Article

Analysis of Trends in Mega-Sized Container Ships Using the K-Means Clustering Algorithm

Woo-Ju Son ¹ and Ik-Soon Cho ^{2,*} ¹ SafeTechResearch Co., Ltd., Yuseong-gu, Daejeon 34050, Korea; wooju@strkorea.co.kr² Division of Navigation Convergence Studies, Korea Maritime and Ocean University, Yeongdo-gu, Busan 49112, Korea

* Correspondence: ischo@kmou.ac.kr; Tel.: +82-51-410-5072

Abstract: The size of ships is increasing rapidly, and over 400 m length overall mega-sized container ships are predicted to appear in the near future. Consequently, studies on large container ships have been conducted; however, based on the 30,000-TEU (twenty-foot equivalent unit) class container ship's length overall, the deviation of the predicted range showed an 83.9 m difference from 453.0 m to 536.9 m. This is because simple linear regression analysis does not consider trends according to the type of cargo or the size of the cargo. In this study, 5497 container ships up to 20 years of age with an International Maritime Organization (IMO) number registered were clustered according to the change in ship dimensions by *k*-means clustering algorithm. Based on the clustered data, deadweight tonnage, TEU, length overall, length between perpendiculars, breadth, and maximum draft of container ships with a coverage rate of 75% were analyzed to predict the change in the main dimensions. The results indicated that for a 30,000-TEU container ship, the predicted length overall is 428.4 m, breadth is 67.6 m, and draft is 17.0 m. This study can help minimize the social costs of designing ports in consideration of future mega-sized container ships.

Keywords: mega-sized container ship; ship main dimensions; trend analysis; *k*-means clustering; coverage rate

Citation: Son, W.-J.; Cho, I.-S.Analysis of Trends in Mega-Sized Container Ships Using the K-Means Clustering Algorithm. *Appl. Sci.* **2022**, *12*, 2115. <https://doi.org/10.3390/app12042115>

Academic Editor: José A. Orosa

Received: 3 December 2021

Accepted: 16 February 2022

Published: 17 February 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

With the rapid development of mega-sized container ships, and over 400 m length overall mega-sized container ships are predicted to appear in the near future, container ships may appear in the near future. According to the Ministry of Oceans and Fisheries of Korea, the second port of Busan Port (Jinhae New Port) will start construction in 2022 and will be capable of berthing 30,000-TEU (twenty-foot equivalent unit) vessels [1]. The size of the target ship is generally considered when constructing ports or designing a water facility such as a navigation route or a breakwater. However, for future large ships, if the length overall (LOA) and breadth cannot secure the safety of passage in the target sea area, costs such as pier extension and breakwater reconstruction will occur, and additional social costs will be incurred during the consultation between related organizations. In consideration of the future size of ships, studies analyzing mega-sized container ships have been conducted based on regression analysis. However, the deviation of the predicted length overall range showed an 83.9 m difference from 453.0 m to 536.9 m in 30,000-TEU class container ships. The Korea Maritime Institute predicted the shape of a mega-sized container ship by deriving a regression equation by simple linear regression analysis using the TEU of the main dimensions [2]. Cho et al. [3] predicted the container ship's main dimension by dividing changes in container ship size according to different scenarios. Park and Suh [4] predicted that the total length of a 30,000-TEU container ship would reach 453 m and that a pier length of 500 m or more must be secured. However, current 15,000-TEU to 24,000-TEU container ships have a total length of 1310 ft (approximately 400 m) and are being enlarged by increasing their breadth.

This is because changes in the overall length of a container ship are determined by the physical limitations of port facilities [5]. Therefore, simple linear regression analysis has limitations in predicting ship size because it does not consider trends according to cargo specifications. In this study, analysis groups were subdivided through clustering to consider the increasing trend due to large size; a size of 30,000-TEU class container ship was predicted considering the cargo characteristics of container ships. Table 1 compares the dimensions provided by previous studies with the dimensions of mega-sized container ships currently in operation [6,7].

Table 1. Comparison of dimensions reported by previous studies and those of mega-sized container ships currently in operation.

Classification		Length Overall (m)	Molded Breadth (m)	Maximum Draft (m)
Previous studies (25,000 TEU)	KMI [2] ¹	462.3	60.7	17.0
	Cho et al. [3]	474.0	61.0	18.3
Previous studies (30,000 TEU)	KMI [2]	536.9	76.1	18.0
	Cho et al. [3]	517.0	65.0	19.4
	Park and Suh [4]	453.0	67.0	17.3
Mega-sized container (15,000–24,000 TEU)	CMA CGM Marco Polo [6]	396.0	54.0	16.0
	MAERSK E-Class [6]	397.0	56.0	15.5
	MAERSK Triple E-Class [6]	400.0	59.0	15.5
	HMM ALGECIRAS [7]	399.9	61.0	16.5

¹ Value calculated by substituting the regression equation derived from the Korea Maritime Institute.

According to the Korean Harbor and Fishery Design Criteria (KDS 64 10 10) [8], standard ship design criteria are based on information from the 2004 Lloyd's Maritime Intelligence Unit and statistical data using ship specifications from Japan from 2004 [9]. However, these are statistical results based on data that are more than 15 years old and are thus not appropriate for presenting a standard design for container ships due to the recent trend of mega-sized container ships.

In Japan, trends in ship size have been updated approximately by a 10-year cycle. Katayama et al. [10] analyzed the interrelationships between various common ship dimensions, such as length overall (LOA), breadth (B), and gross tonnage (GT), for the design and planning of port facilities using regression analysis. Kiyoshi and Yoshida [11] analyzed the interrelationships of various variables for ships in Japan and other countries. To establish standard linear design conditions according to ship type and size for the planning and design of port facilities, Ajiki et al. [12] analyzed the major dimensions using information about ships registered in Japan and Lloyd's Register of Ships. The correlation between the length overall (LOA), breadth (B), draft (d), and tonnage was analyzed for eight types of ships. Akakura et al. [13] performed statistical analysis to determine standard ship designs of the LOA, B, length between perpendiculars (Lpp), displacement, and d using ship specifications from Japan and Lloyd's registered ship information. Takahashi et al. [9] used information from the Lloyd's Maritime Intelligence Unit and ship specifications from Japan. Regression analysis was performed on the correlation between the main dimensions (LOA, Lpp, B, and d), GT, and deadweight tonnage (DWT) according to nine types of ships, including cargo ships, container ships, tankers, roll-on/roll-off ships, automobile carriers, liquefied petroleum gas ships, liquefied natural gas ships, passenger ships, and ferries. Inoue and Akakura [14] analyzed changes in ship design according to the year of construction by subdividing the main dimensions of the target ships. In particular, for container ships, the TEU was presented parallel to the regression trends. Figure 1a presents the DWT-LOA (m) graph, while Figure 1b presents the TEU/DWT-LOA regression graph.

Figure 1. Trends of main dimensions on container ships: (a) deadweight tonnage (DWT)–length overall (LOA) regression graph; (b) twenty-foot equivalent unit (TEU)/DWT–LOA regression trend line graph (Inoue and Akakura [14]).

Iwasaki and Yamagata [15] presented new standard ship dimensions reflecting larger ship size by regression analysis of the main dimensions using ship specifications from Japan and Lloyd’s registered ship information. The new ship dimensions were reflected in the revision of technical standards for port facilities in Japan [16] to handle recently developed large ships.

In the Recommendations for the Design of the Maritime Configuration of Ports, Approach Channels, and Harbor Basins provided by Puertos del Estado [17], which is a port facility management agency in Spain, 95% of the coverage rate of the regression residual distribution was used as the standard of the ship’s main dimensions. Therefore, standard ship specifications were presented by setting 115% (or 110%) as the upper characteristic standard and 85% (or 90%) as the lower characteristic standard.

The World Association for Waterborne Transport Infrastructure MarCOM (PIANC Maritime Navigation Commission), an expert organization in the navigation and port field, also studied standard ship dimensions. In PIANC MarCOM Working Group 121, the classification of ships was further subdivided compared to previous studies using Lloyd’s 2006 registered ship information by referring to previous studies from Japan [9] and Spain [17]. In addition, flexible standard ship dimensions were provided by setting an error range of $\pm 10\%$. Moreover, for the B of a container ship, the size of the container and the cargo space was taken into consideration. For example, 32.2 m ($2.48 \text{ m} \times 13 \text{ rows}$) for Panamax and 59.5 m ($2.57 \text{ m} \times 23 \text{ rows}$) for Triple E-Class (18,000-TEU) were calculated [18]. In Working Group 184 [19], the value corresponding to 90% was presented as the key dimension (upper value) using the real ship’s information rather than a statistical method.

In this study, the automatic identification system (AIS) data of container ships were collected, and regression analysis was performed on container ships operating worldwide to reflect the trend of container ship’s main dimensions and study the increase in the size of container ships. The main dimensions of the ship were classified through the k -mean clustering process through the ship’s length variable, tonnage variable, and ship’s age variable and analyzed for each clustered data. The size of 30,000-TEU ships was estimated according to the changing trends in container ships. By container ship’s standard main dimensions and presenting trends of mega-sized ships, it will be possible to minimize the social cost of designing ports in consideration of the future size of ships.

2. Materials and Methods

In this study, analysis was performed based on the AIS data of ships around the world registered in International Maritime Organization (IMO) number up to 31 December 2020. The data were refined to be suitable for analysis by preprocessing the target data [20]. In addition, using *k*-means clustering, the ship age was added to the relationship with the main dimensions of the ship data. Ships tend to become larger in recent years, so if this is reflected, changes in ship dimensions can be reflected more effectively. Based on the clustered data, a regression analysis of the ship dimensions was performed, and the change in the main dimensions of container ships was derived. A flowchart of this study is presented in Figure 2.

Figure 2. Flowchart of this study.

2.1. Target Data

Due to ship-to-ship communication limitations and the network size of the shipping industry, studies involving big data analysis have been limited due to lack of data [21]. Recently, due to the development of the very-small-aperture terminal and other devices, communication difficulties between ships have been resolved, and amount of data have been collected. Operational data provided by navigation devices, such as Global Positioning System (GPS), AIS, and the Electronic Chart Display and Information System, are valuable for forecasting, decision-making, and preventing accidents in the ship and shipping industry [22]. In particular, AIS data consist of dynamic information, such as GPS position, heading, and speed over ground, and static information, such as the Maritime Mobile Service Identity, dimensions of the ship, type of cargo, and GT. These data can be used for various studies, such as safe navigation range selection, berthing step classification through density-based spatial clustering, and navigation route extraction [23–25]. In this study, analysis was performed using the ship’s AIS information obtained through the IHS Markit Economics and Country Risk, Inc. world fleet of existing ships (December 2020). This includes information on all ships with the IMO number (LR Number) registered.

2.2. Data Preprocessing

Data preprocessing is the step of transforming data into a state suitable for analysis before applying it to an algorithm [26]. In this study, variables necessary for analysis were classified from the entire dataset, and the remaining variables were removed. According to the United Nations Conference on Trade and Development, the average age of all ships in the world merchant fleet is just over 20 years [27]. Therefore, the age of the target ships used for analysis was classified as 20 years. All data with missing values or values of 0 for the ship’s main dimensions, such as the LOA, Lpp, B, and d, were deleted (listwise deletion). For the DWT, a linear regression relationship between the two was obtained for missing values and replaced with the value produced by the expression of relation (missing

data regression imputation). Table 2 presents the number of data points classified through data preprocessing.

Table 2. Results of data preprocessing.

Data Preprocessing	Raw Data	Acquired Data	Outlier & Missing Value Treatment
Count	7791	7730	5497

2.3. Analysis Method and Coverage Rate Concept

Usually, there are ships of the same tonnage but with slightly different main dimensions. This is because the characteristics desired by the shipping company can be achieved with various combinations of dimensions. The regression analysis related to the ship's main dimensions can be divided into three types, and regression analysis suitable for each type must be applied [9]. The first type is logarithmic regression analysis. Ships of the same category are generally spatially similar regardless of their size; thus, their main dimensions are approximately proportional to the $\frac{1}{3}$ (0.33%) power of their size. The relationship between the main dimensions is calculated by the following equation:

$$y = \alpha x^{\beta} \quad (1)$$

where y denotes the main dimensions (LOA, Lpp, B, d), and x is the tonnage variables (DWT, GT). The shape of this graph is presented in Figure 3.

Figure 3. Logarithmic regression analysis graph.

Equation (1) is changed to Equation (2) by transforming both sides into logarithms to make it easy to perform statistical analysis, such as calculating the simple linear regression equation and the standard differential:

$$\log y = \log \alpha + \beta \log x \quad (2)$$

where α denotes the y-intercept value log-transformed graph, and β is the slope of the log-transformed graph (Equations (1) and (2)). The log-transformed graph takes the form of a linear graph, as illustrated in Figure 4.

Figure 4. Linear regression analysis graph (log-transformed).

The second type of regression analysis is average value analysis. The graph of the major specifications of a ship is generally in the form of a logarithmic regression graph, but some have a form in which certain specifications remain constants. The reason for this is that ships are designed according to the width of the navigation passageway, the water depth of the pier, and the length of the berth. The relationship between the main dimensions and the ship size is calculated by the following equation:

$$y = \begin{cases} \alpha_0, & P1 \leq x < P2 \\ \alpha_1, & P2 \leq x < P3 \end{cases} \quad (3)$$

where α_0 is the y -intercept value when $P1 \leq x < P2$, and α_1 is the y -intercept value when $P2 \leq x < P3$. P_n ($n = 1, 2, 3$) is the different values of x . The slope of the target interval is 0. The shape of the graph is displayed in Figure 5.

Figure 5. Average value analysis graph.

The third type of regression analysis is linear regression analysis. This type of analysis is a simple regression graph expressed in the form of a general linear graph; herein, it was used to derive the relationship between the TEU and DWT. The relationship between the main dimensions and the ship size is calculated by the following equation:

$$y = \alpha x + \beta \quad (4)$$

where, x denotes the DWT, y denotes the TEU, α is the slope, and β is the y -intercept. In this study, β was set to 0 because when β is negative, the y value can be obtained (−) if αx is smaller than the absolute value of β . The graph is presented in Figure 6.

Figure 6. Linear regression analysis graph.

The reliability obtained through regression can be obtained using the adjusted R-squared value of the log transformation graph. In the case of a graph with a slope of 0, the accuracy can be judged according to the degree of the standard deviation of the residuals. The regression analysis results were the average ship's main dimensions (50% of residual distribution), but average values do not represent the dimensions of the entire ship. The reason for this is that there are ships with slightly different main dimensions, even among ships of the same tonnage. Currently, a coverage rate of 75% is used in the technical standards for port facilities in Japan [15] and the harbor and fishery design criteria in Korea [8]. A coverage rate of 75% indicates that the representative value of the target ship falls within the 75% confidence interval of the overall residual one-sided distribution. (i.e., that it refers to the 75th largest ship out of 100 vessels of the same tonnage). In general, when presenting representative values, the average value is most often used. However, in terms of the safety design of the port, the average value does not secure the safety of ships with a value larger than that. Since it is economically inefficient to apply the maximum value, an appropriate level of representative value should be presented. Therefore, in this study, the coverage rate of 75%, which design criteria have been used in Korea and Japan, was selected as the representative main dimensions of ships. Here, the coverage rate refers to the coverage range of the distribution according to the confidence interval (one-sided normal curve) of the residual distribution of the regression. When the results of regression analysis follow a linear regression, the distribution of the residual follows a normal distribution according to the characteristics of linear regression analysis [28]. The relationship between the tonnage variables (GT, DWT) and length variables (LOA, Lpp, B, d) of the ship's main dimensions follows a linear regression when the variables are converted to log. The main dimensions were transformed by log to obtain a range according to the 75% coverage rate, and the value was transformed by 10^x again. Figure 7 displays the concept of 75% coverage.

The way to check whether the results of linear regression analysis are appropriate is to check whether the residual distribution follows normality. Verifying that the distribution of residuals is homoscedasticity rather than linear also helps to check the accuracy of the results of a linear regression analysis. In this study, the homoscedasticity of the residuals was reviewed, and the suitability of the linear regression analysis was reviewed by examining whether the residual distribution for the regression line of the linear regression distribution follows normality through a Q-Q plot.

Figure 7. Concept of 75% coverage rate.

2.4. K-Means Clustering

K-means clustering is a clustering method that clusters non-hierarchical data and indicates clusters between data with similar tendencies. This method is the most widely used non-hierarchical cluster classification method and has the advantage of being able to classify many data quickly and easily. In this study, the trend was analyzed through the group of the largest ship among the clustered data. Iwasaki and Yamagata [15] analyzed the ship's main dimensions by classifying them according to the visual form of the graph. However, since the classification method is different depending on the subjective evaluation of the researcher, there is a disadvantage that the classification standard varies depending on the analyst. In this study, the groups were quantitatively classified by performing clustering through the *k*-means clustering technique. Cluster analysis is a method of classifying an initial dataset and finding hidden patterns [29]. *k*-means cluster analysis is a method of relocating data based on similarity by randomly dividing the data into *k* clusters and dividing the mean of the clusters into representative values [30]. Based on the centroid value of each cluster, each object is assigned to the nearest centroid using the Euclidean distance, and data objects are clustered based on the closest *k*-mean value [31]. The data are displayed as a voronoi diagram based on the average value, and the average distance for each *k* cluster division is readjusted and repeated until it converges to a specific *k* value. This process is sensitive to outliers; therefore, it must be performed after data preprocessing, and effective results can be obtained only when an appropriate number of clusters *k* is specified as an input parameter. When the variable with the characteristics $i = 1, 2, \dots, N$, $j = 1, 2, \dots, p$ is x_{ij} , the distance between i and $\bar{i}$ according to the j -th attribute value is derived as follows:

$$d(x_i, x_{\bar{i}}) = \sum_{j=1}^p d_j(x_{ij}, x_{\bar{i}j}) = \sum_{j=1}^p d_j(x_{ij} - x_{\bar{i}j})^2 = ||x_{ij} - x_{\bar{i}j}||^2 \quad (5)$$

The algorithm randomly sets $k = 1, 2, \dots, K$ of the cluster. The center of the k -th cluster is defined as μ_k . In this case, the distance D from the data point to the center of the cluster is as follows:

$$D = \sum_{k=1}^K \sum_I ||x_i - \mu_k||^2 = \sum_{k=1}^K \sum_{i=1}^N a_{ik} ||x_i - \mu_k||^2 \quad (6)$$

where $I : x_i$ is assigned to k

$$a_{ik} = \begin{cases} 1 & \text{if } I \\ 0 & \text{else} \end{cases}$$

In this case, a_{ik} is a binary variable with a value of 1 or 0. The formula for k -means clustering is as follows:

$$\operatorname{argmin}_{a, \mu} \sum_{k=1}^K \sum_{i=1}^N a_{ik} \|x_i - \mu_k\|^2 \quad (7)$$

In this study, the appropriate number of clusters k was derived using the NbClust package [32] of the R program version 4.1.1 and applied to clustering. The NbClust package can test and compare most indices proposed in previous studies and provides indices that determine the number of clusters in the dataset. In addition, it provides the optimal number of clusters k from the results of various methods. This allows multiple clustering schemes to be evaluated simultaneously while changing the number of clusters to determine the most appropriate number of clusters for the dataset of interest. Table 3 presents the 26 indices used in this study. The 26 indices refer to the most appropriate number of k , which is proposed in previous studies [33–57].

Figure 8. Results of k -means clustering on container ship deadweight tonnage–length overall (m). (a) Number of clusters selected by NbClust; (b) Hubert statistic values; (c) D index value; (d) scatter plot clustered into three groups.

Table 3. K -means clustering indices using NbClust package in this study.

Index by NbClust Package	Reference
KL	Krzanowski and Lai [33]
CH	Calinski and Harabasz [34]
Hartigan	Hartigan [35]
CCC	Sarle [36]
Scott	Scott and Symons [37]

Table 3. Cont.

Index by NbClust Package	Reference
Marriot	Marriot [38]
TrCovW	Milligan and Cooper [39]
TraceW	Milligan and Cooper [39]
Friedman	Friedman and Rubin [40]
Rubin	Friedman and Rubin [40]
C Index	Hubert and Levin [41]
DB	Davies and Bouldin [42]
Silhouette	Rousseeuw [43]
Duda	Duda et al. [44]
PseudoT2	Duda et al. [44]
Beale	Beale [45]
Ratkowsky	Ratkowsky and Lance [46]
Ball	Ball and Hall [47]
PtBiserial	Milligan [48,49]; Kraemer [50]
Frey	Frey and Van Groenewoud [51]
McClain	McClain and Rao [52]
Dunn	Dunn [53]
Hubert ¹	Hubert and Arabie [54]
SD index	Halkidi et al. [55]
D index ¹	Lebart et al. [56]
SDbw	Halkidi and Vazirgiannis [57]

¹ Hubert statistic values and D index values are evaluated by a graphical method, and the change in the index causing a significant increase in the second difference (significant peak in second differences plot) is output as the most appropriate cluster (e.g., blue line graphs of Figures 8, 9, 10 and 11b,c).

Figure 9. Results of *k*-means clustering on container ship deadweight tonnage–length between perpendiculars (m). (a) Number of clusters selected by NbClust; (b) Hubert statistic values; (c) D index value; (d) scatter plot clustered into three groups.

Figure 10. Results of *k*-means clustering on container ship deadweight tonnage–breadth (m). (a) Number of clusters selected by NbClust; (b) Hubert statistic values; (c) D index value; (d) scatter plot clustered into three groups.

Figure 11. Results of *k*-means clustering on container ship deadweight tonnage–draft (m). (a) Number of clusters selected by NbClust; (b) Hubert statistic values; (c) D index value; (d) scatter plot clustered into three groups.

3. Results

3.1. Result of K-Means Clustering

As the main target data in this study are the tonnage variables (GT and DWT) and length variables (LOA, Lpp, B, and d), the correlation between each variable is generally high. Therefore, there is a possibility that some predictors in regression analysis have a high degree of correlation with other predictors, which may cause a negative effect (multicollinearity). Considering this, in this study, *k*-means clustering was performed for each dimension by adding an age variable to reflect linear change over time. Figure 8 presents the analysis results of deriving the optimal *k* value of a container ship DWT–LOA derived using the NbClust package. Figure 9 presents the results of DWT–Lpp, Figure 10 presents the results of DWT–B, and Figure 11 presents the result of DWT–d. Since the Hubert statistic values and D index values were evaluated by a graphical method, they were numerically provided as 0, and the results were derived as figures. The most suitable index value was obtained based on the change in the index that caused a significant increase in the second difference. Table 4 presents the *k* analysis results for each index provided by NbClust. The results indicate that a *k* value of 3 was recommended as the most suitable value for container ships; therefore, in this study, regression analysis was performed for the main dimensions by dividing the data into three clusters.

Table 4. Results of optimal value of *k* according to *k*-means clustering.

Classification	Length Overall	Length between Perpendiculars	Molded Breadth	Maximum Draft
KL	7	7	7	7
CH	12	12	12	12
Hartigan	4	4	4	4
CCC	2	2	2	2
Scott	3	3	3	3
Marriot	3	3	6	3
TrCovW	3	3	3	3
TraceW	3	3	3	3
Friedman	6	6	6	6
Rubin	12	12	12	12
C index	14	14	14	14
DB	6	6	6	6
Silhouette	2	2	2	2
Duda	3	3	3	3
PseudoT2	3	3	3	3
Beale	2	2	2	2
Ratkowsky	2	2	2	2
Ball	3	3	3	3
PtBiserial	2	2	2	2
Frey	7	7	7	7
McClain	2	2	2	2
Dunn	3	3	3	3
Hubert	5	5	5	5
SD index	6	6	6	6
D index	6	6	6	8
SDBw	14	14	14	14
Result	3	3	3	3

3.2. Design Criteria of Container Ships

In this study, based on the DWT, regression analysis was performed through clustering by the ship's main dimensions (LOA, Lpp, B, d) to derive the specification range satisfying the 75% coverage rate. Table 5 presents the results of the TEU–DWT correlation analysis for container ships. Because the TEU is smaller than the DWT, the DWT is more suitable than the TEU for subdividing and analyzing the size of container ships. The TEU was used to

derive the specification range according to the coverage rate of 75% through correlation analysis with the DWT.

Table 5. Correlation analysis with deadweight tonnage (DWT) and twenty-foot equivalent unit (TEU) according to 75% coverage rate.

Classification	Cargo Unit	$DWT=x \times \text{Cargo Unit}$	Adj. R^2
Container ship	TEU	0.1006	0.9723

As a result of the analysis for each cluster, 0–55,000 DWT (~5000 TEU) was classified in Cluster 1, 55,000–125,000 DWT (~12,000 TEU) were classified in Cluster 2, and 125,000–241,000 DWT (~12,500–24,000 TEU) were classified in Cluster 3. Table 6 shows the results of the regression equation for each cluster according to the coverage rate 75%. As a result, in the graph of Cluster 1, the variation in LOA due to the increase in DWT was larger than in Cluster 2. Cluster 3 was analyzed as a stepwise graph suitable for average value analysis. Figure 12 is a graph of regression analysis by clustered data for DWT–LOA.

Figure 13 is a graph of regression analysis using clustered data for DWT–Lpp. The trend in ship's dimensions obtained using clustered data was similar to that obtained via the DWT–LOA graph.

Figure 12. Cont.

Figure 12. Results of regression analysis. (a) Deadweight tonnage (DWT)-length overall (LOA) graph of Cluster 1 (0–55,000 DWT); (b) Log(DWT)-Log(LOA) graph of Cluster 1 (0–55,000 DWT); (c) DWT-LOA graph of Cluster 2 (55,000–125,000 DWT); (d) Log(DWT)-Log(LOA) graph of Cluster 2 (55,000–125,000 DWT); (e) DWT-LOA graph of Cluster 3 (125,000–135,000 DWT); (f) DWT-LOA graph of Cluster 3 (135,000–155,000 DWT); (g) DWT-LOA graph of Cluster 3 (155,000–180,000 DWT); (h) DWT-LOA graph of Cluster 3 (180,000–241,000 DWT).

Table 6. Comparison of change in length variable according to DWT increase in Clusters 1 and 2.

Classification		Length Overall		Length between Perpendiculars		Molded Breadth		Maximum Draft	
		α	β	α	β	α	β	α	β
Cluster 1	$y = \alpha x^\beta$	3.90	0.39	3.39	0.39	1.62	0.29	0.43	0.32
	$\log y = \log \alpha + \beta \log x$	0.70	0.36	0.64	0.37	0.25	0.28	−0.48	0.36
Cluster 2	$y = \alpha x^\beta$	12.72	0.28	11.91	0.28	-	-	2.33	0.16
	$\log y = \log \alpha + \beta \log x$	1.10	0.28	1.07	0.29	-	-	0.36	0.16

Figures 14 and 15 are a graph of regression analysis by clustered data for DWT-B. As a result, it can be seen that B increases stepwise as DWT increases continuously after 35,000 DWT, and the reason is related to the cargo unit of the container ship (TEU). An increase in the breadth of a container ship has a close effect on an increase in the loading capacity. Considering the 20-ft (or 40-ft) container's width (2.44 m) and the size of the lashing bridge, which is the cargo loading structure of the container ship, it was analyzed that breadth increases by an average of approximately 3 m each increase stepwise.

Figure 13. Results of regression analysis. (a) Deadweight tonnage (DWT)-length between perpendiculars (Lpp) graph of Cluster 1 (0–55,000 DWT); (b) Log(DWT)-Log(Lpp) graph of Cluster 1 (0–55,000 DWT); (c) DWT-Lpp graph of Cluster 2 (55,000–125,000 DWT); (d) Log(DWT)-Log(Lpp) graph of Cluster 2 (55,000–125,000 DWT); (e) DWT-Lpp graph of Cluster 3 (125,000–135,000 DWT); (f) DWT-Lpp graph of Cluster 3 (135,000–155,000 DWT); (g) DWT-Lpp graph of Cluster 3 (155,000–180,000 DWT); (h) DWT-Lpp graph of Cluster 3 (180,000–241,000 DWT).

Figure 14. Results of regression analysis. (a) Deadweight tonnage (DWT)-breadth (B) graph of Cluster 1 (0–35,000 DWT); (b) Log(DWT)-Log(B) graph of Cluster 1 (0–35,000 DWT); (c) DWT-B graph of Cluster 1 (35,000–50,000 DWT); (d) DWT-B graph of Cluster 1 (50,000–55,000 DWT); (e) DWT-B graph of Cluster 2 (55,000–65,000 DWT); (f) DWT-B graph of Cluster 1 (65,000–90,000 DWT); (g) DWT-B graph of Cluster 2 (90,000–105,000 DWT); (h) DWT-B graph of Cluster 2 (105,000–125,000 DWT).

Figure 15. Results of regression analysis. (a) Deadweight tonnage (DWT)-breadth (B) graph of Cluster 3 (125,000–145,000 DWT); (b) DWT-B graph of Cluster 3 (145,000–170,000 DWT); (c) DWT-B graph of Cluster 3 (170,000–190,000 DWT); (d) DWT-B graph of Cluster 3 (190,000–200,000 DWT); (e) DWT-B graph of Cluster 3 (200,000–241,000 DWT).

Figure 16 is a graph of regression analysis by clustered data for DWT-d. As a result, in the graph of Cluster 1, the variation in LOA due to the increase in DWT was larger than in Cluster 2. In Cluster 3, the variation was analyzed to be almost constant at around 1 m.

In this study, to examine whether the results of the linear regression analysis of the log-transformed graph above were properly derived, the equal variance of the residuals and the normality of the distribution were reviewed. The analysis result of the Normal Q-Q plot is shown in Figure 17. Homoscedasticity refers to the homogeneity of variance, and it means that scores are equally spread across different groups to be compared. As the results show, the residual distribution of Cluster 1 and Cluster 2 was analyzed to have homoscedasticity.

Figure 16. Results of regression analysis. (a) Deadweight tonnage (DWT)-draft (d) graph of Cluster 1 (0–55,000 DWT); (b) Log(DWT)-Log(d) graph of Cluster 1 (0–55,000 DWT); (c) DWT- d graph of Cluster 2 (55,000–125,000 DWT); (d) Log(DWT)-Log(d) graph of Cluster 2 (55,000–125,000 DWT); (e) DWT- d graph of Cluster 3 (125,000–165,000 DWT); (f) DWT- d graph of Cluster 3 (165,000–200,000 DWT); (g) DWT- d graph of Cluster 3 (200,000–241,000 DWT).

Figure 17. Results of log transformed linear regression graph residual homoscedasticity. (a) Dead-weight tonnage (DWT)-Length overall (LOA) graph of Cluster 1; (b) DWT-Length between perpendiculars (Lpp) graph of Cluster 1; (c) DWT-breadth (B) graph of Cluster 1; (d) DWT-draft (d) graph of Cluster 1; (e) DWT-LOA graph of Cluster 2; (f) DWT-Lpp graph of Cluster 2; (g) DWT-d graph of Cluster 2.

Normal Q-Q plot is a method to visually check whether the normality assumption is satisfied. The Q-Q plot can be said to satisfy normality if the values are distributed along the diagonal reference line. Analysis result of Q-Q plot is shown in Figure 18. As the result, the log-transformed graph of this study was analyzed to satisfy normality. According to the results, it was analyzed that both homoscedasticity and normality of the residuals were satisfied through linear regression analysis.

Figure 18. Results of normal Q-Q plot. (a) Deadweight tonnage (DWT)-Length overall (LOA) graph of Cluster 1; (b) DWT-Length between perpendiculars (Lpp) graph of Cluster 1; (c) DWT-breadth (B) graph of Cluster 1; (d) DWT-draft graph of Cluster 1; (e) DWT-LOA graph of Cluster 2; (f) DWT-Lpp graph of Cluster 2; (g) DWT-d graph of Cluster 2.

3.3. Comparison with Previous Study

In this study, the results of the study that classified groups from the subjective view-point of the researcher without clustering [15] were compared with respect to the range of adjusted R-squared value and residual standard deviation value of linear graph of the regression analysis results for each cluster. Table 7 shows the accurate comparison results of the previous study and this study. As a result of obtaining the p-value for the linear regression analysis using the R studio program version 4.1.1, the relationship between the variables was analyzed to have a significant result by showing a p-value smaller than 2.2×10^{-16} in all cases.

Table 7. Comparison of accuracy of results with previous study.

Classification		Adjusted R^2		Residual Standard Deviation		
		This Study	Previous Study [15]	This Study	Previous Study [15]	
0–55,000 DWT (Cluster 1)	$y = ax$	Log(LOA)	0.93	0.94–0.95	0.03	0.02
		Log(Lpp)	0.94	0.94–0.95	0.02	0.02
		Log(B)	0.91	0.88–0.90	0.02	0.02
		Log(d)	0.91	0.46–0.78	0.04	0.04
	$y = a_0$	LOA	-	-	-	10.13–15.12
		Lpp	-	-	-	9.59–14.75
		B	-	-	1.45–1.76	1.00–1.65
		d	-	-	-	0.71–0.73
55,000–125,000 DWT (Cluster 2)	$y = ax$	Log(LOA)	0.63	-	0.03	-
		Log(Lpp)	0.62	-	0.02	-
		Log(B)	-	-	-	-
		Log(d)	0.66	-	0.01	-
	$y = a_0$	LOA	-	-	-	7.98–21.99
		Lpp	-	-	-	7.83–15.97
		B	-	-	1.63–3.48	1.59–3.62
		d	-	-	-	0.59–0.99
125,000–241,000 DWT (Cluster 3)	$y = a_0$	LOA	-	-	3.09–12.93	1.19–21.99
		Lpp	-	-	2.87–12.04	3.30–9.45
		B	-	-	1.18–2.42	1.67–3.29
		d	-	-	0.33–0.38	0.47–0.64

In this study, more than 1896 ships were added and analyzed compared to the previous study. In previous studies, the analysis target sections were divided according to the subjective intuition of the researcher. Therefore, the least adjusted R-squared value of $\log(\text{DWT}) - \log(d)$ was 0.46 m. Further, in the section of 55,000 DWT or more, only the $y = a_0$ graph having a slope of 0 was analyzed and the maximum deviation in the LOA was 21.99 m. A deviation of ~9 m was observed compared to the LOA maximum standard deviation of 12.93 m in this study. The maximum standard deviation corresponding to Cluster 3 in this study is similar to the external standard of a 40-ft container (Length: 12.19 m, width: 2.44 m). The size of the 40-ft container is closely related to the trend in the large ship because the size of the ship is determined according to the cargo characteristics.

3.4. Prediction of Trends for Mega-Sized Container Ship

In this study, the size of future mega-sized container ships was predicted by analyzing the main dimensions of container ships. Table 8 presents the results for a 75% coverage rate calculated through regression analysis.

Table 8. Standard design of container ship according to a 75% coverage rate.

Classification	DWT (ton)	TEU	LOA (m)	Lpp (m)	B (m)	d (m)
Small Feeder	3000	302	86	80	15.9	5.4
	5000	503	105	98	18.4	6.4
	10,000	1006	138	129	22.4	8.0
Large Feeder	20,000	2012	180	169	27.3	10.0
Panamax	30,000	3018	211	198	30.6	11.3
	40,000	4024	235	222	33.1	12.4
	50,000	5030	257	243	33.1	13.3
	55,000	5533	266	252	34.2	13.8
	60,000	6036	275	261	36.9	14.0
Post Panamax	70,000	7042	297	283	40.7	14.0
	85,000	8551	313	299	40.7	14.4
Super Post Panamax	100,000	10,060	328	313	45.1	14.8
	120,000	12,072	336	329	47.8	15.2
Very Large Container Ship	130,000	13,078	348	333	48.9	16.0
	140,000	14,084	368	352	48.9	16.0
	165,000	16,599	378	361	51.2	16.0
	180,000	18,108	378	361	56.1	16.3
Ultra Large Container Ship	200,000	20,120	400	385	59.5	16.3
	240,000	24,144	400	385	61.0	16.7

The results indicate that a larger ship size corresponds to a larger change in B than in LOA and Lpp. Table 9 shows the average and standard deviation values of changes in LOA and B of the ship according to the change in the ship's main dimensions whenever 10,000-DWT is increased for each cluster. As a result, the average change in LOA according to the 10,000-DWT increase in Cluster 1 was 29.29 m, and the average change in B was 2.91 m. It was analyzed that the change according to the 10,000-DWT increase was the largest among all clusters. For Cluster 2, LOA was analyzed as 9.13 m and B as 1.71 m. In the case of Cluster 3, LOA was 4.73 m and B was 1.10 m, showing the smallest change when increasing 10,000-DWT. The coefficient of variation is a measure of the ratio of the standard deviation to the mean value, and since it is a relative value, it can be used to compare variations between distributions. That is, the higher the coefficient of variation, the greater the deviation of the length change. In the case of Cluster 3, the coefficient of variation was analyzed to be high, and this section is a section in which the main dimensions increase step by step as the size of the ship increases.

Draft does not increase further and remains constant at approximately 16.0–16.7 m. This is because major container piers worldwide cannot accommodate ships with a maximum d that exceeds 17.0 m. Table 10 presents the maximum depth at major container ports such as Rotterdam Port, Hong Kong Port, Shanghai Port, Singapore Port, and Busan New Port. According to Park and Suh [4], if a 25,000-TEU container ship has a maximum d of 16.9 m, the only port that would satisfy the water depth requirements even at high tide is Singapore Port.

Table 9. Results of analysis on changes in length overall and breadth of ships according to 10,000.

Classification	Change in Length Overall per 10,000 DWT			Change in Breadth per 10,000 DWT		
	Mean	Standard Deviation	Coefficient of Variation	Mean	Standard Deviation	Coefficient of Variation
Cluster 1	29.29	7.21	0.25	2.91	1.23	0.42
Cluster 2	9.13	3.02	0.33	1.71	1.76	1.03
Cluster 3	4.73	8.19	1.73	1.10	1.65	1.50

DWT increase in each cluster.

Table 10. Terminal water depth at high tide and required water depth for 25,000-TEU container ship (Park and Suh [4]).

Terminal Name	Terminal Water Depth (m)	Time Windows of High Tide	MHHW (High Tide) (m)	Terminal Water Depth during High Tide (m)	Maximum Draft of 25,000-TEU Ship (m)	Required Water Depth (Maximum Draft $\times$ 1.3)
Rotterdam	20.0	3:03 p.m. (CEST)	1.88	21.88	16.9	21.97
Hongkong	15.5	7:11 p.m. (HKT)	1.73	17.23	16.9	21.97
Shanghai	15.0	10:48 p.m. (CST)	2.03	17.03	16.9	21.97
Singapore	20.0	10:11 p.m. (UTC+8)	2.28	22.28	16.9	21.97
Busan	20.0	6:56 p.m. (KST)	1.03	21.03	16.9	21.97

In this study, the maximum d was fixed at 17 m for the trend analysis of mega-sized container ships. The increases in LOA, Lpp, and B were set based on the size of a 40-ft container. This is because 40-ft containers are the most common and constitute ~90% of all shipping containers. Because container shipping transports 90% of the world's freight, over 80% of the world's freight moves via 40-ft containers. The increase in the number of rows, which is affected by B, was optimized for enlargement; the increase in the number of bays, which is affected by the LOA and Lpp, was set as sub-optimal for enlargement. It is assumed that LOA and Lpp increased by 4.73 m for every 1000-TEU increase in consideration of the analysis result of the length variable per 10,000-DWT in Cluster 3, and B increased by 1.10 m (refer to Table 9). However, Cluster 3 maintained a constant value and increased stepwise at once. Therefore, considering 40-ft containers' length and width, (12.19 m $\times$ 2.44 m), LOA and Lpp were set to increase by 14.19 m per 3000-TEU, and B was set to increase by 2.20 m per 2000-TEU. The results are presented in Table 11. The results reveal that a 30,000-TEU container ship is predicted to have a LOA of 428.4 m, B of 67.6 m, and d of 17.0 m.

Table 11. Prediction results for mega-sized container ships.

TEU	LOA (m)	Lpp (m)	B (m)	D (m)	Maximum Bay Number ¹	Maximum Row Number
27,000	414.2	399.2	63.2	17.0	98	25
28,000	414.2	399.2	65.4	17.0	98	26
29,000	428.4	413.4	65.4	17.0	102	26
30,000	428.4	413.4	67.6	17.0	102	27

¹ Bay numbering is based on 40-ft container.

4. Discussion

Currently, the size of ships is increasing at a rapid pace [58]. Many experts in the field have recognized the importance of this development, and various studies on 30,000-TEU container ships have been conducted [2–4]. However, based on the LOA of the 30,000-TEU class container ship, the deviation of 83.9 m in the predicted range was observed from

453.0 m to 536.9 m. Because changes in ship dimensions determine the physical requirements of port facilities and equipment, results of only simple linear regression analysis have limitations in predicting the trend of mega-sized ships. In this study, the standard design of container ships was evaluated through regression analysis, where *k*-means clustering was used to classify 5497 container ships in actual operation into groups. The results revealed that a 30,000-TEU container ship is predicted to have a LOA of 428.4 m, B of 67.6 m, and d of 17.0 m. Compared to the dimensions provided in previous studies, our obtained dimensions indicate a smaller LOA, larger B, and smaller d. Previous studies that suggested the specifications of large container ships only presented predicted values through linear regression analysis, targeting the ship's main dimensions of container ships. This study paid attention to the fact that the amount of change in the ship's main dimensions (LOA, Lpp, B, d) increasing varies according to the increase in the size of container ships. This research was able to obtain more accurate results that fit reality in that it analyzed clustered groups using the *k*-means clustering technique. However, because d was fixed to 17.0 m in this study, this result may change if the dredging of major container ports in the future is considered. Nevertheless, further research on ship dimensions can minimize the social cost of designing ports to accommodate the future enlargement of ships.

5. Conclusions

Because of the increase in port trading volume, ships are being enlarged for efficient cargo transportation. For future large ships, if the LOA and breadth cannot secure the safety of passage in the target sea area, costs such as port extension and breakwater reconstruction will occur, and additional social costs will be incurred during the consultation process. In this study, statistical methods were used to predict future trends for the main dimensions of container ships. The main dimensions of a container ship corresponding to a coverage rate of 75% were selected as standard, and the changes to the main dimensions were analyzed to predict the future trends considering the size of the 40-ft cargo container. Thus, this study effectively predicts the trend in large ships and suggests a standard ship design for accurate marine traffic flow and mooring safety evaluation. Further, this study realizes the efficient arrangement of navigational aids and the installation of an appropriate fender. Future research on the main ship dimensions will focus on considering a wider variety of ship types, not just container ships. This should be expected to obtain appropriate design values for port facilities, which will further strengthen the safety of ships.

Author Contributions: Conceptualization, I.-S.C. and W.-J.S.; methodology, I.-S.C. and W.-J.S.; software, W.-J.S.; validation, I.-S.C. and W.-J.S.; formal analysis, W.-J.S.; investigation, W.-J.S.; resources, W.-J.S.; data curation, W.-J.S.; writing—original draft preparation, W.-J.S.; writing—review and editing, I.-S.C. and W.-J.S.; visualization, W.-J.S.; supervision, I.-S.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: This study was conducted with data provided by the Korea Ports & Harbours Association ship main dimension research in 2021.

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

References

1. MOF (Ministry of Oceans and Fisheries of Korea) Korea. 2030 Port Policy Direction and Implementation Strategy. Available online: https://www.mof.go.kr/jfile/readDownloadFile.do?fileId=MOF_ARTICLE_36220&fileSeq=1 (accessed on 23 November 2021).
2. Korea Maritime Institute. *The Report on Technology Development of Smart Green Container Terminal*; Ministry of Land Transport and Maritime: Seoul, Korea, 2012; pp. 1–343.
3. Cho, S.W.; Won, S.H.; Lee, J.H. The Evolution of Container Vessel Sizes and its Impact on the Vessel Specifications. *JSL* **2015**, *31*, 507–528. [[CrossRef](#)]

4. Park, N.K.; Suh, S.C. Tendency toward Mega Containerships and the Constraints of Container Terminals. *J. Mar. Sci. Eng.* **2019**, *7*, 131. [CrossRef]
5. Sánchez, R.J.; Perrotti, D.E.; Gomez Paz Fort, A. Looking into the future ten years later: Big full containerships and their arrival to south American ports. *J. Shipp. Trade* **2021**, *6*, 2. [CrossRef]
6. Prokopowicz, A.K.; Berg-Andreassen, J. An evaluation of current trends in container shipping industry, very large container ships (VLCSSs), and port capacities to accommodate TTIP increased trade. *Transp. Res. Procedia* **2016**, *14*, 2910–2919. [CrossRef]
7. KR (Korean Register of Shipping). Specification of HMM ALGECIRAS. Available online: http://www.krs.co.kr/eng/ship_as_address/regist_read.aspx?s_code=0103040500&ClassNo=2000023 (accessed on 23 November 2021).
8. MOF. *Harbour and Fishery Design Criteria*; Ministry of Oceans and Fisheries of Korea: Sejong, Korea, 2020.
9. Takahashi, H.; Goto, A.; Abe, M. *Study on Standards for Main Dimensions of the Design Ship*; Technical Note of National Institute for Land and Infrastructure Management Ministry of Land; Infrastructure and Transport, No. 309; NILIM: Tsukuba, Japan, 2006; pp. 1–97.
10. Katayama, T.; Furuhashi, K.; Moto, K.; Hayafuji, H. *Study on the Interrelations among the Several Dimensions of Ships*; Technical Note of Port and Harbour Research Institute, No. 101; Port and Harbour Research Institute: Yokosuka, Japan, 1970; pp. 1–130.
11. Terauchi, K.; Yoshida, Y. *Analysis on the Interrelations among the Several Dimensions of Ships*; Technical Note of Port and Harbour Research Institute, No. 348; Port and Harbour Research Institute: Yokosuka, Japan, 1980; pp. 1–115.
12. Ajiki, K.; Negi, T.; Murata, T. *Statistical Analysis on Staple Dimension of Ship*; Technical Note of Port and Harbour Research Institute; No. 652; Port and Harbour Research Institute: Yokosuka, Japan, 1989; pp. 1–43.
13. Akakura, Y.; Takahashi, H.; Nakamoto, T. *Statistical Analysis of Ship Dimensions for Standard Size of Design Ship*; Technical Note of Port and Harbour Research Institute, No. 910; Port and Harbour Research Institute: Yokosuka, Japan, 1998; pp. 1–23.
14. Inoue, G.; Akakura, Y. *Study on Ship Dimensions by Statistical Analysis*; Technical Note of National Institute for Land and Infrastructure Management Ministry of Land; Infrastructure and Transport, No. 600; NILIM: Tsukuba, Japan, 2010; pp. 1–97.
15. Iwasaki, K.; Yamagata, S. *Study on Ship Dimensions by Statistical Analysis*; Technical Note of National Institute for Land and Infrastructure Management Ministry of Land; Infrastructure and Transport, No. 991; NILIM: Tsukuba, Japan, 2017; pp. 1–152.
16. MLIT. *Technical Standards and Commentaries for Port and Harbour Facilities in Japan*; National Institute for Land and Infrastructure Management: Tsukuba, Japan, 2020.
17. Del Estado, P. *Recommendations for the Design of the Maritime Configuration of Ports, Approach Channels and Harbor Basins*; Ministerio de Fomento: Madrid, Spain, 2007.
18. PIANC (The World Association for Waterborne Transport Infrastructure). *MarCom Working Group 121: Harbour Approach Channels Design Guidelines*; PIANC: Brussels, Belgium, 2014.
19. PIANC (The World Association for Waterborne Transport Infrastructure). *MarCom Working Group 184: Design Principles for Dry Bulk Marine Terminals*; PIANC: Brussels, Belgium, 2019.
20. García, S.; Luengo, J.; Herrera, F. *Data Preprocessing in Data Mining (Vol. 72)*, 1st ed.; Springer International Publishing: Cham, Switzerland, 2015; pp. 1–320.
21. Mirović, M.; Miličević, M.; Obradović, I. Big data in the maritime industry. *Nase More* **2018**, *65*, 56–62. [CrossRef]
22. Lee, H.T. Analysis of Factors Influencing Berthing Velocity of Ship using Machine Learning Prediction Algorithm. Master's Thesis, Korea Maritime and Ocean University, Busan, Korea, 2019.
23. Son, W.J.; Lee, J.S.; Lee, H.T.; Cho, I.S. An investigation of the ship safety distance for bridges across waterways based on traffic distribution. *J. Mar. Sci. Eng.* **2020**, *8*, 331. [CrossRef]
24. Lee, H.T.; Lee, J.S.; Yang, H.; Cho, I.S.; An, A.I.S. An AIS Data-driven approach to analyze the pattern of ship trajectories in ports using the DBSCAN algorithm. *Appl. Sci.* **2021**, *11*, 799. [CrossRef]
25. Lee, J.S.; Son, W.J.; Lee, H.T.; Cho, I.S. Verification of Novel Maritime Route Extraction using Kernel Density Estimation Analysis with Automatic Identification System Data. *J. Mar. Sci. Eng.* **2020**, *8*, 375. [CrossRef]
26. Lee, H.T.; Lee, S.W.; Cho, J.W.; Cho, I.S. Analysis of Feature Importance of Ship's Berthing Velocity Using Classification Algorithms of Machine Learning. *J. Korean Soc. Mar. Environ. Saf.* **2020**, *26*, 139–148. [CrossRef]
27. UNCTAD (United Nations Conference on Trade and Development). *Review of Maritime Transport*; United Nations: Geneva, Switzerland, 2021.
28. Wang, G.C.S.; Jain, C.L. *Regression Analysis: Modeling & Forecasting*; Graceway Publishing Company: New York, NY, USA, 2003; pp. 1–293.
29. Huang, Z. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Min. Knowl. Discov.* **1998**, *2*, 283–304. [CrossRef]
30. Lee, H.T.; Lee, J.S.; Cho, J.W.; Yang, H.; Cho, I.S. A Study on the Pattern of Pilot's Maneuvering using K-means Clustering of Ship's Berthing Velocity. *J. CDP* **2020**, *7*, 221–232. [CrossRef]
31. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed.; Springer: Berlin/Heidelberg, Germany, 2009; p. 509.
32. Charrad, M.; Ghazzali, N.; Boiteau, V.; Niknafs, A. NbClust: An R package for determining the relevant number of clusters in a data set. *J. Stat. Softw.* **2014**, *61*, 1–36. [CrossRef]
33. Krzanowski, W.J.; Lai, Y.T. A Criterion for Determining the Number of Groups in a Data Set Using Sum-of-Squares Clustering. *Biometrics* **1988**, *44*, 23–34. [CrossRef]

34. Calinski, T.; Harabasz, J. A Dendrite Method for Cluster Analysis. *Commun. Stat.-Theory Methods* **1974**, *3*, 1–27. [[CrossRef](#)]
35. Hartigan, J.A. *Clustering Algorithms*; John Wiley & Sons, Inc.: New York, NY, USA, 1975; pp. 1–351.
36. Sarle, W.S. *SAS Technical Report A-108, Cubic Clustering Criterion*; SAS Institute Inc.: Cary, NC, USA, 1983; pp. 1–51.
37. Scott, A.J.; Symons, M.J. Clustering Methods Based on Likelihood Ratio Criteria. *Biometrics* **1971**, *27*, 387–397. [[CrossRef](#)]
38. Marriot, F.H.C. Practical Problems in a Method of Cluster Analysis. *Biometrics* **1971**, *27*, 501–514. [[CrossRef](#)]
39. Milligan, G.W.; Cooper, M.C. An Examination of Procedures for Determining the Number of Clusters in a Data Set. *Psychometrika* **1985**, *50*, 159–179. [[CrossRef](#)]
40. Friedman, H.P.; Rubin, J. On Some Invariant Criteria for Grouping Data. *J. Am. Stat. Assoc.* **1967**, *62*, 1159–1178. [[CrossRef](#)]
41. Hubert, L.J.; Levin, J.R. A General Statistical Framework for Assessing Categorical Clustering in Free Recall. *Psychol. Bull.* **1976**, *83*, 1072–1080. [[CrossRef](#)]
42. Davies, D.L.; Bouldin, D.W. A Cluster Separation Measure. *IEEE PAMI* **1979**, *1*, 224–227. [[CrossRef](#)]
43. Rousseeuw, P.J. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53–65. [[CrossRef](#)]
44. Duda, R.O.; Hart, P.E.; Stork, D.G. *Pattern Classification and Scene Analysis*, 2nd ed.; John Wiley & Sons, Inc.: New York, NY, USA, 1995; pp. 1–69.
45. Beale, E.M.L. *Cluster Analysis*; Scientific Control Systems Ltd.: London, UK, 1969.
46. Ratkowsky, D.A.; Lance, G.N. A Criterion for Determining the Number of Groups in a Classification. *Aust. Comput. J.* **1978**, *10*, 115–117. Available online: <http://hdl.handle.net/102.100.100/300266?index=1> (accessed on 21 November 2021).
47. Ball, G.H.; Hall, D.J. *ISODATA: A Novel Method of Data Analysis and Pattern Classification*; Stanford Research Institute: Menlo Park, CA, USA, 1965; pp. 1–61.
48. Milligan, G.W. An Examination of the Effect of Six Types of Error Perturbation on Fifteen Clustering Algorithms. *Psychometrika* **1980**, *45*, 325–342. [[CrossRef](#)]
49. Milligan, G.W. A Monte Carlo Study of Thirty Internal Criterion Measures for Cluster Analysis. *Psychometrika* **1981**, *46*, 187–199. [[CrossRef](#)]
50. Kraemer, H.C. *Biserial Correlation*; John Wiley & Sons, Inc.: New York, NY, USA, 2006; pp. 276–279.
51. Frey, T.; Van Groenewoud, H. A Cluster Analysis of the D² Matrix of White Spruce Stands in Saskatchewan Based on the Maximum-Minimum Principle. *J. Ecol.* **1972**, *60*, 873–886. [[CrossRef](#)]
52. McClain, J.O.; Rao, V.R. CLUSTISZ: A Program to Test for The Quality of Clustering of a Set of Objects. *J. Mark. Res.* **1975**, *12*, 456–460. Available online: <https://www.jstor.org/stable/3151097> (accessed on 2 December 2021).
53. Dunn, J. Well Separated Clusters and Optimal Fuzzy Partitions. *J. Cybern.* **1974**, *4*, 95–104. [[CrossRef](#)]
54. Hubert, L.J.; Arabie, P. Comparing Partitions. *J. Classif.* **1985**, *2*, 193–218. [[CrossRef](#)]
55. Halkidi, M.; Vazirgiannis, M.; Batistakis, I. Quality Scheme Assessment in the Clustering Process. In *Principles of Data Mining and Knowledge Discovery, Proceedings of the 4th European Conference, Lyon, France, 13–16 September 2000*; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany; Volume 1910, pp. 265–276.
56. Lebart, L.; Morineau, A.; Piron, M. *Statistique Exploratoire Multidimensionnelle*; Dunod: Paris, France, 1995; pp. 1–439.
57. Halkidi, M.; Vazirgiannis, M. Clustering Validity Assessment: Finding the Optimal Partitioning of a Data Set. In *Proceedings of the 2001 IEEE International Conference on Data Mining, San Jose, CA, USA, 29 November–2 December 2001*; pp. 187–194.
58. Lee, Y.S.; Ahn, Y.J. A Study on the Standard Ship's Length of Domestic Trade Port. *J. Korean Soc. Mar. Environ. Saf.* **2013**, *19*, 164–170. [[CrossRef](#)]