Collision risk assessment for ships' routeing waters: an information entropy approach with Automatic Identification System (AIS) data

Abstract: The ship's routing was adopted to organise marine traffic flow and reduce the risk of collision between ships in crowded waters. With the expansion of the world's fleet, ship traffic in shipping bottleneck and chokepoint areas became more and more busy and complex creating serious challenges for navigational safety. Therefore, quantitative collision risk assessment is significantly important for the ships' routeing waters. In this paper, the information entropy method which integrates the K-means clustering based on automatic identification system (AIS) data is introduced to quantitatively evaluate the collision risks in the ships' routeing waters. As a case study, the information entropy of courses over ground (COG) for Ningbo-Zhoushan Port (the largest port in the world since 2009) is calculated by using historical AIS data. Then the K-means clustering is used to group the bytes of information entropy of the different legs in the shipping route. We find that in Ningbo-Zhoushan port Precautionary Area (PA) 2, 4 and 7 are the highest risk legs; PA 1, 5 and 6, Traffic Separation Scheme (TSS) 16, and 17 are medium-high risk areas. Therefore ship collision risk prevention measures should be prioritised in those legs. Our contributions provide a novel approach to quantitatively assess ship collision risks in busy waters.

Keywords: automatic identification system (AIS) data; ship collision risk assessment; information entropy; K-means clustering; Ningbo-Zhoushan Port.

1. Introduction

Since 1967, the ships' routing was widely adopted in busy waters. As an effective control measure, ships' routeing contributed significantly in terms of enhancing navigational safety and promoting marine traffic efficiency. But ships' routeing cannot completely eliminate the marine traffic conflict and ship collisions. For example ship's routing can cause traffic to converge on PAs where traffic density and complexity may increase **(Li *et al.* 2015)**; there are still some vessels crossing the traffic separation zone or separation line for some reasons. With increased trade and an expanding world fleet, traffic density increased along with the risk **(Zhang *et al.* 2021)**. Ships' routeing and traffic separation schemes were formulated after extensive marine surveys and consultation with ship's captains familiar with the area, and various shore-based marine specialists **(Qu *et al.* 2011)**. However, the subjective judgment from experts failed to qualitatively alert visiting vessels as to which legs were high risk or to advise administrative authorities as to where risk reduction measures should be carried out. Consequently, the problem of ship collision risk remains a serious issue arousing widespread attention of all parties. It is therefore meaningful to present an approach to quantitatively estimate collision risk in ships' routeing areas in order to support these judgments.

Up to now, there exist a large number of research to value collision risk in ships' routeing waters, for instance, the Gulf of Finland **(Kujala *et al.* 2009)**, Dover Strait **(Squire 2003, Cockcroft 2004)**,

the Coast of Portugal **(Silveira *et al*. 2013)**, Messina Strait **(Cucinotta *et al*. 2017)**, Istanbul Strait **(Aydogdu *et al*. 2012)**, Baltic Sea **(Kulkarni *et al*. 2020)**, Taiwan Strait **(Chai and Xue 2021)**, Yangtze River **(Li *et al.* 2015, Wang *et al.* 2020, Zhang *et al.* 2020a)**, Singapore Strait **(Kang *et al*. 2018, Kang *et al*. 2019, Zhang and Meng 2019, Zhang *et al*. 2019)** and so on. To some extent, the risk is a qualitative and somewhat fuzzy issue. For quantitative estimation of the collision risk in busy waters, the methodological effort has been made for half a century. In 1974, **Fujii and Tanaka (2010)** proposed the 'ship domain'. After that, ship domain was widely applied and developed to measure marine traffic conflict and ship collision risk, for example **Rawson and Brito (2021)**, **Yu *et al*. (2021)**, **Liu *et al.* (2020a)**, **Zhang *et al*. (2019)** and **Xin *et al*. (2019)**. However, according to **Pietrzykowski (2008)** and **Szlapczynski (2006)**, the ship domain was not fixed but changed with certain subjectivity. Also, the collision risk assessment was usually hindered by the insufficiency of data samples **(Goerlandt and Kujala 2011)**.

In recent years, the availability of AIS data provides an excellent way to overcome this difficulty. AIS is used to automatically exchange static and dynamic data between ships and between ship and shore stations. According to *the International Convention for the Safety of Life At Sea (SOLAS)*, it is mandatory for ships of 300GT or over, on international voyages, and all passenger ships, to be fitted with AIS. With AIS data and on the basis of ship domain, **Qu *et al*. (2011)** added 2 indexes, i.e., index of speed dispersion, degree of acceleration and deceleration to quantitatively measure the collision risks in Singapore Strait. Combing the AIS data, **Li *et al.* (2015)** used the navigational traffic conflict technique to conduct an investigation on traffic safety in the precautionary areas of ships' routeing. **Cucinotta *et al.* (2017)** employed the IWRAP model to calculate the frequency of ship collisions in the Strait of Messina. **Zhang *et al*. (2019)** used origin-to-destination pairs to count and plot hotspot areas and geographical distribution of ship accidents in the Singapore Port. Other methodological approaches include the convolutional neural network model **(Zhang *et al*. 2020b)**, molecular dynamics **(Liu *et al*. 2020b)**, sequence conditional generative adversarial network **(Gao and Shi 2020)**, etc. In this paper, the information entropy which integrates the K-means clustering basing on the AIS data is introduced as a novel approach to quantitatively measure risk of collision in ship routine areas.

Following the Introductory Section, Section 2 introduces the ship collision risk assessment and K-means clustering methods. Section 3 gives a brief description of Ningbo-Zhoushan Port as a case study along with AIS data pre-processing methods. Section 4 includes the results and discussions. Conclusions are finally drawn in Section 5.

2. Methodology

The entropy methodological approach which integrates the K-means clustering based on AIS data is depicted in Figure 1. The method and procedures for AIS data processing is introduced in section 3.3 combining the AIS data of Ningbo-Zhoushan Port.

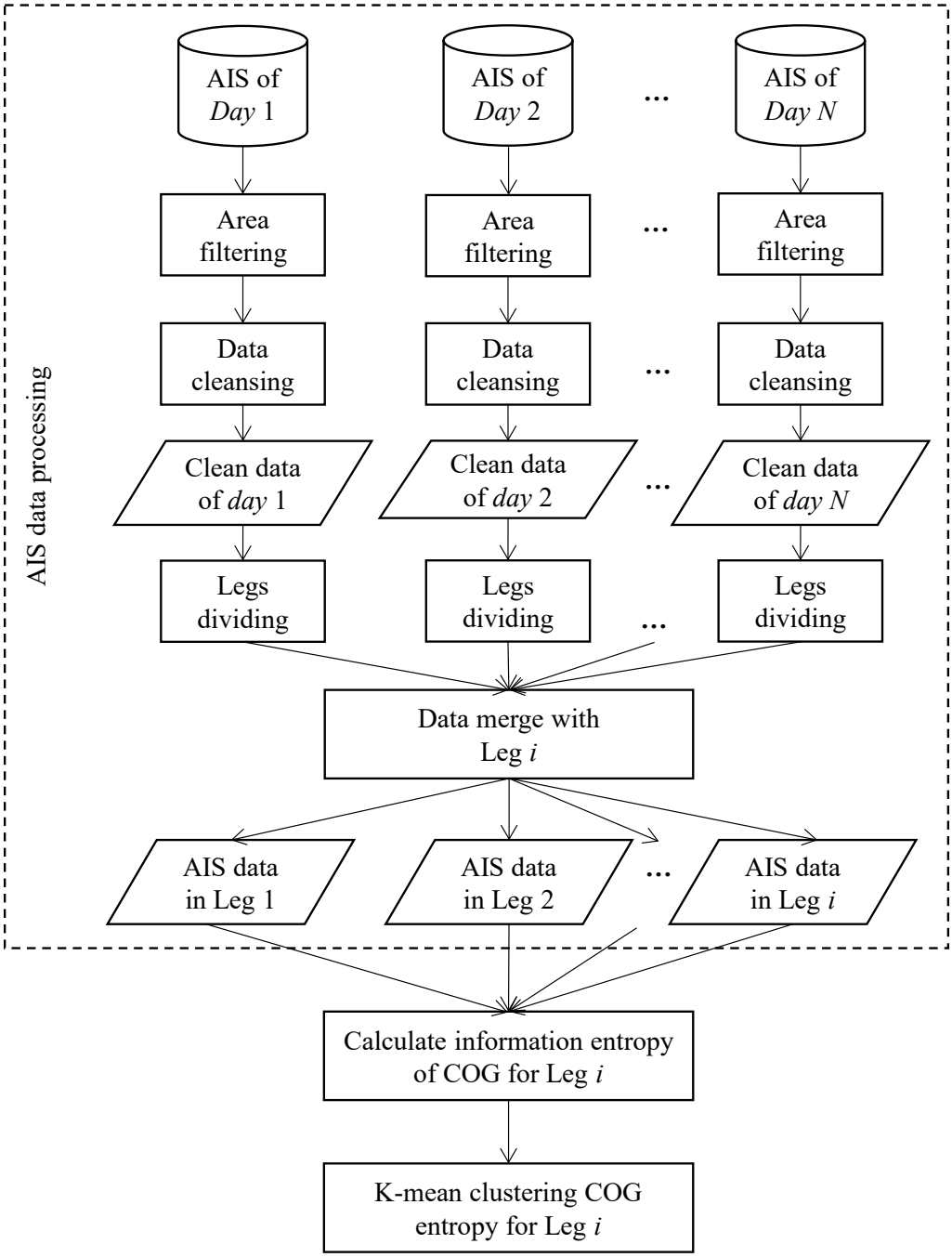

Figure 1. Framework and workflow for this entropy method which integrates the K-means clustering based on AIS data.

2.1. Information Entropy

The concept of information entropy was presented by **Shannon (1948)** to measure the uncertainty of information. According to **Núñez *et al.* (1996)**, the more orderly a system is, the lower the information entropy will be; Conversely, the more chaotic a system is, the higher the information entropy will be. In this sense, information entropy could be used to investigate the degree of navigational order in a ship's routeing. The formula for information entropy is as follows:

$$H(p_1, p_2, \cdots, p_n) = -\sum_{i=1}^{n} p_i(x)\log_2 p_i(x) \quad (1)$$

Where n denotes the number of information records in a system; $p_i(x)$ is the probability of the i^{th} record; and $\log_2 p_i(x)$ means the information offered by the i^{th} record.

According to *Convention on the International Regulations for Preventing Collision at Sea, 1972 (COLREGs)* **(IMO 1972)**, *'a vessel using a traffic separation scheme shall: (i) proceed in the appropriate traffic lane in the general direction of traffic flow for that lane; (ii) so far as practicable keep clear of a traffic separation line or separation zone; (iii) normally join or leave a traffic lane at the termination of the lane'*. Therefore, the distributions of ships' course over ground (i.e., COG) could be used to reflect the degree of order in the routing area.

For a ships' routeing area, if all vessels sailing in the same direction, the information entropy of COG is equal to 0 (see **Figure 2** left); the entropy of COG is equal to $0.5n$ (where n denotes the total quantity of ships) when the same number of ships sail in opposite directions (see **Figure 2** centre); and if n vessels are all navigating with different headings, the entropy of COG is maximum, i.e., $\log_2 n$ (see **Figure 2** right).

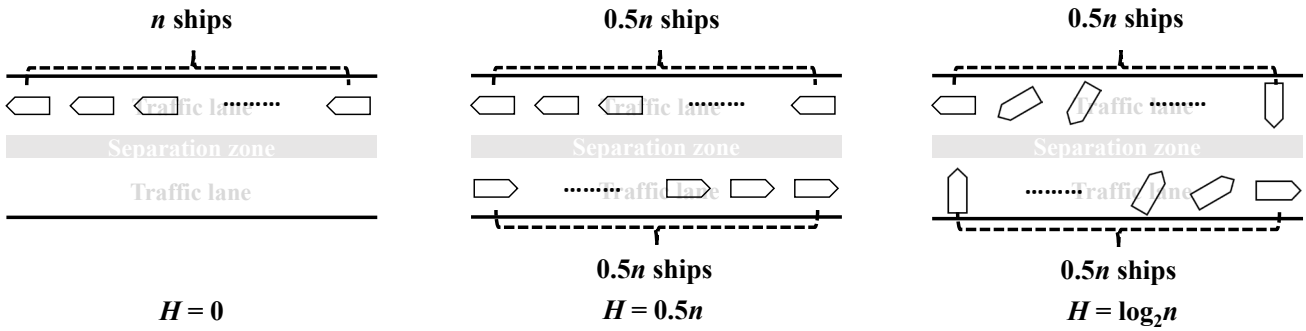

Figure 2. Information entropy of COG in different scenarios: H=0 (left), H=0.5n (center) and H=$\log_2 n$ (right).

2.2. K-means clustering

K-means clustering is a machine learning approach that divides n records into k clusters, and each record falls within the cluster with the minimum mean **(Adamidis *et al.* 2020)**. Each cluster has a barycentre. The barycentre represents the cluster it belongs to and its value is the average of all records. The target of K-means clustering is to obtain the minimized objective function:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} ||r_i^{(j)} - b_j||^2 \quad (2)$$

Where $||r_i^{(j)} - b_j||^2$ is the chosen distance measure between a record $r_i^{(j)}$ and the barycentre b_j of its cluster **(MacQueen 1967)**.

According to **Hartigan and Wong (1979)**, Lloyd's algorithm is usually employed to realize K-means clustering. This algorithm is based on Euclidean distance (ED) **(Hamerly and Elkan 2002)** and follows two steps. Firstly, each record is grouped into a cluster with the minimum ED. Secondly, the new barycentre is calculated as the mean values of the records in the same cluster. When the

distribution of records remains unchanged, the iterative process stops and it is possible to converge to the local minimum.

According to **Bolshakova and Azuaje (2003)**, the Silhouette index (s(i)) is computed to verify the outcomes of K-means clustering, that's:

$$\mathrm{s}(i) = \frac{c(i)-a(i)}{\max(c(i),a(i))} \tag{3}$$

where $a(i)$ and $c(i)$ are respectively the average distance and minimum distance between the record i and other records. s(i) is in the interval $[-1,1]$. If s(i) is approaching 1, then the clustering would be correct. A value close to 0 means the record could be categorized into the cluster with the next closest mean, and a negative s(i) implied the clustering is wrong. The average s(i) is computed Formula (4):

$$S_j = \frac{1}{m}\sum_{i=1}^{m} s(i) \tag{4}$$

The reliability of clustering is measured by Global Silhouette (GS), and GS is calculated by:

$$GS_u = \frac{1}{b}\sum_{j=1}^{b} S_j \tag{5}$$

The K-means clustering is convenient to use. There is a Matlab toolbox accessible at https://ww2.mathworks.cn/help/stats/kmeans.html.

3. Case study

In this section, Ningbo-Zhoushan Port is used as a case study for introducing the entropy method with K-means clustering.

3.1 Brief of Ningbo-Zhoushan Port

Since 2009 Ningbo-Zhoushan Port has been the largest port in the world in terms of port throughput. In 2021, over 3500 ships traversed this port on a daily basis, almost 1.5 times that of the Singapore Straits **(Kang *et al.* 2019).** Huge and intensive marine traffic accompanies by enormous ship collision risks. Therefore, the navigational safety of vessels visiting this port causes great concern for the maritime administrative authorities and relevant stakeholders. Regulations for promoting traffic efficiency and enhancing navigational safety have been put into force over the past decades, for instance, the Traffic Separation Scheme (TSS) and the Mandatory Ship Reporting System (MSRS) were formulated in 2010 and further updated in 2016. The latest TSS and MSRS in Ningbo-Zhoushan Port includes 28 legs, i.e., 18 TSSs, 8 PAs and 2 Two-Way Routes (TWRs) (see **Figure 3**). Ningbo-Zhoushan Port has a complex navigational environment comparable to Singapore, however, few academic concerts are given currently. In this sense, the assessment of ship collision risk at Ningbo-Zhoushan Port is a worthwhile subject to attract our study.

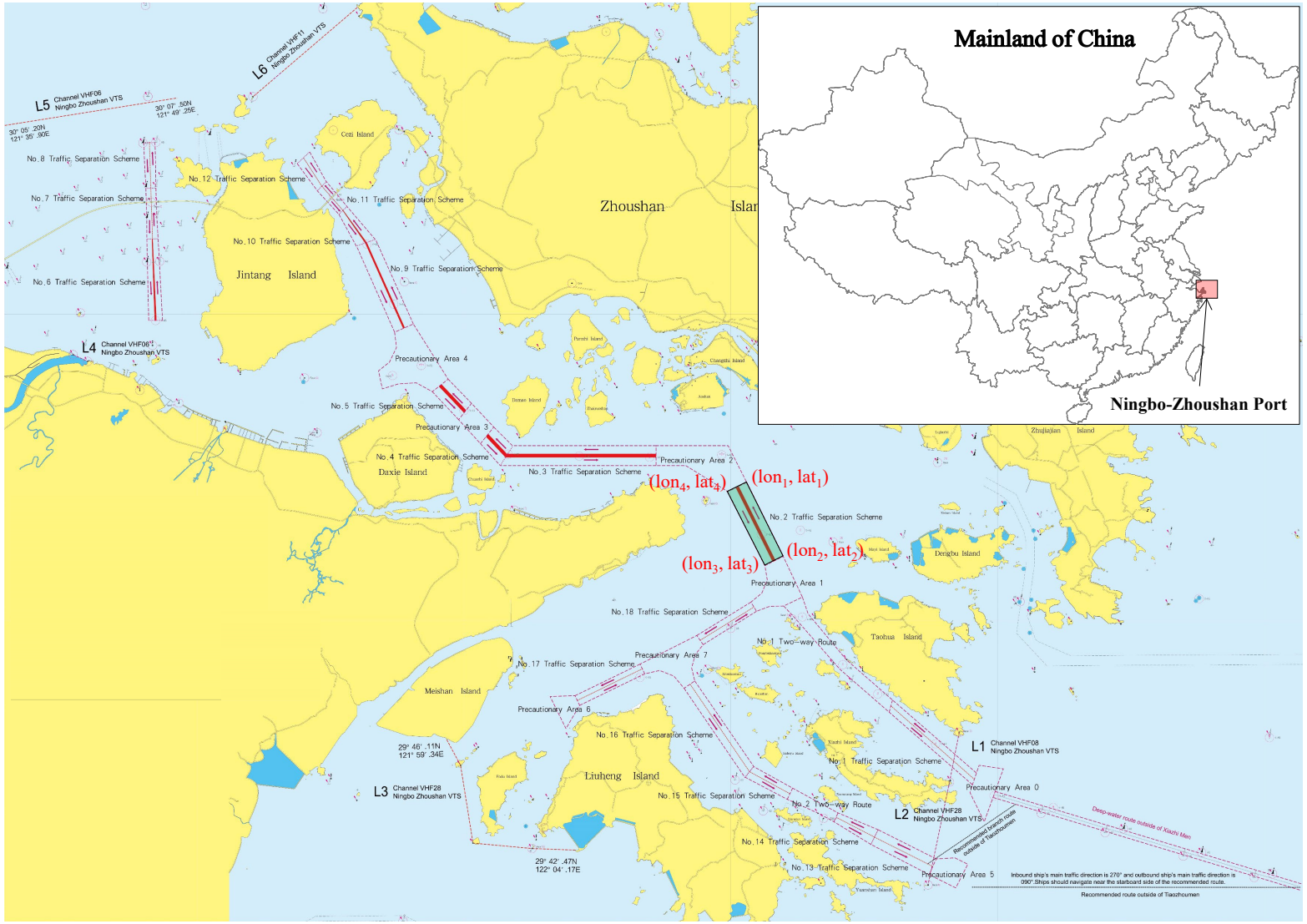

Figure 3. Legs of ships' routeing waters in Ningbo-Zhoushan Port and Leg No.2 TSS (light green rectangle in the centre).

3.2 AIS DATA

China's shore-based AIS network is split into three regions, North, East and South China Sea. Each region is overseen by a special Navigation Service Center. A Navigation Service Center includes several administrative offices and Ningbo-Zhoushan Port is located in the area of Ningbo Office of Eastern Navigation Service Center. Our data is officially provided by the Ningbo Office.

To proceed with a quantitative assessment of ship collision risk for Ningbo-Zhoushan Port, the data from Oct. 1st to Dec. 31st, 2020 was used. In the data set, daily data contained nearly 10 million records, and each record provides dynamic information (for instance, position in latitude and longitude, speed over ground (SOG), COG, reporting time) and statics ones (e.g. MMSI (maritime mobile service identify), ship name, call sign, ship size, draught, destination, ETA (estimated time of arrival).

3.3 AIS data processing

The statics information is highly reliable. However, the dynamics ones will inevitably appear abnormal in the process of generation, transmission and reception. Data noise would make some original AIS data unable to reflect the real movement of ships. For instance, latitude and longitude of ship positions should fall between 90º North and South, 180º East and West. However, some longitudes are 181ºE, and some latitudes are 91ºN. Additionally, the speed limit in Ningbo-Zhoushan Port is 16 knots (kn) downstream and 14 kn upstream. Yet, AIS data indicate that some ships navigate at a speed over 30 kn, even 100 kn. Therefore, we must conduct a data pre-processing prior to the AIS data mining.

3.3.1 Area filtering

Our original data coverage is from 31ºN/120ºE to 28ºN/125ºE, the volume of a whole day's data is about 1.3G. The ships' routeing areas in Ningbo-Zhoushan Port account ~~s~~ for only 0.31% of the total data coverage area. Therefore, we filtered out data that fall outside of the research area to promote calculation efficiency. The procedure is to check record by record whether the ship was sailing in the research area during the period of Oct. 1st to Dec. 31st, 2020. Let *lat* and *lon* are the latitude and longitude values in the AIS records, and the *lat* and *lon* should satisfy:

$$lat_{min} = min(lat_{1,min}, lat_{2,min}, \cdots, lat_{N,min}) \tag{6}$$

$$lat_{max} = min(lat_{1,max}, lat_{2,max}, \cdots, lat_{N,max}) \tag{7}$$

$$lon_{min} = min(lon_{1,min}, lon_{2,min}, \cdots, lon_{N,min}) \tag{8}$$

$$lon_{max} = min(lon_{1,max}, lon_{2,max}, \cdots, lon_{N,max}) \tag{9}$$

$$lat_{min} < lat < lat_{max} \tag{10}$$

$$lon_{min} < lon < lon_{max} \tag{11}$$

Where $lat_{i,min}$, $lat_{i,max}$, $lon_{i,min}$ and $lon_{i,max}$ are the minimum latitude, maximum latitude, minimum longitude and maximum longitude of Leg *i*, $i \epsilon \{1,2,\cdots,N\}$.

After filtering, the volume of a whole day's data in the study area is about 100M. Taking Oct. 30th 2020 as an example, 3,540 ships and 2,650,681 records are caught.

3.3.2 Data cleansing

According to **Zhao *et al.* (2018)**, errors would impact the quality of AIS data. In this study, we conducted data cleansing according to the procedures set by **Feng *et al.* (2021)**

First, we calculated the distance between the positions at time t_j and t_{j+1}under the unique MMSI by:

$$D_{i,t_j} = \sqrt[2]{(lon_{i,t_j} - lon_{i,t_{j+1}})^2 + (lat_{i,t_j} - lat_{i,t_{j+1}})^2} \tag{12}$$

Where $lat_{i,j}$ and lon_{i,t_j} are separately the latitude and longitude of the ship with MMSI i in at time t_j; $lat_{i,t_{j+1}}$ and $lon_{i,t_{j+1}}$ denote the latitude and longitude in at time t_{j+1}.

Secondly, mean value μ and variance s of D_{i,t_j} in the research, areas were obtained according to:

$$\mu = \frac{\sum_{i=1}^{i=I}\sum_{j=1}^{j=J_i-1} D_{i,t_j}}{\sum_{i=1}^{i=I}(J_i-1)} \tag{13}$$

$$s = \sqrt[2]{\frac{\sum_{i=1}^{i=I}\sum_{j=1}^{j=J_i-1}(D_{i,t_j}-\mu)^2}{\sum_{i=1}^{i=I}(J_i-1)}} \tag{14}$$

where $J_i$is the record number of the ship with MMSI i in the research area; I means the total quantity of ships in the selected area.

Finally, delete the abnormal data. According to our statistics, 2, 650, 680 distances are obtained, of which only 2, 245 fall outside the interval of μ+3σ. In another word, 99.9153% of distances are less than μ+3σ. According to **Feng *et al.* (2021)**, points that fall out of μ+3σ are usually abnormal. Therefore, we eliminated that data. **Figure 4** is the AIS trajectory plot with the original data and cleansed data of Oct. 30th, 2020.

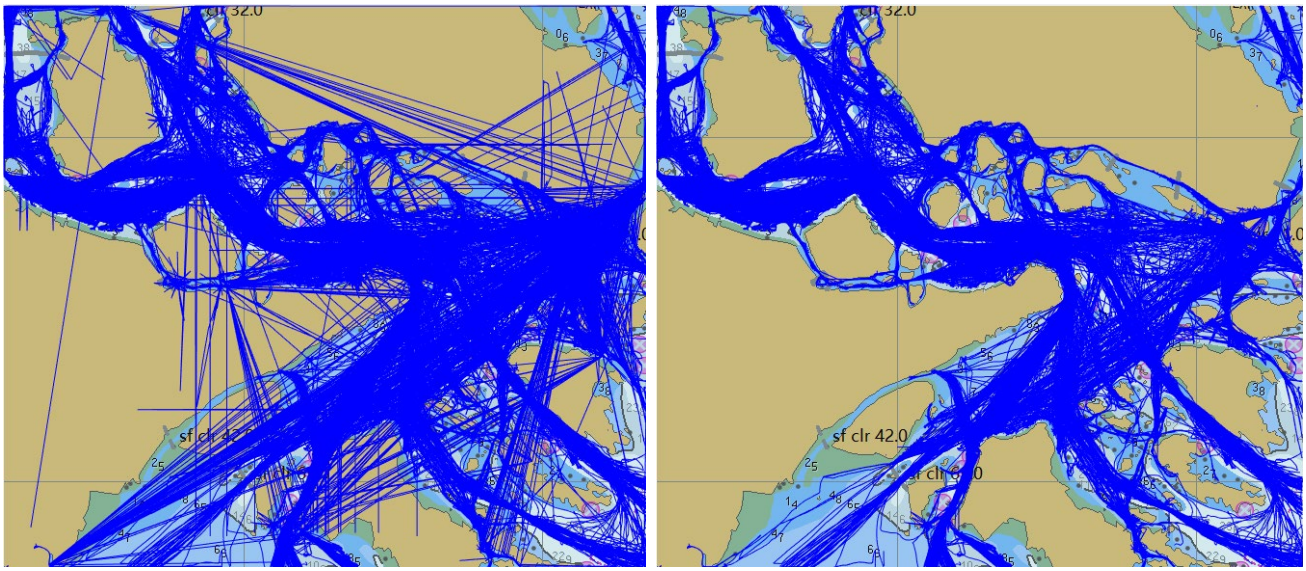

Figure 4. AIS trajectory plot in Ningbo-Zhou Port with original data (left) and with cleansed data of Oct. 30th, 2020 by **Feng *et al.* (2021)**'s algorithm (right).

3.3.3 Legs dividing

There are 28 legs including 18 TSSs, 8 PAs and 2 TWRs in Ningbo-Zhoushan Port. Therefore, we need to group those records obtained by area filtering in section 3.3.1 into 28 groups for further data mining. Firstly, to determine the boundaries of each leg. For example, Leg No.2 TSS is a quadrilateral (see **Figure 3**), and the geographic coordinates of the vertices are listed in **Table 1**. Then, the procedure is proceeded to check (record by record) whether a ship is fallen into this area or not as per Formula (15)-(18):

$$lat - lat_1 \geq \frac{lat_2 - lat_1}{lon_2 - lon_1}(lon - lon_1) \tag{15}$$

$$lat - lat_2 \geq \frac{lat_3 - lat_2}{lon_3 - lon_2}(lon - lon_2) \tag{16}$$

$$lat - lat_3 \geq \frac{lat_4 - lat_3}{lon_4 - lon_3}(lon - lon_3) \tag{17}$$

$$lat - lat_4 \geq \frac{lat_1 - lat_4}{lon_1 - lon_4}(lon - lon_4) \tag{18}$$

where *lat* and *lon* are the latitude and longitude values in the AIS records.

Table 1. The latitude and longitude coordinates of vertices of Leg No.2 TSS.

Vertices	Latitude (N)	Longitude (E)
(lon_1, lat_1)	29.90833333	122.17661111
(lon_2, lat_2)	29.86802778	122.19941667
(lon_3, lat_3)	29.86297222	122.18800000
(lon_4, lat_4)	29.90333333	122.16472222

Following this procedure, all the records of one-day's AIS data are categorized into 28 groups. Next, we further merge 92-day's data (i.e., from Oct. 1st to Dec. 31st, 2020) of each leg for information mining.

4. Results and discussion

4.1 Results

As the first step, following the methodology of information entropy described in Section 2.1, we calculate the COG entropy of different ships' routing legs in Oct.-Dec., 2020. The summary of COG entropy is tabulated in **Table 2** and **Figure 5** (left).

Then, K-means clustering introduced in Section 3.2 is applied to categorize the COG entropy. All the 28 legs in Ningbo-Zhoushan Port are grouped into 5 clusters according to the clustering results, of which Leg PA7, PA4 and PA2 are listed in Q1; Leg PA1, PA6, PA5, TSS16 and TSS17 belong to Q2; Leg TSS18, TSS4, PA3, TSS3 and PA0 are included in Q3; Leg TSS10, TSS12, TSS11, TSS8 and TSS7 fall into Q5; others reside in Q4 (see **Figure 5** (right) and **Table 3**).

Table 2. COG entropy of different ships' routing legs in Oct., Nov., Dec. and Oct.-Dec., 2020.

Area	Entropy in Oct.-Dec.	Entropy in Oct.	Entropy in Nov.	Entropy in Dec.
TSS1	9.437006	9.495599	9.485184	9.293706
TSS2	9.445622	9.485428	9.427324	9.383306
TSS3	9.745884	9.726616	9.771963	9.727595
TSS4	9.962802	9.963339	9.935324	9.955288
TSS5	9.332828	9.352240	9.299271	9.314235
TSS6	9.547908	9.487102	9.495667	9.633948
TSS7	8.460033	8.455798	8.468596	8.428346
TSS8	9.131970	9.176956	9.123440	8.983526
TSS9	9.376190	9.387928	9.363981	9.348116
TSS10	9.164479	9.114813	9.157449	9.179990
TSS11	9.131607	9.096462	9.126073	9.136520
TSS12	9.125986	9.071209	9.137402	9.125853
TSS13	9.446923	9.460766	9.397363	9.394001
TSS14	9.358533	9.291016	9.278039	9.416255
TSS15	9.499747	9.484828	9.411135	9.504675
TSS16	10.23805	10.20241	10.20021	10.26575
TSS17	10.07689	10.10663	10.06030	10.01553
TSS18	9.958244	9.793181	9.954311	9.991887
PA0	9.763028	9.842374	9.699753	9.649678
PA1	10.59500	10.63338	10.58827	10.52276
PA2	10.86911	10.85868	10.84576	10.88601
PA3	9.888699	9.896444	9.891775	9.843703
PA4	11.10966	11.10905	11.09417	11.11066
PA5	10.38200	10.21403	10.02098	10.57486
PA6	10.55678	10.52240	10.50447	10.61616
PA7	11.21362	11.21288	11.20715	11.17656
TWR1	9.521161	9.535013	9.537964	9.476346
TWR2	9.454826	9.458937	9.395435	9.429724

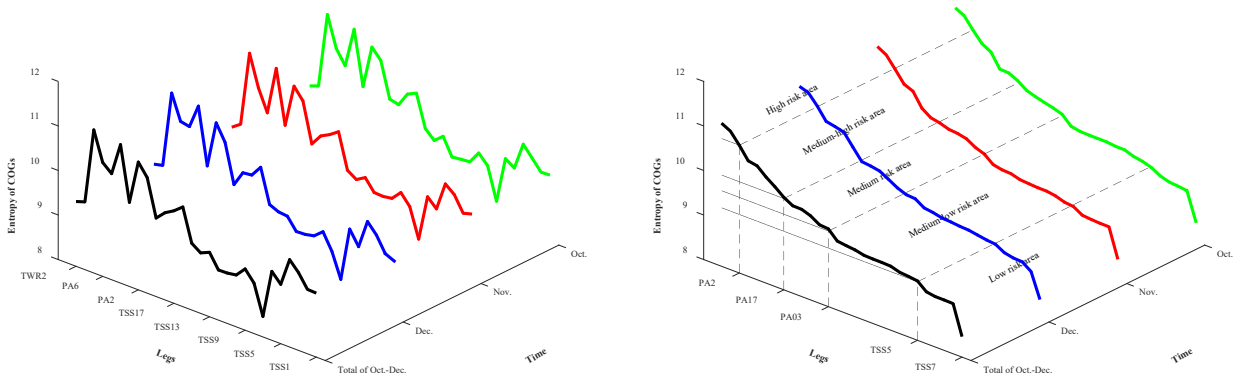

Figure 5. COG entropy (left) and K-means clustering (right) for COG entropy of different ships' routing legs in Oct., Nov., Dec. and Oct.-Dec., 2020.

Table 3. COG entropy and K-means clustering for COG entropy of different ships' routing legs in Oct., Nov., Dec. and Oct.-Dec., 2020.

Clusters	legs	COG entropy				Risk level
		Oct.-Dec.	Oct.	Nov.	Dec.	
Q1	PA7	11.20193	11.21288	11.20715	11.17656	High risk
	PA4	11.10607	11.10905	11.09417	11.11066	
	PA2	10.86964	10.85868	10.84576	10.88601	
Q2	PA1	10.56603	10.63338	10.58827	10.52276	Medium-high risk
	PA6	10.56602	10.52240	10.50447	10.61616	
	PA5	10.40607	10.21403	10.02098	10.57486	
	TSS16	10.24392	10.20241	10.20021	10.26575	
	TSS17	10.04940	10.10663	10.06030	10.01553	
Q3	TSS18	10.00202	9.793181	9.954311	9.991887	Medium risk
	TSS4	9.953289	9.963339	9.935324	9.955288	
	PA3	9.877233	9.896444	9.891775	9.843703	
	TSS3	9.752939	9.726616	9.771963	9.727595	
	PA0	9.698074	9.842374	9.699753	9.649678	
Q4	TSS6	9.573951	9.487102	9.495667	9.633948	Medium-low risk
	TWR1	9.510606	9.535013	9.537964	9.476346	
	TSS15	9.482051	9.484828	9.411135	9.504675	
	TWR2	9.431629	9.458937	9.395435	9.429724	
	TSS13	9.418291	9.460766	9.397363	9.394001	
	TSS2	9.415489	9.485428	9.427324	9.383306	
	TSS1	9.399122	9.495599	9.485184	9.293706	
	TSS14	9.371948	9.291016	9.278039	9.416255	
	TSS9	9.363394	9.387928	9.363981	9.348116	
	TSS5	9.314652	9.352240	9.299271	9.314235	
Q5	TSS10	9.178359	9.114813	9.157449	9.179990	Low
	TSS12	9.142813	9.071209	9.137402	9.125853	
	TSS11	9.139480	9.096462	9.126073	9.136520	
	TSS8	9.081009	9.176956	9.123440	8.983526	
	TSS7	8.456011	8.455798	8.468596	8.428346	

4.2. Discussions

4.2.1 Robustness analysis

To verify the robustness of the employed data, the COG entropy of Oct., Nov. and Dec. 2020 is also calculated (see **Table 2** and **Figure 5** (left)). The correlation coefficient between the data of Oct., Nov., Dec. and the data of Oct.-Dec. are 0.992822263, 0.992584776 and 0.997400626 respectively. In **Figure 5** (left)), high consistency is indicated by the curves of Oct., Nov., Dec and Oct.-Dec, which implies that the traffic flow in Ningbo-Zhoushan Port is temporally stable and regular.

The COG entropy of Oct., Nov. and Dec. are also individually processed with the K-means clustering algorithm. The clustering results of Oct., Nov. and Dec. are the same as those of Oct.-Dec. (see **Table 3** and **Figure 5** (right))., i.e., Leg PA7, PA4 and PA2 are listed in Q1; Leg PA1, PA6, PA5, TSS16 and TSS17 belong to Q2; Leg TSS18, TSS4, PA3, TSS3 and PA0 are included in Q3; Leg TSS10, TSS12, TSS11, TSS8 and TSS7 fall into Q5; others reside in Q4. This means that the clustering of COG entropy for the routing legs of Ningbo-Zhoushan Port is also temporally stable and regular.

4.2.2 Characteristics analysis

According to the clustering results, TSS10, TSS12, TSS11, TSS8 and TSS7 are listed in Q5. **Figure 6** reports the COG probability distribution of each traffic lane in Q5, and a bimodal normal distribution is observed in each of these scenarios (Remarks: in navigation practice, 0 ° is equal to 360 °, therefore, figures of TSS8 and TSS7 are also bimodal normal distributions). This means that the navigation order in these traffic lanes was very good, almost all vessels were proceeding in the general directions of traffic flow for those lanes. Therefore, we could infer that the ship collision risk in Q5 is low.

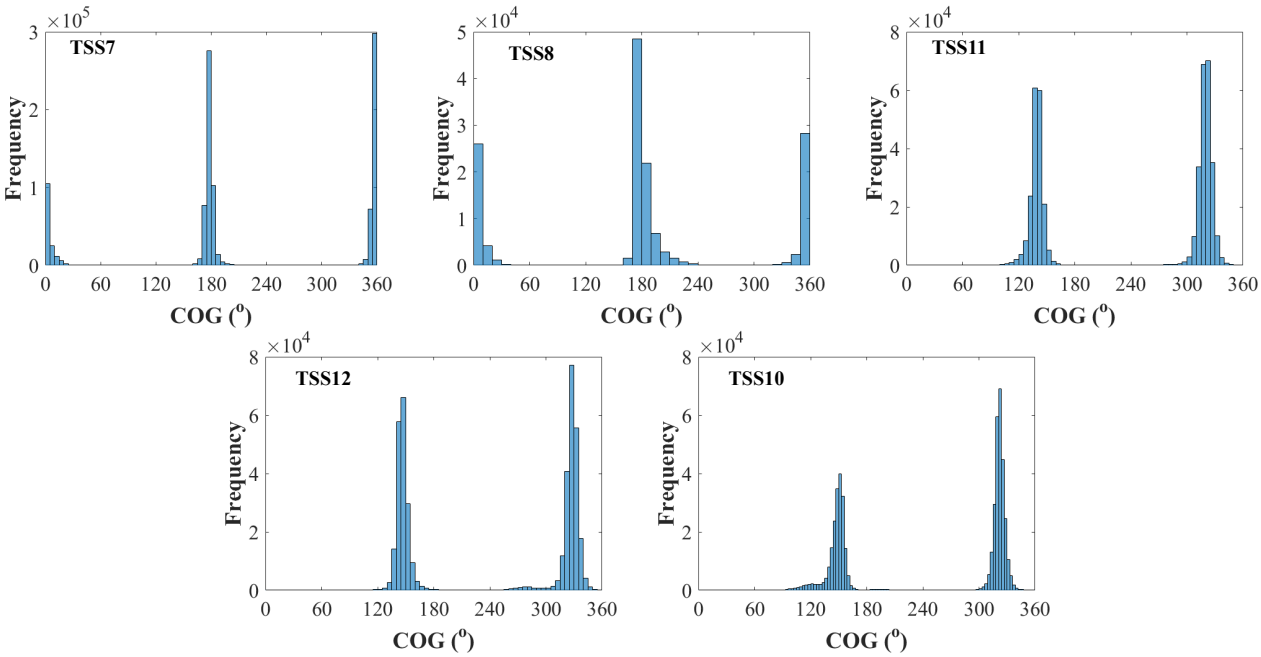

Figure 6. COG probability distributions in Q5.

By K-means clustering, TTS5, TSS9, TSS14, TSS1, TSS2, TSS13, TWR2, TSS15, TWR1 and TSS6 are grouped into Q4. **Figure 7** explicates the COG probability distributions of each traffic lane in Q4, and an approximate bimodal normal distribution with a little noise is observed in each of these scenarios (Remarks: in navigation practice, 0 ° is equal to 360 °, therefore, TSS1 is also an approximate bimodal normal distribution with a little noise). This implies that the navigation order in these traffic

lanes was good, most vessels were proceeding in the general directions of traffic flow for those lanes; however, there are a small number of ships crossing the traffic separation zone or separation line. The ship collision risk was increasing but at a medium-low level.

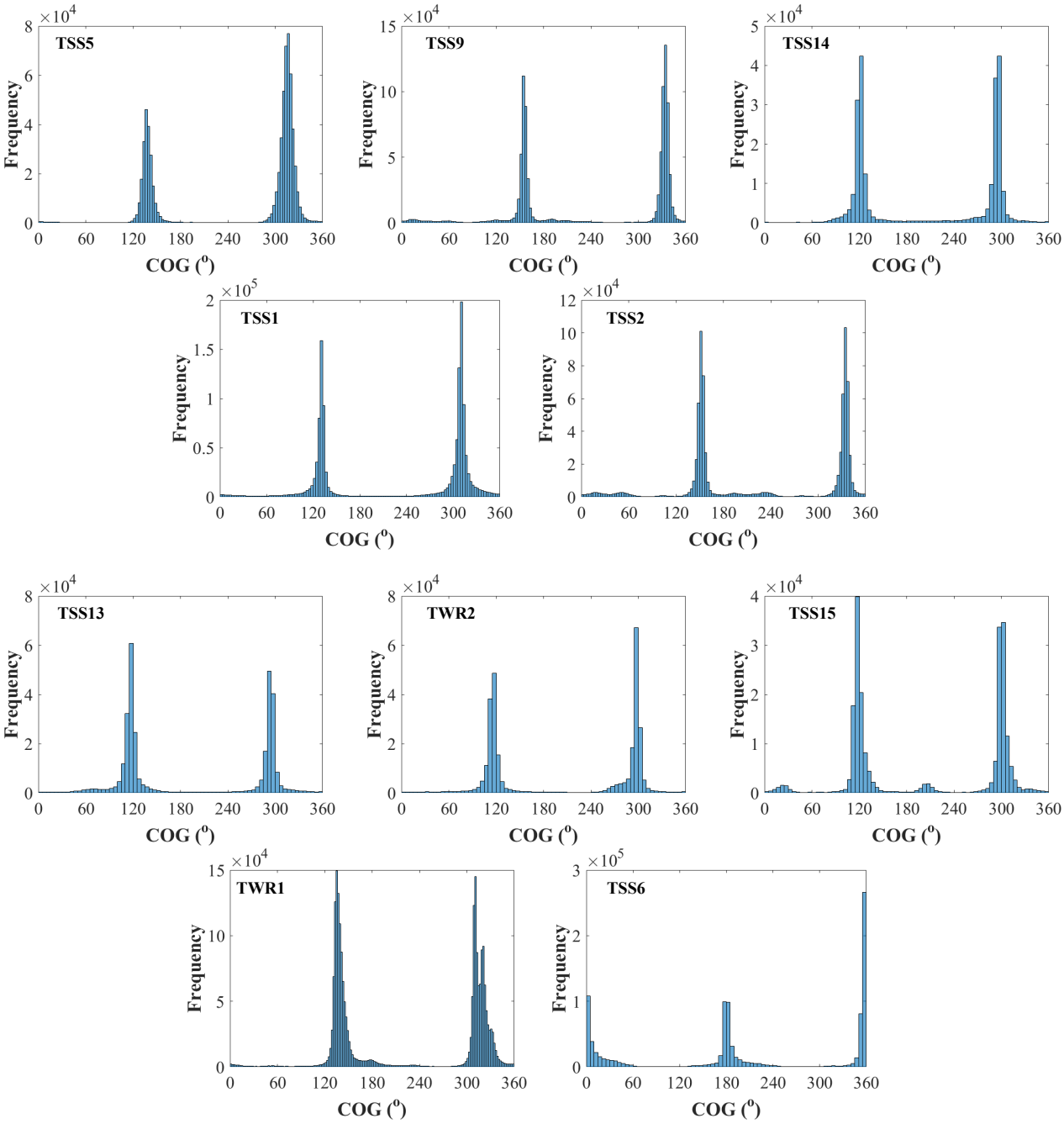

Figure 7. COG probability distributions in Q4.

The clustering results categorize PA0, TSS3, PA3, TSS4 and TSS18 into Q3. **Figure 8** explicates the COG probability distributions of each traffic lane in Q3, and an approximate bimodal normal distribution with significant noise is observed in each of these scenarios. This shows that the navigation order in these traffic lanes was fair, most vessels were proceeding in the general directions of traffic flow for those lanes; however, there are numerous ships crossing the traffic separation zone (for instance in PA3 and TSS18) or many vessels were deviating from the general directions of the traffic

flow by a considerable amount (for instance PA0, TSS3 and TSS4). As a consequence, more crossing situations occurred, and the risk of ship collision increased to a medium level accordingly.

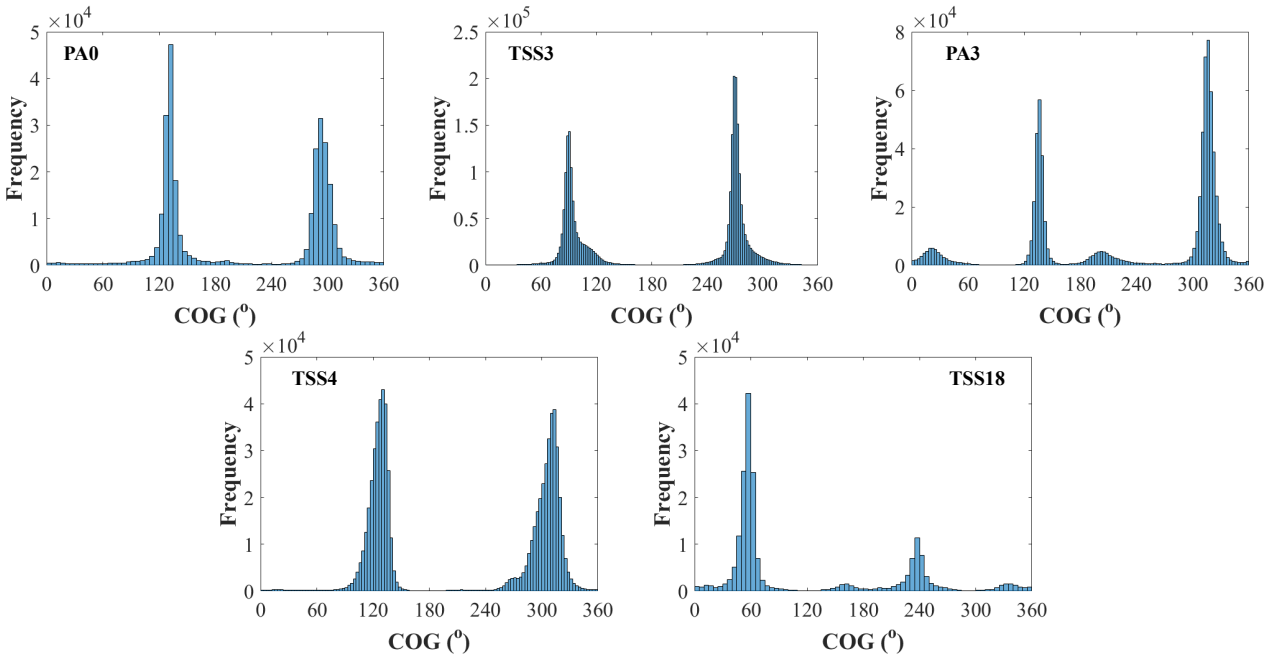

Figure 8. COG probability distributions in Q3.

The K-means clusters TSS17, TSS16, PA5, PA6 and PA1 into Q2. **Figure 9** indicates the COG probability distributions of each traffic lane in Q2, and an approximate multimodal normal distribution with significant noise is observed in each of these scenarios. This shows that the navigation order in these traffic lanes was complex. In these areas, ships were sailing in various directions. Consequently, crossing situations occurred frequently, therefore, the ship collision risk was at a medium-high level.

In Q1 (i.e., PA2, PA4 and PA7, see **Figure 10**), the confusion of course was more obvious, and the navigation order was more complex with a high risk of collision.

In this way, the calculation of COG entropy divides the ships' routeing areas of Ningbo-Zhoushan Port into different risk levels (see **Figure 11**). Of them, PA7, PA4 and PA2 are in the high-risk level; PA1, PA6, PA5, TSS 16 and TSS17 are medium-high risky. The precautionary areas where vessels meet and cross are always higher risk areas. Therefore, the results of K-means clustering match the actual situation. And then, why TSS 16 and TSS17 are marked as medium-high risky? We plot the AIS trajectory of Oct. 30th, 2020 near TSS16 and TSS17. From **Figure 12** (left) it could be seen that a dense traffic flow that should pass through PA6 actually crossed the west area of TTS 17. In addition, there existed a number of shipyards to the south of TTS 17. As vessels entered and departed, they would cross the traffic separation line. Therefore, the COG probability distribution of TSS 17 presented an approximate four modal normal distribution with significant noise (see **Figure 9** (TSS17)) and a high entropy value. From **Figure 12** (left) it could be seen that a dense traffic flow that should pass through PA7 actually crossed the west area of TTS 16. Additionally, there was an inner-islands fixed voyage that crossed the TTS 17. Therefore, the COG probability distribution of TSS 16 was similar to an approximate six modal normal distribution with significant noise (see **Figure 9** (TSS16))

and a high entropy value.

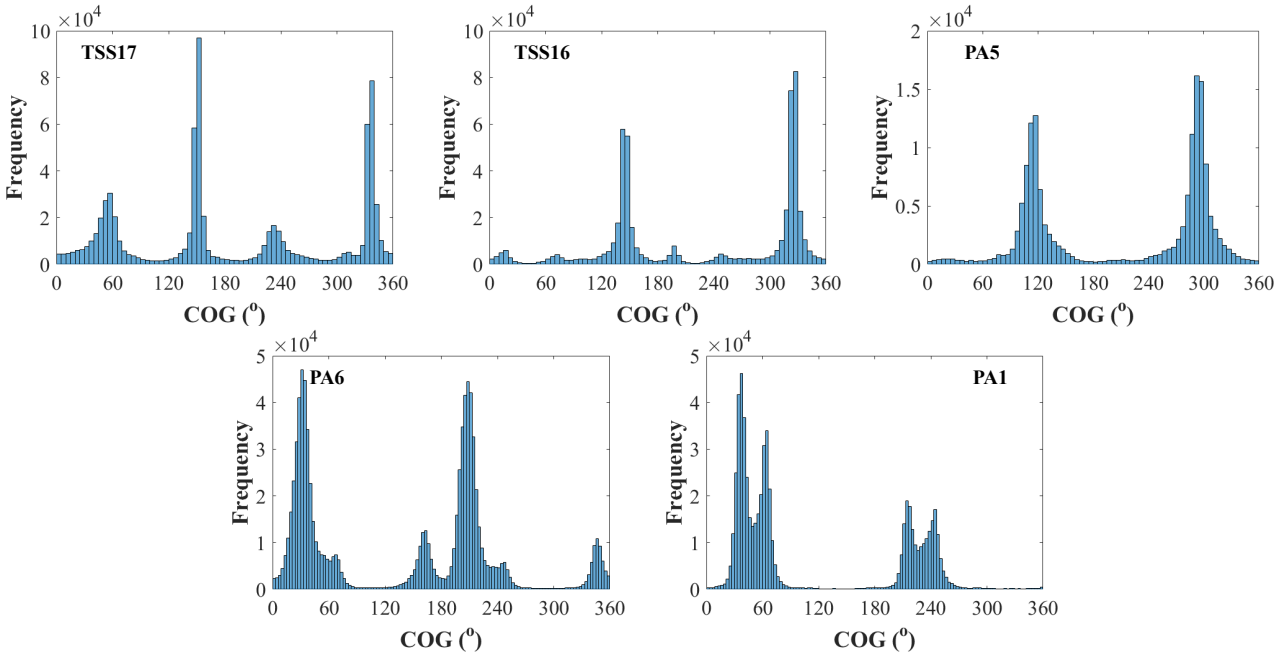

Figure 9. COG probability distributions in Q2.

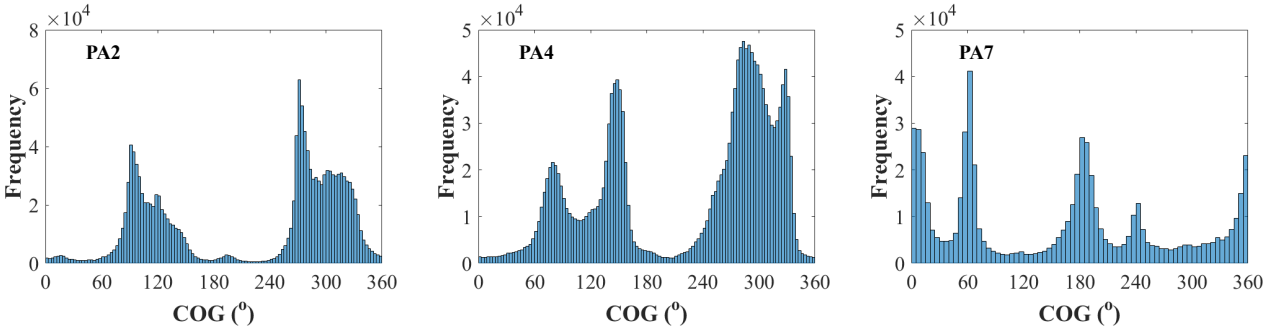

Figure 10. COG probability distributions in Q1.

It should be emphasized that PA0 and PA3 are all busy precautionary areas, however, their entropy levels are not very high, and they are included in medium-risk groups. We plot the AIS trajectory of Oct. 30th, 2020 near PA3 (see **Figure 12** (central)) and PA0 (see **Figure 12** (right)). From **Figure 12** (central) it could be noticed a fixed traffic flow that passed through PA3 (actually a scheduled passenger voyage between Ningbo and Zhoushan), and there were no other crossing routes. This made the navigation order in PA3 relatively simple, therefore, the entropy of PA3 is not very high. PA0 is the most important entrance of Ningbo-Zhoushan Port, this area is a key point that Maritime Safety Administration. From **Figure 12** (left) it could be seen that most vessels were sailing along the Deep-water route outside of Xiazhi Men. This made the navigation order in PA0 also relatively simple, therefore, the entropy of PA0 is not very high.

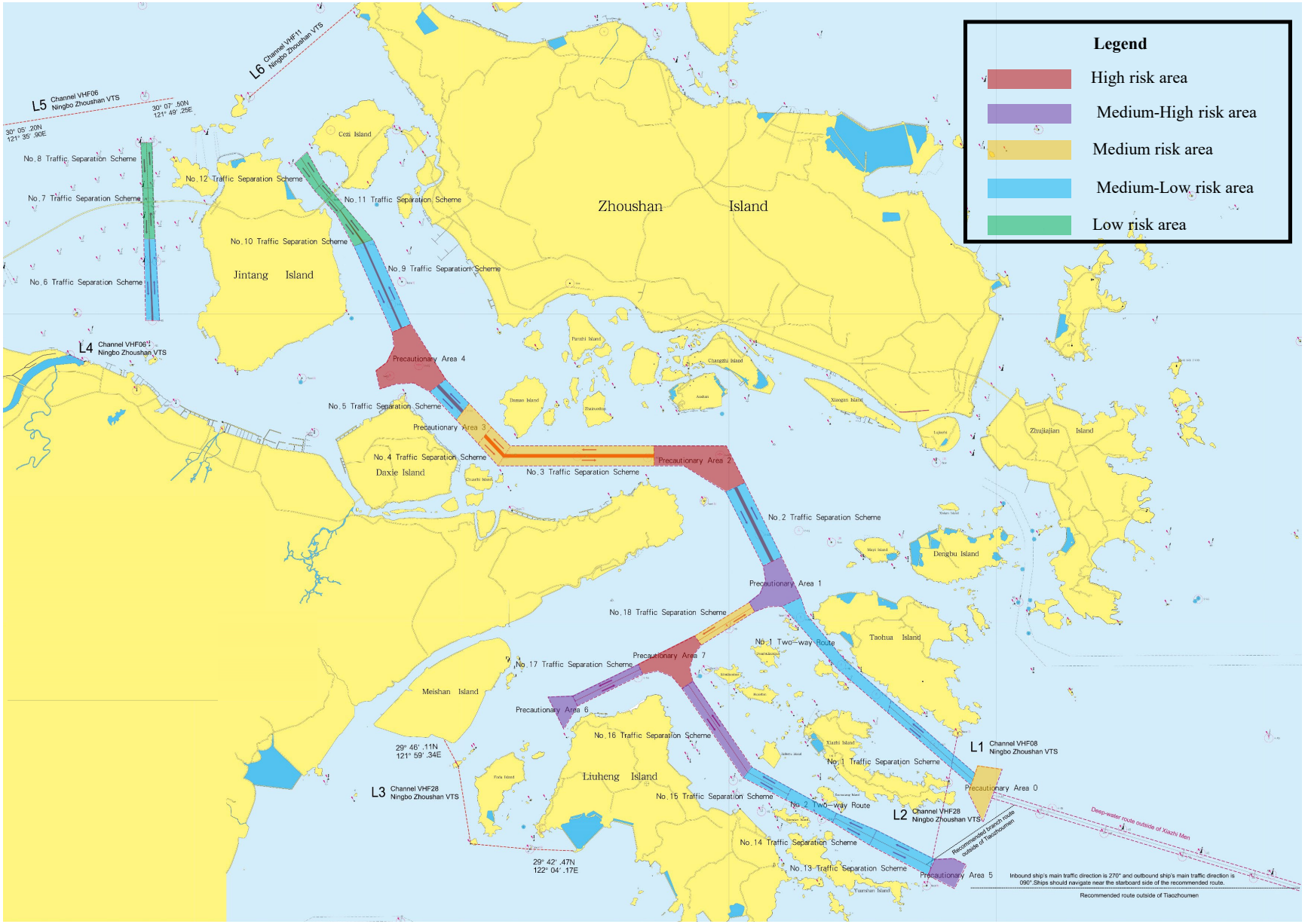

Figure 11. Collision risk division for the ships' routeing areas in Ningbo-Zhoushan Port.

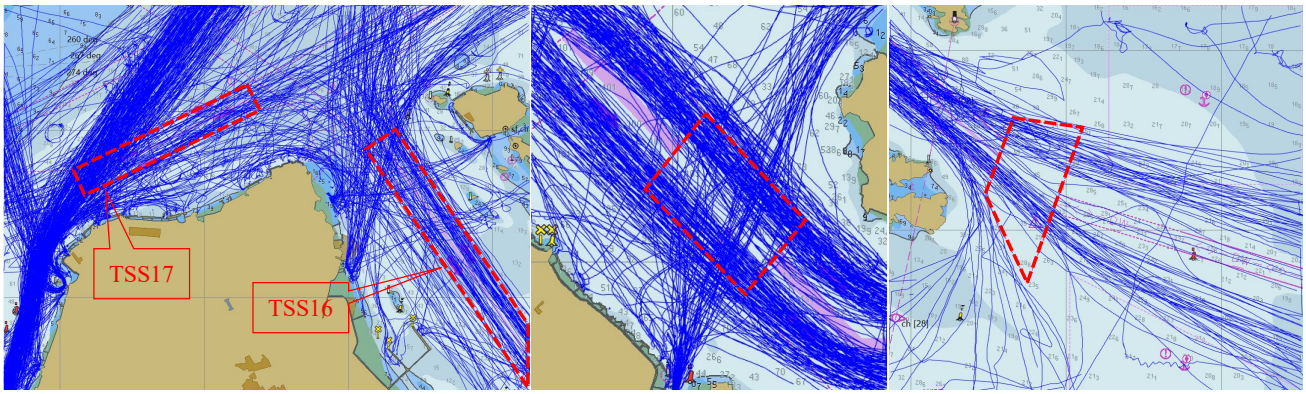

Figure 12. AIS trajectory plot in TSS16, TSS17 (left), PA3 (central) and PA0 (right) of Ningbo-Zhoushan Port with the data of Oct. 30th, 2020.

5. Conclusions

Ships' routeing was adopted to guide marine traffic flow and reduce the risk of collision in the crowded waterways. With the expansion of the world's fleet, marine traffic in bottleneck and chokepoint areas became more and more complex creating serious challenges for navigational safety. Therefore, quantitative collision risk assessment is significantly important for the ships' routeing areas. In this contribution, information entropy is proposed to develop indices to quantitatively measure the risk of collision. We use Ningbo-Zhoushan Port as the case study. The AIS data of Ningbo-Zhoushan Port is used to offer the course over ground (COG) information after data-cleansing processing and then the information entropy of COGs in each leg of the ships' routeing in Ningbo-Zhoushan Port is calculated. Next, the K-means clustering is applied to group the SOGs entropy. According to the clustering results, PA7, PA4 and PA2 are categorized as the most high-risk legs in the ships' routeing areas of Ningbo-Zhoushan Port. TSS17, TSS16, PA5, PA6 and PA1 follow with medium-high risk. Therefore, visiting shipmasters should proceed with caution in these areas. Priority should be given to collision risk prevention in these areas. This study also finds that numerous ships cross the traffic separation zone or separation line, which results in higher entropy in those areas and greater potential for ship collision. The safety level would be significantly improved if the rules on navigation are abided according to the entropy reduction principle.

Ships' routeing areas in Ningbo-Zhoushan Port also include several inshore traffic zones and the Deep-water route outside of Xiazhi Men. Considering the inshore traffic zones are usually used by small vessels, and the Deep-water route is specially designed for ultra-large vessels with draught more than 22.5m (few in number), these areas were not covered in this study.

Declaration of competing interest

None.

Reference

Adamidis, F.K., Mantouka, E.G., Vlahogianni, E.I., 2020. Effects of controlling aggressive driving behavior on network-wide traffic flow and emissions. International Journal of Transportation Science and Technology 9 (3), 263-276.
Aydogdu, Y.V., Yurtoren, C., Park, J.-S., Park, Y.-S., 2012. A study on local traffic management to improve marine traffic safety in the istanbul strait. Journal of Navigation 65 (1), 99-112.
Bolshakova, N., Azuaje, F., 2003. Cluster validation techniques for genome expression data. Signal Processing 83 (4), 825-833.
Chai, T., Xue, H., 2021. A study on ship collision conflict prediction in the taiwan strait using the emd-based lssvm method. PloS one 16 (5), e0250948.
Cockcroft, A.N., 2004. The dover strait traffic separation scheme. Journal of Navigation 57 (1), 161-161.
Cucinotta, F., Guglielmino, E., Sfravara, F., 2017. Frequency of ship collisions in the strait of messina through regulatory and environmental constraints assessment. journal of navigation 70 (5), 1-21.
Feng, H.-X., Mujal-Colilles, A., Yang, Z.-Z., 2021. A method for pre-processing ais abnormal data using distance Navigation of China, in press.
Fujii, Y., Tanaka, K., 2010. Traffic capacity. Journal of Navigation 24 (4), 543-552.
Gao, M., Shi, G.-Y., 2020. Ship collision avoidance anthropomorphic decision-making for structured learning based on ais with seq-cgan. Ocean Engineering 217, 107922.
Goerlandt, F., Kujala, P., 2011. Traffic simulation based ship collision probability modeling. Reliability Engineering & System Safety 96 (1), 91-107.
Hamerly, G., Elkan, C., 2002. Alternatives to the k-means algorithm that find better clusterings. Proceedings of the eleventh international conference on Information and knowledge management. Association for Computing Machinery, McLean, Virginia, USA, pp. 600–607.
Hartigan, J.A., Wong, M.A., 1979. Algorithm as 136: A k-means clustering algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics) 28 (1), 100-108.
Imo, 1972. Convention on the international regulations for preventing collisions at sea, 1972 (colregs).
Kang, L., Meng, Q., Liu, Q., 2018. Fundamental diagram of ship traffic in the singapore strait. Ocean Engineering 147, 340-354.
Kang, L., Meng, Q., Zhou, C., Gao, S., 2019. How do ships pass through l-shaped turnings in the singapore strait? Ocean Engineering 182, 329-342.
Kujala, P., Hänninen, M., Arola, T., Ylitalo, J., 2009. Analysis of the marine traffic safety in the gulf of finland. reliability engineering & system safety 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. Safety Science 129, 104798.
Li, S., Zhou, J., Zhang, Y., 2015. Research of vessel traffic safety in ship routeing precautionary areas based on navigational traffic conflict technique. Journal of Navigation 68 (3), 589-601.
Liu, C., Liu, J., Zhou, X., Zhao, Z., Wan, C., Liu, Z., 2020a. Ais data-driven approach to estimate navigable capacity of busy waterways focusing on ships entering and leaving port. Ocean Engineering 218, 108215.
Liu, Z., Wu, Z., Zheng, Z., 2020b. A molecular dynamics approach for modeling the geographical distribution of ship collision risk. Ocean Engineering 217, 107991.
Macqueen, J.B., 1967. Some methods for classification and analysis of multivariate observations. Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press, pp. 281-297.
Núñez, J., Cincotta, P., Wachlin, F., 1996. Information entropy an indicator of chaos. Celestial Mechanics & Dynamical Astronomy - CELEST MECH DYNAM ASTRON 64, 43-53.
Pietrzykowski, Z., 2008. Ship's fuzzy domain – a criterion for navigational safety in narrow fairways. Journal of Navigation 61 (3), 499-514.
Qu, X., Meng, Q., Suyi, L., 2011. Ship collision risk assessment for the singapore strait. Accident Analysis & Prevention 43 (6), 2030-2036.
Rawson, A., Brito, M., 2021. A critique of the use of domain analysis for spatial collision risk assessment. Ocean Engineering 219, 108259.
Shannon, C.E., 1948. A mathematical theory of communication. The Bell System Technical Journal 27 (3), 379-423.
Silveira, P.a.M., Teixeira, A.P., Soares, C.G., 2013. Use of ais data to characterise marine traffic patterns and ship collision risk off the coast of portugal. journal of navigation 66 (6), 879-898.

Squire, C.D., 2003. The hazards of navigating the dover strait (pas-de-calais) traffic separation scheme. Journal of Navigation 56 (2), 195-210.
Szlapczynski, R., 2006. A unified measure of collision risk derived from the concept of a ship domain. Journal of Navigation 59 (3), 477-490.
Wang, N., Chang, D., Yuan, J., Shi, X., Bai, X., 2020. How to maintain the safety level with the increasing capacity of the fairway: A case study of the yangtze estuary deepwater channel. Ocean Engineering 216, 108122.
Xin, X., Liu, K., Yang, X., Yuan, Z., Zhang, J., 2019. A simulation model for ship navigation in the "xiazhimen" waterway based on statistical analysis of ais data. Ocean Engineering 180, 279-289.
Yu, H., Fang, Z., Murray, A.T., Peng, G., 2021. A direction-constrained space-time prism-based approach for quantifying possible multi-ship collision risks. IEEE Transactions on Intelligent Transportation Systems 22 (1), 131-141.
Zhang, J., He, A., Fan, C., Yan, X., Soares, C.G., 2020a. Quantitative analysis on risk influencing factors in the jiangsu segment of the yangtze river. Risk Analysis (in press), https://doi.org/10.1111/risa.13662.
Zhang, L., Meng, Q., 2019. Probabilistic ship domain with applications to ship collision risk assessment. Ocean Engineering 186, 106130.
Zhang, L., Meng, Q., Fang Fwa, T., 2019. Big ais data based spatial-temporal analyses of ship traffic in singapore port waters. Transportation Research Part E: Logistics and Transportation Review 129, 287-304.
Zhang, W., Feng, X., Goerlandt, F., Liu, Q., 2020b. Towards a convolutional neural network model for classifying regional ship collision risk levels for waterway risk analysis. Reliability Engineering & System Safety 204, 107127.
Zhang, X., Wang, C., Jiang, L., An, L., Yang, R., 2021. Collision-avoidance navigation systems for maritime autonomous surface ships: A state of the art survey. Ocean Engineering 235, 109380.
Zhao, L., Shi, G., Yang, J., 2018. Ship trajectories pre-processing based on ais data. Journal of Navigation 71 (5), 1210-1230.