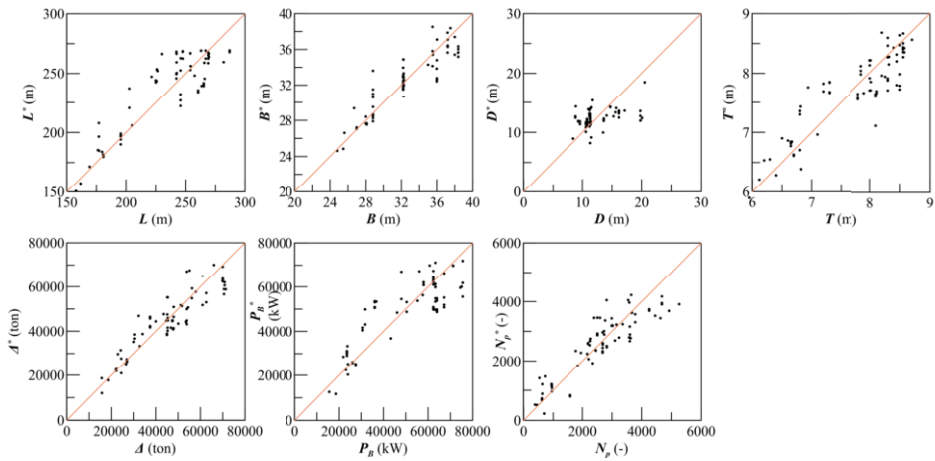


Figure 2. Training set values vs predicted values for the multiple linear regression models as a function of V_s and DWT .

Table 3. Quality of fit of the regressions as a function of V_s and DWT .

	R^2	R^2_{adj}	$MAPE$	$RMSE$	$RRMSE$	Prs
L	0.725	0.719	0.063	1.9E1	0.119	0.851
B	0.821	0.801	0.031	1.538	0.025	0.906
D	0.238	0.197	0.159	2.803	0.077	0.487
T	0.760	0.740	0.038	0.374	0.012	0.872
Δ	0.806	0.789	0.119	6.6E3	3.004	0.897
P_B	0.652	0.638	0.191	1.0E4	4.582	0.808
N_p	0.771	0.756	0.202	5.5E2	1.035	0.878

shows the regression coefficients for model *mlr-1*.

Figure 3 shows the predicted values of the regressions against the training set values for all the developed regression models. The figure illustrates the extent of data spread relative to the bisector (highlighted in red), providing an overview of the deviation between the predicted values and the training dataset. Table 5 shows the quality of fit indicators for all the dependent variables regressions. It can be observed that, except for D , all the regressions have a good fit quality.

4.0.3. Model *mlr-2*: regressions as a function of N_p and DWT

This set of regression has been performed by considering N_p and DWT as independent variables. Then, regressions have been performed for L , B , D , T , Δ , P_B and V_s . Table 6 shows the regression coefficients for model *mlr-2*.

Figure 4 shows the predicted values of the regressions against the training set values for all the developed regression models. The figure shows how much is the spreading of data with respect to the bisector (highlighted in red), giving an overview of the deviation of the predicted data from the training set. The quality of fit has been calculated, as for the previous cases, by evaluating the R^2 , R^2_{adj} , $MAPE$, $RMSE$, $RRMSE$ and Prs quality of fit indicators. Table 7 shows the quality of fit indicators for all the dependent variables

Table 4. Regression coefficients for model *mlr-1*.

Length L		Breadth B		Depth D		Draught T	
Variable	Coefficient	Variable	Coefficient	Variable	Coefficient	Variable	Coefficient
-	3.954E+01	-	0.000E+00	-	5.911E+03	-	1.114E+01
Δ	3.299E-03	Δ	-5.654E-02	Δ	-1.633E-01	Δ	1.016E-04
V_s	4.054E+00	V_s	1.278E+02	V_s	-9.238E+02	V_s	-7.458E-01
Δ^2	-1.387E-08	Δ^2	-2.293E-08	ΔV_s	2.510E-02	ΔV_s	-2.910E-06
		ΔV_s	8.817E-03	V_s^2	4.803E+01	V_s^2	2.357E-02
		V_s^2	-1.917E+01	ΔV_s^2	-1.282E-03		
		ΔV_s^2	1.137E-09	V_s^3	-8.289E-01		
		$\Delta^2 V_s$	-4.477E-04	ΔV_s^3	2.174E-05		
		V_s^3	9.592E-01				
		ΔV_s^3	7.445E-06				
		V_s^4	-1.587E-02				
Deadweight DWT		Installed power P_B		Number of passengers N_p			
Variable	Coefficient	Variable	Coefficient	Variable	Coefficient		
-	-5.846E+06	-	0.000E+00	-	6.114E+04		
Δ	-1.441E+02	Δ	3.689E+02	Δ	-1.948E+00		
V_s	1.380E+06	V_s	-8.914E+05	V_s	-6.368E+03		
ΔV_s	2.144E+01	Δ^2	3.240E-04	ΔV_s	2.013E-01		
V_s^2	-1.189E+05	ΔV_s	-5.770E+01	V_s^2	1.638E+02		
ΔV_s^2	-1.058E+00	V_s^2	1.355E+05	ΔV_s^2	-5.014E-03		
V_s^3	4.455E+03	$\Delta^2 V_s$	-1.734E-05				
ΔV_s^3	1.734E-02	ΔV_s^2	2.928E+00				
V_s^4	-6.152E+01	V_s^3	-6.769E+03				
		ΔV_s^3	-4.789E-02				
		V_s^4	1.108E+02				

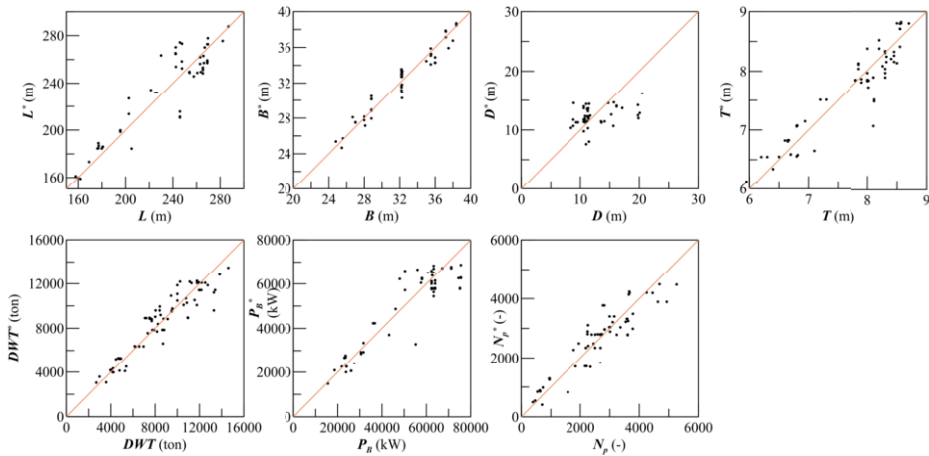


Figure 3. Training set values vs predicted values for the multiple linear regression models as a function of V_s and Δ .

Table 5. Quality of fit of the regressions as a function of V_s and Δ .

	R^2	R^2_{adj}	$MAPE$	$RMSE$	$RRMSE$	Prs
L	0.855	0.851	0.044	1.3E1	0.087	0.925
B	0.934	0.928	0.022	0.923	0.015	0.966
D	0.233	0.176	0.165	2.840	0.077	0.483
T	0.877	0.872	0.028	0.262	0.009	0.936
DWT	0.799	0.779	0.110	1.3E3	1.352	0.894
P_B	0.850	0.835	0.115	7.1E3	3.015	0.921
N_p	0.852	0.844	0.169	4.3E2	0.832	0.923

Table 6. Regression coefficients for model $mlr-2$.

Length L		Breadth B		Depth D		Draught T	
Variable	Coefficient	Variable	Coefficient	Variable	Coefficient	Variable	Coefficient
-	1.04E+02	-	1.94E+01	-	-3.52E+00	-	5.29E+00
DWT	4.82E-03	DWT	3.01E-03	DWT	7.43E-03	DWT	4.17E-04
N_p	7.23E-02	N_p	-1.11E-03	DWT^2	-1.02E-06	N_p	-5.98E-04
DWT^2	3.28E-06	DWT^2	-2.32E-07	DWT^3	4.20E-11	$DWTN_p$	-1.04E-07
$DWTN_p$	-1.81E-05	$DWTN_p$	-4.37E-07			N_p^2	5.37E-07
N_p^2	8.44E-06	N_p^2	1.81E-06			N_p^3	-4.83E-11
DWT^2N_p	-1.40E-09	DWT^2N_p	8.64E-11				
$DWTN_p^2$	7.73E-09	$DWTN_p^2$	-2.01E-10				
N_p^3	-9.13E-09						
Displacement Δ		Installed power P_B		Vessel speed V_s			
Variable	Coefficient	Variable	Coefficient	Variable	Coefficient		
DWT	1.03E-02	-	5.84E+03	-	0.00E+00		
DWT	4.17E+00	DWT	-1.57E+01	$DWTN_p$	3.53E-06		
N_p	8.02E-01	N_p	4.15E+01	N_p^2	-5.76E-06		
DWT^2	-2.08E-04	DWT^2	5.60E-03	DWT^3	1.70E-10		
$DWTN_p$	7.05E-04	$DWTN_p$	-2.34E-02	DWT^2N_p	3.11E-11		
		N_p^2	2.93E-02	$DWTN_p^2$	-1.13E-09		
		DWT^3	-3.17E-07	N_p^3	2.32E-09		
		DWT^2N_p	1.56E-06	DWT^3N_p	-5.32E-14		
		$DWTN_p^2$	-2.30E-06	$DWT^2N_p^2$	2.17E-13		
				$DWTN_p^3$	-2.79E-13		

Table 7. Quality of fit of the regressions as a function of N_p and DWT .

	R^2	R^2_{adj}	$MAPE$	$RMSE$	$RRMSE$	Prs
L	0.796	0.778	0.054	1.6E1	0.102	0.892
B	0.823	0.809	0.030	1.503	0.025	0.907
D	0.185	0.159	0.159	2.868	0.079	0.430
T	0.782	0.770	0.035	0.351	0.012	0.884
Δ	0.868	0.862	0.096	5.3E3	2.475	0.931
P_B	0.715	0.690	0.165	9.7E3	4.148	0.845
V_s	0.635	0.594	0.028	0.874	0.017	0.799

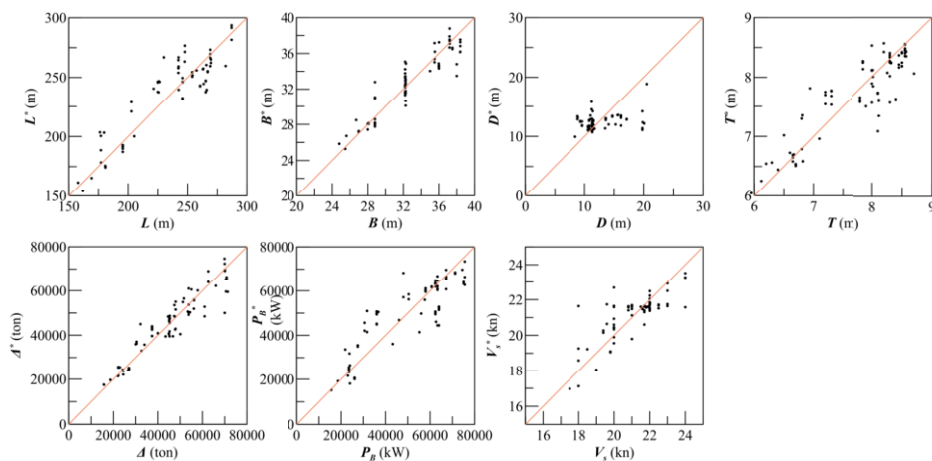


Figure 4. Training set values vs predicted values for the multiple linear regression models as a function of N_p and DWT .

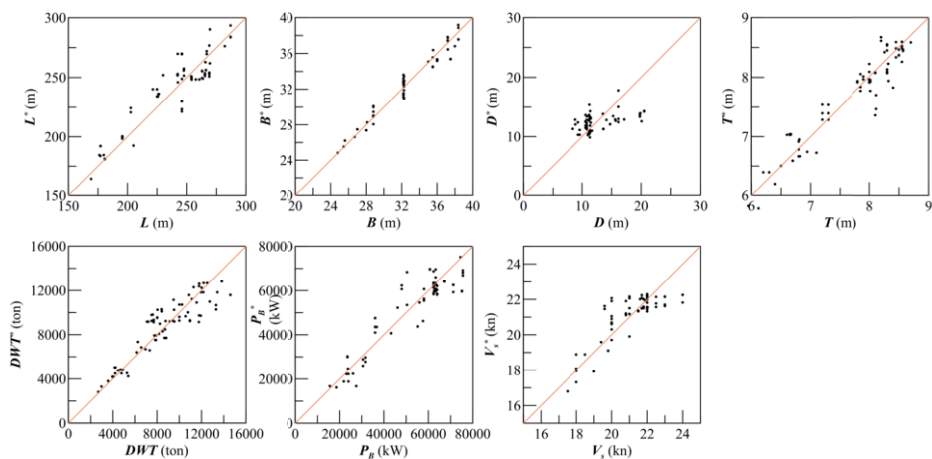


Figure 5. Training set values vs predicted values for the multiple linear regression models as a function of N_p and Δ .

regressions. It can be observed that, except for D , all the regressions have a good fit quality.

4.0.4. Model $mlr-3$: regressions as a function of N_p and Δ

This set of regression has been performed by considering N_p and Δ as independent variables. Then, regressions have been performed for L , B , D , T , DWT , P_B and V_S . Table 8 reports the regression coefficients for model $mlr-3$.

Figure 5 shows the predicted values of the regressions against the training set values for all the developed regression models. The figure shows how much is the spreading of data with respect to the bisector (highlighted in red), giving an overview of the deviation of the predicted data from the training set. Table 9 shows the quality of fit indicators

Table 8. Regression coefficients for model *mlr-3*.

Length <i>L</i>		Bradth <i>B</i>		Depth <i>D</i>		Draught <i>T</i>	
Variable	Coefficient	Variable	Coefficient	Variable	Coefficient	Variable	Coefficient
-	0.00E+00	-	0.00E+00	-	0.00E+00	-	0.00E+00
Δ	1.20E-02	Δ	2.81E-03	Δ	1.65E-03	Δ	6.10E-04
N_p	-2.51E-02	Δ^2	-9.78E-08	Δ^2	-1.33E-08	Δ^2	-1.77E-08
Δ^2	2.96E-08	ΔN_p	1.72E-07	ΔN_p	-2.74E-06	ΔN_p	-2.14E-08
ΔN_p	-8.66E-06	N_p^2	-3.07E-06	N_p^2	2.67E-05	N_p^2	-3.89E-07
N_p^2	1.08E-04	Δ^3	5.28E-13	Δ^3	-1.89E-12	$\Delta^2 N_p$	1.17E-11
Δ^3	-7.90E-12	$\Delta^2 N_p$	3.26E-11	$\Delta^2 N_p$	1.68E-10	ΔN_p^2	-1.56E-10
$\Delta^2 N_p$	4.09E-10	ΔN_p^2	-3.67E-10	ΔN_p^2	-2.40E-09	N_p^3	8.46E-10
ΔN_p^2	-5.59E-09	N_p^3	1.82E-09	N_p^3	9.47E-09	$\Delta^2 N_p^2$	-2.18E-15
N_p^3	1.17E-08	$\Delta^2 N_p^2$	-8.95E-15	$\Delta^2 N_p^2$	-2.09E-14	ΔN_p^3	4.21E-14
N_p^4	1.19E-12	ΔN_p^3	1.59E-13	ΔN_p^3	4.56E-13	N_p^4	-2.42E-13
		N_p^4	-8.30E-13	N_p^4	-2.60E-12		
Deadweight <i>DWT</i>		Installed power <i>P_B</i>		Vessel speed <i>V_s</i>			
Variable	Coefficient	Variable	Coefficient	Variable	Coefficient		
-	0.00E+00	-	0.00E+00	-	0.00E+00		
Δ	1.97E-01	Δ	1.84E+00	Δ	1.73E-03		
Δ^2	2.16E-05	N_p	-2.98E+01	Δ^2	-4.46E-08		
ΔN_p	-1.03E-03	Δ^2	-4.42E-05	ΔN_p	-7.33E-08		
N_p^2	1.08E-02	ΔN_p	2.65E-03	N_p^2	-3.27E-07		
Δ^3	-9.10E-10	N_p^2	-1.66E-02	Δ^3	-7.05E-13		
$\Delta^2 N_p$	4.39E-08	Δ^3	-2.76E-09	$\Delta^2 N_p$	6.43E-11		
ΔN_p^2	-4.93E-07	$\Delta^2 N_p$	1.26E-07	ΔN_p^2	-1.06E-09		
N_p^3	4.73E-07	ΔN_p^2	-2.34E-06	N_p^3	6.09E-09		
$\Delta^2 N_p^2$	-1.90E-12	N_p^3	1.39E-05	$\Delta^2 N_p^2$	-4.99E-15		
ΔN_p^3	2.89E-11			ΔN_p^3	1.15E-13		
				N_p^4	-7.57E-13		

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

	R^2	R^2_{adj}	<i>MAPE</i>	<i>RMSE</i>	<i>RRMSE</i>	<i>Prs</i>
<i>L</i>	0.887	0.876	0.043	1.2E1	0.076	0.942
<i>B</i>	0.949	0.943	0.019	0.821	0.013	0.974
<i>D</i>	0.189	0.099	0.178	2.969	0.079	0.435
<i>T</i>	0.906	0.897	0.022	0.234	0.007	0.952
<i>DWT</i>	0.876	0.843	0.085	1.1E3	1.062	0.936
<i>P_B</i>	0.848	0.835	0.124	7.1E4	3.023	0.921
<i>V_s</i>	0.685	0.650	0.027	0.812	0.016	0.830

for all the dependent variables regressions. It can be observed that, except for *D*, all the regressions have a good fit quality.

4.0.5. Model *mlr-4*: regressions as a function of N_p , V_s and *DWT*

This set of regression has been performed by considering N_p , V_s and *DWT* as independent variables. Then, regressions have been performed for *L*, *B*, *D*, *T*, Δ and P_B . Table 12

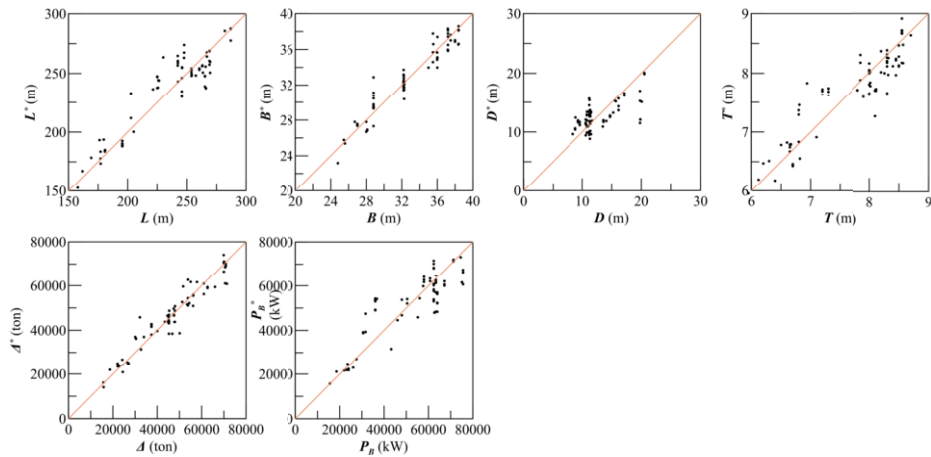


Figure 6. Training set values vs predicted values for the multiple linear regression models as a function of N_p , V_s and DWT .

shows the regression coefficients for model *mlr-4*.

Figure 6 shows the predicted values of the regressions against the training set values for all the developed regression models. The figure shows the extent to which the data deviate from the bisector (in red), offering insight into the accuracy of the predictions relative to the training set. The quality of fit has been calculated, as for the previous cases, by evaluating the R^2 , R^2_{adj} , $MAPE$, $RMSE$, $RRMSE$ and Prs quality of fit indicators. Table 11 shows the quality of fit indicators for all the dependent variables regressions. It can be observed that, except for D , all the regressions have a good fit quality.

4.0.6. Model *mlr-5*: regressions as a function of N_p , V_s and Δ

This set of regression has been performed by considering N_p , V_s and Δ as independent variables. Then, regressions have been performed for L , B , D , T , DWT and P_B . Table 12 reports the regression coefficients for model *mlr-5*.

Figure 7 shows the predicted values of the regressions against the training set values for all the developed regression models. The figure shows how much is the spreading of data with respect to the bisector (highlighted in red), giving an overview of the deviation of the predicted data from the training set. Table 13 shows the quality of fit indicators for all the dependent variables regressions. It can be observed that, except for D , all the regressions have a good fit quality.

5. Models verification

The previous analysis is worthy for determining the quality of fit of the regressions, establishing the intrinsic quality of the models. However, the analysis is not giving an indication on the possible generalisation of the model to a different set of ships. To overcome this issue, the present section presents the quality of fit analysis on the hold-out test set, which means considering a set of ships not used for the models development. The evaluation of the quality of fit indices on this new set of ships may allow for a more

Table 10. Regression coefficients for model *mlr-4*

Length <i>L</i>		Bradth <i>B</i>		Depth <i>D</i>		Draught <i>T</i>	
Variable	Coefficient	Variable	Coefficient	Variable	Coefficient	Variable	Coefficient
-	0.00E+00	-	0.00E+00	-	0.00E+00	-	7.16E+01
<i>DWT</i>	4.36E-01	DWT^2	-1.60E-05	<i>DWT</i>	-9.83E-02	<i>DWT</i>	-2.59E-03
V_s	-2.55E+01	$DWTV_s$	1.06E-03	N_p	-1.01E+00	V_s	-7.51E+00
N_p	-1.49E+00	$DWTN_p$	1.83E-04	DWT^2	2.93E-05	N_p	3.85E-04
DWT^2	-1.17E-05	V_sN_p	-5.50E-03	$DWTV_s$	4.61E-03	DWT^2	-6.85E-07
$DWTV_s$	-2.64E-02	N_p^2	-3.63E-04	V_s^2	5.50E-01	$DWTV_s$	5.31E-04
V_s^2	2.18E+00	DWT^3	-2.32E-10	$DWTN_p$	-1.20E-04	V_s^2	2.08E-01
V_sN_p	7.68E-02	DWT^2V_s	1.77E-06	V_sN_p	1.53E-01	$DWTN_p$	2.99E-06
N_p^2	5.44E-04	$DWTV_s^2$	-1.03E-04	N_p^2	2.26E-04	V_sN_p	-3.06E-05
DWT^3	-4.61E-10	V_s^3	7.01E-03	DWT^3	-2.15E-09	N_p^2	-5.18E-06
$DWTV_s^2$	1.18E-06	DWT^2N_p	-7.75E-10	DWT^2V_s	-1.35E-06	DWT^2V_s	3.19E-08
$V_sN_p^2$	-2.76E-05	$DWTV_sN_p$	-1.70E-05	V_s^3	-2.49E-02	$DWTV_s^2$	-1.85E-05
N_p^3	-6.80E-08	$V_s^2N_p$	3.40E-04	DWT^2N_p	1.22E-08	$DWTV_sN_p$	-1.42E-07
$V_sN_p^3$	3.42E-09	$DWTN_p^2$	6.21E-10	$DWTV_sN_p$	5.42E-06	$V_sN_p^2$	2.50E-07
		$V_sN_p^2$	3.42E-05	$V_s^2N_p$	-7.63E-03		
		N_p^3	1.29E-09	$DWTN_p^2$	-2.31E-08		
		DWT^3V_s	7.92E-12	$V_sN_p^2$	-1.03E-05		
		$DWT^2V_s^2$	-4.23E-08	DWT^3N_p	9.87E-11		
		$DWTV_s^2N_p$	4.07E-07	$DWT^2V_sN_p$	-5.55E-10		
		$DWT^2N_p^2$	1.83E-13	$V_s^3N_p$	1.26E-04		
		$V_s^2N_p^2$	-8.77E-07	$DWTV_sN_p^2$	1.06E-09		
		$DWTN_p^3$	-5.88E-13				
		$V_sN_p^3$	2.57E-10				
Displacement Δ		Installed power P_B					
Variable	Coefficient	Variable	Coefficient				
-	0.00E+00	-	2.01E+06				
<i>DWT</i>	1.16E+02	<i>DWT</i>	4.97E+02				
V_s	-1.66E+04	V_s	-4.64E+05				
N_p	2.78E+02	N_p	1.84E+03				
DWT^2	-5.87E-03	$DWTV_s$	-5.20E+01				
$DWTV_s$	-7.01E+00	V_s^2	3.21E+04				
V_s^2	1.04E+03	$DWTN_p$	-1.97E-01				
$DWTN_p$	4.63E-02	V_sN_p	-1.90E+02				
V_sN_p	-4.15E+01	$DWTV_s^2$	1.36E+00				
N_p^2	-1.65E-01	V_s^3	-6.92E+02				
DWT^3	-1.22E-07	$DWTV_sN_p$	2.06E-02				
DWT^2N_p	4.43E-04	$V_s^2N_p$	4.88E+00				
$DWTV_sN_p$	-2.14E-03	$DWTV_s^2N_p$	-5.36E-04				
$V_s^2N_p$	1.35E+00						
$DWTN_p^2$	-8.63E-06						
$V_sN_p^2$	1.84E-02						
$DWTV_sN_p^2$	4.03E-07						
$V_s^2N_p^2$	-5.00E-04						

Table 11. 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	0.824	0.800	0.050	1.5E1	0.095	0.908
B	0.920	0.899	0.022	1.094	0.016	0.959
D	0.418	0.281	0.151	2.652	0.067	0.646
T	0.810	0.782	0.034	0.342	0.011	0.900
Δ	0.912	0.896	0.077	4.6E3	2.011	0.955
P_B	0.727	0.690	0.153	9.7E3	4.059	0.852

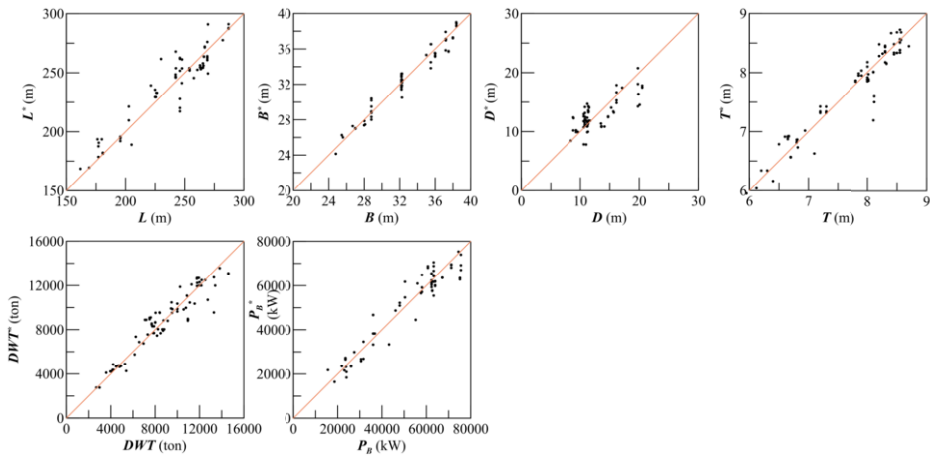


Figure 7. Training set values vs predicted values for the multiple linear regression models as a function of N_p , V_s and Δ .

comprehensive understanding of the significance of the presented models, establishing a ranking between the proposed solutions. 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.

5.1. Length L

All the developed multiple linear regressions models describe the length L . The quality of fit results, obtained on the test set, are shown in Table 14.

From a first observation of the obtained quality of fit indicators, it can be stated that the regression models are significant for the dependent variable L . In conclusion, the model with the highest fitting capabilities are $mlr-5$, $mlr-3$ and $mlr-2$, with R^2 and Prs coefficients all above 0.88. Therefore, the developed multiple linear regression models can be employed by designers for the feasibility study of a new cruise ship.

5.2. Breadth B

All the developed multiple linear regression models describe B as dependent variable. The results of the quality of fit analysis for B is shown in Table 15. From a first observation of the obtained quality of fit indicators, it can be stated that the regression models are significant for the dependent variable B . In fact, the quality of fit indicators R^2 and Prs

Table 12. Regression coefficients for model *mlr-5*

Length <i>L</i>		Breadth <i>B</i>		Depth <i>D</i>		Draught <i>T</i>	
Variable	Coefficient	Variable	Coefficient	Variable	Coefficient	Variable	Coefficient
-	0.00E+00	-	0.00E+00	-	0.00E+00	-	5.45E+01
Δ	1.33E-01	Δ	-6.86E-03	Δ	-1.30E-02	Δ	2.77E-05
Δ^2	-4.35E-06	Δ^2	4.75E-07	N_p	-4.59E-01	V_s	-5.51E+00
ΔV_s	-8.43E-03	ΔV_s	6.73E-04	Δ^2	1.02E-07	N_p	-4.33E-03
V_s^2	-5.07E-01	V_s^2	1.18E-01	ΔV_s	1.54E-03	Δ^2	-1.38E-09
ΔN_p	-1.11E-04	ΔN_p	1.55E-06	V_s^2	7.54E-02	ΔV_s	7.14E-06
$V_s N_p$	-1.16E-02	$V_s N_p$	-2.77E-03	ΔN_p	-4.64E-07	V_s^2	1.49E-01
N_p^2	2.17E-03	N_p^2	-8.63E-05	$V_s N_p$	4.31E-02	$V_s N_p$	9.71E-04
$\Delta^2 V_s$	1.85E-07	$\Delta^2 V_s$	-3.75E-08	N_p^2	2.11E-04	N_p^2	-1.82E-06
ΔV_s^2	1.83E-04	ΔV_s^2	-2.27E-05	ΔV_s^2	-4.96E-05	$V_s^2 N_p$	-4.10E-05
$\Delta^2 N_p$	2.46E-09	$\Delta^2 N_p$	-1.28E-10	$\Delta^2 N_p$	-3.07E-11	$V_s N_p^2$	1.17E-07
$\Delta V_s N_p$	1.21E-05	$\Delta V_s N_p$	-2.96E-08	$\Delta V_s N_p$	-8.46E-07	N_p^3	-6.10E-11
$V_s^2 N_p$	-7.20E-05	$V_s^2 N_p$	1.95E-04	$V_s^2 N_p$	-9.73E-04		
ΔN_p^2	-5.54E-08	ΔN_p^2	2.21E-09	ΔN_p^2	3.74E-09		
$V_s N_p^2$	-1.94E-04	$V_s N_p^2$	6.44E-06	$V_s N_p^2$	-9.63E-06		
N_p^3	3.01E-07	N_p^3	-8.89E-09	N_p^3	-4.88E-08		
$\Delta^2 V_s N_p$	-1.08E-10	$\Delta^2 V_s N_p$	7.90E-10	$\Delta V_s^2 N_p$	3.78E-08		
$\Delta V_s^2 N_p$	-3.47E-07	$\Delta V_s^2 N_p$	5.38E-12	$\Delta V_s N_p^2$	-1.46E-10		
$\Delta V_s N_p^2$	2.56E-09	$\Delta V_s N_p^2$	-1.00E-10	$V_s N_p^3$	2.12E-09		
$V_s^2 N_p^2$	4.95E-06	$V_s^2 N_p^2$	-1.66E-07				
$V_s N_p^3$	-1.67E-08	$V_s N_p^3$	6.22E-10				
N_p^4	4.32E-12	N_p^4	-3.32E-13				
Deadweight <i>DWT</i>		Installed power <i>P_B</i>					
Variable	Coefficient	Variable	Coefficient				
-	0.00E+00	-	0.00E+00				
Δ	-5.44E+00	Δ	4.53E+01				
Δ^2	-5.57E-04	N_p	-3.48E+02				
ΔV_s	4.85E-01	Δ^2	-2.53E-04				
V_s^2	4.41E+01	ΔV_s	-6.68E+00				
ΔN_p	3.02E-02	V_s^2	1.19E+03				
$V_s N_p$	-2.45E+00	ΔN_p	-8.74E-02				
N_p^2	-2.70E-01	$V_s N_p$	6.81E+01				
Δ^3	-1.25E-09	N_p^2	1.27E+00				
$\Delta^2 V_s$	7.06E-05	$\Delta^2 V_s$	1.09E-04				
ΔV_s^2	-1.53E-02	ΔV_s^2	2.53E-01				
$\Delta^2 N_p$	-9.27E-08	V_s^3	-7.31E+01				
$\Delta V_s N_p$	-3.23E-03	$\Delta^2 N_p$	1.27E-07				
$V_s^2 N_p$	1.54E-01	$\Delta V_s N_p$	4.33E-03				
ΔN_p^2	3.06E-06	$V_s^2 N_p$	-2.70E+00				
$V_s N_p^2$	2.86E-02	ΔN_p^2	6.11E-06				
N_p^3	-2.41E-05	$V_s N_p^2$	-8.74E-02				
$\Delta^2 V_s^2$	-1.93E-06	N_p^3	-9.42E-05				
$\Delta^2 V_s N_p$	6.15E-09	$\Delta^2 V_s^2$	-5.53E-06				
$\Delta V_s^2 N_p$	8.31E-05	$\Delta V_s N_p^2$	-4.17E-07				
$\Delta V_s N_p^2$	-1.66E-07	$V_s^2 N_p^2$	1.29E-03				
$V_s^2 N_p^2$	-7.43E-04	$V_s N_p^3$	5.13E-06				
$V_s N_p^3$	1.24E-06						

Table 13. Quality of fit of the regressions as a function of N_p , V_s and Δ .

	R^2	R^2_{adj}	MAPE	RMSE	RRMSE	Prs
L	0.916	0.896	0.033	1.1E1	0.065	0.957
B	0.233	0.176	0.165	2.840	0.077	0.483
D	0.577	0.490	0.123	2.233	0.057	0.759
T	0.933	0.925	0.017	0.194	0.006	0.966
Δ	0.805	0.788	0.119	6.6E3	3.004	0.897
P_B	0.919	0.899	0.090	5.5E3	2.208	0.958

Table 14. Quality of fit on the test set for the L prediction models.

Model	R^2	MAPE	RMSE	RRMSE	Prs
<i>mlr-0</i>	0.863	0.041	1.9E1	0.176	0.929
<i>mlr-1</i>	0.879	0.041	1.4E1	0.165	0.938
<i>mlr-2</i>	0.888	0.034	1.7E1	0.159	0.943
<i>mlr-3</i>	0.916	0.039	1.3E1	0.138	0.962
<i>mlr-4</i>	0.872	0.039	1.6E1	0.171	0.934
<i>mlr-5</i>	0.944	0.027	1.2E1	0.113	0.974

Table 15. Quality of fit on the test set for the B prediction models.

Model	R^2	MAPE	RMSE	RRMSE	Prs
<i>mlr-0</i>	0.713	0.042	1.538	0.068	0.851
<i>mlr-1</i>	0.950	0.019	0.923	0.028	0.976
<i>mlr-2</i>	0.759	0.043	1.503	0.062	0.879
<i>mlr-3</i>	0.964	0.017	0.821	0.024	0.985
<i>mlr-4</i>	0.783	0.034	1.094	0.059	0.893
<i>mlr-5</i>	0.970	0.014	0.736	0.021	0.986

are all above 0.7. The remaining coefficients are quite low, confirming the good fitting of the variable on the test set. Also in this case, the quality of fit indicators are different from the ones obtained on the training set, but the models remains significant for the selected variable.

In conclusion, the model with the highest fitting capability is *mlr-5*, followed by *mlr-1*. Therefore, the developed multiple linear regressions models can be employed by designers in the feasibility study of a new cruise ship.

5.3. Depth D

All the developed multiple linear regression models describe D as dependent variable. The quality of fit results, obtained on the test set, are shown in Table 16.

From a first overview of the quality of fit indicators, it can be stated that the developed regression models are not significant for this dependent variable. The quality of fit indicators like R^2 and Prs are all below 0.7, and in particular the R^2 is never above 0.4. This trend is confirming the results observed during the fitting analysis on the training set.

The best fitting performances are represented by model *mlr-4*, but also in this case, the low levels of the indicators suggest that this model could be at most moderate sig-

Table 16. Quality of fit on the test set for the D prediction models.

Model	R^2	$MAPE$	$RMSE$	$RRMSE$	Prs
<i>mlr-0</i>	0.298	0.111	2.803	0.149	0.652
<i>mlr-1</i>	0.105	0.146	2.840	0.164	0.395
<i>mlr-2</i>	0.209	0.120	2.868	0.157	0.569
<i>mlr-3</i>	0.082	0.177	2.969	0.165	0.311
<i>mlr-4</i>	0.366	0.132	2.652	0.141	0.653
<i>mlr-5</i>	0.011	0.154	2.233	0.191	0.492

Table 17. Quality of fit on the test set for the T prediction models.

Model	R^2	$MAPE$	$RMSE$	$RRMSE$	Prs
<i>mlr-0</i>	0.815	0.035	0.373	0.023	0.907
<i>mlr-1</i>	0.891	0.030	0.262	0.017	0.945
<i>mlr-2</i>	0.876	0.027	0.351	0.018	0.936
<i>mlr-3</i>	0.908	0.022	0.234	0.016	0.957
<i>mlr-4</i>	0.823	0.033	0.342	0.022	0.912
<i>mlr-5</i>	0.922	0.022	0.199	0.015	0.963

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

Model	R^2	$MAPE$	$RMSE$	$RRMSE$	Prs
<i>mlr-1</i>	0.750	0.129	1.1E3	3.285	0.874
<i>mlr-3</i>	0.791	0.111	1.3E3	3.048	0.889
<i>mlr-5</i>	0.857	0.094	1.1E3	2.487	0.927

nificant for the selected dependent variable. Then, it is not advisable to use this set of regressions for a reliable estimation of D in a feasibility study of a new cruise ship.

5.4. Draught T

All the developed multiple linear regression models describe T as dependent variable. The quality of fit results, obtained on the test set, are shown in Table 17. From a first overview of the quality of fit indicators, it can be stated that the regression models are significant for the selected dependent variable. The quality of fit indicators like R^2 and Prs are all above 0.75, while the remaining coefficients remains relatively low. This is confirming the good fitting capabilities of the model on the test set.

The model with the highest fitting capability is *mlr-5*, having an R^2 above 0.9. Also model *mlr-3* has high capability performance. It can be concluded that the developed models for estimating T can be employed in a feasibility study of a new cruise ship.

5.5. Deadweight DWT

Between the developed multiple linear regressions models, only *mlr-1*, *mlr-3* and *mlr-5* allow for DWT estimation. The quality of fit results, obtained on the test set, are shown in Table 18. From a first overview of the fitting data on the test set, it can be stated that the developed regression models are significant for this dependent variable. The quality of fit indicators R^2 and Prs are both above 0.7 for all the tested case, while the remaining

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

Model	R^2	$MAPE$	$RMSE$	$RRMSE$	Prs
<i>mlr-0</i>	0.784	0.105	6.6E3	6.426	0.894
<i>mlr-2</i>	0.856	0.098	5.3E3	5.200	0.926
<i>mlr-4</i>	0.843	0.093	4.6E3	5.479	0.920

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

Model	R^2	$MAPE$	$RMSE$	$RRMSE$	Prs
<i>mlr-0</i>	0.774	0.187	1.1E4	8.445	0.898
<i>mlr-1</i>	0.822	0.139	7.1E4	7.440	0.919
<i>mlr-2</i>	0.835	0.134	9.7E4	7.199	0.932
<i>mlr-3</i>	0.852	0.141	7.1E4	6.788	0.937
<i>mlr-4</i>	0.829	0.139	9.7E4	7.321	0.927
<i>mlr-5</i>	0.892	0.118	5.5E4	5.829	0.960

coefficients are relatively low. The model with the highest fitting capabilities is *mlr-5*, the sole with an R^2 and Prs coefficients both above 0.85. In conclusion, the developed models for estimating DWT can be employed in the feasibility study of a new cruise ship.

5.6. Displacement Δ

The displacement Δ is not represented by all the developed models. More specifically, only *mlr-0*, *mlr-2* and *mlr-4* are available. The quality of fit results, obtained on the test set, are shown in Table 19. From a first overview of the fitting data on the test set, it can be stated that the developed regression models are significant for this dependent variable. The quality of fit indicators R^2 and Prs are both above 0.75 for all the tested case, while the remaining coefficients are relatively low. This is confirming the good fitting capabilities of the developed regressions.

The model with the highest fitting capabilities is *mlr-2*, the sole with an R^2 and Prs coefficients both above 0.85. In conclusion, the developed models for estimating Δ can be employed in the feasibility study of a new cruise ship.

5.7. Installed power P_B

All the developed multiple linear regression models describe P_B as dependent variable. The quality of fit results, obtained on the test set, are shown in Table 20. From a first overview of the quality of fit indicators, it can be concluded that the developed models are significant for the present variables. The quality of fit indicators R^2 and Prs for this specific model are ranging from 0.7 to 0.9. The overall value of the coefficients is different from the one obtained with the training set, but are still confirming the overall significance of the developed models for this dependent variable.

The model with the highest quality of fit is *mlr-5*, followed by *mlr-3* and *mlr-2*, all with R^2 and Prs coefficients above 0.825. Therefore, the developed models for the P_B estimation can be employed in a feasibility study for a new cruise ship.

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

Model	R^2	$MAPE$	$RMSE$	$RRMSE$	Prs
<i>mlr-2</i>	0.714	0.032	0.874	0.036	0.853
<i>mlr-3</i>	0.517	0.039	0.812	0.047	0.732

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

Model	R^2	$MAPE$	$RMSE$	$RRMSE$	Prs
<i>mlr-0</i>	0.826	0.180	5.4E2	1.840	0.912
<i>mlr-1</i>	0.856	0.165	4.3E2	1.683	0.926

5.8. Vessel speed V_s

The vessel speed V_s is represented only by *mlr-2* and *mlr-3* multiple linear regression models. Table 21 shows the quality of fit indicators obtained for the test set. From a first overview of the quality of fit indicators, it can be stated that the models to estimate V_s are moderately significant for the present dependent variable. The quality of fit coefficients R^2 and Prs are ranging between 0.45 and 0.85, while the remaining coefficients remains relatively high.

The model with the highest fitting performances is *mlr-2*, having Prs and R^2 indicators above 0.7. This is the sole model that can be considered significant for the reproduction of the vessel speed V_s . Therefore, it can be concluded that only model *mlr-2* can be employed for V_s estimation in the feasibility study of a new cruise ship.

5.9. Number of passengers N_p

The number of passengers N_p is represented only by *mlr-0* and *mlr-1* multiple linear regression models. Table 22 shows the quality of fit coefficients obtained on the test set. From a first overview of the quality of fit indicators, it can be stated that the models to estimate N_p are significant for the present dependent variable. The quality of fit indicators R^2 and Prs are all above 0.79. The other coefficients remains relatively low, confirming the good fitting performances of the developed model on the test set.

The model with the highest fitting performances is *mlr-1*, having R^2 and Prs coefficients above 0.85. Therefore, it can be concluded that the presented models can be applied for N_p reproduction in the feasibility study of a new cruise ship.

6. Conclusions

This paper presents a dataset of 126 cruise vessels, introducing multiple linear regressions to estimate the main dimensions and the general particulars of the ships. The dataset was split into training and test subsets to enable model development and validation, respectively.

The paper provides 40 multiple linear regression models for the main dimensions and general particulars of Cruise vessels, employing pairs or triplets of independent variables for the analysis. The analysis of the regressions highlight the quality of the provided models, allowing designers to incorporate more than one independent variable as input parameters for estimating the main dimensions of a ship.

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.

The present study focused on the applicability of a standard regression technique for the estimation of main dimensions and general particulars of cruise ships. The result are promising, as they achieve good fitting level on the training and validation set employed. However, further studies could be performed applying machine learning techniques, to understand whether more modern fitting techniques could enhance the prediction of the analysed database.

References

- [1] Andrews DJ. A comprehensive methodology for the design of ships and other complex systems. *Proceedings of Mathematical, Physical and Engineering Sciences*. 1998; 454(1968): 187-211. DOI:10.1098/rspa.1998.0154
- [2] Caprace JD, Rigo P. Ship complexity assessment at the concept design stage. *Journal of Marine Science and Technology*. 2011; 16: 68-75. DOI: 10.1007/s00773-010-0107-9
- [3] Watson D. *Practical Ship Design*. Elsevier, 1998.
- [4] Papanikolaou A, Harries S, Hooijmans P, Marzi J, Nena R, Torben S, Yrjan A, Boden B. A holistic approach to ship design: tools and applications. *Journal of Ship Research*. 2022; 66: 25-63
- [5] Žanić V, Grubišić I, Trincas G. Multiattribute cecision-making system based on random generation of non dominated solutions: an application to fishing vessel design. *Proceedings of the PRADS, 5th International symposium on the practical design of ships and mobile units*; 1992, p. 17-22
- [6] Kolokarinos E, Mavroedis T, Radou G, Zachariou Z. Regression analysis on basic ship design values for merchant ships. National University of Athens, 2005
- [7] Grubišić I, Begović E. Multi-attribute concept design model of the adriatic type of fishing vessels. *Brodogradnja*. 2001; 49(1): 39-54
- [8] Papanikolaou A. *Ship Design. Methodologies of Preliminary Design*. Springer, 2014
- [9] Abramowski T, Cepowsky T, Zvolensky P. Determination of regression formulas for key design characteristics of container ships at preliminary design stage. *New Trends in Production Engineering*. 2018; 1(1): 247-57. DOI:10.2478/ntpe-2018-0031
- [10] Aarnio M. *Cruise Ship Handbook*. Springer, 2023
- [11] Brunello A, Mauro F, Braidotti L, Bucci V. Surrogate models for propulsive performance in the feasibility study of cruise ships. *Proceedings of the 21st International Conference on Ship and Maritime Research, NAV 2025*, 2025 Jun 18-20, Messina, Italy. DOI: 10.3233/PMST250089
- [12] Mocerino L, Rizzuto E. Statistical Analysis of Installed Power on Board Modern Cruise Ships. *Lecture Notes in Civil Engineering*. 2021; 64: 769-83. DOI:10.1007/978-981-15-4672-3_48
- [13] Utzeri S, Braidotti L, Norcia N, A. Marinó. A Methodology for Predicting Vertical Motions in Ship Concept Design. *Progress in Marine Science and Technology*. 2024; 8: 127-37
- [14] Mauro F, Salem, A. Development of regression models for estimating main particulars of RoPax vessels in the conceptual design stage. *Ocean Engineering*. 2025; 333: 121407. DOI:10.1016/j.oceaneng.2025.121407
- [15] Piko DP. *Regression Analysis of Ship Characteristics*. Australian Government Publishing Service. 1980, Canberra, Australia.
- [16] Friis AM, Andersen P, Jensen JJ. *Ship Design (Part I and II)*. Technical University of Denmark, 2002
- [17] Kristensen HO. *Analysis of Technical Data of Ro-ro Ships*. Konges Lyngby, HOK Marineconsult Aps., 2016
- [18] Clarkson. Shipping and trade data analysis and research services. <https://www.clarksons.com>, December 2024
- [19] Clausen HB, Lützen M, Hansen AF, Bjorneboe N. Bayesian and neural networks for preliminary ship design. *Mar. Technol*. 2001; 38(4): 268-77

- [20] Gurgen S, Altin I, Ozkok M. Prediction of main particulars of a chemical tanker at a preliminary ship design using artificial neural network. *Ship and Offshore Structures*. 2018; 13 (5):459-65. DOI:10.1080/17445302.2018.1425337
- [21] Ekinci S, Celabi UB, Bal M, Amasyali MF, Boyaci UK. Predictions of oil/chemical tanker main design parameters using computational intelligence techniques. *Appl. Soft. Comput.* 2011; 11(2): 2356-66. DOI:10.1016/j.asoc.2010.08.015
- [22] Rinauro B, Begovic E, Mauro F, Rosano G. Regression analysis for container ships in the early design stage. *Ocean Engineering*. 2024; 292: 116499. DOI:10.1016/j.oceaneng.2023.116499
- [23] Freedman DA. *Statistical Models: Theory and Practice*. Cambridge University Press, 2009
- [24] Frank E, Harrell JR. *Regression Modeling Strategies: with applications to linear models, logistic and Ordinal Regression, and Survival Analysis*. Springer, 2015