The Journal of Navigation (2023), **76**:4–5 525–544
doi:10.1017/S0373463323000231

CAMBRIDGE
UNIVERSITY PRESS

RESEARCH ARTICLE

A framework of marine collision risk identification strategy using AIS data

Xiaofei Ma,[1,2] Guoyou Shi,[1,2]* Jiahui Shi,[1] and Jiao Liu[1,2]

[1] Navigation College, Dalian Maritime University, Dalian, China
[2] Key Laboratory of Navigation Safety Guarantee of Liaoning Province, Dalian, China.
*Corresponding author: Guoyou Shi; Email: sgydmu@163.com

Received: 15 January 2022; **Accepted:** 8 August 2023; **First published online:** 16 November 2023

Keywords: Collision risk identification; AIS data; ship domain; maritime safety

Abstract
Collisions are one of the major accidents in the shipping industry, causing significant losses. In this work, a framework of marine collision risk identification strategy was developed to quantitatively analyse collision risks and provide an easy and convenient way to monitor traffic flow in relevant waters to mitigate the chances of collision. The model was verified by using automatic identification system data obtained from Tianjin Port. When compared to previous research, the proposed model can identify risks earlier and give people more time to analyse and take action. The results indicate that it also can provide a visual display to alert relevant personnel. The model can be used as a reference to identify potential collision risks or as an information source for future research.

1. Introduction

Safety has always been the top priority in the shipping industry, particularly in recent years. Prioritising safety and preventing accidents will boost industry personnel confidence and save many lives. Accidental costs to personnel, equipment, the economy and the environment, are often unaffordable. Marine traffic accidents are less common than road traffic accidents and have been significantly reduced in recent years because of using advanced technology (Luo and Shin, 2019). However, the frequency of marine accidents, particularly accidents between ships, has no possibility of declining to zero due to human error. Various reports, such as those from the Marine Accident Investigation Branch or the European Maritime Safety Agency, have shown that collision is the most common type of marine accident. Therefore, scholars have paid much attention to collision avoidance research and many related models and techniques have been developed. Some focused on the collision accident per se, while others chose to study non-accident events (i.e. collision risk, near misses, close-quarter situations and immediate danger). These non-accident events are critical to accident prevention research because scholars (Majumdar et al., 2021) have proved that numerous near-misses will eventually lead to an actual accident. Some shipping companies rewarded $10 per near-miss event to any crew who reported it, but only a maximum of five such reports are rewarded monthly (i.e. Thome Ship Management Pte Ltd., http://thome.com/). Because the company also believed that these research findings are effective in reducing accidents, it is of crucial importance for ship owners and other stakeholders to have effective ways of analysing the collision risks and improving the safety of maritime transportation in different waters.

https://doi.org/10.1017/S0373463323000231 Published online by Cambridge University Press

1.1. *Literature review*

Although potential collision risk is a relatively new field for marine traffic research, it has a long history in other fields, such as road traffic. The traffic conflict technique (TCT) was developed by road traffic research, with an early foundation laid by Chin and Quek (1997). Later, Debnath and Chin (2010) proposed a nautical traffic conflict technique as an improved approach to the TCT. For marine traffic, collision risk or ‘risk of collision’ appeared in the Convention on the International Regulations for Preventing Collisions at Sea (COLREGS)[1] many times, but no specific explanations about the concept of collision risk exist. Alternatively, some articles outlined the relevant definitions. Li and Pang (2013) proposed a vessel collision risk assessment model based on the Dempster–Shafer theory of evidence. The collision risk was explained as a combination of the probability and consequence of ship’s collision. A multi-radar network was used to manage the complex relationship between collision risk and the distance to the closest point of approach (DCPA), the time to the closest point of approach (TCPA) and other distances. Simsir et al. (2014) presented a decision support system to help the vessel traffic service (VTS) to monitor and guide vessels transiting the Istanbul Strait. The data from passing vessels were used to train an artificial neural network that can predict the subsequent 3-min trajectory of each vessel. Goerlandt and Montewka (2015a) created a two-stage risk assessment model to analyse maritime transportation risks. First, they use a Bayesian network model to quantify the risk probability. Then, they give an assessment of uncertainty based on the first stage. This model helped to evaluate tanker collision oil spill risks after the case study was implemented. Moreover, the risk was believed as a concept acted on a system in the evaluator’s mind, instead of a physical property that exists in the system itself. Zhang et al. (2016) proposed a model that uses automatic identification system (AIS) data to differentiate the severity of a ship’s encounter. Szlapczynski and Szlapczynska (2016) gave analytical formulas for domain-based collisions that can upgrade the use of the ship domain in real-time systems and significantly reduce the processing time. Chai et al. (2017) developed a quantitative risk assessment model to evaluate ship collision risks. This model considers the frequency and consequences of possible accidents and includes a collision frequency estimation model, an event tree and a consequence estimation model. The accident assessment is well represented by the consequence estimation model. Zhang et al. (2018) created a collision probability model using Bayesian rules and the least-squares estimation method to assess the risk-influencing factors that cause collisions. Seven factors were determined from the model’s analysis and test. Zhang and Meng (2019) proposed a probabilistic ship domain with a vague domain boundary rather than a crisp value with the desired boundary. Jon et al. (2021) proposed a safety criterion based on the risk assessment of marine accidents. The assessment was done by combining a Markov model and a Markov chain Monte Carlo simulation. The criteria were determined based on the earlier accident data. Zhao et al. (2021) used a data-driven Bayesian network based on accident data from the Yangtze River to determine the risk factors for marine accidents involving unmanned surface vessels (USVs). They built a network to determine how the future occurrence of maritime accidents, such as collisions, can be reduced as crews are removed from USVs, but accidents, such as fire and extreme weather, can be even worse. Relevant literature reviews (Goerlandt and Montewka, 2015b; Lim et al., 2018; Čorić et al., 2021) provide a more comprehensive view of this field. It is worth noting that some research approach the issues qualitatively, while others address them quantitatively.

A word cloud (Figure 1) was created as a hot index from the 103 papers we collected on this topic. The larger the word, the more frequently it appears. Therefore, the top keyword is ‘collision risk’, and other collision-related terms are also frequently used. This image highlights the current state of the research and indicates the research enthusiasm for collision risk.

The model proposed by Zhang et al. (2016) was considered adequate for detecting potential collision risks between vessels. However, the adopted ship domain, named the Fujii model, was deemed too risky for navigators and supervisors to use when they needed to take action to avoid potential collisions (Wang et al., 2009). In the Fujii model, the major axis of the ellipse domain is only 4 times the ship’s length,

[1] The COLREGS are the International Regulations for Preventing Collisions at Sea. They are the maritime traffic rules that shall apply to all vessels upon the high seas and in all waters connected therewith navigable by seagoing vessels.

https://doi.org/10.1017/S0373463323000231 Published online by Cambridge University Press

Figure 1. Word cloud of collected papers.

and the minor axis is 1 · 6 times the ship's length. In other words, this model does not give sufficient time for navigators and/or supervisors to react before the collision; hence, despite being suitable for analysis, this model does not practically improve safety.

By addressing the aforementioned problem, this study has made some improvements to the research conducted by Zhang et al. (2016) by using a different ship domain, the Goodwin model. Unlike traditional risk assessment models often involved the probability and consequence of an accident, this model does not use accident probability as an input. Some further improvements included the AIS data-cleaning process, the analysis of overtaking situations with the new model and considering the situation where the observing vessel is located in the boundary of the observed vessel's domain (where $d = 0$, see Section 3.2).

The rest of the paper is arranged as follows: Section 2 presents the associated conceptual basis. Section 3 describes the mathematical model using the Goodwin domain. Section 4 outlines the testing and verification of the collision risk model. Section 5 presents the discussion and concluding remarks.

2. Conceptual basis

2.1. Automatic identification system

AIS is a commonly used system in the shipping industry because it is a major navigational element. AID data can provide nearby vessels' navigational status (e.g. visible/invisible), particularly occluded

https://doi.org/10.1017/S0373463323000231 Published online by Cambridge University Press

Table 1. *Major topics using AIS data.*

Major topic	Research articles
Ship trajectory optimisation	(Wei et al., 2020; Gao et al., 2021)
Marine accident investigation	(Bye and Aalberg, 2018; Bye and Almklov, 2019)
Collision avoidance optimisation	(Gao and Shi, 2020; Murray and Perera, 2021)
Collision probability research	(Altan and Otay, 2018; Chen et al., 2019)
Maritime risk model	(Hörteborn and Ringsberg, 2021; Zhao and Fu, 2021)
Ship domain development	(Du et al., 2021; Liu et al., 2021)
Potential risk detection	(Zhang et al., 2015; Szlapczynski and Szlapczynska, 2016; Zhang et al., 2016)

ships' information, which is considered an important supplement to automatic radar plotting aid (ARPA) function. AIS data are used for a variety of research purposes, as shown in Table 1.

AIS data are vital information for the duty officer, the Maritime Safety Administration (MSA), VTS and port control. Vessels equipped with AIS can detect other vessels equipped with AIS because AIS data are automatically exchanged between the AIS equipment. AIS data contain two kinds of information: static information and dynamic information. Static information includes the ship's name, maritime mobile service identity (MMSI) number, call sign, IMO number, length, width and so on. Dynamic information includes the ship's position, heading, speed, course, cargo type, onboard crew number and destination, amongst other things. VTS, as a shore facility, can store the AIS data in a particular area which will be helpful for future studies and relevant investigations.

However, a large number of illogical and abnormal data are generated while AIS is in operation, whether on the transmitting side or the receiving side. The reasons can be summarised as follows (Liu et al., 2020): (a) signal error, (b) equipment failure or (c) network error. Because noisy data affects the accuracy of subsequent calculations, AIS data processing must be performed prior to analysis.

This paper uses AIS data from Tianjin Port for the model's calculations and evaluation. The data set comprises 1044857 original AIS data from 3383 ships in January 2015. Figure 2 shows the distribution of the original AIS data. There are many abnormal AIS data values; some data points are even spread across the land, which is clearly an error. As a result, data cleaning is required for accurate calculations.

Since the original AIS data were not properly organised, we sorted them based on the database as follows:

$$D = \{V_1, V_2, V_3, \cdots V_i\}, i = 1, 2, \cdots, n \tag{1}$$

$$V = \{Name, MMSI, Type, Length, Width, t_j, c_j, s_j, p_j\}, j = 1, 2, \cdots, k \tag{2}$$

$$p_j = (Lon_j, Lat_j) \tag{3}$$

where D is the ship's database; V represents one of the ship's AIS data points; i is the number of ships; j refers to the number of AIS data for this ship; and *Name*, *MMSI, Type, Length, Width,* $t_j, c_j, s_j, p_j, Lon_j$, Lat_j are the ship's name, MMSI, type, length, width, present time, course, speed, position, longitude and latitude, respectively. Although the AIS data contain more information, we will only focus on this list.

The MMSI is a unique number assigned to each ship and is not editable from the AIS equipment onboard. If the MMSI is incorrect, it is impossible to ensure that the data are correct. Therefore, data with an incorrect MMSI number (e.g. <9 digits, empty or '888888888'type) should be deleted. Notably, the ship's name is rarely wrong and corresponds to the MMSI number. Similarly, the length and width are not editable onboard. Normal vessels do not exceed a length of 450 m or a width of 100 m, and data with abnormal values should be removed. The course has a range of 360° and cannot exceed it; hence, data outside of this range should be deleted. The speed must be considered because obtaining data from a ship that never moves is pointless. Considering both the current and the wind, the lower and upper

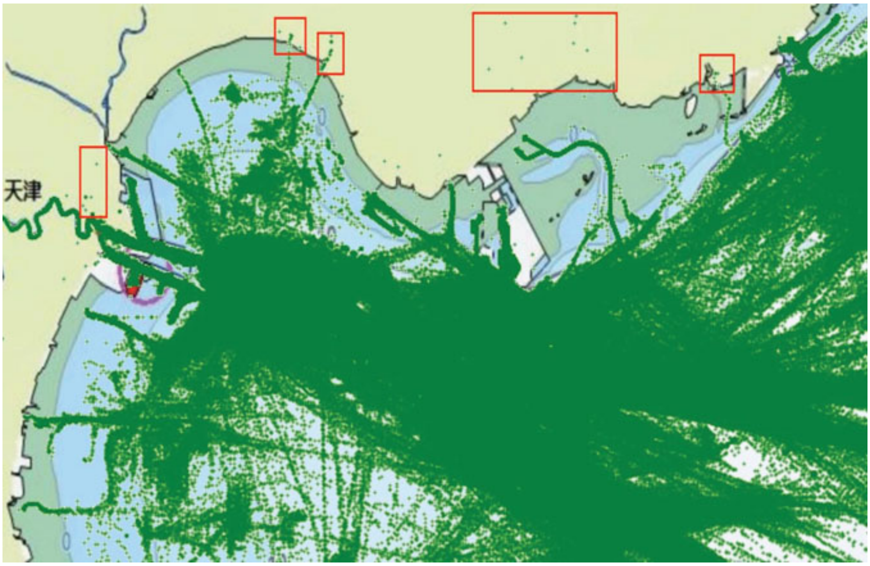

Figure 2. *Original AIS data distribution: the red square marks the abnormal data.*

speed thresholds have been determined to be 3 and 30 kn, respectively. Therefore, if the speed is <3 or >30 kn, it should be deleted, similar to previous research (Liu et al., 2020).

Furthermore, the completeness of the AIS track should be considered (Zhao et al., 2018). There are several isolated AIS data tracks, including one where the number of the AIS data is less than 10, indicating that it is incomplete. These AIS tracks cannot characterise the ship's motion or identify features of interest. Based on the work of Zhao et al. (2018), track points with <80 should be removed.

2.2. Ship domain

The ship domain is defined as 'the waters around the vessel which the navigator would like to keep free of other vessels or objects for safety reasons' (Goodwin, 1975). Since 1975, around 10 typical ship domains have been proposed; they are not entirely distinct. The domain can be distinguished based on its circular, elliptical or polygonal shapes. The circular ship domain is represented by the Goodwin model (Goodwin, 1975) and the Davis model (Davis et al., 1980). The elliptical ship domain is described by the Fujii model (Fujii and Tanaka, 1971), the Coldwell model for head-on situations, the Coldwell model for the overtaking situation (Coldwell, 1983) and the Kijima model (Kijima and Furukawa, 2001). The polygonal domain can be represented by the Smierzchalski model (Smierzchalski, 2001) and the Pietrzykowski model (Pietrzykowski and Uriasz, 2004). Wang (2013) proposed a dynamic quaternion ship domain (DQSD) that combines ship, human and situational submodels. When simulating the Esso Osaka tanker, the DQSD outperformed the previous ship domain in terms of performance and accuracy. Wang and Chin (2016) proposed an empirically calibrated elliptical ship domain. When compared to the existing model, it is a free-form ship domain that can represent the non-typical encounters. Zhang et al. (2016) developed an AIS-based ship collision detection model. The Fujii model was used to improve differentiation. Zhang et al. (2021) presented a dynamically established ship domain for inland waters. This domain divides the surrounding waters of the ship into grids and calculates the grid densities to determine shape and size of the ship domain. It is influenced by the water level and the season (wet or dry).

The circular ship domain proposed by Goodwin is applied in this study because it takes into account the COLREGS. The boundary of this domain is divided into three sectors based on the arcs of the ship's sidelights and stern light, as shown in Figure 3. Typically, the domain parameters are set to $r1 = 0 \cdot 85$ nm,

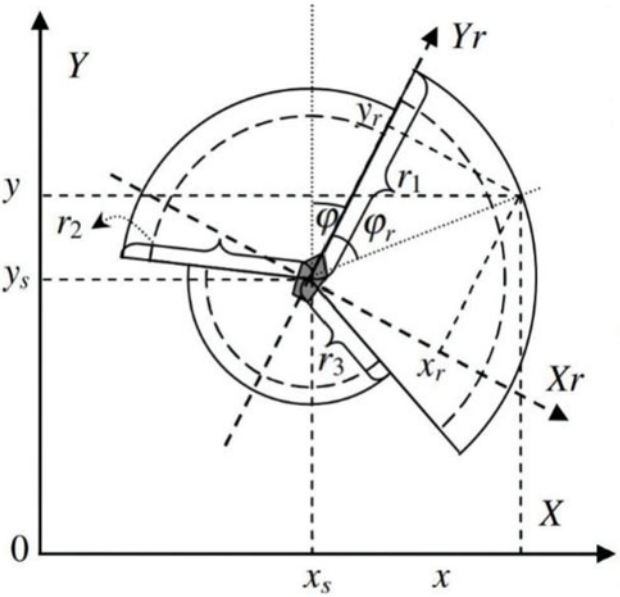

Figure 3. *Goodwin ship domain.*

r2 = 0 · 70 nm and r3 = 0 · 45 nm. Jingsong et al. (1993) proposed a fuzzy ship domain based on fuzzy set theory, which is shown as broken lines in Figure 3. The following are some comments regarding this domain:

1. The domain looks like the arcs of a lighthouse that can display different colours of light simultaneously based on its direction. The arcs have different radii. The arc radius of the starboard is the longest, while the arc radius of the stern is the shortest. The COLREGS are the main reason for this arrangement;
2. The size of this domain is obtained from the research (Wang et al., 2009); it is not influenced by the ship's size;
3. The broken arcs within the domain form a fuzzy ship domain, as proposed by Jingsong et al. (1993). The corresponding parameters are r1 = 0 · 68 nm, r2 = 0 · 56 nm and r3 = 0 · 35 nm.

The Goodwin domain can be described as follows (Wang et al., 2009):

$$\begin{cases} f_{\text{circle}}(x,y) > 0, \text{while } (x,y) \text{ is out of domain} \\ f_{\text{circle}}(x,y) = 0, \text{ while } (x,y) \text{ is on the boundary} \\ f_{\text{circle}}(x,y) < 0, \text{ while } (x,y) \text{ is in the domain} \end{cases} \tag{4}$$

$$f_{\text{circle}}(x,y) = \begin{cases} (x-x_t)^2 + (y-y_t)^2 - r_1^2, \text{ if } 0^\circ \leq \varphi_t \leq 112.5^\circ \\ (x-x_t)^2 + (y-y_t)^2 - r_2^2, \text{ if } 247.5^\circ \leq \varphi_t < 360^\circ \\ (x-x_t)^2 + (y-y_t)^2 - r_3^2, \text{ if } 112.5^\circ < \varphi_t < 247.5^\circ \end{cases} \tag{5}$$

$$\varphi_t = \begin{cases} \arccos \dfrac{y_{tr}}{\sqrt{{x_{tr}}^2 + {y_{tr}}^2}}, & x_{tr} \geq 0 \\ 360^\circ - \arccos \dfrac{y_{tr}}{\sqrt{{x_{tr}}^2 + {y_{tr}}^2}}, & x_{tr} < 0 \end{cases} \tag{6}$$

$$\begin{cases} x_{tr} = (x-x_t)\cos\varphi - (y-y_t)\sin\varphi \\ y_{tr} = (x-x_t)\sin\varphi + (y-y_t)\cos\varphi \end{cases} \tag{7}$$

https://doi.org/10.1017/S0373463323000231 Published online by Cambridge University Press

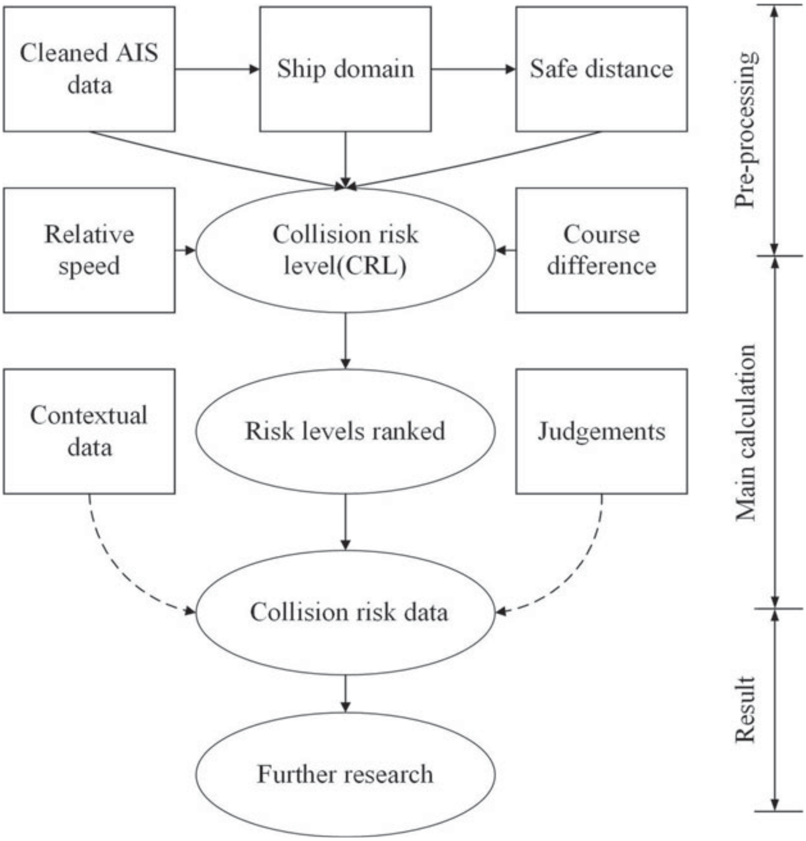

Figure 4. *Framework of the model.*

$$\begin{cases} x_t = x_s + d_t \sin(\varphi + 19^\circ) \\ y_t = y_s + d_t \cos(\varphi + 19^\circ) \end{cases} \tag{8}$$

where r_i ($i = 1, 2, 3$) is the radius of each domain, d_t is 0, φ represents the course of the ship and $(x_{s,}\ y_s)$ are coordinates of the ship in the Earth reference coordinate system.

2.3. *Framework of the proposed model*

Figure 4 shows the framework of the marine collision risk identification strategy. Its primary steps include preprocessing, main calculation and result presentations. The kernel of this framework should be the collision risk level (CRL), which was developed with several variables based on a set of logic (see details in Section 3). The variables comprise AIS data, ship domain, distance to the domain, relative speed and course difference of the two vessels. The first two were covered in Sections 2.1 and 2.2. The distance to the domain was explained in Section 3.1.1 and the relative speed and course difference were explained in Sections 3.1.2 and 3.1.3, respectively. CRLs are calculated and classified in Section 4.

In large or busy sea areas, ship encounters happen all the time; judging every one of them would require a consistent effort. This study outlines a filtering procedure that can reduce the effort required to determine the collision risk severity. Compared with previous work (Zhang et al., 2016), this method provides the concerned parties with more evaluation time to take action.

https://doi.org/10.1017/S0373463323000231 Published online by Cambridge University Press

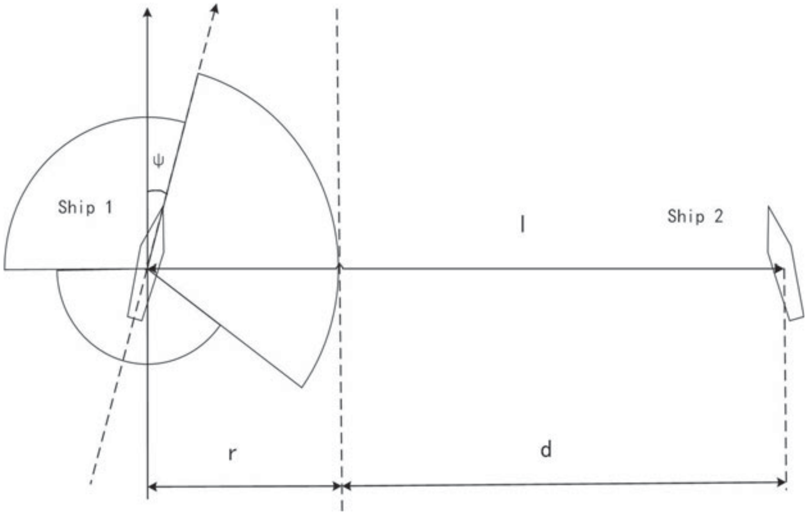

Figure 5. *Distance between the two ships.*

3. The collision risk identification model

3.1. Model contents

Collision risk research (Mou et al., 2010; Goerlandt and Montewka, 2015b) has shown that the DCPA and TCPA are not the only factors that can be used to assess a collision situation. The relative azimuth of the encountered ships is also an important factor. Furthermore, different meeting scenarios require different timescales for actions; for example, a head-on situation is more urgent than a crossing encounter; however, a crossing encounter is more urgent than an overtaking encounter. Considering these facts, we propose the CRL model, which combines expert knowledge and the following relevant factors:

1. The distance *d* between the target ship and the domain boundary of the own ship, as shown in Figure 5
2. The relative speed *v* of the encountered ships
3. The course difference *h* between the encountered ships

These factors are considered the key elements that can identify the collision risk severity during the encounter and can be used as a reference to determine the CRL.

3.1.1. Safe distance d

The safe distance in the model is not the distance between two vessels' centres, but the distance between the domain boundaries of the observing ship and the observed ship. Moreover, a longer safe distance implies a lower collision risk between vessels, but there is no clear evidence of the relationship between the safe distance and the collision risk. However, Zhang et al. (2016) reasonably assumed that the safe distance is inversely proportional to the conflict severity; we adopted this strategy and treated the relationship between the safe distance and the CRL as inversely proportional:

$$CRL \sim f\left(\frac{1}{d}\right) \tag{9}$$

$$d = l - r \tag{10}$$

where *r* is the radius of the safety domain of the observing ship and *l* is the distance between the encounter ships (Figure 5). The distance *l* is calculated from the positions of the two vessels. It differs from the distance between two known points in the Cartesian coordinate system and, therefore, the calculation procedure must be illustrated. The positions of the two vessels are ($L1$, $B1$) and ($L2$, $B2$),

https://doi.org/10.1017/S0373463323000231 Published online by Cambridge University Press

respectively, and the safe distance between them can be calculated as:

$$\begin{cases} \Delta L = L2 - L1 \\ \Delta B = B2 - B1 \end{cases} \tag{11}$$

$$S = a(1 - e^2)\int_0^B (1 - e^2\sin^2 B)^{-1.5}\mathrm{d}B \tag{12}$$

$$l = \begin{cases} \dfrac{S_{B_2} - S_{B_1}}{\cos A_1} & (\Delta B \neq 0) \\ r_1|\Delta L| & (\Delta B = 0) \end{cases} \tag{13}$$

$$r_1 = \frac{a\cos B_1}{\sqrt{1 - e^2\sin^2 B_1}} \tag{14}$$

where ΔL is the difference in longitude, ΔB is the difference in latitude, a is the Earth's ellipsoid long radius, e refers to the eccentricity of the Earth, S is the arc length from the point to the equator, r_1 is the latitude radius at the point ($L1$, $B1$) and A_1 is the ship's orientation at ($L1$, $B1$).

3.1.2. Relative speed v

A higher relative speed implies a faster rate of change in the distance between encountered ships. If the v value is negative, the ships are not in danger because they are moving away from each other. However, if the v value is positive, there is a collision risk. Therefore, relative speed should be considered when assessing the collision risk between ships. Based on the senior expert's experience, the relative speed positively correlates with the distance change, and the distance change has a linear relationship with the collision risk. Combining these two facts, we propose a relationship between the relative speed and the CRL:

$$CRL \sim f(v) \tag{15}$$

3.1.3. Course difference h *between the encounter ships*

The course difference h is a dynamic angle that changes with the movement of the ships. It has a significant impact on the vessel's collision-avoidance capacity and is influenced by the frequency of manoeuvring actions. The range of h can be defined as $[-\pi, \pi]$, where a positive value indicates that the two vessels are moving closer and a negative value indicates that they are moving apart (Figure 6). Zhang et al. (2020) showed that h can be expressed as an odd periodic function with a period of 2π using a Fourier series as:

$$CRL \sim f(h) \tag{16}$$

$$\varphi(h) = \sum_{i=-\infty}^{\infty} a_i \cdot e^{ji(2\pi/T)h} \tag{17}$$

where *φ(h)* is the Fourier series, T is the period and coefficients a_i can be expressed as:

$$a_i = \frac{1}{T}\int_T \varphi(h) \cdot e^{-ji(2\pi/T)h} \tag{18}$$

$$e^{ji} = \cos i + j \cdot \sin i \tag{19}$$

https://doi.org/10.1017/S0373463323000231 Published online by Cambridge University Press

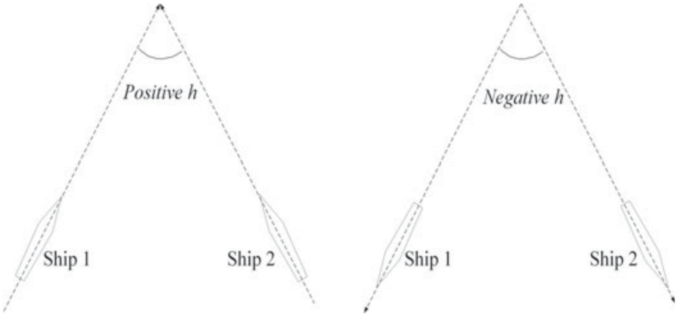

Figure 6. *Course difference h.*

Thus, $f(h)$ can be represented using a sine series as:

$$f(h) = \sum_{-\infty}^{\infty} b_i \cdot \sin(i \cdot h) \tag{20}$$

$$b_i = a_i \cdot j \tag{21}$$

3.2. The collision risk level formulation

The influencing factors of the model were discussed in Section 3.1. Generally, encountered ships with a large d and small v should be at very low risk of collision, and their CRL value should be small, while a higher CRL value represents an increased collision risk severity, and vice versa. This aligns with the findings in Sections 3.1.1 and 3.1.2 that a combination of the two factors (d and v) can better reflect the risk scenario. Therefore, CRL could be formulated as:

$$CRL(d, v, h) = k \cdot \frac{1}{d} \cdot v \cdot \left(\sum_{-\infty}^{\infty} p_i \cdot \sin(i \cdot h) \right) \tag{22}$$

where k is the final coefficient that considers all the factors simultaneously, i is the number of sine series and p_i is the coefficients of the sine series. It can be observed that a greater i value implies more precision for $f(h)$. However, a larger i value indicates a more complex calculation and increased calculation time. A good balance must be maintained between precision management and complex calculation. Previous researchers (Wang et al., 2014; Zhang et al., 2016) set the example by computing with a fixed i in the Fourier series, and the result remained with enough precision. The formulation can then be modified as:

$$CRL(d, v, h) = k \cdot \frac{1}{d} \cdot v \cdot \left(\sum_{i=1}^{n} p_i \cdot \sin(i \cdot h) \right) \tag{23}$$

$$CRL(v, h) = k \cdot v \cdot \left(\sum_{i=1}^{n} p_i \cdot \sin(i \cdot h) \right) \tag{24}$$

$$CRL(d, v) = k \cdot \frac{1}{d} \cdot v \tag{25}$$

When v equals zero, the encountered ships are relatively stationary, implying that there is no collision risk because the CRL value is zero. When d equals zero, the observed vessel is located in the boundary of the ship domain of the observing vessel; Equation (23) is not suitable for expressing the CRL in that case. Therefore, Equation (24) can be used in place of Equation (23). When h equals zero, the encountering ships are moving in the same direction (e.g. overtaking), the collision risk still exists, and

https://doi.org/10.1017/S0373463323000231 Published online by Cambridge University Press

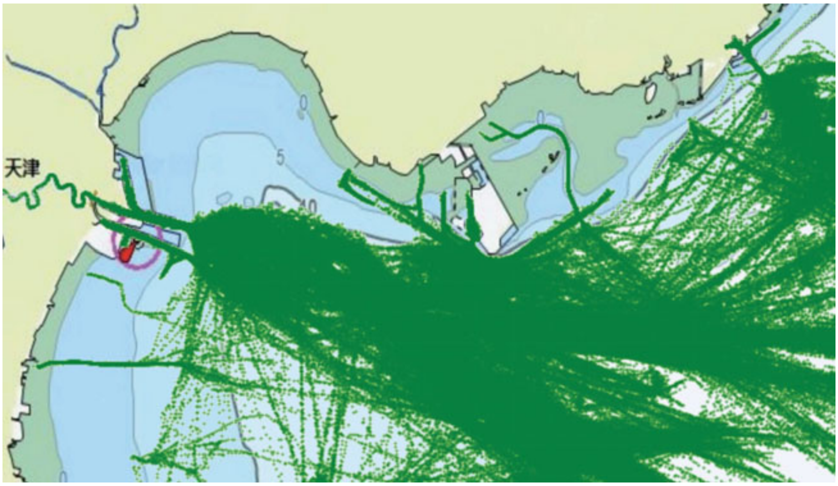

Figure 7. *Cleaned AIS data distribution.*

the formulation must be transformed again. Given that the safe distance and relative speed are the main factors influencing this situation, Equation (25) is a suitable choice.

3.3. Model parameter estimation

Two standard CRL values are predefined to provide reference CRL values. As stated in Section 1.1, the two preset values are referred to Zhang et al. (2016) to better make the comparison. The first value occurs when the distance between encountering ships equals 6 nm; this range will draw the attention of the duty deck officers because any vessel entering this range should be monitored, particularly in coastal waters. The second value occurs when d is equal to 1 nm; the officers must thoroughly evaluate the collision risk situation and be prepared to take evasive action to avoid the collision. The equations are:

$$E(CRL(d,v,h))_{l=6\text{nm}} = 5 \tag{26}$$

$$E(CRL(d,v,h))_{d=1\text{nm}} = 100 \tag{27}$$

$$V = \left(\left(\sum\nolimits_{i=1}^{m} k\frac{1}{d_i}v_i \sum\nolimits_{j=1}^{n} p_j \sin(j \cdot h_i)\right)^2\right)_{\min} \tag{28}$$

where E is the expectation of the formulation, m is the number of sampled encounters and n is the Fourier series number. Both m and n influence the accuracy of the formulation. The AIS data used in this paper will help to determine the m value, and n can be determined from relevant references. Wang et al. (2014) found that sufficient accuracy will be obtained when n is >5; Zhang et al. (2016) used $n = 17$ to ensure sufficient precision. Herein, we also selected 17 as the n value.

To determine the parameters of Equations (23)–(25), Equations (26) and (27) were applied using the AIS data of Tianjin Port. The least square method was used to calculate the parameters [Equation (28)]. It should be noted that the predefined values serve as a reference for identifying the different encounter situations. Only after connecting the predefined values to the collision risk situations, can the degree of collision risk severity level be determined; thereafter, these values become meaningful.

https://doi.org/10.1017/S0373463323000231 Published online by Cambridge University Press

Table 2. Sample AIS data structure.

MMSI	Ship type	L (m)	W (m)	Timestamp (s)	C (°)	Speed (kn)	Lon (°)	Lat (°)
249655000	cargo	289	44	1421092980	112	3 · 2	117 · 77128	38 · 97528
249655000	cargo	289	44	1421093160	105	5 · 8	117 · 77745	38 · 97345
249655000	cargo	289	44	1421093340	105	7 · 3	117 · 78310	38 · 97223
249655000	cargo	289	44	1421093460	104	8 · 1	117 · 78887	38 · 97098
249655000	cargo	289	44	1421093580	104	8 · 9	117 · 79467	38 · 96980
370299000	cargo	199	32	1421107500	167	8 · 2	117 · 77435	39 · 01000
370299000	cargo	199	32	1421107620	166	8 · 9	117 · 77750	39 · 00468
370299000	cargo	199	32	1421107740	163	9 · 2	117 · 78098	38 · 99873
370299000	cargo	199	32	1421107920	161	9 · 3	117 · 78322	38 · 99327

4. Calculation and evaluation

This section analyses some randomly selected encounter cases to validate the proposed model. On the basis of the aforementioned model, a number of results are shown in this section, including the cleaned AIS data, the CRL model parameters, the discriminant evaluation and the concurrent evaluation.

4.1. Calculation

As mentioned, AIS data from Tianjin Port in January 2015 are used for calculation and evaluation. The original messy data have been cleaned, as shown in Figure 7, and the sample AIS data structure is listed in Table 2. The parameters of CRL are presented in Table 3.

When discussing the manoeuvrability of different vessels, ship size should be considered. Generally, a larger vessel (excluding warships) has a worse manoeuvrability than a small vessel. In other words, bigger ships require more time to evaluate and take action to avoid collisions. Therefore, it is necessary to classify the vessels based on their size so that an accurate analysis can be carried out. The cleaned AIS data contains 1354 ships, as shown in Figure 8. Using the K-means clustering method, we classified the ships into three groups: small, medium and large (Table 4).

4.2. Evaluation

4.2.1. Discriminant evaluation

In this section, we randomly analysed five encounter cases from the database to evaluate the CRL model. Similar encounter scenarios were avoided during the process. This measure ensures that the following evaluations have improved discriminant, coherence and result. The encounter situations and corresponding calculation results are shown in Figures 9–13. The highest CRL value is used as a reference because it indicates whether or not there is a potential collision. In the figures, ‘Lond’ represents the longitude, ‘Latd’ represents the latitude and ‘Time’ represents the timestamp. The dashed line is the projection contour of the solid line (not the exact number of positions of the solid line), ‘d’ is the distance to the closest point of the ship domain and ‘T’ means the time to the closest point of the domain.

Figure 9 shows the first case, a crossing encounter between a cargo ship (MMSI:371811000) of 229 m length and another cargo ship (MMSI: 373140000) of 189 m length. After one course change, they were travelling in opposite directions. ‘Zhang et al.’ represents the model proposed by Zhang et al. (2016), ‘Proposed’ represents the model developed in this research, same as in Figures 10–13. The minimum values of d from the two models are 0 · 708 nm and 0 · 864 nm. The values of T are 1 · 678 min and 2 · 059 min, and the CRL values are 45 · 679 and 47 · 605, respectively. Based on the results, the proposed method can identify the potential risk 0 · 381 min earlier than the model developed by Zhang

https://doi.org/10.1017/S0373463323000231 Published online by Cambridge University Press

Table 3. *Parameters of CRL.*

k	p1	p2	p3	p4	p5	p6	p7	p8
52 · 58912	0 · 13550	−0 · 00184	−0 · 02656	0 · 00452	0 · 01945	−0 · 00514	−0 · 00080	0 · 00410
p9	p10	p11	p12	p13	p14	p15	p16	p17
0 · 00602	−0 · 00293	−0 · 00021	0 · 00299	0 · 00030	−0 · 00185	0 · 00427	0 · 00082	−0 · 00429

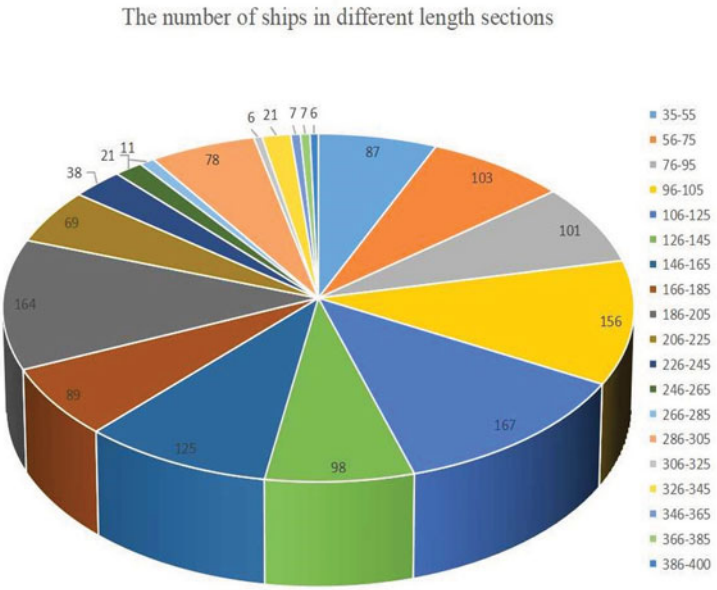

Figure 8. *Number of ships in different length sections.*

Table 4. *Clustering results based on ship length.*

Type	Ship length (m)	Average (m)	Number	Percentage
Small	[35, 137]	93 · 90	453	33 · 46%
Medium	[138,237]	181 · 07	711	52 · 51%
Large	[240, 399]	300 · 02	190	14 · 03%

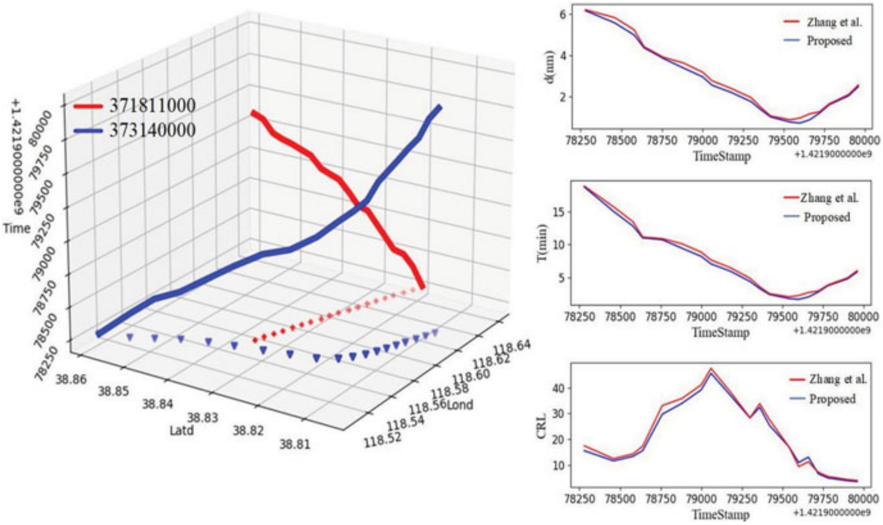

Figure 9. *Encounter between two medium vessels.*

https://doi.org/10.1017/S0373463323000231 Published online by Cambridge University Press

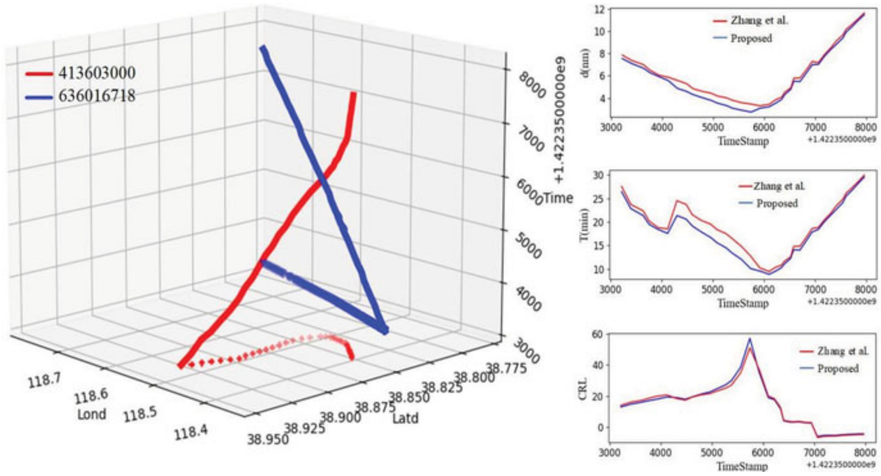

Figure 10. *Encounter between a medium and a small vessel.*

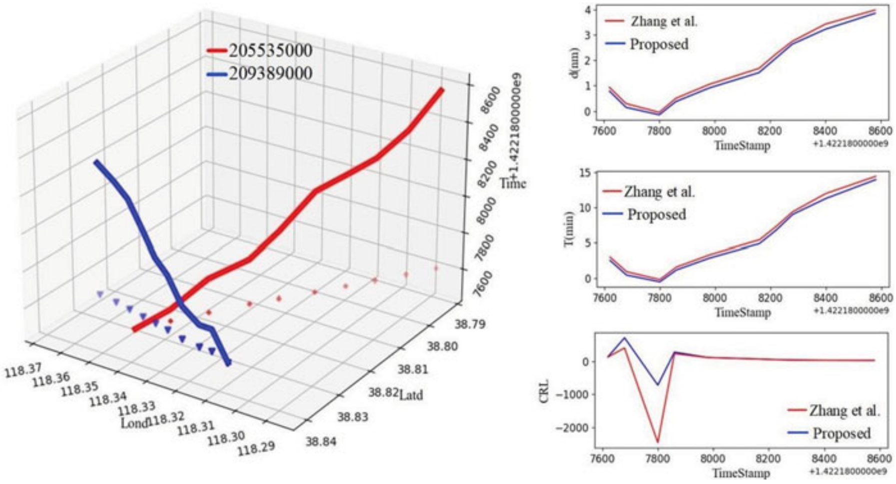

Figure 11. *Encounter between a large and a medium vessel, with high risk.*

et al. (2016). The second case is shown in Figure 10. At first, the two vessels (MMSI: 413603000 and MMSI: 636016718) were in a crossing situation before one of them altered course. Both were cargo vessels, one with a length of 151 m and the other with a length of 108 m. The minimum *d* values obtained from the two models are 2 · 740 nm and 3 · 305 nm, respectively, and the corresponding T values are 8 · 869 min and 9 · 460 min with the CRL values being 57 · 068 and 50 · 934, respectively. Based on the results, the proposed method can identify the potential risk 0 · 591 min earlier than the other one.

Figure 11 shows a crossing situation between the two vessels (MMSI: 205535000 and MMSI: 209389000). Both are cargo vessels, one with a length of 289 m and the other 184 m. The minimum *d* values of the two models are −0 · 144 nm and −0 · 047 nm, indicating that the observing vessel is within the safety domain of the observed vessel. The corresponding T values are −0 · 466 min and −0 · 154 min with the CRL values of 711 · 606 and 402 · 033, respectively. The results indicate that the proposed method can identify the collision risk 0 · 312 min earlier than the other method.

The fourth case is shown in Figure 12. At first, the two vessels are nearly on reciprocal courses, where the small vessel (MMSI: 372229000) is a cargo ship of 80 m length and the large vessel (MMSI: 356984000) is a cargo vessel of 399 m length. Later, both vessels altered their course. The minimum *d*

https://doi.org/10.1017/S0373463323000231 Published online by Cambridge University Press

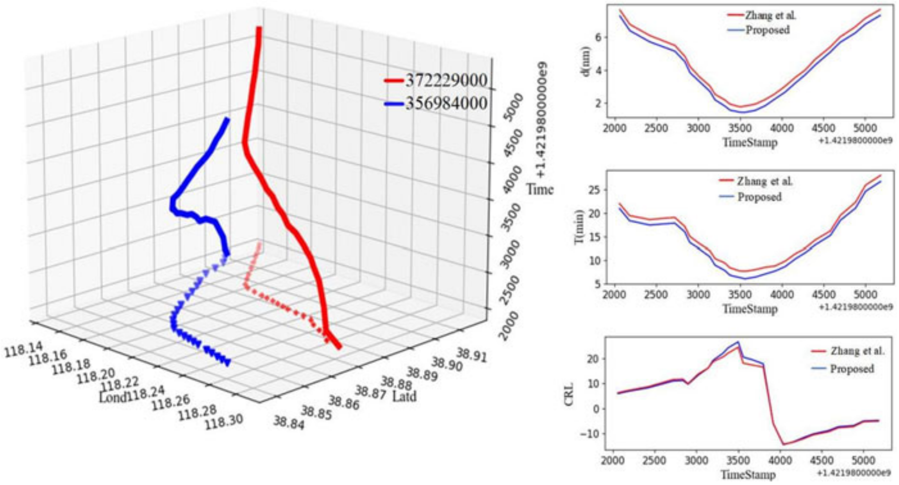

Figure 12. Encounter between a large and a small vessel.

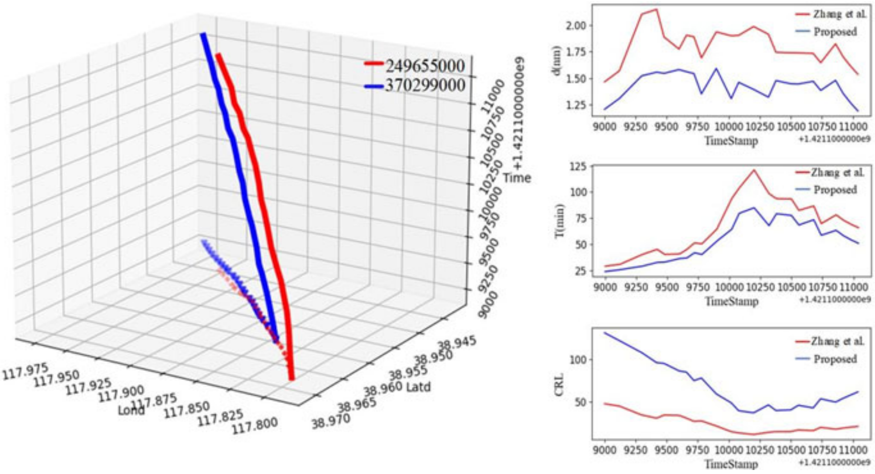

Figure 13. Encounter between a large and a medium vessel.

values of the two methods are 1 · 442 nm and 1 · 773 nm, respectively. The values of T are 6 · 079 min and 7 · 680 min, and the CRL values are 26 · 670 and 24 · 511, respectively. Based on the results, the proposed method can identify the potential collision risk 1 · 601 min earlier compared to the previous model developed by Zhang et al. (2016).

The fifth case is shown in Figure 13. An overtaking situation occurs between the two vessels: a cargo ship of length 289 m (MMSI: 249655000) and another cargo ship of length 199 m (MMSI: 370299000). The minimum *d* values of the two models are 1 · 189 nm and 1 · 464 nm; the relevant T values are 23 · 955 min and 29 · 103 min and the CRL values are 131 · 719 and 47 · 870, respectively. Based on the results, the proposed method can identify the collision risk 5 · 148 min earlier than the method of Zhang et al. (2016).

From the five encounter scenarios, we determined the following:

1. The proposed CRL model can identify the collision risk severity of encountering ships, allowing for a better understanding of the situation and more time to take action. It has been proven to be reasonable.

https://doi.org/10.1017/S0373463323000231 Published online by Cambridge University Press

Table 5. *Distribution of CRL and encounter numbers.*

Collision risk level	No risk	Low risk	Medium risk	High risk	Collision risk	Total
Encounter number	24152	36528	9375	1264	89	71408
Percentage	33 · 82%	51 · 16%	13 · 13%	1 · 77%	0 · 12%	100%

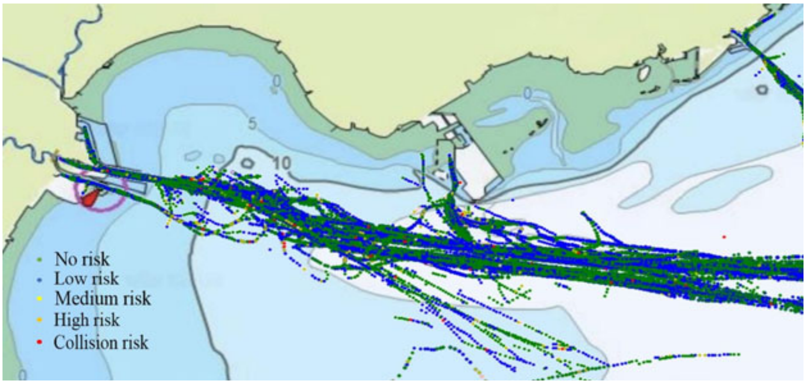

Figure 14. *Distribution of different CRL values.*

2. The CRL model can identify the potential risks earlier than the method developed by Zhang et al. (2016), enabling more time for subsequent analysis and response. This is crucial for maritime safety because more time means better chance to survive.
3. The CRL value positively correlates with the potential collision risk. A higher CRL value indicates a higher probability of collision. This can provide another reference for a vessel's safety, in addition to the CPA and TCPA.

4.2.2. Concurrent evaluation

To execute the concurrent evaluation, all the cleaned AIS data from Tianjin Port in January 2015 were used to calculate the CRL values. The calculated CRL values were clustered into several groups and plotted on a nautical chart, each group representing a level of collision risk severity. Collision risk severity levels can vary in different studies. Baldauf et al. (2011) proposed four levels: safe, caution, warning and alarm. Zhang et al. (2016) proposed five levels: regular encounter low, regular encounter high, collision avoidance, close encounter and risk encounter. Zhang et al. (2020) proposed seven levels: impossible, improbable, uncertain, fifty–fifty, expected, probable and certain. Zhao and Fu (2021) proposed the margin of projected collision index, a binary judgement of the collision situation consisting of margin of projected collision in angle, margin of projected collision in speed and margin of projected collision in time. After checking the related research, no criteria concerning the collision risk severity levels were found. Herein, to make better comparison, we chose five levels: no risk, low risk, medium risk, high risk and collision risk. The K-means clustering method was used to divide all the CRL values into five groups corresponding to the five levels. The results are presented in Table 5.

As seen from Table 5, our results are generally in line with the hierarchical pyramid, which has been proved by Majumdar et al. (2021) and Debnath and Chin (2010). Figure 14 shows the spatial chart of the clustered CRL values, with different colour representing different collision risk severity level.

The lowest level, which corresponds to no risk, happens in most areas within the figure boundary, followed by the low risk level. These two levels are regarded as safe and account for the majority of all encounters. This phenomenon corresponds to the real conditions in navigable waters.

The CRL values of medium risk have 13 · 13%, which mostly occurred in fairways to the harbour and the main traffic lanes. This situation is reasonable because the fairways and the main traffic lanes have major traffic flow and limited navigable space. The amount of high risk and collision risk encounters have 1 · 77% and 0 · 12%, respectively. Most of these values occurred in the outer waterway of Tianjin Port, the southern waters of the Cao Fei-Dian Port and their crossing area, and the fairway off Jing Tang Port. Some high-risk spots also appear in the harbour waters, which is understandable, as vessels pass each other at close distances, and the ship domain probably impacts this situation. Almost no encounters are detected in other sea areas because they are far from the main traffic lanes and ther are no harbours nearby. This is consistent well with the previous research of Majumdar et al. (2021). Many CRL values of high risk and collision risk were located close to the harbour; this is reasonable because ships have to pass each other at a close distance in this area. Therefore, it is necessary to note that the CRL model is not intended for use in harbour areas but for coastal water and open sea (Hansen et al., 2013).

5. Conclusions

The framework proposed in this paper to identify the ship collision risk is an improvement over previous research Zhang et al. (2016). This model identified the collision risk of encountered ships based on their influencing factors, which include ship domain, safe distance, relative speed and course difference. Adopting the Goodwin domain in this framework has prolonged the reaction time to deal with potential risks. Furthermore, the domain has considered the COLREGs, which is critically important for maritime safety. The Fujii domain used in the previous study (Zhang et al., 2016) was a symmetrical elliptical structure in the fore-and-aft direction. This arrangement puts the head-on and overtaking situations in equal positions, which is unsuitable for real navigation. The Goodwin domain has a different radius of boundary based on the scope of the vessel's light arcs. This can differentiate the risk level of vessels approaching from different directions. This is beneficial and convenient for VTS and other parties in monitoring waters and implementing effective collision prevention measures.

Although the judgement from the collision risk severity model is not perfect, it is reasonable and can be used as a reference for detecting potential collisions or monitoring traffic information. The results given by the system are based on the hypothesis that all vessels will continue with their present status. This hypothesis, however, is not perfect. Some additional contextual factors, for example good seamanship, machine malfunction and environmental disorder, also have an impact on the situation and should be the focus of future research.

Funding. This research was funded by the National Natural Science Foundation of China, grant number (52101399).

Competing interests. The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

Altan, Y. C. and Otay, E. N. (2018). Spatial mapping of encounter probability in congested waterways using AIS. *Ocean Engineering*, **164**, 263–271.

Baldauf, M., Benedict, K., Fischer, S., Motz, F. and Schröder-Hinrichs, J. U. (2011). Collision avoidance systems in air and maritime traffic. *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, **225**, 333–343.

Bye, R. J. and Aalberg, A. L. (2018). Maritime navigation accidents and risk indicators: An exploratory statistical analysis using AIS data and accident reports. *Reliability Engineering & System Safety*, **176**, 174–186.

Bye, R. J. and Almklov, P. G. (2019). Normalization of maritime accident data using AIS. *Marine Policy*, **109**, 103675.

Chai, T., Weng, J. and De-qi, X. (2017). Development of a quantitative risk assessment model for ship collisions in fairways. *Safety Science*, **91**, 71–83.

Chen, P., Mou, J. and Van Gelder, P. H. A. J. M. (2019). Integration of individual encounter information into causation probability modelling of ship collision accidents. *Safety Science*, **120**, 636–651.

Chin, H. C. and Quek, S. T. (1997). Measurement of traffic conflicts. *Safety Science*, **3**, 169–185.

Coldwell, T. G. (1983). Marine traffic behaviour in restricted waters. *The Journal of Navigation*, **36**, 430–444.

Čorić, M., Mandžuka, S., Gudelj, A. and Lušić, Z. (2021). Quantitative ship collision frequency estimation models: A review. *Journal of Marine Science and Engineering*, **9**, 533.

Davis, P. V., Dove, M. J. and Stockel, C. T. (1980). A computer simulation of marine traffic using domains and arenas. *The Journal of Navigation*, **33**, 215–222.

Debnath, A. K. and Chin, H. C. (2010). Navigational traffic conflict technique: A proactive approach to quantitative measurement of collision risks in port waters. *The Journal of Navigation*, **63**, 137–152.

Du, L., Banda, O. A. V., Huang, Y., Goerlandt, F., Kujala, P. and Zhang, W. (2021). An empirical ship domain based on evasive manoeuvre and perceived collision risk. *Reliability Engineering & System Safety*, **213**, 107752.

Fujii, Y. and Tanaka, K. (1971). Traffic capacity. *The Journal of Navigation*, **24**, 543–552.

Gao, M. and Shi, G. Y. (2020). Ship collision avoidance anthropomorphic decision-making for structured learning based on AIS with Seq-CGAN. *Ocean Engineering*, **217**, 107922.

Gao, D. W., Zhu, Y. S., Zhang, J. F., He, Y. K., Yan, K. and Yan, B. R. (2021). A novel MP-LSTM method for ship trajectory prediction based on AIS data. *Ocean Engineering*, **228**, 108956.

Goerlandt, F. and Montewka, J. (2015a). A framework for risk analysis of maritime transportation systems: A case study for Oil spill from tankers in a ship–ship collision. *Safety Science*, **76**, 42–66.

Goerlandt, F. and Montewka, J. (2015b). Maritime transportation risk analysis: Review and analysis in light of some foundational issues. *Reliability Engineering & System Safety*, **138**, 115–134.

Goodwin, E. M. (1975). A statistical study of ship domains. *The Journal of Navigation*, **28**, 328–344.

Hansen, M. G., Jensen, T. K., Lehn-Schiøler, T., Melchild, K., Rasmussen, F. M. and Ennemark, F. (2013). Empirical ship domain based on AIS data. *The Journal of Navigation*, **66**, 931–940.

Hörteborn, A. and Ringsberg, J. W. (2021). A method for risk analysis of ship collisions with stationary infrastructure using AIS data and a ship manoeuvring simulator. *Ocean Engineering*, **235**, 109396.

Jingsong, Z., Zhaolin, W. and Fengchen, W. (1993). Comments on ship domains. *The Journal of Navigation*, **46**, 422–436.

Jon, M. H., Kim, Y. P. and Choe, U. (2021). Determination of a safety criterion via risk assessment of marine accidents based on a Markov model with five states and MCMC simulation and on three risk factors. *Ocean Engineering*, **236**, 109000.

Kijima, K. and Furukawa, Y. (2001). Design of automatic collision avoidance system using fuzzy inference. *IFAC Proceedings Volumes*, **34**, 65–70.

Li, B. and Pang, F. W. (2013). An approach of vessel collision risk assessment based on the D–S evidence theory. *Ocean Engineering*, **74**, 16–21.

Lim, G. J., Cho, J., Bora, S., Biobaku, T. and Parsaei, H. (2018). Models and computational algorithms for maritime risk analysis: A review. *Annals of Operations Research*, **271**, 765–786.

Liu, J., Shi, G. and Zhu, K. (2020). Online multiple outputs least-squares support vector regression model of ship trajectory prediction based on automatic information system data and selection mechanism. *IEEE Access*, **8**, 154727–154745.

Liu, K., Yuan, Z., Xin, X., Zhang, J. and Wang, W. (2021). Conflict detection method based on dynamic ship domain model for visualization of collision risk hot-spots. *Ocean Engineering*, **242**, 110143.

Luo, M. and Shin, S. H. (2019). Half-century research developments in maritime accidents: Future directions. *Accident Analysis & Prevention*, **123**, 448–460.

Majumdar, A., Manole, I. and Nalty, R. (2021). Analysis of port accidents and calibration of heinrich's pyramid. *Transportation Research Record*, **2676**, 476–489.

Mou, J. M., Van Der Tak, C. and Ligteringen, H. (2010). Study on collision avoidance in busy waterways by using AIS data. *Ocean Engineering*, **37**, 483–490.

Murray, B. and Perera, L. P. (2021). An AIS-based deep learning framework for regional ship behavior prediction. *Reliability Engineering & System Safety*, **215**, 107819.

Pietrzykowski, Z. and Uriasz, J. (2004). The Ship Domain in A Deep-Sea Area. *Proceeding of the 3rd International Conference on Computer and IT Applications in the Maritime Industries*, Siguenza, Spain.

Simsir, U., Amasyalı, M. F., Bal, M., Çelebi, U. B. and Ertugrul, S. (2014). Decision support system for collision avoidance of vessels. *Applied Soft Computing*, **25**, 369–378.

Smierzchalski, R. (2001). On-Line trajectory planning in collision situations at sea by evolutionary computation-experiments. *IFAC Proceedings Volumes*, **34**, 407–412.

Szlapczynski, R. and Szlapczynska, J. (2016). An analysis of domain-based ship collision risk parameters. *Ocean Engineering*, **126**, 47–56.

Wang, N. (2013). A novel analytical framework for dynamic quaternion ship domains. *Journal of Navigation*, **66**, 265–281.

Wang, Y. and Chin, H. C. (2016). An empirically-calibrated ship domain as a safety criterion for navigation in confined waters. *Journal of Navigation*, **69**, 257–276.

Wang, N., Meng, X., Xu, Q. and Wang, Z. (2009). A unified analytical framework for ship domains. *Journal of Navigation*, **62**, 643–655.

Wang, J., Deng, W. and Guo, Y. (2014). New Bayesian combination method for short-term traffic flow forecasting. *Transportation Research Part C: Emerging Technologies*, **43**, 79–94.

Wei, Z., Xie, X. and Zhang, X. (2020). AIS trajectory simplification algorithm considering ship behaviours. *Ocean Engineering*, **216**, 108086.

Zhang, L. and Meng, Q. (2019). Probabilistic ship domain with applications to ship collision risk assessment. *Ocean Engineering*, **186**, 106130.

Zhang, W., Goerlandt, F., Montewka, J. and Kujala, P. (2015). A method for detecting possible near miss ship collisions from AIS data. *Ocean Engineering*, **107**, 60–69.

Zhang, W., Goerlandt, F., Kujala, P. and Wang, Y. (2016). An advanced method for detecting possible near miss ship collisions from AIS data. *Ocean Engineering*, **124**, 141–156.

https://doi.org/10.1017/S0373463323000231 Published online by Cambridge University Press

Zhang, J., Teixeira, ÂP, Soares, C. G. and Yan, X. (2018). Quantitative assessment of collision risk influence factors in the Tianjin port. *Safety Science*, **110**, 363–371.

Zhang, W., Feng, X., Goerlandt, F. and Liu, Q. (2020). Towards a convolutional neural network model for classifying regional ship collision risk levels for waterway risk analysis. *Reliability Engineering & System Safety*, **204**, 107127.

Zhang, F., Peng, X., Huang, L., Zhu, M., Wen, Y. and Zheng, H. (2021). A spatiotemporal statistical method of ship domain in the inland waters driven by trajectory data. *Journal of Marine Science and Engineering*, **9**, 410.

Zhao, L. and Fu, X. (2021). A novel Index for real-time ship collision risk assessment based on velocity obstacle considering dimension data from AIS. *Ocean Engineering*, **240**, 109913.

Zhao, L., Shi, G. and Yang, J. (2018). Ship trajectories Pre-processing based on AIS data. *Journal of Navigation*, **71**, 1210–1230.

Zhao, X., Yuan, H. and Yu, Q. (2021). Autonomous vessels in the Yangtze river: A study on the maritime accidents using data-driven Bayesian networks. *Sustainability*, **13**, 9985.

Cite this article: Ma X, Shi G, Shi J, Liu J (2023). A framework of marine collision risk identification strategy using AIS data. *The Journal of Navigation* **76**: 4–5, 525–544. https://doi.org/10.1017/S0373463323000231

https://doi.org/10.1017/S0373463323000231 Published online by Cambridge University Press