한국항해항만학회지 제46권 제2호
J. Navig. Port Res. Vol. 46, No. 2 : 92-98, April 2022 (ISSN:1598-5725(Print)/ISSN:2093-8470(Online))
DOI : http://dx.doi.org/10.5394/KINPR.2022.46.2.92

Compression of the Variables Classifying Domestic Marine Accident Data

Deuk-Jin Park · † Hyeong-Sun Yang · ‡ Jeong-Bin Yim*

**Professor, Division of Marine Production System Management, Pukyong National University, 45 Yongso-Ro, Nam-Gu, Busan 48513, Republic of Korea*
† Professor, Division of Navigation Convergence Studies, Korea Maritime and Ocean University, 727 Taejong-Ro, Yeongdo-Gu, Busan 49112, Republic of Korea
‡ Professor, Division of Navigation Science, Mokpo National Maritime University, Mokpo, Jeollanam-do 58628, Republic of Korea

Abstract : *Maritime accidents result in enormous economic loss and loss of life; thus, such accidents must be prevented, and risks must be managed to prevent these occurrences Risk management must be based on statistical evidence such as variables. Because calculating when variables increase statistically can be difficult, compressing the designated variables is necessary to use the maritime accident data in Korea. Thus, in this study, variables of marine accident data are compressed using statistical methods. The date, ship type, and marine accident type included in all maritime accident data were extracted, the number of optimal variables was confirmed using the hierarchical clustering analysis method, and the data were compressed. For the compressed variables, the validity of the data use was statistically confirmed using analysis of variance, and the data of the variables identified using the variable compression method were designated. Consequently, among the monthly and yearly data, statistical significance was confirmed in yearly data, and compression was possible. The significance of the data was confirmed in six and eight types of ships and accidents, respectively, and these were compressed. These results can be directly used for prevention or prediction based on past maritime accident data. Additionally, the data range extracted from past maritime accidents and the number of applicable data will be studied in the future.*

Key words : *maritime accident, risk management, data compression, hierarchical clustering, ANOVA*

1. Introduction

Because maritime accidents cause enormous economic and human losses, their prevention is necessary(Chauvin et al., 2013). To prevent accidents, risks that likely result in maritime accidents must be managed and evaluated (Montewka et al., 2011; Kaplan and Garrick, 1981). To manage the causes of maritime accidents, analyzing and evaluating the accidents that have occurred in past is necessary(Cacciabue, 2004; Senders et al., 1991).

Human error, which accounts for 80% of the causes of accidents in the ocean, exists in various forms(Qiao et al., 2021). York University, UK, applied human reliability analysis technique for human error assessment and reduction technique(HEART) to describe the method of trusting humans. Using this technique, procedures and results for describing and quantitatively evaluating human errors have been described (Smith and Harrison, 2002, Kirwan, 1992). For maritime accident analysis, the International Maritime Organization(IMO) published a guide for human reliability analysis on formal safety assessment(Hu et al., 2007; IMO, 2001).

Furthermore, to analyze the causes of marine accidents by factors, the HEART methodology was applied to identify significant causes(BOWO and Furusho, 2018). Cacciabue(2004) used an engineering system management methodology to solve methodological problems related to accident analysis. Thus, an evaluation method and system are required to manage maritime accidents. Besides, the strength of the probability calculation was presented when risks were being evaluated using a Bayesian network.(Abbassinia et al., 2020).

In addition, variables used in the system should be considered in various ways, such as situations and causes.(Rothblum, 2000). Moreover, because using all data is practically difficult(Friedman et al., 1997), compressing the variables using a statistical method is necessary. In the maritime field, when an accident occurs, there are many causal variables that affect the accident. A large number of variables increases the difficulty in analyzing or predicting

† Corresponding author, epikyang@mmu.ac.kr 061)240-7170
‡ Corresponding author, jbyim@kmou.ac.kr 051)410-4246
* pdj@pknu.ac.kr 051)629-5887
Note) This paper was presented on the subject of "A Study on the construction method of marine accident qualification data" in 2016 Joint Conference KINPR proceedings(BEXCO, 19th May-20th May, 2016, pp.194-195).

marine accidents. In the study conducted by Yim(2017), the cause and cause judgment keyword was classified, and the optimized common word was extracted and reduced for simplicity.

Therefore, in this study, to manage maritime accidents using objective data, the variables of maritime accident data for maritime accident situations are compressed using statistical methods.

This paper is structured as follows:

Chapter 2 deals with the collection of maritime accident data and describes the methodology for variable compression. In Chapter 3, the results of the hierarchical clustering analysis of the variables and the statistical significance of the results were verified using analysis of variance(ANOVA). Finally, Chapter 4 summarizes the results.

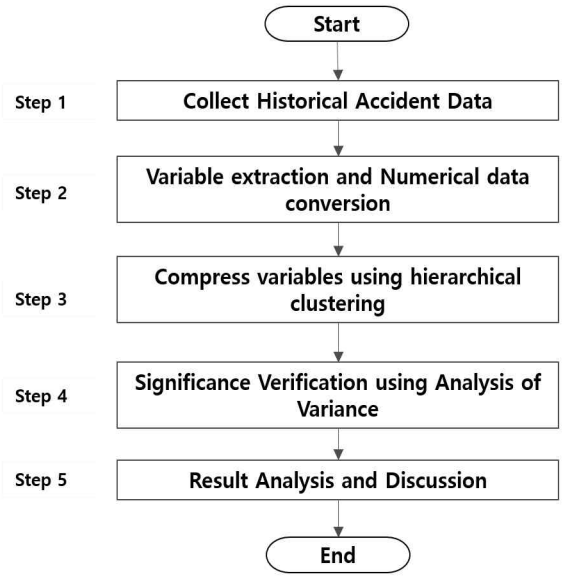

Fig. 1 Study Procedure

2. Method

2.1 Research approach procedure

The research approach procedure is shown in Figure 1. In the first step, maritime accident rulings were collected from the Korean Maritime Safety Tribunal website to acquire the historical maritime accident data. Maritime accidents that occurred over the past decade were confirmed and collected regardless of the region of the judges.

In the second step, data regarding the situation usually including in all rulings, maritime accident type, time, and three types of ships were converted to a numerical database(NDB). The quantification is performed for calculation(Yim 2009) and to avoid the increase in calculation time with the increase in the character data(Yim, 2017).

In the third step, the variables were compressed using hierarchical clustering. The optimal number of variables was confirmed and a bottom-up approach was applied in an agglomerative format.

In the fourth step, ANOVA was used to statistically verify the clustering results. Thus, the validity of data use was statistically verified.

In the last step, the research was comprehensively discussed by analyzing and discussing the results.

2.2 Hierarchical clustering

Variable compression is essential for Bayesian computation and estimation.(Wasserman, 2013). Variables related to maritime accidents are diverse, and limited data distribution exists for data with several variables(Yim, 2009). For example, if “0” data are generated, the probability calculation does not proceed, and an error occurs. Moreover, if there are excess variables, handling the amount of calculation may be difficult(Bradley and Carlin, 2011). Thus, it is necessary to compress variables as little as possible. Hence, the optimal number of variables was confirmed using cohesive hierarchical clustering analysis for variable usability among condition variables.

Compression using clustering was developed by Cilibrasi and Vitány(2005). They reported that if a variable can be moderately compressed when enough relevant information is available, it indicates that the two objects are close(Cilibrasi and Vitány, 2005). If the two objects are close according to any practical similarity, all practical similarities are found considering that they are close according to the normalized information distance.

Herein, we used the method in which if the Euclidean distance of the cluster is small, higher-level cluster is included. Furthermore, here, the commercial program MATLAB, 2018b.(MATLAB, 2019) was used.

2.3 Analysis of Variance

One-way ANOVA was used to test whether the means of two or more populations are equal and evaluate the mean difference between two or more observations. ANOVA uses the sample data as basic data to draw general conclusions about populations. This analysis method investigates the effect of a single factor on the observed value, such as machine productivity, through an experiment based on the principle of randomization. The independent t-test is a special case of the one-way ANOVA for situations in which only the mean values of two means exist. Therefore, when there are only two groups, an independent t-test uses one-way ANOVA but provides the same results. Moreover, the sum of squares is divided into the overall variability of the continuous dependent variable. The two components of variance are variabilities between and within groups, which can be calculated as the sum of squares between groups or the sum of the mean difference and the mean squared difference for each group Eq.(1).

$$\sigma^2 = \sum_{i=1}^{n} I\frac{(X-\overline{X})^2}{n} = \frac{SS}{df} \quad (1)$$

where $\overline{X}$ represents the mean of the group, the measure of X object, SS is sum of squares, and df = degrees of freedom.

3. Variable compression result of maritime accident data

3.1 Maritime accident data collection and extraction results

For maritime accident data, the Korean maritime accident ruling was used. The ruling and its summary provided by the Korea Maritime Safety Tribunal were used(KMST, 2021), and maritime accident data for approximately a decade were collected.

The source of marine accident data is the decision of the Maritime Safety Tribunal. About 10% of marine accidents are judged. In other words, although the characteristics of the entire maritime accident cannot be included in the data of the ruling, the use of the decision of KMST can be limited to domestic cases as all official data.

Table 1 presents the results. Various variables exist in maritime accidents, and these variables are described in the ruling. For example, maritime accident type, ship type, date, position, nationality, license, speed, detection range, weather, and laws are included. Because these variables are excluded in all rulings, only the maritime accident type, date, and ship type included in all rulings were extracted and used.

To accept as data, all data, including character data, were converted into NDB to be used as a variable. For example, for the quantification method, for the monthly data, January was used as 1 and February as 2; for the year, the numbers were used in their original form. Among the types of vessels, 1, 2, 3, 4, 5, and 6 were classified into a fishing vessel, a cargo ship, a passenger ship, an oil tanker, a tugboat, and a special vessel, respectively. Variables with indeed data were used here.

Table 1 Maritime accident data collection and extraction results

Classification	Number of Data
Number of Maritime Accident	1,745
Number of Accidents's Type Variable	17
Number of Ship's Type Variable	25
Number of Month Variable	12

3.2 Hierarchical clustering analysis results

Hierarchical clustering was performed on the collected data for maritime accident types, ship types, and dates.

First, 17 types of maritime accidents were present in the maritime accident report, including collision, rollover, sinking, fire, explosion, safety accidents, contact, stranded, engine damage, suspended wind, obstruction, flooding, propulsion system damage, and helm damage. Figure 2 shows the results of clustering 17 maritime accidents, and Figure 3 shows the maximum number of clusters of eight. Clustering results show eight accidents, including collision, grounding, contact, fire, casualties, sinking, and hull damage(Figure 3). The figure shows the average, maximum, and minimum values of the types of maritime accidents. According to the statistical analysis results, the average of collisions was the highest at 91, and in the box plot graph, there was no significant difference between 2 and 7, but there was a large difference with collisions. Accordingly, the overall analysis results showed that there was a difference in the mean.

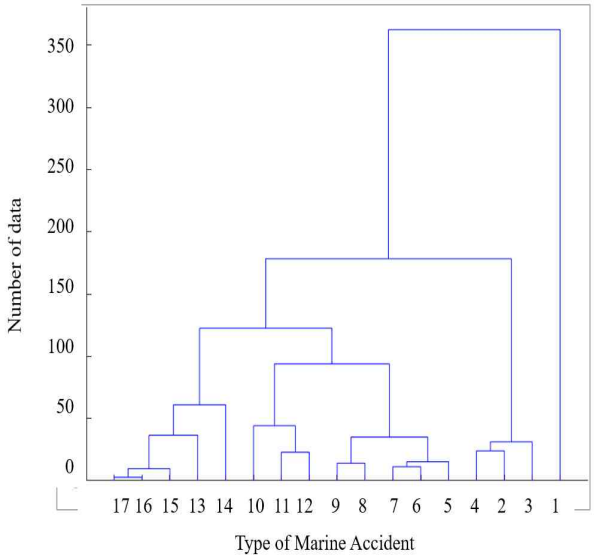

Fig. 2 Clustering result of maritime accident-type variable data

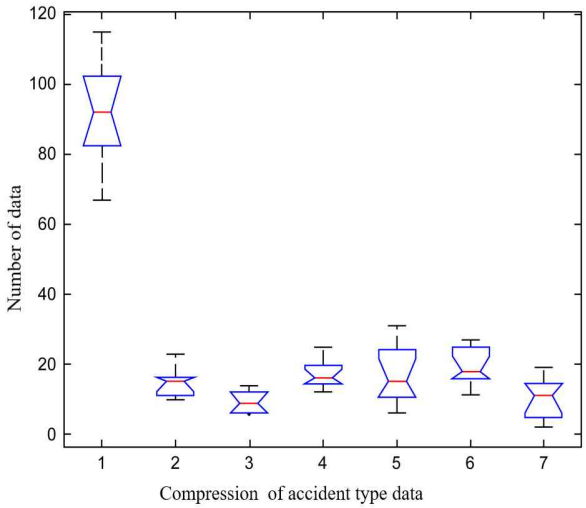

Fig. 3 General statistics of maritime accident-type variable data

Ship types collected 25 types of vessels reported in the maritime accident report. First, fishing vessels were classified into coastal combined fishing vessels, coastal gillnet fishing boats, offshore fishing boats(with reel and line with multiple hooks), and fishing boats. Passenger ships were classified into general passenger ships, car ferry passenger ships, high-speed passenger ships, and conducting vessels with more than 13 passengers.

Cargo ships were classified into general cargo ships, car ferry ships, automobile carriers, refrigeration carriers, container carriers, crude oil carriers, liquefied natural gas carriers, liquefied petroleum gas carriers, and oil-refueling vessels. Tugboats were classified into push-type, towing, and berthing tugs, and other ships included barges, dredgers, fishing guidance ships, test survey ships, fishing ground purification ships, guiding boats, and yachts. Figure 4 shows the clustering result. The 25 types of vessels were divided into six types, and the general statistics are shown in Figure 5. Numbers 1, 2, 3, 4, and 5 represent a fishing vessel, general cargo ship, passenger ship, tanker, and tugboat, respectively. In contrast, all other ships were classified into 6, and the remaining ship data were used. According to the statistical analysis result, the average of fishing vessels was 96, which was the highest, and in the box plot graph, the highest ranking was shown in the order of tug boats and general cargo ships.

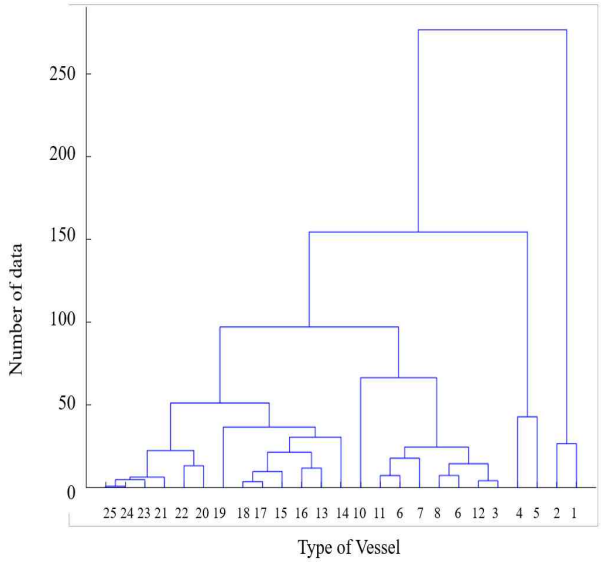

Fig. 4 Clustering results of vessel-type variable data

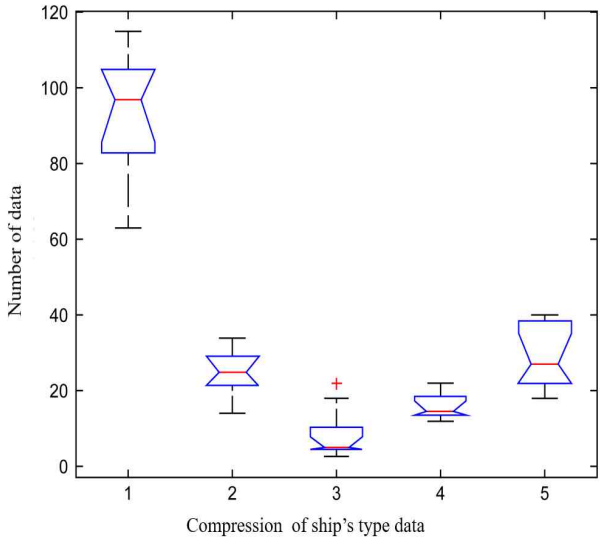

Fig. 5 General statistics of vessel-type variable data

Finally, the yearly and monthly data were clustered for the date, but hierarchical clustering was not established; thus, the data were used unmodified. The collected yearly and monthly data were used unmodified, and as shown in Figure 6, the figures above and below show the yearly and monthly data, respectively. According to the statistical analysis result, the value in 2011 was the highest, and in the box plot graph, it was confirmed that the difference between the years was larger than the difference between the months.

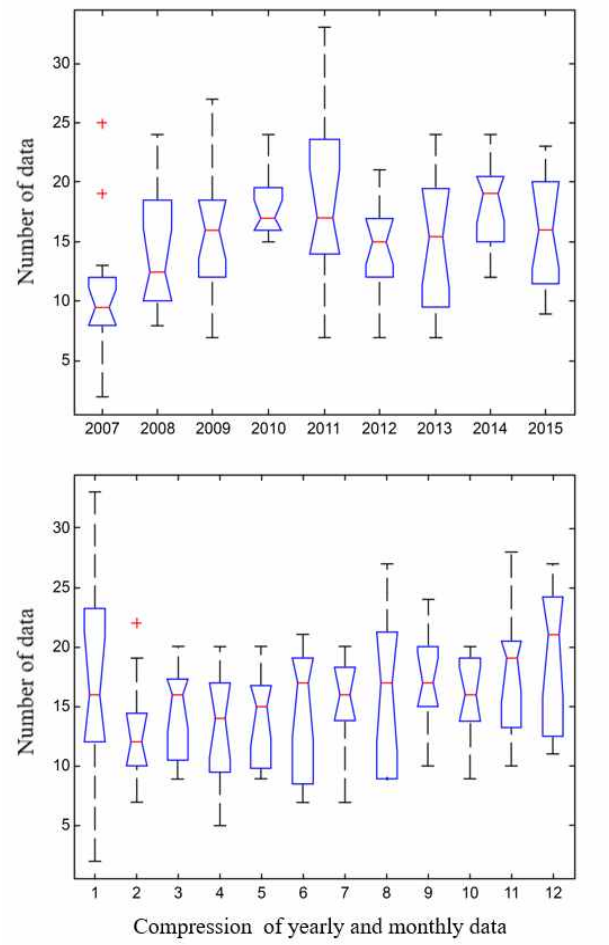

Fig. 6 Clustering results of yearly and monthly variable data

3.3 ANOVA Results

The clustering results were statistically validated using ANOVA to validate the data use. These results were also analyzed using the maritime accident type, vessel type, and date data.

First, for the types of maritime accidents, other accidents were added to seven types to use the excluded data, and the ANOVA results are shown in Table 2. Statistical significance was found because the F and p values were 136.12 and 0, respectively.

Table 2 ANOVA result of maritime accident-type variable data

Source	Sum Square	df	Mean Square	F	p
Type of Accidents	47800.8	6	7966.79	136.12	0.00
Error	3277.6	56	58.53		
Total	51078.3	62			

df, degree of freedom; p, p-value

Second, other types were added to the five types of vessels to use the excluded data, and Table 3 presents the ANOVA results. Statistical significance was found because the F and p values were 120.33 and 0, respectively.

Table 3 ANOVA result of vessel-type variable data

Source	Sum Square	df	Mean Square	F	p
Type of Ships	42066.8	4	10516.7	120.33	0.00
Error	3496.0	40	87.4		
Total	45562.8	440			

df, degree of freedom; p, p-value

Finally, the yearly and monthly data confirmed the statistical significance of the data without compression. Table 4 presents the results. Here statistical significance was found because the F and p values were 2.78 and 0.0087, respectively, for the data per year. However, no significance was found in the monthly data with F and p values of 1.26 and 0.2641, respectively.

Table 4 ANOVA result of yearly and monthly variable data

Source	Sum Square	df	Mean Square	F	p
Month	360.74	11	32.7946	1.26	0.2641
Year	581.02	8	72.6273	2.78	0.0087
Error	2299.43	88	26.1298		
Total	3241.19	107			

df, degree of freedom; p, p-value

3.4 Analysis and discussion of results

Because various variables affect maritime accidents, a study was conducted to compress these variables. In the data collection and classification, a problem occurred with the nonexistent data, “0.” Because all data must be objective, classifying it subjectively is difficult. Solving this problem requires limiting the conditional items to the extent that “0” data do not occur. Another approach is the possibility of emergence using cumulative fractional inclusion(Yim, 2017).

The data used here comprise a maritime accident investigation report by the Maritime Safety Tribunal(KMST, 2021). The maritime accident investigation report provides common data regarding the types of maritime accidents, vessel types, and dates covered herein. Various variables related to other accidents are not provided in common. This part should be improved through a common form of investigation or model of accidents, and further research is required.

Various reports have been presented on assessing the current level of accident or crisis, and various studies have been published to establish standards(Kirwan et al., 2008; Kang, 2014). For probabilistic evaluation of risks related to maritime accidents, because the crisis level cannot be easily known if the crisis is expressed only by probability, several research methods suggest this criterion(Yim, 2009). Furthermore, in response to the opinion that the quantitative evaluation of the crisis is necessary owing to the recent emergence of autonomous ships(Yildiz et al., 2021), discussing how much it can reduce the leading causes of maritime accidents is necessary. Therefore, this study compresses the data to be used to prevent and predict maritime accidents based on past data so that it can be practically used. It can also be directly used in a model or probability because statistical significance has been verified.

However, because variables need to be added or changed depending on the data period or the model or calculation formula to be used later, additional research is required depending on the method to be applied.

4. Conclusion

Herein, the variables of maritime accident data were compressed using a statistical method. The hierarchical clustering analysis method was used to compress the maritime accident types, vessel types, and dates by clustering, and the statistical significance was confirmed through ANOVA. The analysis found significance in eight and six types of maritime accidents and vessels, respectively. Statistical significance was confirmed in the yearly data, thus, the compression was possible. By contrast, no significance was found in the monthly data. This study provides a basis that can be directly considered for prevention or prediction based on past maritime accident data. Securing additional data that exclude the three data types covered herein will increase the reliability of maritime accident prevention or prediction research, which is left for further study.

Acknowledgements

This work was supported by Pukyong National University Research Fund in 2021 (CD20210999)

References

[1] Abbassinia, M., Kalatpour, O., Motamedzade, M., Soltanian, A. and Mohammadfam, I.(2020), Dynamic human error assessment in emergency using fuzzy bayesian cream, Journal of Research in Health Sciences, Vol. 20, No. 1, p. e00468.

[2] Bradley P. and Carlin Thomas,(2011), Baysian network analysis, second edition, andrew gelman et al; baysian models for categorical data, peter congdon; baysian methods for data analysis, third edition.

[3] Bowo, L. P. and Furusho, M.(2018), Human error assessment and reduction technique for reducing the number of marine accidents in Indonesia, In Applied Mechanics and Materials, Vol. 874, pp. 199-206.

[4] Cacciabue, P. C.(2004), Human error risk management for engineering systems: A methodology for design, safety assessment, accident investigation and training, Reliability Engineering System and Safety, Vol. 83, pp. 229-240.

[5] Chauvin, C., Lardjane, S., Morel, G., Clostermann, J. P. and Langard, B.(2013), “Human and organisational factors in maritime accidents: Analysis of collisions at sea using the HFACS”, Accident Analysis and Prevention, Vol. 59, pp. 26-37.

[6] Cilibrasi, R. and Vitányi, P. M.(2005), Clustering by compression, IEEE Transactions on Information theory, Vol. 51, No. 4, pp. 1523-1545.

[7] Friedman, N., Geiger, D. and Goldszmidt, M.(1997), Bayesian network classifiers,. Machine learning, Vol. 29, No. 2, pp. 131-163.

[8] IMO(2001), Formal Safety Assessment, Report on the Joint MSC/MEPC Working Group on the Human Element and Formal Safety Assessment, MSC 74/WP.19.

[9] Kaplan, S. and Garrick, B. J.(1981), On the quantitative definition of risk. Risk analysis, Vol. 1, No. 1, pp. 11-27.

[10] Kirwan, B.(1992), Human error identification in human reliability assessment, Part 1: Overview of approaches. Applied ergonomics, Vol. 23, No. 5, pp. 299-318.

[11] KMST, 2021 Written Verdict. [Online], Available at: https://www.kmst.go.kr/kmst/verdict/writtenVerdict/selectWrittenVerdict.do [Accessed at 10th Mar. 2021].

[12] MATLAB(2019), MATLAB and Statistical Toolbox Release 2018b, The MathWorks, Inc., Natick, Massachusetts, United States.

[13] Montewka, J., Krata, P., Goerlandt, F., Mazaheri, A. and Kujala, P.(2011), Marine traffic risk modelling – an innovative approach and a case study, Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability, Vol. 225, No. 3, pp. 307-322.

[14] Park, D. J., Yang, H. S. and Yim, J. B.(2019), A Study on the Estimation of Optimal Probability Distribution Function for Seafarers' Behavior Error, Journal of Navigation and Port Research, Vol. 43, No. 1, pp. 1-8.

[15] Qiao, W., Liu, Y., Ma, X. and Lan, H.(2021), Cognitive Gap and Correlation of Safety-I and Safety-II: A Case of Maritime Shipping Safety Management. Sustainability, Vol. 13, No. 10, p. 5509.

[16] Rothblum, A. M.(2000), Human error and marine safety, In: National Safety Council Congress and Expo, Orlando, Florida.

[17] Senders, John, W. and Neville, Moray, P.(1991), Human error: cause, prediction, and reduction, Lawrence Erlbaum Associates, New Jersey, USA. p. 25.

[18] Smith, S. P. and Harrison, M. D.(2002), Improving hazard classification through the reuse of descriptive arguments, In International Conference on Software Reuse, Springer, Berlin, Heidelberg.

[19] Wasserman, L.(2013), All of statistics: a concise course in statistical inference, Springer Science & Business Media.

[20] Yildiz, S., Uğurlu, Ö., Wang, J. and Loughney, S.(2021), Application of the HFACS-PV approach for identification of human and organizational factors (HOFs) influencing marine accidents, Reliability Engineering & System Safety, Vol. 208, p. 107395.

[21] Yim, J. B.(2009), Development of Quantitative Risk Assessment Methodology for the Maritime Transportation Accident of Merchant Ship, Journal of Navigation and Port Research, Vol. 33, No. 1, pp. 9-19.

[22] Yim, J. B.(2017), A Study on the Reduction of Common Words to Classify causes of Marine Accidents, Journal of Navigation and Port Research, Vol. 41, No. 3, pp. 109-118.

Received 02 December 2021
Revised 20 December 2021
Accepted 22 December 2021