Characteristics analysis of intercontinental sea accidents using weighted association rule mining: Evidence from the Mediterranean Sea and Black Sea

Daozheng Huang[1*], Tiantian Liang[1], Shenping Hu[2], Sean Loughney[3], Jin Wang[3]

1. College of Transport & Communications, Shanghai Maritime University, Shanghai, 201306, PR China;
2. Merchant Marine College, Shanghai Maritime University, Shanghai, 201306, PR China;
3. Liverpool Logistics, Offshore and Marine, Research Institute, Liverpool John Moores University, Liverpool L3 3AF, UK;

*Corresponding author and E-mail address: dzhuang@shmtu.edu.cn

Abstract

With the continuous growth of international trade and the number of ships, the pressure on marine safety is increasing. To strengthen the efficiency and pertinence of accident prevention and control measures, in-depth research on the characteristics of marine traffic accidents is of great significance. This study uses a Weighted Association Rule Mining (WARM) approach to investigate the association between marine traffic accidents characteristics. The Mediterranean Sea and Black Sea play a crucial role in global maritime transportation. The marine traffic accidents in the Mediterranean Sea and Black Sea region from 2006 to 2020 are collected from the Lloyd's List Intelligence (LLI) database. Then, WARM is applied to obtain the characteristics of marine traffic accidents. The findings show that Flag of Convenience (FOC) vessels in the Mediterranean have a lower accident rate than non-FOC vessels do. Moreover, the results reveal a strong relationship between accidents and the age, gross tonnage, and ship type. Older general cargo ships with a gross tonnage between 500 and 3,000 tons are more prone to accidents. This research provides insights for authorities to develop targeted measures for preventing specific types of accidents and enhancing marine safety.

Keywords: Marine traffic; Association rules; Accidents prevention; Data mining; Weighted transaction data; Weighted Association Rule Mining

1 Introduction

The shipping industry is one of the most important sectors of the global economy, accounting for more than 80% of world trade by volume (UNCTAD, 2022). Due to the significant increase in trade transportation demand, the shipping industry has experienced rapid development in recent years (Chen et al., 2022; Weng and Yang, 2015). This significant transportation demand inevitably leads to more ships engaged in maritime trade, which further leads to an increase in the potential for vessel traffic accidents (Wang et al., 2021). The continued growth of marine traffic has thus led to increasing concerns about shipping safety (Bakdi et al., 2020; Christensen et al., 2022). Furthermore, the increasing vessel traffic accidents may cause severe consequences, such as human injury/fatality, economic loss, property damage and environmental pollution, which put pressure on maritime administration departments (Goerlandt and Kujala, 2011; Wang et al., 2023; Yu et al., 2021). Marine traffic accidents are intricate and complex, and involve various characteristics of the ship and the accident, such as Gross Tonnage (GT), age, ship type, flag state and time of day. However, the

relationship between these characteristics needs to be discussed in terms of the causes of vessel traffic accidents, and how the officers of maritime departments develop effective policies to prevent vessel traffic accidents. Therefore, to provide references for maritime departments to formulate targeted accident prevention measures and establish a sound safety management mechanism, it is worth analysing how these characteristics affect vessel traffic accidents and finding the internal mechanism of vessel traffic accidents.

The Mediterranean and Black Sea waters are intercontinental waters, surrounded by three continents - Europe, Asia, and Africa. They are essential nodes located on the route from the Far east to Europe, one of the three busiest container shipping routes in the world. The waters also play an important role in the 21st century maritime silk road, which is a part of China's Belt & Road initiative. China is actively investing ports around the Mediterranean and the Black seas, such as the Piraeus Port. The region has some of the most hectic shipping lanes in the world. Approximately 20% of ocean traffic and 10% of container traffic passes through here each year (Leone, 2017). According to the Lloyd's List Intelligence (LLI) database, over 5,000 ships were involved in accidents in these waters between 2006 and 2020. Figure 1 shows the distribution characteristics of ship accidents in this region. It can be seen that the Eastern Mediterranean is a highly accident-prone area.

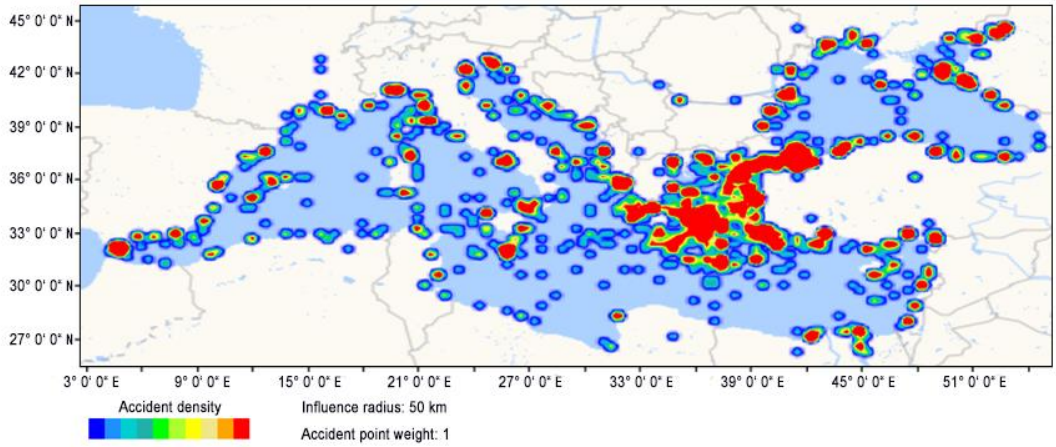


Figure 1: Heat map of accident distribution in the Mediterranean Sea and Black Sea

In recent decades, the analysis on the causes of ship traffic accidents has been a focus of maritime research. A variety of models and methods have been applied to analyse vessel traffic accidents, which provide reasonable suggestions for maritime administration departments to reduce the occurrence of ship traffic accidents (Fu et al., 2018; Liu et al., 2022; Ugurlu et al., 2020; Yang et al., 2009; Zhang et al., 2021). These methods can be used to identify the characteristics that affect the severity of ship accidents excellently, and are mainly based on the logit model, probit model and regression model (Wang et al., 2022; Weng, 2015; Weng et al., 2018; Yang et al., 2018; Zheng et al., 2016). These methods can effectively identify the characteristics that influence the severity of marine traffic accidents, but they do not reveal the associations among the characteristics in maritime accidents (Weng and Li, 2017). Hence, specific attention should be paid to the identification of ship accident characteristics and the association relationship between characteristics. Association analysis is a useful method to find interesting relationships in large-scale data sets. To find the meaningful relationships between the characteristics hidden in the data of marine traffic accidents, association rule mining was applied to the analysis of ship accidents by

Huang and Hu (2018). When analysing marine traffic accidents, traditional association rule mining, like the "apriori" algorithm, assumes that each event in the database has the same importance (Datta et al., 2020). In fact, different ship incidents should have different levels of importance in the database. In terms of the application of association rules to maritime accidents, to the best of the authors' knowledge, no investigation has been conducted to apply weights to each maritime accident to study the association between accident characteristics yet. Furthermore, whether the association characteristics between maritime and intercontinental sea accidents are consistent remains to be discussed. Therefore, to fill this gap in the literature, this study proposes an improved Weighted Association Rule Mining (WARM) approach to the association between characteristics influencing ship accidents to understand the connection between different characteristics of marine traffic accidents. To compute the weight of each accident, Hyperlink-Induced Topic Search (HITS) algorithm is applied. The Mediterranean and Black Sea waters are chosen to illustrate the effectiveness of the approach. In this way, a detailed association of the attributes of the accidents is obtained, which provides a basis for the formulation of prevention and accident control measures.

2 Literature review

2.1 Methods in marine traffic accidents analysis

Novel models and methods have been applied to maritime accident analysis, which have made substantial contributions to knowledge and accident mitigation. Parametric methods are widely used to analyse maritime accidents, such as classical econometric methods (e.g., Probit, Poisson regression, Negative binomial regression and Logit model) to implement maritime traffic accident analysis with improvements and innovations. Some other scholars investigated maritime accidents using non-parametric models (e.g., Fuzzy Analytic Network Process (FANP), Artificial Neural Network (ANN), Fault Tree Analysis (FTA), Human Factor Analysis and Classification System (HFACS), Systems Theoretic Process Analysis (STPA), Systems-Theoretic Accident Model and Processes (STAMP)) (Luo and Shin, 2019). In addition, methods such as deep learning have also been continuously applied to maritime safety research in recent years (Park and Kim, 2022).

Logit regression models are suitable for predicting correlations between a set of independent variables and discrete targets, and therefore, such regression models are often used by scholars to model the probability of ship accidents and estimate the marginal effects of each influencing factor. Weng and Yang (2015) analysed a decade of global ship accidents and developed a binary logistic regression model to predict the probability of fatal ship accidents. The probit model is similar to the logit model and can usually be interchanged. However, if the dependent variable is ordinal, only the ordered probit model can be used for regression. Zheng et al. (2016) developed a probit regression model to study the determinants of the probability of crew injury in fatal and non-fatal accidents of a container ship. Similarly, Wang et al. (2022) proposed a zero-inflated ordered probit model to analyse the severity of maritime accidents. They found that the type of accidents, flag country, and distance from the coast affected the severity of accidents. In addition, some other regression models have been applied in the analysis of shipping accidents. A zero-inflated negative binomial regression model was used by Weng et al. (2016) to assess the factors of human life loss in ship accidents. Li et al. (2022) made use of the Tobit regression model to evaluate the economic loss of maritime accidents. In addition, the fit of the zero-inflated negative binomial regression model has been shown to outperform other regression models in predicting the probability of ship accidents (Wang

et al., 2022). However, it is worth noting that while parametric methods can analyse the effects of multiple factors on ship safety, they often require a large amount of accident data to ensure the accuracy of the fit. The maritime accident database is one of the important sources to obtain information, however, compared to maritime accident reports, such an information base contains less in-depth information. It would be a laborious task to obtain information about the navigational environment at the specific time of the accident, the status of the crew, and the indirect or direct causes of the accident from the accident investigation reports. More importantly, parametric methods usually use a model-based approach in which we have to make assumptions about the form of the estimated function, but the assumptions we make are not necessarily correct.

The non-parametric methods have also been widely applied in maritime accident studies. For example, Chang et al. (2014) used a random occupancy approach to analyse the relationship between each risk factor of container ships and accidents. Qiao et al. (2020) presented an accident analysis model that integrates FTA and ANN for the analysis of maritime accidents. Uğurlu et al. (2020) analysed accidents in the Black Sea based on HFACS and Bayesian Network (BN) model. Kim et al. (2016) presented an analysis of the cruise ship accident based on a STAMP, providing extensive insights into the causal relationship of the accident. To investigate the influence of human factors on maritime traffic accidents, Kokotos and Linardatos (2011) utilized the classification and regression tree method, and Fan et al. (2020) analysed the impact of human factors on maritime safety with data-driven BNs.

Association Rule Mining (ARM) is the analytical process of discovering association relationships between factors. It provides an opportunity to discover item-to-item relationships from a data set containing a large number of variables. ARM can discover relationships between influencing factors in maritime accidents and is applicable to small data sets. It is not essential to assume in advance that the data set fits a certain model. The association rule was originally created to solve the shopping basket problem (Agrawal et al., 1993), and now it has been widely used in risk and safety analysis (Cakir et al., 2021; Montella et al., 2012; Ozaydin et al., 2022; Yang et al., 2018). In order to analyse the connection between the factors influencing marine traffic accidents and to discover the characteristics of marine traffic accidents, ARM has become increasingly popular in the past few years. Huang and Hu (2018) performed association rule analysis on maritime accident data to identify potential relationships between causal factors. With the widespread application of association rules in maritime transportation, they have been combined with other methods to study maritime safety issues. Yang et al. (2018) combined the “apriori” algorithm (an algorithm for association rules) and k-medoids to focus on the risk of ships in Zhejiang waters and its influencing factors. Ozaydin et al. (2022) applied a combination of BNs and association rules to study the factors influencing fishing vessel accidents.

ARM is a competitive approach to maritime accident analysis due to its ability to explore ship accident characteristics using small data sets and its ability to study the links between different influencing factors. Nevertheless, it should be noted that the traditional ARM considers every event in the database to be equally important, and this approach tends to ignore events that occur less frequently but are relatively important. WARM is capable of weighting each maritime incident in the database, preventing some low-frequency but important incidents from being overwhelmed by the large incident data. Therefore, this study uses a WARM approach to study maritime incident characteristics and provide new insights for mining maritime incident characteristics.

2.2 Geographic characteristics in marine traffic accidents analysis

There are mainly two types of studies according to the research scope. One is research associated with global maritime accidents and the other focuses on a particular water area. Studies focusing on global maritime accidents provide a macro-level summary analysis of maritime accidents. Huang et al. (2013) explored the spatial distribution of global maritime accidents using a geographic information system. The area around the United Kingdom, the coastal areas around East Asian countries (such as China, Japan, and South Korea), and the Mediterranean Sea are identified as hot spots. Zhang et al. (2021) also applied geospatial techniques to describe the global maritime accident landscape. The macro analysis with global maritime accidents as the object can summarize the distribution characteristics of maritime accidents and the high incidence areas of accidents. However, the characteristics of maritime accidents in various waters are often different due to distinct natural environments, vessel characteristics and other factors in different waters. Therefore, the macroscopic perspective analysis of maritime accidents does not explain the characteristics of specific maritime accidents exhaustively, which means that a micro-level analysis of specific water areas is needed to understand the causes and consequences of maritime accidents and to provide useful insights for accident prevention.

When researchers focus on analysing the accident characteristics of a particular water area, they can often find interesting phenomena. In Arctic waters, severe weather condition is a critical factor affecting shipping safety (Fu et al., 2018). Dobrzycka-Krahel and Bogalecka (2022) revealed a paradox in the Baltic Sea: despite the heavy Baltic shipping traffic, there are only about 100 maritime accidents per year. In the waters of the South China Sea, shipwrecks caused the most damage compared to other accident types (Weng et al., 2016). Wang et al. (2020) found that vessel types had a significant effect on accident probability in Hong Kong waters. Specific water areas studied in previous research tend to focus on strategic transport passages, such as the Mediterranean Sea and the Arctic routes (Huang et al., 2021). The Mediterranean and Black Sea regions serve as major shipping routes with numerous ports and high ship traffic, and more ships mean higher risk. According to the LLI database, more than 5,000 ship accidents occurred in these regions from 2006 to 2020. It is therefore meaningful to study the characteristics of maritime accidents in these regions in order to prevent the occurrence of accidents and ensure safe transit.

3 Methodology

3.1 Weighted association rule

Association rule is a widely used approach to mine the relationship between different variables in a data set, which was first proposed by Agrawal and Swami (1993). Compared to conventional association rule method, the weighted association rule has significant advantages in exploring the relationship among items in transaction database (Ramkumar, 1998). Firstly, it does not require independent variables or hypothetical models, and it is applicable to small data sets. Secondly, it takes full account of the fact that transactions in the database have different importance (Datta et al., 2020). Different transactions have different importance in the shipping accident dataset. The shipping accident dataset does not have pre-assigned weights for each accident. Using weighted association rules, weights can be computed for each ship accident and some interesting relationships between accident characteristics can be found. According to the obtained maritime accident

association rules, some specific measures can be taken to reduce the probability of accidents.

Definitions: The mathematical expression of the association rule is as follows: Let $I = \{i_1, i_2, i_3, \ldots i_n\}$ be a set of n variables, called items. Let $D = \{t_1, t_2, t_3, \ldots t_m\}$ be a set of m accidents called the transactions, and, each transaction in D has a unique transaction identification (ID) that consists of a subset of the items in I.

In addition, X and Y are defined as the sets of items (itemsets for short) respectively. Supposing A is an association rule, it is defined as an implication of the form $X \Rightarrow Y$, and $X \subset I, Y \subset I, X \cap Y = \emptyset$. X and Y are called antecedent (the Left Hand Side of an association rule or LHS) and consequent (the Right Hand Side of an association rule or RHS) of the rule, respectively (Agrawal, 1994). For example, this is an association rule for ship accident {Flag of Convenience=FALSE, Vessel Type = General cargo $\Rightarrow$ Cause = Machinery damage}. In this rule, {Flag of Convenience = FALSE, Vessel Type = General cargo} and {Cause = Machinery damage} are the LHS and RHS, respectively.

3.2 The flow of rules generation

The flow chart in Figure 2 demonstrates the stages of weighted association rule mining (WARM). The steps of the WARM method in this study are outlined as follows. The code for implementing the WARM method using the R language software is shown in Table 1.

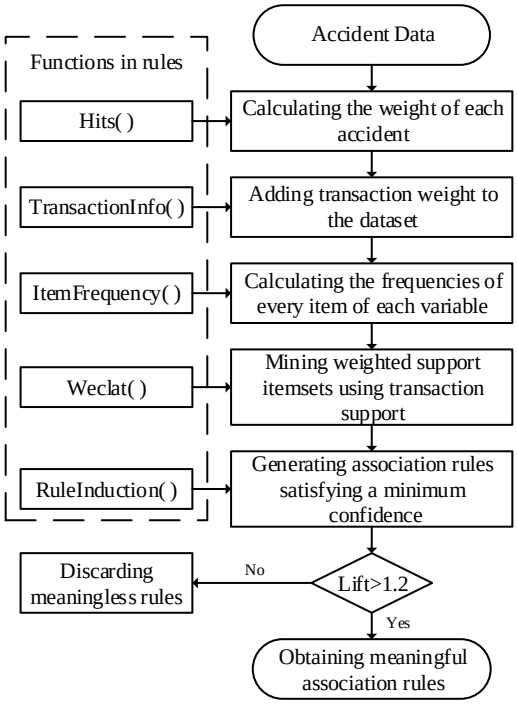


Figure:2 Flow chart of the weighted association rule mining (WARM)

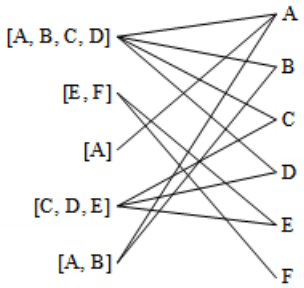
Table 1: Code of R software solving

WARM:
1: Convert accident data sets to transaction sets: Define Data as transaction sets. Data <- transactions(accident database)
2: Calculate the weight of each transaction: Define WI as the weight sets. WI <- hits(Data)
3: Assign the calculated weights to the corresponding transactions: transactionInfo(Data)[["weight"]] <- WI
4: Generate frequent itemsets with support: Define F_itemsets as frequent itemsets. F_itemsets <- weclat(data, support = 0.1)
5: Generate association rules with a confidence level greater than 0.7: Define R_rules as rules sets. R_rules <- ruleInduction(F_itemsets , confidence = 0.7)
6: Inverse order of association rules based on Lift: Define SortLift as sorted set of association rules. sortlift <- arules::sort(R_rules,by = "Lift")

Step 1 - Calculating the weight of each accident: The "HITS" (Hyperlink-Induced Topic Search) algorithm was applied to maritime accident analysis to calculate the weight of each ship accident, which is regarded as a transaction (Sun and Feng, 2008). A ship accident database (transactions) can be described as a bipartite graph, as shown in Figure 3(a). Each transaction in Figure 3 (a) corresponds to a ship accident, and the symbols (items) in the transaction corresponds to the ship accident characteristic.

ID	Transaction
1	[A, B, C, D]
2	[E, F]
3	[A]
4	[C, D, E]
5	[A, B]

(a) Database



(b) Bipartite graph

Figure 3. The bipartite graph representation of a database.

The main idea of the HITS model is that a good transaction which is highly weighted, should contain excellent items. Similarly, a good item should be contained by many good transactions as shown Figure 3 (b). The reinforcing relationship between transactions and items is just like the relationship between hubs and authorities in the HITS model (Kleinberg, 1999). Regarding the transactions as “pure” hubs, and the items as “pure” authorities, the hits function in the R software was applied to implement this iteration. Each iteration is performed with Eq. (1). When the HITS model eventually converges and the iteration ends, the hub weights of all transactions can be obtained (Sun and Feng, 2008). Transactions with higher weights contain items of higher value. A transaction with fewer items may still have a larger weight if the items contained by the transaction rank higher. In contrast, if a transaction contains many ordinary items, it may also have a lower weight.

$$auth(i) = \sum_{t:i \in t} hub(t), hub(t) = \sum_{i:i \in t} auth(i) \quad (1)$$

where $i \in I = \{i_1, i_2, i_3, \dots i_n\}$ and $t \in D = \{t_1, t_2, t_3, \dots t_m\}$.

Step 2 - Adding transaction weight to the dataset: After calculating the weights for each transaction, the transaction weights are stored in the transaction as a column called weight in TransactionInfo. Table 2 shows some sample transactions with computed weights. It should be noted that the FOC (Flag of Convenience), GT (Gross Tonnage), Year, Season and Cause in the header of Table 2 respectively represent whether a ship is registered as a flag of convenience (FALSE or TRUE), the total gross tonnage range of the ship, the occurrence year, the occurrence season, and the cause of the accident. The value of GT column indicates the corresponding gross tonnage range of the ship, and the value of Year column indicates the corresponding occurrence year of the accident. Details of the variables in the analysis are given in Table 3. Furthermore, the value of the weight column in Table 2 represents the absolute weight value of each transaction. It can be seen that after the calculation of the weight using the HITS algorithm, different ship accidents have different levels of importance so that some ship accidents that occur less frequently but are more important will not be overwhelmed in the database.

Table 2: Weighted transactions display

ID	FOC	GT	Year	Season	Cause	Weight
1	FALSE	2	3	Win.	Machinery damage	2.133
2	FALSE	4	3	Win.	Machinery damage	2.117
3	FALSE	2	3	Win.	Collision	2.203
4	FALSE	3	3	Win.	Collision	2.19
5	TRUE	3	3	Win.	Collision	1.353
6	TRUE	4	3	Win.	Collision	1.266
7	TRUE	5	3	Win.	Machinery damage	1.765
8	TRUE	5	3	Fal.	Machinery damage	1.711
9	TRUE	2	3	Fal.	Machinery damage	1.813
10	FALSE	2	3	Fal.	Machinery damage	2.139

Step 3 - Calculating the frequencies of every item of each variable: The itemFrequency( ) function in the R language software is capable of calculating the frequency of all the value of items. This approach allows us to find out which characteristics of maritime accidents are more frequent and draw our intense attention.

Step 4 - Mining weighted support itemsets using transaction support: Before mining the frequent itemsets in Transactions, an introduction to the concept of weighted support is in order. Support is the measure of the significance of association rules. The weighted support of an itemset is the sum of the weights of the transactions that contain the itemset. $Wsupp(X)$ denotes the probability that the set of items (X) occurs in the total set of items. It denotes the probability that X occurs simultaneously in the total number I, and the formula is shown in Eq. (2).

$$Wsupp(X) = \frac{w}{W} \tag{2}$$

where w indicates the weights of transactions containing X, and W indicates the total weights of transactions.

The frequent itemsets can be found by using the Weclat function. This implementation uses optimized tidlist joins and transaction weights to implement WARM. Before mining association rules, the minimum weighted support threshold (min_$Wsupp$) is defined in advance. An itemset is

frequent if its weighted support is equal or greater than the threshold specified by support.

Step 5 - Generating association rules satisfying a minimum confidence: After obtaining frequent itemsets, all association rules satisfying the confidence threshold can be induced using the ruleInduction function. The confidence $Conf(X \Rightarrow Y)$ of the rule "$X \Rightarrow Y$" indicates the probability that Y will occur simultaneously in the transactions in which X occurs, *i.e.,* the ratio of the weighted support of the itemset X together with Y to the weighted support of the itemset X only, and the formula is shown in Eq. (3).

$$Conf(X \Rightarrow Y) = \frac{Wsupp(X \cup Y)}{Wsupp(X)} \tag{3}$$

where $X \cup Y$ represents transactions in which X and Y occur simultaneously.

Step 6 - Obtaining meaningful association rules: During the association rule generation process, there would be too many rules that would satisfy the support and confidence levels. A practical measure to filter or rank the found rules is the *Lift* (Brin, 1997). The *Lift* reflects the correlation between LSH X and RSH Y in the association rule, and it suggests the deviation of the support of the whole rule from the support expected under independence given the supports of both sides of the rule. This is demonstrated by Eq. (4).

$$Lift(X \Rightarrow Y) = \frac{Wsupp(X \cup Y)}{Wsupp(X)Wsupp(Y)} \tag{4}$$

A higher *Lift* value indicates a stronger association between X and Y. $Lift > 1$ indicates a higher positive correlation, $Lift < 1$ indicates a negative correlation, and $Lift = 1$ indicates no correlation (Montella et al., 2011). In this study, *Lift* equal to 1.2 is taken as the threshold to determine whether an association rule is strong (Hahsler et al., 2009).

4 Data

Ship accident data in the Mediterranean and Black Sea regions from 2006 to 2020 in the LLI database is collected to implement the WARM. Each accident record includes the following information: ship information and accident characteristics. Ship information includes vessel type, GT, Age and Flag of Convenience (FOC). The accident information includes Year, Season, Cause, Location and Pollution Indicator. 4,933 vessel accidents were recorded and sorted in this database and Table 3 shows the detailed description and mathematical statistics of the ship accident information. Accident characteristics may differ for different accident levels, so the characteristics of serious and non-serious accidents are studied separately based on the accident level indicators in the accident database. Of these accidents, 3,148 vessel accidents are non-serious and 1,785 vessel accidents are serious. As can be seen from Table 3, the number of ships whose flags were no- FOC accounts for 66.27%. Ships of 0~500 GT are less likely to be involved in accidents. In addition, general cargo vessels are significantly more likely to be involved in accidents. Significantly, the percentage of serious accidents on passenger ships is lower than that of non-serious accidents. Regarding the age of the ship, the statistical accident data indicates that the probability of the ship being involved in an accident increases with the age of the ship. The percentage of accidents for ships between 0~5 years old is 9.77%, while the percentage of accidents for ships over 30 years old is 36.41%.

This paragraph analyzes the accident data of different types of ships based on Table 3. It shows that general cargo vessels have the highest accident rate, while fishing vessels have the lowest. It

also reveals that ship traffic accidents have an upward trend over time. Moreover, in terms of seasonal weather, there is no significant association in the proportion of accidents occurring in different seasons, but to be precise, the proportion of accidents in spring is 22.71% and the largest proportion is in winter with 26.51%. It should be noted that the proportion of serious accidents on ships in winter is significantly higher than that of non-serious accidents. The main cause of accidents on ships is Machinery damage (57.66%), while the number of vessels suffering from Foundering (1.95%) and Hull damage (2.96%), is relatively small. Finally, it is also worth noting that among the 4,933 ship accidents, the percentage of pollution accidents is relatively low.

With the help of the "Wayhe" cloud platform (https://services.wayhe.com/), the spatial distribution profile for ship accidents is presented in Figure 4 where the higher density of accidents can be seen in the Eastern Mediterranean and the Northern Mediterranean. Moreover, it demonstrates that accidents are more likely to occur near the coastline. Figure 4 (a) and Figure 4 (b) show the distribution of serious and non-serious accidents, respectively. They have similar patterns, with most accidents concentrated in the south and east of Greece and the Turkish Strait.

Table 3: The definition of items in each variable and their proportions

Attribute	Variables	Variable classification and descriptions	Proportion		
			Total	Serious accidents	Non-serious accidents
Information of Vessel	Flag of Convenience	False	66.27%	63.40%	67.91%
		True	33.73%	36.60%	32.09%
	GT	1: (0, 500]t	3.1%	3.81%	2.7%
		2: (500, 3000]t	34.39%	34.92%	34.10%
		3: (3000, 10000]t	33.85%	34.36%	33.56%
		4: (10000, 30000]t	18.50%	18.05%	18.75%
		5: (30000, +∝]t	10.16%	8.86%	10.90%
	Age	1: (0, 5] years old	9.77%	9.14%	10.14%
		2: (5, 10] years old	10.14%	10.65%	9.85%
		3: (10, 20] years old	19.14%	19.96%	18.68%
		4: (20, 30] years old	24.80%	25.17%	24.59%
		5:(30,+∝) years old	36.14%	35.09%	36.73%
	Vessel Type	Containership	6.65%	6.95%	6.48%
		Dry bulk carrier	10.10%	9.87%	10.23%
		Fishing	0.83%	1.35%	0.54%
		General cargo	46.14%	49.33%	44.53%
		Liquid bulk ship	10.16%	11.77%	9.25%
		Passenger Ro/Ro	26.12%	20.74%	29.17%
Information of Accident	Year	1: [2006-2010] year	29.79%	30.44%	29.42%
		2: [2011-2015] year	32.53%	23.21%	37.81%
		3: [2016-2020] year	37.68%	46.36%	32.76%
	Season	1: Spring	22.71%	21.69%	23.29%
		2: Summer	25.78%	24.10%	26.72%
		3: Autumn	25.01%	23.99%	25.58%
		4: Winter	26.51%	30.21%	24.40%
	Cause	Collision	20.30%	19.17%	20.94%
		Contact	10.55%	7.23%	12.42%
		Fire	6.59%	9.87%	4.73%
		Foundered	1.95%	5.27%	0.06%
		Hull damage	2.96%	4.71%	1.97%
		Machinery damage	57.66%	53.76%	59.87%
	Location	Black Sea	15.94%	16.65%	15.54%
		East Mediterranean	66.03%	60.59%	69.11%
		West Mediterranean	18.03%	22.76%	15.35%
	Pollution Indicator	True	0.91%	1.96%	0.32%
		False	99.09%	98.04%	99.68%

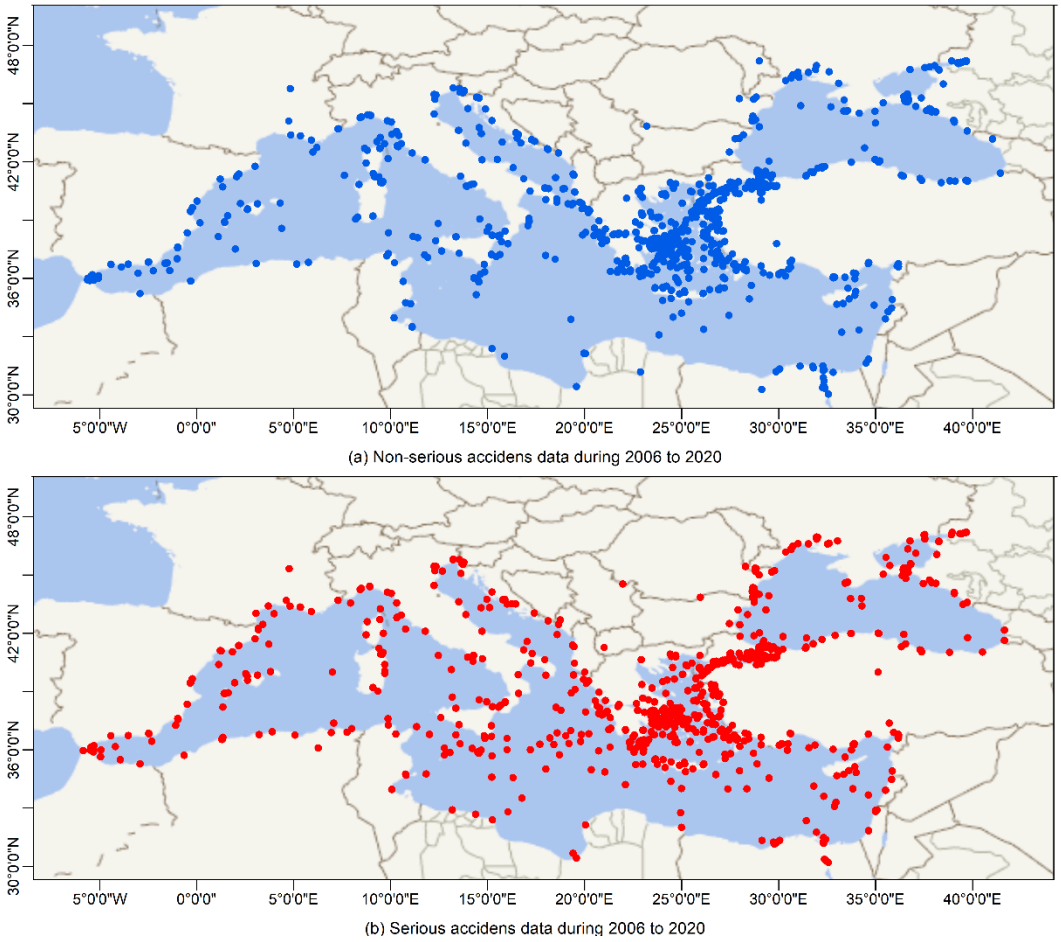


Figure 4: Spatial distribution of accidents in the Mediterranean Sea and Black Sea

To better understand the relationship among the occurrence time (year and season) of ship accidents, the causes of ship accidents and the severity of ship accidents, a mosaic plot was developed, as shown in Figure 5.

In order to understand the accident characteristics from this mosaic diagram, some explanatory notes are needed, which can help to get the information clearly. To simplify the diagram, some abbreviations of the labels have been made. The labels 1, 2 and 3 on the left side of the figure indicate that the accidents occurred in 2006-2010, 2011-2015 and 2016-2020, respectively. The top of the chart indicates the severity category of ship accidents, the left side indicates non-serious accidents and the right side indicates serious accidents. On the right side of the graph, the labels 1, 2, 3 and 4 indicate that the accident occurred in spring, summer, autumn and winter, respectively. The labels Con, Col, Fd, Fire, Hd, and Md on the lower side of the figure indicate the causes of ship accidents: Contact, Collision, Foundered, Fire, Hull damage, and Machinery damage. In addition, the low proportion of Fd and Hd in non-serious accidents leads to overlapping labels for Fd, Fire and Hd in the lower part of Figure 5, but their information can be inferred from the labels in the serious accidents part. Most importantly, the key feature of the mosaic plot is that the area of the nested rectangles is proportional to the frequency of the corresponding accident characteristics.

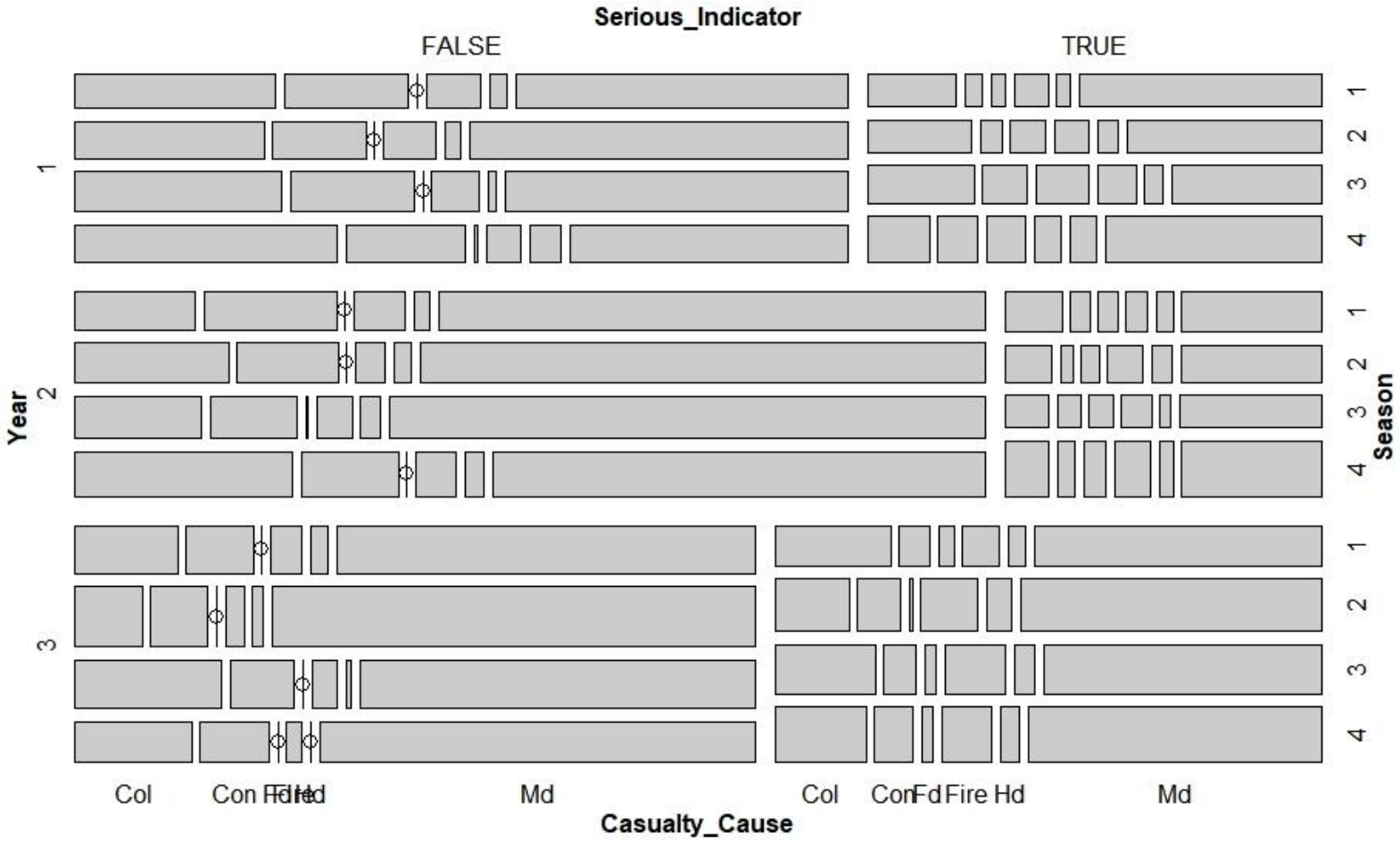


Figure 5: Mosaic image of accidents informatio

Four features can be obtained from Figure 5. The first one is that the frequency of ship accidents shows a gradual increase over time. It can be seen from the height of the rectangle on the left that the number of accidents in 2006-2010 is lower than that in 2011-2015, and the number in 2011-2015 is also less than that in 2016-2020. This means that the situation of ship navigation safety is getting more and more severe, which is worth drawing the attention of all management departments to improve the level of maritime safety supervision. The second feature is that the proportion of serious accidents on ships increases significantly in 2016-2020, presumably due to the growing transport demand and the increasing number of large and medium-sized ships, as well as the higher probability of serious accidents on large ships. Third, the probability of machinery damage accidents is much higher than other accidents, and machinery damage accidents are more likely to occur in winter. Finally, the right side of the figure shows that accidents occurring in spring and summer are less than those in autumn and winter, which may be related to seasonal environmental factors in the Mediterranean and Black Sea.

5 Rule generation in marine traffic accidents

5.1 Single-factor characteristics in marine traffic accidents

According to the WARM working flow in Section 3.2, association rules were mined from a database of ship accidents including 3,148 non-serious accidents and 1,785 serious accidents. Before generating the association rule, support and confidence thresholds need to be determined. In order to obtain a more frequent set of items, referring to previous studies (Ozaydin et al., 2022; Weng and Li, 2017; Yang et al., 2018), the support threshold is set to be 0.1 and the confidence threshold is set to be 0.7. In addition, the lift threshold is set as 1.2 to obtain the strong association rule according to a previous study conducted by Hahsler et al. (2006). If the generated association rule has a *Lift* greater than 1.2, then the association rule is considered to be a strong association rule and is retained. In the process of calculating the frequent item set, the frequent item set of 1,079 for non-serious accidents and the frequent item set of 844 for serious accidents were obtained. After weighted association rule mining, 82 association rules for serious accidents and 115 association rules for non-serious accidents, which meet the given threshold of support and confidence, are found. Table 4 and Table 5 show the association rules for some of the serious incidents sorted by *Lift*, and the association rules for non-serious incidents sorted by *Lift*, respectively.

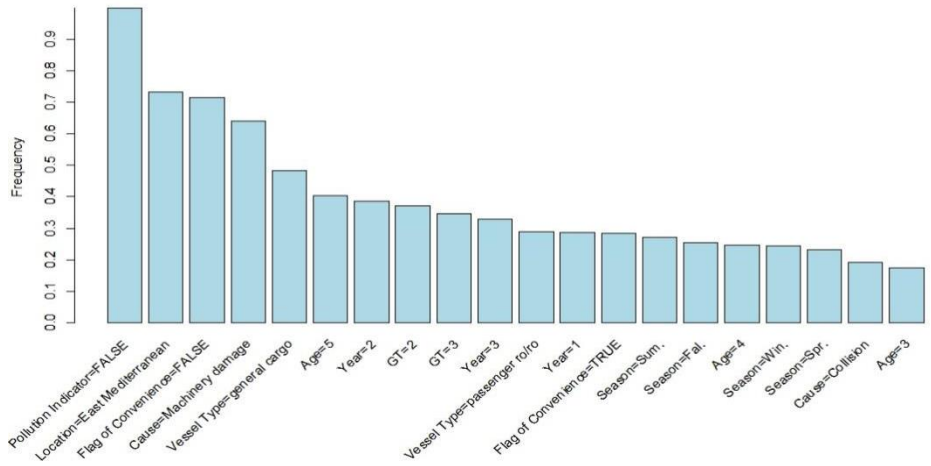
Table 4: Association rules of serious accidents

	LHS	RHS	Weighted support of LHS	Confidence	Lift
1	Flag of Convenience=FALSE, GT=2, Location=East Mediterranean, Vessel Type=General cargo	Age=5	0.11	0.73	1.93
2	Vessel Type=General cargo, Pollution Indicator=FALSE, Age=5	GT=2	0.18	0.71	1.89
3	Flag of convenience=FALSE, GT=2, Vessel Type=general cargo	Age=5	0.16	0.71	1.89
4	GT=2, Location=East Mediterranean, Pollution Indicator=FALSE, Age=5	Vessel Type= General cargo	0.12	0.84	1.57
5	GT=2, Age=5	Vessel Type= General cargo	0.18	0.83	1.56
6	GT=2, Season=Win.	Vessel Type= General cargo	0.10	0.81	1.51
7	GT=2, Cause=Machinery damage	Vessel Type= General cargo	0.17	0.77	1.43
8	GT=2, Location=East Mediterranean	Vessel Type= General cargo	0.19	0.76	1.42
9	Flag of Convenience=FALSE, GT=2, Cause=Machinery damage	Vessel Type= General cargo	0.12	0.76	1.42
10	Flag of Convenience=FALSE, GT=2, Cause=Machinery damage, Pollution Indicator=FALSE	Vessel Type= General cargo	0.12	0.76	1.42

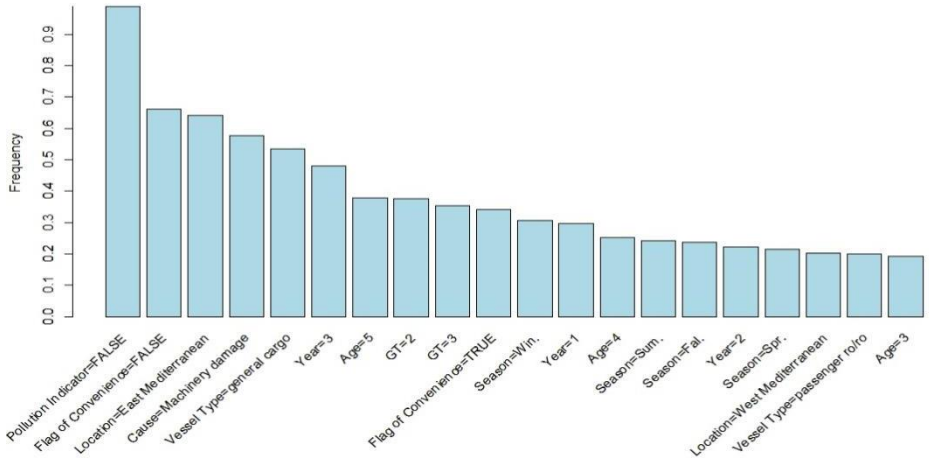
According to step 2 in Section 3.2, the frequency (support) of frequent items was obtained and Figure 6 was created to graphically show which characteristics occur frequently in maritime accidents. Figure 6 describes the frequency of the top 20 items for all accidents in the study, split into Figure 6(a) for Non-Serious and Figure 6(b) for Serious. The frequency (support) of the top 20 items is calculated by Eq. (2). In addition, the frequency of the top 20 items is determined by their frequency of occurrence in accidents and the weight of the corresponding accidents. Before explaining this graph, some clarifications are needed: GT denotes Gross Tonnage and Age denotes the age of the ship, and the abbreviations of these necessary variables and their indicated ranges can also be found in Table 3. From this graph, we know that in the case of non-serious accidents, the two most frequent items are {Pollution Indicator = false}, which has a frequency of 0.99, {Location = Eastern Mediterranean}, which has a frequency of 0.73, and the two least frequent items are {Cause = Collision}, with a frequency of 0.19, and {Age = 3}, with a frequency of 0.18 in Mediterranean and Black Sea. However, in the case of serious accidents, the two most frequent items are {Pollution Indicator = FALSE} whose value is 0.99 and {Flags of Convenience=FALSE} whose value is 0.66, while the two least frequent items are {Vessel Type=Passenger/Roll-on/Roll-off} whose value is 0.20 and {Age=3} whose value is 0.19. Moreover, it can also be seen that the older the ship, the more likely it is to be involved in an accident, whether serious or not.

Table 5: Association rules of non-serious accidents

	LHS	RHS	Weighted support of LHS	Confidence	Lift
1	Flag of Convenience=FALSE, Cause=Machinery damage, Vessel Type=General cargo, Age=5	GT=2	0.12	0.70	1.90
2	GT=2, Cause=Machinery damage, Age=5	Vessel Type= General cargo	0.13	0.88	1.82
3	Flag of Convenience=FALSE, GT=2, Location=East Mediterranean	Vessel Type= General cargo	0.17	0.72	1.50
4	Flag of Convenience=FALSE, Year=2, Pollution Indicator=FALSE, Age=5	Vessel Type= General cargo	0.10	0.70	1.46
5	Flag of Convenience=FALSE, Year=2, Location=East Mediterranean, Vessel Type=General cargo, Pollution Indicator=FALSE	Cause= Machinery damage	0.10	0.84	1.32
6	Year=2, Location=East Mediterranean, Vessel Type=General cargo	Cause= Machinery damage	0.13	0.83	1.30
7	Year=2, Vessel Type=General cargo, Age=5	Flag of Convenience =FALSE	0.10	0.92	1.29
8	GT=2, Location=East Mediterranean, Age=5	Flag of Convenience =FALSE	0.15	0.91	1.27
9	GT=3, Location=East Mediterranean, Age=5	Flag of Convenience =FALSE	0.10	0.91	1.27
10	Year=3, Vessel Type=General cargo	Cause=Machinery damage	0.11	0.81	1.27



(a): non-serious accidents



(b): serious accidents

Figure 6: Item frequency of accidents

5.2 Multi-factor association characteristics

In this study, the relationship between the support, confidence and lift of all generated association rules is visualized as a scatter plot using the "arulesViz" package in R language. Figure 7 and Figure 8 represent scatter plots of association rules for non-serious and serious accidents, respectively. In these two graphs, the left side indicates the confidence level, and the bottom side indicates the support level. Each dot in the graph represents an association rule, and the size of the dot indicates the strength of the rule. The larger the size of the points, the better the quality of the association rule, provided that the confidence level is satisfied. From Figure 7, it can be seen that the confidence level of the strong association rules among the 649 association rules for non-serious accidents is mainly concentrated in the range of 0.75~0.95, and the support level is mainly concentrated in the range of 0.1~0.4. For the association rules of serious accidents, it can be seen from Figure 8 that the confidence level of strong association rules among 317 association rules is mainly concentrated in 0.7~0.92 and the support level is mainly concentrated in 0.1~0.3. This means that setting the support threshold at 0.1 and the confidence threshold at 0.7 will not miss most of the association rules, which also indicates that the initial settings of support and confidence thresholds in the experiment are reasonable. If the confidence threshold is set too high, a large number of strong

association rules concentrated around 0.7~0.8 will be lost, and the association rules obtained with too low a confidence threshold are unreliable.

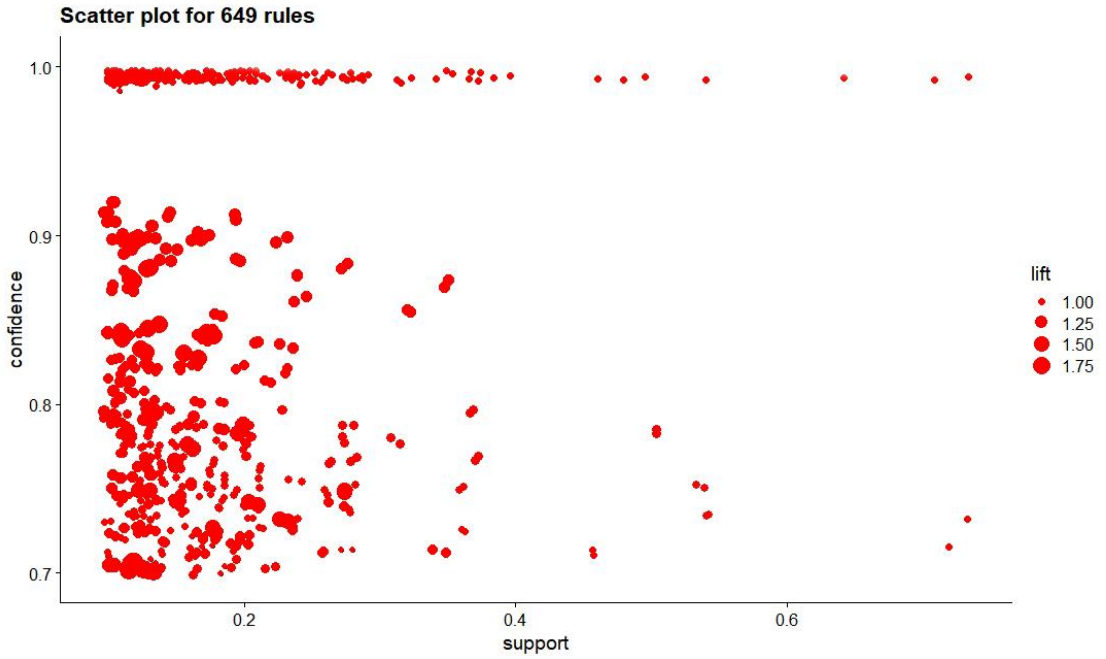


Figure 7: Scatter plot for non-serious accidents

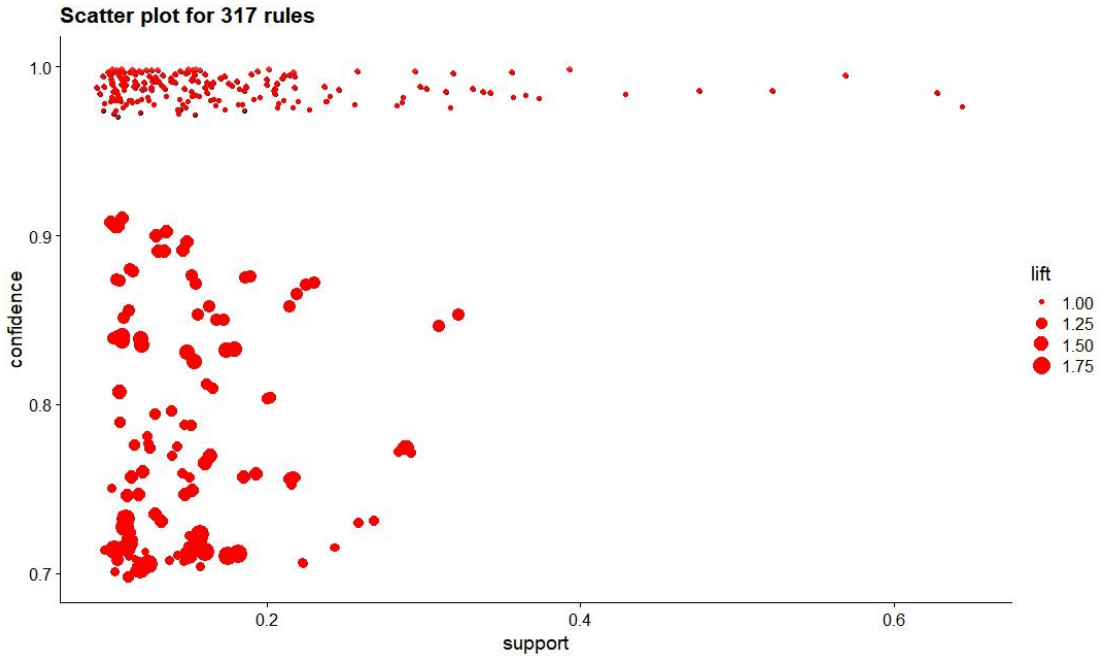


Figure 8: Scatter plot for serious accidents

When visualizing association rules, balloon diagrams like Figure 9 and Figure 10 were generated using the "arulesViz" package in R. The k-means clustering technique was used to cluster the antecedents of the association rules into 20 groups. Each column in the figure represents the most frequent items in the LHS of the grouped association rules, respectively, and each row represents the RSH of all association rules. Figures 9 and 10 depict the relationship between the grouped items in LHS and all association results for non-serious and serious incidents, respectively. In these two figures, the colour shades of the balloons indicate the lift values of the association rule

groups. The light-coloured balloons indicate that the clustered association rule groups have smaller lifts (i.e., weak association rules) and the dark-coloured balloons indicate that the clustered association rule groups have larger lifts (i.e., strong association rules). The size of the balloon represents the support of the association rule set, and a larger balloon indicates higher support of the association rule set. In addition, the lift values represented by the balloon colours in the figure are reordered so that the lift values of the associated rule groups are decreasing from top left to bottom right. The association rule groups with high lift values are displayed in the upper left corner. Such a balloon diagram can describe association rules more intuitively and simplify association rules, which can remove redundant terms from a large number of association rules to find closely related items. For illustration, the balloon in the upper left corner of Figure 10 indicates that there are 12 rules whose LSH contains these items {Age=5, Vessel Type=general cargo}, while the RSH is {GT=2}, and four other items are also involved in the RHS for non-serious accidents.

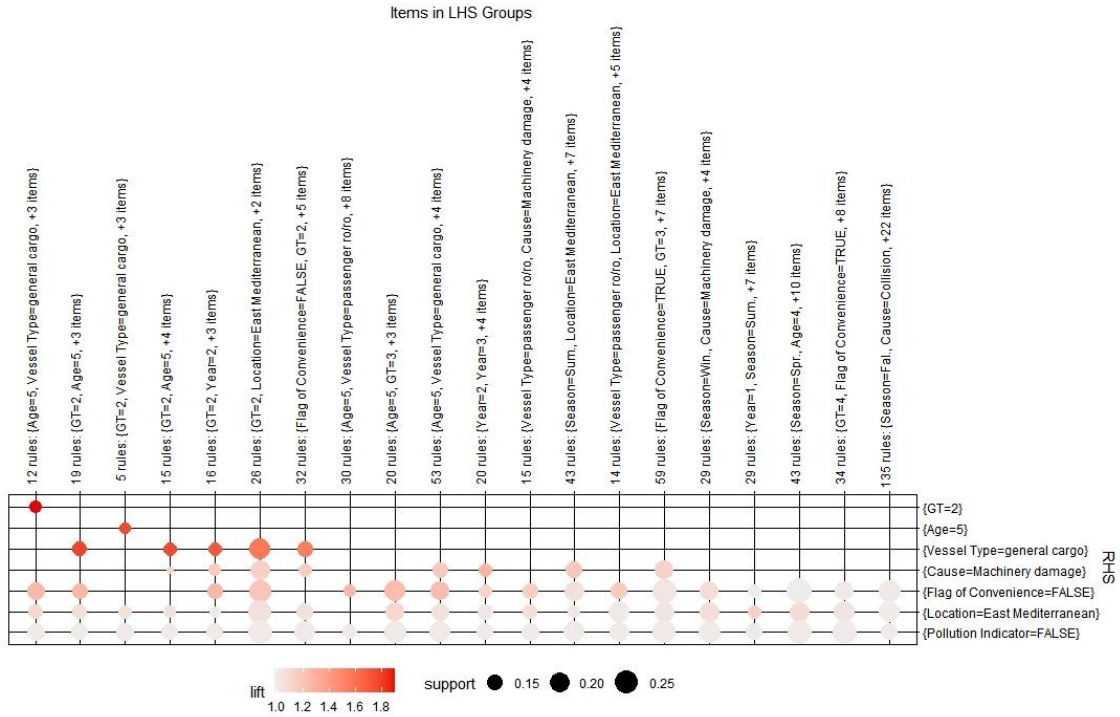


Figure 9: Association rule clustering matrix of non-serious accidents

Furthermore, with the help of "Gehpi" software, two network relationship diagrams in Figure 11 were created to show the features and details of the association rules based on the strong association rules that had been generated. Figure 11(a) represents the relationships between LHS and RHS for the association rules of non-serious accidents; Figure 11(b) represents serious accidents. Each node of the network graph represents the RHS of the rule, and the number of edges connected to the node represents the strength of the node, which is represented by the size of the node in the graph. From Figure 11(a), it can be determined that the three largest nodes are {Flag of Convenience = FALSE}, {Vessel Type = General cargo} and {Cause = Machinery damage} for non-serious accidents. As illustrated in Figure 11(b), unlike non-serious accidents, {Location = East Mediterranean} ranks among the top three with an intensity higher than {Cause = Machinery damage}.

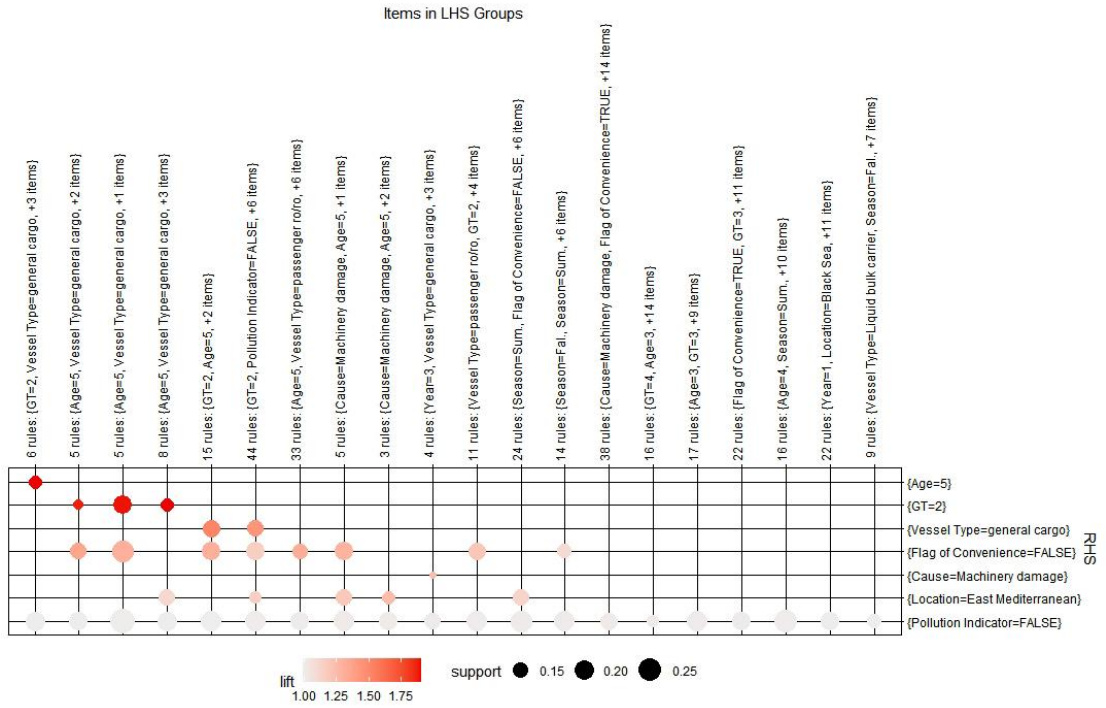
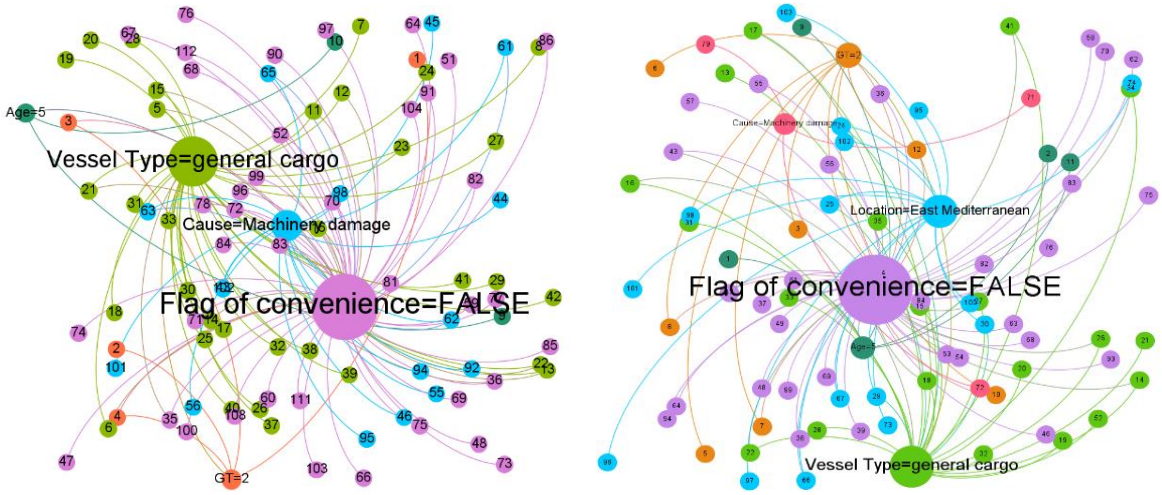


Figure 10: Association rule clustering matrix of serious accidents



(a) non-serious accidents *(b) serious accidents*

Figure 11: Network relationship diagram of association rules

Based on the generated association rules, the relationship between the cause of the accident with the ship information and other variables of the ship accident is further analysed by taking the cause of the accident as the RHS. Under the constraints of a support threshold of 0.1, a confidence threshold of 0.7, and a lift threshold of 1.2, the 8 association rules for non-serious incidents shown in Table 6 were generated. It can be seen from Table 6 that most of the non-serious accidents in the Eastern Mediterranean waters are caused by machinery damage, and most of the ships involved in these accidents are general cargo ships which, in most instances, don't result in pollution. However, for serious accidents, only three association rules with the cause of the accident as the RHS were generated as shown in Table 7. In addition, this study analysed the differences between machinery

damage accidents and other accidents in terms of spatial location, using accident causes as the classification criteria. As shown in Figure 12, the spatial distribution of machinery damage accidents and other accidents is approximately in the same order, but the number of machinery damage accidents is much larger than that of other accidents. It can also be seen that accidents are mostly concentrated in coastal waters.

Table 6: Association rule with accident cause as RHS of non-serious accidents

	LHS	RHS	Weighted support of LHS	Confidence	Lift
1	Flag of Convenience=FALSE, Year=2, Location=East Mediterranean, Vessel Type=General cargo	Cause= Machinery damage	0.10	0.84	1.32
2	Year=2, Location=East Mediterranean, Vessel Type=General cargo	Cause= Machinery damage	0.13	0.83	1.30
3	Year=3, Vessel Type=General cargo, Pollution Indicator=FALSE	Cause= Machinery damage	0.11	0.81	1.27
4	GT=2, Year=2, Vessel Type=General cargo	Cause= Machinery damage	0.10	0.81	1.26
5	GT=2, Year=2, Vessel Type=General cargo, Pollution Indicator=FALSE	Cause= Machinery damage	0.10	0.81	1.26
6	Year=2, Location=East Mediterranean, Age=5	Cause= Machinery damage	0.10	0.79	1.23
7	Year=2, Vessel Type=General cargo, Pollution Indicator=FALSE	Cause= Machinery damage	0.16	0.79	1.23
8	Year=3, Location=East Mediterranean	Cause= Machinery damage	0.18	0.78	1.22

Table 7: Association rule with accident cause as RHS of serious accidents

	LHS	RHS	Weighted support of LHS	Confidence	Lift
1	Year=3, Location=East Mediterranean, Vessel Type=General cargo, Pollution Indicator=FALSE	Cause= Machinery damage	0.11	0.73	1.26
2	Year=3, Location=East Mediterranean, Vessel Type=General cargo	Cause= Machinery damage	0.11	0.72	1.25
3	Flag of Convenience=FALSE, Year=3, Vessel Type=General cargo, Pollution Indicator=FALSE	Cause= Machinery damage	0.12	0.7	1.22

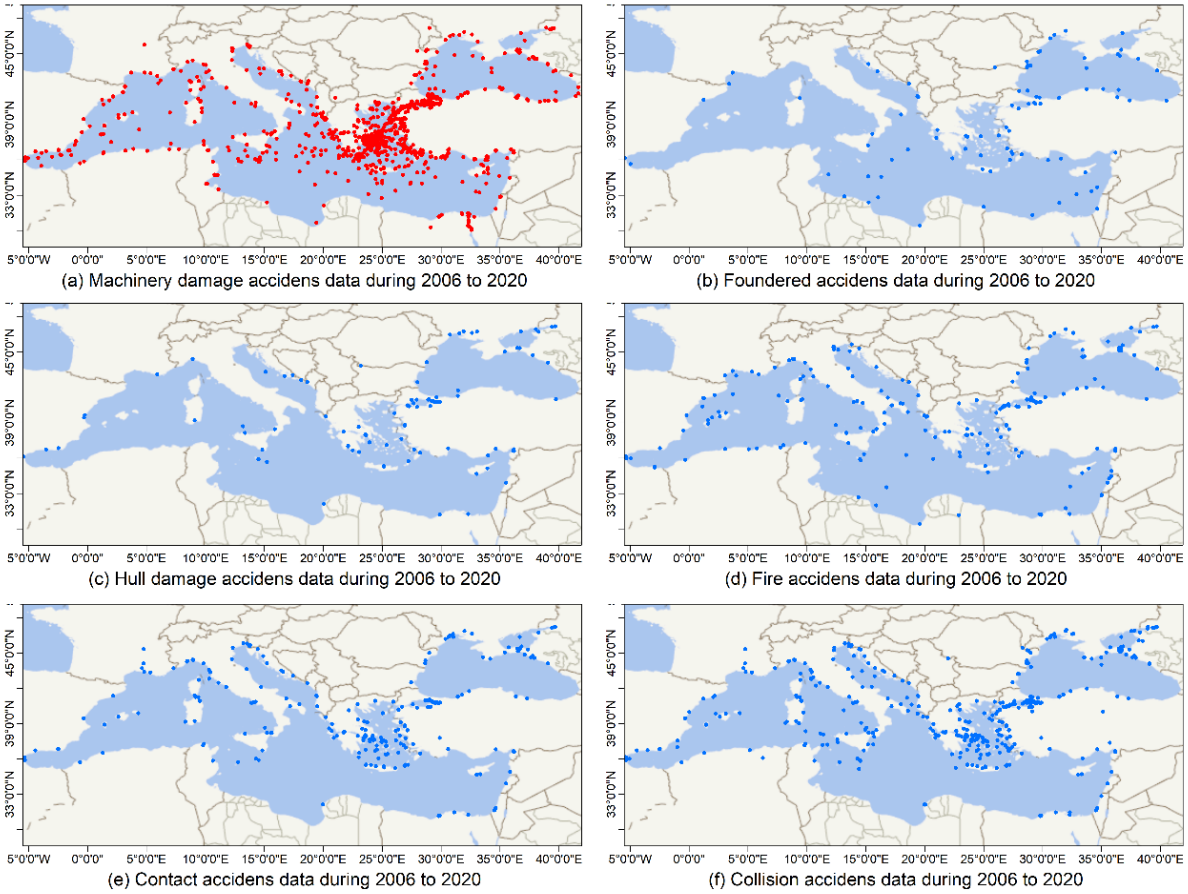


Figure 12: Spatial distribution of accidents from 2006 to 2020 based on accident causes

In summary, this section has generated association rules by running the WARM algorithm in R language based on the ship accident data that have been pre-processed. Moreover, the detailed features of the generated association rules are presented in graphs and tables through certain visualization and statistical techniques to facilitate better observation of the association between the causes of ship accidents. The results obtained in this section will be discussed and summarized in Section 6.

6 Discussion

6.1 Single item characteristics for accidents

6.1.1 Similarities between serious accidents and non-serious accidents

In the Mediterranean and Black Sea regions, Figure 6 shows that items {Flag of Convenience = FALSE} and {Pollution Indicator = FALSE} are quite frequent items for non-serious and serious accidents. This indicates that pollution accidents are less frequent in these regions and non-Flag of Convenience (non-FOC) vessels are more likely to be involved in an accident. In addition, general cargo ships are more likely to be involved in accidents. This study also found that both serious and non-serious accidents are closely associated with ships with a GT of 500~1000 tons. It can also be seen from Figure 6 that the proportion of Machinery damage accidents is relatively high. From Figure 6(a) and Figure 6(b), it can be seen that the frequency of the item {Season=Spr.} for both serious and non-serious accidents is the lowest, which indicates that the probability of accidents is the lowest in spring. As can be observed from these two diagrams, vessels over 20 years old, were more likely to be involved in accidents. This means that the older the ship, the higher the accident rate.

In our common sense, FOC ships should be more likely to be involved in an accident because the requirement of FOC condition is less than that of non-FOC condition. However, according to Figure 6, a higher percentage of non-FOC ship accidents occurred in both non-serious and serious accidents. This phenomenon could be explained in two aspects. On one hand, possibly due to the loose Port State Control of non-FOC vessels that have a better reputation than FOC vessels, the probability of accidents on such vessels is greater than that on FOC vessels. On the other hand, it is evident from Figure 1 that in the Mediterranean and Black Sea regions, ship accidents are mainly concentrated in ports and coastal waters. These regions have a windy coastline and many ports, and the coastal countries generally have the restriction of coastal transportation rights, i.e., the fleet with the national flag enjoys transportation rights between its coastal ports, and the fleet with a foreign flag is restricted to operate the transportation services in the coastal ports. Therefore, more non-FOC vessels are found in the accidents of these regions.

Non-serious and serious accidents share a similar spatial distribution. If classified based on the proximity to the coast, accidents are mainly distributed along the coastline and in ports. According to the sea area, the occurrence rate in the East Mediterranean is higher than that in the West Mediterranean and Black Sea, and the accident rate is higher in the northern Mediterranean than in the southern part. The northern and eastern coastlines are winding, with numerous bays, islands, and ports, while the southern coastline is straighter and has fewer ports than the north and east. This geography and shipping conditions result in far more ships sailing in the north and northeast than in the south. Winds play a non-negligible role in the variability of the uppermost marine layer, especially in the Eastern Mediterranean (Martinez et al., 2022). Similarly, the direction of Mediterranean waves is predicted to have a slight eastward trend between 2006 and 2010 (Leone, 2017). Additionally, the sea between Greece and Turkey has the largest number of islands in the entire Mediterranean and Black Sea regions. This geographic and navigational environment has led to the Eastern Mediterranean region being the area with the highest number of maritime accidents.

6.1.2 Differences between serious and non-serious accidents

From the above generation process of association rules, 3,148 transactions of non-serious accidents generated 107 frequent items and 115 rules, and 1,785 transactions of serious accidents generated 386 frequent items and 82 rules. The number of rules for serious accidents was comparable to that of non-serious accidents, despite the smaller dataset size. This suggests that the causal factors of serious accidents may be more complex and interrelated. For example, as can be seen in Figure 9, 12 rules contain the item {Vessel type=General cargo} in the LHS, while the RHS is {GT=2} in the non-serious accidents. However, Figure 10 shows that there are 18 different combinations of items on the RHS of the rules with {Vessel type=General cargo} in the LHS for serious accidents. As seen in Figures 5(a) and 5(b), the frequency of the item {Season=Win.} for serious accidents is higher than non-serious accidents for the seasons in which ship accidents occur, which indicates that serious accidents are more likely to occur in winter. Poor sailing conditions in winter make serious accidents more likely to occur.

In addition, comparing Figure 11(a) and Figure 11(b), for non-serious accidents, the three nodes with higher intensity are {Flag of Convenience = FALSE}, {Vessel Type = General cargo} and {Cause = Machinery damage}, and for serious accidents, the three nodes with higher intensity are {Flag of Convenience = FALSE}, {Vessel Type = General cargo} and {Location=East Mediterranean}. The nodes with higher intensity represent ship traffic accidents closely related to the variables affecting ship traffic accidents represented by these nodes. For serious accidents, attention should be paid to inspect and maintain the ship to avoid the occurrence of machinery damage accidents.

6.2 Multi-item association rules for accidents

One of the benefits of the association rule approach is that it can capture the interactions among multiple factors that influence ship accidents and assess their joint impact on ship accidents. For example, in section 6.1 single item characteristics analysis, it is known from Figures 5(a) and 5(b) that the term {Cause=Machinery damage} is more frequent in the ship accident database. Nevertheless, it is not known which characteristics are strongly associated with this type of ship accidents. In order to understand which characteristics are strongly associated with {Cause=Machinery damage}, the {Cause=Machinery damage} is taken as the RHS of association rules. After putting together several characteristics of the accident as the LHS, it can be seen that the items {Vessel Type=General cargo}, {Location=East Mediterranean} are closely related to the item {Cause=Machinery damage} in both non-serious and serious accidents from Table 6 and Table 7, which means that general cargo ships in the Eastern Mediterranean are vulnerable to machinery damage accidents. In Table 6 it can also be found that the item {GT=2} has an effect on accidents with {Cause=Machinery damage} as the RHS in non-serious accidents, but only when the item {Year=2} appears. The effect of variable {Year} on accidents has no greater effect on the prevention of future ship safety accidents, so it can be concluded that item {GT=2} is not closely related to item {Cause=Machinery damage}.

According to Table 4 and Table 5, an interesting phenomenon is revealed, which shows that the item {Flag of Convenience = FALSE} can demonstrate an influence on the occurrence of both serious and non-serious accidents in some specific cases. Nevertheless, comparing the severity level of each accident in Table 6 and Table 7, it can be seen that the item {GT=2} is strongly associated with the non-serious accident but is not associated with the serious accident under the situation

characterized by {Vessel Type = General cargo, Year=2}. This indicates that the gross tonnage of a ship has some influence on the severity level of ship accidents.

Another interesting rule can be found in Figure 9 and Figure 10. For non-serious and serious accidents, the item {GT=2}, the item {Vessel Type = General cargo} and the item {Age=5} always appear together. Such a phenomenon indicates that in the Mediterranean and Black Sea regions the probability of a ship accident is higher when a general cargo ship is 30 years old or more and has a gross tonnage between 500 and 3,000 tons. Therefore, the maritime administration should strengthen the safety supervision of general cargo ships that are more than 30 years old and have a gross tonnage of 500~3,000 tons.

6.3 Uncertainty and Sensitivity analysis

Uncertainty analysis of data, model and results is a technique for assessing the data quality, the model assumptions, and the validity of results (Tirthajyoti, 2021). Goerlandt et al. (2017b) stated that uncertainty should be identified and quantified, and sensitivity analysis should be performed in the risk analysis of maritime accidents, in order to evaluate the impact of uncertainty on the risk analysis outcomes. It can enable us to comprehend the constraints of data and models, the variation and sensitivity of results and augment the confidence of results. Accordingly, this can furnish more robust and transparent evidence for decision making, and also elicit more pertinent recommendations for future research (Goerlandt et al., 2017a).

6.3.1 Basis of uncertainty and sensitivity analysis

Flage and Aven (2009) investigated the appropriate treatment of uncertainty in quantitative risk analysis and suggested some principles and guidelines about classifications of uncertainty and sensitivity. The categorization scheme of uncertainty and sensitivity devised by Flage and Aven was employed in this study, and the detailed description is as follows.

A. Uncertainty rating and criteria

i. Low uncertainty

All following conditions are met:

- The phenomena involved are well understood, models used are well validated and can give predictions with the desired accuracy.
- Sufficient reliable data are available.
- The assumptions made are very rational.
- There is wide consensus among experts.

ii. High uncertainty

Conditions that are contrary to those defining low uncertainty.

iii. Moderate uncertainty

Conditions that lie between Low and High uncertainty.

B. Sensitivity rating and criteria

i. Low sensitivity

Base case values require to be changed unrealistically large for altered conclusion.

ii. High sensitivity

Base case values require to be changed relatively small for altered conclusion.

iii. Moderate sensitivity

Base case values require to be changed relatively large for altered conclusion.

6.3.2 Evaluation of uncertainty and sensitivity for maritime accident analyses

Uncertainty and sensitivity arise from multiple aspects, such as the quality, quantity and representativeness of data sources, and the assumptions, parameters, and structure of the model (Bostelmann et al. 2022). Due to the uncertainty of the model's input, parameters and structure, this uncertainty propagates through the model and affects the output. Therefore, to improve the reliability and transparency of the results a comprehensive uncertainty and sensitivity evaluation of data, model and results is necessary. Similarly, the outcomes of the preliminary subjective evaluation of uncertainty and sensitivity of the data, model and results are shown in Table 8.

Table 8: Evaluation of Uncertainty and Sensitivity rating

	Uncertainty rating			**Sensitivity rating**		
	Low	**Moderate**	**High**	**Low**	**Moderate**	**High**
Accident data	√			√		
HITS algorithm	√				√	
Parameters of WARM		√				√

Lloyd's List Intelligence (LLI) provides maritime data to professionals around the globe to make confident decisions that drive the safe, efficient, and lawful movement of trade by sea. Their data collection, cleaning, enhancement, and output processes are managed through COACT (Consistency, Origin, Accuracy, Completeness, Timeliness). It ensures that everything they do with the data meets the same high standards. During the data preparation stage, Mediterranean and Black Sea data was screened from the LLI database, coded and organized according to the needs of the study. Therefore, it can be considered that the uncertainty of the data related to the maritime accident characteristics in the LLI database is very low. There is a considerable connection between environmental factors, ship management factors, human factors, and maritime accidents. However, it is difficult to obtain this information from existing accident databases, which has a significant impact on the analysis of maritime accidents.

The application of the HITS algorithm in WARM is mainly based on two assumptions: a high-quality item will be pointed to by many high-quality transactions, and a high-quality transaction will point to many high-quality items. These two assumptions are reasonable for maritime accidents. Factors that are more likely to affect maritime transportation safety will appear in many maritime accidents, and a maritime accident will always contain many unsafe accident influencing factors. However, the HITS algorithm calculates the weight of each accident based on the entire accident database. If several data points are deleted from the database, it will cause a slight change in the weight value of each accident, which may affect the results of the generated association rules. To verify this, 147 accidents were removed from the end of the list of 3,147 non-serious accidents, and the weight of each accident was recalculated using the HITS algorithm. The weight values of the top ten accidents are shown in Table 9. Comparing Table 2 and Table 9, it can be found that the weights of each incident have changed slightly after deleting some data. 3,000 accident data were used to generate 125 association rules using the WARM method and compared with 115 association rules generated from 3,147 accidents. After comparison, it was found that although the number of association rules increased by ten, the main association rules did not change. Therefore, this performance can explain that the sensitivity of the HITS algorithm is relatively low and has little effect on the quality of association rules.

Table 9: Weighted transactions display after delete 147 accidents

ID	FOC	GT	Year	Season	Cause	Weight
1	FALSE	2	3	Win.	Machinery damage	2.149
2	FALSE	4	3	Win.	Machinery damage	2.122
3	FALSE	2	3	Win.	Collision	2.210
4	FALSE	3	3	Win.	Collision	2.193
5	TRUE	3	3	Win.	Collision	1.346
6	TRUE	4	3	Win.	Collision	1.261
7	TRUE	5	3	Win.	Machinery damage	1.768
8	TRUE	5	3	Fal.	Machinery damage	1.718
9	TRUE	2	3	Fal.	Machinery damage	1.826
10	FALSE	2	3	Fal.	Machinery damage	2.158

In association rule mining, support and confidence are two important indicators. The settings of these two parameters directly affect the number of association rules generated. As shown in Table 10, different settings of support and confidence thresholds will produce different numbers of association rules. It can be considered that the results of association rules generated by the WARM method are highly sensitive to the influence of support and confidence parameters. However, by comparing the specific content of the generated association rules, it can be found that when the support threshold is less than 0.15 and the confidence threshold is less than 0.7, although there is a large difference in the number of rules generated, there is not much change in the content of association rules. Generally speaking, in practical applications, the settings of support and confidence thresholds need to be adjusted according to specific situations. If the thresholds are set too high, the rules mined would be few and homogenous. If they are set too low, some meaningless rules would also be identified. The confidence and support threshold values in this study mainly refer to previous research (e. g. Hahsler et al., 2009; Weng and Li, 2017; Yang et al., 2018).

Table 10: Rules number with different support and confidence thresholds

Support	Confidence	Lift	Rules number
0.1	0.5	1.2	265
0.1	0.6	1.2	191
0.1	0.7	1.2	115
0.1	0.8	1.2	79
0.15	0.5	1.2	109
0.15	0.6	1.2	79
0.15	0.7	1.2	43
0.15	0.8	1.2	25
0.2	0.5	1.2	39
0.2	0.6	1.2	25
0.2	0.7	1.2	17
0.2	0.8	1.2	11

6.4 Performance with association rule mining methods

In order to reflect the different importance of ship accidents, it is crucial to apply weights to the accident transactions, for which different accidents have different importance. An accident containing fewer items may still be a more important event if all the items it contains are ranked high in importance. Conversely, an accident may have low importance even if it contains many items, because it contains items that are ranked lower in importance. The weighting of accidents allows accidents that occur less frequently but contain more important items to be taken into account, thus increasing the number of association rules to be mined. The Apriori algorithm and Eclat algorithm are two other well-known algorithms for association rule mining methods, which are used to mine association rules in unweighted transactions. This study applied the Apriori algorithm and the Eclat algorithm to mine association rules and compare the number of association rules generated with WARM and the results are presented in Table 11. The results show that, with other conditions being equal, the number of association rules generated by WARM is more comparable to Apriori algorithm and Eclat algorithm for both serious and non-serious accidents. Furthermore, it can be found from Table 11 that the improvement in the number of association rules mined using the WARM method for severe accidents is more significant. This may be due to the higher complexity of severe accidents, which is consistent with the findings of Weng et al. (2017).

Table 11: Statistics of the generated association rules

Method	Support level	Confidence level	Lift level	Number of association rules	
				Serious accident	Non-serious accident
Apriori	0.1	0.7	1.2	58	101
Eclat	0.1	0.7	1.2	58	101
WARM	0.1	0.7	1.2	82	115

In an attempt to compare the impact of different association rule mining methods on the number of specific association rules, {Flag of Convenience=FALSE}, {Vessel Type= General cargo} and {Cause=Machinery damage} were taken as the RHS of association rules according to Figure 11, and the corresponding proportion and number were counted, with the results shown in Table 12. Compared with association rules generated by the unweighted association rules mining methods, the proportion of association rules generated by the WARM method with {Flag of Convenience=FALSE} as the RHS has decreased in serious accidents. This situation indicates that after weighted calculation, the weight of serious accidents containing {Flag of Convenience=FALSE} has decreased, which means that {Flag of Convenience=FALSE} is a relatively common item in serious accidents. Conversely, the proportion of association rules generated by the WARM method with {Vessel Type= General cargo} as the RHS has increased in non-serious accidents, indicating that {Vessel Type= General cargo} is a relatively important item in non-serious accidents. Similarly, {Cause=Machinery damage} is a relatively important item in serious accidents and a relatively common item in non-serious accidents. The performance shows that different accidents should have different weights in the association rule mining process. The WARM method differentiates the importance of different accidents while increasing the number of association rules.

Table 12: Proportion and number of specific RHS

Method	Flag of Convenience=FALSE		Vessel Type=General cargo		Cause=Machinery damage	
	Proportion (Number) of RHS		Proportion (Number) of RHS		Proportion (Number) of RHS	
	Serious accident	Non-serious accident	Serious accident	Non-serious accident	Serious accident	Non-serious accident
Apriori	0.48(28)	0.48(48)	0.31 (18)	0.26 (26)	0.00 (0)	0.21 (21)
Eclat	0.48(28)	0.48(48)	0.31 (18)	0.26 (26)	0.00 (0)	0.21 (21)
WARM	0.44(36)	0.48(56)	0.31 (26)	0.31 (53)	0.04 (3)	0.14 (16)

7 Conclusions

To strengthen the efficiency and pertinence of accident prevention and control measures, the weighted association rule approach is applied to study the causal characteristics of ship traffic accidents. Based on ship traffic accidents in the Mediterranean and Black Sea regions from 2006 to 2020, 115 association rules for non-serious accidents and 82 association rules for serious accidents are generated using the WARM algorithm. Uncertainty and sensitivity analysis are carried out. The performance of WARM was evaluated by comparing it with the unweighted association rule mining method. The association rules were also analyzed and discussed. The findings provide a basis for the formulation of prevention and accident control measures.

WARM is an effective and efficient method to analyze the characteristics of marine traffic accidents. It can generate more accurate and relevant association rules than the unweighted method by considering the weights of accident attributes according to their importance and frequency. The association rules reveal some interesting and useful patterns and insights about marine traffic accidents. The single characteristic results show that the items {Flag of Convenience=FALSE}, {Cause=Machinery damage} and {Vessel Type=General cargo} are the most common in serious and non-serious accidents. Similarly, the older a ship is, the more frequently the ship is involved in accidents. Furthermore, ships are more likely to have serious accidents in winter. The results show that general cargo ships are prone to machinery damage accidents when sailing in the Eastern Mediterranean. The probability of an accident is higher when the gross tonnage of a general cargo ship is within 500 to 3,000 tons and the ship is older than 30 years.

Thus, the objective of analyzing the characteristics of marine traffic accidents using WARM has been met. This study contributes to the literature on marine traffic accident analysis by proposing a novel method and providing new insights. It also benefits the practitioners and policymakers in the maritime industry by offering useful information and suggestions for improving maritime safety. It is worth stating that although the WARM method has improved the disadvantages of non-weighted association rule mining methods to some extent, it also has certain limitations. When mining rules, the model generates association rules based on a series of parameters set by the researcher. Although the generated rules meet the setting of the parameters, this does not mean that there is a real connection between certain accident characteristics. That is, many rules are generated, but some of them are not applicable to reality. Therefore, researchers need to screen the generated rules one by one, ensuring cautious interpretation and application of association rules. Furthermore, when the association rule is applied to maritime accident analysis, it cannot quantify the correlation between

accident characteristics, but only reveals that certain accident characteristics are related.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that might influence the work described herein.

Acknowledgments

This work is supported by the National Key R&D Program of China (Grant No.2021YFC 2801005) and Humanities and Social Sciences Foundation of Ministry of Education of the people's republic of China [Grant No. 19YJCGJW003].

References

Agrawal, R., Imielinski, T., Swami, A., 1993. Mining association rules between sets of items in large databases. Proc. ACM SIGMOD'93, 207-216.

Agrawal, R., Srikant, R., 1994. Fast algorithms for mining association rules. In: Proc. 20th Int. Conf. Very large data bases. VLDB, 487-499.

Bakdi, A., Glad, I. K., Vanem, E., Engelhardtsen, O., 2020. Ais-based multiple vessel collision and grounding risk identification based on adaptive safety domain. J. Mar. Sci. Eng. 8(1), 5.

Bostelmann, F., Wiarda, D., Arbanas, G., Wieselquist, W. A., 2022. Extension of scale/sampler's sensitivity analysis. Ann. Nucl. Energy, 165, 108641.

Brin, S., Motwani, R., Ullman, J., Tsur, S., 1997. Dynamic itemset counting and implication rules for market basket data. Proc. ACM SIGMOD'97, 255-264.

Cakir, E., Fiskin, R., Sevgili, C., 2021. Investigation of tugboat accidents severity: An application of association rule mining algorithms. Reliab. Eng. Syst. Saf. 209, 107470.

Chang, C. H., Xu, J., Song, D. P., 2014. An analysis of safety and security risks in container shipping operations: A case study of Taiwan. Saf. Sci. 63, 168-178.

Chen, J. H., Zhuang, C. L., Xu, H., Xu, L., Ye, S. M., Rangel-Buitrago, N., 2022. Collaborative management evaluation of container shipping alliance in maritime logistics industry: CKYHE case analysis. Ocean Coast. Manag. 225, 106176.

Christensen, M., Georgati, M., Arsanjani, J.J., 2022. A risk-based approach for determining the future potential of commercial shipping in the Arctic. J. Mar. Eng. Technol. 21(2), 82-99.

Datta, S., Mali, K., Ghosh, S., 2020. Weighted association rule mining over unweighted databases using inter-item link based automated weighting scheme. Arab. J. Sci. Eng. 46(4), 3169-3188.

Dobrzycka-Krahel, A., Bogalecka, M. 2022. The baltic sea under anthropopressure-the sea of paradoxes. Water. 14(22), 3772.

Fan, S. Q., Zhang, J. F., Blanco-Davis, E., Yang, Z. L., Yan, X. P., 2020. Maritime accident prevention strategy formulation from a human factor perspective using bayesian networks and topsis. Ocean Eng. 210, 107544.

Flage, R., Aven, T., 2009. Expressing and communicating uncertainty in relation to quantitative risk analysis (QRA). Reliab. Risk Anal. Theory Appl. 2, 9-18.

Fu, S. S., Yan, X. P., Zhang, D., Zhang, M. Y., 2018. Risk influencing factors analysis of Arctic maritime transportation systems: a Chinese perspective. Marit. Policy Manag. 45(4), 439-455.

Goerlandt, F., Kujala, P., 2011. Traffic simulation based ship collision probability modeling. Reliab. Eng. Syst. Saf. 96(1), 91-107.

Goerlandt, F., Khakzad, N., Reniers, G., 2017a. Validity and validation of safety-related quantitative risk analysis: a review. Saf. Sci. 99, 127-139.

Goerlandt, F., Montewka, J., Zhang, W. B., Kujala, P., 2017b. An analysis of ship escort and convoy operations in ice conditions. Saf. Sci. 95, 198-209.

Hahsler, M., Bettina, Gr., Hornik, K., 2006. Introduction to arules-mining association rules and frequent item sets. Available: https://www.researchgate.net/publication/246525355.

Hahsler, M., Grün, B., Hornik, K., Buchta, C., 2009. Introduction to arules-a computational environment for mining association rules and frequent item sets. Compr. R. Arch. Netw.

Huang, C. H., Hu, S. P., 2018. Factors correlation mining on maritime accidents database using association rule learning algorithm. Cluster Comput. 22, 4551-4559.

Huang, D. Z., Hu, H., Li, Y. Z., 2013. Spatial analysis of maritime accidents using the geographic information system. Transport. Res. Rec. 2326, 39-44.

Huang, D. Z., Loughney, S., Wang, J., 2021. Identification of China's strategic transport passages in the context of the belt and road initiative. Marit. Policy Manag. 1-26.

Kim, T. E., Nazir, S., Overgard, K. I., 2016. A STAMP-based causal analysis of the Korean sewol ferry accident. Saf. Sci. 83, 93-101.

Kleinberg, J. M., 1999. Authoritative Sources in a Hyperlinked Environment. J. ACM, 46(5), 604-632.

Kokotos, D. X., Linardatos, D. S., 2011. An application of data mining tools for the study of shipping safety in restricted waters. Saf. Sci. 49(2), 192-197.

Leone, G., 2017. The 2017 Mediterranean quality status report un environment programme. Available: https://www.medqsr.org/socioeconomic-characteristics.

Li, G. R., Weng, J. X., Wu, B., Hou, Z. Q., 2022. Incorporating multi-scenario underreporting rates into MICE for underreported maritime accident record analysis. Ocean Eng. 246, 110620.

Liu, W.W., Liu, Y.C., Bucknall, R., 2022. Filtering based multi-sensor data fusion algorithm for a reliable unmanned surface vehicle navigation. J. Mar. Eng. Technol. 22(2), 67-83.

Luo, M., Shin, S. H., 2019. Half-century research developments in maritime accidents: Future directions. Accid. Anal. Prev. 123, 448-460.

Martinez, J., Garcia-Ladona, E., Ballabrera-Poy, J., Isern-Fontanet, J., Gonzalez-Motos, S., Allegue, J. M., Gonzalez-Haro, C., 2022. Atlas of surface currents in the Mediterranean and Canary-Iberian-Biscay waters. J. Oper. Oceanogr. 1-23.

Michael, H., Christian, B., Bettina, G., Kurt, H., Ian, J., Christian, B., 2021. Mining association rules and frequent itemsets. Available: https://github.com/mhahsler/arules.

Montella, A. 2011. Identifying crash contributory factors at urban roundabouts and using association rules to explore their relationships to different crash types. Accid. Anal. Prev. 43(4), 1451-1463.

Ozaydin, E., Fiskin, R., Ugurlu, O., Wang, J., 2022. A hybrid model for marine accident analysis based on bayesian network (BN) and association rule mining (ARM). Ocean Eng. 247, 110705.

Park, H. M., Kim, J. H., 2022. Multi-Task deep learning model with an attention mechanism for ship accident sentence prediction. Appl. Sci-Basel. 12(1), 233.

Qiao, W. L., Liu, Y., Ma, X. X., Liu, Y., 2020. A methodology to evaluate human factors contributed to maritime accident by mapping fuzzy FT into ANN based on HFACS. Ocean Eng. 197, 106892.

Ramkumar, G. D., Ranka, S., Tsur, S., 1998. Weighted association rules: model and algorithm. Proc. ACM SIGKDD'1998, 145-154.

Sun, K., Feng, S. B., 2008. Mining weighted association rules without preassigned weights. IEEE T. Knowl. Data En. 20(4), 489-495.

Tirthajyoti, S., 2021. Performing uncertainty analysis in three steps: a hands-on guide. Available: https://towardsdatascience.com/performing-uncertainty-analysis-in-three-steps-a-hands-on-guide.

Ugurlu, F., Yildiz, S., Boran, M., Ugurlu, O., Wang, J., 2020. Analysis of fishing vessel accidents with bayesian network and chi-square methods. Ocean Eng. 198, 106956.

Ugurlu, O., Yildiz, S., Loughney, S., Wang, J., Kuntchulia, S., Sharabidze, I., 2020. Analyzing collision, grounding, and sinking accidents occurring in the Black Sea utilizing HFACS and bayesian networks. Risk Anal. 40(12), 2610-2638.

UNCTAD, 2022. Review of maritime transport 2022. Available: https://unctad.org/system/files/official-document/rmt2022_en.pdf

Wang, H., Liu, Z., Wang, X., Graham, T., Wang, J., 2021. An analysis of factors affecting the severity of marine accidents. Reliab. Eng. & Syst. Saf. 210, 07513.

Wang, H., Liu, Z., Wang, X., Huang, D. Z., Cao, L., Wang, J., 2022. Analysis of the injury-severity outcomes of maritime accidents using a zero-inflated ordered probit model. Ocean Eng. 258, 111796.

Wang, K., Su, M. Y., 2002. Item selection by "Hub-Authority" profit ranking. Proc. ACM SIGKDD'02, 254-260.

Wang, X., Xia, G., Zhao, J., Wang, J., Yang, Z., Loughney, S., Fang, S., Zhang, S., Xing, Y., Liu, Z., 2023. A novel method for the risk assessment of human evacuation from cruise ships in maritime transportation. Reliab. Eng. Syst. Saf. 230, 108887.

Wang, Y. F., Wang, L. T., Jiang, J. C., Wang, J., Yang, Z. L., 2020. Modelling ship collision risk based on the statistical analysis of historical data: A case study in Hong Kong waters. Ocean Eng. 197, 106869.

Weng, J. X., Yang, D., 2015. Investigation of shipping accident injury severity and mortality. Accid. Anal. Prev. 76, 92-101.

Weng, J. X., Ge, Y. E., Han, H., 2016. Evaluation of shipping accident casualties using zero-inflated

negative binomial regression technique. J. Navigation. 69(2), 433-448.

Weng, J. X., Li, G. R., 2017. Exploring shipping accident contributory factors using association rules. J. Transp. Saf. Secur. 11(1), 36-57.

Weng, J. X., Li, G. R., Chai, T., Yang, D., 2018. Evaluation of two-ship collision severity using ordered probit approaches. J. Navigation. 71(4), 822-836.

Yang, B., Zhao, Z., Ma, J., 2018. Marine accidents analysis based on data mining using K-medoids clustering and improved a priori algorithm. IOP Conf. Ser.: Earth Environ. Sci. 189, 042006.

Yang, Z. L., Wang, J., Bonsall, S., Fang, Q. G., 2009. Use of fuzzy evidential reasoning in maritime security assessment. Risk Anal. 29(1), 95-120.

Yu, Y., Chen, L., Shu, Y., Zhu, W. 2021. Evaluation model and management strategy for reducing pollution caused by ship collision in coastal waters. Ocean Coast. Manag. 203, 105446.

Zhang, Y., Sun, X. K., Chen, J. H., Cheng, C., 2021. Spatial patterns and characteristics of global maritime accidents. Reliab. Eng. Syst. Saf. 206, 107310.

Zheng, Y., Talley, W. K., Jin, D., Ng, M., 2016. Crew injuries in container vessel accidents. Marit. Policy Manag. 43(5), 541-551.