Research on Time Characteristics of Near Miss in Bohai Sea

YANGYU ZHOU1,2, JIAXUAN YANG1,2, YUAN LIU1,2, AND JIAGUO LIU3

1 Navigation College, Dalian Maritime University, Dalian 116026, China
2 Key Laboratory of Navigation Safety Guarantee of Liaoning Province, Dalian Maritime University, Dalian 116026, China
3 Shipping Economics and Management College, Dalian Maritime University, Dalian 116026, China

Corresponding author: Jiaxuan Yang (yangjiaxuan@dlmu.edu.cn)

ABSTRACT It is an effective method to analyze the maritime traffic risk using near miss model, but there are no works from the temporal perspective. The duration of near miss shows the maritime traffic risk. The longer the duration is, the higher the risk is. The influence of the ship maneuvering capability, speed, size, and Convention on the International Regulations for Preventing Collisions at Sea (COLREGS) on the safety of ship navigation are considered in this article, the navigation safety domain is used to detect near misses, and the duration of a single near miss is defined, which is the time from target ship sailing into the navigation safety domain of own ship to leaving, and then an algorithm for calculating duration of near misses is proposed. To measure the similarity of two ships’ lengths, the similarity coefficient model of ships’ lengths is defined. Based on the $k$-means algorithm, different parameters $k$ are selected to cluster the duration of near misses. According to the clustering results, the connections between the duration of near misses and the ship type, length, and speed were analyzed to identify a series of time characteristics of near misses, that is high frequencies of short-duration near misses and the strong connections between the long-duration near misses, ship length, and ship speed. The method is first proposed for doing research of near misses in temporal dimension. The findings of this article help researchers and maritime traffic management agencies to better understand the importance of maritime traffic risk from the temporal perspective.

INDEX TERMS Maritime traffic risk, near miss, AIS, navigation safety domain, duration of near miss.

I. INTRODUCTION

With the rapid development of modern economy, water transport has become increasingly busy and the occurrence of ship collision has become more and more frequent [1], [2]. The safe navigation of ships is of high societal concern[3]. It is difficult to evaluate the overall risk of waters based on actual ship collisions because of few ship collision accidents[4]. In recent years, it is an effective method to evaluate the maritime traffic risk by replacing the actual ship collisions with near misses[3]. The greater the number of near misses is, the higher the risk of maritime traffic is.

Scholars have considered many factors and proposed a variety of conditions for detecting near misses. Berglund and Huttunen believed that distance to close point of approaching (DCPA) ≤ 0.3nm can be used as a detection condition for near misses [5]. Fukuto and Imazu regarded DCPA of 1 nm and time to close point of approaching (TCPA) of 5 min as the conditions for detecting near misses from the perspective of reducing the frequency of ship collision warnings and reducing the workload of pilots[6]. Park set DCPA ≤ 0.15nm and TCPA ≤ 3min as the thresholds for VTS to detect the safe speed [7]. Langard et al. found that the average DCPA of ships taking collision avoidance actions was 0.64nm, TCPA was 26min, and the distance between ships was 1.81nm [8]. Kim and Jeong defined the invasion of the ship domain as a detection condition for near misses, to distinguish between encountering and near miss [9]. Sang-Lok Yoo thought that DCPA of 0.1nm, TCPA of 3min, and ships distance of 0.3nm can be used as detection conditions for near misses, and then developed a charting system based on near miss density [4]. To find the critical encountering, Goerlandt et al. defined overlapping of ship domain and ship contour was near miss, and analyzed the risk of collision between ships in open waters of the Gulf of Finland [10]. Tan et al. believed that overlapping of ship domain and ship contour was not suitable for detecting the near misses of merchant ships and fishing ships in restricted waters, and then defined “shortest passing
distance” ≤ 500m and “time difference” ≤ 3min as the detection conditions [11]. Zhang et al. believed that DCPA and TCPA could not fully reflect the risks of encountering. The Vessel Conflict Ranking Operator (VCRO) model was built by using the distance between ships, relative speed and phase. The distance between ships of 1nm and 6nm was set as high-risk and low-risk standards respectively, and then the risks of Finnish waters were analyzed [12]–[14].

At present, most of the works of maritime traffic risk based on near misses are carried out from the spatial perspective, mainly focusing on detection model of near misses and analysis of the frequency and geographical distribution characteristics of near misses to evaluate the maritime traffic risk. However, there is no work focus on the maritime traffic risk based on near misses from temporal perspective.

Near miss is a critical situation that ships encounter each other [10], and the duration of near miss shows the characteristics of maritime traffic risk from temporal perspective. In this article, the navigation safety domain is used to detect near misses, and then the model for calculating duration of near misses is proposed, based on the processing for time synchronization of AIS data. Cargo ships, oil tankers, and passenger ships in Bohai Sea are selected as the research objects, and then the duration of near misses is counted. The k-means clustering model is used to cluster the duration of near misses. The results of clustering are analyzed, and then the correlations between the duration of near misses, the ship type, length, and speed are obtained.

The main contributions of this article are as follows:

1) The influence of the ship maneuvering capability, speed, size, and COLREGS on the safety of ship navigation are considered in this article, and then the navigation safety domain is used to detect near misses, and the model for detecting near misses based on navigation safety domain is proposed.

2) The absolute difference of two ships’ lengths cannot fully show the similarity of those, the similarity coefficient model of ships’ lengths is defined, which could better measure the similarity of two ships’ lengths.

3) A method for studying maritime traffic risk from temporal perspective is first provided. The duration of a single near miss is defined, which is the time from target ship sailing into the navigation safety domain of own ship to leaving, and then the model for calculating duration of near miss is proposed.

4) Findings, the high frequencies of short-duration near misses and the strong connections between long-duration near misses, ship length, and ship speed, could help officers to detect the high potential risk encountering.

The rest of the paper is organized as follows. Section II outlines the conceptual basis of AIS data and k-means clustering algorithm, and presents the model for detecting near misses based on the navigation safety domain, the model for calculating duration of near miss, the similarity coefficient model of ships’ lengths. Section III analyzes the connections between the duration of near misses and the ship type, length, and speed. Section IV gives the conclusion.

| TABLE 1. The Types of AIS Data Used in this Article. |
|-----------------------------------------------------|
| Data Type | Accuracy | Unit |
| Maritime Mobile Service Identity | - | - |
| Timestamp: UTC time | 1 | S |
| Longitude | 1/100000 | degree |
| Latitude | 1/100000 | degree |
| Heading | 1/10 | degree |
| Ship type | - | - |
| Ship length | 1 | M |

| TABLE 2. Sample Rate of Variable Data From AIS [18]. |
|-----------------------------------------------------|
| Manoeuvring Situation | Sample Rate |
| At Anchor | 3min |
| Speed 0-14 knots | 12s |
| Speed 0-14 knots and changing course | 4s |
| Speed 14-23 knots | 6s |
| Speed 14-23 knots and changing course | 2s |
| Speed >23 knots | 3s |
| Speed >23 knots and changing course | 2s |

II. BACKGROUND KNOWLEDGE

A. AIS DATA

The determination of near misses is based on an extensive analysis of data obtained from the Automatic Identification System (AIS). The system transmits certain data about the navigational status of the ship to other ships and onshore receiving facilities continuously and automatically [10], [15]. The application of AIS is meaningful in protecting the marine environment, ensuring the safety of life at sea, improving the safety of navigation, and developing the shipping industry efficiently [16], [17]. The types of AIS data used by this article are summarized in Table 1.

There are various types of error in AIS data, such as data corruption, faulty position reports and erroneous Maritime Mobile Service Identity (MMSI) numbers [18], [19] because of sensor signal misalignment, AIS equipment fault, etc. These errors have been filtered out before analyzing the data. The system is self-organized by a time division multiplexing algorithm and provides updates to variable data at rates depending on the ship speed and manoeuvre situation [20]. The rate of data transmission is seen in Table 2. Specifically, the frequency with which AIS data is transmitted depends on whether a ship is underway (intervals of 2–12s depending on speed) or at anchor (intervals of 3min).

Goerlandt believed that for a given ship, the sample rate of about 5 minutes is enough to roughly estimate the number of near misses [10]. However, the sample rate of about 5 minutes is not sufficient for the detailed study of the duration of near misses. Generally, the higher the sample rate is, the more accurate the ship motion characteristics are. This necessitated the interpolation of raw AIS data according to a sample rate interval of 1min.

In practice, the data of ships’ trajectories may not be received synchronously. The AIS trajectory sequence needs a fixed sample rate to detect the near misses, so the
time-synchronization interpolation is needed to obtain the time synchronous and continuous ship trajectory data. It is necessary to segment the ship trajectory to ensure the correctness of the interpolation. The trajectory will be separated between the two points if the time interval between those two points is too long. The subtrajectory data after segmentation will be deleted when it is too little to provide important information [21].

It is necessary to evaluate whether or not the trajectories of the two considered ships occur in an overlapping timeframe because the two selected ships may sail in same water area in a different period of time. Fig.1 is the data chart of time points of two ships’ trajectories. The number of time points in different period of time. The trajectories of the two ships do not occur in an overlapping timeframe, the first time point is $t_1$ and the last time point is $t_2$; $t_1 > t_2$ represents the trajectories of the two ships do not occur in an overlapping timeframe.

B. THE MODEL FOR DETECTING NEAR MISSES BASED ON THE NAVIGATION SAFETY DOMAIN

As described in the introduction of this article, scholars have considered many factors to detect near misses, such as DCPA and TCPA, but DCPA and TCPA do not fully reflect the severity level of encountering [12]. The distance between the two ships is also an important factor to detect near misses, but there are no conclusions about the relationship between near misses and the distance between the two ships. Therefore, of simultaneous sailing between two ships respectively. The start time and end time are calculated by

$$T_1 = \begin{cases} t_{j-start} & t_{i-start} \geq t_{j-start} \\ t_{i-start} & t_{i-start} < t_{j-start} \end{cases}$$

$$T_2 = \begin{cases} t_{i-end} & t_{i-end} \geq t_{j-end} \\ t_{j-end} & t_{i-end} < t_{j-end} \end{cases}$$

where $T_1 \leq T_2$ represents the trajectories of the two ships occur in an overlapping timeframe, the first time point is $T_1$ and the last time point is $T_2$; $T_1 > T_2$ represents the trajectories of the two ships do not occur in an overlapping timeframe.

$$QSD_k = \{(x, y) | f_k (x, y; Q) \leq 1, Q = \{R_{fore}, R_{aft}, R_{starb}, R_{port}\}, k \geq 1\}$$

$$f_k (x, y; Q) = \left(\frac{2x}{(1 + sgnx) R_{fore} - (1 - sgnx) R_{aft}}\right)^k + \left(\frac{2y}{(1 + sgny) R_{starb} - (1 - sgny) R_{port}}\right)^k$$

$$sgnx = \begin{cases} 1 & x \geq 0 \\ -1 & x < 0 \end{cases}$$

$$sgny = \begin{cases} 1 & y \geq 0 \\ -1 & y < 0 \end{cases}$$

$$R_{fore} = \left(1 + 1.344\sqrt{k_{AD}^2 + (k_{DT}/2)^2}\right)L$$

$$R_{aft} = \left(1 + 0.676\sqrt{k_{AD}^2 + (k_{DT}/2)^2}\right)L$$

$$R_{starb} = (0.2 + k_{DT})L$$

$$R_{port} = (0.2 + 0.75k_{DT})L$$

$$k_{AD} = A_D/L = 10^{0.3591lgV_{own}+0.0952}$$

$$k_{DT} = D_T/L = 10^{0.5441lgV_{own}-0.0795}$$
the relationship between the distance and near miss is further analyzed.

Ship domain is the area around ship where operators want to avoid other ships for safety reasons. Ship domain is a special form of the safe distance, because the distance between ship and domain boundary in various bearing is different. Since Fujii introduced the concept of ship domain, researchers have been making statistical analysis on the distances between ships in scene of encountering, and calculating the size of domain according to expert experience [22]–[25]. At present, concept of ship domain has been used in maritime research widely, including recent applications to various traffic engineering problems: waterway capacity analysis [24], waterway collision risk analyses [27]–[30], near miss detection [13], [31]–[34], ship traffic simulation development [35], and ship collision avoidance analysis [36], [37].

Zhang used AIS data-driven approach to discover that there is an area, named as probabilistic ship domain [22], around own ship that no ship enters. It should be noted that the size of probabilistic ship domain is basically similar to that of Fujii ship domain, which proves that it is reliable to use the ship domain to detect near misses. The sizes of domains vary quite significantly [38], and the application of different domains to detect near misses affects the detection results. The probabilistic domain seems preferable in terms of safety analysis of marine traffic, but is heavily dependent on the set of AIS data and still under development. Fujii ship domain is small, which can be used to detect the most critical scene of encountering effectively. However, a number of comments should be made in the use of this domain [10]:

(1) The domain is symmetric, which implies that the possible influence of COLREGS is not considered.

(2) It is unrealistic that passing behind the stern is considered as dangerous as passing in front of the bow as the result of symmetry.

(3) The largest ship has the largest domain, which means the encountering is classified as danger for the largest ship, whereas safety for the smallest ship.

Considering the influence of the ship maneuvering capability, speed, size, and COLREGS on ship safety, Wang proposed a navigation safety domain named quaternion ship domain (QSD) [39], shown in Fig.2. The boundary functions are described as Eqs (3) - (7), as shown at the bottom of the previous page. The power \( k \) determines the shape of the domain while the \( \{R_{\text{fore}}, R_{\text{aft}}, R_{\text{starb}}, R_{\text{port}}\} \) identifies the ship domain size. To facilitate the study, \( k \) is usually set to 2 [40]. The navigation safety domain is larger than the Fujii ship domain, but the encountering is still so critical that it is worthy of the operators’ attention when target ship is sailing into the navigation safety domain of own ship. The comments (1) and (2) can be solved effectively by using the navigation safety domain to detect near misses.

where \( R_{\text{fore}} \) and \( R_{\text{aft}} \) are the longitudinal radius of the QSD in fore and aft directions respectively; \( R_{\text{port}} \) and \( R_{\text{starb}} \) are the lateral radius of the QSD in starboard and port directions respectively; \( L \) is the own ship length, \( k_{AD} \) and \( k_{DT} \) are the gains of the advance \( A_D \) and the tactical diameter \( D_T \) respectively; \( V_{\text{own}} \) is the own ship speed represented in knots.

Position relationship between target ship and own ship is represented in Fig.3, and the model for detecting near misses is described using Eqs (9) - (13), where \( OS \) is own ship; \( TS \) is target ship; \( \theta \) is the bearing of target ship relative to own ship; \( \beta \) is heading of own ship; \( \lambda_1 \) is the longitude of own ship; \( \phi_1 \) is the latitude of own ship; \( \lambda_2 \) is the longitude of target ship; \( \phi_2 \) is the latitude of target ship; \( L \) is own ship length; \( r \) is the distance from the center of own ship to the boundary of the domain. \( f \leq 0 \) represents own ship is involved in near miss, and \( f > 0 \) represents own ship is not in near miss.

\[
\theta = \arccos \left( \frac{\phi_2 - \phi_1}{\sqrt{(\lambda_2 - \lambda_1)^2 + (\phi_2 - \phi_1)^2}} \right) - \beta
\] (9)
Y. Zhou et al.: Research on Time Characteristics of Near Miss in Bohai Sea

FIGURE 4. The encountering formed between ship A and ship B.

\[
r = \sqrt{1 + \tan^2 \theta + \tan^2 \phi} \left( \frac{R_{\text{fore}} - (1 - \alpha) \beta}{2} \right)^2 + \left( \frac{R_{\text{after}} - (1 - \alpha) \beta}{2} \right)^2.
\]  
(10)

\[
d = \sqrt{(\lambda_2 - \lambda_1)^2 + (\varphi_2 - \varphi_1)^2}
\]  
(11)

\[
f = d - r
\]  
(12)

\[
\begin{cases}
  f \leq 0 & \text{near miss} \\
  f > 0 & \text{no near miss}
\end{cases}
\]  
(13)

In the scene of encountering, one ship is usually taken as the research object, which is defined as own ship, to analyze the relationship between other ships. For the two ships involved in the same encountering, either ship can be selected as own ship, and the other ship is target ship. The encountering formed between ship A and ship B is shown in Fig 4. The length of ship A is shorter than that of ship B. In Fig 4 (a), the scene is named as A-B scene if A is selected as own ship. In this scene, the position of B is not in the domain of ship A because length of ship A is short, so ship A is not in near miss. In Fig 4 (b), the scene is named as B-A scene if ship B is selected as own ship. In this scene, the position of A is in the domain of ship B because of ship B’s large domain, so ship B is in near miss.

According to the above-mentioned example, the scene of near miss with two ships encountering should be divided into four types: both A and B are in near miss, only A is in near miss, only B is in near miss, neither A nor B is in near miss. The status of A and B are shown in Table 3. The flow diagram of the model for detecting near misses is shown in Fig 5.

The data structure is represented in Eqs. (14) and (15), as shown at the bottom of the next page, where Dataset \((i, j)\) is the data set of the scenes of near misses which the \(i\)th trajectory and the \(j\)th trajectory formed, \(\text{traj}(i, j, t)\) is the data of the \(i\)th trajectory and \(j\)th trajectory at time \(t\).

C. THE ALGORITHM FOR CALCULATING DURATION OF NEAR MISSES

In this article, the process of a single near miss is defined, which is the process from target ship sailing into the ship domain of own ship to leaving. The duration of the process is defined as the duration of a single near miss. This article presents the algorithm for calculating duration of near miss, in order to get the duration of near misses happened between two selected ships. Algorithm 1 presents the algorithm’s corpus. Because the sample rate is 1 minute in this article, 1 minute is also selected as the time threshold.

D. SIMILARITY COEFFICIENT MODEL OF SHIPS’ LENGTHS

Due to the complex patterns of maritime traffic, the absolute length difference between own ship and target ship cannot fully show the similarity of two ships’ lengths [41]. Therefore, the similarity coefficient model of ships’ lengths is defined as

\[
L_{\text{diff}} = L_{\text{OS}} - L_{\text{TS}}
\]  
(16)

\[
L_{\text{sim}} = \frac{L_{\text{diff}}}{L_{\text{OS}}}
\]  
(17)

where \(L_{\text{OS}}\) is the own ship length, \(L_{\text{TS}}\) is the target ship length, \(L_{\text{diff}}\) is the absolute difference of the own ship length and the target ship length, \(L_{\text{sim}}\) is the similarity coefficient of the target ship length and the own ship length. \(L_{\text{sim}} > 0\) represents the own ship length is longer than the target ship length.
Algorithm 1 Algorithm for Calculating Duration of Near Misses

Input: The dataset \( \{\text{traj}_1, \text{traj}_2, \text{traj}_3, \cdots, \text{traj}_n\} \) of near misses happened between two selected ships, which of traj\(_n\) consisting two selected ships’ dynamic information at the time point \( t \) of near miss.

Output: A list of Duration of near miss

1. \( n \leftarrow \text{length of dataset} \)
2. \( k \leftarrow 1 \)
3. Duration of near miss \( \leftarrow [ ] \)
4. while \( k < n \) do
5.     if \( k == 1 \) then
6.         \( t_{\text{start}} \leftarrow \text{traj}_k.t \)
7.     else
8.         if \( \text{traj}_{k+1}.t - \text{traj}_k.t \leq \Delta t \)
9.             \( k \leftarrow k + 1 \)
10.        \( \text{if } k == n \) then
11.            \( t_{\text{end}} \leftarrow \text{traj}_k.t \)
12.        \( T \leftarrow t_{\text{end}} - t_{\text{start}} \)
13.        \( \text{Duration of near miss.append}(T) \)
14.     else
15.        \( t_{\text{end}} \leftarrow \text{traj}_k.t \)
16.        \( T \leftarrow t_{\text{end}} - t_{\text{start}} \)
17.        \( \text{Duration of near miss.append}(T) \)
18.    \( k \leftarrow k + 1 \)
19. \( t_{\text{start}} \leftarrow \text{traj}_{k+1}.t \)
20. end while
21. return Duration of near miss

length. \( L_{\text{sim}} < 0 \) represents the own ship length is shorter than the target ship length. \( L_{\text{sim}} = 0 \) represents the own ship length is equal to the target ship length. The closer \( L_{\text{sim}} \) is to 0, the more similar the lengths of the two ships.

Given two groups of data, one with a length of own ship is 400 m and target ship is 350 m, and the other with 100 m and 50 m, respectively, \( L_{\text{diff}} \) will have the same value, which is 50 m. In fact, the first group is more similar in length. In the above-mentioned example, \( L_{\text{sim}} \) has values of 0.125 and 0.33, respectively. The closer the value of \( L_{\text{sim}} \) is to 0, the more the similarity of the ships’ lengths.

Algorithm 1: Algorithm for Calculating Duration of Near Misses

Input: The dataset \( \{\text{traj}_1, \text{traj}_2, \text{traj}_3, \cdots, \text{traj}_n\} \) of near misses happened between two selected ships, which of traj\(_n\) consisting two selected ships’ dynamic information at the time point \( t \) of near miss. Time Threshold \( \Delta t \)

Output: A list of Duration of near miss

1. \( n \leftarrow \text{length of dataset} \)
2. \( k \leftarrow 1 \)
3. Duration of near miss \( \leftarrow [ ] \)
4. while \( k < n \) do
5.     if \( k == 1 \) then
6.         \( t_{\text{start}} \leftarrow \text{traj}_k.t \)
7.     else
8.         if \( \text{traj}_{k+1}.t - \text{traj}_k.t \leq \Delta t \)
9.             \( k \leftarrow k + 1 \)
10.        \( \text{if } k == n \) then
11.            \( t_{\text{end}} \leftarrow \text{traj}_k.t \)
12.        \( T \leftarrow t_{\text{end}} - t_{\text{start}} \)
13.        \( \text{Duration of near miss.append}(T) \)
14.     else
15.        \( t_{\text{end}} \leftarrow \text{traj}_k.t \)
16.        \( T \leftarrow t_{\text{end}} - t_{\text{start}} \)
17.        \( \text{Duration of near miss.append}(T) \)
18.    \( k \leftarrow k + 1 \)
19. \( t_{\text{start}} \leftarrow \text{traj}_{k+1}.t \)
20. end while
21. return Duration of near miss

E. K-MEANS CLUSTERING ALGORITHM

k-means clustering is a classic clustering algorithm, which is widely used in data analysis because of its good clustering effect. The mathematical function model is shown as follow:

\[
J = \min \left( \sum_{j=1}^{a} \sum_{p=1}^{k} \left( \|x_j - c_p\|^2 \right) \right)
\]

where \( \|x_j - c_p\|^2 \) is the squared Euclidean distance between \( p \)th cluster centroid and \( j \)th data point; \( j = 1, 2, \cdots, n; p = 1, 2, \cdots, k; n \) is the total number of data points and \( k \) is the number of clusters. The aim of \( k \)-means algorithm is to minimize the objective function \( J \).

In this article, the \( k \)-means clustering algorithm is used to cluster the duration of near misses, and then the relationship between the duration of near misses and the similarity coefficient of ships’ lengths is analyzed. The data set \( X = \{\text{Duration}_1, \text{Duration}_2, \cdots, \text{Duration}_{n-1}, \text{Duration}_n\} \) is the duration of near misses, where the number of near misses is \( n \). The \( k \)-means clustering can be divided into the following steps:

Step1: Randomly select \( k \) initial cluster center sample points \( \{c_1, c_2, \cdots, c_k\} \) from the data set \( X \);

Step2: Cluster the data set \( X \) according to the clustering center \( \{c_1, c_2, \cdots, c_k\} \) to obtain \( k \) clusters \( \{C_1, C_2, \cdots, C_k\} \), for any \( x_j \in X \), if \( \|x_j - c_{p1}\|^2 \leq \|x_j - c_{p2}\|^2, p1 \neq p2 \), \( p1, p2 = 1, 2, \cdots, k \), then \( x_j \) is classified as \( C_{p1} \);

Step3: Update the center of the adjustment cluster to get the new cluster center, which is recorded as \( c_{1}^*, c_{2}^*, \cdots, c_{k}^* \), \( c_p^* = \frac{1}{n_p} \sum_{j \in C_p} x_j \), where \( n_p \) is the number of samples contained in \( C_p \);

Step4: Judge whether the iteration termination condition is satisfied: the iteration stops when \( c_p^* = c_p, p = 1, 2, \cdots, k \), and the clustering result \( \{C_1, C_2, \cdots, C_k\} \) and cluster center \( \{c_1, c_2, \cdots, c_k\} \) are output when the value of clustering error is minimum; otherwise, Step2 is returned.

k-means clustering algorithm requires determining the optimal parameter \( k \) artificially. However, it is difficult to determine the appropriate value of \( k \), when the dataset is large or there is short of prior knowledge. Elbow method is effective to assess the validity of clustering results, which focuses on the percentage of variants as the function of the number of clusters. There should be an optimal value of

\[
\text{Dataset } (i, j) = \text{traj}(i, j, T_1), \text{traj}(i, j, T_1 + 1), \cdots, \text{traj}(i, j, T_2 - 1), \text{traj}(i, j, T_2)
\]

\[
\text{traj}(i, j, t) = \{\text{MMSI}_i, \text{MMSI}_j, \text{LAT}_i, \text{LON}_i, \text{heading}_i, \text{speed}_i, \text{LAT}_j, \text{LON}_j, \text{heading}_j, \text{speed}_j, \text{relative_speed}, t\}
\]
$k$-means algorithm, adding the value of $k$ will not contribute to the Sum Square Error (SSE) value significantly. The value of $k$ is added and the SSE value is recorded.

$$SSE = \sum_{j=1}^{n} \sum_{p=1}^{k} \left( \| x_j - c_p \|^2 \right)$$  \hspace{1cm} (19)

SSE is the sum of the average Euclidean Distance of each point against the centroid. We can try to find the optimal natural number of clusters in a data set by looking for the number of clusters at which there is a knee in the plot of the evaluation measure when it is plotted against the number of clusters. Starting from $k = 2$, the SSE value is increased with $k$. The largest difference between the SSE at $k$ and the SSE at $k - 1$ is the point in which the optimal $k$ value is selected. When the value of $k$ is re-added, the new cluster is similar to the previous cluster and the number of errors does not change significantly which resulted in the value of $k$.

### III. EXPERIMENTAL RESULTS AND DISCUSSIONS

#### A. TIME CHARACTERISTICS OF MARITIME TRAFFIC RISK BASED ON NEAR MISSES

To study the characteristics of maritime traffic risk based on near miss from temporal perspective, the expert questionnaire method was adopted in the process of data collection. 30 interviewees included engineers, professors engaged in the research of maritime traffic risk, captains, chief officers with practical collision avoidance experience and doctoral students with relevant research interests. After ensuring fully understand the purpose of this study, they were asked to fill in the questionnaire according to their knowledge. The maritime traffic risk is divided into 5 levels, and the higher the level is, the greater the risk is. The interviewees judged the level of maritime traffic risk according to the different duration of near misses, and the statistical result is shown in Fig 6. The number in grid represents the expert scoring results.

It can be seen that in 0-30 minutes, the longer the duration of near miss is, the greater the maritime traffic risk is; after 30 minutes, the maritime traffic risk decreases, but still remains at a high level. As a whole the long-duration near miss means high maritime traffic risk. Generally, the short-duration near miss means the distance of two ships is so far that the ships can escape danger through maneuvering quickly, whereas the long-duration near miss means two ships are in danger situation for a long time, and the operator must focus on the dynamics of the two ships to avoid collisions until the ships are out of danger, which increases the workload of the operators. The operators adapt to the new navigation state which involved in near miss lasting more than 30 minutes, so the risk level for operators is reduced, but it is still high at this time. Although the characteristics are easy to understand, there has not been a detailed quantitative analysis of the characteristics. The accurate research results are helpful to the collision avoidance decision of ships. Therefore, the researches of maritime traffic risk from temporal perspective are of high theoretical and practical value.

#### B. DATA SOURCE

The analyzed data is the period 01.04.2017 - 28.02.2017 in Bohai Bay waters (138°E-139°E, 38.15°N-39.05°N). Raw AIS data consists of millions of data points with information given in Table 4. Because the cooperative operations between merchant ships and engineering ships may be wrongly detected as near misses or collisions[4], engineering ships such as tugs are not considered in this article. Only cargo ships, oil tankers, and passenger ships are selected for this article. Finally, 16,419 trajectories are obtained, which include the information of 2,224 cargo ships, 333 oil tankers, and 14 passenger ships.

#### C. THE NEAR MISS FORMED BETWEEN TWO SHIPS

A statistical study is made of the duration of both two ships are in near miss and only one ship is in near miss in the study waters, and the results are shown in Table 5.
It can be seen from Table 5, the number of both two ships are in near miss is smaller, and the average duration is shorter. When both two ships are in near miss, the position of each ship is in the domain of the other, both ships have collision risk. Only one ship has collision risk when only one ship is in near miss. Therefore, the scenes of two ships are in near miss should have higher risk, and the number of near misses is smaller. High-risk scenes should be avoided, and the ships should leave the danger as soon as possible when the high-risk scenes are unavoidable. Therefore, the average duration of near misses is short when both two ships are in near miss.

D. CHARACTERISTICS OF DURATION OF NEAR MISSES BASED ON SHIP TYPE

Cargo ships, oil tankers, and passenger ships are selected as the research objects to count the number of ships which involved in near misses, and the result is shown in Table 6. It can be seen that the number of cargo ships is largest and that of the passenger ships is smallest. According to the different scenes, the number of near misses and the total duration of near misses are calculated, and then the average duration of near misses is also calculated. The results are shown in Table 7.

It can be seen from Table 7 that the longest average duration of near misses is 444.74s, which is formed between two cargo ships. Due to the number of cargo ships in study waters is largest, the number of near misses formed between cargo ships is the largest, which is 5419. Observing the scenes of cargo-cargo, cargo-tanker, and cargo-passenger, the average duration is longer than 350s when the type of own ship is the cargo ship. Observing the scenes of passenger-cargo and passenger-tanker, the average duration of near misses is shorter than 300s when the type of own ship is the passenger ship.

In the study waters, the two scenes with the smallest number of near misses are passenger-cargo and passenger-tanker, wherein the number of near misses is 14 and 2 respectively. Since the high safety requirements of the passenger ships, the operators must be very cautious and keep sufficient safety distance to ensure safety. Therefore, the number of near misses is the smallest and the duration is the shortest if own ship is the passenger ship.

Table 8 is the statistical table of near misses of cargo-cargo, cargo-tanker, tanker-cargo, cargo-passenger, tanker-passenger, passenger-cargo, and passenger-tanker. Observing the scenes of cargo-cargo, cargo-tanker, tanker-cargo, the number of long-duration near misses is smaller and the number of short-duration near misses is larger. Observing the scenes of cargo-passenger, tanker-tanker, passenger-cargo, and passenger-tanker, the numbers of near misses are all small and the duration of each near miss is short. The duration of each near miss is no longer than 1000s, which formed between passenger ship and cargo ship, as well as passenger ship and oil tanker, and the duration of each near miss is no longer than 1500s, which formed between oil tankers. On the whole, the number of long-duration near misses is smaller and the number of short-duration near misses is larger.

The advantage of box-plot is that it is not affected by outliers and can describe the discrete distribution of data in a relatively stable way, so it is used to analyze the duration of near misses which formed in different ship types. Fig.7 is the box-plot of the duration of near misses for cargo-cargo, tanker-cargo, cargo-tanker, cargo-passenger, tanker-tanker, passenger-cargo, and passenger-tanker, where “ship type1-ship type2” represents the near miss which is formed between ship type1 and ship type2, such as “cargo-cargo” represents the near miss which is formed between cargo ships, ○ is the long-duration near miss, green ▲ is the average duration of near misses, and orange line is median value of the duration of near misses.
It can be seen that there are significant differences between the number of long-duration near misses which formed between different ship types. The long-duration near misses which formed in cargo-cargo, tanker-cargo, cargo-tanker, are more than those formed in cargo-passenger, tanker-tanker, passenger-cargo, and passenger-tanker.

The number of cargo ships in Bohai Sea is the largest, the chance of forming long-duration near misses is greater, so the number of long-duration near misses which formed in cargo ships is the largest. The upper edge of each box is lower than 1500s (25min), indicating that the duration of most near misses is shorter than 1500s, and the number of short-duration near misses is largest.

The above-mentioned results indicate that cargo ship is more often than others involved in near miss; passenger ships have higher safety requirements, so the number of near misses is small, and the duration is short; the high frequencies of short-duration near misses are found; the longer the duration of near misses is, the smaller the number of those is. The statistical results of the questionnaire show that when the duration of near misses is less than 30min, it is generally considered that the longer the duration of near misses is, the higher the risk is. However, the duration of most near misses in Bohai Sea is less than 25min, so it is thought that the results of the questionnaire are applicable in the study water.

E. CHARACTERISTICS OF DURATION OF NEAR MISSES BASED ON SHIP LENGTH

Ship length is an important factor that affects the safety of ships. The lengths of all ships which involved in near misses in the study waters are counted in this article, and the results are shown in Fig.8. The number of ships with length of 220-230m is the largest, followed by 190-200m and 290-300m, all the numbers are larger than 600.

Fig.9 is the scatter diagram of duration of near misses with own ship length and target ship length. Each point represents a single near miss. The red point indicates that duration of a single near miss is greater than 2000s, and the blue point indicates that duration of a single near miss is less than or equal to 2000s. It can be seen that the distribution of points
representing duration of near misses shorter than 2000s are very dense, while the distribution of points representing duration of near misses longer than 2000s are sparse. According to statistics, the number of the scenes of near misses with the duration less than 2000s is 6132, and the ratio of that to the number of all near misses is 98.1%. It indicates that the number of long-duration near misses is smaller and the number of short-duration near misses is larger.

Fig. 10 is the scatter diagram of the own ship length and the target ship length in each near miss, and the detailed data are presented in Table 9. It can be seen that the points are dense, and the number of near misses is large, due to the large number of the ships with length is 220-230m, 190-200m, and 290-300m in the study waters. When the target ship length is shorter than 200m, the points are dense, and the number of those is 4621. When the target ship length is longer than 200m, the points are sparse, and the number of those is 1627. Especially, there are only 10 points when the own ship length is shorter than 100m and the target ship length is longer than 200m. It means that own ship is more likely to be involved in near miss if the target ship length is short than 200m.

Fig. 11 is the scatter diagram of the own ship length and the duration of near misses and Fig. 12 is the scatter diagram of the target ship length and the duration of near misses. It can be seen from Table 9, Fig. 11, and Fig. 12, the number of near misses is only 163, when the own ship length is shorter than 100m, which is smaller than that when the target ship length is short than 100m. However, the number of near misses is
TABLE 9. Number of Near Misses in Connection With Own Ship Length (OSL) and Target Ship Length (TSL).

| TSL | OSL | 0<=OSL<100 | 100<= OSL <200 | 200<= OSL <300 | 300<= OSL <400 |
|-----|-----|------------|----------------|----------------|----------------|
| 0<= TSL <100 | 63  | 458        |                |                |                |
| 100<= TSL <200 | 90  | 1554       |                |                |                |
| 200<= TSL <300 | 7   | 520        |                |                |                |
| 300<= TSL <400 | 3   | 74         |                |                |                |

FIGURE 11. The scatter diagram of the own ship length and the duration of near misses.

FIGURE 12. The scatter diagram of the target ship length and the duration of near misses.

950 when the own ship length longer than 300m, which is larger than that when the target ship length longer than 300m.

Using Eqs.(16) and (17) to calculate the similarity coefficient of ships’ lengths between the own ship length and the target ship length, the results are presented in Fig.13 (a). The number of near misses is 1651 when $L_{sim} < 0$, and the number of near misses is 4597 when $L_{sim} \geq 0$. There are 28 points when $L_{sim} \geq 0$ and the duration is longer than 3000s. However, there are only 5 points when $L_{sim} < 0$ and the duration larger than 3000s.

$k$-means clustering algorithm is used to cluster the duration of near misses in Bohai Sea. The appropriate value of $k$ is determined by Elbow method. Fig.14 and Table 10 show that the value of SSE at $k = 2$ to $k = 10$. The value of SSE is 1378912 at $k = 2$ then 6544511 at $k = 3$, which create a difference of 7244612. The others are 4174451, 2872259, 1798883, 1154094, 849204, 617525, 473775 at $k = 4$ to $k = 10$ and the difference gradually decrease, whose ranges from 2370059 to 143750. It can be seen that the change is quite drastic at the value of $k = 3$ wherein SSE is 6544511, which means that there is a difference of 7244612, higher than 2370059 at $k = 4$. It can be concluded that the optimal $k$ for the clustering data in Bohai Sea is 3 based on elbow method.

To analyze the overall variation of the data, different parameters $k$ are selected to compare the results of clustering. Due to different value is large when $k = 3, 4, 5, and 6$, $k$ is set to 3, 4, 5, and 6 respectively. According to the clustering results, the duration range of each cluster, the range of $L_{sim}$, the number of data and ratio of each cluster are counted, and then the average duration, the span of duration and the span of $L_{sim}$ are also calculated. The results are presented in Fig.13 (b) (c) (d) (e), and the detailed data are given in Tables 11, 12, 13, and 14 respectively.

Observing the last cluster of each table, in Table 11 and 12, the ranges of $L_{sim}$ both are $[-2.81, 0.87]$, and the spans of $L_{sim}$ both are 3.68; in Table 13 and 14, the ranges of $L_{sim}$ both are $[-2.81, 0.85]$, and the spans of $L_{sim}$ both are 3.66; in each table, the average duration is shorter than 200s, the range of $L_{sim}$ is the largest, the span of $L_{sim}$ is larger than 2900s, the number is larger than 3000, the radio is higher than 50%. Observing the cluster 1 of each table, the average duration is longer than 2300s, the range of $L_{sim}$ is the smallest, the span of $L_{sim}$ is smaller than 2, the number is smaller than 250, and the radio is lower than 4%. It means that the number of short-duration near misses is the largest; the frequencies of long-duration near misses are low; and the frequencies of short-duration near misses are high regardless of $L_{sim}$.

TABLE 10. Result of $K$-Means Clustering.

| $k$ | SSE       | Difference of SSE |
|-----|-----------|-------------------|
| 2   | 1378912   | 1378912           |
| 3   | 6544511   | 7244612           |
| 4   | 4174451   | 2370059           |
| 5   | 2872259   | 1302192           |
| 6   | 1798883   | 1073376           |
| 7   | 1154094   | 644789            |
| 8   | 849204    | 304890            |
| 9   | 617525    | 231679            |
| 10  | 473775    | 143750            |

VOLUME 8, 2020
Observing the data from the first cluster to the last cluster, it can be seen that the shorter the average duration is, the larger the span of $L_{sim}$ is. However, in Table 14, the span of $L_{sim}$ of cluster 3 is 2.3, which is larger than that of cluster 4.

### TABLE 11. Result of K-Means Clustering ($K = 3$).

| Cluster | Range of duration (s) | Span of duration (s) | Average duration (s) | Range of $L_{sim}$ | Span of $L_{sim}$ | Number | Ratio |
|---------|-----------------------|----------------------|----------------------|-------------------|-----------------|--------|-------|
| 1       | [1620,6180]           | 4560                 | 2315.36              | [-0.94,0.81]      | 1.75            | 224    | 3.6%  |
| 2       | [540,1560]            | 1020                 | 861.71               | [-1.47,0.85]      | 2.32            | 1592   | 25.5% |
| 3       | [60,480]              | 420                  | 194.68               | [-2.81,0.87]      | 3.68            | 4452   | 70.9% |

**FIGURE 13.** Results of k-means clustering.
TABLE 12. Result of K-Means Clustering ($K = 4$).

| Cluster | Range of duration (s) | Span of duration (s) | Average duration (s) | Range of $L_{sim}$ | Span of $L_{sim}$ | Number | Ratio |
|---------|-----------------------|----------------------|----------------------|-------------------|------------------|--------|-------|
| 1       | [2160,6180]           | 4080                | 2852.14              | [-0.63,0.81]      | 1.44             | 112    | 1.8%  |
| 2       | [960,2040]            | 1080                | 1290.55              | [-1.47,0.83]      | 2.3              | 656    | 10.5% |
| 3       | [420,9090]            | 480                 | 611.31               | [-1.48,0.85]      | 2.33             | 1565   | 25.0% |
| 4       | [60,3600]             | 300                 | 161.06               | [-2.81,0.87]      | 3.68             | 3915   | 62.7% |

TABLE 13. Result of K-Means Clustering ($K = 5$).

| Cluster | Range of duration (s) | Span of duration (s) | Average duration (s) | Range of $L_{sim}$ | Span of $L_{sim}$ | Number | Ratio |
|---------|-----------------------|----------------------|----------------------|-------------------|------------------|--------|-------|
| 1       | [2820,6180]           | 3360                | 3590.45              | [-0.63,0.80]      | 1.43             | 44     | 0.7%  |
| 2       | [1560,2760]           | 1200                | 1949.56              | [-1.12,0.81]      | 1.93             | 205    | 3.3%  |
| 3       | [840,1500]            | 660                 | 1082.69              | [-1.47,0.83]      | 2.3              | 736    | 11.8% |
| 4       | [360,780]             | 420                 | 529.03               | [-1.48,0.87]      | 2.35             | 1674   | 26.8% |
| 5       | [60,3000]             | 240                 | 142.99               | [-2.81,0.83]      | 3.66             | 3589   | 57.4% |

TABLE 14. Result of K-Means Clustering ($K = 6$).

| Cluster | Range of duration (s) | Span of duration (s) | Average duration (s) | Range of $L_{sim}$ | Span of $L_{sim}$ | Number | Ratio |
|---------|-----------------------|----------------------|----------------------|-------------------|------------------|--------|-------|
| 1       | [3240,6180]           | 2940                | 4026.92              | [-0.63,0.72]      | 1.35             | 26     | 0.4%  |
| 2       | [1920,3180]           | 1260                | 2384.59              | [-0.94,0.81]      | 1.75             | 109    | 1.7%  |
| 3       | [1140,1860]           | 720                 | 1399.71              | [-1.47,0.83]      | 2.3              | 414    | 6.6%  |
| 4       | [660,1080]            | 420                 | 825.49               | [-1.03,0.84]      | 1.87             | 881    | 14.1% |
| 5       | [300,6000]            | 300                 | 425.89               | [-1.62,0.87]      | 2.49             | 1589   | 25.4% |
| 6       | [60,2400]             | 180                 | 125.48               | [-2.81,0.85]      | 3.66             | 3229   | 51.7% |

Observing the Fig.13 (d), there is only one point in cluster 3 with the $L_{sim}$ which is about -1.5. It means that this point is a special value point that should be ignored, and the law what is got in the former is still valid.

Comparing the span of $L_{sim}$, the minimum of $L_{sim}$ in each cluster changed a lot, the value of the largest difference can reach 2.31 (cluster 1 and cluster 6 in Table 14), but the maximum of $L_{sim}$ changed little, which is larger than 0.72 and smaller than 0.82. It can be seen that the longer the average duration is, the larger the value of $L_{sim}$ is, which means the long-duration near misses generally have the large $L_{sim}$.

The above-mentioned results indicate that the frequencies of short-duration near misses are high regardless of $L_{sim}$, the frequencies of long-duration near misses are low, and the long-duration near misses generally have the large $L_{sim}$.

F. CHARACTERISTICS OF DURATION OF NEAR MISSES BASED ON SHIP SPEED

In this article, the average speed is defined as the average value of the ship speed during the whole process of a single near miss. The average speed of own ship and that of target ship are calculated respectively. The results are shown in Fig.15, and the detailed data are given in Table 15. When the average speed of the own ship and the average speed of target ship are both lower than 15kn, the number of near misses is 5905, and the ratio is 94.5%. It means that, the ships are more likely to be involved in near miss in case the average speed of the own ship and the average speed of target ship are both lower than 15kn.

Fig.16 is the scatter diagram of the duration of near misses, the average speed of own ship and the average speed of target ship. The red point indicates that duration of a single near miss is greater than 2000s, and the blue point indicates that duration of a single near miss is less than or equal to 2000s. It can be seen that the number of long-duration near misses is smaller and the number of short-duration near misses is
larger. There are few ships involved in long-duration near misses when the average speed of own ship and the average speed of target ship are both high.

Fig.17 is the scatter diagram related to the average speed of own ship and the duration of near misses, and Fig.18 is the scatter diagram related to the average speed of target ship and the duration of near misses. It can be seen from Table 15, Fig.17, and Fig.18, the number of near misses is only 134, which is smaller than that when the average speed of target ship is higher than 15kn in case the average speed of own ship is higher than 15kn. However, in case the average speed of own ship is higher than 5kn, the number of near misses is 2384, which is larger than that when the average speed of target ship is less than 5kn.
TABLE 18. Result of K-Means Clustering ($K = 5$).

| Cluster | Range of duration (s) | Span of duration (s) | Average duration (s) | Range of average relative speed (kn) | Span of average relative speed (kn) | Number | Ratio |
|---------|-----------------------|----------------------|----------------------|---------------------------------------|-----------------------------------|--------|-------|
| 1       | [2820,6180]           | 3360                 | 3590.45              | [0.34,6.52]                           | 6.18                              | 44     | 0.7%  |
| 2       | [1560,2760]           | 1200                 | 1949.56              | [0.27,13.38]                          | 13.11                             | 205    | 3.3%  |
| 3       | [840,1500]            | 660                  | 1082.69              | [0.16,17.48]                          | 17.64                             | 736    | 11.8% |
| 4       | [360,780]             | 420                  | 529.03               | [0.17,19.76]                          | 19.59                             | 1674   | 26.8% |
| 5       | [60,300]              | 240                  | 142.99               | [0.11,29.05]                          | 28.94                             | 3589   | 57.4% |

TABLE 19. Result of K-Means Clustering ($K = 6$).

| Cluster | Range of duration (s) | Span of duration (s) | Average duration (s) | Range of average relative speed (kn) | Span of average relative speed (kn) | Number | Ratio |
|---------|-----------------------|----------------------|----------------------|---------------------------------------|-----------------------------------|--------|-------|
| 1       | [3240,6180]           | 2940                 | 4026.92              | [0.34,6.52]                           | 6.18                              | 26     | 0.4%  |
| 2       | [1920,3180]           | 1260                 | 2384.59              | [0.27,13.38]                          | 13.11                             | 109    | 1.7%  |
| 3       | [1140,1860]           | 720                  | 1399.71              | [0.27,13.24]                          | 12.97                             | 414    | 6.6%  |
| 4       | [660,1080]            | 420                  | 825.49               | [0.16,17.64]                          | 17.48                             | 881    | 14.1% |
| 5       | [300,600]             | 300                  | 425.89               | [0.14,19.62]                          | 19.62                             | 1589   | 25.4% |
| 6       | [60,240]              | 180                  | 125.48               | [0.11,29.05]                          | 28.94                             | 3229   | 51.7% |

FIGURE 16. The scatter diagram of the duration of near misses with the average speed of own ship and the average speed of target ship.

FIGURE 17. The scatter diagram of the average speed of own ship and the duration of near misses.

The average relative speed is defined as the average value of relative speed between own ship and target ship during the whole process of a single near miss, which can show the overall situation of the relative speed of the two ships.
in this scene. The average relative speed in each near miss is calculated, and the results are presented in Fig.19 (a). It can be seen that, there are a few cases that the duration of near miss is longer than 4000s when the average relative speed is less than 5kn. However, in case the average relative speed is more than 15kn, there are almost none situation that the duration of near miss longer than 1000s. As the average relative speed increases, the number of near misses decreases.

According to the $k$-means clustering results in Section III, the range of average relative speed, the number of data and ratio of each cluster are counted, and then the average duration, the span of duration and the span of average relative speed are calculated. The results are presented in
Fig. 20. The scatter diagram of the duration of near misses, the similarity coefficient of ships’ lengths and the average relative speed.

Fig. 19 (b) (c) (d) (e), and the detailed data are presented in Tables 16, 17, 18, and 19 respectively.

Observing the last cluster of each table, the range of average relative speed and the span of average relative speed are both largest, which are [0.11, 29.05] and 28.94 respectively; the average duration is shorter than 200s, the number is larger than 3000, and the ratio is higher than 50%. Observing the cluster 1 of each table, the average duration is longer than 2300s, the range of average relative speed is the smallest, the span of average relative speed is smaller than 2, the number is smaller than 250, and the radio is lower than 4%. It means that the number of short-duration near misses is largest; the frequencies of long-duration near misses are low; and the frequencies of short-duration near misses are high whatever relative speed.

Observing the data from the first to the last cluster, it can be seen that the shorter the average duration is, the larger the span of average relative speed is. Comparing the span of average relative speed, the maximum of average relative speed in each cluster changed a lot, the value of the largest difference can reach 22.76kn (cluster 1 and cluster 6 in Table 19), but the minimum of average relative speed changed little, which is larger than 0.1 and smaller than 0.35. It can be seen that the longer the average duration is, the smaller the value of maximum of average relative speed is, which means the long-duration near misses generally happen with the small relative speed.

The above-mentioned results indicate that the frequencies of short-duration near misses are high whatever relative speed, the frequencies of long-duration near misses are low, and the long-duration near misses generally have the small relative speed.

The duration, $L_{sim}$, and the average relative speed of each near miss are presented in Fig. 20. The red point indicates that duration of a single near miss is greater than 2000s, and the blue point indicates that duration of a single near miss is less than or equal to 2000s. It can be seen that the longer the duration is, the sparser the data points are, which means the number of long-duration near misses is smaller; the long-duration near misses generally have large $L_{sim}$ and small average relative speed; the short-duration near misses exist, regardless of the value of $L_{sim}$ and average relative speed.

IV. CONCLUSION

The duration of near misses shows the maritime traffic risk, the long-duration near miss means high maritime traffic risk. The navigation safety domain is used to detect near misses accurately from AIS database. The influence of ship maneuverability, size, speed, COLREGS on ship safety are considered, and then the model for detecting near misses based on the navigation safety domain is proposed. Concept about duration of a single near miss is defined as time from target ship sailing into the ship domain of own ship to leaving, and then the model for calculating the duration of near misses is proposed. Due to the absolute difference of two ships’ lengths which cannot fully show the similarity of those, the similarity coefficient model of ships’ lengths is proposed. The duration of each near miss in the study waters is calculated, and the clustering results of $k$-means for the duration of all near misses are analyzed. Finally, some connections between duration of near misses, ship type, length and speed are found in this article.

Passenger ships have high safety requirements, so the number of near misses is small, and the duration is short; the
frequencies of short-duration near misses are high regardless of ship type, the value of $L_{sim}$ and the value of average relative speed; the connections between long-duration near misses, ship length and speed are strong, the frequencies of long-duration near misses are low, and the long-duration near misses generally have large $L_{sim}$ and small average relative speed.

Nonetheless, the model can be further refined to account for factors such as hydrometeorological conditions. The connections between the duration of near misses and the traffic environment are also worth studying. This article provides a method for studying maritime traffic risk from temporal dimension perspective, the connections found in this article can help operators to automatically detect the high potential risk scenarios and help researchers and maritime traffic management agencies to better understand the maritime traffic risk from temporal dimension perspective.

REFERENCES

[1] D. Chen, C. Dai, X. Wan, and J. Mou, “A research on AIS-based embedded system for ship collision avoidance,” in Proc. Int. Conf. Transp. Inf. Syst. (ICTIS), 2015, pp. 512–517.

[2] T. J. Moore, J. V. Redfern, M. Carver, S. Hastings, J. D. Adams, and G. K. Silber, “Exploring ship traffic variability off California,” Ocean Coastal Manag., vol. 163, pp. 515–527, Sep. 2018.

[3] L. Du, F. Goerlandt, and P. Kujala, “Review and analysis of methods for assessing maritime waterway risk based on non-accident critical events detected from AIS data,” Rel. Eng. Syst. Saf., vol. 200, Aug. 2020, Art. no. 106933.

[4] S.-L. Yoo, “Near-miss density map for safe navigation of ships,” Ocean Eng., vol. 163, pp. 15–21, Sep. 2018.

[5] R. Berglund and M. Huttunen, “AIS data analysis for identification of close encounter situations of vessels,” Sci. Activities Saf. Secur., pp. 106–107, 2009.

[6] J. Fukuto and H. Imazu, “New collision alarm algorithm using obstacle zone by target (OZT),” IFAC Proc. Volumes, vol. 46, no. 33, pp. 91–96, 2013.

[7] J. W. Park, “A study on the estimation of a safe speed for prevention of ship’s collision in VTS area,” M.S. thesis, Navigation, College, Mokpo Maritime Univ., Mokpo, South Korea, 2014.

[8] B. Langard, G. Morel, and C. Chauvin, “Collison risk management in passenger transportation: A study of the conditions for success in a safe shipping company,” Psychologie Francaise, vol. 60, no. 2, pp. 111–127, Jun. 2015.

[9] K.-I. Kim and I. S. Jeong, “Visualization of ship collision risk based on near-miss accidents,” in Proc. Joint 8th Int. Conf. Soft Comput. Intell. Syst. (SCIS), 17th Int. Symp. Adv. Intell. Syst. (ISIS), Sapporo, Japan, Aug. 2016, pp. 323–327.

[10] F. Goerlandt, J. Montewka, H. Lammi, and P. Kujala, “Analysis of near collisions in the Gulf of Finland,” Astron. Nachr., vol. 326, no. 18, pp. 2880–2886, 2012.

[11] Z. R. Tan, Q. L. Zhang, W. Wang, and Z. Z. Fan, “Visualization of traffic conflict risk in Chengshantou waters based on near misses,” China Navigation, vol. 39, no. 4, pp. 43–46 and 107, 2016.

[12] W. Zhang, F. Goerlandt, J. Montewka, and P. Kujala, “A method for detecting possible near miss ship collisions from AIS data,” Ocean Eng., vol. 107, pp. 60–69, Oct. 2015.

[13] W. Zhang, F. Goerlandt, P. Kujala, and Y. Wang, “An advanced method for detecting possible near miss ship collisions from AIS data,” Ocean Eng., vol. 124, pp. 141–156, Sep. 2016.

[14] W. Zhang, C. Kopca, J. Tang, D. Ma, and Y. Wang, “A systematic approach for collision risk analysis based on AIS data,” J. Navigat., vol. 70, no. 5, pp. 1117–1132, Sep. 2017.

[15] L. Le Tixerant, D. Le Guayder, F. Gourmelon, and B. Queffelec, “How can automatic identification system (AIS) data be used for maritime spatial planning?” Ocean Coastal Manage., vol. 166, pp. 18–30, Dec. 2018.
YANGYU ZHOU received the B.E. degree in navigation technology from Jimei University, Xiamen, China, in 2018. He is currently pursuing the master’s degree in navigation science and technology with the Navigation College, Dalian Maritime University, Dalian, China. His current research interests include ocean big data analysis, AIS data mining, and marine traffic engineering.

JIAXUAN YANG was born in Shandong, China, in 1981. He received the Ph.D. degree in traffic information engineering and control from Dalian Maritime University, Dalian, China, in 2008. He is currently an Associate Professor of transportation engineering with the Navigation College, Dalian Maritime University, and also with the Key Laboratory of Navigation Safety Guarantee of Liaoning Province. His research interests include marine traffic engineering, maritime safety evaluation, vessel traffic organization and management, AIS data mining, and the collision avoidance of USVs.

YUAN LIU received the B.E. degree in industrial engineering from the Guangdong University of Technology, Guangzhou, China, in 2018. He is currently pursuing the master’s degree in traffic and transportation engineering with Dalian Maritime University, Dalian, China. His current research interests include ocean big data research and analysis, and AIS data mining.

JIAGUO LIU was born in Zaoyang, China, in 1979. He received the B.S. degree in management business administration from Northeast University, Shenyang, China, in 2003, and the Ph.D. degree in management engineering from Harbin Engineering University, Harbin, China, in 2010. He is currently a Professor with Dalian Maritime University, China. His research interests include supply chain management, procurement management, and maritime operations management.