Real-time collision risk based safety management for vessel traffic in busy ports and waterways

Li, Mengxia; Mou, Junmin; Chen, Pengfei; Chen, Linying; van Gelder, P. H.A.J.M.

DOI
10.1016/j.ocecoaman.2022.106471

Publication date
2023

Document Version
Final published version

Published in
Ocean and Coastal Management

Citation (APA)
Li, M., Mou, J., Chen, P., Chen, L., & van Gelder, P. H. A. J. M. (2023). Real-time collision risk based safety management for vessel traffic in busy ports and waterways. *Ocean and Coastal Management*, *234*, Article 106471. https://doi.org/10.1016/j.ocecoaman.2022.106471

Ocean and Coastal Management 234 (2023) 106471

Contents lists available at ScienceDirect

Ocean and Coastal Management

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

Real-time collision risk based safety management for vessel traffic in busy ports and waterways

Mengxia Li [a,b,c], Junmin Mou [a,b], Pengfei Chen [a,b,*], Linying Chen [a,b], P.H.A.J.M. van Gelder [c]

[a] School of Navigation, Wuhan University of Technology, Wuhan, China
[b] Hubei Key Laboratory of Inland Shipping Technology, Wuhan, China
[c] Safety and Security Science Group, Faculty of Technology, Policy and Management, Delft University of Technology, Delft, the Netherlands

ARTICLE INFO

Keywords:
Maritime safety management
Ship collision accidents
Non-accident critical events
Risk assessment and management
NLVO
AIS

ABSTRACT

Regional risk analysis and management of maritime accidents is one of the fundamental tasks for maritime safety management. With the heavy and complicated maritime traffic in the ports and waterways, accidents, especially ship collision accidents, have been continuously posing threats to the maritime transportation system. To achieve effective and prompt identification of collision risk and to facilitate the stakeholders such as Maritime Safety Administration, this paper proposes an integrated approach for regional collision risk analysis and maritime safety management in busy ports and waterways. Firstly, regional gridding is used to link accident data and traffic data based on geographical location; Secondly, the risk model based on accident data is established. The reliability of the accident risk model is verified by data feature analysis. Finally, non-accident critical events are mined from historical accident data and traffic data as surrogate indicators of collision accidents. A regional real-time risk model is developed for integrating the accident risk model and non-accident critical events risk model by using random forest. A case study in Shenzhen port indicates that the proposed collision risk model can identify high-risk areas and facilitates maritime safety management to improve the safety level of vessel traffic in these areas. In this paper, the regional grid is used to overcome the shortcomings of different scales between data, and a real-time risk model is established by combining accidents and traffic. The 15-year maritime collision accidents are used for collision risk modeling, which improves the performance of the model.

1. Introduction

Marine transportation has been an ideal way to transport large amounts of cargo across the world for thousands of years. In recent decades, the maritime transportation system has been continuously contributing to world trade and the development of the global economy. With the expansion of shipping, the hub ports and waterways have witnessed tremendous growth in ship visits, and have been getting busier and busier. However, with the growth of vessel traffic, the risk of collision will rising(Mou et al., 2010). Collision accidents will cause severe consequence to human and economic loss, and environmental (Chen et al., 2019a; Li et al., 2019a; Yu et al., 2021).

Maritime risk is an important factor in port and ship safety management. For this reason, considerable efforts have been devoted to developing risk models for navigation safety and maritime safety management (Goerlandt and Montewka, 2015). reviewed the risks in the maritime transportation system, and the definitions of risks are various. Maritime risk can be divided into macro risk and micro risk from different applicable objects. For a single ship, micro risk can provide support for ship collision avoidance decision-making. However, for maritime safety-related stakeholders such as Maritime Safety Administrations, macro risk can provide a basis for risk-based management and policy-making (Mou et al., 2010).

There are many methods of maritime accident risk modeling, and the collision risk analysis between ships is an important aspect of maritime safety analysis. In maritime collision risk research, the most commonly used is to use the collision accident data to analyze the possibility of collision accidents and the possible consequences of collision accidents to evaluate the ship collision risk in the sea area (Montewka et al., 2012). With the popularization and application of the ship automatic identification system (AIS), ship traffic flow data are used to analyze collision risk (Qu et al., 2011; Zhang et al., 2017).

However, the current risk models have some limitations. For example, the risk model based on accident data is highly dependent on

* Corresponding author. School of Navigation, Wuhan University of Technology, Wuhan, China.
E-mail address: chenpf@whut.edu.cn (P. Chen).

https://doi.org/10.1016/j.ocecoaman.2022.106471
Received 26 September 2022; Received in revised form 28 November 2022; Accepted 23 December 2022
Available online 31 December 2022

accident data, and its modeling accuracy is greatly affected by data quality. The reliability and validity of risk models obtained by different methods are different (Goerlandt and Kujala, 2014). For this reason, how to establish an objective and reliable real-time risk model for identifying high-risk waters is the main problem to be solved in this article. To this end, this paper proposes a new collision risk modeling method that combines historical accidents with AIS data. This paper extracts a multi-factor nonparametric model from historical data. Real-time environmental data and AIS data are used as risk model inputs to predict the real-time risk of water areas. This is a macro and real-time risk model, which can identify high-risk waters and provide risk control and decision support for port safety management.

Out of this objective, the contributions of this work are as follows: 1) The concept of the geographical grid is introduced to associate traffic and collision accident data by geographical location, which overcomes the disadvantage of different scales between data; 2) Based on historical accident and AIS data, a multi-factor nonlinear nonparametric real-time risk model is established; 3) The ship encounter that has potential for the accident is identified and drawn into risk modeling to integrate the microscopic perspective of risk modeling into the model; 4) 15 years of maritime collision accidents are used in collision risk modeling to improve the performance of the model with long term observation.

The contents of the paper are arranged as follows: Section 2 illustrates a brief review on state-of-art of methods for collision risk; Section 3 illustrates the general methodology, followed by the design of the risk assessment model in Section 4; Based on the historical accident data and AIS data in the Shenzhen area, a risk model is established; real-time environment and AIS data are used for risk assessment of Shenzhen port as a case study in Section 5. A critical analysis of the proposed model and its application in the field of maritime safety management are presented in Section 6. Section 7 concludes the paper.

2. Literature review

Various research on maritime accident risk assessment has been conducted so far to reduce or control risk. The review research related to maritime risk summarizes the existing research from different perspectives such as risk definition and modeling methods (Chen et al., 2019a; Goerlandt and Montewka, 2015; Kulkarni et al., 2020; Li et al., 2012; Lim et al., 2018). According to the different data of risk assessment, the current maritime risk assessment methods can be divided into two categories: One is the risk analysis model focusing on maritime accidents; the other is the risk analysis model based on non-accident critical events (Du et al., 2020) as surrogate indicators of collision accidents.

The first kind of risk modeling method mainly relies on historical maritime accident data. Based on the accident data, the probability and density of the accident can be analyzed, and the factors affecting the accident can also be mined. There are various forms of risk based on accident data, such as the annual accident probability (Jin et al., 2002); relative accident frequency (Bye and Almklov, 2019); the combination of probability accident and consequence (Wu et al., 2019); density of accidents (Zhang et al., 2021) and so on. The number of accidents can be used as an indicator of risk model verification (Rawson and Brito, 2021). Accident statistics analysis is one of the most traditional and common methods (Kujala et al., 2009; Kum and Sahin, 2015; Rezaee et al., 2016). The regression model has also received a lot of attention in risk modeling. By analyzing the factors that may affect the occurrence of accidents, negative binomial regression (Yip, 2008) and logistic regression (Bye and Aalberg, 2018; Rezaee et al., 2016) models are used for risk modeling (Kum and Sahin, 2015). introduced Fuzzy Fault Tree Analysis (FFTA) to clarify the causes and prevent future incidents from happening. Besides (Köse et al., 1998), and (Uğurlu et al., 2013) also used the fault tree analysis method for risk analysis. Bayesian is an important tool for accident risk modeling (Hänninen and Kujala, 2012; Liu et al., 2021). The risk model based on the accident has been widely studied, which has been accepted by most researchers. The accident data itself contains rich information, so it can review the risk distribution of the water area in the past period of time. Therefore, the risk modeling based on accident data can accurately express history risk and is easy to be verified. In addition, the accident data contains rich information, which can identify the factors affecting the accident. However, solely relying on accident analysis for maritime risk modeling has the following problems: (1) Although there are many accident data in the world when focusing on a certain water area, the accident data is limited, and the risk model cannot truly reflect the regional risk level. (2) Currently, the collection and process of historical accident data are conducted manually, which leads to the inconsistency of data standards and the omission of key information; (3) The factors leading to the accident are highly coupled and nonlinear, it is difficult to use these factors from the accident to regress the risk.

To alleviate some of the limitations of the risk analysis method based on accident data, the second kind of non-accident critical events risk modeling method has attracted more and more researchers' interest (Du et al., 2020). Many non-accident critical events-related terminologies and the corresponding methods have been proposed to analyze maritime traffic risk (Lei, 2019; Zhang et al., 2015). Among the literature, one can find that non-accident critical events are often related but not limited to traffic conflict (Debnath, 2009; Lei, 2019), near-miss (Szłapczyński and Niksa-Rynkiewicz, 2018), near-collision (Watawana and Caldera, 2018; Zhang et al., 2015), collision candidate (Chen et al., 2018, 2019a, 2019b) and critical encounter (Hassel et al., 2019). These key events can indicate the potential for collision. There are many methods to detect non-accident critical events. The ship domain concept is adopted to detect the events (Chen et al., 2018, 2019a; Li et al., 2019b; Wu et al., 2016). The ship domain is the area around the ship that avoids the entrance of other obstacles for navigational safety (Li et al., 2021). When the ship domain is violated or will be violated by other ships within a certain time, a potential collision risk occurs (Kim and Jeong, 2016; Weng et al., 2012; Weng and Xue, 2015). Another method is to use the Closest Point of Approach (CPA) method to determine the conflict (Debnath and Chin, 2015). The collision risk is existing if the Distance at the Closest Point of Approach (DCPA) is less than the safe distance and Time to the Closest Point of Approach (TCPA) is positive, and vice versa. In addition, similar to the CPA method, the risk is determined by comparing with the set threshold, such as the relative distance between two ships and the time of the collision (Lei, 2019). When the relative distance is less than the threshold, the time of collision is positive, and it is also considered that there is a conflict (Li et al., 2019b). used a new distance definition as the risk judgment index, which combines factors such as length overall, distance, movement trend, and crossing angle (Gan et al., 2022). constructed a navigation risk model from four aspects: human, ship, environment, and management. When the relative distance is less than the threshold, there is a risk. Furthermore, the velocity obstacle (VO) method is a popular risk detection method. VO is the conflicting velocity leading to the collision. If the own ship's speed falls into this VO set, the collision risk occurs (Chen et al., 2018; Du et al., 2019). Differently (Zhang and Meng, 2019), proposed a probabilistic ship domain to evaluate the risk of collision. The development of these models actually benefits from the application of big AIS data in maritime safety. These methods alleviate the shortcomings of poor quantity and quality of accident data. However, whether the process from non-accident critical events to risk transformation is reasonable? It is difficult to verify the risk model based on non-accident critical events (Du et al., 2020).

To sum up, the reliability and validity of risk models obtained by different methods are different (Goerlandt and Kujala, 2014). To build a real-time and reliable risk model, this paper proposes a new risk model, which combines the above two methods, to make up for these limitations such as lack of verification, subjectivity, and unreliable. To build a real-time collision risk model, one key issue should be addressed:

How to link collision accidents with traffic to better identify and predict the collision risk in the interested areas?

Therefore, this paper proposes a real-time collision risk modeling method based on accident and non-accident critical events. This paper uses the historical data as the model input of random forest to train the real-time risk model and then uses the real-time data as the risk model input to evaluate the real-time risk of the whole water area. The specific steps of the model establishment are as follows: Firstly, traffic and collision accident data are matched by geographical location through gridding water area; Secondly, the risk model based on accident data needs to be established, respectively. The reliability of the accident risk model is verified by data feature analysis. Finally, the risk model based on non-accident critical events is established, and two models are linked by the same historical traffic and accident data through the random forest method, and a new calibration risk model is trained.

3. Methodology

Analyzing the collision risk from a macro perspective can help the Maritime Safety Administration (MSA) to understand the current level of maritime traffic risk from a management perspective and facilitate them to adopt effective maritime safety supervision methods (Chen et al., 2019a; Mou et al., 2019). Therefore, this paper aims to establish a risk model using historical accident data and traffic data, which provides an integrated risk analysis tool for the MSA and stakeholders of maritime safety. The risk model takes the real-time external environment information and traffic information as the model input, and the real-time risk value is calculated. The technical framework of this research is shown in Fig. 1, which is divided into three parts.

The first part is the basic data preparation and analysis, which is the blue part of the figure. This part is regional gridding. The grid can be used as a bridge between collision accidents and traffic data. The purpose of regional gridding is to effectively associate collision accident data with traffic data through the geographical location. The research area is divided into a rectangular grid based on the document of the ministry of transport (China, 2015a). In the meantime, the design of the grid can be based on the regional regulations for management.

The second part is the establishment of the collision risk model based on accidents, which is the green part in the figure. This part contains two important elements for the research: one is collision risk modeling, and the other is data feature analysis. The first is to establish an accident-based risk model for the collision risk assessment of each grid. The establishment of an accident-based risk model requires the consequences and collision probability of the collision accident. The consequence calculation model refers to the author's previous research and uses the concept of accident hazardous degree (Li et al., 2019a). The set pair analysis method (Li et al., 2019a) was applied to map the consequence of each accident into a determined value, defined as the

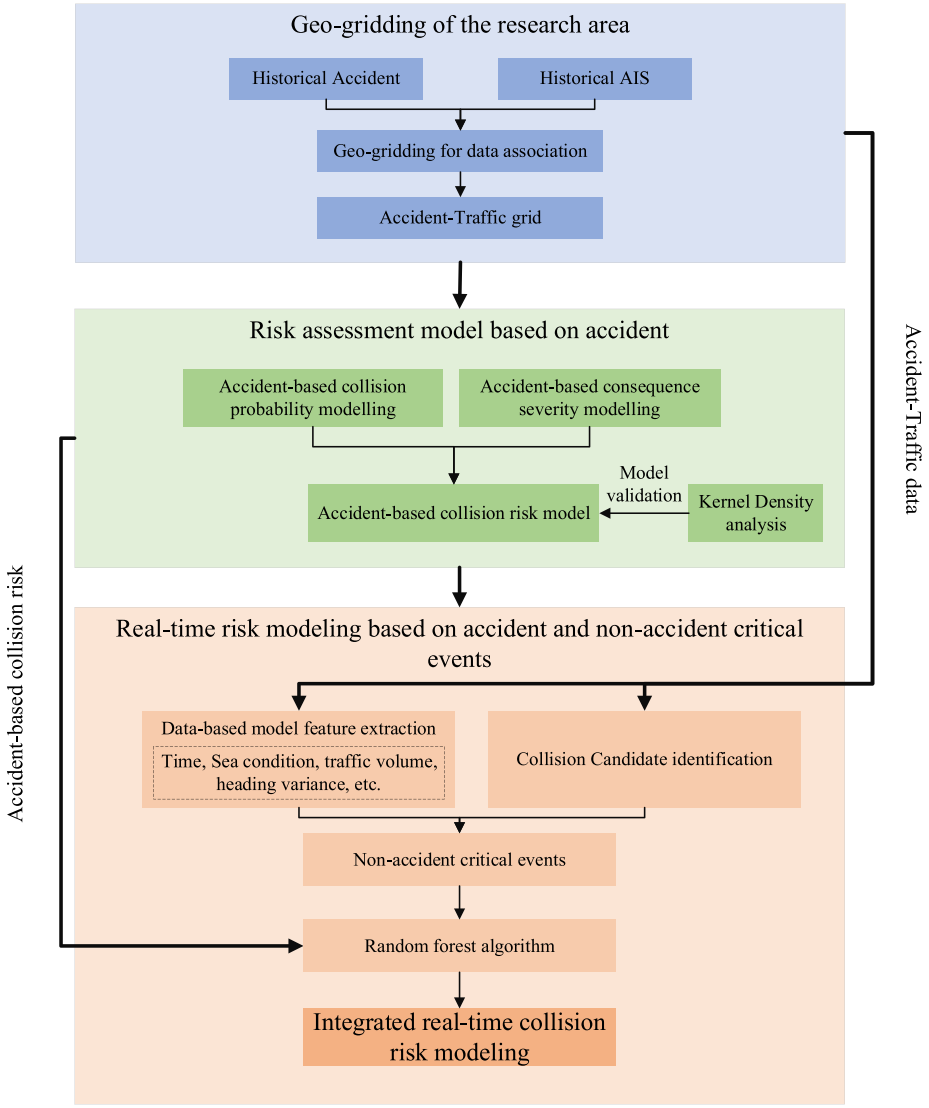

Fig. 1. Framework of the research.

hazardous degree of an accident viewed from the aspects of hull loss, fatalities, and direct economic losses. The collision probability is the ratio of the frequency of collision accidents to the traffic volume. The second is kernel density analysis through ArcGIS software, which is to provide a basis for the establishment and verification of the risk model. Through kernel density analysis of accidents and traffic, we can find out the areas where accidents occur frequently and traffic flow is dense. The reliability of the model is verified by comparing the high accident incidence areas, traffic-intensive areas, and high-risk areas calculated by the collision risk model based on accidents.

The third part is the establishment of the real-time collision risk model, which is the orange part in the figure. Firstly, we need to identify and extract the critical factors related to the collision risk from the data. Combined with the research (Li et al., 2019a) and expert experience, six factors are selected from the traffic and collision accidents data. In addition to the collision candidate set, the other factors of each grid can be obtained by statistical methods. The collision candidate set factors can be calculated by Non-linear Velocity Obstacle (NLVO) method (Chen et al., 2018). Secondly, the same historical accident and traffic are used in those two risk models, then the two models are linked by the random forest algorithm, and the accident-based risk model is used to calibrate the risk model based on non-accident critical events. Finally, the real-time collision risk model is obtained by training the random forest algorithm.

4. Risk modeling

4.1. Regional gridding

Grid management has a wide range of applications in various fields (Zheng et al., 2005). The purpose of gridding in maritime safety management is to divide the research area into geographical grids, which is efficient for MSA to manage regional safety. Besides, it also enables collision data and traffic data to be linked by geographical location. The size and shape of the grid is the key to gridding. In order to make the results of risk assessment can be directly applied to maritime grid management. The grid division standard in this research refers to the documents of the Ministry of transport (China, 2015a). The water area is divided into small squares, each of which takes 1 min in longitude and latitude as the length. To facilitate the fast location and search of the grid, this paper will adopt the encoding rules in Table 1 to code the grid. The grid number consists of a letter and four-digit numbers. The letter numbers of the grid are A, B, and C, which represent the global grid, local grid, and unit grid, respectively. The grid digit number refers to the longitude and latitude in the lower-left corner of the grid, with the latitude number in the front and the longitude number in the back. Longitude and latitude consist of degrees and minutes. The four-digit numbers of the global grid are composed of the tens and ones of degrees in the position coordinates of the lower-left corner. The four-digit numbers of the local grid are composed of units digit of degrees and tens digit of minutes in the position coordinates of the lower-left corner. The four-digit numbers of the unit grid are composed of tens and ones in the position coordinates of the lower-left corner. The unit grid is the smallest grid, and the local grid is composed of 100 unit grids. The global grid consists of 3600 unit grids.

Table 1
Encoding rule.

Grid type	code	Rule
Global grid	A	code + The tens and units digit of the degrees in the latitude and longitude coordinates of the lower-left corner of the grid
Local grid	B	code + The ones of the degree and the tens of the minutes in the latitude and longitude coordinates of the lower-left corner of the grid
Unit grid	C	code + The tens and ones of the minutes in the latitude and longitude coordinates of the lower-left corner of the grid

4.2. Risk assessment model based on accident

4.2.1. Risk modeling based on accident

Risk takes on many forms but is broadly accepted as the likelihood of danger (loss) together with an indication of how serious that danger (loss) could be (Aven, 2012; Goerlandt and Montewka, 2015; Li et al., 2021). Different definitions of risk make the scope of risk application different. In this paper, the risk is the combined value of collision accident probability and collision accident consequence, and its formula is shown in Eq. (1).

$$r_{acc}^{j} = \sum_{i}^{n} P_i \times C_{ij} \tag{1}$$

where: r_{acc}^{j} represents the risk value based on collision accident in the$j-$th grid; P_i represents the collision probability of ith year; C_{ij} represents the consequences of the collision accident in year $i-$th in the$j-$th grid.

The probability of ship collision is calculated as follows:

$$P_i = \frac{N_{acc}^{i}}{N_{traff}^{i}} \tag{2}$$

where: N_{acc}^{i} represents the number of collision accidents in year ith; N_{traff}^{i} represents the vessel traffic volume of $i-$th year.

As for the accident consequence model, there are various forms of accident consequences, such as the losses caused by accidents, the number of accidents, and so on. The loss caused by each accident is different. Only depending on the number of accidents to express the consequences of accidents may not be a good way to quantify the consequences of accidents. Therefore, the accident hazard degree based on the set pair analysis method is introduced to express the accident consequence in this paper. The accident hazard degree model is a composite severity rating system to include fatalities, injuries, property damages, hull loss, and time loss to evaluate the overall impact of an accident, which can better represent the consequences of the accident. Because the dimensions of each indicator are different, each indicator will be converted into the same type of data according to Table 3. According to the criteria for the classification of accidents (China, 2015b), 10 serious injuries are normally equal to 3 deaths. In this context, a case of severe injury is converted to 0.3 deaths, thus the indicator of fatality in Table 3 is not an integer (Li et al., 2019a). The purpose of conversion is to convert all casualty accidents to a comparable and measurable value. Therefore, the fatality here is not the death toll caused by the actual accident.

After data conversion, the accident consequence can be estimated following the equations as follows (Li et al., 2019a).

$$C_{ij} = \sum_{k}^{n} \left(w_h l_{hk} r_{hk} + w_f l_{fk} r_{fk} + w_d l_{dk} r_{dk} \right) \tag{3}$$

where: $l = \{l_{hk}, l_{fk}, l_{dk}\}$ is each evaluation indicator grade of the $k-$th accident in $i-$th year in the $j-$th grid;

$w = \{w_h, w_f, w_d\}$ is the weight of the evaluation indicator;
$r = \{r_{hk}, r_{fk}, r_{dk}\}$ is the degree of contact corresponding to level l and can be expressed as

$$r_k = \begin{cases} 0 & , p_i \le L_a \\ 1 - \dfrac{p_k - L_a}{L_b - L_a} & , L_a \le p_i \le L_b \\ 1 & , p_i \ge L_b \end{cases} \tag{4}$$

where.

$p_i = \{p_{hk}, p_{fk}, p_{dk}\}$ is the value of each evaluation indicators grade of the $k-$th accident;
L_a and L_b are the standard values of evaluating indicator p_i at a and b levels.

To discriminate between the importance of the indicators, the expert's opinions are adopted. The weight of each indicator can be obtained as: w = [hull loss, fatality, direct economic losses] = $[w_h, w_f, w_d]$ = [0.25, 0.41, 0.34].

4.2.2. Feature analysis for accident-based risk model validation

Feature analysis refers to analyzing the characteristics of collision accidents and vessel traffic data in the time and space domain. In this paper, spatial autocorrelation analysis and Kernel Density analysis are mainly used. Spatial autocorrelation analysis can determine whether there is a spatial similarity between collision accidents. Kernel density analysis can be used to determine the hot spots of collision accidents and traffic. This feature analysis can determine the qualitative relationship between accident and traffic, as well as the high-risk area of water area, and provide a reference for the validation of the risk modeling.

4.2.2.1. Spatial autocorrelation analysis. Spatial autocorrelation is a method used to analyze whether the observed value of a point is correlated with its adjacent points. Spatial autocorrelation is characterized by a correlation in a signal among nearby locations in space. Spatial autocorrelation of collision accidents means that the closer two accident points are in spatial position, the more similar they are. This paper uses Moran's I for correlation analysis.

In statistics, Moran's I index is a measure of spatial autocorrelation developed by Patrick Alfred Pierce Moran (1950). Moran's I is a widely used global index that measures the similarity for values in neighboring places from an overall mean value and reflects a spatially weighted form of Pearson's correlation coefficient. Spatial autocorrelation has been applied in maritime field (Shahrabi, 2004; Zhang et al., 2019). The Moran's I spatial autocorrelation methods are used to determine whether near collisions show spatial clustering from global perspectives (Rong et al., 2021).

Moran's I index is defined as:

$$I = \frac{N}{W} \frac{\sum_i \sum_j w_{i,j}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2} \tag{5}$$

where N is the number of spatial unit grids indexed by i and j ; x is the variable of interest; $\bar{x}$ is the mean of x; $w_{i,j}$ is a matrix of spatial weights with zeroes on the diagonal (i.e., $w_{i,i} = 0$); and W is the sum of all $w_{i,j}$.

The global Moran's I index is within $[-1, 1]$ and indicates the spatial distribution pattern. Positive values of Moran's I are associated with strong geographic patterns of spatial clustering, negative values of Moran's I are associated with a regular pattern, and Moran's I value close to zero represents complete spatial randomness (Jackson et al., 2010). If $I > 0$, the collision accidents are positively correlated in space, and the values approaching 1 indicate a strong clustering. If $I < 0$, the collision accidents have negative correlation in space, which means the collision accidents are distributed dispersedly. If $I = 0$, the collision accidents are truly randomly dispersed (perfect randomness).

In order to realize spatial autocorrelation analysis, data projection needs to be transformed into WGS_1984_UTM_Zone_50N. Then the data format is transformed from point to Polygon (where the Polygon size is the same as the unit grid size). Finally, the autocorrelation toolbox (Moran's I) of ArcGIS software is used to analyze, a spatial autocorrelation report is formed.

In the output report, in addition to Moran's I, two indicators p and z are added. the p-value represents the probability that the observed spatial pattern is created by a random process. z scores indicate the standard deviation multiple. The standard deviation can reflect the dispersion degree of a data set. According to Table 2, the confidence of Moran's I can be determined by p values and z scores.

4.2.2.2. Kernel density analysis. Spatial distribution characteristics can reveal the degree of data aggregation in space. Kernel density analysis is a tool for mining spatial distribution characteristics. This paper will use kernel density to deeply analyze the spatial distribution characteristics of traffic and collision accident data, and find their spatial hot spots.

Kernel Density calculates the density of point features around each grid. Conceptually, a smoothly curved surface is fitted over each point. The surface value is highest at the location of the point and diminishes with increasing distance from the point, reaching zero at the Search radius distance from the point. The density at each grid is calculated by adding the values of all the kernel surfaces where they overlay the raster cell center. The kernel function is based on the quartic kernel function described in Silverman (Dehnad, 2012; Luter and Silverman, 2010). So, the search radius is an important parameter in kernel density analysis. The search radius can be calculated by Eq. (6).

$$\text{Search}Radius = 0.9 * \min\left(SD, \sqrt{\frac{1}{\ln(2)}} * D_m\right) * n^{-0.2} \tag{6}$$

where: SD is the standard distance of points;D_m is the median distance of points; n is the number of points if no population field is used, or if a population field is supplied, n is the sum of the population field values.

In this study, Kernel Density analysis was carried out with the tools included in ArcGIS software. The population is set to NONE, the output pixel is set to the default value, the search radius is 0.01, the output value is DENSITIES, and the method is selected as PLANAR. Finally, the Kernel Density analysis chart is formed, and a good visualization effect is achieved by adjusting image attributes.

Finally, by comparing the areas with high accident incidence waters, high traffic density waters and high-risk waters, when the areas match, this collision risk model based on accident can be laterally verified to be effective.

4.3. Real-time risk modeling based on accident and non-accident critical events

According to Section 3, the real-time risk model will integrate the accident-based risk model and the non-accident critical events-based risk model by using historical collision accident and traffic data. This section is divided into two parts: the risk model based on non-accident critical events, the combination method.

4.3.1. Risk assessment modeling based on non-accident critical events

In the previous research by the authors(Li et al., 2019a), the traffic volume, heading variance, and speed in the traffic are identified as positively correlated with the occurrence of accidents. Therefore, in this research, such factors are analyzed and further integrated into the collision risk analysis model. The collision candidate set refers to the collision candidate set factor refers to the number of possible collisions in each grid, which is also closely related to the occurrence of accidents (Chen et al., 2018). And external environmental factors such as time and sea state also play a certain role in promoting the occurrence of collisions (Rezaee et al., 2016). Therefore, combined with the previous research and expert experience, the factors that affect the collision risk are identified, which are as follows: time, sea state, traffic volume, speed

Table 2
P values and Z scores table.

Z score (standard deviation)	p-value (probability)	Confidence level
$< -1.65, > +1.65$	<0.10	90%
$< -1.96, > +1.96$	<0.05	95%
$< -2.58, > +2.58$	<0.01	99%

Table 3
Evaluation standards.

Accident Severity	1	2	3	4	5
Hull Loss (Ship)	<0	–	–	–	≥ 1
Fatality (Number of people)	0–0.4	0.4–0.8	0.8–1.2	1.2–1.6	≥ 1.6
Direct Economic Losses (Unit: Ten Thousand CNY/ EUR)	0–10/ 0–1.3	10–20/ 1.3–2.6	20–30/ 2.6–3.9	30–40/ 3.9–5.2	≥ 40/≥ 5.2

variance, heading variance, collision candidate set, and so on.

Based on the non-accident critical events, the non-accident critical events risk can be obtained through Eq. (7). However, there is a highly nonlinear coupling relationship between these factors, which is difficult to calibrate with a simple function. Therefore, this paper will use the random forest to model and establish this risk model. See the next section for specific methods.

$$r^{j}_{\mathrm{non-acc}} = \mathrm{f}\left(\mathrm{F}^{j}_{1}, \mathrm{F}^{j}_{2}, \mathrm{F}^{j}_{3}, \mathrm{F}^{j}_{4}, \mathrm{F}^{j}_{5}, \mathrm{F}^{j}_{6}\right) \tag{7}$$

Where: $r^{j}_{\mathrm{non-acc}}$ represents the risk value based non-accident critical events in the $j-$ th grid; F^{j}_{i} represents the $i-$ th factor that affects the risk.

In order to facilitate calculation and understanding, these factors are transformed into a value in the grid through certain transformation methods. The specific transformation methods of each factor are as follows.

4.3.1.1. Factor1 (F_1): Time. Time factor refers to time. The probability of accident occurrence is different at different times of the day. This is because the navigational competency levels of crews and the environment are different at different times. Therefore, through the statistical analysis of the time of historical accident data, the frequency of accidents in different periods is calculated. Then according to the frequency of the accident, each time period is classified. In the time period of the high incidence of accidents, the grade is high; In the time period of the low incidence of accidents, the grade is low. According to the statistical results of accident data, the classification standard adopts the method of uniform distribution, which is generally divided into six levels.

4.3.1.2. Factor2 (F_2): Sea state. Similar to the time factor, the sea state factor is also one of the factors leading to the accident.

Sea state is to estimate the roughness of the sea for navigation. The sea state estimation will use the Douglas Sea scale, also known as the "international sea and swell scale", which was designed by Captain H.P. Douglas (Owens, 1984). In this paper, the sea conditions will be classified according to the Douglas Sea scale, and the classification standards can be referred to Table 4. The sea state can be determined by wave height.

Table 4
Sea state classification standard.

Name of sea surface condition	Wave height range	Sea state grade
CALM-GLASSY	0 FT (0 METERS)	0
CALM-RIPPLED	0-1/3 FT (0-.1METERS)	1
SMOOTH-WAVELET	1/3-1 2/3 FT (.1-.5 METERS)	2
SLIGHT	1 2/3–4 FT(.5–1.25 METERS)	3
MODERATE	4-8 FT(1.25–2.50 METERS)	4
ROUGH	8-13 FT(2.50–4.0 METERS)	5
VERY ROUGH	13-20 FT(4–6 METERS)	6
HIGH	20-30 FT(6–9 METERS)	7
VERY HIGH	30-45 FT(9–14 METERS)	8
PHENOMENAL	>45 FT (>14 METERS)	9

4.3.1.3. Factor3 (F_3): speed variance/Factor4 (F_4): Volume/(F_5):Heading variance. Traffic is closely related to the occurrence of accidents. Traffic volume, speed variance, heading variance are positively correlated with the occurrence of accidents. These three factors are defined as follows: Speed variance is the standard deviation of the speeds of all ships in a grid. Volume is the number of all ships in a grid. Heading variance is the standard deviation of the course of all ships in a grid. These factors need to be extracted from AIS data.

4.3.1.4. Factor6 (F_6): Collision candidate set. Collision candidate is the pair of ships in an encounter process where their Spatio-temporal relationships satisfy certain criteria that have the potential for collision (Chen et al., 2018). The criterion of collision candidate detection can be determined by Velocity Obstacle (VO) method. Velocity Obstacle sets refer to a set of velocities that can lead to a collision between two objects in the future. The specific criteria are as follows: if the velocity of one ship falls into its own VO sets induced by the other ship during the encounter process, this pair of ships will be deemed as collision candidates. This paper adopts a Non-linear Velocity Obstacle (NLVO) to detect collision candidates. The NLVO sets of Ship i induced by Ship j are denoted as $\mathrm{NLVO}_{i|j}$. After $\mathrm{NLVO}_{i|j}$ is calculated, the next step is to determine whether the velocity of the Ship i at any time falls into this set. If so, it is considered that there is a potential collision between the two ships. These potential collision locations are considered as collision candidate sets.

The specific calculation steps of the collision candidate set are as follows:

Step 1. set up the initial parameters $i = 1, j = 1$; establish AIS database, and number all the ship trajectory, and record the total number of ship trajectory n;

Step 2. extract the ith trajectory data of the Ship i from the AIS database and record it as $ShipT_i[L_i, P_i, V_i]$;

Step 3. judge if $i \leq$ n?, If so, go to next step; otherwise, go to Step 8;

Step 4. extract the jth trajectory data from the AIS database and record it as $ShipT_j[L_j, P_j, V_j]$;

Step 5. judge if $j \leq$ n?, If so, go to next step; Otherwise $i = i + 1$, and return to step 2;

Step 6. calculate $\mathrm{NLVO}_{i|j}$ by the method proposed by (Chen et al., 2018);

Step7. determine whether the velocity of the Ship i at any time falls into this set. If so, the trajectory of two ships is output, and $j = j + 1$, return to step 4; Otherwise $j = j + 1$, return to step 4;

Step 8. calculate Closest Point of Approach between two ships, and use all the Closest Points of Approach as collision candidate sets;

Step 9. end.

4.3.2. Collision risk model calibration

4.3.2.1. Model training. The final step is to link the non-accident critical events to the accident-based risk measurement. In the previous sections, the contributing factors of the non-accident-based risk analysis model have been identified. However, as aforementioned, the relationship between the factors and their influence on the collision risk is highly non-linear and is of significant difficulty to be quantified with classic formula manner. Therefore, in this section, the random forest model is utilized to identify and calibrate the complicated parameters in the risk analysis model.

The key to the new model lies in the risk assessment of the same environment and location. No matter what method is used for risk assessment, the characteristics of risk, such as temporal and spatial distribution and evolution trend, are consistent. Based on this feature,

we integrate the accident and non-accident critical events, evaluate the risk in the same location and environment. Therefore, we will replace non-accident critical events-based risk value with accident-based risk value, as shown in Eq. (8). F_i^j as the independent variable, r_{acc}^j as the dependent variable.

$$r_{acc}^j = r_{non-acc}^j = f(F_1^j, F_2^j, F_3^j, F_4^j, F_5^j, F_6^j) \tag{8}$$

It is worth noting that there are multiple accident points in a grid, the sea condition and time of the accident need to be converted into a value, and Eq. (9) is as follows:

$$\left[F_1^j, F_2^j\right] = \left[\frac{\sum_{m=1}^{M} F_1^{mj}}{M}, \frac{\sum_{m=1}^{M} F_2^{mj}}{M}\right] \tag{9}$$

where: F_1^{mj} is the collision time factor of the mth accident in the $j-$ th grid; F_2^{mj} is the sea condition at the time of collision of the mth accident in the $j-$ th grid; M is the number of accidents in the $j-$ th grid.

When the input and output of the model have been obtained, the most important step is to obtain the risk model. To better build the risk model, this paper will use the machine learning method to obtain the risk model. This method is relatively objective and can well simulate the coupling between various factors.

There are many methods of machine learning. Random forest is a machine learning method proposed by (Breiman, 2000). Random forest is a combination of multiple decision trees, each tree depends on the value of random vectors sampled independently, and all the trees in the forest have the same distribution. The generalization error tends to converge with the increase of trees in the forest. The steps of modeling and forecasting are as follows:

① Firstly, N_{RF} is used to represent the number of samples in the original training sample set, and M_{RF} is used to represent the number of attributes.
② Secondly, determine a fixed value m_{RF} ($m_{RF} < N_{RF}$), which is used to determine how many attributes will be selected when making decisions on a node.
③ Bootstrap resampling technology method is used to randomly extract K_{RF} training data sets from the original training sample set, and K_{RF} decision trees are constructed. The samples that are not extracted each time form out of bag data, that is, out of bag data, or OOB for short, which can be used to predict the accuracy of classification.
④ Each training data set grows into a single decision tree. In each node of the tree, m attributes are randomly selected from m attributes. According to the principle of minimum node impure, one of the M features is selected for branch growth. Let the tree grow sufficiently to minimize the impurity of each node, and do not prune in this process.
⑤ According to the trained random forest algorithm, the risk can be obtained by inputting the influencing factors.

4.3.2.2. Accuracy evaluation of collision risk model. After training the collision risk model with a random forest algorithm, we need to evaluate the accuracy of the model. The test data will be input into the trained collision risk model, the predicted collision risk value will be output. By comparing the predicted risk value with the actual test risk value, the smaller the difference is, the better the performance of the training model is.

In this paper, the Goodness of Fit (R^2), Mean Absolute Error (MAE), and Root Mean Square Error (RMSE) are selected to evaluate the accuracy of the model.

The R^2 can be used to measure whether the data not involved in training can be well predicted by the risk model. The R^2 can be calculated by Eq. (10). The range of R^2 is [0,1]. The larger the R^2 is, the better the fitting between the predicted value and the actual value of the training risk model is.

$$R^2 = \frac{\sum_{i=1}^{n}(r_i - \bar{r})^2 - \sum_{i=1}^{n}(r_i - \widehat{r})^2}{\sum_{i=1}^{n}(r_i - \bar{r})^2} \tag{10}$$

where: r_i is the predicted risk value; $\widehat{r}_i$ is the actual risk value.

MAE is the average value of the absolute value of the deviation between the predicted risk value and the real risk value. The MAE can be calculated by Eq. (11). The weights of all the differences in the average value are equal, which can reflect the actual situation of the difference between the predicted risk value and the real risk value.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|r_i - \widehat{r}_i\right| \tag{11}$$

RMSE is the square root of the ratio of the square of the deviation between the predicted risk value and the real risk value and the number of observations. The RMSE can be calculated by Eq. (12). RMSE measures the deviation between the predicted risk value and the real risk value. The smaller the RMSE is, the smaller the deviation between the predicted value and the real value is, the higher the accuracy of the model prediction is.

$$RMSE = \left(\frac{1}{n}\sum_{i=1}^{n}\left(r_i - \widehat{r}_i\right)^2\right. \tag{12}$$

5. Case study

5.1. Research area gridding

Shenzhen is located on the east bank of the estuarine of the Pearl River and it is adjacent to Hong Kong, China. The Port of Shenzhen is a collective name of a number of ports along with parts of the coastline of Shenzhen, Guangdong Province, China. These ports, as a whole, form one of the busiest and fastest-growing container ports in the world. With expansion of the shipping business, more and more attentions have been directed to the safety management of the waterways and ports. The Shenzhen Port can be conducted as a good case study (Fig. 2). According to the principle of grid division criteria, combined with the characteristics of the waters in Shenzhen, the Shenzhen waters were divided into a total of 428-unit grids, as shown in Fig. 2.

After the area is gridded, AIS and accident data need to be linked by geographical location. This case study needs to use historical collision accidents and AIS data in the research waters to establish a collision risk model. Working with the Shenzhen Maritime Bureau, which is the main authority for port management in the area, the authors gained access to the collision accident records for the years from 2002 to 2017. Therefore, the collision accident data have high accuracy and can be directly analyzed. AIS data is provided by the traffic flow Laboratory of Wuhan University of technology. Due to a large amount of AIS data, one day of AIS data was randomly extracted from each month in 2019. After training the risk model, real-time AIS data and environmental data are needed to analyze the risk. In addition, AIS data has some abnormal data. Before using these data, it is necessary to detect and remove abnormal data. Firstly, the abnormal data are directly eliminated according to MMSI format errors, out-of-range latitude and longitude, and other illegal errors. Secondly, the abnormal data are further eliminated based on the average speed and the average change rate of course over the ground between two adjacent points (Guo et al., 2021).

In this paper, a total of 129 collision accidents occurred in Shenzhen waterways from 2002 to 2017. The time distribution of the collision accident is shown in Fig. 3. The frequency of collision accidents shows a downward trend on the whole.

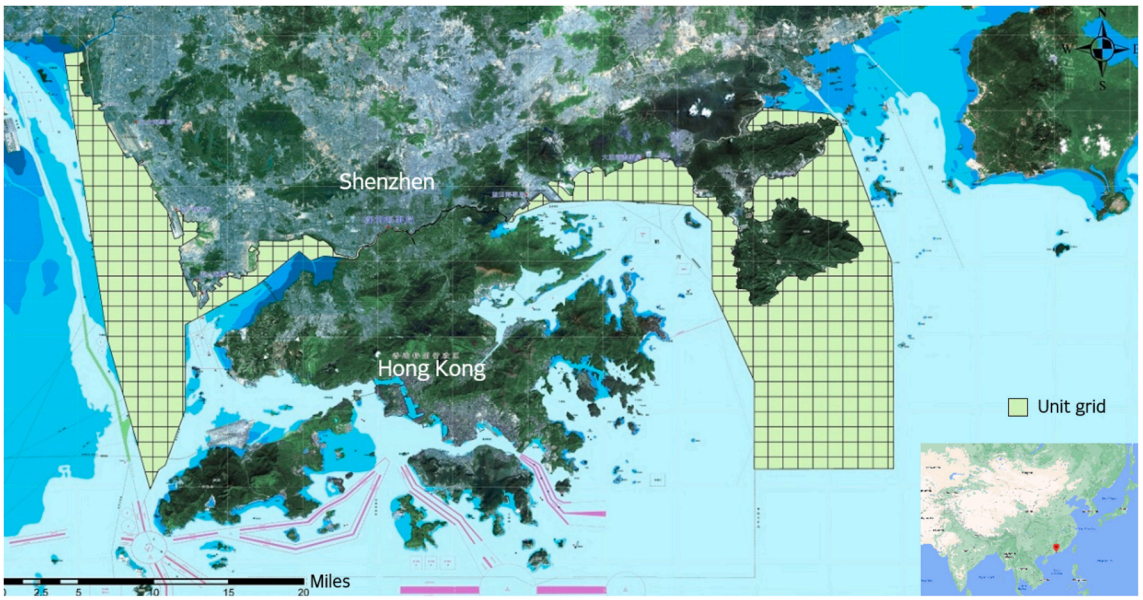

Fig. 2. Research area.

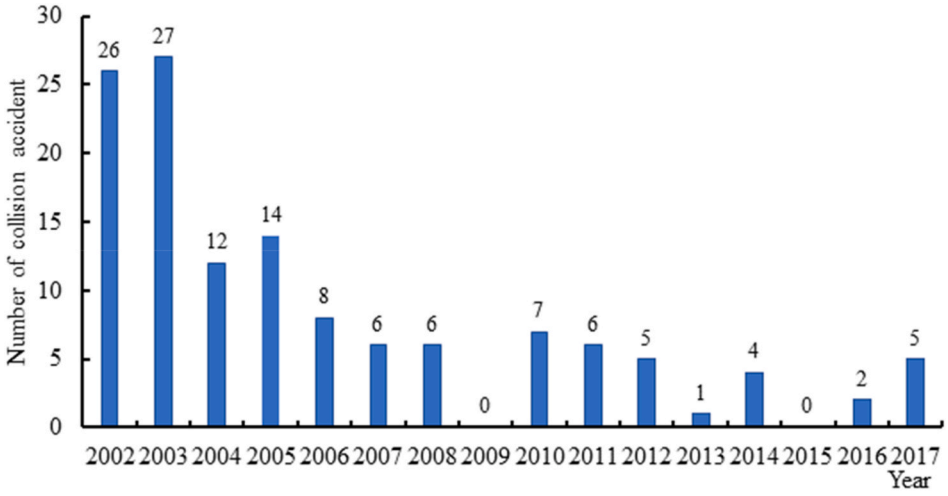

Fig. 3. Time distribution of collision accidents.

The traffic volume in Shenzhen from 2002 to 2017 was selected as the statistical object, and the statistical results are shown in Fig. 4. From 2002 to 2007, the total traffic volume of Shenzhen continued to grow, and from 2007 to 2017, the traffic volume tended to be stable, close to 500,000 ships.

5.2. Collision risk based on observed accidents

5.2.1. Collision risk based on observed accidents

The risk model based on accident data has been proposed in Section 2.3.1. According to the model, we need to calculate the probability of an accident every year and the consequences of each accident in each grid according to Eq. (3), the results are as shown in Appendix I. Finally, we use Eq. (1) to calculate the risk value of each unit grid, as shown in Fig. 5. The four-unit grids with the highest risk are C3150, C3051, C2752, and C2753, and their positions are shown in Fig. 6.

5.2.2. Data feature analysis

Feature analysis is to better grasp the characteristics of data, which is to provide a basis for the establishment and verification of the risk model.

5.2.2.1. Accident spatial autocorrelation analysis. The purpose of accident spatial autocorrelation analysis is to analyze whether there is a correlation between accidents in a space unit and accidents in other space units around it. In other words, whether there is interdependence between adjacent accidents.

Firstly, the collision accidents are plotted on the GIS map, which can have a preliminary understanding of the spatial distribution of the collision accident, as shown in Fig. 7. The collision accidents are mainly concentrated in the western waters, and there are relatively few accidents in the eastern waters. The collision accidents are mainly concentrated in the western waters (such as Shekou, Dachan Bay, Chiwan and Mawan, etc.).

Moran's I spatial autocorrelation analysis tool provided by ArcGIS is used to analyze spatial correlation. The Moran index is 0.66, which indicates that the collision accidents have a positive spatial correlation, and the data set for analysis is proportional to the spatial aggregation. Conforming to Table 2, the p-value is less than 0.01, the z score is greater than 2.58, confidence is 99%. That means the probability of random

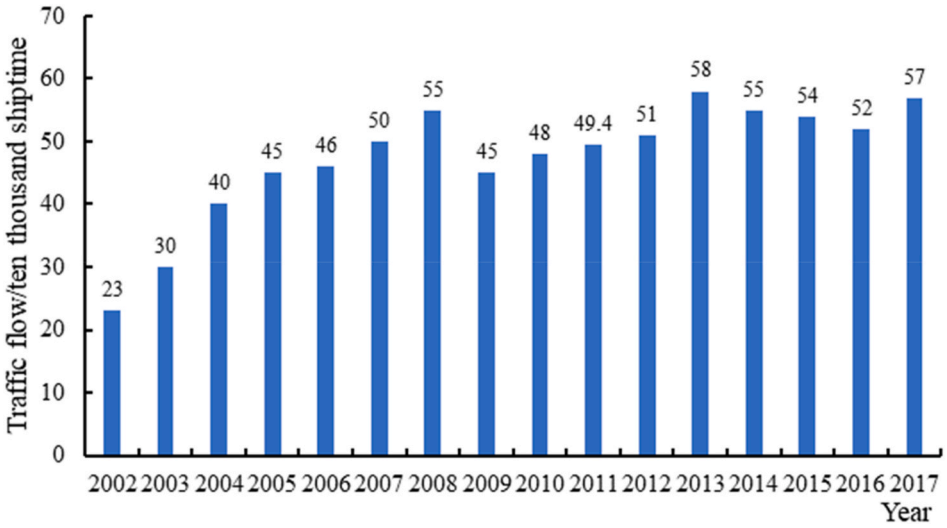

Fig. 4. Shenzhen port traffic volume statistics.

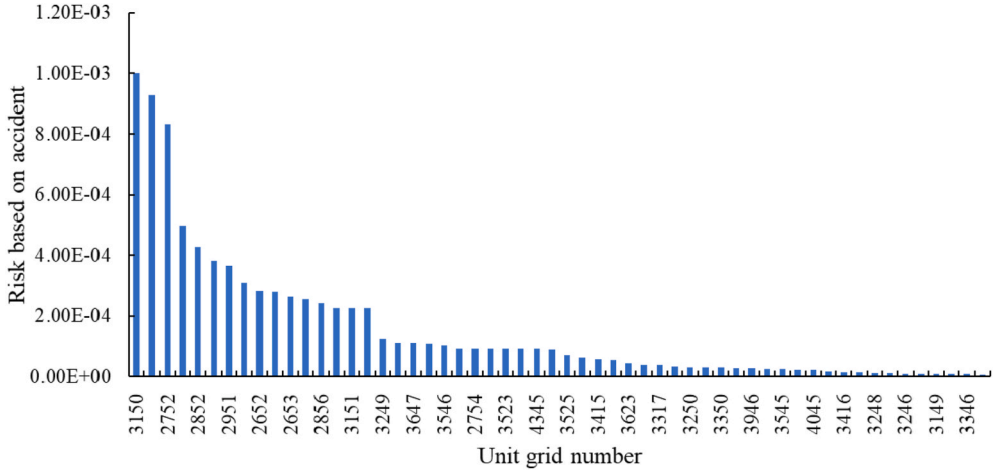

Fig. 5. Collision risk based on accidents in-unit grids.

generation of this data is only 1%, which shows significant clustering characteristics. Moran's I > 0 represents a positive spatial correlation. The collision accidents are positively correlated in space, which further shows the necessity of regional grid management and the reliability of regional grid risk assessment.

5.2.2.2. Kernel density analysis of collision accidents. With the kernel density analysis of the collision accidents, one can intuitively grasp the spatial distribution of collision accidents in waters and explore the waters with a high occurrence of collision accidents. In this paper, *129 collision accidents from 2002 to 2017 were selected*. Based on the gridding of the water area, the kernel density of accident data is analyzed, and the spatial distribution of collision accidents in the water area is obtained and shown in Fig. 8. As shown in Figure, the accident density distribution is uneven, and the eastern waters are significantly lower than the western waters.

In order to analyze the traffic spatial distribution in the water area, this paper selected the AIS data for one day in 2019. On the basis of the grid of the water area, kernel density analysis was used to analyze the spatial distribution of the traffic volumes in the water area, as shown in Fig. 9. The distribution of vessel traffic volumes density is uneven, and the high-density area in the eastern waters is significantly less than that in the western waters.

From the above analysis of the density of accidents and traffic flow, we can see that the hot spots of accidents and traffic flow show a high level of similarity. This also shows that there is a strong correlation between collision accidents and traffic volume. The more intensive the traffic volume is, the more likely the collision accident will occur.

Comparing Figs. 6, 8 and 9, the location with high risk is consistent with the location with high accident frequency and dense traffic. To make the risk results more reliable, we consulted the staff of the Maritime Safety Administration in the area, and the actual grid they focused on monitoring was consistent with our calculation results. Therefore, from this point of view, the collision risk model in this paper is relatively reliable.

5.3. Real-time risk model in Shenzhen waters

5.3.1. Risk-based on non-accident critical events

According to the above, the non-accident critical events will be extracted from historical accident and AIS data, respectively. Time factor and sea state factor will be obtained from historical accident data, while the other four factors will be extracted from AIS data. However, since the AIS data at the time of the accident cannot be obtained, the pattern of vessel traffic is assumed as evenly distributed during the accident occurrences every year. So the AIS data in 2019 will be assumed

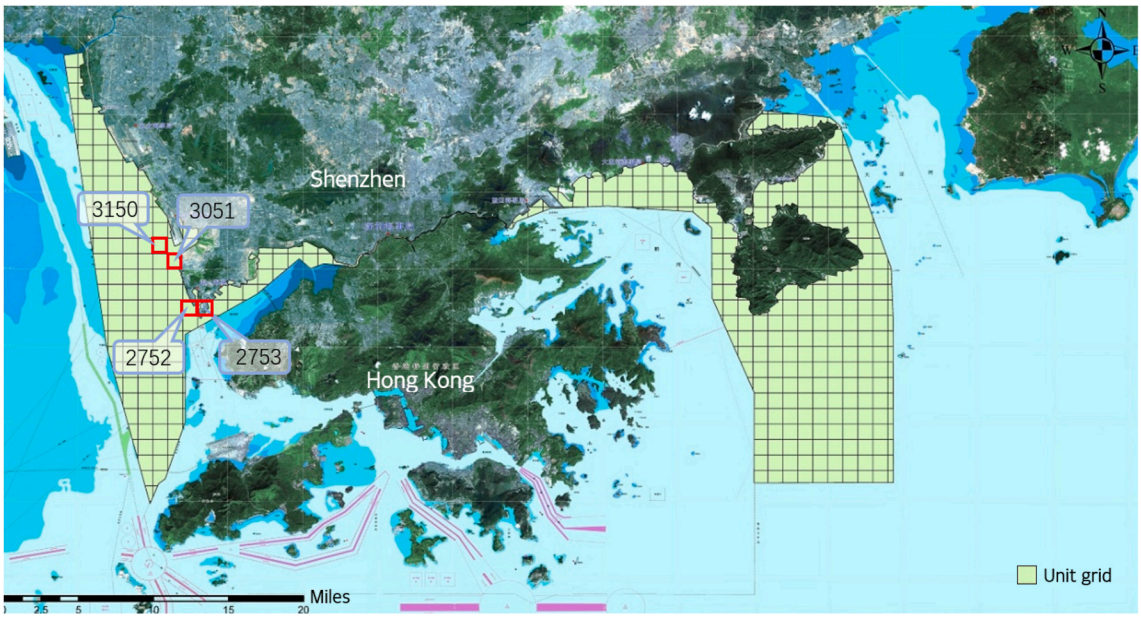

Fig. 6. Risk grid geography distribution map.

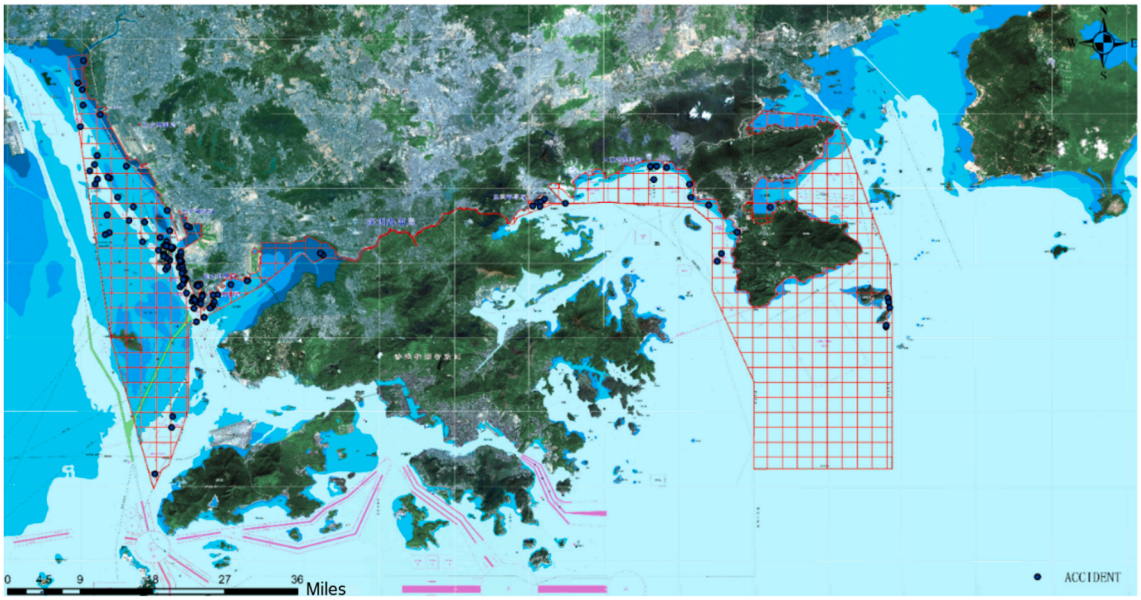

Fig. 7. Spatial distribution of collision accidents indicated by blue dots.

to be the AIS data at the time of the accident.

F_1. Time: Firstly, the time of 129 collision accidents is counted, as shown in Table 5. Then, the time factor is classified according to the frequency of accidents at different times, which is shown in Table 6. Finally, the occurrence time of 129 collision accidents is converted into the corresponding grade. Since there are multiple accidents in a grid, the corresponding accident time also has multiple values, so it needs to be integrated into one value according to Eq. (11).

F_2. Sea state: According to Table 4 sea state classification standard, the sea state at the time of the accident is classified. Since the sea state data in a grid has multiple values as time data, a value is converted according to Eq. (11).

F_3. Speed variance: Speed variance is the standard deviation of speed in a grid. Theoretically, it should be the standard deviation of velocity in the grid when the accident occurred. Therefore, based on the 12 day AIS data, the speed variance per day in each grid is calculated, and then the 12-day data are averaged.

F_4. Volume: Volume is the number of ships in a grid. Theoretically, it should be the number of ships in the grid when the accident occurred. Therefore, the average ship volume in each grid per day is calculated based on the AIS data of twelve days. To make the traffic volume closer to the traffic volume at the time of the accident, it is necessary to transform the traffic volume. The premise of traffic flow transformation is that the distribution of traffic flow in the waters is relatively stable. The paper assumes that the dense distribution of traffic flow in the study area is consistent during the study period. N^{ij}_{traff} can be calculated by Eq. (16).

$$N^{ij}_{traff} = \frac{N^{uj}_{traff}}{N^{u}_{straff}} \times N^{i}_{straff} \tag{13}$$

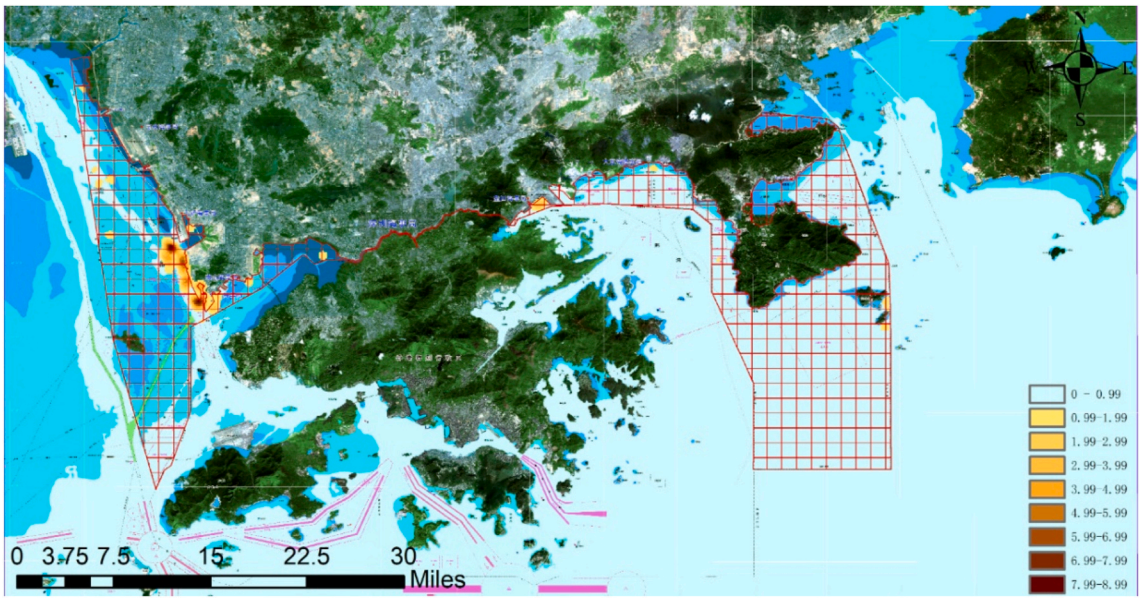

Fig. 8. Accident kernel density analysis.

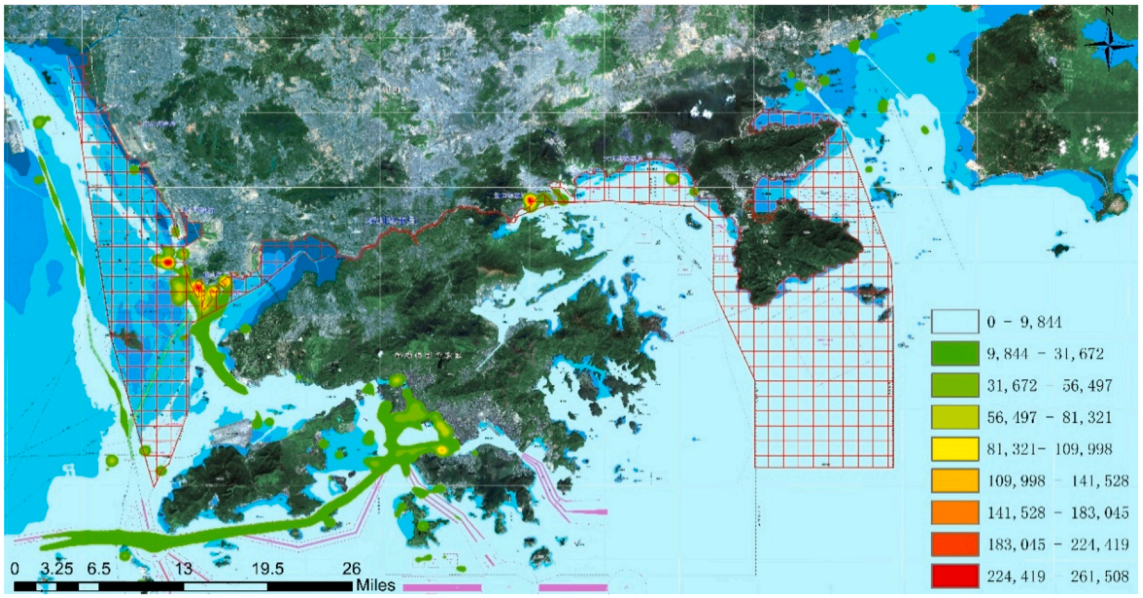

Fig. 9. Traffic kernel density analysis.

Table 5
Statistics of accident occurrence time.

Time	Frequency	Time	Frequency
0:00–1:00	3	12:00–13:00	1
1:00–2:00	3	13:00–14:00	2
2:00–3:00	5	14:00–15:00	1
3:00–4:00	3	15:00–16:00	2
4:00–5:00	6	16:00–17:00	2
5:00–6:00	4	17:00–18:00	4
6:00–7:00	2	18:00–19:00	1
7:00–8:00	1	19:00–20:00	3
8:00–9:00	4	20:00–21:00	3
9:00–10:00	2	21:00–22:00	3
10:00–11:00	4	22:00–23:00	3
11:00–12:00	3	23:00–24:00	6

Table 6
Statistics of accident occurrence time.

Time of accident	Grade
23:00–24:00,4:00–5:00	6
2:00–3:00	5
5:00–6:00,8:00–9:00 10:00–11:00,17:00–18:00	4
0:00–2:00, 3:00–4:00, 19:00–23:00, 11:00–12:00	3
6:00–7:00,9:00–10:00, 13:00–14:00, 15:00–17:00	2
7:00–8:00, 12:00–13:00, 14:00–15:00, 18:00–19:00	1

where: N^{tij}_{traff} represents the traffic volume of year u in the jth grid; N^{i}_{straff} represents the total traffic volume of the i year; N^{u}_{straff} represents the total traffic volume of the u year.

F_5. Heading variance: Heading variance refers to the course standard deviation of all ships in a grid. Therefore, according to the twelve

days AIS data, the heading variance in each grid of each day is calculated, and then the twelve days data is averaged.

F_6. Collision candidate set: The collision candidate set refers to the number of ships that may collide in each grid. Theoretically, it is also calculated by AIS data at the time of the accident. Therefore, according to the twelve days AIS data, the number of ships that may collide in each grid is calculated by the NLVO method, and then the number of ship accidents that may collide in each grid every day is calculated as the input of F_6.

5.3.2. Real-time risk model calibration

After obtaining the factors and risk of each unit grid, the spatial location of the grid is used to associate the factors with the risk based on collision accidents. Random forest algorithm is used to integrate accident-based risk and non-accident critical events-based risk. Each factor calculated in section 5.3.1 and the risk value calculated in Section 5.2.1 is taken as the input and output of the random forest model respectively. The real-time collision risk model can be obtained by training of random forest algorithm. The parameter setting of the random forest model is shown in Table 7.

80% of the data is used as training data and 20% as test data. The training results are shown in Fig. 10. Fig. 10(1) shows the out-of-bag error for training result 1. It can be seen from the figure that with the increase of decision tree, the out of bag error becomes smaller and smaller, and finally stabilized at 3.07E-08. Fig. 10 (2) shows the importance of each factor in the risk model, F_4 has the highest importance, about 0.47, and F_1 has the lowest importance, about −0.04. There was no significant difference in the importance of the other factors. The importance of F_6, F_3, F_2 and F_5 were 0.29, 0.16, 0.11 and 0.03 respectively. Fig. 10 (3) shows the goodness of fit between actual train risk values and predicted values, the R^2 is 0.78. Fig. 10 (4) shows the goodness of fit between actual test risk values and predicted values, R^2 is 0.65. Fig. 10 (5) and Fig. 10 (6) show the comparison between the training risk data and the real risk value, and the comparison between the test risk data and the real risk value, respectively. From the trend point of view, the trend is basically the same. MAE is 1.87E-04, RMSE is 2.67E-04. In general, the fitting between the predicted value and the actual value of the training risk model is good.

According to the above training result 1, F_1 has the lowest contribution to the risk model. Therefore, to further improve the accuracy of the risk model, F_1 will be removed for risk model training, and the results are shown in Fig. 11. Fig. 11(1) shows out-of-bag error for training result 1. It can be seen from the figure that with the increase of decision tree, the out of bag error becomes smaller and smaller, and finally stabilized at 4.76E-08. Fig. 11 (2) shows the importance of each factor in the risk model, F_4 has the highest importance, about 0.40. The importance of F_6, F_3, F_2, and F_5 were 0.37, 0.30, 0.21 and 0.06 respectively. Fig. 11 (3) shows the goodness of fit between actual train risk values and predicted values, the R^2 is 0.83. Fig. 11 (4) shows the goodness of fit between actual test risk values and predicted values, R^2 is 0.88. Fig. 11 (5) and Fig. 11 (6) show the comparison between the training risk data and the real risk value, and the comparison between the test risk data and the real risk value, respectively. From the trend point of view, the trend is basically the same. MAE is 6.56E-05, RMSE is 8.18E-05. In general, the fitting between the predicted value and the actual value of the training risk model is good.

Table 7
The parameter setting of random forest model.

Parameter	**Value**	
Method	regression	Method used by trees. The possible values are 'classification' for classification ensembles, and 'regression' for regression ensembles.
Surrogate	on	A matrix of size Nvars-by-Nvars with predictive measures of variable association, averaged across the entire ensemble of grown trees. If you grew the ensemble setting 'surrogate' to 'on', this matrix for each tree is filled with predictive measures of association averaged over the surrogate splits. If you grew the ensemble setting 'surrogate' to 'off' (default), SurrogateAssociation is diagonal.
minleaf	5	Minimum number of observations per tree leaf. By default, MinLeafSize is 5 for regression.
NumTrees	3000	Scalar value equal to the number of decision trees in the ensemble.

5.4. Real-time risk assessment result

The real-time AIS data and environmental data of this water area from 10:40 to 11:40 on January 6, 2019, are selected for risk assessment to identify the water area with relatively high risk. According to the collected environmental data, the sea condition is grade 3. Input the environmental data and AIS data into the risk model trained in section 5.3. The grid with the top 20 risk values is shown in Table 8, and its spatial distribution is shown in Fig. 12. And nineteen grids with relatively high-risk areas in the western waters of Shenzhen and one in the eastern waters of Shenzhen. The relative risk of C3050, C2653, C2854, and C2752 is high.

In order to compare the identification results of this high-risk water area, we collected the accident data of this water area at the same time. About at 12:00 on January 6, 2019, a ship collision accident occurred in the C2652 grid. C2652 is the high-risk grid identified in this risk assessment, which needs to be taken risk mitigation measures. This shows that the high-risk waters identified by this model have reference values.

6. Discussion

6.1. Case result discussion

This paper proposes a real-time regional risk modeling method based on long-term collision accident and traffic data. The steps of real-time risk modeling are as follows: firstly, the research area is gridded; then the collision risk based on accident data is evaluated for each grid; secondly, the factors in different grids are extracted from the same collision accident and traffic data; finally, a real-time risk model is obtained by correlating risk and factors with random forest. Taking Shenzhen port as a case study, this paper uses 129 collision accidents from 2002 to 2017 and AIS data from 2019 to establish a real-time risk model for Shenzhen port. The real-time AIS data of a day is selected to evaluate the collision risk of Shenzhen port.

The results show that: (1) the historical high-risk areas of Shenzhen port are located in Shekou, Dachan Bay, Chiwan, and Mawan; (2) The factors that contribute the most to the risk of Shenzhen Port are traffic volume and collision candidate set; (3) There is a positive correlation between collision accidents and traffic; (4) These results coincided with the previous knowledge and experience of experts in vessel traffic management; (5) Risk situation of Shenzhen port on one day is evaluated, and the high-risk location is matched with the accident location on that day.

(1) When other parameters in the random forest model remain unchanged, the indexes of the risk model trained before and after removing F_1 factor are shown in Table 9. It can be seen from the table that after removing factor F_1, Out of Bag Error is stable earlier, and the stable value is lower. From the contribution of factors to the risk model, F_4, F_6, F_3, F_2, and F_5 are in the same order. From the perspective of data goodness of fit, the goodness of fit is higher after removing factor F_1. From MAE and RMSE, the error is smaller after removing factor F_1. Generally, the result of the training risk model is better after removing factor F_1 (Nielsen and Jungnickel, 2003). also proposed the view that the impact of

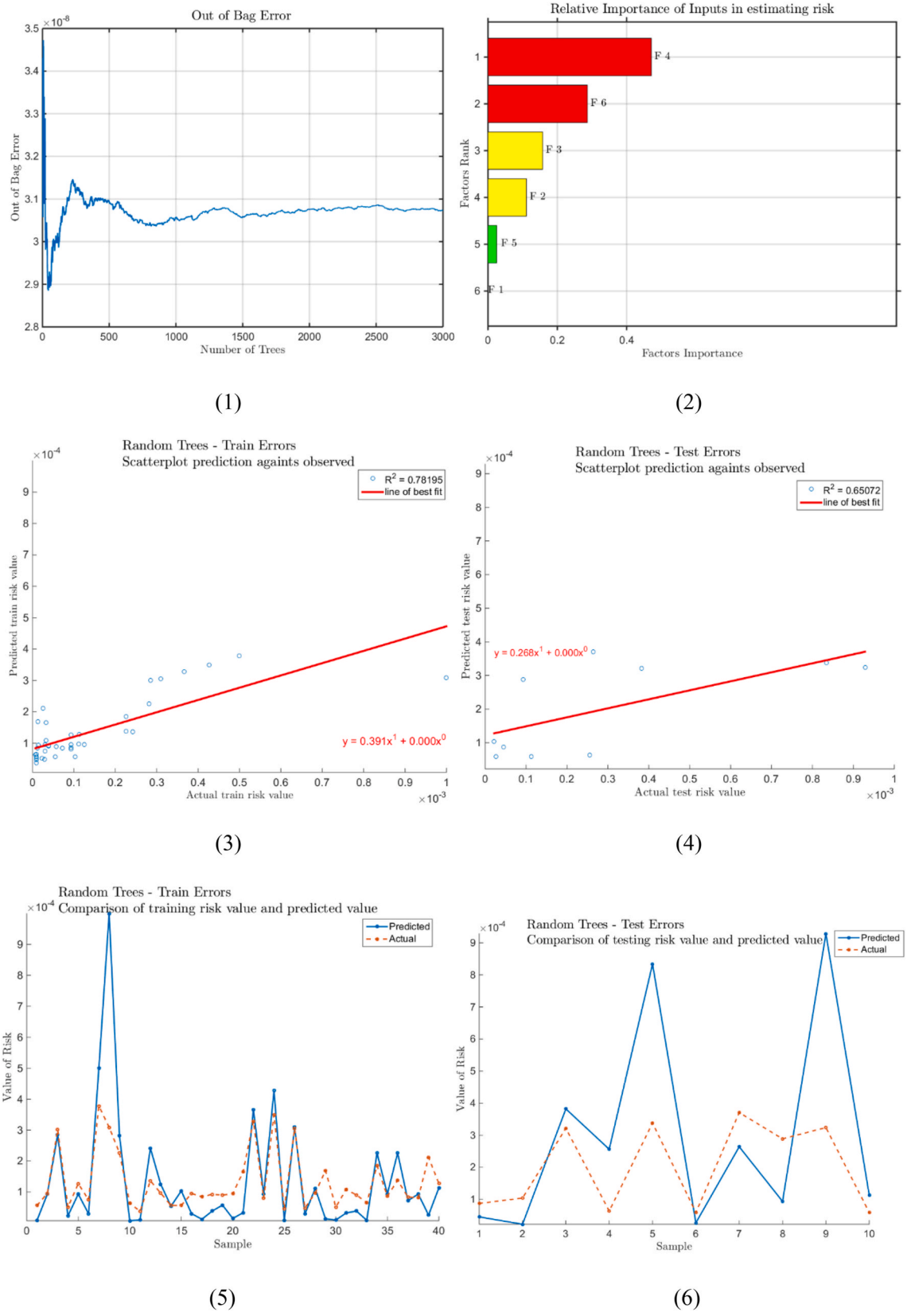

Fig. 10. Risk model training result 1.

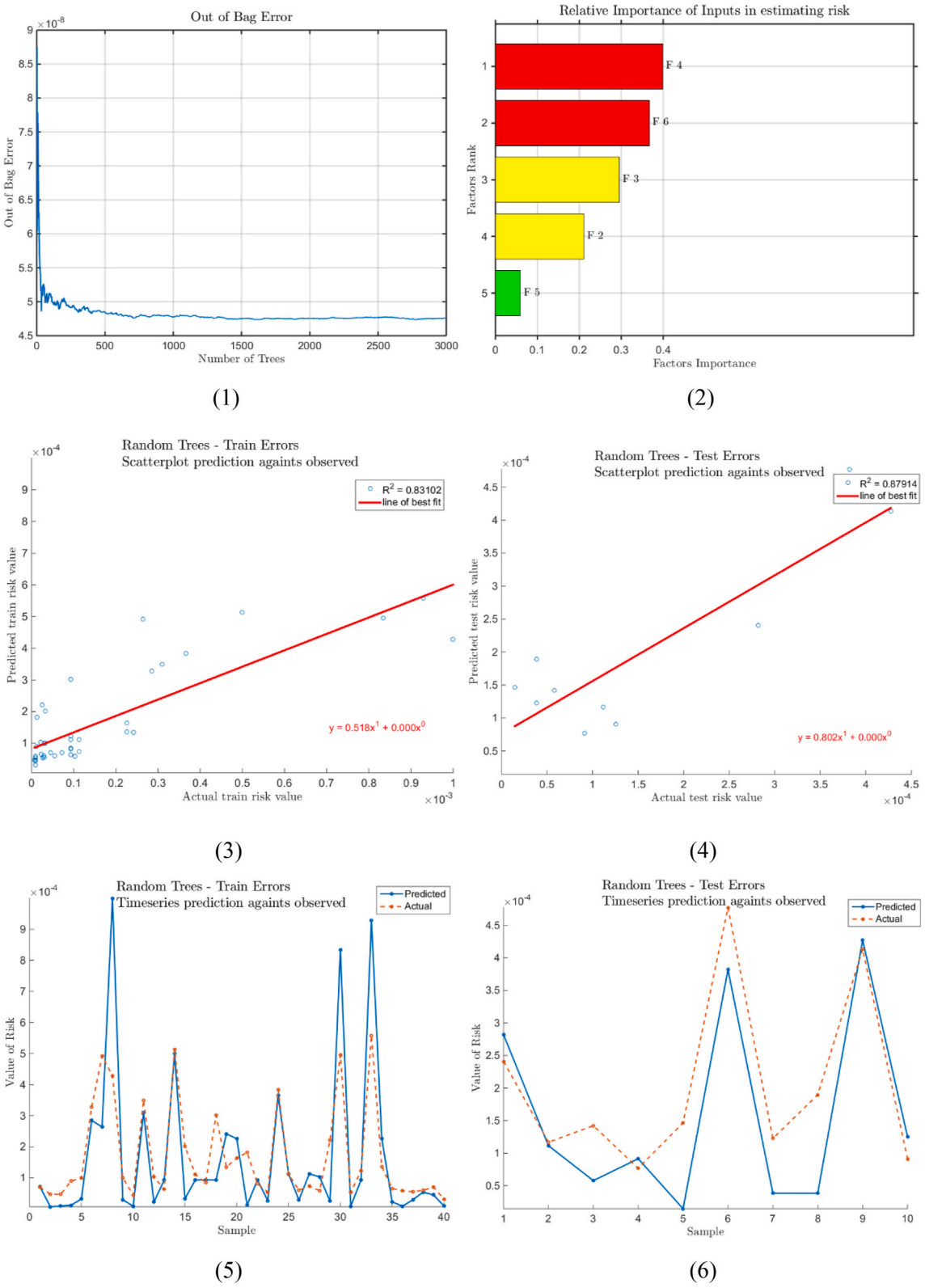

Fig. 11. Risk model training result 2.

the time factor on the accident is limited. This may be related to the collected accident data itself.

From the change of contribution degree of each factor, the interaction between various factors for the collision risk model is too complex and may not be a linear correlation model. To further confirm this point, we analyzed the linear correlation analysis between collision risk and factors, as shown in Fig. 13. It can be seen from the figure that the linear correlation between collision risk and each factor is low. Heading variance, Traffic volume, and collision candidate set are positively correlated with risk. With the increase of heading variance, traffic volume, or collision candidate set, the risk also increases. For the sea

Table 8
Real-time risk value (Top 20).

Grid Number	Real-time risk	Grid Number	Real-time risk
C3050	1.46E-04	C2451	1.29E-04
C2653	1.46E-04	C2149	1.28E-04
C2854	1.45E-04	C2652	1.28E-04
C2754	1.44E-04	C2250	1.28E-04
C2752	1.30E-04	C2351	1.27E-04
C2450	1.29E-04	C1951	1.25E-04
C2851	1.29E-04	C2248	1.25E-04
C2450	1.29E-04	C2835	1.25E-04
C2951	1.29E-04	C3248	1.24E-04
C3051	1.29E-04	C2852	1.24E-04

conditions factors, the relationship shows that most of the accidents in our training data set to occur in the case of low sea state. This may be because when the sea state is high, the traffic volume at sea will be reduced, and the crew will navigate more carefully, so the accident rate will be less. From the performance of speed variance on risk, the value of speed variance is mainly concentrated between 0 and 5. The velocity in each grid is average and the variance is small. According to the existing data, the greater the speed variance, the greater the risk is not necessarily. This also indicates that the speed variance may not be able to express the degree of traffic disorder. Heading variance can indicate the degree of traffic disorder.

(2) From the case study of Shenzhen port, the two factors that contribute most to the risk are traffic volume and collision candidate set. This shows that the larger the volume is, the more likely accidents will occur. The collision candidate set is based on AIS data to predict the number of possible collisions in the grid, which is also the most intuitive parameter to represent the possibility of collision (Chen et al., 2018). proposed a collision candidate detection method, but the research did not explore the relationship between collision candidate set and risk. In this paper, the contribution value of collision candidate set to risk is obtained. In addition, heading variance and speed variance also contributes to the risk. All those factors are extracted from AIS data. It also proves that collision accidents are closely related to traffic.

(3) According to the analysis of the characteristics of accidents and traffic, it is known that there is a certain spatial correlation between collision accidents and traffic. According to the risk assessment results of Shenzhen port, the high-risk area is the area with a high probability of accident occurrence and dense traffic. This directly shows that there is a correlation between collision accidents, traffic, and risks.

(4) The grid with the highest risk does not necessarily have an accident. The occurrence of accidents is accidental. However, the water area with high risk indicates dense traffic and a high possibility of accidents, which need to provide to MSA. If MSA can take measures in advance, they may reduce the occurrence of accidents to some extent.

6.2. Advantages of the model

(1) The risk model proposed in this paper can not only do the real-time risk assessment but also assess the historical risk of a certain region in a certain period of time. Based on the accident risk model, the historical risk of the region can be evaluated, and the high-risk waters can be analyzed from the historical risk. Besides, the trained risk model can be used for real-time risk assessment according to the real-time input of each factor value in a grid. Historical risk distribution can give maritime management agencies a risk warning, clearly focusing on high-risk waters. Real-time risks can make it possible for maritime management agencies to dynamically monitor maritime traffic safety. And the maritime management agencies can timely give effective risk control measures to reduce the risk.

(2) This paper introduces the concept of grid management, which is consistent with the practice of maritime management. Since 2015, maritime grid management has been implemented in the maritime supervision of Chinese waters. The grid generation

Table 9
Comparison of training models.

Index		Result	Result 1
Out of Bag Error		**3.07E-08**	4.76E-08
Factor Rank		**F_4, F_6, F_3, F_2, F_5,** F_1	**F_4, F_6, F_3, F_2, F_5**
R^2	Training dataset	0.78	**0.83**
	Testing dataset	0.65	**0.88**
MAE		1.87E-04	**6.56E-05**
RMSE		2.67E-04	**8.18E-05**

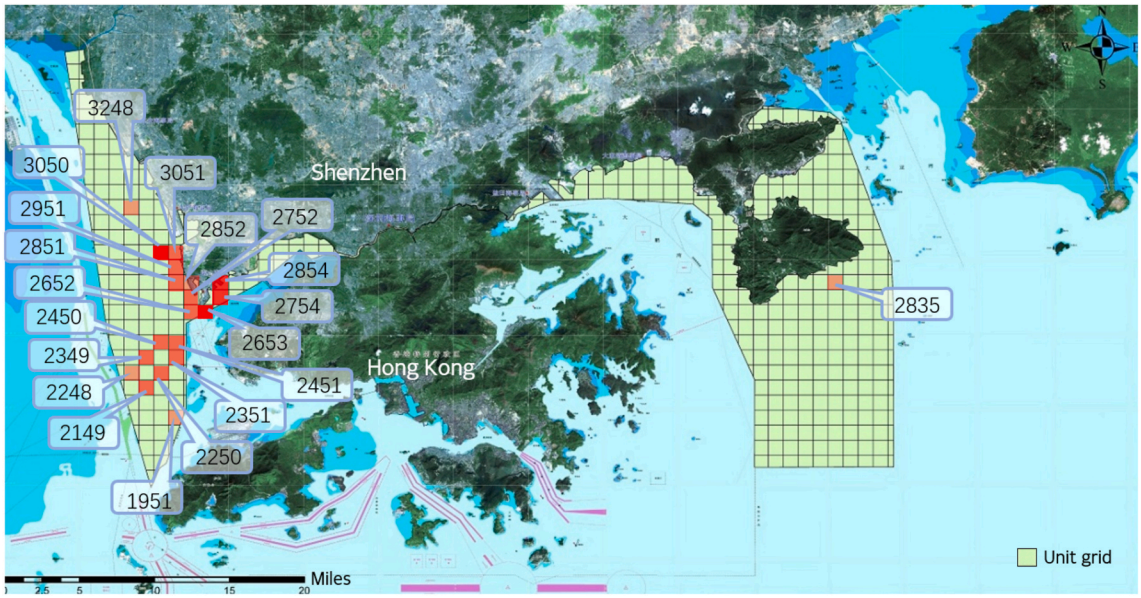

Fig. 12. Real-time risk result.

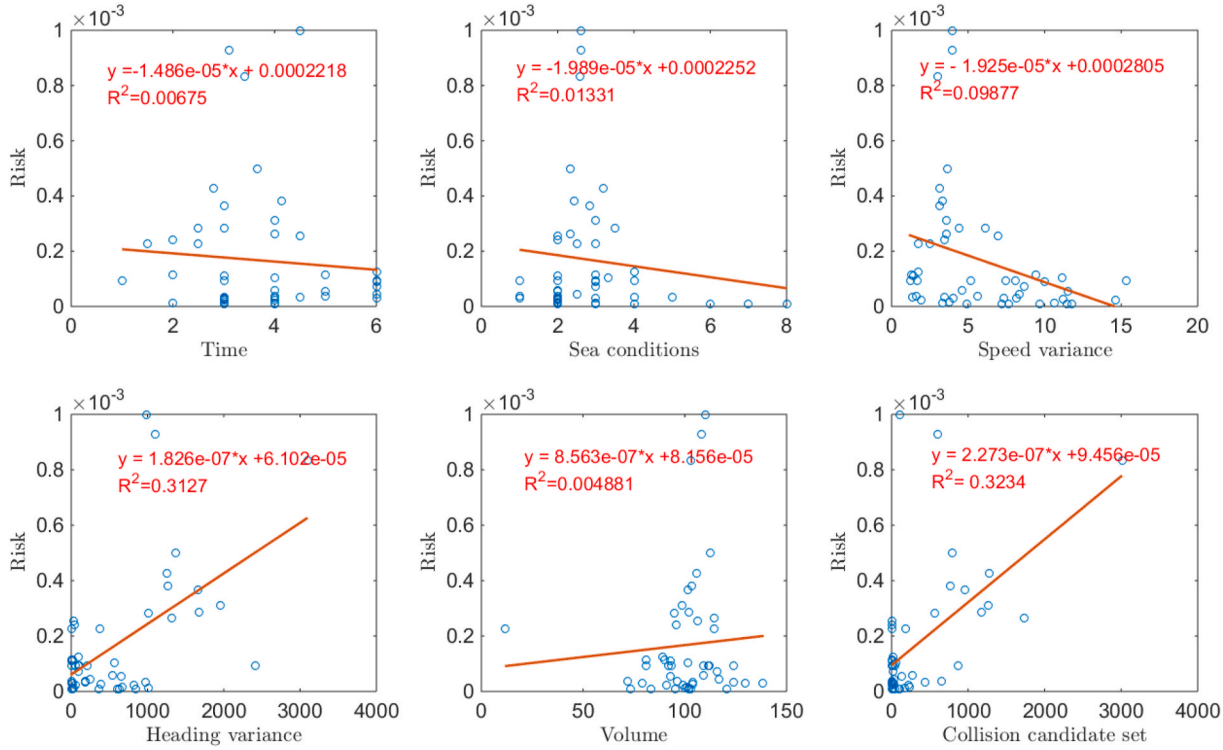

Fig. 13. Linear correlation analysis between risk and factors.

standard of this paper is consistent with that of maritime practice. The real-time risk calculated by this model can be directly used for maritime traffic safety management, which has a very important guiding role for regional maritime traffic management.

(3) The risk assessment model based on non-accident critical events can generally obtain the weight of each factor through the expert experience method, and then establish the corresponding risk model through the comprehensive evaluation of each factor. However, the weight of the expert experience method is subjective. This paper combines the advantages of the accident-based risk model and the non-accident critical events-based risk model and proposes a real-time risk modeling method. Compared with the accident-based risk model, this model makes up for the defects of insufficient accident numbers and inaccurate accident data in the accident-based risk model. Compared with the risk model based on non-accident critical events, this model makes up for the lack of verifiability and subjectivity of the non-accident critical events risk model.

6.3. Limitations of the model

As the time span of accident data is from 2002 to 2017, the corresponding AIS data collection is difficult. Although AIS data collection is easier at this stage than in previous years, the collection of AIS data with the time span of ten years ago was not so easy. Due to the characteristic of the low frequency of maritime accidents, the accuracy of the risk assessment model could be low even when with the traffic data. In order to improve the accuracy of risk, to make the traffic data closer to the traffic at the time of the accident, we converted the traffic flow data to a certain extent. In the meantime, we have introduced the data on near-miss incidents, which is collision candidates into the risk modeling. With such an introduction, the accuracy of the output of the risk assessment model can be improved to some extent, as the risk here also refers to the encounters that have the potential for accident, not only the accident. However, from the analysis of the results, there are still some errors. But the purpose of this paper is to provide a way, combined with the two models, to provide a modeling method that can be used for real-time risk assessment. And the same technique has also been utilized in the research on traffic simulations. Furthermore, the objective of the risk analysis model we proposed in this research is to identify the area with higher risk and facilitate the decision-making of the MSA. From this perspective, we think the relative relationships between the risk values of each geogrids have a higher value than its absolute numerical values. In the future, we will collect enough AIS data when the accident occurs and establish a risk model for more accurate risk modeling by analyzing the traffic characteristics when the accident occurs.

When the risk model proposed in this paper is applied to different port waters, it needs to be trained according to the actual port data to make it suitable for different waters.

From the model training results, we can see that due to the limited number of data, the accuracy of the model still has room to improve. In addition, we can get the collision risk model and the contribution of each factor to the collision risk model through a random forest algorithm, but we can't get the complex relationship between each factor.

On the other hand, this paper analyzes the relationship between six factors and collision risk. In fact, there are many factors that affect the collision risk. In the future, more risk-related factors can be taken into account, and more refined risk modeling can be provided.

6.4. Application

Mou et al. (2019) presented the safety index as a simple but effective method to evaluate and manage the safety status of vessel traffic in busy waterways and conducted risk analysis for vessel traffic transiting the western Shenzhen port as a case study. It is only based on the accident data with a span of 20 years to examine the actual risk level and the safety indexes. The indexes consist of Safety Evaluation Indexes (SEI) and Safety Warning Indexes (SWI). SEI work as a ruler to measure the safety status in last year and give the direct answer of 'safe' or 'unsafe', while the SWI can act as another safety threshold and provide early warnings for the risk control. Since 2005, the indexes have been widely implemented in safety management for vessel traffic control by the

Maritime Safety Administration of Shenzhen (MSA).

However, due to the inherent weakness of these indexes, they are only functioning after accidents but very weak to predict the real time collision risk inadvance or other this area. It is demanding to develop a real time collision risk model, which can be input to Vessel Traffic Services (VTS) or benefit basic situation aware for vessel traffic safety management in this area. The software provider of the VTS, Saab Technolgies (Hong Kong) Limited, welcomed to input the real time collision risk model to the system and financially supported the study of intelligent perception for ship collision risk in the Guangdong-Hong Kong-Macau Greater Bay Area.

7. Conclusion

Large hub ports and busy waterways are frequently visited by high densities of vessel traffic. Such a heavy and complicated maritime traffic situation has been continuously posing threat to the safe operation of the regional and global maritime transportation networks. As for one of the major stakeholders, it is of great significance for the maritime safety administrations to obtain insights on the real-time navigation risk characteristics in the area, to better perform navigational management and improve the safety level of the area.

To facilitate the task of real-time collision risk analysis in the busy ports and waterways, this paper proposed a grid-based collision risk identification and prediction model via integrating the historical accident data and maritime traffic data. The geographical grid, which is a GIS-based tool is utilized here as the key element to connect the accident and traffic data, based on which, the spatial-temporal characteristics and the accident contributing factors of a maritime accident, especially ship collision are analyzed. By applying the random forest tree, we have successfully established an accident risk prediction model based on the analysis of the accident data and the integration of maritime traffic data in the model. A case study focusing on one of the busiest ports in China-Shenzhen port was conducted, and the results compared with the historical accident data indicate that the model can effectively identify the region of high risk and their spatial-temporal characteristics. Such results show that the proposed method has important value in identifying and profiling the real-time collision risk in the regions integrating the historical information and also the traffic data, which can be an effective tool for maritime safety management in the interested areas.

In this research we have demonstrated the feasibility of utilizing the historical accident data and maritime traffic data at the same time to identify and predict the collision risk in real-time, using a geographical approach. It is extended our former study for vessel traffic management in Shenzhen Ports and can provide explicit display of collision risk in real time. In the meantime, more work could be conducted, to furtherly strengthen the link between the big data of vessel traffic, environmental contributors, and human factors, as well as to establish a more integrated and accurate risk model.

Author statement

Mengxia Li: Conceptualization, Methodology, Investigation, Formal analysis, Writing - Original Draft, **Junmin Mou:** Methodology, Resources, Supervision, **Pengfei Chen:** Conceptualization, Methodology, Supervision, Writing-Review & Editing, **Linying Chen:** Model, Methodology, Visualization, **P.H.A.J.M. van Gelder:** Supervision, Writing-Review & and Editing.

Declaration of competing interest

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

Data availability

The authors do not have permission to share data.

Acknowledgments

The work presented in this study is financially supported by National Natural Science Foundation of China (Grant Number: 52271367, 52101402, 52001242, 52071249).

Appendix I. Collision risks

The following table presents the collision risks based on the observed number of accidents. Consequences are determined according to Eqn. (5).

Table a
Collision risk

Grid Number	Year of accident	Consequence	Probability	Collision risk
C1449	2004	1.00	3.00E-05	3.00E-05
C1851	2003	1.00	9.33E-05	9.33E-05
C1951	2006	1.34	1.74E-05	2.32E-05
C2538	2004	3.35	3.00E-05	1.11E-04
	2008	1.00	1.09E-05	
C2652	2002	2.34	1.13E-04	2.85E-04
	2011	1.67	1.21E-05	
C2653	2003	1.34	9.33E-05	2.64E-04
	2003	1.34	9.33E-05	
	2010	1.00	1.46E-05	
C2739	2006	1.34	1.74 E−05	6.33 E−05
	2007	1.34	1.20E-05	
	2007	2.01	1.20E-05	
C2752	2002	1.00	1.13E-04	8.34E-04
	2002	1.00	1.13E-04	
	2002	1.00	1.13E-04	
	2003	1.00	9.33E-05	
	2003	1.67	9.33E-05	
	2004	1.00	3.00E-05	
	2004	3.35	3.00E-05	
	2005	1.00	2.89E-05	

(*continued on next page*)

Table a (*continued*)

Grid Number	Year of accident	Consequence	Probability	Collision risk
	2005	1.00	2.89E-05	
	2005	1.00	2.89E-05	
	2006	1.00	1.74E-05	
	2008	1.00	1.09E-05	
C2753	2002	1.67	1.13E-04	4.99E-04
	2002	1.00	1.13E-04	
	2003	1.34	9.33E-05	
	2005	1.00	2.89E-05	
	2006	1.00	1.74E-05	
	2010	1.82	1.46E-05	
C2754	2003	1.00	9.33E-05	9.33E-05
C2851	2003	1.00	9.33E-05	3.82E-04
	2003	1.00	9.33E-05	
	2003	1.00	9.33E-05	
	2004	1.00	3.00E-05	
	2004	1.00	3.00E-05	
	2004	1.00	3.00E-05	
	2007	1.00	1.20E-05	
C2852	2002	1.00	1.13E-04	4.27E-04
	2002	1.00	1.13E-04	
	2003	1.00	9.33E-05	
	2003	1.00	9.33E-05	
	2010	1.00	1.46E-05	
C2855	2002	1.00	1.13E-04	1.13E-04
C2856	2003	2.34	9.33E-05	2.42E-04
	2006	1.34	1.74E-05	
C2950	2012	2.65	9.80E-06	3.32E-05
	2014	1.00	7.27E-06	
C2951	2002	2.34	1.13E-04	3.66E-04
	2004	1.00	3.00E-05	
	2004	1.00	3.00E-05	
	2004	1.00	3.00E-05	
	2012	1.00	9.80E-06	
	2013	1.00	1.72E-06	
C3027	2003	2.65	9.33E-05	2.56E-04
	2016	2.34	3.85E-06	
C3050	2002	1.00	1.13E-04	3.10E-04
	2002	1.00	1.13E-04	
	2005	1.00	2.89E-05	
	2007	1.67	1.20E-05	
	2007	1.00	1.20E-05	
	2007	1.00	1.20E-05	
	2008	1.00	1.09E-05	
C3051	2002	1.34	1.13E-04	9.29E-04
	2002	1.00	1.13E-04	
	2003	1.00	9.33E-05	
	2003	1.00	9.33E-05	
	2003	1.00	9.33E-05	
	2003	2.34	9.33E-05	
	2003	1.00	9.33E-05	
	2005	1.00	2.89E-05	
	2010	2.01	1.46E-05	
	2017	1.67	8.77E-06	
C3061	2002	1.00	1.13E-04	2.26E-04
	2002	1.00	1.13E-04	
C3146	2010	1.82	1.46E-05	2.66E-05
C3149	2017	1.00	8.77E-06	8.77E-06
C3150	2002	1.00	1.13E-04	1.00E-03
	2002	1.00	1.13E-04	
	2002	1.00	1.13E-04	
	2002	1.00	1.13E-04	
	2003	1.00	9.33E-05	
	2003	1.00	9.33E-05	
	2003	1.00	9.33E-05	
	2003	1.82	9.33E-05	
	2003	1.00	9.33E-05	
	2016	1.00	3.85E-06	
C3151	2002	1.00	1.13E-04	2.26E-04
	2002	1.00	1.13E-04	
C3228	2005	1.00	2.89E-05	2.89E-05
C3246	2012	1.00	9.80E-06	9.80E-06
C3248	2012	1.34	9.80E-06	1.31E-05
C3249	2004	4.18	3.00E-05	1.25E-04
C3250	2011	2.65	1.21E-05	3.22E-05
C3252	2002	1.00	1.13E-04	2.26E-04
	2002	1.00	1.13E-04	

(*continued on next page*)

Table a (*continued*)

Grid Number	Year of accident	Consequence	Probability	Collision risk
C3315	2002	2.34	1.13E-04	2.82E-04
	2006	1.00	1.74E-05	
C3317	2011	3.16	1.21E-05	3.84E-05
C3327	2017	1.00	8.77E-06	8.77E-06
C3331	2017	1.82	8.77E-06	1.60 E−05
C3346	2017	1.00	8.77E-06	8.77E-06
C3348	2003	1.00	9.33E-05	9.33E-05
C3350	2004	1.00	3.00E-05	3.00E-05
C3415	2005	1.00	2.89E-05	5.78E-05
	2005	1.00	2.89E-05	
C3416	2014	2.01	7.27E-06	1.46E-05
C3425	2005	3.16	2.89E-05	9.14E-05
C3447	2014	1.00	7.27E-06	7.27E-06
C3523	2003	1.00	9.33E-05	9.33E-05
C3525	2005	2.49	2.89E-05	7.21E-05
C3545	2008	2.34	1.09E-05	2.55E-05
C3546	2006	1.34	1.74E-05	1.03E-04
	2010	1.34	1.46E-05	
	2011	5.00	1.21E-05	
C3623	2008	1.00	1.09E-05	4.50E-05
	2010	2.34	1.46E-05	
C3624	2014	1.34	7.27E-06	9.71E-06
C3645	2006	1.34	1.74E-05	5.42E-05
	2012	3.16	9.80E-06	
C3647	2002	1.00	1.13E-04	1.13E-04
C3746	2005	1.34	2.89E-05	3.86E-05
C3944	2011	1.00	1.21E-05	1.21E-05
C3946	2005	1.00	2.89E-05	2.89E-05
C4045	2011	1.82	1.21E-05	2.22E-05
C4145	2008	1.33	1.09E-05	1.46 E−05
C4244	2003	1.00	9.33E-05	9.33E-05
C4345	2003	1.00	9.33E-05	9.33E-05

References

Aven, T., 2012. The risk concept—historical and recent development trends. Reliab. Eng. Syst. Saf. 99, 33–44.

Breiman, L., 2000. Some Infinite Theory for Predictor Ensembles.

Bye, R.J., Aalberg, A.L., 2018. Maritime navigation accidents and risk indicators: an exploratory statistical analysis using AIS data and accident reports. Reliab. Eng. Syst. Saf. 176, 174–186.

Bye, R.J., Almklov, P.G., 2019. Normalization of maritime accident data using AIS. Mar. Pol. 109.

Chen, P., Huang, Y., Mou, J., van Gelder, P.H.A.J.M., 2018. Ship collision candidate detection method: a velocity obstacle approach. Ocean Eng. 170, 186–198.

Chen, P., Huang, Y., Mou, J., van Gelder, P.H.A.J.M., 2019a. Probabilistic risk analysis for ship-ship collision: state-of-the-art. Saf. Sci. 117, 108–122.

Chen, P., Mou, J., van Gelder, P.H.A.J.M., 2019b. Integration of individual encounter information into causation probability modelling of ship collision accidents. Saf. Sci. 120, 636–651.

China, M.o.t.o.t.p.s.R.o., 2015a. In: China, M.o.t.o.t.p.s.R.o (Ed.), Implementation Guide of Maritime Dynamic Supervision Grid. Beijing, China.

China, M.o.T.o.t.P.s.R.o., 2015b. In: China, S.C.o.t.P.s.R.o (Ed.), Statistics on Water Traffic Accidents. Peking, China.

Debnath, A., 2009. Traffic-conflict-based Modeling of Collision Risk in Port Waters.

Debnath, A.K., Chin, H.C., 2015. Modelling collision potentials in port anchorages: application of the navigational traffic conflict technique (NTCT). J. Navig. 69 (1), 183–196.

Dehnad, Khosrow, 2012. Density estimation for statistics and data analysis. Technometrics 29 (4), 495-495.

Du, L., Goerlandt, F., Kujala, P., 2020. Review and analysis of methods for assessing maritime waterway risk based on non-accident critical events detected from AIS data. Reliab. Eng. Syst. Saf. 200.

Du, L., Valdez Banda, O.A., Kujala, P., 2019. An Intelligent Method for Real-Time Ship Collision Risk Assessment and Visualization, Developments in the Collision and Grounding of Ships and Offshore Structures, pp. 293–300.

Gan, L., Yan, Z., Zhang, L., Liu, K., Zheng, Y., Zhou, C., Shu, Y., 2022. Ship path planning based on safety potential field in inland rivers. Ocean Eng. 260.

Goerlandt, F., Kujala, P., 2014. On the reliability and validity of ship–ship collision risk analysis in light of different perspectives on risk. Saf. Sci. 62, 348–365.

Goerlandt, F., Montewka, J., 2015. Maritime transportation risk analysis: review and analysis in light of some foundational issues. Reliab. Eng. Syst. Saf. 138, 115–134.

Guo, S., Mou, J., Chen, L., Chen, P., 2021. Improved kinematic interpolation for AIS trajectory reconstruction. Ocean Eng. 234.

Hänninen, M., Kujala, P., 2012. Influences of variables on ship collision probability in a Bayesian belief network model. Reliability Engineering [?]. System Safety 102, 27–40.

Hassel, M., Aalberg, A., Nordkvist, H., 2019. An Advanced Method for Detecting Exceptional Vessel Encounters in Open Waters from High Resolution AIS Data.

Jackson, M.C., Huang, L., Xie, Q., Tiwari, R.C., 2010. A modified version of Moran's I. Int. J. Health Geogr. 9, 33.

Jin, D., Kite-Powell, H., Thunberg, E., Solow, A., Talley, W., 2002. A model of fishing vessel accident probability. J. Saf. Res. 33, 497–510.

Kim, K.I., Jeong, J.S., 2016. Visualization of ship collision risk based on near-miss accidents. In: 2016 JOINT 8TH INTERNATIONAL CONFERENCE ON SOFT COMPUTING AND INTELLIGENT SYSTEMS (SCIS) AND 17TH INTERNATIONAL SYMPOSIUM ON ADVANCED INTELLIGENT SYSTEMS (ISIS), pp. 323–327.

Köse, E., Dinçer, A.C., Durukanoğlu, H.F., 1998. Risk assessment of fishing vessels 22, 417–427.

Kujala, P., Hänninen, M., Arola, T., Ylitalo, J., 2009. Analysis of the marine traffic safety in the Gulf of Finland. Reliab. Eng. Syst. Saf. 94 (8), 1349–1357.

Kulkarni, K., Goerlandt, F., Li, J., Banda, O.V., Kujala, P., 2020. Preventing shipping accidents: past, present, and future of waterway risk management with Baltic Sea focus. Saf. Sci. 129.

Kum, S., Sahin, B., 2015. A root cause analysis for Arctic Marine accidents from 1993 to 2011. Saf. Sci. 74, 206–220.

Lei, P.-R., 2019. Mining maritime traffic conflict trajectories from a massive AIS data. Knowl. Inf. Syst. 62 (1), 259–285.

Li, M., Mou, J., He, Y., Chen, L., Huang, Y., 2021. A rule-aware time-varying conflict risk measure for MASS considering maritime practice. Reliab. Eng. Syst. Saf. 215, 107816.

Li, M., Mou, J., Liu, R., Chen, P., Dong, Z., He, Y., 2019a. Relational model of accidents and vessel traffic using AIS data and GIS: a case study of the western port of shenzhen City. J. Mar. Sci. Eng. 7 (6).

Li, S., Meng, Q., Qu, X., 2012. An overview of maritime waterway quantitative risk assessment models. Risk Anal. 32 (3), 496–512.

Li, Y.P., Liu, Z.J., Kai, J.S., 2019b. Study on complexity model and clustering method of ship to ship encountering isk. J. Mar. Sci. Technol.-Taiwan 27 (2), 153–160.

Lim, G.J., Cho, J., Bora, S., Biobaku, T., Parsaei, H., 2018. Models and computational algorithms for maritime risk analysis: a review. Ann. Oper. Res. 271 (2), 765–786.

Liu, K., Yu, Q., Yuan, Z., Yang, Z., Shu, Y., 2021. A systematic analysis for maritime accidents causation in Chinese coastal waters using machine learning approaches. Ocean Coast Manag. 213.

Luter, H., Silverman, B.W., 2010. Density Estimation for Statistics and Data Analysis. Chapman & Hall, London – New York, p. 175, 1986, 12.—. Biometrical Journal 30 (7).

Montewka, J., Ehlers, S., Goerlandt, F., Hinz, T., Kujala, P., 2012. A Model for Risk Analysis of RoPax Ships -the Gulf of Finland Case.

Moran, P.A., 1950. Notes on continuous stochastic phenomena. Biometrika 37 (1–2), 17–23.
Mou, J.M., Chen, P.F., He, Y.X., Yip, T.L., Li, W.H., Tang, J., Zhang, H.Z., 2019. Vessel traffic safety in busy waterways: a case study of accidents in western shenzhen port. Accid. Anal. Prev. 123, 461–468.
Mou, J.M., Tak, C.v.d., Ligteringen, H., 2010. Study on collision avoidance in busy waterways by using AIS data. Ocean Eng. 37 (5–6), 483–490.
Nielsen, D., Jungnickel, D., 2003. Maritime accident investigation and temporal determinants of maritime accidents: a case study. WMU J. Marit Affairs 2 (1), 49–59.
Owens, E.H., 1984. SEA CONDITIONSDouglas scale; Peterson scale; Sea stateSea conditions. In: Schwartz, M. (Ed.), Beaches and Coastal Geology. Springer US, New York, NY, 722-722.
Qu, X., Meng, Q., Suyi, L., 2011. Ship collision risk assessment for the Singapore Strait. Accid. Anal. Prev. 43 (6), 2030–2036.
Rawson, A., Brito, M., 2021. A critique of the use of domain analysis for spatial collision risk assessment. Ocean Eng. 219.
Rezaee, S., Pelot, R., Finnis, J., 2016. The effect of extratropical cyclone weather conditions on fishing vessel incidents' severity level in Atlantic Canada. Saf. Sci. 85, 33–40.
Rong, H., Teixeira, A.P., Guedes Soares, C., 2021. Spatial correlation analysis of near ship collision hotspots with local maritime traffic characteristics. Reliab. Eng. Syst. Saf. 209.
Shahrabi, J., 2004. Spatial and Temporal Analyses of Maritime Fishing and Shipping Traffic and Incidents. Dalhousie University.
Szłapczyński, R., Niksa-Rynkiewicz, T., 2018. A framework of A ship domain-based near-miss detection method using Mamdani Neuro-Fuzzy classification. Pol. Marit. Res. 25 (s1), 14–21.
Uğurlu, Ö., Köse, E., Yıldırım, U., Yüksekyıldız, E., 2013. Marine accident analysis for collision and grounding in oil tanker using FTA method. Marit. Pol. Manag. 42 (2), 163–185.
Watawana, T., Caldera, A., 2018. Analyse Near Collision Situations of Ships Using Automatic Identification System Dataset.
Weng, J., Meng, Q., Qu, X., 2012. Vessel collision frequency estimation in the Singapore strait. J. Navig. 65 (2), 207–221.
Weng, J., Xue, S., 2015. Ship collision frequency estimation in port fairways: a case study. J. Navig. 68 (3), 602–618.
Wu, B., Tian, H., Yan, X., Guedes Soares, C., 2019. A probabilistic consequence estimation model for collision accidents in the downstream of Yangtze River using Bayesian Networks. Proc. Inst. Mech. Eng. O J. Risk Reliab. 234 (2), 422–436.
Wu, X., Mehta, A.L., Zaloom, V.A., Craig, B.N., 2016. Analysis of waterway transportation in Southeast Texas waterway based on AIS data. Ocean Eng. 121, 196–209.
Yip, T.L., 2008. Port traffic risks – a study of accidents in Hong Kong waters. Transport. Res. E Logist. Transport. Rev. 44 (5), 921–931.
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.
Zhang, L., Meng, Q., 2019. Probabilistic ship domain with applications to ship collision risk assessment. Ocean Eng. 186, 106130.
Zhang, L.M., Qiang, Fang Fwa, Tien, 2019. Big AIS data based spatial-temporal analyses of ship traffic in Singapore port waters. Transport. Res. Part E 129, 287–304.
Zhang, W., Goerlandt, F., Montewka, J., Kujala, P., 2015. A method for detecting possible near miss ship collisions from AIS data. Ocean Eng. 107, 60–69.
Zhang, W., Kopca, C., Tang, J., Ma, D., Wang, Y., 2017. A systematic approach for collision risk analysis based on AIS data. J. Navig. 70, 1–16.
Zhang, Y., Sun, X., Chen, J., Cheng, C., 2021. Spatial patterns and characteristics of global maritime accidents. Reliab. Eng. Syst. Saf. 206.
Zheng, S.Y., Hui, X.U., Wang, H.C., 2005. A survey of grid and grid management. Syst. Eng.