Zhang, Mingyang; Montewka, Jakub; Manderbacka, Teemu; Kujala, Pentti; Hirdaris, Spyros

A Big Data Analytics Method for the Evaluation of Ship - Ship Collision Risk reflecting Hydrometeorological Conditions

Published in:
Reliability Engineering and System Safety

DOI:
10.1016/j.ress.2021.107674

Published: 01/09/2021

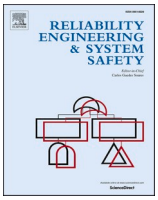
Document Version
Publisher's PDF, also known as Version of record

Reliability Engineering and System Safety 213 (2021) 107674

Contents lists available at ScienceDirect

Reliability Engineering and System Safety

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



A Big Data Analytics Method for the Evaluation of Ship - Ship Collision Risk reflecting Hydrometeorological Conditions

Mingyang Zhang [a], Jakub Montewka [a,b], Teemu Manderbacka [c], Pentti Kujala [a], Spyros Hirdaris [a,*]

[a] Department of Mechanical Engineering, Marine Technology Group, Aalto University, Espoo, Finland
[b] Department of Transport and Logistics, Gdynia Maritime University, Gdynia, Poland
[c] NAPA Ltd., Helsinki, Finland

ARTICLE INFO

Keywords:
Ship safety
Maritime operations
Collisions
Big data analytics
Machine learning
Gulf of Finland

ABSTRACT

This paper presents a big data analytics method for the evaluation of ship-ship collision risk in real operational conditions. The approach makes use of big data from Automatic Identification System (AIS) and nowcast data corresponding to time-dependent traffic situations and hydro-meteorological conditions respectively. An Avoidance Behavior-based Collision Detection Model (ABCD-M) is introduced to identify potential collision scenarios and Collision Risk Indices (CRIs) are quantified when evasive actions are taken for each detected collision scenario in various voyages. The method is applied on Ro-Pax ships operating over 13 months of the ice-free period in the Gulf of Finland. Results indicate that collision risk estimates may be extremely diverse among voyages, and in 97.5% of potential collision scenarios the evasive actions are triggered only when risk is at 45% or more of its maximum value. The overall CRI for ships operating over the given area tends to be lower for adverse hydro-meteorological conditions. It is therefore concluded that the proposed method may assist with the (1) identification of critical scenarios in various voyages not currently accounted for by existing accident databases, (2) definition of commonly agreed risk criteria to set off alarms, (3) the estimation of risk profile over the life cycle of fleet operations.

1. Introduction

Ship collisions and groundings are the most frequent maritime traffic accidents globally [38]. They often result in unwanted and devastating consequences such as oil spills, severe ship flooding or loss of human life [33]. Their effect is especially critical for passenger shipping operations [52]. To mitigate risks associated with such events it is necessary to develop maritime risk management tools.

To date, research on risk management of ship collisions focuses on (a) semi - empirical and (b) probabilistic risk analysis models. The former help estimate the probability and consequence of accidents on the basis of accident data statistics and expert judgment (e.g., [7,11,15]). Common modelling tools include: Fault Tree Analysis (e.g., [2,47,48,87]); Bayesian Networks (e.g., [13,14,26,37,54,55,89,100]); Hybrid Causal Logic (e.g., [60,71,72]); Event Trees (e.g., [9,33]) and traffic simulation methods (e.g., [3,20,28,50,51,70]). These approaches are useful in terms of assessing collision risk in a specific sea area. Notwithstanding, they fail to suggest reliable risk mitigation measures during shipping operations [15,21,59]. This is because it is challenging to provide a convincing justification for Risk Control Options (RCOs) in complex traffic situations pertaining to real hydro-meteorological conditions (e.g., [22,23,25,76,77,81]).

Research in probabilistic risk analysis may help to overcome problems associated with traffic complexity by utilizing openly available big data (e.g., Automatic Identification System data – AIS; Gridded Bathymetry data – GEBCO, etc.). The algorithms or models adopted are known as: 'Vessel Conflict Ranking Operat or' (e.g., [16,82,88]), 'Ship Safety Domain' (e.g., [67,83,85] and [68]), 'Velocity Obstacle' (e.g., [10,12,32,99]), 'DCPA and TCPA' (e.g., [1,6,44,66,94]). Nevertheless, similarly to empirical methods they may lead to underestimation of accidental risk indices as they do not account for real environmental conditions or traffic uncertainty (e.g., [36,39,53,61]). Furthermore, the difference in complex traffic scenarios in various voyages is often underestimated in the existing methods. To explore collision risk in more detail, the ship trajectories should be grouped using similarity measurement at first.

Ship trajectories data streams incorporate multiple parameters

* Corresponding author: Otakaari 4, 02150, Koneteknikka 1, Espoo, Finland; Tel: +358-50-349-9903
E-mail address: spyros.hirdaris@aalto.fi (S. Hirdaris).

https://doi.org/10.1016/j.ress.2021.107674
Received 26 November 2020; Received in revised form 16 March 2021; Accepted 5 April 2021
Available online 10 April 2021

Nomenclature

Variable	Definition
ABCD-M	Avoidance Behavior-based Collision Detection Model
ARPA	Automatic Radar Plotting Aids https://www.wartsila.com/encyclopedia/term/automatic-radar-plotting-aids-(arpa)
β	Relative bearing angle
C	Cluster
CRI	Collision Risk Index
cog	Course Over Ground
C_r	The relative angle
COLREGs	Convention on the International Regulations for Preventing Collisions at Sea
$d(p_j,p_{j+1})Dist_{ij}$	The distance between the point $p_j$andp_{j+1} obtained from AIS
$d_{\min}$, d_2	The minimum distance between ship trajectories
d_{CPA},t_{CPA}	Distance/ time to the Closest Point of Approach
d_{ij}, D	Distance between two ships
d_1	The minimum safe meeting distance
θ	The course of the striking ship and struck ship
E	The performance of ship trajectories clustering
ε	A spatial distance threshold to delimit the neighborhood of a ship trajectory
$h(Tr_i, Tr_{i+y})$	The Hausdorff distance between two ship trajectories
Hydro	The hydro-meteorological condition
JCOMM	Joint Technical Commission for Oceanography and Marine Meteorology
K	Number of clusters using K-means clustering
k,t,n,m	The timestamp
(lon_1, lat_1)	Longitude and latitude of the departure port
(lon_n, lat_n)	Longitude and latitude of the destination port
L_i L_j	Ship length
MinLns	The minimum number of ship trajectories required to form a dense cluster
N	The maximum number of iterations
NOAA	National Oceanic and Atmospheric Administration
p_j^i	A point of ship trajectory
p^{k+t}	Ship trajectories of the struck and striking ship when evasive actions are taken
p_1,p_n	The departures /destinations of ship trajectory
r, ROT	Rate of turn
RMS	Root Mean Square
sog	Speed Over Ground
SM	The ship trajectories similarity estimation matrix
S_{dd}^p,S_{dd}^{st},S_l	The similarity parameters of voyage details
S_h,S_{mpv},S_{sog},S_{cog}	The similarity parameters of navigation features
ST, Tr	Ship trajectory
TC	Ture Course
TS, OS	Target ship (Striking ship), Own ship (Struck ship)
V	Ship speed
v_{ij},S_r	Relative speed
w	Weight coefficient
XOY	WGS 1984 Coordinate System
xoy	Coordinate system fixed to the struck ship

related to static voyage features (e.g., departures/destinations, voyage length) and dynamic navigation features (e.g., speed, course, motion parameter variation, and ship trajectory spatial distance). However, it may be challenging to handle all available information using the available labels (i.e., MMSI, IMO number, call signs) delivered from AIS data [90]. An alternative could be to use unsupervised machine learning theory and apply clustering analysis of big data analytics with the aim to classify complex traffic scenarios preferably in real hydro-meteorological conditions. Typical unsupervised machine learning methods, clustering algorithms can automatically cluster ship traffic data by similarity measurements. They can be classified into three groups, namely: (a) distance partition methods (e.g., K-means algorithm; see [8,90,95]); (b) hierarchy methods (e.g., Balanced Iterative Reducing and Clustering using Hierarchies – BIRCH algorithm; see [43,92]); (c) density methods (e.g., Density-Based Spatial Clustering of Applications with Noise – DBSCAN: see [63,93]). A suitable selection of one of those could help classify ship trajectories and detect anomalies based on the maneuvering behavior of ships under real operational conditions (e.g., [5,8,10,43,62,69,97]). Distance partition methods have been adopted in ship trajectories clustering due to their high-efficiency performance [90]. Hierarchical methods suffer from the fact that once merge or split is done, it is not reversible [43]. Density methods are of great representativeness owing to their superiority in clustering ship trajectories with arbitrary shapes [5,63]. To date, the mentioned algorithms have been successfully used to cluster simplistic ship trajectories in open seas. Nevertheless, they fail in restricted waters where operational paths are more complex (e.g., [5,43,97]). This is because it is challenging to handle all available information of complex ship trajectories delivered from AIS data using a single algorithm.

Therefore, it is desirable to develop a big data analytics method for evaluation of ship-ship collision risk in various voyages using now-cast data and AIS data, by recovering detailed time-dependent traffic situations and the hydro-meteorological conditions at the times. This would allow insight to be gained into collision risk reflecting real operational conditions, as well as exploring the time to trigger evasive actions in various voyages [53].

This paper introduces a data mining method for ship collision avoidance behavior. The method detects collision scenarios based on clustered ship trajectories encompassing AIS and hydro-meteorological big data streams at the time of collision avoidance maneuvers in various routes (see Section 2). Consequently, the time during evasive actions taken is analyzed using a multi-criteria-based CRI. The practical application of the approach is demonstrated by the use of data covering a 13-month ice-free period in the Gulf of Finland, considering all large RoRo/Passenger ships (RoPax) (46,124 GT > Gross tonnage > 10,000 GT; 218.8 m > Length > 120 m) as the struck ships (see Section 3). The paper concludes on the potential of the method to develop intelligent decision support systems to mitigate collision risk by inspecting traffic patterns in various voyages and ship- ship collision risk (see Section 4).

2. Methods

2.1. Machine learning methods

AIS is an automatic tracking system that may be used to identify and locate ships through data exchange with nearby ships, AIS base stations and satellites. The use of this system has been required by the International Maritime Organization (IMO) since 2004 and to date transponders have been installed in more than 400,000 ships. AIS big data streams contain multiple parameters related to static voyage features (e.g., departures /destinations, voyage length) and dynamic navigation features (e.g., speed, course, motion parameter variation, and ship trajectory spatial distance). Although IMO number/call signs can be used as labels to separate ship trajectories (STs) of various ships, existing methods do not offer automatic means for ship trajectories clustering in various voyages. This is because it is difficult to derive available labels to fully explore both static voyage and dynamic navigation features of STs in real environmental conditions and complex traffic scenarios. Thus,

Table 1
K-Means algorithm for STs clustering.

Algorithm 1:K-Means algorithm
Input: Dataset$D = \{x_1, x_2, ..., x_m\}$, clustering number K, the maximum number of Niterations **Output**: Clustering division $C = \{c_1, c_2, ..., c_k\}$ **Process:** 1. Select K trajectories as the center trajectories $\{\mu_1, \mu_2, ..., \mu_k\}$; 2. Initially cluster division $C_t = \{c_1, c_2, ..., c_k\}$; 3. For $n = 1, 2, ..., N$: 4. For $i = 1, 2, ..., m$: 5. Calculate distance between trajectory x_i and $\mu_j (j = 1, 2, ..., k) d_{ij} = \left\|x_i - \mu_j\right\|$; 6. Mark category as j corresponding the smallest d_{ij}; 7. End for 8. For $j = 1, 2..., K$: 9. Calculate the center trajectories based on new clustering result $\mu^j = \frac{1}{\|\mu^j\|} \sum x(x \in \mu^j)$ 10. End for 11. If the clustering result remains consistent: 12. Go to line 17; 13. Else: 14. Go to line 4; 15. End if 16. End for 17. Output $C = \{c_1, c_2, ..., c_k\}$. 18. End procedure

Table 2
DB-SCAN algorithm for STs clustering.

Algorithm 2:DB-SCAN algorithm
Input: Dataset$D = \{x_1, x_2, ..., x_m\}$ **Output:** Clustering division $C = \{c_1, c_2, ..., c_k\}$ **Process:** 1. Mark the D as unprocessed trajectories; 2. For i $= 1, 2, ..., m$; 3. Check the neighborhood$\varepsilon(x_i)$; 4. If the number of objects in $\varepsilon(x_i) \geq$ *MinLns*: 5. Mark x_i as core point and set up a new class c and add objects in $\varepsilon(x_i)$ to N; 6. For p in N: 7. Check the neighborhood$\varepsilon(p)$; 8. If the number of objects in $\varepsilon(p) \geq$*MinLns*; 9. Add objects not be classified in $\varepsilon(x_i)$ to N and add p to c; 10. Else: 11. Add p to c; 12. End if 13. End for 14. End if 15. If the number of objects in $\varepsilon(x_i) <$*MinLns*: 16. Mark x_i as boundary point or noise point; 17. End if 18. End for 19. Output$C = \{c_1, c_2, ..., c_k\}$. 20. End procedure

when using information directly from historical AIS data (i.e., MMSI, IMO number, call signs) ship voyages cannot be separated automatically. Big data clustering may be useful in terms of grouping STs by measuring the similarity between available data streams [63]. Clustering algorithms, as typical unsupervised machine learning methods, can automatically cluster ship trajectories through similarity measurements of ship trajectory feature. However, toward to massive and complex ship trajectories in restricted waters, they are difficult to be clustered in more detail using a unique algorithm. To evaluate ship-ship collision risk in various voyages associated hydro-meteorological data in time-dependent traffic scenarios, the ship trajectories of struck ships should be classified in more detail, according to the similarity in both static voyage features and dynamic navigation features. With the latter in mind in this work K-means and DB-SCAN are selected and employed to cluster STs. This is because the k-means algorithm is high-efficiency performance in clustering ship trajectories using static voyage features, and DB-SCAN is of great representativeness owing to their superiority in clustering ship trajectories using dynamic navigation features. Accordingly, the complex ship trajectories can be clustered in more detail combining K means and DB-SCAN.

2.1.1. K-means algorithm

K-means is a clustering algorithm that distance partitions data points into groups based on Euclidean Distances – e.g., [90] – as presented in Table 1. It is easy to understand, implement and can handle large datasets. It requires clear specification of the desired number of clusters, which is easy to determine based on static voyage features (e.g., departure and destination points, voyage length). However, it may be sensitive to the number of clusters and the presence of noise in big data streams (e.g., outlying points in the trajectories as explained in Section 2.2.1). In K-means, similarity denotes the degree of similar trajectories measured. Accordingly, two STs are similar if their departure, destination, and voyage length are similar. The K-means algorithm can be efficiently used to cluster trajectories of ships navigating in a specific voyage route (i.e., in between the same departure and destination points).

However, even though the ships navigate in a specific voyage route, dynamic navigation features (e.g., speed, course, motion parameter variation, and ship trajectory spatial distance) may be diverse. Clustering test shows that if we consider more than three parameters (departure and destination points, voyage length) for STs clustering, the performance of k-means is not worked well. This is because the K-means algorithm is difficult to handle all available information (both static voyage features and dynamic navigation features) of complex ship trajectories. Thus, dynamic navigation features also should be mined to explore the difference of ship trajectories using DB-SCAN following K-means.

2.1.2. DB-SCAN algorithm

In contrast to K-means method that applies to static points datasets DB-SCAN is an algorithm that helps to form data clusters based on regular and irregular dense data. Those data may be associated with dynamic navigation features following K - means clustering. But DB-SCAN algorithms may not work well with static voyage features (distance points datasets) of STs. This is reason why the both K-means algorithm and DB-SCAN algorithm are used to cluster. STs in the paper. In the process of DB-SCAN clustering, data are divided into three categories, namely: (a) core, (b) border, and (c) noise; the latter ones associated with low-density data streams [93]. The algorithm does not require specifying the number of clusters in advance, as presented in Table 2. STs are similar if their voyage/navigation features and spatial distance have similar data densities (see Section 2.2.1). So the DB-SCAN algorithm is employed to cluster STs with similar motion parameters in the same voyage route after K- means clustering, like speed, course, and their variations, as well as spatial trajectory distance between the same departure and destination points (See ST 3,4 and ST 5,6 after DB-SCAN clustering in Fig. 3).

2.2. Big data analytics framework

The collision risk evaluation framework (Fig. 1) comprises of three steps:

- **Step (i)** where STs are reconstructed using AIS data that contain static voyage and dynamic navigation details. The process is used to cluster ship trajectories of the struck ships. Static voyage details (departure and destination points, voyage length) are illustrated to cluster ship trajectories using K-means if their departure, destination and voyage length are similar. Then, DB-SCAN is used to re-cluster results based on dynamic navigation features (speed, course, motion parameter variation, and spatial ship trajectory distances). STs

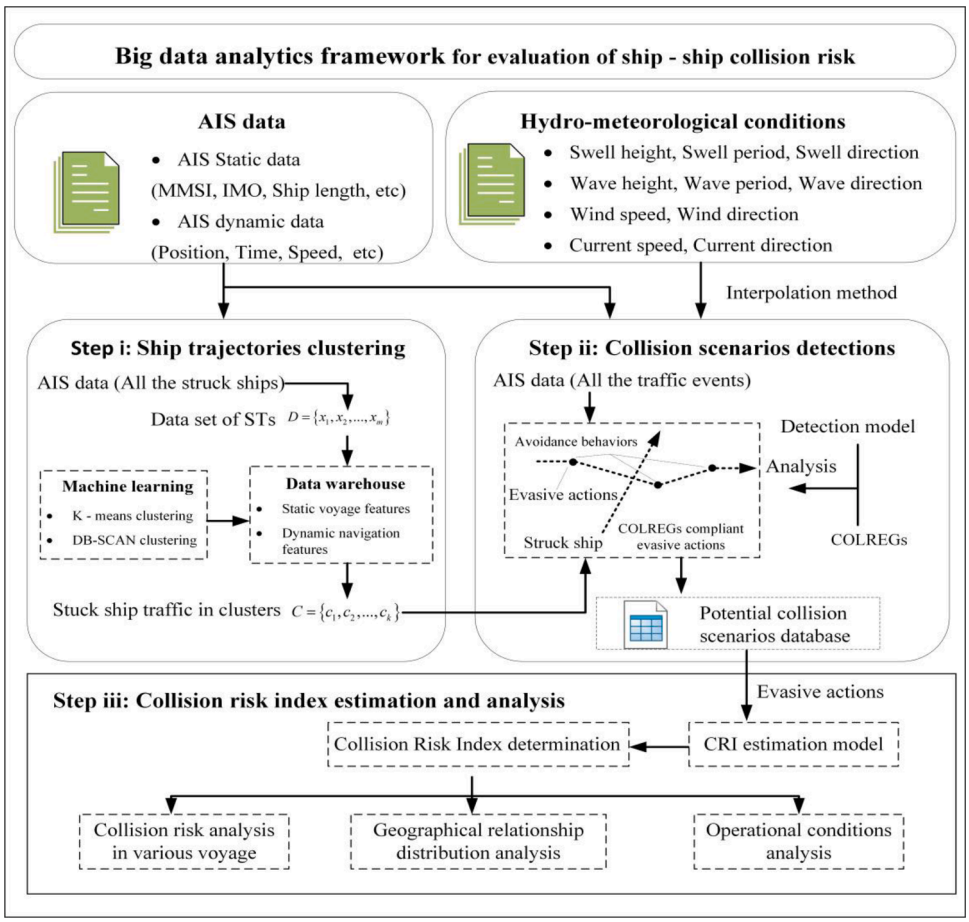


Fig. 1. The logic framework of collision risk evaluation using big data analytics

can be clustered into various voyages for ship-ship collision evaluation in more detail. The paper focuses on ship-ship collisions, with RoRo/Passenger ships (RoPax) being considered as the struck ships. So, Step (i) is only applied to cluster ship trajectories of the struck ships.

- **Step (ii)** - for each cluster identified under Step (i) collision scenarios are identified using the proposed avoidance behavior-based collision detection model. This part of the analysis considers evasive actions as per COLREGs [91]. Then, collision scenarios and hydro-meteorological data at that time associated with each cluster are stored in a database for further collision risk analysis in more detail.
- **Step (iii)** - for each collision scenario detected under Step (ii) the collision risk when evasive actions are taken is evaluated using a CRI estimation model. More specifically, the risk profiles of ships are analyzed for each cluster by a method accounting for potential collision events over a pre-defined period corresponding to specific ship type operations in an area of reference. The results of CRIs are explored by statistical analysis accounting for real hydro-meteorological conditions.

2.2.1. *Step i: Clustering of ship trajectories*

The flowchart of AIS trajectory clustering using K-means and DBSCAN is depicted in Fig. 2. It consists of three steps, namely: (a) reconstruction of STs; (b) grouping of static data by K-means and (c) clustering of dynamic data by DB-SCAN. For step (a), throughout the clustering process uncertainties in AIS big data streams may relate to collection, transmission and reception errors [86]. AIS data may also not be transmitted at the same time. This may cause data streams of different ships to be out of sync [74,87]. Thus, AIS data reconstruction requires trajectory separation, data filtering (i.e. outliers removal), and interpolation over 20 s intervals [30,79,80,84].

Using the proposed unsupervised machine learning method based on K-means and DB-SCAN algorithms, complex traffic scenarios can be explored in more detail in various voyages. An example of the ST clustering process for one ship with 6 STs (voyages) sailing in a given area is depicted in Fig. 3. Therein the direction of ST1 is opposite to ST2, likewise ST3,4 are opposite to ST5,6. Despite ST3 and ST4 describe trajectories of ships navigating between the same departure and destination points, these are different. In a similar manner, ST 5 and 6 head in the same direction, but the speeds of the ships along the trajectories are different – ships on ST5 is faster than a ship on ST6. Separation of the STs and exploration of the collision risk is achieved as follows:

- K-means algorithm is used to classify STs into 4 clusters using static voyage features (departure, destination, voyage length). In this way, ST1, ST2, ST3,4,and ST5,6 should be positioned in different clusters.
- DB-SCAN algorithm is employed to re-cluster results using dynamic navigation data (ship speed, course, motion parameter variation and trajectory spatial distance). In this way, ST3, and ST4 (ST5, and ST6) should be positioned in different sub-clusters.

The STs are clustered into 6 clusters. Similar ship trajectories are grouped into the same cluster. Thus, a cluster may contain more than one similar ship trajectories/ voyages.

The adequacy of the approach depends on the availability of AIS data (Fig. 4 and Table 3). Along with a trajectory, paths are defined as follows:

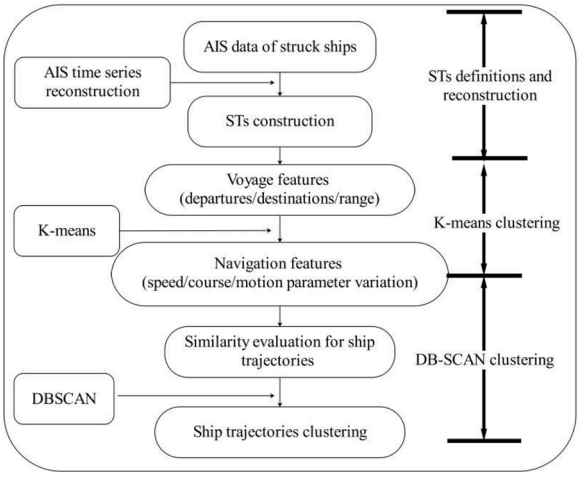


Fig. 2. The flowchart for ST clustering

$$Tr_i = p_1^i, p_2^i, p_3^i, \ldots p_j^i, \ldots, p_n^i, (1 \le j \le n) \tag{1}$$

for

$$p_j^i = \{MMSI, TIMESTAMP, LON, LAT, SOG, COG, ROT, H, SS, L, W, D\} \tag{2}$$

where p_j^i is a point in 2D space that contains MMSI number of the ship, timestamp, geographical position, speed, course, heading, ship type, ship length, ship width, and draft; j is the timestamp of this point; n is the total number of the points in the trajectoriesTr_i and p_1^i, p_n^irepresent ship departure and destination points.

The clustering of STs between the same departure/destination points is defined by the similarity parameter S_{dd}^p. This is a set including the distance between ship departure ($p_1{}^i$,p_1^{i+y}) and destination points ($p_n{}^i$, p_n^{i+y}). It is used to identify the STs sharing the departure points (p_1), destinations (p_n), and vice versa as follows:

$$S_{dd}^p = \left\{ \begin{matrix} \{dist(p_1^i(lon_1, lat_1), p_1^{i+j}(lon_1, lat_1))\} \\ \{dist(p_n^i(lon_n, lat_n), p_n^{i+j}(lon_n, lat_n))\} \end{matrix} \right\} \tag{3}$$

Table 3
Description of parameters for a point in ST.

	Description
MMSI	Maritime Mobile Service ID (MMSI) and location of the system's antenna on board
TIMESTAMP	The timestamp of AIS data
LON	Longitude of the position
LAT	Latitude of the position
SOG	Speed over ground
COG	Course over ground
ROT	Right or left (ranging from 0 to 720° per minute)
H	heading of the ship
SS	Ship Specification
W	Width of the ship
L	Length of the ship
D	Draught ranges from 0.1 m to 25.5 m

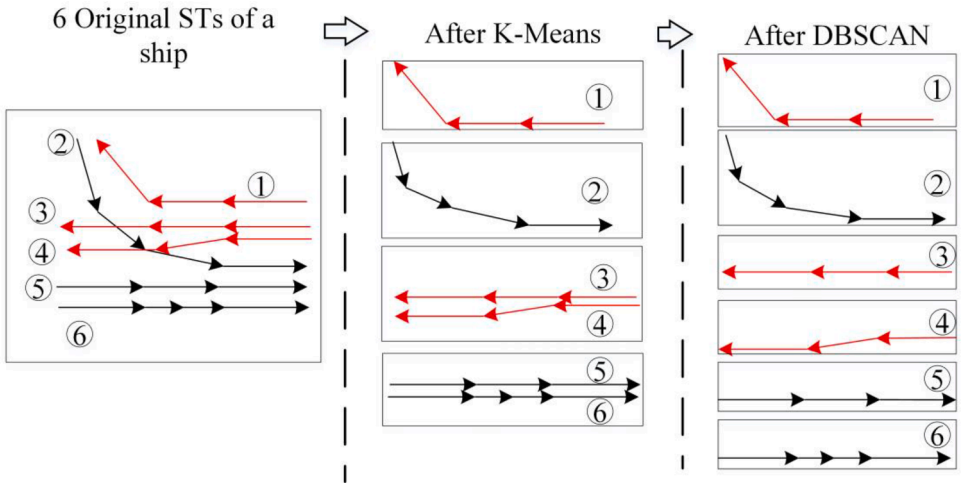


Fig. 3. Process of trajectories clustering using the K-Means algorithm and DB-SCAN algorithm

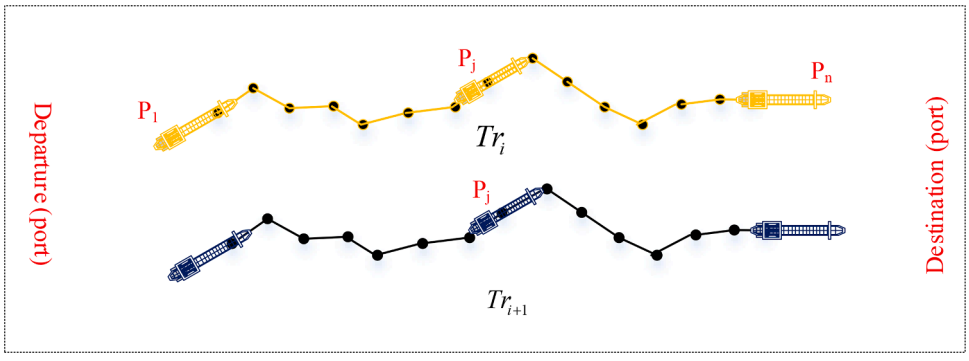


Fig. 4. Schematic diagram of ST

where $(lon_1, lat_1)$and (lon_n, lat_n) denote longitude and latitude of the departure and destination points, respectively; n is the total number of the waypoints of the ST i.Voyage length is defined as:

$$d(p_j, p_{j+1}) = dist(lon_j, lat_j, lon_{j+1}, lat_{j+1}) \tag{4}$$

$$T_{length} = \sum_{j=1}^{n-1} d(p_j, p_{j+1}) \tag{5}$$

K-Means clusters the similarity of voyage features of different STs based on the main difference between departure p_1, destination p_n and voyage length. The similarity parameterS_{dd}^{p} denotes the main difference between alternative departure $(p_1{}^i, p_1^{i+y})$ or destination points $(p_n{}^i, p_n^{i+y})$. On the other hand S_{dd}^{st} is defined as a similarity set that uses the sum of distances of the same departure and destination points Tr_i and Tr_{i+y} according to the equation:

$$\begin{aligned} S_{dd}^{st}(Tr_i, Tr_{i+y}) &= dist(lon^i{}_1, lat^i{}_1, lon_1^{i+y}, lat_1^{i+y}) \\ &+ dist(lon^i{}_n, lat^i{}_n, lon_n^{i+y}, lat_n^{i+y}) \end{aligned} \tag{6}$$

The similarity parameter S_l denotes the difference in the voyage length of different trajectories defined by Equations (5, 7). If the value of the similarity parameter $S_l$is small, and STs of ships navigating between the same departure points and same destination points, then:

$$S_l(Tr_i, Tr_{i+y}) = \left| T_{length}^{i} - T_{length}^{i+y} \right| \tag{7}$$

Consequently, STs can be clustered using K-mean algorithm based on the following three factors defined as points in three-dimensional space: similarity parameter $S_l$and similarity parameters S_{dd}^{p} and S_{dd}^{st}. Additionally, if we consider more than three above parameters for STs clustering using K-means, the performance is not worked well. Thus, dynamic navigation features also should be mined to explore the difference of ship trajectories in more detail using DB-SCAN in the same voyage route.

The navigation features of STs consider AIS data, including SOG, COG, and variations of those (e.g., average value, median value, and variance). The average and median value of COG are used for determining the course feature defined by similarity parameters:

$$S_{sog} = \{sog_{mean}, sog_{median}\} \tag{8}$$

$$S_{cog} = \{cog_{mean}, cog_{median}\} \tag{9}$$

The motion parameter variation features are defined as follows:

$$S_{mpv} = \{sog_{interval}, sog_{std}, cog_{interval}, cog_{std}\} \tag{10}$$

To present the difference of navigation features of various trajectories, S_{sog}, S_{cog} and $S_{mpv}(Tr_i, Tr_{i+y})$ are defined as:

$$S_{cog}(Tr_i, Tr_{i+y}) = \left| cog_{mean}^{i} - cog_{mean}^{i+y} \right| \tag{11}$$

$$S_{sog}(Tr_i, Tr_{i+y}) = \left| sog_{mean}^{i} - sog_{mean}^{i+y} \right| \tag{12}$$

$$\begin{aligned} S_{mpv}(Tr_i, Tr_{i+y}) &= \left| sog_{interal}^{i} - sog_{interal}^{i+y} \right| + \left| sog_{std}^{i} - sog_{std}^{i+y} \right| \\ &+ \left| cog_{interal}^{i} - cog_{interal}^{i+y} \right| + \left| cog_{std}^{i} - cog_{std}^{i+y} \right| \end{aligned} \tag{13}$$

where, the sog_{mean} and sog_{median} represent the average and median values of SOG, respectively; the cog_{mean} and cog_{median} represent the average and median values of COG, respectively; the$sog_{interval}$, sog_{std}, $cog_{interval}$, and cog_{std} denote variable interval and standard deviation of SOG and COG; Tr_i and Tr_{i+y}represent different STs.

Voyage details and navigation features are delivered from tempo-spatial AIS data. To calculate the spatial distance of two STs using discrete AIS points of STs, the spatial similarity of STs is calculated using the Hausdorff distance algorithm [40]:

$$h(Tr_i, Tr_{i+y}) = \max_{p_j^i \in Tr_i} \left(\min_{p_j^{i+y} \in Tr_{i+y}} \left(d\left(p_j^i, p_j^{i+y}\right) \right) \right) \tag{14}$$

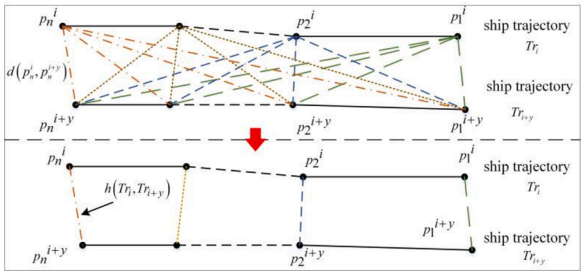


Fig. 5. Illustration of the Hausdorff distance algorithm for the spatial similarity calculation of STs

$$h(Tr_{i+y}, Tr_i) = \max_{p_j^{i+y} \in Tr_{i+y}} \left(\min_{p_j^i \in Tr_i} \left(d\left(p_j^{i+y}, p_j^i\right) \right) \right) \tag{15}$$

The spatial similarity parameter of two different STs is defined as:

$$S_h = \max\{h(Tr_i, Tr_{i+y}), h(Tr_{i+y}, Tr_i)\} \tag{16}$$

where, $h(Tr_i, Tr_{i+y})$ denotes the Hausdorff distance of trajectory Tr_i to Tr_{i+y} and the $h(Tr_{i+y}, Tr_i)$ denotes the Hausdorff distance (see Fig. 5) of ST Tr_{i+y} and Tr_i; S_h is the spatial similarity parameter of different STs.

Clustering of voyage features (e.g., departure/destination, voyage), navigation features (e.g., speed, course, and ship motion parameter variation, spatial distance), and spatial distance of trajectories by the DB-SCAN method is achieved by:

$$S = \sum w_i * S_i, S_i \in \left[S_{dd}^{p}, S_{dd}^{st}, S_l, S_{sog}, S_{cog}, S_{mpv}, S_h \right] \tag{17}$$

where, Sdenotes the multi-criteria feature of ST, ω_i indicates the weight of the above-mentioned feature parameters. The weights ω_i of the feature parameters are tested using a small sample based on the evaluation equation (31). Experience shows that when the weights ω_i are determined as $[0.13, 0.16, 0.21, 0.12, 0.12, 0.09, 0.17]$ the performance of STs clustering is best. Due to their different dimensions features and spatial trajectory distances must be normalized according to the similarity estimation matrix [65]:

$$SM = \begin{bmatrix} S_{11} & S_{12} & \cdots & S_{1n} \\ S_{21} & \ddots & & S_{1n} \\ \vdots & & S_{kk} & \vdots \\ S_{n1} & S_{n1} & \cdots & S_{nn} \end{bmatrix} \tag{18}$$

where, n is the number of the STs for clustering, S_{kk} is the multi-criteria feature of ST Tr_i and ST Tr_{i+j}.

2.2.2. Step ii: Collision detection

During this stage a database utilizing global now-cast data from different providers is developed. Wind data are obtained from US NOAA (https://www.noaa.gov/); Wave and tide data are based on Tidetech (https://www.tidetech.org/) and Ocean currents information is described as per Mercator Ocean (https://www.mercator-ocean.fr). The applicability of now-cast data is confirmed by comparisons against on-board measurements [27]. In these records, swell and wind wave components are presented by significant wave height, wave zero-crossing period and wave direction over 60 minutes. The spatial resolution of 1.25 km is used [34]. From now-casts, wave heights can be obtained within 0.3 meters of uncertainty (globally). Wave periods are estimated within 2s (e.g., [4,46]). The accuracy of main sea weather forecast models is evaluated by comparing records against data collected on weather buoys using RMS error estimators [27,30]. The hydro-meteorological data are interpolated to the ship position and time delivered from AIS data. The interpolation process follows the principles outlined in Appendix A and comprises of the following steps (1)

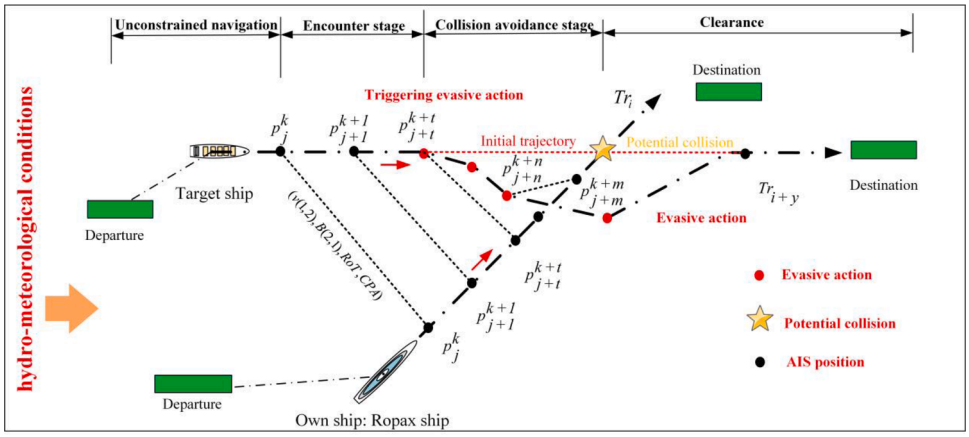


Fig. 6. Potential collision detection process from the struck ship perspective (crossing between struck ship and target ship as an example)

worldwide AIS database extract including positions with respect to the timestamp of all struck ships; (2) extract of hydro-meteorological data that includes information on weather, ship position and time; (3) a trilinear interpolation procedure to find the link between operational and hydro-meteorological conditions under which ships operate.

In Fig. 6, as potential collision scenario is defined a ship - ship encounter that comprises of four stages, namely (a) unconstrained navigation; (b) encounter; (c) collision avoidance; (d) clearance. Ship evasive actions take place when a ship performs course or speed alterations or both. In Fig. 6, STs Tr_i and Tr_{i+y}relate to struck and target ships. During stages (a) and (d) the risk of collision between the two ships is negligible, either because of the distance between two ships (stage a) or their diverging courses (stage d). At encounter stage (b) when the rate of change of relative bearing angle $\Delta\beta$ relative to struck ship falls within [-2.00 to +2.00], the risk of collision is defined by COLREGs [35], and a collision may occur unless evasive action is taken. If the distance between two ships reduces but the rate of change of relative bearing angle $\Delta\beta$ exceeds the range of [-2.00 to +2.00], this indicates the striking ship (give-way ship) changes her course to avoid collision. The critical point associated with maximum rate of relative bearing angles $\Delta\beta$ is defined as the time of evasive action taken. Thus, she enters the collision avoidance stage c (see timestamp $k+t$ in Fig. 6). At this stage, ships converge and the minimum distance between STs of striking and struck ships is below 3 nm, the minimum DCPA is below 1 nm and the minimum TCPA is located within (0 to 30) mins. The end point of the collision avoidance stage is defined as the point where TCPA becomes 0. If TCPA is below 0, there is no collision risk, the distance between two ships is increasing, and the stage of clearance begins.

The Avoidance Behavior-based Collision Detection Model (ABCD-M) used to detect collision can be described as follows: **Part A** where the coordinate system is converted from the earth-fixed (AIS) to struck ship-fixed status (see more in Appendix B). In this part we determine the minimum distance between two STs. This requires that STs of potential striking ships keep clear from the struck ship to minimize the potential of collision. The minimum ship distance d_{min} is defined at timestamp $k+i$ corresponding to STs as $Tr_i^{[k,k+m]}$and $Tr_{i+y}^{[k,k+m]}$, where $[k,k+m]$ denotes the timestamp interval of the two series (see more in Appendix B); **Part B** during which we determine collision avoidance behaviors during ship encounters based on ship course, relative bearing angles $\Delta\beta$, rate of turn (ROT), TCPA, DCPA. The calculation process in Appendix B), and the difference between the headings (Fig. 6); **Part C** where we classify collision scenarios as per COLREGs (Fig. 7 and Fig. 8).

For an encounter stage defined during the time interval $[k,k+t]$ threshold conditions of DCPA, TCPA, distance, $\Delta\beta$ within [-2.00 to +2.00] and the observation range of 6 nm are:

- $Dist \leq 6nm$ [21];
- $d_{\min}(p_{j+i}^{k+i}) \leq 3nm$ [55,75];
- $\Delta\beta \in [-2^{\circ}, +2^{\circ}]$at the time interval $[k,k+t]$ [55];
- $\min(d_{CPA}(p_T^{[k+t,k+n+1]})) \leq 1$nm during collision avoidance stage and for the time interval $[k+t,k+n+1]$ [42]; $0 < \min(t_{CPA}(p_T^{[k+t,k+n+1]})) \leq 30mins$during collision avoidance stage and for the time interval $[k+t,k+n+1]$ [42];
- $t_{CPA}(p_T^{k+n+1}) \leq 0mins$ at clearance stage [71,74].

To analyze the collision avoidance behaviors, the STs $Tr_T$and Tr_O during evasive action are defined by:

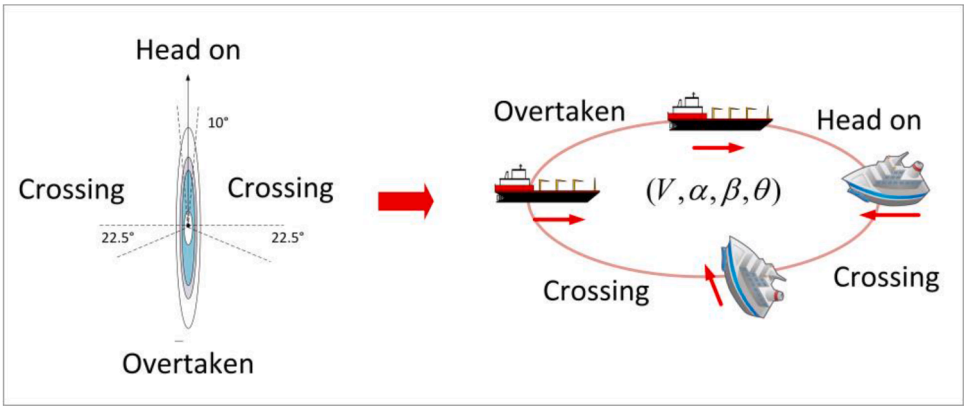


Fig. 7. Three analyzed collision types

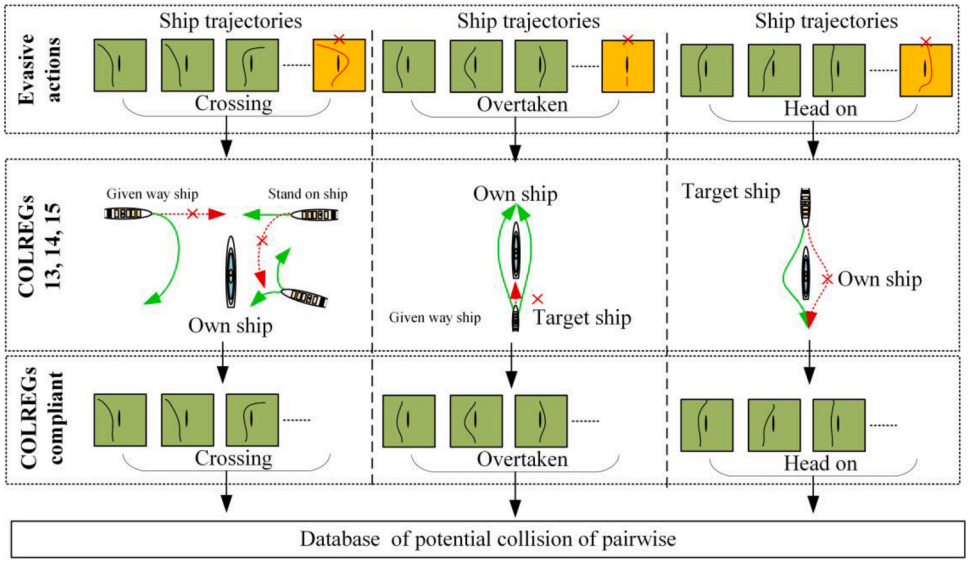


Fig. 8. The evasive actions in real operations (Legal: green track; illegal: red track)

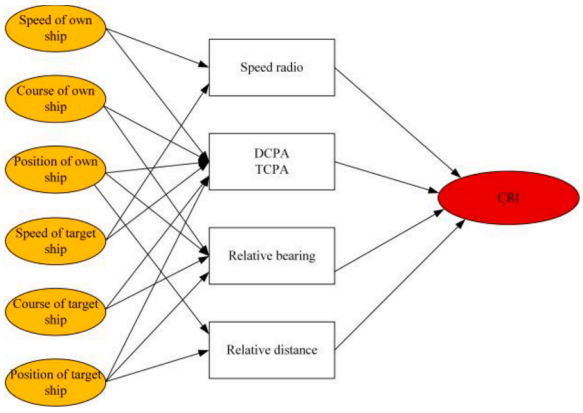


Fig. 9. Influencing factors of CRI and process of calculation

$$Tr_T = \left\{ p_{j+t}^{k+t}, p_{k+t+1}^{k+t+1}, p_{k+t+2}^{k+t+2}, \cdots p_{j+n}^{k+n} \right\} \tag{19}$$

$$Tr_O = \left\{ p_{j+t}^{k+t}, p_{k+t+1}^{k+t+1}, p_{k+t+2}^{k+t+2}, \cdots p_{j+n}^{k+n} \right\} \tag{20}$$

For each point of a ST p_{j+t}^{k+t} defined in Table 3 hydro-meteorological conditions $Hydro_{j+t}^{k+t}$ are implemented as:

$$\mathrm{p}_{j+t}^{k+t} = \begin{bmatrix} MMSI_{j+t}^{k+t}, TIMESTAMP_{j+t}^{k+t}, LON_{j+t}^{k+t}, LAT_{j+t}^{k+t}, SOG_{j+t}^{k+t}, COG_{j+t}^{k+t}, \cdots \\ \cdots, ROT_{j+t}^{k+t}, H_{j+t}^{k+t}, ST_{j+t}^{k+t}, L_{j+t}^{k+t}, W_{j+t}^{k+t}, D_{j+t}^{k+t}, Hydro_{j+t}^{k+t} \end{bmatrix} \tag{21}$$

Based on the detected potential collision scenarios, to analyze the geographical relationship of potential conflicts from struck ships perspective, the coordinate system should be converted from the earth-fixed (AIS) to struck ship-fixed (see Appendix B). Whereas potential collision scenarios are classified into three types: head-on, crossing and overtaking, as depicted in Fig. 7.

Collision avoidance maneuvers that do not comply with COLREGs are usual during navigation [64]. Those are so-called cooperative collision avoidance maneuvers of two ships, which indicates that two ships understand the collision situations through communication and work out jointly the solution. A demonstration of such scenario is given in Fig. 8 from struck ships perspective, those according to COLREGs. Ideally the maneuvers of the give-way ship should be along the green track. However, some give-way ships may take the evasive actions along the red track defined as the illegal evasive actions that should be culled. This is because non-compliant to COLREGs evasive actions cannot be used to define commonly agreed risk thresholds for intelligent decision support system development. In such encounters communication between the vessels involved may lead to accident resolution. The detected illegal evasive actions (cooperative collision avoidance scenarios) may be analyzed separately for the research regarding ship collision avoidance under human-machine interaction.

Illegal evasive actions are detected using the relative bearing angle β from striking to struck in collision avoidance stage (Fig. 6) from struck ship perspective, shown Fig. 8 according to COLREGs Rule 13, 14, and 15 [29]. The pseudocode for COLREGs Rule 15 (crossing collision scenario) is summarized in So, the relative bearing angle β will decrease to less than 270^0. Otherwise, the evasive actions will be defined as COLREGs uncompliant in head on situation [73].

Table 4. The crossing collision scenarios are classified into three cases based on COLREGs Rule 15 (Fig. 8).

- If $5^0 < \beta < 67.50$ the struck ship is the give-way ship, and the striking ship should pass from the bow during the collision avoidance stage. The relative bearing angle β from struck ship perspective will increase to more than 180^0.
- If $67.5^0 < \beta < 112.5^0$ the struck ship is the give-way ship and she is obliged to alter her course to the starboard side or reduce speed during the collision avoidance stage. In this situation, all the evasive actions are COLREGs compliant.
- If $247.5^0 < \beta < 355^0$ the striking ship is the give-way ship and the striking ship should pass from the stern during the collision avoidance stage. In this situation the relative bearing angle β of the struck ship will decrease to less than 180^0.

Examples of potential collisions during the encounter stage from struck ship perspective are shown in Fig. B.1 of Appendix B. Notably, if the collision avoidance behaviors violate the above terms from struck ship perspective, the evasive actions will be defined as COLREGs uncompliant focusing on crossing collision scenarios.

The pseudocode for overtaking collision scenario (COLREGs Rule 13) is shown in Table 5. In this case if $112.5^0 < \beta < 247.5^0$ the speed ratio of striking/struck ships is more than 1 (i.e., the striking ship is faster than the struck ship). Thus, the striking ship is the give-way ship, and should overtake the struck ship during the collision avoidance stage (Fig. 8).

Table 4
The procedure pseudocode for culling the illegal evasive actions focus on crossing situation.

Algorithm 3:COLREGs 15: Crossing situation
Input: Ship trajectories Tr_iandTr_{i+y};
Output: Ship trajectories $Tr_T$and Tr_Owithout illegal evasive actions;
Process:
1. Procedure pseudocode for crossing situation
2. Initial:struck ship = power-driven ship && striking ship = power-driven ship;
3. Input:struck ship's positions (x_0, y_0), course(θ_0), and velocity(v_0) from ship trajectories Tr_i;
4. Input:striking ship's positions (x_T, y_T), course(θ_T), and velocity(v_T) from ship trajectories Tr_{i+y};
5 5. Calculate the minimum distance $d_{\min}(p_{j+i}^{k+i})$;
6. Calculate the relative bearing angle β;
7. Calculate CPA;
8. if CPA, $\Delta\beta <$ Thresholds then
9 . return no risk of collision;
10. end
11. if ($\beta \in$ [5°, 67.5°]) then
12. Output: struck ship *is give-way ship && turn starboard*;
13. else if ($\beta \in$ [67.5°, 112.5°]) then
14. Output: struck ship *is give-way ship && (turn starboard side || speed alteration);*
15. else if ($\beta \in$ [247.5°, 355°]) then
16. Output: *struck ship is stand-on vessel && keep course;*
17. End procedure

Table 5
The procedure pseudocode for culling the illegal evasive actions focus on the overtaken situation.

Algorithm 4:COLREGs 13: Overtaken situation
Input: Ship trajectories Tr_iandTr_{i+y};
Output: Ship trajectories $Tr_T$and Tr_Owithout illegal evasive actions;
Process:
1. Procedure pseudocode for overtaken ships
2. Input:struck ship's positions (x_0, y_0), course(θ_0), and velocity(v_0) from ship trajectories Tr_i;
3. Input:striking ship's positions (x_T, y_T), course(θ_T), and velocity(v_T) from ship trajectories Tr_{i+y};
4. Calculate the minimum distance $d_{\min}(p_{j+i}^{k+i})$;
5. Calculate the relative bearing angle β;
6. Calculate the ratio of speed k (k= v_0 / v_T);
7. Calculate the CPA;
8. if CPA, $\Delta\beta <$ Threshold then
9. return no risk of collision;
10. end
11. if ($\beta \in$ [112.5°, 247.5°]) && (k<1) then
12. Output: *the struck ship is stand-on ship;*
13. End procedure

The relative bearing angle β from struck ship perspective will increase to more than 270^0 or decrease to less than 90^0. Otherwise, the evasive actions will be defined as COLREGs uncompliant, focusing on overtaken scenarios. On the contrary, if the struck is the overtaking ship, all the evasive actions are legal (turn to the port side or starboard side). Besides, Ro-Pax ships are defined as struck ship in the paper. We only consider the overtaken cases. Finally, the pseudocode for the head-on collision scenario (COLREGs Rule 14) is shown in

Table 6. If $0^0 < \beta < 5^0$ or $355^0 < \beta < 360^0$ the struck ship is the give-way ship, and the ships should pass each other port-to-port during the collision avoidance stage (Fig. 8). So, the relative bearing angle β will decrease to less than 270^0. Otherwise, the evasive actions will be defined as COLREGs uncompliant in head on situation.

Table 6
The procedure pseudocode for culling the illegal evasive actions focus on the head-on situation.

Algorithm 5:COLREGs 14: Head-on situation
Input: Ship trajectories $Tr_i$and Tr_{i+y};
Output: Ship trajectories $Tr_T$and Tr_Owithout illegal evasive actions;
Process:
1. Procedure pseudocode for head-on situation
2. Initial struck ship = power-driven ship && striking ship = power-driven ship;
3. Input:struck ship's positions (x_0, y_0), course(θ_0), and velocity(v_0) from ship trajectories Tr_i;
4. Input:striking ship's positions (x_T, y_T), course(θ_T), and velocity(v_T) from ship trajectories Tr_{i+y};
5. Calculate the minimum distance $d_{\min}(p_{j+i}^{k+i})$;
6. Calculate the relative bearing angle β;
7. Calculate the CPA;
8. if CPA, $\Delta\beta <$ Threshold then
9. return no risk of collision;
10. end
11. if ($\beta \in [0°, 5°] \cup [355°, 360°]$) then
12. Output: *struck ship is give-way ship && turn starboard;*
13. End procedure

2.2.3. Step iii: CRI estimation

CRI presents the risk of ship - ship collision by evaluating the geographical relationship of potential conflicts. The applications of CRI method can be classified into two groups: (a) a specific value of CRI defined by expert' knowledge is used as risk criteria to detect ship conflicts (e.g., [17,71]); (b) the CRI model is employed to quantify collision risk for collision avoidance (e.g., [31,66]). However, the former lacks commonly agreed on risk criteria to show what is the real dangerous situation or what is time to take collision avoidance [53]. Thus, using the detected potential collision scenarios in real operational conditions under Step (ii), the CRI method adopted in this paper is used to quantify collision risk when the give way ships take the evasive actions under COLREGs compliant in real operations. The wide set of data can be used to calibrate a commonly agreed risk criteria value by statistical analysis, which is defined by expert' knowledge in previous research. The CRI method is represented as:

$$CRI_i = \{DCPA,\ TCPA,\ \mathrm{D},\ \beta,\ \mathrm{K}\} \tag{22}$$

The risk value for DCPA is defined as:

$$u(d_{CPA}) = \begin{cases} 1, d_{CPA} < d_1 \\ 0.5 - 0.5\sin\left[\dfrac{\pi}{d_2 - d_1}\left(d_{CPA} - \dfrac{d_1 + d_2}{2}\right)\right], d_1 < d_{CPA} \le d_2 \\ 0, d_2 < d_{CPA} \end{cases} \tag{23}$$

where d_1 is the minimum safe meeting distance, and d_2 is the minimum distance between the striking ship and struck ship. In practice, to avoid a collision accident a striking ship should not pass the struck ship at a distance shorter than the one that is considered safe [18,19,52,56]. According to Gang et al. [17] such distance can be calculated as follows:

$$d_1 = \begin{cases} 1.1 - \dfrac{0.2\beta}{180^\circ}, 0^\circ \le \beta < 112.5^\circ \\ 1.0 - \dfrac{0.4\beta}{180^\circ}, 112.5^\circ \le \beta < 180^\circ \\ 1.0 - \dfrac{0.4 \times (360^\circ - \beta)}{180^\circ}, 180^\circ \le \beta < 247.5^\circ \\ 1.1 - \dfrac{0.2 \times (360^\circ - \beta)}{180^\circ}, 247.5^\circ \le \beta \le 360^\circ \end{cases} \tag{24}$$

where β is the relative bearing angle from striking ship to struck ship.

The risk value for TCPA is defined as:

$$u(t_{CPA}) = \begin{cases} 0, t_2 < t_{CPA} \\ \left(\dfrac{t_2 - TCPA}{t_2 - t_1}\right)^2, t_1 < t_{CPA} < t_2 \\ 1, 0 < t_{CPA} < t_1 \end{cases} \tag{25}$$

where $t_1 = \frac{\sqrt{{d_1}^2 - DCPA^2}}{S_r}$ denotes the time it takes for the ship to sail from the point of evasive actions to the closest point of approach; $t_2 = \frac{\sqrt{{d_2}^2 - DCPA^2}}{S_r}$ is the collision avoidance time, referring to the time it takes for the ship to sail from the current location to the point with minimum distance and S_r is the relative speed between two ships.

The risk value for the distance between the striking and struck ships (D) is defined as:

$$u(D) = \begin{cases} 0, \mathrm{d}_2 < D \\ \left(\frac{d_2 - D}{d_2 - d_1}\right)^2, \mathrm{d}_1 < D \le d_2 \\ 1, D \le d_1 \end{cases} \quad (26)$$

The risk value for the relative bearing angle β between striking and struck ships is defined as:

$$\mathrm{u}(\beta) = 0.5 \times \left[\cos(\beta - 19^\circ) + \sqrt{\frac{440}{289} + cos^2(\beta - 19^\circ)}\right] - \frac{5}{17}, \quad 0^\circ \le \beta \le 360^\circ \quad (27)$$

The risk value for the ship speed ratio of striking and struck ships is defined as:

$$u(K) = \frac{1}{1 + \frac{2}{K\sqrt{K^2 + 1 + 2K\sin C}}} \quad (28)$$

where $\sin C = |\sin(|\theta_T - \theta_0|)|$, $\mathrm{k} = (\mathrm{V}_0 / \mathrm{V}_T)$, θ_T and θ_0 are the course of the striking ship and struck ship.

The mentioned factors above contribute to CRI. But the degree of influence in CRI is different. According to Gang et al. [17] and Hu et al. [31], the degree of the mentioned five factors influencing in CRI is defined as per equation (29), and the weighting factors are determined as presented in the equation (30).

$$DCPA > TCPA > D > \beta = K \quad (29)$$

$$CRI = 0.40u(d_{CPA}) + 0.367u(t_{CPA}) + 0.167u(D) + 0.033u(\beta) + 0.033u(K) \quad (30)$$

Overall, CRI is a single crisp value reflecting the risk of collision with other ships, which summarizes the mentioned five factors influencing in collision risk by equation (30). Usually the CRI for two ships is a cost-like value. It trends to be higher for the higher of the collision risk (the higher CRI value, the higher of maneuvering difficulty of ship avoidance). CRI weighting factors usually are set as 0.40, 0.367, 0.167, 0.033 and 0.033, respectively, which are determined by quantifying the difficulty of ship avoidance in various conflicts using the navigation simulator [78,96]. The CRI calculated using equation (30) usually is used in collision risk evaluation and collision avoidance research. (e.g., [17,31,66,71]).

3. Case study

As part of a practical case study we analyzed more than 4 billion AIS and hydro-meteorological data records describing various conditions over 13 months of ice-free navigation period of atypical Ro-Pax ships steaming through the Gulf of Finland. As a result, the estimates describing the risk of collision for this ship are derived. The information on ship specification and study area are presented in Table 7. Fig. 10

Table 7
Information on ship specification and study area.

Study area	Longitude: 23.57 E and 27.64 ELatitude: 58.99 N and 60.59 N
Ship specification	Gross tonnage: 10,000 GT and 46,124 GT Length: 120 m and 218.8 m
Period	Ice-free period from 2018 to mid-2019

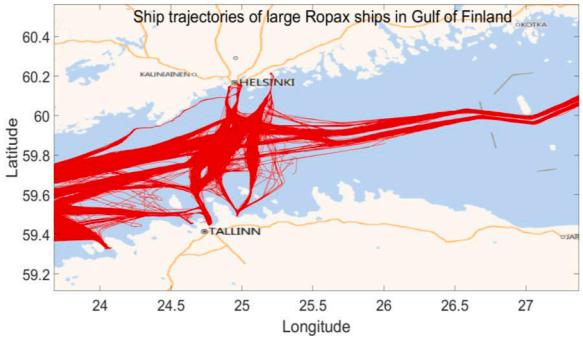


Fig. 10. STs of struck ships (Ro-Pax ships) in the Gulf of Finland (Period: 2018-2019)

shows that the STs of the mentioned Ro-Pax ships are complex and irregular.

3.1. Ship trajectories clustering into various voyages

The K-Means algorithm was used to cluster voyage details of STs (Fig. 2). As part of this process STs were reconstructed and then separated based on the distribution of time between voyages. Then, STs were separated for those cases the time interval between two ships exceeded 360 s. The 12,214 ship voyages of struck ships were divided into 8 clusters using the proposed method in Section 2.1 and Section 2.2.1 (see Fig. 11 and Table 8). Detailed analysis confirmed that K-Means can help classify STs using static voyage details.

The DB-SCAN algorithm was used to classify STs based on dynamic big data streams. The algorithm contains two threshold parameters, namely, *MinLns* and ε [92,93], where ε denotes a spatial distance threshold delimiting the neighborhood of a ST and *MinLns* denotes the minimum number of STs required to form a dense cluster. Formula (31) is often used to evaluate the performance of the clustering method. The parameters of clustering methods can be evaluated according to [45] as:

$$E = \sum_{i=1} \left(\frac{1}{2|C_i|} \sum_{x \in C_i} \sum_{y \in C_i} dist\left(Tr_i, Tr_{i+y}\right)^2 \right) + \frac{1}{2|N|} \sum_{w \in N} \sum_{z \in N} dist(w, z)^2 \quad (31)$$

Where C and N represent normal categories and abnormal results, $dist(Tr_x, Tr_y)$ represents the distance between trajectories Tr_i and Tr_{i+y}.

Theoretically, the lower the value E is, the better performance of clustering becomes. In this paper, several groups of *MinLns* (1 to 9) were compared with ε between 0.001 to 0.01. The experiences show that when the *MinLns* and ε are determined as 6 and 0.006, the value$E$ is lowest, showing that the performance of STs clustering is the best.

The results illustrated in Table 9 (see also Fig. 12 and Fig. 13) contain 16 sub-clusters on top of those initially identified by the K-means method (Fig. 3 and Table 8). Sub-clusters (1), (2), and (3) represent ship traffic behaviors for trips from the port of Helsinki to West Baltic and Russia. Sub-clusters (7), (9), (11), (13) and (16) represent ship traffic behaviors entering the port in Helsinki to Baltic Sea, Russia and Tallinn. Sub-clusters (4), (6), (14), and (15) represent the ship traffic behaviors of entering the port of Tallinn from the Baltic Sea, Russia, and Helsinki. Sub-cluster (5), (8), (9), (10) and (13) represent the ship traffic behaviors of leaving the port in Tallinn to the Baltic Sea, Russia, Helsinki. Sub-cluster (12) represents the ship traffic behaviors from the Baltic Sea through the Gulf of Finland, heading directly to Russia. In addition, some incomplete STs are classified under cluster 17. STs belonging to the same sub-clusters are similar to each other in the navigation features. The results show that the proposed method exhibits effective performance associated with marine traffic pattern recognition using massive STs.

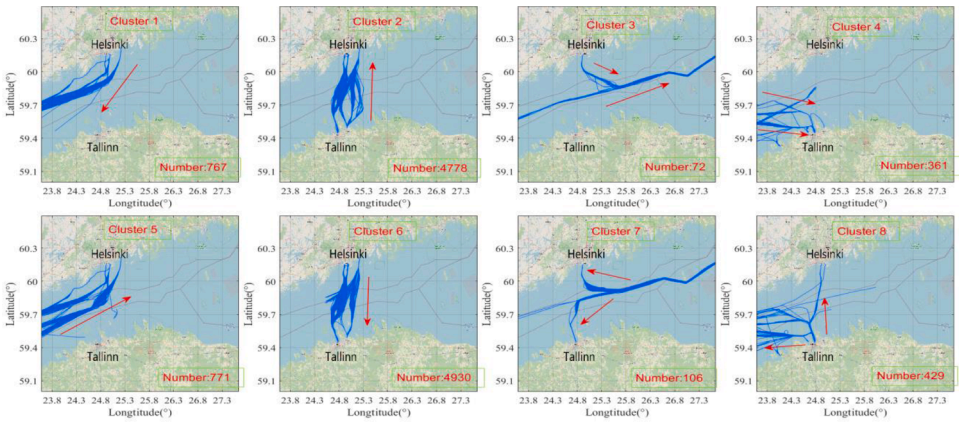


Fig. 11. The STs of clustered results after K-Means

Table 8
Cluster descriptions after K-Means.

No.	Cluster descriptions	Number
1	Westbound from Helsinki.	767
2	From Tallinn to Helsinki.	4778
3	Eastbound to Russia.	72
4	Westbound from Tallinn.	361
5	Eastbound to Helsinki.	771
6	From Helsinki to Tallinn.	4930
7	Westbound from Russia.	106
8	Westbound from Tallinn.	429

Table 9
Sub-cluster descriptions after DB-SCAN.

No.	Sub-cluster descriptions	Number of ship trajectories
1	After leaving the port from Helsinki, heading directly to the Baltic Sea in the western in coastal waters.	37
2	After leaving the port from Helsinki, heading directly to the Baltic Sea in the western in open sea.	712
3	After leaving the port from Helsinki heading directly to Russia.	41
4	After leaving the port from Helsinki heading directly to Tallinn.	570
5	After leaving the port from Tallinn, heading directly to the Baltic Sea in the western in coastal waters.	354
6	After leaving the port from Helsinki heading directly to Tallinn.	4127
7	From the Baltic Sea to the port in Helsinki.	362
8	After leaving the port from Russia to Tallinn.	28
9	After leaving the port from Tallinn, heading directly to the Helsinki.	571
10	After leaving the port from Tallinn to the Baltic.	40
11	From the Baltic Sea to the port in Helsinki in coastal waters.	375
12	Form the Baltic Sea through the Gulf of Finland, heading directly to Russia.	20
13	After leaving the port from Tallinn to Helsinki.	4098
14	From the Baltic Sea to the port in Tallinn in coastal waters.	319
15	From the Baltic Sea to the port in Tallinn in open sea.	10
16	After leaving the port from Russia, heading directly to the Helsinki.	84
17	Incomplete ship trajectories	467

3.2. Statistical analysis

STs were compared in terms of shape, speed and course (see Fig. 13, Fig. 14 and Fig. 15). From an overall perspective, the clustered STs are different (Table 9). However, within the same cluster, STs show a high similarity when it comes to voyage and navigation features. The reason for the latter is that struck ships encounter different traffic densities associated with different collision scenarios in different clusters.

The available weather database is listed in Appendix A. Hydro-meteorological data history records for STs of all clusters at different locations and global ocean now-cast records were reviewed using online weather database and the trilinear interpolation method of Appendix A.

Table 10 and Fig. 16 demonstrate the hydro-meteorological parameters cumulative distributions for the 2-year operations of struck ships in the Gulf of Finland.

Analysis of the results shows that in the Gulf of Finland in the ice-free period and for 99% of the time, all Ro-Pax ships navigate in wave heights smaller than 3.24 m, the swell height of less than 1.49 m, wind speed conditions that are less than 17.91 m/s over ground and currents are less 0.51 m/s over-ground. However, the combination of these conditions does not reflect the hydro-meteorological data encountered in one area of operation over the same time of the year. They rather reflect extreme encounters in different areas of operation during different times.

3.3. Collision scenarios

Potential collision scenarios were detected by applying the approach of Section 2.2.2 (see Fig. 6). To present the relationship between struck and striking ship using AIS, the origin of the original WGS-84 coordinate system was converted to a struck ship-fixed system (see Fig. 17(a)). For unconstrained navigation 266,666 pairs of STs were merged within the 6 nm conventional radar range [24,82]. Furthermore 138,973 pairs of STs were selected, and the minimum distance within each pair was found under 3 nm (see Fig. 17b).

31,079 pairs of STs were obtained over the two years of maritime operations. The relative bearing angles between ships involved in different scenarios varied from [-2.00 to +2.00] over 6 min observations. During the collision avoidance and clearance stages DCPA and TCPA threshold conditions were applied. This resulted in 10,781 potential collision scenarios. Of those 9,240 were COLREGs compliant. The remaining were assumed to be illegal evasive actions (cooperative anti-collision behaviors) and were culled according to COLEGs Rules 13, 14, and 15 (Table 11).

Fig. 18 demonstrates the locations of striking ships triggering evasive actions for 12,214 voyages of struck ships during ice-free operations. Most potential collision scenarios were located between Helsinki –Tallinn because of high traffic complexity. A radar display is shown in Fig. 19, where the struck ship is in the center, therein the blue scatter denotes the positions of the striking ship taking evasive actions. As the speed of Ro - Pax ships is high, most striking ships were located in the 1^{st} and 4^{th} quadrants ahead of the struck ships (see Fig. 19) and their density lied within [1 km to 4 km] radius. Higher density areas were visible for relative angles 10^0, 80^0, and 280^0 in relation to the struck ship. These can provide essential guidance to crews to understand the striking ships distribution from own ship perspective. They may also be used to identify higher collision risk areas while onboard. A summary of hydro-meteorological data accounting for evasive actions during collision encounters is shown in

Table 12. The results are based on the method presented in section 2.2.2 (see also Fig. 14 and Table 10).

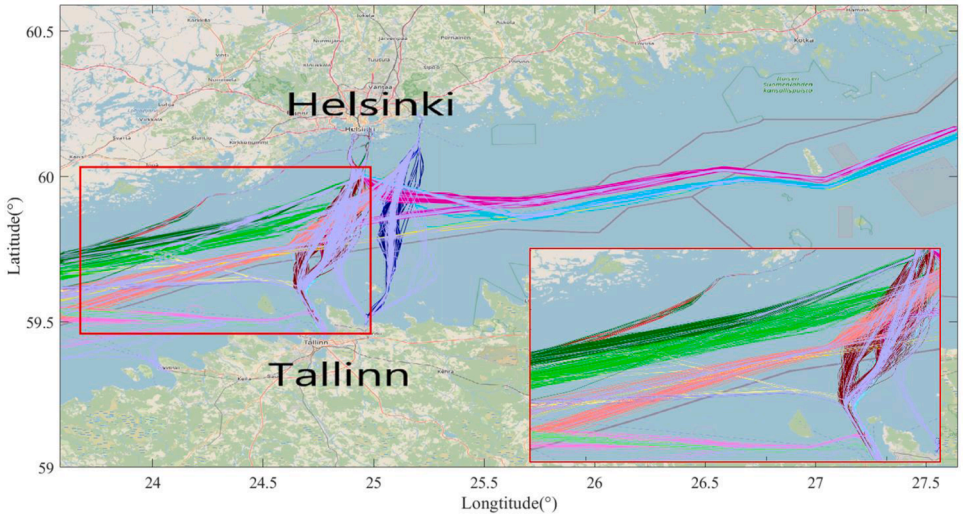


Fig. 12. All the sub-clusters in the Gulf of Finland

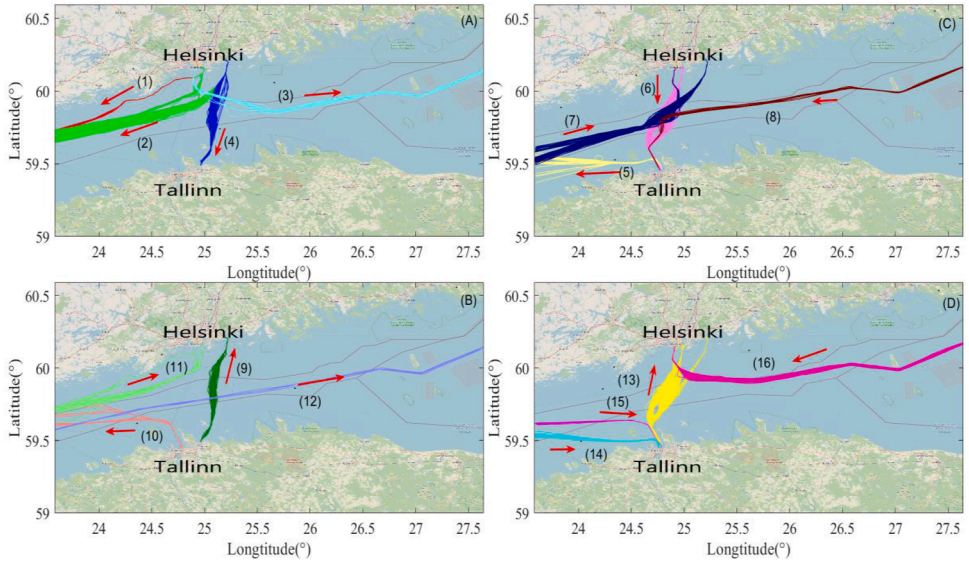


Fig. 13. Results of all the sub-clusters in the Gulf of Finland

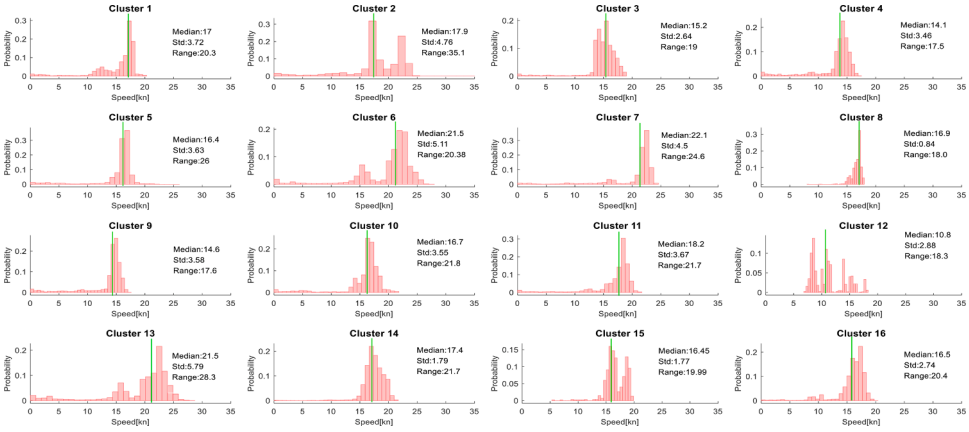


Fig. 14. The ship speed distributions of all the sub-clusters in the Gulf of Finland

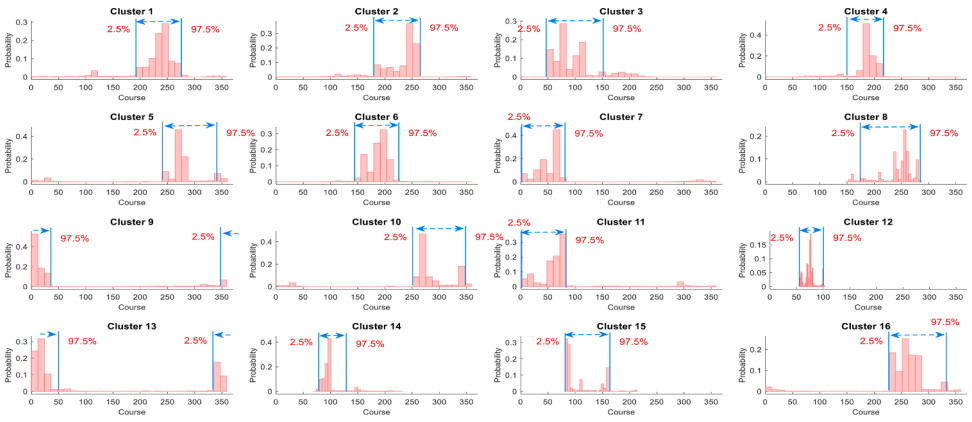


Fig. 15. The course distributions of all the sub-clusters in the Gulf of Finland

Table 10
Hydro-meteorological parameters cumulative distributions for all struck ships over different seasons (Spring, Summer, and Autumn).

	Spring (March, April, May)			Summer (June, July, August)				Autumn (September, October, November)				Winter*	
	25%	50%	75%	99%	25%	50%	75%	99%	25%	50%	75%	99%	
Wave height	0.148	0.341	0.644	2.318	0.125	0.324	0.625	3.240	0.384	0.991	1.662	2.890	none
Current speed	0.011	0.031	0.066	0.404	0.009	0.027	0.065	0.510	0.003	0.024	0.051	0.280	
Wind speed	3.940	5.617	7.561	14.85	3.871	5.657	7.656	17.91	6.756	8.950	11.69	16.21	
Swell height	0.109	0.181	0.299	1.490	0.101	0.175	0.283	1.379	0.123	0.191	0.447	1.306	

*Note: In Winter, the Gulf of Finland may be ice-covered for several weeks. Thus the ice-free period is considered here as dominating in this area.

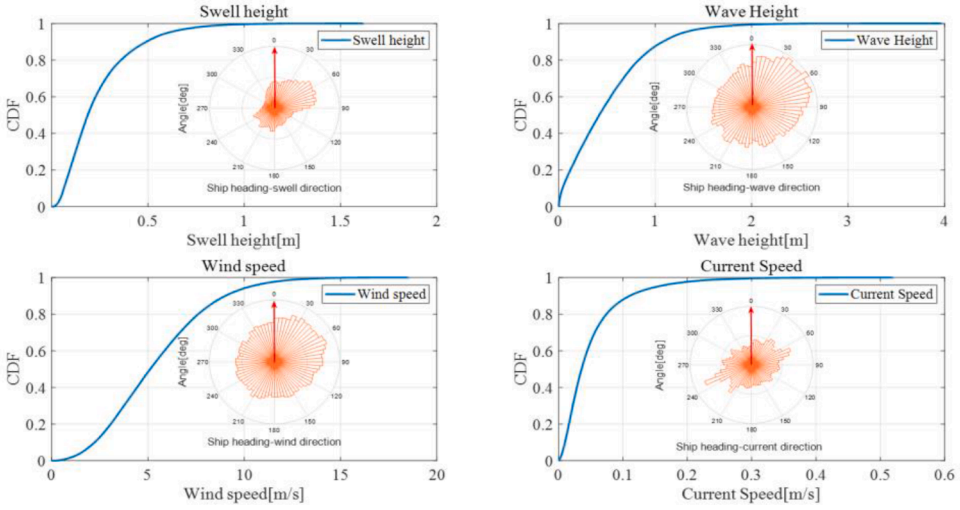


Fig. 16. Hydro-meteorological parameters cumulative distributions for the 2-year operations

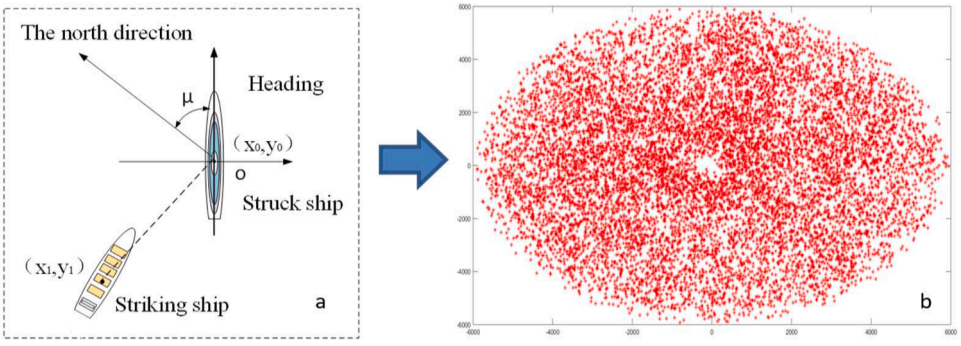


Fig. 17. The minimum distance cloud between two ships

Table 11
*The potential collision events of all struck ship voyages (Results based on 6,213 crossing; 125 overtaking and 2,902 head on collisions).

Navigation stages in trajectories of struck ships	Threshold Criteria	Scenarios* including one struck ship
Unconstrained	Ship – ship distance	266,666 138,973
Encounter	Relative bearing angle	31,079
Collision avoidance and clearance	$\Delta\beta$, ROT, TCPA and DCPA	10,781
	COLREGs compliance	9,240

*Note: An entire ST of a struck ship often encounters more than one striking ship resulting in more pairs of STs of the struck ships and striking ships available.

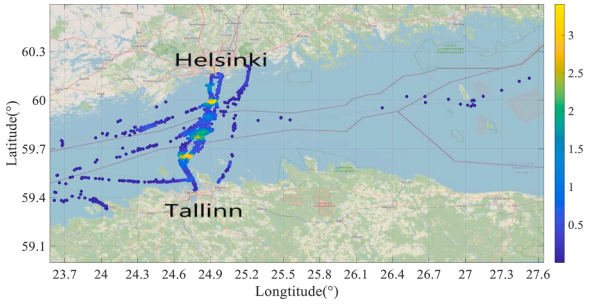


Fig. 18. The locations of the mentioned potential collision scenarios (the scatters denote the positions of Ro-Pax ships encountering potential collisions)

The analysis identified 16 clusters containing complete STs between departure and destination points and 1 cluster containing incomplete STs (see Fig. 13 and Table 9). Collision scenarios (Fig. 18) were classified into 16 clusters (Fig. 20 and Table 13). Consequently, it was found that 50% of the potential collisions occur in cluster 13 (i.e., after leaving the port of Tallinn and towards Helsinki). The mentioned clusters in Fig. 20 denote the grouped STs (see Fig. 13 and Table 9). This observation leads to the conclusion that potential scenarios can be evaluated, focusing on various clusters (voyages) in more detail. The frequency denotes the number of occurrences of potential collisions per journey during the period, calculated using the Formula (32).

$$f_i = \frac{N^i_{\text{potentialcollision}}}{N^i_{\text{shiptrajectories}}}, i = 1, 2, 3\cdots, 16 \tag{32}$$

Table 13 and Fig. 20 show that the number of potential collisions per journey is at its highest in cluster 11 (3.03 potential collisions per journey). Notably, in clusters 12, 15 Ro-Pax ships do not encounter other vessels. Clusters 6 and 13 are located in the same route, but the voyage is reversed between Helsinki and Tallinn. In cluster 6, 0.25 potential collisions per journey or one potential collision per 4.0 Ro-Pax journeys occur. However, 1.13 potential collisions per journey in cluster 13 are 4.52 times that those observed in cluster 6. The results show that the collision frequency is diverse in various voyages, even though they navigate in the same route.

3.4. Risk assessment

3.4.1. Collision risk index analysis during evasive action triggered

Potential collisions are detected based on grouped STs using the proposed method presented in Section 3.3. The aim of this section is to calibrate risk criteria to trigger evasive actions by quantitatively assessing CRI. An example is presented in Fig. 21. At point 29 of Fig. 21 DCPA, TCPA, and CRI are 0.78 nm, 7.8 min, 0.68 respectively. The probability density of the collision risk index is presented while evasive actions are triggered in Fig. 22. Rare evasive actions are taken with CRI smaller than 0.43. Only in 2.5% of the cases evasive actions are taken with CRI smaller than 0.45. Thus in 97.5% of the potential collisions, evasive actions are taken only when CRI reached at least 0.45. Fig. 23 shows that the intervals of CRI for 16 clusters lie within the interval [0.43 to 0.96].

This information could help provide essential guidance for triggering evasive actions in time. To validate the results of the detected potential collision scenarios, the TCPA and DCPA distributions are analyzed as shown in Fig. 24. Results confirm that if a struck ship's course falls into these eventualities, action should be taken to avoid collision (e.g., [57, 58], and [98]).

Fig. 25 shows the CRI distribution during evasive actions taken, indicating that most of the striking ships with the highest collision risk

Table 12
Hydro-meteorological parameters cumulative distributions during the potential collision.

	Hydro-meteorological conditions			
	25%	50%	75%	99%
Wave height [m]	0.1	0.4	0.9	2.5
Current speed[m/s]	0.01	0.02	0.05	0.3
Wind speed[m/s]	4.4	7.2	12.4	16.1
Swell height[m]	0.2	0.3	0.4	1.0

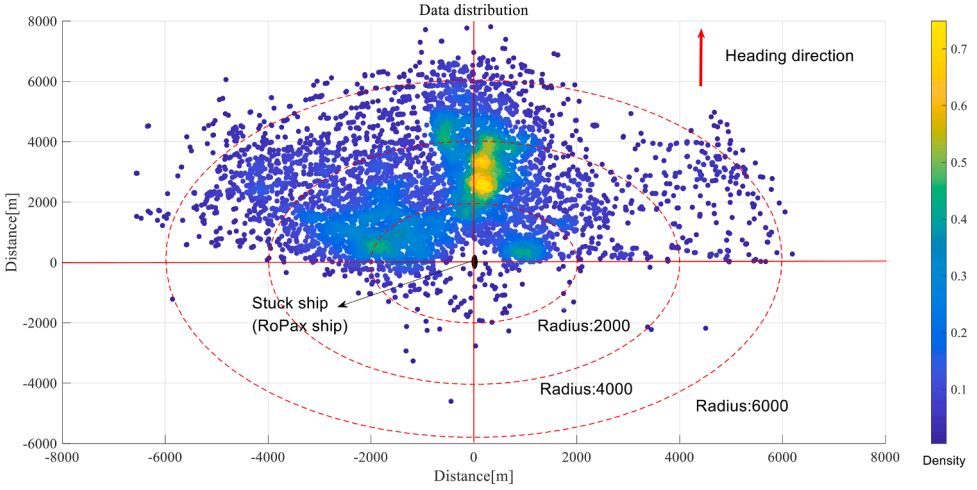


Fig. 19. The locations of striking ships, corresponding to potential collisions while obvious evasive actions are taken (blue scatters denote the relative locations of striking ships)

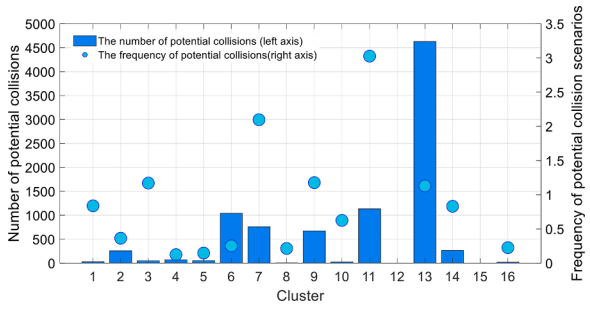


Fig. 20. Number and frequency of the potential collision scenarios by identified clusters

are located in front of the struck ships within [1 km to 2 km] radium. It can also be seen that most of striking ships with higher collision risk are located in front of the struck ships within [2 km to 4 km] radius in the heading direction, and some striking ships with lower collision risk are located in the stern of the struck ships and are located in front of the struck ships within [4 km to 6 km] radius. When comparing Fig. 19 and Fig. 25 the collision risk level distribution appears to be different in relation to the location density of striking ships. It may be therefore concluded that a higher collision risk area may lead to more serious accidents, and the location density of striking ships influences the number of potential collision locations related to the struck ship.

3.4.2. *Collision risk relationship among hydro-meteorological conditions*

To understand the dependence of CRI with evasive actions and hydro-meteorological conditions, correlation analysis was carried out using the approach of Pearson Correlation Coefficient (γ) [41] and Mutual Information (U) [49]. The method of the Pearson coefficient assumes normal data distributions. It is therefore thought to be sufficiently representative of positive or negative correlations and assumes linear relationship between CRI and hydro-meteorological conditions. On the other hand, MI is a measure of the mutual dependence between the two variables, which is more general and helps determine joint distributions. Not limited to real-valued random variables and linear dependence like the Pearson Correlation Coefficient. by using the MI test, the uncertainty coefficient (U) is calculated here that determines how large a proportion of the uncertainty about collision risk can be decreased by observing the hydro-meteorological condition variables. Table 14 summarizes statistical correlations.

Negative γ correlations imply that adverse hydro-meteorological conditions may be associated with decreased CRI value during the evasive actions triggered. The negative statistical correlations between CRI and wave height, wind speed, and swell height imply lower risk for encounters under adverse weather conditions when the bridge crew may

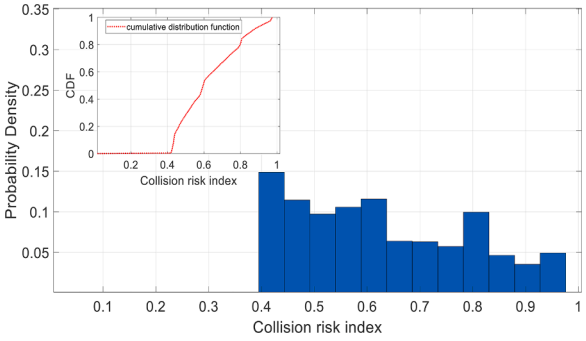


Fig. 22. The probability density of collision risk index

Table 13
The number of potential collision scenarios per journey based on the STs during the time period.

.	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8
Number of STs	37	712	41	570	354	4127	362	28
Number of potential collisions	31	259	48	71	52	1041	759	6
Frequency of potential collision scenarios*	0.84	0.36	1.17	0.12	0.15	0.25	2.10	0.21
Rank	7	9	4	14	13	10	2	12
No.	Cluster 9	Cluster 10	Cluster 11	Cluster 12	Cluster 13	Cluster 14	Cluster 15	Cluster 16
Number of STs	571	40	375	20	4098	319	10	84
Number of potential collisions	672	25	1135	0	4629	265	0	19
Frequency of potential collision scenarios*	1.18	0.63	3.03	0	1.13	0.83	0	0.23
Rank	3	8	1	/	5	6	/	11

*Note: The frequency of potential collision scenarios denotes the number of occurrences of potential collisions per voyage during the period.

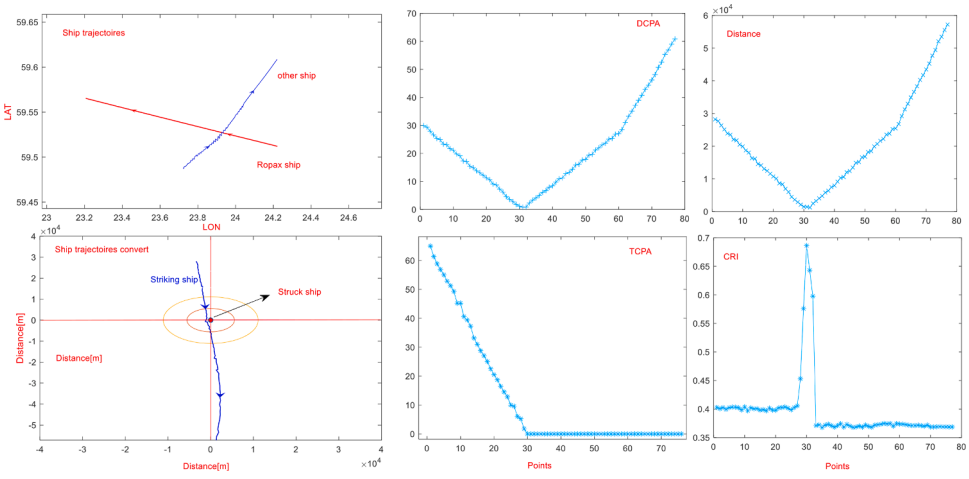


Fig. 21. Potential collision scenarios associated with DCPA, TCPA, distance and CRI

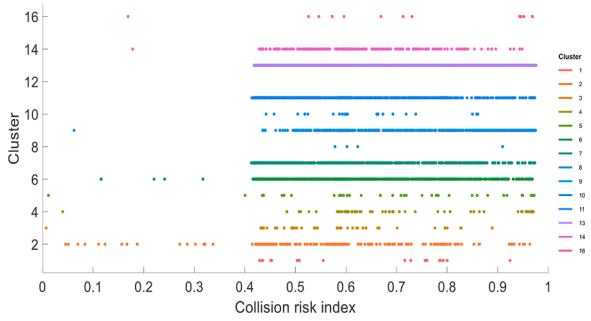


Fig. 23. Collision risk index mapping for 16 clusters

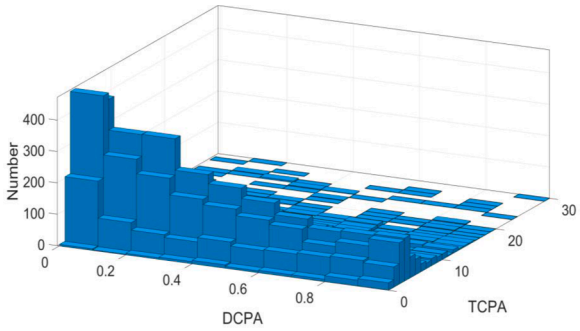


Fig. 24. The distributions of DCPA [nm] and TCPA [min] during evasive action taken

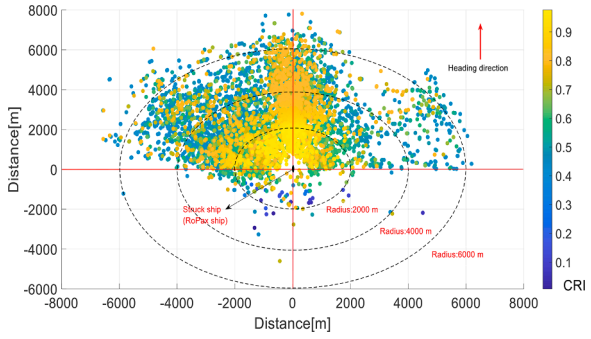


Fig. 25. Collision risk index distribution during evasive actions taken (the color denotes the value of collision risk index; the colorful scatter denotes the striking ships related to struck ship)

wish to initiate collision evasive actions at longer distance to the target, accounting for the effect of wave and wind on ship maneuverability. The value of γ correlations is low, showing that the correlation between collision risk and hydro-meteorological conditions is more complex instead of linear. Therefore, the MI test is employed, which reveals that by getting to know the hydro-meteorological condition in more detail. Thus, based on the results of this study and within its boundaries, the negative γ correlations and positive U correlation variation show are influencing factors related to the CRI value, affirming that adverse hydro-meteorological conditions evasive actions are associated with lower CRI in real operations. This finding may be supported by triggering the evasive actions in various hydro-meteorological conditions, showing that the give way ships should trigger the evasive actions with lower CRI value in adverse hydro-meteorological conditions. Notwithstanding, further studies are needed to quantify the effect of hydro-meteorological conditions on CRI in more detail.

Table 14
Correlations between collision risk and hydro-meteorological condition variables.

		Hydro-meteorological condition variables				Accepted Hypothesis
		Wave height	Current speed	Wind speed	Swell height	
Coefficient	γ	-0.0479	0.0644	-0.0292	-0.0847	(α=0.05)
	U	0.6892	0.6725	0.6946	0.6924	

4. Conclusions

The paper introduces a big data analytics method for evaluation ship-ship collision risk based on collision avoidance behaviors, with a RoRo/Passenger ship (RoPax) being considered as the struck ship. The big data analytics method introduced accounted for (1) A data mining model to cluster STs of struck ships using unsupervised machine learning algorisms (K-means and DB-SCAN); (2) the identification of time-dependent traffic situations and associated hydro-meteorological conditions at the times of potential collision in the different clusters; (3) ship collision risk assessment using CRI model during evasive action taken. The method is demonstrated using data covering a 13-month ice-free period in the Gulf of Finland, considering all large Ro-Pax ships (46,124 GT > Gross tonnage > 10,000 GT; 218.8 m > Length > 120 m). Key conclusions may be summarized as follows:

- The innovative use of the data mining method combining K-means and DB-SCAN for clustering struck STs is promising and useful for collision risk evaluation in more detail.
- Now-cast data and AIS data are useful for recovering detailed time-dependent traffic situations and the hydro-meteorological conditions at the times of unwanted events.
- The voyage may be the key influential factor contributing to collision risk, which is ignored in the traditional models (Fig. 20, and Table 13).
- Big data analytics help understand the location distribution of striking ships (Fig. 19) and the degree of collision risk during evasive actions taken in real operational conditions (Fig. 23,25), indicating that both higher collision risk hotspot areas and higher density hotspot areas should be considered to design remedial steps for collision avoidance.
- 97.5% of mentioned scenarios account for evasive actions when CRI is greater than 0.45 (Fig. 22). The CRI criteria outlined may provide important support to the master on Ro-Pax ships, as part of an intelligent decision support system for collision avoidance. However, the right time to take any evasive action is also influenced by other factors, e.g., hydro-meteorological conditions, ship navigation systems (specifically the autopilot and the ARPA radar), operational instructions, and procedures by the shipping company.
- Adverse hydro-meteorological conditions seem to decrease the CRI, indicating that the give way ships tend to take evasive actions earlier that in favorable hydro-meteorological conditions (see Table 14).
- **Mingyang Zhang:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing - original draft, Writing - review & editing, Visualization, project administration.
- **Spyros Hirdaris:** Conceptualization, Methodology, Writing - review & editing, Supervision, resources, visualization, funding acquisition
- **Teemu Manderbacka:** review, data curation
- **Jakub Montewka:** review, data curation, supervision.
- **Pentti Kujala:** supervision

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence

the work reported in this paper.

Acknowledgments

The work presented in this paper has been carried out within EU Horizon 2020 project FLARE (Grant No.: 814753-2). All authors acknowledge NAPA Ltd. for the provision of the AIS data used in the study case presented under Section 3 of this paper. The views set out in this paper are those of the authors and do not necessarily reflect the views of their respective organizations or of the FLARE consortium. Dr. Hirdaris acknowledges the financial support received from the Academy of Finland University competitive funding award (SA Profi 2-T20404) at the early stages of this work. Mr. Mingyang Zhang acknowledges the support of the Finnish Maritime Technology Foundation (Merenkulun Säätiö).

Appendix A: Interpolation method of hydro-meteorological data associated with AIS data

Hydro-meteorological data history records for each ship at different locations and global ocean now-cast records are reviewed. As part of this process we captured data streams with information on swell, wind, waves and sea currents. Swell and wind wave components are presented by significant wave height, wave zero-crossing period and wave direction. The trilinear interpolation method can be applied as appropriate, which contains the bilinear and linear interpolation using the equation (a 1-a 2). Fig. A 1 shows 3D view of this trilinear interpolation process. In the time dimension, the linear interpolation method is used to fit the timestamp of the hydro-meteorological stream delivered from Weather now-cast data database linking to the timestamp of the AIS data stream. Furthermore, in the space dimension, the hydro-meteorological data could be interpolated on the ship point of ST based on the latitude and longitude of the hydro-meteorological stream and AIS data stream, using bilinear interpolation.

$$\begin{aligned}\Delta \mathrm{T}_j &= \left(AIS_t^j - \mathrm{Time}_i\right)/\left(\mathrm{Time}_{i+1} - \mathrm{Time}_i\right)\\ \Delta Lon_j &= \left(AIS_{lon}^j - \mathrm{Lon}_i\right)/\left(\mathrm{Lon}_{i+1} - \mathrm{Lon}_i\right)\\ \Delta Lat_j &= \left(AIS_{lat}^j - \mathrm{Lat}_i\right)/\left(\mathrm{Lat}_{i+1} - \mathrm{Lat}_i\right)\end{aligned} \tag{a 1}$$

$$Hydro_j = \begin{cases} Hydro_{(i,i,i)}\left(1-\Delta \mathrm{T}_j\right)\left(1-\Delta \mathrm{Lon}_j\right)\left(1-\Delta \mathrm{Lat}_j\right)+\\ Hydro_{(i,i,i+1)}\Delta \mathrm{T}_j\left(1-\Delta \mathrm{Lon}_j\right)\left(1-\Delta \mathrm{Lat}_j\right)+\\ Hydro_{(i+1,i,i)}\left(1-\Delta \mathrm{T}_j\right)\Delta \mathrm{Lon}_j\left(1-\Delta \mathrm{Lat}_j\right)+\\ Hydro_{(i,i+1,i)}\left(1-\Delta \mathrm{T}_j\right)\left(1-\Delta \mathrm{Lon}_j\right)\Delta \mathrm{Lat}_j+\\ Hydro_{(i,i+1,i+1)}\Delta \mathrm{T}_j\left(1-\Delta \mathrm{Lon}_j\right)\Delta \mathrm{Lat}_j+\\ Hydro_{(i+1,i+1,i)}\left(1-\Delta \mathrm{T}_j\right)\Delta \mathrm{Lon}_j\Delta \mathrm{Lat}_j+\\ Hydro_{(i,i+1,i+1)}\Delta \mathrm{T}_j\Delta \mathrm{Lon}_j\left(1-\Delta \mathrm{Lat}_j\right)+\\ Hydro_{(i+1,i+1,i+1)}\Delta \mathrm{T}_j\Delta \mathrm{Lon}_j\Delta \mathrm{Lat}_j \end{cases} \tag{a 2}$$

Where, $\Delta \mathrm{T}_j$,ΔLon_j,ΔLat_jdenote the amount change of time, longitude, and latitude of the hydro-meteorological data stream, respectively; $Hydro_{(i,i,i)}$ presents the hydro-meteorological data stream at the location (Lon_i,Lat_i) at the time i; $Hydro_j$ presents the hydro-meteorological data stream at ship point p_j.

In specific, weather records included data in the following format:

- Wind speed and direction from US NOAA - https://www.noaa.gov/
- Wave height, period and direction, tidal current, water level from Tidetech - https://www.tidetech.org/
- Ocean current from Mercator Ocean - https://www.mercator-ocean.fr

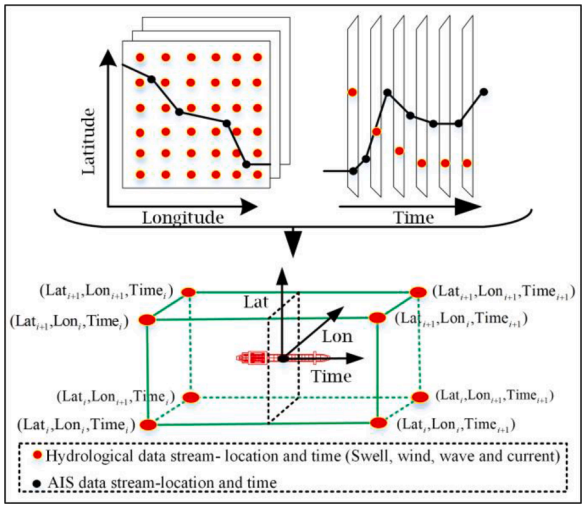


Fig. A1. Interpolation method of hydro-meteorological data from the ship perspective

Appendix B: Ship trajectory distance and CPA measures

Firstly, the coordinate system must be converted from the earth-fixed (AIS) to struck ship-fixed (see Fig. B 1).

Then, the dynamic data (location and speed) of the striking ships can be converted from in the new coordinate system defined by:

$$[x, y] = [X - X_0, Y - Y_0] \times \begin{bmatrix} \cos(TC) & \sin(TC) \\ -\sin(TC) & \cos(TC) \end{bmatrix} \tag{b 1}$$

$$[\dot{x}, \dot{y}] = \left[\dot{X}, \dot{Y}\right] \times \begin{bmatrix} \cos(TC) & \sin(TC) \\ -\sin(TC) & \cos(TC) \end{bmatrix} \tag{b 2}$$

$$A^{-} = \begin{bmatrix} \cos(TC) & \sin(TC) \\ -\sin(TC) & \cos(TC) \end{bmatrix} \tag{b 3}$$

$$\begin{cases} [u, v] = [U, V] \times A^{-} \\ r = ROT \end{cases} \tag{b 4}$$

where, $[x, y]$ and $[X, Y]$ are the coordinates of the struck ship in coordinate system XOY and xoy. $[X_o, Y_o]$ represents the origin of the coordinate system xoy. r denotes the value of ROT. TC denotes true course.

Finally, the minimum distance $d_{\min}(p_{j+i}^{k+i})$ is determined and calculated using the STs Tr_iandTr_{i+y}within the timestamp interval$[k-m,k+m]$(see in

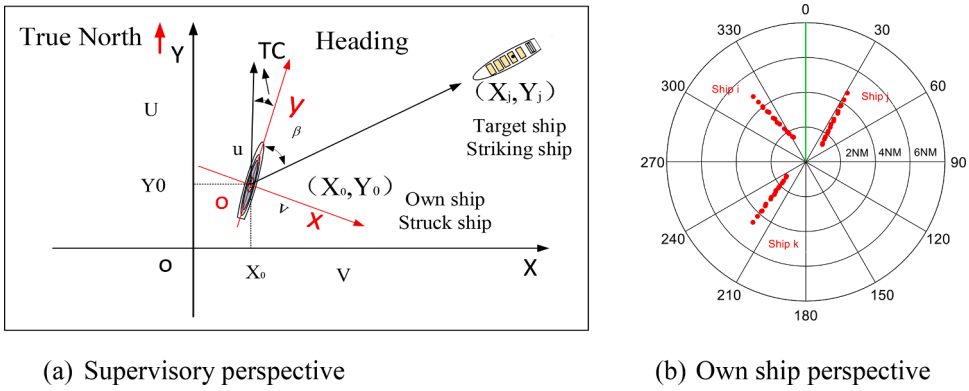


Fig. B1. The coordinate system of striking ship and struck ship

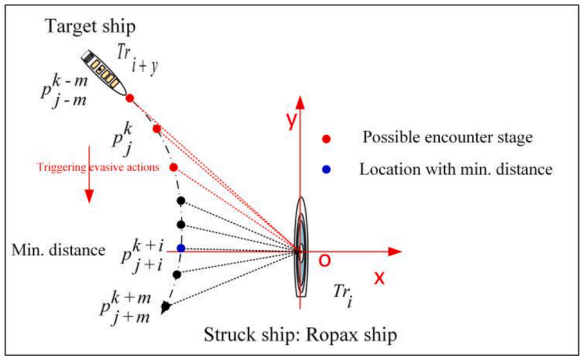


Fig. B2. Relations between value to be optimized and the value of p_{j+i}^{k+i} that is to be found a minimum

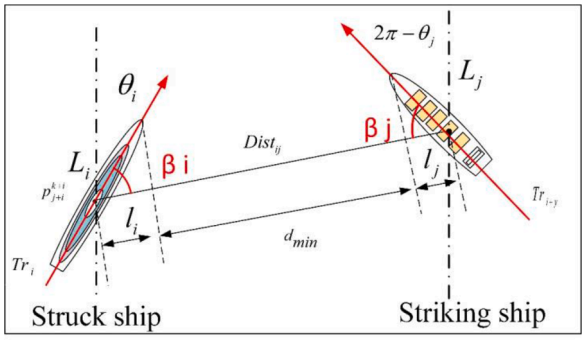


Fig. B3. The distance of the striking ship and struck ship

Fig. B 2). Aforementioned, the possible encounter stages are identified based on STs $p_{j\pm m}^{k\pm m}$from Tr_iandTr_{i+y}at the time $k \pm m$associated with the minimum distance $d_{\min}(p_{j+i}^{k+i})$ between pairs of ships.

Distances are calculated considering the location of AIS onboard (Fig. B 3) and the length of ships, as follows:

$$\beta_i = \arccos\left(\frac{y_j - y_i}{\sqrt{(x_j - x_0)^2 + (y_j - y_0)^2}}\right) - \theta_i \tag{b 5}$$

$$\beta_j = \theta_j - \theta_i - \beta_i - \pi \tag{b 6}$$

$$l_i = \frac{4}{5} L_i COS(\beta_i) \tag{b 7}$$

$$l_j = \frac{4}{5} L_j COS(\theta_j - \theta_i - \beta_i - \pi) \tag{b 8}$$

$$d_{\min} = Dist_{ij} - l_i - l_j \tag{b 9}$$

where β_i is the relative bearing angle from striking to struck; θ_i is the course of the encountered ships; (x_0, y_0) and (x_j, y_j) are the locations of two ships; $Dist_{ij}$ is the distance between the reference of AIS positions of two ships. The coefficients of equations (b 7) -(b 8) are defined based on AIS positions on the ship [30].

The DCPA and TCPA can be calculated based on the equations (b 10) -(b 13).

$$S_r = \sqrt{sog_o^2 + sog_T^2 + 2sog_o \times sog_T cos(cog_T - cog_0)} \tag{b 10}$$

$$C_r = \begin{cases} cog_O - arcos\left(\frac{S_r + sog_o^2 - sog_T^2}{2S_r \times sog_o}\right), cog_O < cog_T \\ cog_O + arcos\left(\frac{S_r + sog_o^2 - sog_T^2}{2S_r \times sog_o}\right), cog_O > cog_T \end{cases} \tag{b 11}$$

$$d_{CPA} = Dist \times (sin(C_r - cog_O - \beta - \pi)) \tag{b 12}$$

$$t_{CPA} = Dist \times \left(\frac{cos(C_r - cog_O - \beta - \pi)}{S_r}\right) \tag{b 13}$$

Where, $C_r$represents the relative angle. S_r denotes the relative speed.

References

[1] Ahn JH, Rhee KP, You YJ. A study on the collision avoidance of a ship using neural networks and fuzzy logic. Applied Ocean Research 2012;37:162–73.
[2] Antao P, Soares CG. Fault-tree models of accident scenarios of RoPax vessels. International Journal of Automation and Computing 2006;3(2):107–16.
[3] Bergström M, Erikstad SO, Ehlers S. Assessment of the applicability of goal-and risk-based design on Arctic sea transport systems. Ocean Engineering 2016;128:183–98.
[4] Bidlot J-R. Intercomparison of operational wave forecasting systems against buoys: data from ECMWF, MetOffice, FNMOC, MSC, NCEP, MeteoFrance, DWD, BoM, SHOM, JMA, KMA, Puerto del Estado, DMI, CNR-AM, METNO, SHN-SM January 2016 to December 2016. European Centre for Medium-range Weather Forecasts (ECMWF) 2017.
[5] Bouveyron C, Celeux G, Murphy TB, Raftery AE. Model-based clustering and classification for data science: with applications in R (Vol. 50). Cambridge University Press; 2019.
[6] Bukhari AC, Tusseyeva I, Kim YG. An intelligent real-time multi-vessel collision risk assessment system from VTS view point based on fuzzy inference system. Expert systems with applications 2013;40(4):1220–30.
[7] Bye RJ, Aalberg AL. Maritime navigation accidents and risk indicators: An exploratory statistical analysis using AIS data and accident reports. Reliability Engineering & System Safety 2018;176:174–86.
[8] Cai W, Zhao J, Zhu M. A real time methodology of cluster-system theory-based reliability estimation using k-means clustering. Reliability Engineering & System Safety 2020:107045.
[9] Chauvin C, Lardjane S, Morel G, Clostermann JP, Langard B. Human and organizational factors in maritime accidents: Analysis of collisions at sea using the HFACS. Accident Analysis & Prevention 2013;59:26–37.
[10] Chen P, Huang Y, Mou J, van Gelder PHAJM. Ship collision candidate detection method: A velocity obstacle approach. Ocean Engineering 2018;170:186–98.
[11] Chen P, Huang Y, Mou J, van Gelder PHAJM. Probabilistic risk analysis for ship-ship collision: state-of-the-art. Safety science 2019;117:108–22.
[12] Chen P, Huang Y, Papadimitriou E, Mou J, van Gelder PHAJM. An improved time discretized non-linear velocity obstacle method for multi-ship encounter detection. Ocean Engineering 2020;196:106718.
[13] Christian R, Kang HG. Probabilistic risk assessment on maritime spent nuclear fuel transportation (Part II: Ship collision probability). Reliability Engineering & System Safety 2017;164:136–49.
[14] Dinis D, Teixeira AP, Soares CG. Probabilistic approach for characterising the static risk of ships using Bayesian networks. Reliability Engineering & System Safety 2020;203:20.
[15] Du L, Goerlandt F, Kujala P. Review and analysis of methods for assessing maritime waterway risk based on non-accident critical events detected from AIS data. Reliability Engineering & System Safety 2020:106933.
[16] Fang Z, Yu H, Ke R, Shaw SL, Peng G. Automatic identification system-based approach for assessing the near-miss collision risk dynamics of ships in ports. IEEE Transactions on Intelligent Transportation Systems 2018;20(2):534–43.
[17] Gang L, Wang Y, Sun Y, Zhou L, Zhang M. Estimation of vessel collision risk index based on support vector machine. Advances in Mechanical Engineering 2016;8(11):1687814016671250.
[18] Gil M, Montewka J, Krata P, Hinz T, Hirdaris S. Semi-dynamic ship domain in the encounter situation of two vessels. In: Developments in the Collision and Grounding of Ships and Offshore Structures: Proceedings of the 8th International Conference on Collision and Grounding of Ships and Offshore Structures (ICCGS 2019), 21-23 October, 2019. Lisbon, Portugal: CRC Press; October, 2019. p. 301.
[19] Gil M, Montewka J, Krata P, Hinz T, Hirdaris S. Determination of the dynamic critical maneuvering area in an encounter between two vessels: Operation with negligible environmental disruption. Ocean Engineering 2020;213:107709.
[20] Goerlandt F, Kujala P. Traffic simulation based ship collision probability modeling. Reliability Engineering & System Safety 2011;96(1):91–107.
[21] Goerlandt F, Kujala P. On the reliability and validity of ship–ship collision risk analysis in light of different perspectives on risk. Safety science 2014;62:348–65.
[22] Goerlandt F, Montewka J. Maritime transportation risk analysis: Review and analysis in light of some foundational issues. Reliability Engineering & System Safety 2015;138:115–34.
[23] Goerlandt F, Montewka J, Kuzmin V, Kujala P. A risk-informed ship collision alert system: framework and application. Safety Science 2015;77:182–204.
[24] Goerlandt F, Montewka J, Zhang W, Kujala P. An analysis of ship escort and convoy operations in ice conditions. Safety science 2017;95:198–209.
[25] Graziano A, Teixeira AP, Guedes Soares C. Classification of human errors in grounding and collision accidents using the TRACEr taxonomy. Safety Science 2016;86:245–57.

[26] Hanninen M, Kujala P. Influences of variables on ship collision probability in a Bayesian belief network model. Reliability Engineering & System Safety 2012; 102:27–40.
[27] Haranen M, Myöhänen S, Cristea DS. The Role of Accurate Now-Cast Data in Ship Efficiency Analysis. In: 2nd Hull Performance & Insight Conference; 2017. p. 25–38.
[28] Harrald JR, Mazzuchi TA, Spahn J, Van Dorp R, Merrick J, Shrestha S, Grabowski M. Using system simulation to model the impact of human error in a maritime system. Safety Science 1998;30(1-2):235–47.
[29] He Y, Jin Y, Huang L, Xiong Y, Chen P, Mou J. Quantitative analysis of COLREG rules and seamanship for autonomous collision avoidance at open sea. Ocean Engineering 2017;140:281–91.
[30] Hirdaris SE, Zhang M, Montewka J, Manderbacka T. Analysis of Routing and Traffic Data. EU Project FLARE WP2 2019. Deliverable Report D2.4.
[31] Hu Y, Zhang A, Tian W, Zhang J, Hou Z. Multi-Ship Collision Avoidance Decision-Making Based on Collision Risk Index. Journal of Marine Science and Engineering 2020;8(9):640.
[32] Huang Y, van Gelder PHAJM. Collision risk measure for triggering evasive actions of maritime autonomous surface ships. Safety science 2020;127:104708.
[33] Huang Y, Chen L, Chen P, Negenborn RR, van Gelder PHAJM. Ship collision avoidance methods: State-of-the-art. Safety science 2020;121:451–73.
[34] Hulkkonen T, Manderbacka T, Sugimoto K. Digital Twin for Monitoring Remaining Fatigue Life of Critical Hull Structures. In: 18th Conference on Computer Applications and Information Technology in the Maritime Industries (COMPIT2019), 25-27 March 2019; 2019.
[35] IMO. COLREG: Convention on the International Regulations for Preventing Collisions at Sea, 1972. International Maritime Organization; 2003. Available at, http://books.google.com/books?id=_ZkZAQAAIAAJ&pgis=1. Accessed: 2 April 2020.
[36] Jiang D, Wu B, Cheng ZY, Xue J, van Gelder P. Towards a probabilistic model for estimation of grounding accidents in fluctuating backwater zone of the Three Gorges Reservoir. Reliability Engineering & System Safety 2021;205:16.
[37] Kelangath S, Das PK, Quigley J, Hirdaris SE. Risk analysis of damaged ships–a data-driven Bayesian approach. Ships and Offshore Structures 2012 Sep 1;7(3): 333–47.
[38] Kim SJ, Kõrgersaar M, Ahmadi N, Taimuri G, Kujala P, Hirdaris S. The influence of fluid structure interaction modelling on the dynamic response of ships subject to collision and grounding. Marine Structures 2021;75:102875.
[39] Kujala P, Hanninen M, Arola T, Ylitalo J. Analysis of the marine traffic safety in the Gulf of Finland. Reliability Engineering & System Safety 2009;94(8):1349–57.
[40] Kumar KS, Manigandan T, Chitra D, Murali L. Object recognition using Hausdorff distance for multimedia applications. Multimedia Tools and Applications 2020;79 (5):4099–114.
[41] Kumar M. Measuring Pearson's correlation coefficient of fuzzy numbers with different membership functions under weakest t-norm. International Journal of Data Analysis Techniques and Strategies 2020;12(2):172–86.
[42] Langard B, Morel G, Chauvin C. Collision risk management in passenger transportation: A study of the conditions for success in a safe shipping company. Psychologie française 2015;60(2):111–27.
[43] Li H, Liu J, Wu K, Yang Z, Liu RW, Xiong N. Spatio-temporal vessel trajectory clustering based on data mapping and density. IEEE Access 2018;6:58939–54.
[44] Lopez-Santander A, Lawry J. An ordinal model of risk based on mariner's judgement. The Journal of Navigation 2017;70(2):309–24.
[45] Łukasik S, Kowalski PA, Charytanowicz M, Kulczycki P. Clustering using flower pollination algorithm and Calinski-Harabasz index. In: 2016 IEEE Congress on Evolutionary Computation (CEC). IEEE; July, 2016. p. 2724–8.
[46] Manderbacka T. On the uncertainties of the weather routing and support system against dangerous conditions. In: Proceedings of the 17th International Ship Stability Workshop (ISSW2019), 10-12 June 2019, Helsinki, Finland; 2019.
[47] Martins MR, Maturana MC. Human error contribution in collision and grounding of oil tankers. Risk Analysis: An International Journal 2010;30(4):674–98.
[48] Martins MR, Maturana MC. Application of Bayesian Belief networks to the human reliability analysis of an oil tanker operation focusing on collision accidents. Reliability Engineering & System Safety 2013;110:89–109.
[49] Mazaheri A, Montewka J, Kotilainen P, Sormunen OVE, Kujala P. Assessing grounding frequency using ship traffic and waterway complexity. The Journal of Navigation 2015;68(1):89–106.
[50] Mazurek J, Montewka J, Smolarek L. A simulation model to support planning of resources to combat oil spills at sea. In Developments in the Collision and Grounding of Ships and Offshore Structures. In: Proceedings of the 8th International Conference on Collision and Grounding of Ships and Offshore Structures (ICCGS 2019), 21-23 October, 2019. Lisbon, Portugal: CRC Press; October, 2019. p. 355.
[51] Merrick JR, Van Dorp JR, Blackford JP, Shaw GL, Harrald J, Mazzuchi TA. A traffic density analysis of proposed ferry service expansion in San Francisco Bay using a maritime simulation model. Reliability Engineering & System Safety 2003;81(2):119–32.
[52] Montewka J, Ehlers S, Goerlandt F, Hinz T, Tabri K, Kujala P. A framework for risk assessment for maritime transportation systems—A case study for open sea collisions involving RoPax vessels. Reliability Engineering & System Safety 2014; 124:142–57.
[53] Montewka J, Gil M, Wróbel K. Discussion on the article by Zhang & Meng entitled "Probabilistic ship domain with applications to ship collision risk assessment. Ocean Engineering; 2020, 107527.
[54] Montewka J, Goerlandt F, Innes-Jones G, Owen D, Hifi Y, Puisa R. Enhancing human performance in ship operations by modifying global design factors at the design stage. Reliability Engineering & System Safety 2017;159:283–300.
[55] Montewka J, Hinz T, Kujala P, Matusiak J. Probability modelling of vessel collisions. Reliability Engineering & System Safety 2010;95(5):573–89.
[56] Montewka J, Krata P, Goerlandt F, Mazaheri A, Kujala P. Marine traffic risk modelling–an innovative approach and a case study. Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability 2011; 225(3):307–22.
[57] Ożoga B, Montewka J. Towards a decision support system for maritime navigation on heavily trafficked basins. Ocean Engineering 2018;159:88–97.
[58] Perera LP, Soares CG. Collision risk detection and quantification in ship navigation with integrated bridge systems. Ocean Engineering 2015;109:344–54.
[59] Ramos MA, Thieme CA, Utne IB, Mosleh A. Human-system concurrent task analysis for maritime autonomous surface ship operation and safety. Reliability Engineering & System Safety 2020;195:21.
[60] Ramos MA, Utne IB, Mosleh A. Collision avoidance on maritime autonomous surface ships: Operators' tasks and human failure events. Safety science 2019; 116:33–44.
[61] Rawson A, Brito M. A critique of the use of domain analysis for spatial collision risk assessment. Ocean Engineering; 2020, 108259.
[62] Rong H, Teixeira AP, Soares CG. Ship trajectory uncertainty prediction based on a Gaussian Process model. Ocean Engineering 2019;182:499–511.
[63] Rong H, Teixeira AP, Soares CG. Data mining approach to shipping route characterization and anomaly detection based on AIS data. Ocean Engineering 2020;198:106936.
[64] Statheros T, Howells G, Maier KM. Autonomous ship collision avoidance navigation concepts, technologies and techniques. The Journal of Navigation 2008;61(1):129–42.
[65] Su H, Liu S, Zheng B, Zhou X, Zheng K. A survey of trajectory distance measures and performance evaluation. The VLDB Journal 2020;29(1):3–32.
[66] Szlapczynski R. A unified measure of collision risk derived from the concept of a ship domain. The Journal of navigation 2006;59(3):477.
[67] Szlapczynski R, Krata P, Szlapczynska J. A ship domain-based method of determining action distances for evasive manoeuvres in stand-on situations. Journal of Advanced Transportation 2018. 2018.
[68] Szlapczynski R, Krata P, Szlapczynska J. Ship domain applied to determining distances for collision avoidance manoeuvres in give-way situations. Ocean Engineering 2018;165:43–54.
[69] Talavera A, Aguasca R, Galván B, Cacereño A. Application of Dempster–Shafer theory for the quantification and propagation of the uncertainty caused by the use of AIS data. Reliability Engineering & System Safety 2013;111:95–105.
[70] Van de Wiel G, van Dorp JR. An oil outflow model for tanker collisions and groundings. Annals of Operations Research 2011;187(1):279–304.
[71] Wang S, Zhang Y, Li L. A collision avoidance decision-making system for autonomous ship based on modified velocity obstacle method. Ocean Engineering 2020;215:107910.
[72] Wang T, Wu Q, A Diaconeasa M, Yan X, Mosleh A. On the use of the hybrid causal logic methodology in ship collision risk assessment. Journal of Marine Science and Engineering 2020;8(7):485.
[73] Wang T, Wu Q, Zhang J, Wu B, Wang Y. Autonomous decision-making scheme for multi-ship collision avoidance with iterative observation and inference. Ocean Engineering 2020;197:106873.
[74] Wang Y, Zhang J, Chen X, Chu X, Yan X. A spatial–temporal forensic analysis for inland–water ship collisions using AIS data. Safety science 2013;57:187–202.
[75] Wen Y, Huang Y, Zhou C, Yang J, Xiao C, Wu X. Modelling of marine traffic flow complexity. Ocean Engineering 2015;104:500–10.
[76] Woerner K, Benjamin MR, Novitzky M, Leonard JJ. Quantifying protocol evaluation for autonomous collision avoidance. Autonomous Robots 2019;43(4): 967–91.
[77] Wu B, Cheng T, Yip TL, Wang Y. Fuzzy logic based dynamic decision-making system for intelligent navigation strategy within inland traffic separation schemes. Ocean Engineering 2020;197:106909.
[78] Xie S, Chu X, Zheng M, Liu C. Ship predictive collision avoidance method based on an improved beetle antennae search algorithm. Ocean Engineering 2019;192: 106542.
[79] Yoo SL. Near-miss density map for safe navigation of ships. Ocean Engineering 2018;163:15–21.
[80] Yu Q, Liu KZ, Chang CH, Yang ZL. Realising advanced risk assessment of vessel traffic flows near offshore wind farms. Reliability Engineering & System Safety 2020;203:18.
[81] Zhang D, Yan XP, Yang ZL, Wall A, Wang J. Incorporation of formal safety assessment and Bayesian network in navigational risk estimation of the Yangtze River. Reliability Engineering & System Safety 2013;118:93–105.
[82] Zhang W, Goerlandt F, Montewka J, Kujala P. A method for detecting possible near miss ship collisions from AIS data. Ocean Engineering 2015;107:60–9.
[83] b Zhang W, Montewka J, Goerlandt F. Semi-qualitative method for ship collision risk assessment. In: Nowakowski In:, editor. Safety and Reliability; Methodology and Applications. London: Taylor and Francis Group; 2015. p. 1563–72.
[84] Zhang L, Wang H, Meng Q. Big data–based estimation for ship safety distance distribution in port waters. Transportation research record 2015;2479(1):16–24.
[85] Zhang W, Goerlandt F, Kujala P, Wang Y. An advanced method for detecting possible near miss ship collisions from AIS data. Ocean Engineering 2016;124: 141–56.
[86] Zhang L, Meng Q. Probabilistic ship domain with applications to ship collision risk assessment. Ocean Engineering 2019;186:106130.

[87] Zhang M, Zhang D, Goerlandt F, Yan X, Kujala P. Use of HFACS and fault tree model for collision risk factors analysis of icebreaker assistance in ice-covered waters. Safety science 2019;111:128–43.
[88] Zhang W, Feng X, Goerlandt F, Liu Q. Towards a Convolutional Neural Network model for classifying regional ship collision risk levels for waterway risk analysis. Reliability Engineering & System Safety 2020:107127.
[89] Zhang M, Zhang D, Yao H, Zhang K. A probabilistic model of human error assessment for autonomous cargo ships focusing on human–autonomy collaboration. Safety Science 2020;130:104838.
[90] Zhang M, Montewka J, Manderbacka T, Kujala P, Hirdaris S. Analysis of the Grounding Avoidance Behavior of a Ro-Pax Ship in the Gulf of Finland using Big Data. In: The 30th International Ocean and Polar Engineering Conference. International Society of Offshore and Polar Engineers; 2020.
[91] Zhao L, Roh MI. COLREGs-compliant multi-ship collision avoidance based on deep reinforcement learning. Ocean Engineering 2019;191:106436.
[92] Zhao L, Shi G. A trajectory clustering method based on Douglas-Peucker compression and density for marine traffic pattern recognition. Ocean Engineering 2019;172:456–67.
[93] Zhao L, Shi G, Yang J. An adaptive hierarchical clustering method for ship trajectory data based on DBSCAN algorithm. In: 2017 IEEE 2nd International Conference on Big Data Analysis; 2017. p. 329–36.
[94] Zhao Y, Li W, Shi P. A real-time collision avoidance learning system for Unmanned Surface Vessels. Neurocomputing 2016;182:255–66.
[95] Zhen R, Jin Y, Hu Q, Shao Z, Nikitakos N. Maritime anomaly detection within coastal waters based on vessel trajectory clustering and Naïve Bayes Classifier. The Journal of Navigation 2017;70(3):648.
[96] Zheng Z, Wu Z. New model of collision risk between vessels. Journal of Dalian Maritime University 2002;28(2):1–5.
[97] Zhou Y, Daamen W, Vellinga T, Hoogendoorn SP. Ship classification based on ship behavior clustering from AIS data. Ocean Engineering 2019;175:176–87.
[98] Zhang J, Zhang D, Yan X, Haugen S, Soares CG. A distributed anti-collision decision support formulation in multi-ship encounter situations under COLREGs. Ocean Engineering 2015;105:336–48.
[99] Du L, Banda OAV, Goerlandt F, Huang Y, Kujala P. A COLREG-compliant ship collision alert system for stand-on vessels. Ocean Engineering 2020;218:107866.
[100] Kelangath Subin, Das Purnendu K, Quigley John, Hirdaris Spyros. Risk analysis of damaged ships – a data-driven Bayesian approach. Ships and Offshore Structures 2012;7(3):333–47. https://doi.org/10.1080/17445302.2011.592358.