Xin, X, Liu, K, Loughney, S, Wang, J and Yang, Z

Maritime traffic clustering to capture high-risk multi-ship encounters in complex waters

Article

Xin, X, Liu, K, Loughney, S, Wang, J and Yang, Z (2022) Maritime traffic clustering to capture high-risk multi-ship encounters in complex waters. Reliability Engineering and System Safety, 230. ISSN 0951-8320

Maritime traffic clustering to capture high-risk multi-ship encounters in complex waters

Xuri Xin[a,b], Kezhong Liu[a,c,d*], Sean Loughney[b], Jin Wang[b], Zaili Yang[b]

[a] *School of Navigation, Wuhan University of Technology, Wuhan, Hubei, China*

[b] *Liverpool Logistics, Offshore and Marine (LOOM) Research Institute, Liverpool John Moores University, Liverpool, UK*

[c] *Hubei Key Laboratory of Inland Shipping Technology, Wuhan, Hubei, China*

[d] *National Engineering Research Center for Water Transport Safety, Wuhan University of Technology, Wuhan, Hubei, China*

* Corresponding author.
E-mail address: kzliu@whut.edu.cn (K. Liu).

Abstract:

Maritime traffic situational awareness is fundamental to the safety of maritime transportation. The state-of-the-art research primarily attaches importance to collision risk estimation and evaluation between/among ships but encounters the challenges of capturing the high-risk traffic clusters in complex waters. This paper develops a systematic traffic clustering approach to enhance traffic pattern interpretability and proactively discover high-risk multi-ship encounter scenarios, in which both the conflict connectivity and spatial compactness of encounter ships are considered. Specifically, a novel hybrid clustering approach that integrates a composite distance measure, a constrained Shared Nearest Neighbour clustering, and a fine-tuning strategy is developed to segment maritime traffic into multiple conflict-connected and spatially compact clusters. Meanwhile, a hierarchical bi-objective optimization algorithm is introduced to search for optimal clustering solutions.

Through maritime traffic data obtained from the Ningbo-Zhoushan Port, a thorough methodology performance evaluation is carried out through application demonstration and validation. Experiment results reveal that the new approach: 1) can effectively capture the high-risk/density traffic clusters; 2) is robust with respect to various traffic scenarios; and 3) can be extended to assist in collision risk management. It therefore offers new insights into enhancing maritime traffic surveillance capabilities and eases the design of risk management strategy.

Keywords: Maritime safety, intelligent maritime monitoring, traffic cluster identification, clustering technique, AIS data

1 Introduction

Maritime safety management has always been regarded as one of the essential concerns due to the intolerable ramifications (e.g., loss of life, economic damage, and/or environmental pollution) when maritime traffic accidents occur. Economic globalization associated with the rapid boom in transportation demand has made maritime traffic more sophisticated, especially in restricted waterways and heavy-traffic ports [1–3]. This change brings significant challenges to maritime operational authorities on maritime traffic safety management, particularly when the fast development of emerging autonomous ships is considered, which could potentially increase the occurrence likelihood of ship collisions without an effective solution to be found. In response to such challenges, advanced equipment and systems, such as maritime traffic service networks, geographical information systems, and vessel traffic services (VTS), have been employed to assist operators in maritime traffic monitoring and surveillance [4–6]. Although presenting a great variety of intelligent functionalities in monitoring and regulating maritime traffic behaviours, the currently used systems still reveal some drawbacks in rationally interpreting maritime traffic pattern complexity and adaptively capturing real-time high-risk traffic clusters. Accordingly, maritime traffic controllers often need to face difficulties of capturing the high-risk ship areas by their intuition and experience, significantly increasing their workload and hindering the timely implementation of anti-collision risk control strategies. Thus, it is essential to develop advanced techniques and tools to further enhance Maritime Situational Awareness (MSA) to aid the development of intelligent transportation service systems.

To improve the operational safety of ship traffic in complex waters, various approaches have been developed and implemented to quantitatively analyse ship collision risk [7–11]. These methods are a prerequisite for maritime traffic safety monitoring and management, allowing the operators to raise early collision alarms and ensure ship anti-collision safety. In particular, the considerable development of Automatic Identification System (AIS) has significantly led to abundant AIS-based trajectory data for enhancing vessel collision risk estimation and assessment [12]. In the meantime, detecting clusters of encounter vessels based on real-time AIS-based trajectory information has become an emerging research topic [13–15]. It plays a significant role in improving maritime surveillance capabilities and identifying potentially multiple ship encounters. However, the existing studies suffer from some drawbacks, such as ignoring or simplifying ship dynamics, only concerning traffic density, and having difficulty in discovering the traffic clusters with varying densities. Similarly, the ever-growing ship spatiotemporal movement uncertainty and maritime traffic complexity further influence the state-of-the-art approaches' effectiveness and applicability, especially in complex traffic scenarios involving changeable traffic behaviour. To identify the encountering traffic clusters, it is of paramount importance to fully consider the spatiotemporal dynamics of ship movements and the multiple dependent conflict-related interrelationships (e.g., spatiotemporal proximity and conflict severity) of encounter ships. Therefore, these research gaps must be filled to ensure ship anti-collision safety at sea.

This paper aims to develop a systematic traffic clustering methodology to adaptively discover the high-risk/density multi-ship encounters. It can be used to enhance intelligent situational awareness by decomposing the whole traffic complexity within a surveillance area, facilitating anti-collision decision-making to control multi-ship collision risks. To achieve this, an improved novel density-based clustering approach that incorporates the multi-attribute interactions (i.e., conflict relation and spatial distance) among ships is designed to partition the regional ship traffic into lots of conflict-connected and spatially compact clusters. In the meantime, a hierarchical bi-objective optimization algorithm is formulated as an integral part of the methodology framework to generate robust clustering solutions resilient to the variability of ship traffic situations. The main originality and contributions of this paper are summarized as follows.

1) The proposed traffic clustering methodology captures and recognizes the high-risk/density traffic clusters effectively and reliably by considering both the conflict connectivity and density compactness of encounter vessels. It also accounts for the influence of the ships' stochastic and uncertain behaviour on collision detection, facilitating the operators to better understand and reveal the actual traffic conflict patterns.
2) A novel density clustering approach is developed by synergizing a composite distance measure, a constrained Shared Nearest Neighbour (SSN) clustering, and a fine-tuning strategy. Compared with the traditional traffic clustering approaches, the proposed clustering approach can handle traffic scenarios with varying densities and find traffic clusters with strong conflict connectivity at high density, making it feasible and applicable in complex maritime traffic waters.
3) A hierarchical bi-objective optimization algorithm is designed to find the optimal clustering solutions based on a grid-search strategy. This algorithm is proven effective and robust with respect to various traffic scenarios by explicitly concerning the hierarchical priorities among different objectives.

The rest of the paper is organized as follows. Section 2 summarizes the state-of-the-art research on ship collision risk and AIS data applications in maritime surveillance. In Section 3, the proposed traffic clustering approach is introduced and described. Section 4 describes the proposed hierarchical bi-objective optimization procedure for super-parameter determination. The parameter sensitivity analysis, application demonstration, model comparison and validation, and discussion are provided in Section 5. Conclusions and potential future directions are summarized in Section 6.

2 Literature review

2.1 Ship collision risk evaluation and estimation

Ship collision risk has long been among the most critical concerns in marine transportation research. Numerous publications have focused on evaluating and quantifying the occurrence likelihood and consequence severity of ship collisions. For example, a complete and recent survey can be found in [16–18]. Among such literature, collision risk detection and forecasting are one of the most widely studied subjects. An abundance of non-accidental definitions, such as traffic conflict [19] and near-miss [20], have been proposed to characterize and estimate the collision risk. They constitute an integral part of maritime traffic safety management and serve as a prerequisite for potential real-time collision risk detection. Two main types of relevant research favoured by many scholars are ship domain-based and synthetic index methods.

The ship domain is a unique property in the maritime traffic domain that refers to a safe space around the vessel that the drivers wish to keep free from other vessels. It is employed to detect and estimate potentially dangerous encounter events based on the overlaps or violations of the approaching ships' domain regions. More recently, various ship domain models have been designed to identify the candidates with collision potential and undesired consequences. The most important concerns of these models are rationally screening the influential factors, specifying the domain shapes, and ascertaining the employed methodological techniques [21–23]. They have been applied to a variety of issues and demonstrate their strengths in quantitatively examining waterway capacity [24], discovering collision risk hot-spots [23], and exploring interrelations between collision candidates and historical accident databases [25]. Despite this, the application performance of ship domain models for

Conflict Detection (CD) heavily relies on trajectory prediction techniques. Combining these models with refined trajectory estimation approaches therefore offers an appealing way to enhance the practical CD capabilities.

Synthetic index methods formulate mathematical or black-box models to synthesize the indices that reveal the spatiotemporal proximity level between encountering ships. The most favoured approaches are deploying the two commonly accepted indices, Distance to Closest Point of Approach (DCPA) and Time to Closest Point of Approach (TCPA), to calibrate the collision risk. On this basis, further improvements of such approaches are implemented by including more essential factors [26–29], employing reliable multi-index integration methods [30,31], and assuring their evaluation performance in disparate ship encounter scenarios [32]. These models have reliable and practical performance in assisting in noticing potential collisions and issuing an earlier alert in open sea. However, they hold a critical assumption that the ships will keep an unchanged speed when encountering others, which limits their practical applicability in complex waters. For example, this type of research may lack the desired accuracy when the encounter ships take some manoeuvres (e.g., turning) because of the constrains of waterway geometry. This suggests that further investigation on the ship traffic's spatiotemporal dynamic patterns is essential for accurate and reliable collision risk estimation.

Generally, vessel collision risk estimation has always been an active direction of research while at the same time the increasingly complicated traffic conditions have required researchers to develop new advanced technologies. This can be proved from the following aspects. First, there has been little collision risk estimation research which accounts for the dynamics and uncertainty involved in ship motion. Most studies are highly dependent on the assumptions that the encounter ships would sail linearly, or their future trajectories are able to be determined in advance, overlooking the influence of various uncertainty sources. These assumptions are the detriments of discovering the actual traffic conflict pattern, resulting in their inapplicability in complex waters. Many studies in the aviation research domain suggest that incorporating traffic dynamics and uncertain behaviours is essential to detect and resolve conflicts [33,34]. They estimate the collision risk by developing probabilistic CD models to incorporate the impact of trajectory uncertainty. Second, the studies that can automatically capture the real-time high-risk multi-ship clusters are limited in the reported literature, but they are crucial for decomposing the global collision risk in a given high-traffic water area and relieving surveillance operators' management pressure. In heavy-traffic waters, the collision avoidance measures performed by one ship to resolve its conflict with another may bring about a higher risk with other nearby ships. Therefore, more attention should be given to the difficulties caused by the spatiotemporal interactions of multiple ships for conflict resolution rather than the collision risk between ship pairs. Correspondingly, many researchers have focused on collision risk avoidance in multi-ship encounters [35,36]. To summarize, improving the applicability and practical usability of collision risk estimation models and proactively promoting maritime traffic monitoring through the development of traffic cluster recognition technologies are beneficial and promising.

2.2 AIS data applications in maritime transportation surveillance

Owing to its high sampling frequency, wide coverage, and accessibility of rich information, the applications of AIS data have attracted growing attention from academic circles and bring great potential to maritime traffic behaviour analysis and ship collision risk characterization. Within this context, advances in computer science and artificial intelligence technologies have resulted in its broader and more practical applications in various directions, such as maritime anomaly detection [37,38], spatiotemporal ship traffic correlation analysis [39,40], ship behaviour modelling and recognition [41–44], and vessel path planning [45]. A detailed literature review of AIS data applications has been documented in [46–48]. These studies in the literature fall into two main groups: maritime traffic pattern mining and maritime traffic prediction.

Maritime traffic pattern mining relies on various data mining techniques to undertake maritime traffic analytics, traffic pattern exploration, and knowledge extraction. Classical solutions to traffic pattern mining involve vector-based, grid-based, and statistics-based approaches [40,48,49]. The vector-based methods extract the network waypoints (i.e., nodes) and routes (i.e., edges) to formulate the maritime traffic network, allowing the vessel motions and traffic patterns over busy waters of interest to be characterized as a high compactness graph-based representation [48]. Typically, the pre-processing of vessel trajectories is a prerequisite for traffic network construction through clustering algorithms. Theoretical maritime traffic network modelling involves two important components. One component is to adopt clustering techniques such as Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [50] and Ordering Points To Identify the Clustering Structure (OPTICS) [49,51] to extract the waypoints, including static points (e.g., port and anchorage area), and entry and exit points. The other component is to use a maritime route learning method to detect the vessel trajectories following identical itineraries. Leveraging the established maritime geographical networks, they contribute to supporting maritime traffic surveillance [52], assisting in anomaly detection [49], facilitating route planning [53], and helping to understand maritime traffic patterns [50]. However, maritime traffic following regular behaviour patterns is the basic premise for the applications of these methods, and it is highly problematic. They reveal the weaknesses in terms of configuring the water areas where the traffic patterns are hard to categorize [54]. Further detailed modelling that incorporates more traffic features, such as course and speed distribution, should be developed to effectively differentiate the traffic features and improve the accuracy of geographical networks.

The grid-based methods discretize the target maritime traffic area into indexed grids. Each grid is attached with essential property statistics (e.g., traffic density, course, and speed) to characterise the maritime traffic scenarios. The intention is to construct the gridded database to reduce the data scale and facilitate efficient retrieval and search operations of maritime knowledge. Based on the gridded database, various maritime traffic layers such as traffic route information and traffic distribution information can be established to identify the traffic spatial-temporal patterns [55], differentiate the anomalous behaviours [56], investigate the traffic motion mechanism for maritime situation prediction [57], and discover correlations between the local traffic pattern and near collision hotspots [40]. For example, Xiao et al. [58] populated the AIS data into the structured grids to support the application of a clustering algorithm to extract the waterway and waypoint patterns. However, these methods are only suitable for small-scale waters and are unable to tackle the intense computational load required to support the analysis in large-scale water areas [59]. Additionally, the prior determination of the grid size is a problematic issue that highly depends on the local traffic features.

The statistics-based methods analyse the traffic characteristics and conduct quantitative modelling to reveal the distribution profile of traffic properties. Examples include the identification of distribution characteristics of ship traffic [60,61], the capturing of hot-spot water areas [62], the investigation of temporal variations of density maps [63], the correlation examination between different traffic attributes [64], and the visualization of maritime traffic situations [65]. These studies set the foundations for enhancing maritime traffic situation interpretation, determining important traffic parameter thresholds, and facilitating anti-collision decision-making. Notably, they need to work with other advanced technologies to support high-level MSA. For example, Rong et al. [38] developed an uncertain ship motion prediction approach for trajectory anomaly detection by combining the ship acceleration distributions with a data-driven non-parametric Bayesian model. Overall, the above methods provide essential knowledge for maritime surveillance and management. Unfortunately, these studies are highly dependent on the batch analysis of historical maritime datasets (e.g., months or years), thereby providing fewer guidelines for maritime management authorities and operators to comprehend traffic situations in real time.

Maritime traffic forecasting employs reasonable input and output to construct mathematical functions or models for prediction applications. It is among the hottest research topics because it is one of the indispensable components for proactive traffic surveillance and management. The relevant traffic prediction research can be divided into physics-based, manoeuvre-based, and interaction-aware methods in terms of application methodologies [18]. Many types of techniques and algorithms, such as Constant Velocity Model (CVM) [66], neural network [67], support vector machine [68], random forest [69], Kalman Filter [70], and Particle Filter [71], have been applied for the trajectory prediction. These works demonstrate their merits in terms of prediction accuracy, efficiency, and practical usage, enhancing MSA and safety management capabilities to a large extent. Apart from trajectory prediction, some research studies [49,58] have attached importance to estimating traffic hot-spots (e.g., traffic speed and density) to assist in collision alerts and route planning. In the context of maritime intelligence surveillance, traffic prediction, collision detection, and conflict resolution constitute the base of the operational authorities' task, while traffic prediction is the first basic module and is fundamental to providing precise collision estimations and supporting practical collision evasion actions. Therefore, traffic prediction provides a strong foundation for further studies that facilitate the perception of the forthcoming traffic situations.

While pioneering efforts have been made in AIS data surveillance, there are still unresolved issues to be addressed. Very few studies [13–15] focus on capturing the real-time high-risk multi-ship encounters and fewer, if not none, on coping with the ship movement dynamics and uncertainty while simultaneously considering both the conflict relations and spatial compactness of traffic clusters. However, failure to address these two issues often leads to a negative impact on identifying actual traffic conflict patterns and decomposing regional collision risk. Hence, much potential remains for improvement by integrating the AIS-based trajectory information into recognizing high-risk traffic clusters. In fact, some recent research studies [72–74] have started considering detection of large-scale traffic congestion in both air and maritime networks. They have developed risk-based interdependency analysis methods to analyse the congestion dependencies and investigate the delay risk propagation, thereby providing valuable insights into relieving traffic delays. However, these studies analysed cascading congestions and identified key nodes/edges based on dependency risk graphs from a network analysis perspective rather than a traffic partitioning perspective. On the other hand, some research communities [75–78] have applied clustering techniques to partition heterogeneous road traffic networks into a small number of spatially compact, connected, and homogeneous regions. This can facilitate the discovery of congested areas and assist in implementing congestion mitigation strategies. Inspired by these studies, an improved clustering approach is developed to simultaneously consider multiple ships' conflict connectivity and spatial compactness. Extensions are also performed to consider the influence of the ships' stochastic and uncertain behaviour on collision detection. As a result, the proposed methodology offers the potential for capturing real high-risk traffic clusters in complex waters.

3 Methodology: a novel density-based clustering approach

As previously stated, this study aims to develop an approach to partition the whole ship traffic over busy waters of interest into multiple clusters, achieving two essential objectives: (1) conflict connectivity and (2) spatial compactness. A conflict-connected ship traffic cluster implies that the ships with high conflict criticality are assigned to the same cluster, whereas a spatial compact cluster suggests that the ships are organized together based on their closeness in space. The two outlined objectives are crucial for detecting multiple ship encounters.

In essence, the two objectives can conflict with each other, as the conflict severity between ships is not entirely related to their spatial distance. It is also influenced by their converging trend, spatial approaching rate, sizes, speeds, etc. Therefore, a new clustering approach is introduced to address this issue. First, a composite distance measure is designed to incorporate conflict criticality into the spatial distance between ship pairs (Section 3.1). Based on the distance measure results, a constrained SSN clustering algorithm is further employed to group the ship traffic under various traffic situations (Section 3.2). Finally, a fine-tuning strategy is applied to guarantee that all ships are in the proper clusters by repeatedly adjusting the boundaries of the clusters (Section 3.3). In addition, a hierarchical bi-objective optimization approach is proposed to adaptively determine the optimal combinations of super parameters to generate clustering solutions robust to the variability of ship traffic situations (Section 4). Detailed descriptions of the critical techniques in each module are shown in Fig. 1 and explicitly highlighted as follows.

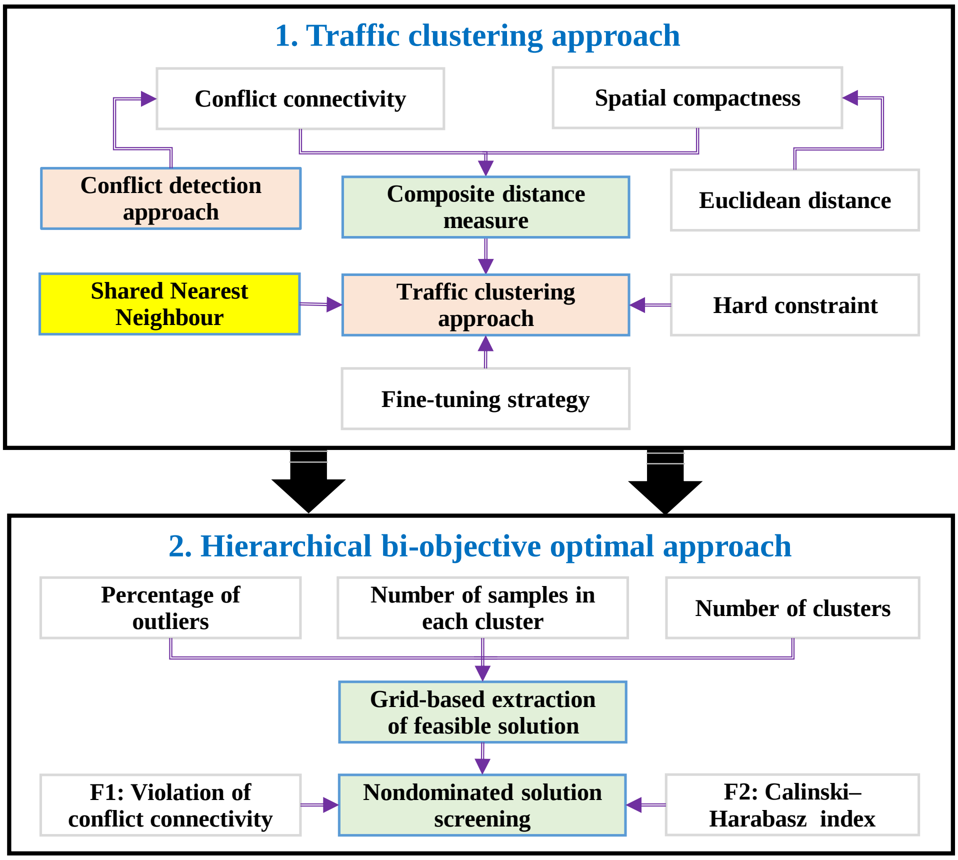


Fig. 1. The research framework.

3.1 Distance measures

Distance/similarity measure is a critical component when implementing clustering algorithms. Therefore, it is necessary to develop a powerful distance measure to assist the clustering approaches to fulfil the two mentioned criteria. In the following subsection, the probabilistic conflict criticality model is first introduced. Then a composite distance measure model that incorporates the conflict relation between ship pairs into their spatial distance is described.

3.1.1 ***Probabilistic conflict characterization and estimation***

A CD approach is designed from a probabilistic perspective to make collision detection adaptive to complex waters. It considers the influence of dynamic and uncertain characteristics inherent to the ships' spatiotemporal movements, providing a quantitative basis for discovering the actual traffic conflict patterns. The proposed CD approach involves two blocks: 1) conflict definition and characterization, and 2) conflict criticality estimation.

Ship conflict is a critical situation where the encountering ships are predicted to violate the minimum allowed distances in the future. This study applies a typical ship domain model [79] to declare the conflict between ships. An example of conflict declaration is displayed in Fig. 2. A potential conflict is deemed to exist when the following equation is held over the look-ahead time horizon.

$$Dist_{AB}(t) \le SD_A(t) + SD_B(t) \tag{1}$$

where $Dist_{AB}$ represents the distance between the ships, and SD_A denotes the distance from the centre of ship A to its domain boundaries.

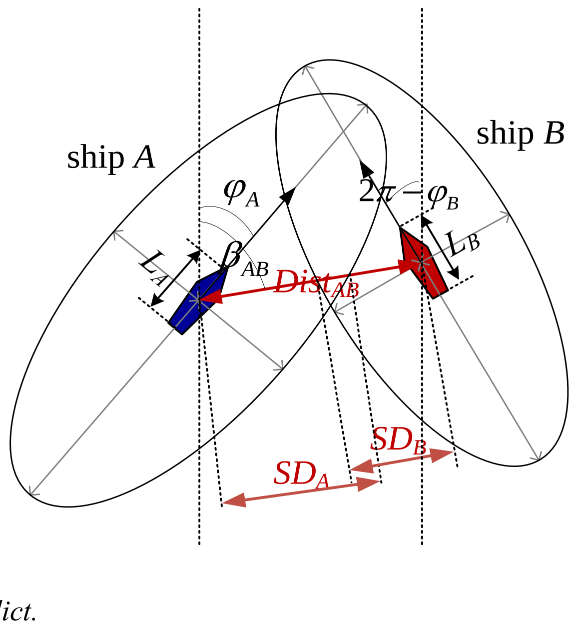


Fig. 2. Definition of ship conflict.

As various uncertainty sources impact future ship trajectories, whether Eq. (1) is satisfied is probabilistic. Assume that the probability density function of the unsafe separations, i.e., $Dist_{AB}(t) - SD_A(t) - SD_B(t) \le$

0, is given by $f_{L(t)}$, the instantaneous occurrence probability of a conflict at a moment *t* is expressed in Eq. (2).

$$PC(t) = \Pr[L(t) \leq 0] = \int_{-\infty}^{0} f_{L(t)}(\rho) d\rho \tag{2}$$

where *L*(*t*) represents $Dist_{AB}(t) - SD_A(t) - SD_B(t)$.

To effectively quantify the expected collision risk between the encounter ships over the look-ahead horizon, a practical conflict criticality measure is designed by incorporating the maximum conflict probability and its occurrence time. These two indices are similar to DCPA and TCPA and play equally important roles in maritime safety navigation [35,80]. One represents the maximum intensity of a conflict, while the other reveals the difficulty level of a conflict resolution as the less time left, the fewer opportunities to address the conflict. Hence, they are combined by an exponential function [81], as shown in Eq. (3).

$$C(\gamma) = MPC^{1+(t_{MPC}/T)} \tag{3}$$

where *MPC* is the maximum *PC*(*t*) over the prediction time horizon, t_{MPC} represents the time associated with the highest conflict probability, and *T* denotes the predicted time horizon.

In addition to the definition of conflict criticality, probabilistic CD estimation encounters two issues for its practical applications: 1) Modelling the future ship movement uncertainty, i.e., the probability density functions of future prediction trajectories; and 2) calculating real-time conflict probability. To address these issues, a probabilistic CD framework is introduced, in which a ship motion model is employed to address the uncertain trajectory forecasting and a two-stage Monte Carlo (TSMC) algorithm is applied to compute the conflict criticality precisely and efficiently. For more details about the probabilistic CD process, one can refer to the work in [82].

3.1.2 *Composite distance measure model*

The composite distance measure model can be further designed based on the conflict criticality measure model. In practice, conflict connectivity is gaining more attention from supervisors or navigators as opposed to density compactness. This is because high conflict criticality explicitly means the potential collision risk, while high traffic density merely implies the busy and sophisticated traffic situation. Additionally, the risk states can be categorized into three groups in terms of probability and/or consequences: intolerable, As Low As Reasonably Practicable (ALARP), and negligible [83]. Here, the conflict criticality is also divided into three states, i.e., intolerable, ALARP, and negligible. The intolerable state indicates that ship pairs must be assigned to the same traffic cluster. The ALARP state represents that the ship pairs should be arranged to the same cluster as much as possible. Finally, the negligible state suggests that the ship pairs are not required to be arranged to the same cluster if the conflict criticality is maintained during the operation.

Based on the risk property and our purpose of revealing the actual conflict patterns, a composite model that incorporates the conflict criticality into the spatial distance is proposed to support a robustness distance measure, which is defined as follows:

$$Dist_{ij} = \begin{cases} 0, \ C(\gamma)_{ij} \geq R_1 \\ Dist_{ij} \times (1 - C(\gamma)_{ij}), \ R_2 \leq C(\gamma)_{ij} < R_1 \\ Dist_{ij}, \ C(\gamma)_{ij} < R_2 \end{cases} \tag{4}$$

where $Dist_{ij}$ denotes the spatial distance between ships i and j, and R_1 and R_2 represent the user-specified parameters to divide the conflict criticality into three states. The first part in Eq. (4) helps merge the ships with intolerable conflict criticality together (0 distance between them), the second part uses weights to strengthen the connection relations between ships with ALARP conflict criticality, while the third part implies that the negligible conflict criticality does not influence the distance measure. The motivation for constructing such an improved distance measure is inspired by the simi-supervised clustering work in [84], in which both the hard and soft constraints are used to spread the influence of the known information. It is noteworthy that the two user-specified parameters (i.e., R_1 and R_2) require to be set reasonably in terms of both the operational authorities' practical demands and the clustering performance, which will be analysed in Section 5.2.

3.2 Constrained Shared Nearest Neighbour (SNN) algorithm

Clustering is a typical unsupervised learning technique that partitions a dataset into multiple clusters based on similarity/distance metrics. It has become vital across many domains, including environmental science, geographical science, information techniques, and business intelligence. Among the existing approaches, DBSCAN is among the most widely used data mining means in the maritime domain because of its desirable properties, such as the ability to cluster datasets with arbitrary shapes, identify and eliminate the noise samples, and automatically discover the number of clusters. For example, it is adopted to extract marine waterway patterns [58], identify maritime traffic turning sections [37], and detect abnormal maritime behaviours [85]. Hence, it is used to perform the traffic clustering task.

Nevertheless, maritime traffic has complex and dynamic features, and directly applying DBSCAN may not produce a desirable clustering solution. On the one hand, DBSCAN cannot handle datasets with varying densities because it uses a fixed density criterion to find all clusters globally [86,87]. It is therefore challenging for DBSCAN to enumerate all ship traffic scenarios, especially for the complicated scenarios with variant traffic densities. On the other hand, DBSCAN might not always guarantee that the high-conflict ship pairs reside in the same clusters, even though combining it with the designed composite distance measure. It is however essential to group the ship pairs with intolerable conflict criticality into the same cluster to capture actual traffic conflict patterns.

As for the first issue, an alternative way to deal with it is to use the Shared Nearest Neighbour (SNN) similarity measure [88]. It is relatively insensitive to variations in density and works well in discovering clusters in uneven regions. Unlike the distance metric in DBSCAN, the SNN focuses on the local configuration of the sample points in the data space. It computes the similarity between two individuals in terms of the number of common elements between their *k*-nearest neighbour lists, which is defined by the following equation:

$$Simlalirity_{ij} = intersect(NN_i(k) \cap NN_j(k)) \tag{5}$$

where $NN_i(k)$ is the subset of k nearest neighbours of individual i. Since the distance (i.e., dissimilarity) is adopted as the input of DBSCAN, the dissimilarity/distance between objects i and j is represented by k minus $Simlarity_{ij}$. Another density-based algorithm, i.e., OPTICS, has been considered for traffic clustering. This algorithm can also handle data with varying densities, however, it is better suited to offline traffic clustering due to its limitations in adaptively identifying the super parameters in real-time [49].

As for the second issue, a hard constraint is integrated into the DBSCAN algorithm to ensure that the ship pairs with intolerable conflict criticality belong to one cluster [89]. The must-link constraints are identified by employing the graph theory. The nodes represent ships, and the edges are used to connect the ship pairs with intolerable conflict criticality. Then a must-link closure (*MLC*) comprising of icc_1, icc_2, …, icc_i, …, icc_s can be obtained, in which icc_i represents a subgraph. Each pair of ships in icc_i can be reached through at least one sequence of edges.

A detailed description of the constrained SNN is depicted in Algorithm 1. Its procedure is an improvement based on the standard DBSCAN algorithm. Three super-parameters (i.e., *EPS*, *MinPts*, and *k*) and a must-link closure (*MLC*) are newly added as the inputs. The super-parameters *EPS* and *MinPts* come from DBSCAN, denoting the maximum radius of neighbours from an observing point and the minimum number of points contained in such neighbours. Based on the two parameters, the relations between two points can be defined as: Directly density-reachable, Density-reachable, and Density-connected. Furthermore, all points are classified as core points, border points, and noise, in which the core points are determined in terms of the number of points that are directly density-reachable (lines 6 and 17). The super-parameter *k* refers to the *k* most similar neighbours of each point being kept for the SNN similarity measure. The improvements of the constrained SNN compared with DBSCAN involve two aspects:

1) A SNN measure is embedded to support a more robust and flexible similarity measure that adapts to traffic scenarios with varying densities (line 3).

2) During the loop iteration, if a point falls into *MLC*, all points belonging to the same subgraph in *MLC* are arranged to the current cluster to satisfy the must-link constraints (lines 10 and 15).

Algorithm 1: Constrained SNN

Input: Given parameter k, *EPS*, and *MinPts*, $I = \{1, 2, \cdots, N\}$, *MLC*

Output: $C = \{c_i\}_{i=1:N}$ // clusters that each point belongs to

// Initialization

1. $N_c = 1$; $v_i = 0$, $N_\varepsilon(i) = []$, $i = 1, 2, \cdots, N$; $\{c_i\}_{i=1:N} = -1$;

// Distance measure

2. Compute the distance matrix based on the composite similarity measure

3. Obtain the SNN measure results based on the parameter k

4. Identify the $N_\varepsilon(i)$ of each point based on the SNN measure results and the parameters *EPS* and *MinPts*

// Loop

5. **For** $i = 1: N$ **do**

6.   **If** $v_i = 0$ && $|N_\varepsilon(i)| \geq$ *MinPts* **then**

7. $v_i = 1$, $c_i = N_c$, *seedlist* = [], and let $I := I\backslash\{i\}$

8.   Insert $N_\varepsilon(i) \cap I$ into the *seedlist*, $v_{N_\varepsilon(i)\cap I} = 1$, $c_{N_\varepsilon(i)\cap I} = N_c$, and let $I := I\backslash\{N_\varepsilon(i) \cap I\}$

9.   **If** point $i \in$ *MLC* **then**

10.   Insert $MLC(i) \cap I$ into the *seedlist*, $v_{MLC(i)\cap I} = 1$, $c_{MLC(i)\cap I} = N_c$, and let $I := I\backslash\{MLC(i) \cap I\}$

11.   **End**

12.   **While** NOTEMPTY(*seedlist*) **do**

13.   Get an element j from *seedlist*, and let *seedlist* := *seedlist*$\backslash\{j\}$

14.   **If** point $j \in$ *MLC* **then**

15.   Insert $MLC(j) \cap I$ into the *seedlist*, $v_{MLC(j)\cap I} = 1$, $c_{MLC(j)\cap I} = N_c$, and let $I := I\backslash\{MLC(j) \cap I\}$

16.   **End**

17.   **If** $|N_\varepsilon(j)| \geq$ *MinPts* **then**

18.   Insert $N_\varepsilon(j) \cap I$ into the *seedlist*, $v_{N_\varepsilon(j)\cap I} = 1$, $c_{N_\varepsilon(j)\cap I} = N_c$, and let $I := I\backslash\{N_\varepsilon(j) \cap I\}$

19.   **End**

20.   **End**

21. $N_c = N_c + 1$ // update the serial number of clusters

22.   **End**

23. **End**

I: sample dataset; N_c: serial number of clusters; $N_\varepsilon(i)$: neighbours of point i in *Eps*-based Radius-neighbourhood; v_i: 1 represents point i is touched and 0 otherwise; $MLC(j)$: subgraph in *MLC* that point i belongs to.

3.3 Fine-tuning

During the clustering process, ships sometimes are assigned to improper clusters. This is due to the fact that: 1) the distribution of ship traffic in space is random and subject to various disturbances, which has exposed unobvious cluster distribution characteristics in some cases, and 2) the SNN measure considers the overlap between the objects' k-nearest neighbours lists instead of the density values to find clusters in varying density regions. It could produce clusters with significant density differences for some complex traffic scenarios. Therefore, a fine-tuning strategy is presented to further adjust the ships on the boundary to the proper clusters. It consists of two phases: 1) deleting the points far away from the cluster and 2) adding the outliers into the proper clusters.

Fig. 3 illustrates three typical clusters produced that can be improved using outlier deletion. It is apparent that the clustering performance can be enhanced by eliminating the red points. To achieve this goal, each point's distance from its *2^nd^* nearest neighbour (*2*-nn) in the same cluster is identified and ranked. The *2*-nn distance of each point is shown at the bottom of Fig. 3. According to these diagrams in Fig. 3, the red points have a substantially larger *2*-nn distance than other points. Therefore, these red points are deleted in an iteration way. In each iteration, if the point with the largest *2*-nn distance satisfies the following conditions, it will be set as an outlier:

(1) Its *2*-nn distance is larger than $th1$ (a coefficient larger than 1 to help eliminate the points far away from the clusters) times the third largest *2*-nn distance.

(2) It does not have an intolerable or ALARP conflict with the points in the same cluster.

The process will continue for the current cluster until no point is deleted. The reason for using the *2*-nn instead of *1*-nn is that the latter cannot work properly for case 3, where two points far away from the cluster are close. A detailed algorithmic step of the outlier deletion is provided in Algorithm 2.

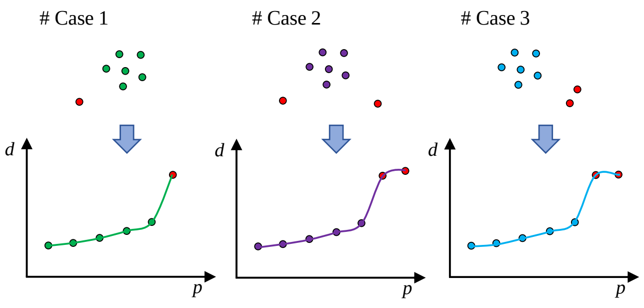


Fig. 3. Graphical illustration of typical clusters produced that can be improved using outlier deletion. (The longitudinal axis 'd' represents the point's distance from its 2^{nd} nearest neighbour, and the horizontal axis 'p'

represents the point label.)

Algorithm 2: Outlier deletion

Input: $C = \{c_i\}_{i=1:N}$, *MHCC*
Output: $C = \{c_i\}_{i=1:N}$

```
For i = 1: N_C do
    While |C_i| ≥ min_np do
        Determine each point's distance (Dist) with its 2nd
            nearest neighbour in C_i
        Rank the points as {i1, i2, ⋯ iC_i} in descending
            order of the 2-nn distances
        If i1 ∉ MHCC && Dist_i1 > th1 × Dist_i3 then
            c_i1 = -1;
        Else
            break
        End
    End
End
```

C_i: subset of points that belong to ith cluster; *MHCC*: dataset of ships that have intolerable or ALARP conflict with others.

Fig. 4. displays two typical clusters produced that can be improved using outlier addition. The red points represent the outliers marked by the constrained SNN. It is shown that the outliers' distances with the points in the cluster are relatively close. The outliers are assumed to belong to the cluster, and the average distance between each point and other points in the assumed new cluster is calculated. From the right diagrams in Fig. 4, the average distances of outliers are not the largest. Therefore, the outliers are added to make the cluster more compact. The outliers that have intolerable or ALARP conflict relations (case 2) can be added only if they satisfy the above distance requirements simultaneously. The procedure of outlier addition is codified in Algorithm 3. Note that the outliers are preferentially added to the closer clusters (lines 3-4).

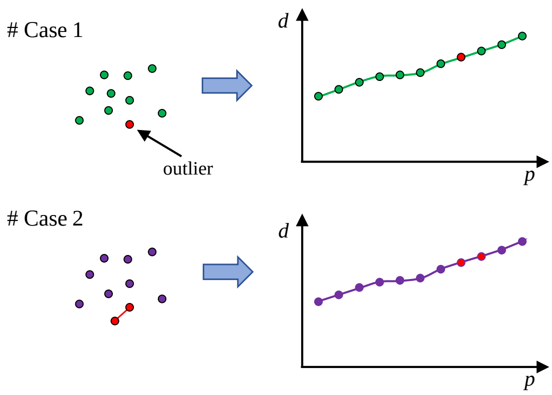


Fig. 4. Graphical illustration of typical clusters produced that can be improved using outlier addition. (The longitudinal axis 'd' represents the point's average distance from other points in the same cluster, and the horizontal axis 'p' represents the point label.)

Algorithm 3: Outlier addition

Input: $C = \{c_i\}_{i=1:N}$; $v_{oi} = 0$, $i = 1, 2, \cdots\cdots, |O|$.
Output: $C = \{c_i\}_{i=1:N}$

For $i = 1: |O|$ **do**
  If $v_{O_i} = 0$ && $MHP_{O_i} \notin O$ **then**
    $v_{O_i} = 1$ and calculate the distance between O_i with each cluster
    Rank these clusters as $\{i1, i2, \ldots, iN_C\}$ in ascending order of distances
    For $k = 1: N_c$ **do**
      Assume that O_i belongs to the ikth cluster and calculate average distance ($Dist$) between each point and other points in the cluster
      If $Dist_{O_i}$ is not the largest **then**
        Let c_{o_i} equal to the mark of the ikth cluster and break
      End
  Elseif $v_{o_i} = 0$ && $MHP_{O_i} \in O$ **then**
    Let $X_i = [O_i, MHP_{O_i} \cap O]$, $v_{X_i} = 1$, and calculate the average distance between X_i with each cluster
    Rank these clusters as $\{i1, i2, \ldots, iN_C\}$ in ascending order of distances
    For $k = 1: N_c$ **do**
      Assume that X_i belong to the ikth cluster and calculate average distance ($Dist$) between each point and other points in the cluster
      If $\max\{Dist_{X_i}\}$ is not the largest **then**
        Let c_{X_i} equal to the mark of the ikth cluster and break
      End
    End
  End
End

O: dataset of outliers; MHP_{O_i}: points that have intolerable or ALARP conflict with the ith outlier; v_{oi}: 1 represents the ith outlier is touched and 0 otherwise.

It is worth mentioning that the above steps play an essential but modest role as the points with an intolerable or ALARP conflict are rarely processed. Despite this, the fine-tuning strategies help to yield more reasonable clustering results. This is further demonstrated in detail in Section 5.4.

4 Hierarchical bi-objective optimization approach

One direct obstacle to implementing the proposed clustering approach lies in how to adaptively determine the super-parameters to ensure the clusters' conflict connectivity while keeping their spatial compactness. More importantly, the two criteria conflict with each other. Hence, their non-dominant relations need to be considered to allow simultaneous optimization [80]. Additionally, using different combinations of super-parameters leads to a variety of clustering results, many of which are infeasible solutions. This study employs a hierarchical bi-objective optimization algorithm for a solution.

Algorithm 4: Hierarchical bi-objective optimization

Input: Dataset, value range of k, $MinPts$ and EPS.
Output: Non-dominated solution set Ψ.
// Candidate pool generation
1. Perform the extended clustering for different combinations of k, $MinPts$, and EPS based on a grid-search strategy
2. Store all clustering results into a candidate pool
// Feasible solution extraction
3. **For** $i = 1: N_k \times N_{MinPts} \times N_{EPS}$ **do**
4. **If** $NP_{CR_i} < noise\_percent$ && $NC_{CR_i} \geq min\_nc$ && $MNP_{CR_i} \geq min\_np$ **then**
5. Add the ith clustering result CR_i into Γ
6. **End**
7. **End**
// Nondominated solution screening
8. **While** NOTEMPTY(Γ) **do**
9. **If** CR_1 is not dominated by any CR_i **then**
10. Add CR_1 into Ψ
11. **End**
12. **If** any CR_i in Γ is dominated by CR_1 **then**
13. Let $\Gamma := \Gamma \backslash \{CR_i\}$
14. **End**
15. Let $\Gamma := \Gamma \backslash \{CR_1\}$
16. **End**

NP_{CR_i}: outlier ratio of CR_i; NC_{CR_i}: number of clusters of CR_i;
MNP_{CR_i}: number of points of the minimal cluster in CR_i.

Algorithm 4 presents the overall procedure of hierarchical bi-objective optimization. The proposed algorithm is first performed based on a grid-search strategy to generate a candidate pool. Then the pre-processing phase eliminates the infeasible solutions as feasibility takes priority over the two optimization criteria. Three constraints are considered: 1) the noise/outlier ratio of the dataset should be less than *noise_percent* because a larger number of outliers imply that the ship traffic clusters are not accurately detected; 2) the number of clusters generated should be larger than *min_nc* because a smaller number of clusters cannot effectively decompose the whole collision risk for high traffic waters; and 3) the number of points in each cluster should be not less than *min_np* because a group with a certain number of points can be regarded as a cluster/group. To guarantee that the formulated clusters are conflict-connected and spatially compact, the non-dominant solution set is further searched for from the feasible solution pool. With respect to the performance of the two criteria, two indexes are used to evaluate and examine the effects. The conflict connectivity is measured by the degree that the conflicting ship pairs are not arranged into the same cluster, which is defined as：

$$V_{cc} = \sum_{i=1}^{N_{vc}} C(\gamma)_i \tag{6}$$

where N_{vc} represents the number of conflicting ship pairs that are not in the same cluster, $C(\gamma)_i$ represents the associated conflict criticality. The smaller the V_{CC} index, the better the clustering performance in terms of conflict connectivity. The spatial compactness is calibrated by the Calinski–Harabasz (*CH*) index [90], which is defined as:

$$CH = \frac{[trace\ B/(c-1)]}{[trace\ W/(n-c)]} = \frac{[\sum_{k=1}^{c} n_k ||z_k - z||^2]/(c-1)}{[\sum_{k=1}^{c}\sum_{j=1}^{n_k} ||x_k - z_k||^2]/(n-c)} \tag{7}$$

where n is the number of samples, c denotes the number of clusters, n_k represents the number of samples belonging to the kth cluster, z_k is the centre of the kth cluster, and z is the centre of the entire dataset. A larger *CH* index indicates that the inter-cluster distance is larger while the intra-cluster distance is smaller. It implies that the spatial compactness of produced clusters is well satisfied.

Based on the above algorithm, the two optimization objectives can be simultaneously considered, and the generated clustering solutions can be robust and resilient to the variability of ship traffic situations.

5 Case study

5.1 Study area and data description

This study chooses the Ningbo-Zhoushan Port, the world's largest port according to cargo throughput, as the case to test the applicability and effectiveness of the proposed traffic clustering approach. It exhibits heavy traffic, restricted navigable areas, and dynamic and uncertain traffic behaviours, hence being an ideal representative of complex waters. Fig. 5 shows the port region, which is bounded by latitudes between 29°43’N and 30°02’N, and by longitudes between 121°52’E and 122°22’E. One-month AIS data records during 01/11/2018–30/11/2018 are collected from the region for real-time traffic cluster identification. Indeed, the raw AIS data received is highly informative, but its inherent weaknesses like data noise and position/speed information errors caused by data transmission or technical failures are often criticized. Hence, it is crucial to conduct data cleaning and filtering. In this study, a systematic cleaning procedure is employed by referring to the pre-processing methods in [64,91]. The procedure can enhance the data quality and reconstruct clean trajectory data.

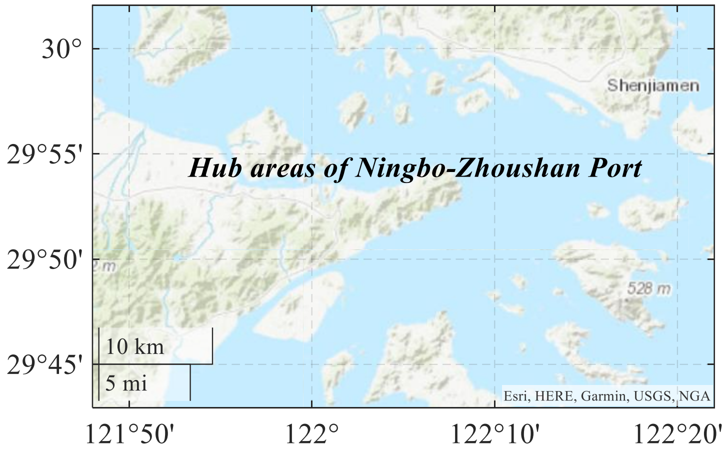


Fig. 5. Hub area of Ningbo-Zhoushan Port, China

5.2 Parameter sensitivity analysis

According to the methodologies in Section 3.2, three super-parameters (i.e., *EPS*, *MinPts*, and *k*) directly influence the model performance and need to be calibrated to determine the optimal clustering structure. This can be addressed by a grid-search strategy to find the best combinations of super-parameters. However, a large parameter value range will lead to a higher exploration of the search space and increase computational costs. As a result, it is crucial to search for the best clustering performance within reasonable ranges of parameter search space.

The influence of *EPS* and *MinPts* is first investigated by using the traffic scenarios extracted every moment from one week of AIS data. Because of the spatial heterogeneity, different traffic scenarios may correspond to different optimal parameter combinations. Therefore, the traffic scenarios are divided into different groups to determine the reasonable super-parameter search ranges under different scenarios. They are categorized into four groups based on the number of ships, which are 30-60, 60-90, 90-120, and >120. Fig. 6 shows the probability density distribution of the number of ships. The traffic scenarios with a number of ships less than 30 are not analysed because they are associated with low traffic complexity. It is not necessary to capture the corresponding real-time traffic clusters. Fig. 7 (a) illustrates the percentage distribution of the optimal *EPS* values under different traffic scenarios. It is seen that almost all the optimal *EPS* values are located within [2, 8]. Besides, a traffic scenario with a larger number of ships has a higher probability of corresponding to a larger optimal *EPS* value. The same phenomenon can be found for *MinPts*. According to Fig. 7 (b), the optimal *MinPts* values mainly fall within [1, 11] and disappear when *MinPts* is larger than 11. These results allow us to search for the optimal *EPS* between 2-8 and *MinPts* between 1-11, respectively.

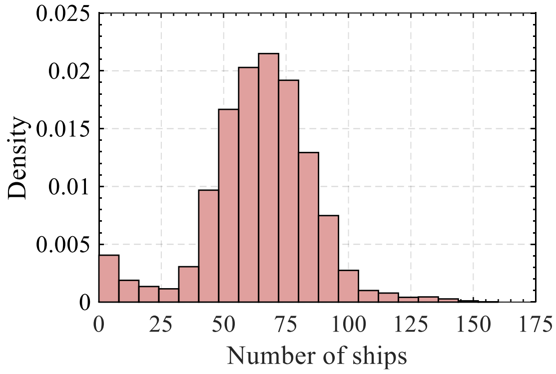


Fig. 6. Probability density distribution of the number of ships at each time slice in the investigated waters.

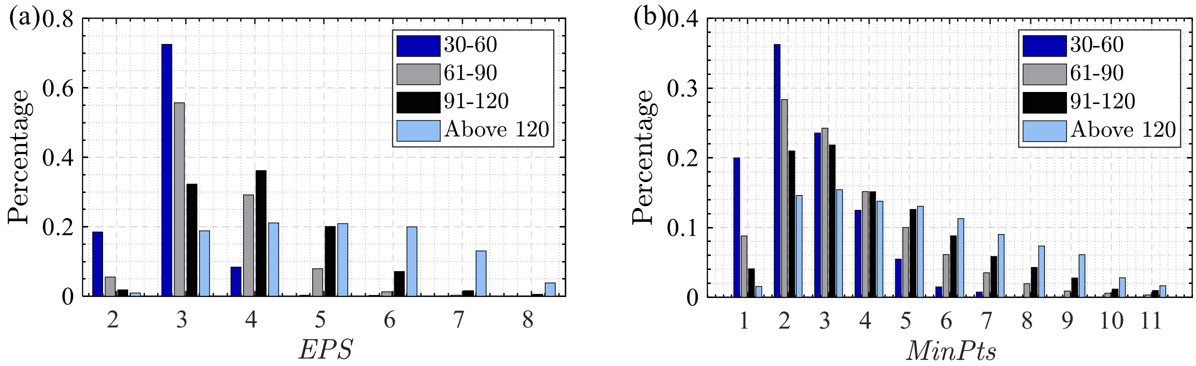


Fig. 7. (a) Percentage distribution of the optimal EPS values; (b) percentage distribution of the optimal Minpts values. (The legend represents the number of ships for clustering.)

Fig. 8 illustrates the impact of *k* on the performance of the clustering results. The ratio of "*k* to the number of ships" (k/N) is used as the variable considering the high fluctuation of the dataset size. From the figure, the optimal k/N values are mainly concentrated around 0.1. The percentage of the optimal k/N accounts for 99% between 0-0.65, and the influence becomes negligible when the value exceeds 0.65. Thus, the search range of k/N is set to be [0, 0.65].

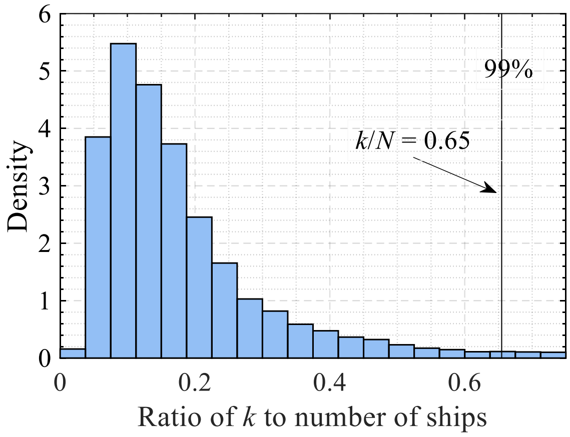


Fig. 8. Probability density distribution of the ratio of "k to the number of ships" that corresponds to the optimal clustering results.

In addition, the two specified parameters that divide conflict criticality into three states play a vital role in the clustering performance. In fact, the conflict model measures the probability of ship domain overlap, and it is far more frequent than collision [25,92,93], suggesting that a higher conflict criticality is acceptable compared with a higher collision probability. Therefore, R_1 and R_2 are set as [0.05, 0.1, 0.2, 0.3] and [0.3, 0.5, 0.7, 0.9] to compare the performance of these combinations. As two clustering objectives are considered simultaneously, it must determine which parameter combination is better based on the dominant relations of the clustering results instead of using one specific indicator to describe their performance. As a result, the number of times that each combination is superior to other combinations is counted, as shown in Fig. 9. It is noticeable that the combination of R_1 = 0.05 and R_2 = 0.5 has the best performance as it dominates all other combinations. This enables us to determine the critical parameter combination to achieve desirable traffic partitioning.

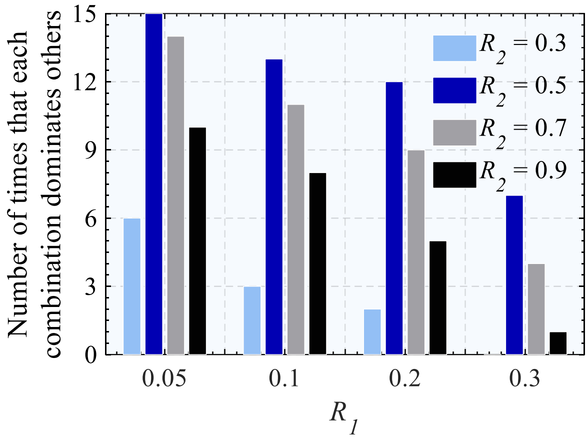


Fig. 9. Clustering performance comparison of various combinations of R_1 and R_2.

5.3 Clustering performance demonstration

To test the performance of our proposed approach, its online application in high-risk traffic cluster identification is illustrated through a case study. An example of the traffic clustering performance within an hour in the study area is depicted in Fig. 10. In Fig. 10 (a)-(d), the visualization of ship traffic clustering at time t = 0, 20, 40, and 60 min is presented, where the points represent the spatial distribution of ship traffic, the red lines represent the intolerable conflict between ships, and blue lines represent the ALARP conflict. It is found that most ship pairs with intolerable or ALARP conflict are arranged in one cluster, and the ships in the same cluster have a spatial compact shape. This implies that the high-risk/density traffic clusters are effectively detected by partitioning the ship traffic into clusters. In addition, it is observed that there are lots of multi-ship encounters with close distances. Hence, the traditional models focusing on the collision risk of ship pairs provide little insight into the MSA for this type of heavy-traffic waters. However, the proposed approach decomposes the whole ship traffic by considering the multi-ship relations, which offer a higher resolution to the identification of regions with high risks.

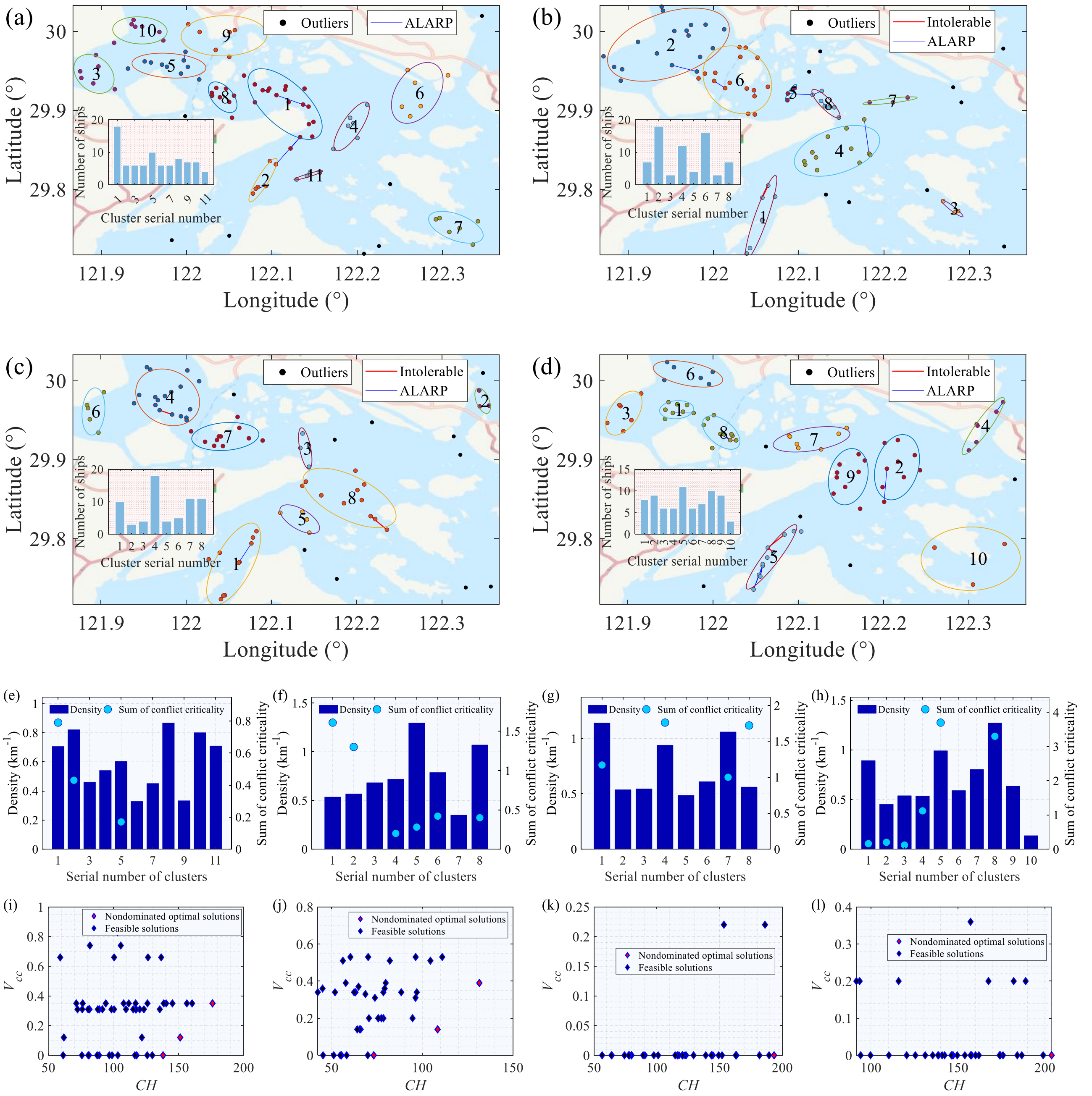


Fig. 10. Illustration of the evolution of ship traffic clustering results, including ship traffic cluster visualization, traffic density, sum of conflict criticality of each cluster, and objective function value distribution of feasible solutions at time t = 0, 20, 40, and 60 min.

Based on the detected ship traffic clusters, one can further rank and identify the critical traffic clusters. Two indices are utilized to describe each cluster's characteristics. One index is the sum of conflict criticality, and the other is the average density of the ships in one cluster (see [94] for the density definition). Fig. 10 (e)-(h) displays each cluster's two indices for the above four scenarios. According to these figures, it can be clearly found which clusters should be more concerned. For example, it is observed that Clusters 1 and 2 in Fig. 10 (e) have high conflict criticality and Cluster 5 in Fig. 10 (f) has high traffic density. Additionally, the traffic density and conflict criticality of each cluster are significantly unbalanced and heterogeneous, which shows the necessity of identifying the key clusters. The proposed traffic clustering approach offers valuable insights into where to enhance maritime surveillance and offer hazard warnings.

To help better understand the proposed hierarchical bi-objective optimization strategy, the sets of feasible clustering solutions of the four scenarios (see Fig. 10 (i)-(l)) are plotted. The red diamonds stand for optimal solutions that are not dominated by others. In Fig. 10 (i) and (j), more than one optimal solution is non-dominated by others. These non-dominant solutions constitute the Pareto front. For these cases, the intermediate solution in the Pareto front is chosen as the final clustering scheme. In this way, the optimal solutions that achieve a sensible trade-off between conflict connectivity and density compactness are retained.

Fig. 11 further shows the conflict criticality and density evolution of the traffic clusters and whole traffic. Note that the maximum conflict criticality and density of the traffic clusters produced at each time slice are presented. Two phenomena can be derived from the figure. Frist, the global traffic density varies slightly with time, whereas the maximum traffic cluster density has considerable fluctuations. Such results suggest that the global traffic density provides little insight into enhancing maritime traffic surveillance capabilities. In contrast, detecting high-density traffic clusters can assist the operators in better understanding the traffic situation and improve their working ability to deal with complex traffic scenarios. Second, it is observed that the conflict criticality of global traffic and traffic clusters display consistent trends. This is of great practical significance to anti-collision risk management because the high-conflict clusters are explicitly captured. It can help determine the conflict resolution constraints and facilitate the implementation of risk mitigation measures. In summary, the proposed approach shows potential to be applied to the intelligent transportation service system to relieve the operator’s safety monitoring pressure in the future.

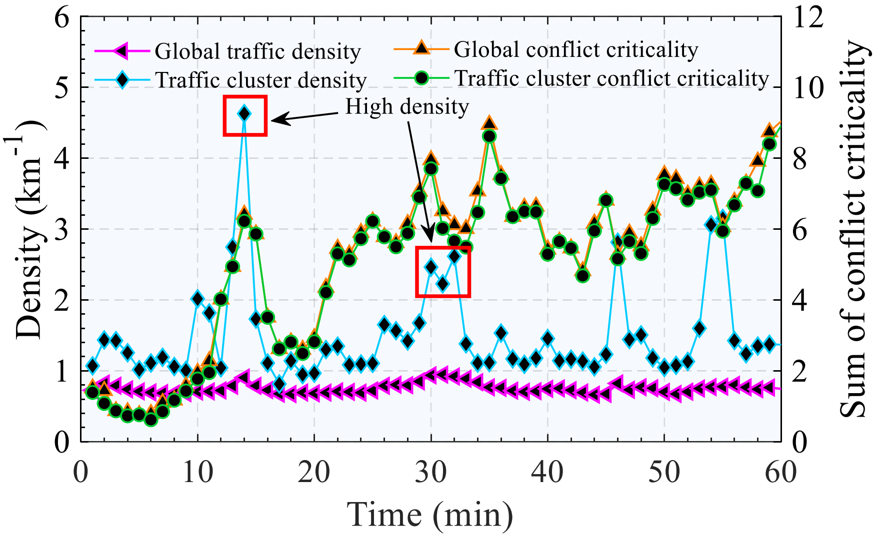
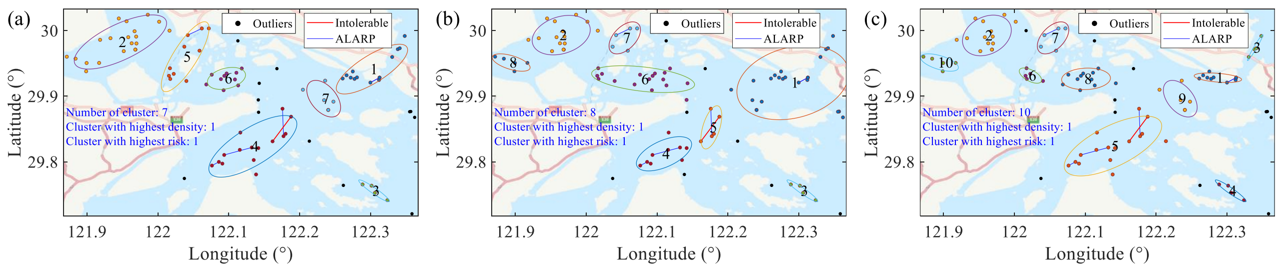


Fig. 11. Density and conflict criticality evolution of ship traffic within one hour.

Additionally, the clustering approach's performance under different clusters is also investigated. Fig. 12 depicts one traffic scenario’s clustering results with varying numbers of clusters. Table 1 presents the objective values of the clustering results with different numbers of clusters. The optimal number of clusters is 13/14 in terms of the dominant relations of objective values. However, the operators can adaptively specify the desired number of clusters for the practical scenarios according to their individual preferences and the clustering performance. Therefore, the proposed approach can offer a multi-view analysis of the traffic scenario and provide a scalable clustering solution.



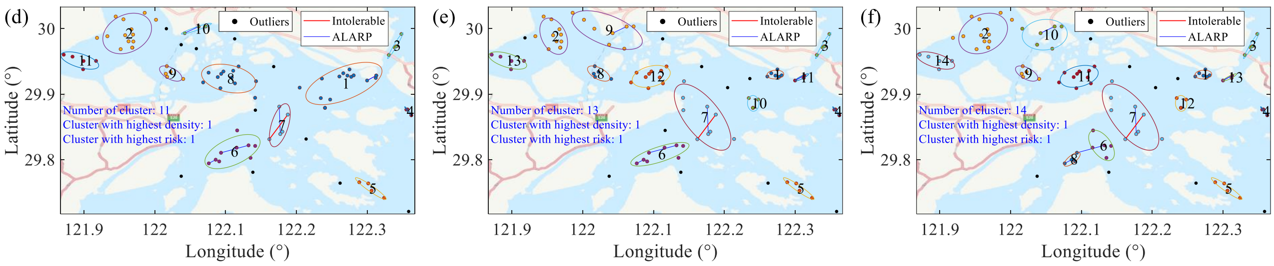


Fig. 12. Clustering result illustrations with different numbers of clusters.

Table 1. Optimal values of two objectives with respect to different numbers of clusters

Objective function	Number of clusters					
	7	8	10	11	13	14
V_{cc}	0	0	0	0	**0**	**0.3**
CH	210.5	162.9	239.5	277.0	**292.9**	**298.3**

5.4 Model performance comparison and validation

The methodological comparison and validation are critical to ensure the superiority and effectiveness of the model results. Therefore, it starts by examining the performance of the proposed traffic cluster approach by comparing it with the DBSCAN algorithm in [13–15]. All configurations of the two algorithms are identical except that DBSCAN does not perform the SNN measure. Two super-parameters (i.e., *EPS* and *MinPts*) in DBSCAN are taken as [1, 1.5, …, 6 km] and [2, 3, …, 10] to search for the best clustering results. Fig. 13 shows the percentage of the traffic scenarios with feasible clustering solutions when using the two algorithms. It is observed that almost all traffic scenarios can obtain feasible solutions through the proposed approach. In contrast, only 33.39% of traffic scenarios using DBSCAN have feasible solutions when adopting the same constraint settings (see Section 4). Even though the tolerance of the outlier percentage is adjusted up to 20%, there are still a large proportion of traffic scenarios without feasible solutions based on DBSCAN. These results suggest that DBSCAN has limitations in handling complex traffic scenarios with varying densities.

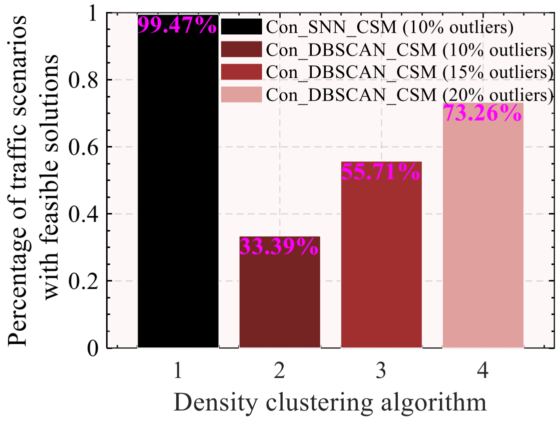


Fig. 13. Percentage of ship traffic scenarios with feasible clustering solutions.

In contrast, the proposed clustering approach overcomes the weaknesses by self-adaptively determining the density threshold rather than depending on a constant global value. For example, Fig. 14 illustrates the DBSCAN's optimal clustering performance for the first case in Fig. 10. As shown in the figure, the size of the clusters produced in Fig. 14 is highly uneven, and the outlier percentage reaches 14.3%. However, the clusters produced in Fig. 10 (a) are balanced in size, showing better traffic cluster identification performance. In addition, the dominant relations of the two algorithms are compared (see Fig. 15). The percentage of the traffic scenarios in which the proposed approach outperforms DBSCAN takes up 93.83%. These analyses validate the proposed approach's applicability in dealing with traffic scenarios in complex waters.

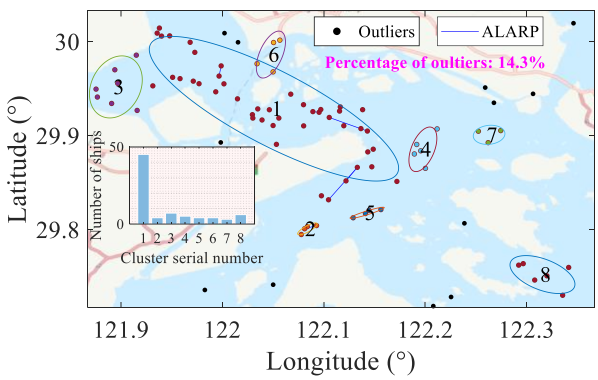


Fig. 14. Clustering result visualization using the DBSCAN algorithm at t = 0 min.

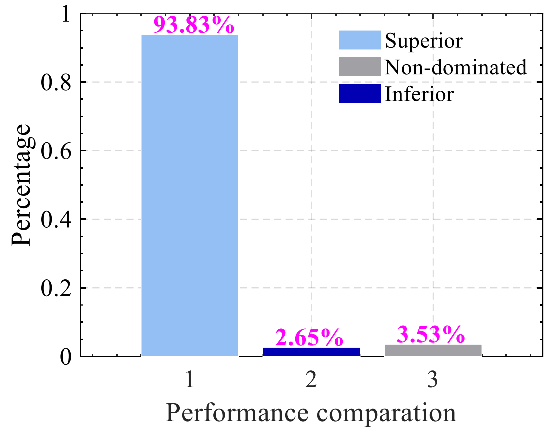


Fig. 15. Performance comparison between the proposed approach and DBSCAN, where "Superior" represents that the proposed method outperforms DBSCAN and "Inferior" otherwise.

To prove the functionality and utility of the key modules in the proposed clustering approach, it is compared with the following algorithms: the standard SNN (SNN), the standard SNN employing the composite distance measure (SNN_CDM), and the constrained SNN without employing the composite distance measure (Con_SNN). Fig. 16 provides the comparison results in terms of their dominant relations. These results show that the proposed approach significantly outperforms the SNN and SNN_CDM. This implies that constructing hard connection constraints for intolerable conflicting ship pairs leads to a substantial improvement in the clustering performance. In the meantime, it can be observed that the performance of the proposed clustering model is better than that of SNN_CDM to some extent. This is because the composite distance measure model is designed in a more precise way to allow for the simultaneous consideration of the conflict connectivity and spatial compactness of clustered ship traffic. However, the traditional similarity/distance measures cannot cope with multiple clustering objectives. Their effects are thereby poor when several clustering properties need to be considered. Therefore, the hard constraints and the composite distance measure are critical to improving the clustering performance.

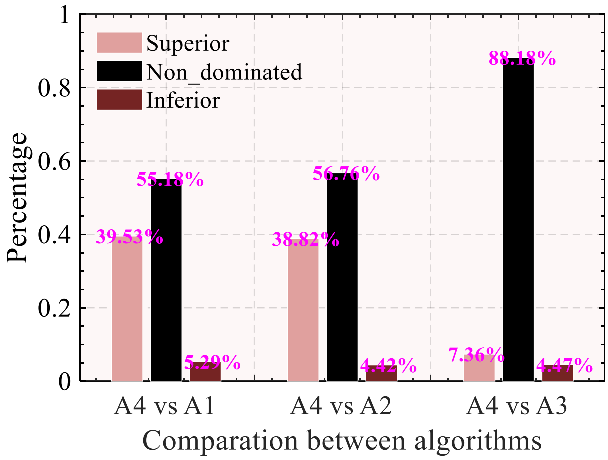


Fig. 16. Comparison results between the proposed approach and other algorithms. A1, A2, A3, and A4 represent SNN, SNN_CDM, Con_SNN, and the proposed method (Con_SNN_CDM).

Furthermore, the performance of the different models under various traffic scenarios is compared to evaluate the generalization ability and stability of the proposed clustering approach. As shown in Fig. 17, the performance of the proposed approach is better than that of all other models for the scenarios with different numbers of ships. These results confirmed the robustness of the proposed clustering approach.

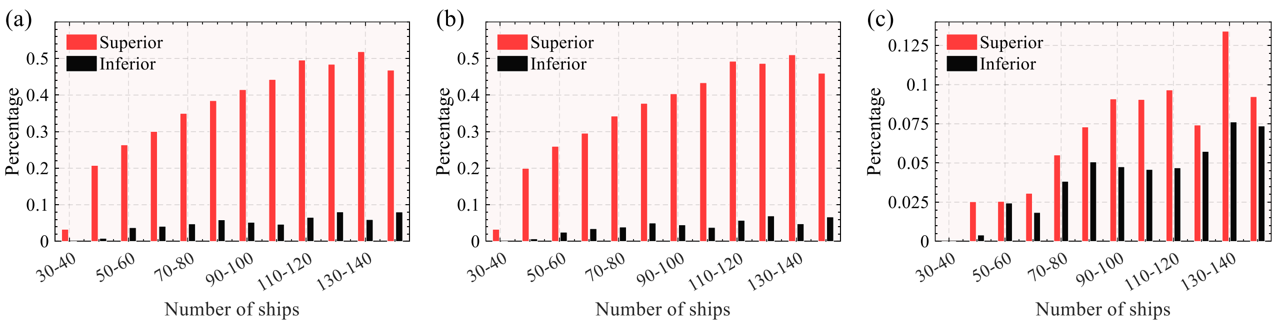


Fig. 17. Comparison between the proposed algorithm with other algorithms under different traffic scenarios. (a) the proposed algorithm vs. SNN; (b) the proposed algorithm vs. SNN_CDM; (c) the proposed algorithm vs. Con_SNN.

In addition, a typical example of the traffic clustering performance improvement using the fine-tuning strategy is displayed in Fig. 18. As shown in Fig. 18, ships *i*, *j*, and *o* are assigned to the clusters, but they are far away from the ships in the same clusters. This is mainly because the complete robustness of the constrained SNN algorithm cannot be guaranteed due to the traffic scenarios' stochastic characteristics and the SNN measure's inherent properties. However, after performing the outlier deletion step, these ships are set as outliers (see Fig. 18 (b)). As for ship *p*, it is close and then added to Cluster 6 based on the outlier addition step (see Fig. 18 (c)). The distance relations between these adjusted ships with others in the same clusters are shown in Fig. 18 (d). The above results indicate that the fine-tuning step enables the ships to be adjusted to the proper clusters. It should be noted that a traffic scenario associated with underperforming clustering results, without using the fine-tuning strategy, is utilized to show how the fine-tuning strategy further adjusts the clustering results. It is also significant that most traffic scenarios could correspond to a good clustering performance even without the fine-tuning strategy. However, the necessity and benefit of the newly introduced strategy could supplement and ensure good clustering performance in every possibility. Applying the fine-tuning strategy therefore renders improvements in the clustering performance.

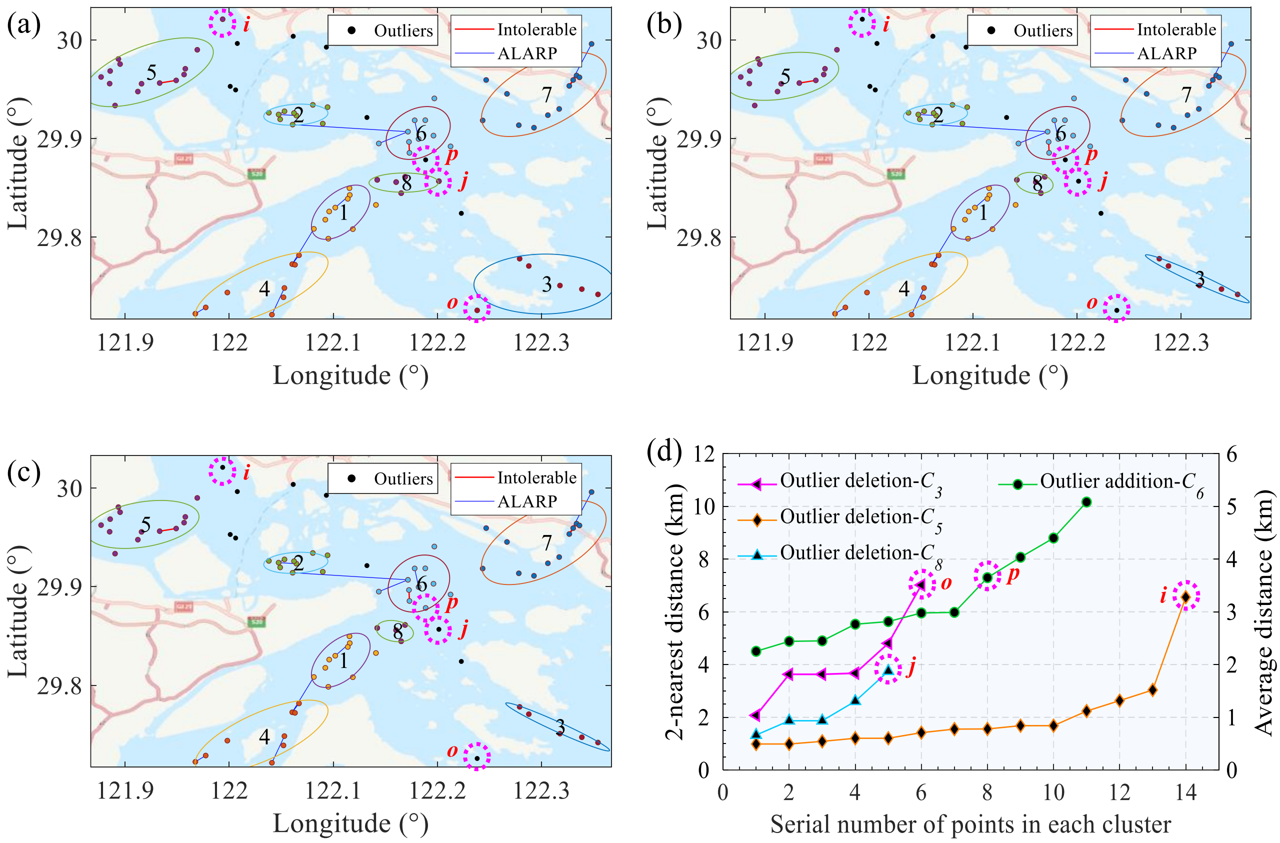


Fig. 18. Performance demonstration of the fine-tuning strategy. (a) clustering result visualization without performing the fine-tuning step; (b) clustering result visualization after performing the outlier deletion; (c)

clustering result visualization after performing the outlier addition; (d) distance relations between the ships with others in the same traffic clusters.

5.5 Discussion

Currently, a great variety of clustering techniques have been used to analyse AIS data. However, the most existing research is focused on traffic feature extraction and knowledge discovery, such as ship motion behaviour recognition [41] and ship traffic network construction [50]. The identification of real-time high-risk multi-ship encounters based on AIS trajectory data is yet to be investigated and addressed. It is an emerging topic under development in the maritime traffic domain, as claimed in the work in [13–15]. Therefore, a theoretical comparison analysis is conducted with these relevant studies to illustrate the differences and superiority of the proposed approach. Basically, the multi-ship encounter recognition using clustering involves three new modules: modelling of similarity measure, selection of clustering algorithms, and identification of super parameters.

With respect to the similarity measure modelling, for the first time, it considers both the conflict criticality and spatial distance among ships for maritime traffic partition. Existing research identified the real-time traffic clusters merely based on the spatial distance attribute of ship traffic, which fails to effectively incorporate the multi-attribute dependencies among ships. This study adopts the risk states to divide the conflict criticality into three states and then refers to a constrained Tired Random Walk (TRW) in [84] to combine the two dependence measures to generate a new similarity model. Additionally, the conflict detection model integrates the ship motion dynamics and uncertainty to ensure its applicability to complex waters.

Regarding the selection of clustering algorithms, this study also adopts DBSCAN due to its three desirable properties previously mentioned. Notably, the direct application of DBSCAN may result in undesirable outputs due to its limitation in handling datasets with varying densities and high sensitivity to the super parameters. Therefore, compared with these existing works, two improvements are undertaken in this study. One is to incorporate an SNN similarity measure to make it insensitive to the variations in traffic density. The other is to integrate a hard constraint to ensure that the ships with high conflict criticality are assigned to the same cluster. In this way, its robustness and scalability with respect to various traffic scenarios are guaranteed.

In terms of the identification of super parameters, a hierarchical bi-objective optimization algorithm is proposed. In [13], the traffic grid with different sizes is used to optimize the super parameters to adapt to different density surfaces. However, it cannot explicitly address the hierarchical priorities among different objectives. The newly proposed algorithm in this research allows the different objectives to be hierarchically addressed and strikes a balance between the conflict connectivity and spatial compactness of traffic partition results.

Despite the proposed analytical approach showing improved performance and superiorities over traditional methods in the field, it has some limitations, including the following.

1. The effects of water topography on spatial distance calculation need to be explored. In this study, the spatial distance between ship pairs is measured based on a Euclidean distance model. However, the two spatially close ships (of a short distance) may not necessarily be at collision risk if there are obstacles between them (e.g., islands and Skerrys). Hence, further effort could be made to construct the maritime traffic network by extracting the network nodes (waypoints including entry, exit, and turning points) and the traffic routes, and thus a reasonable spatial distance measure model can be designed in terms of the extracted vessel movement patterns.
2. The dynamic evolution characteristics of ship traffic clusters could be further investigated. Incorporating traffic dynamic evolution characteristics is critical for applying risk management strategies for the traffic clusters, particularly if they rapidly change over time. Therefore, it is essential to construct a new traffic clustering framework that can offer to incorporate temporal smoothness to generate stable and consistent

traffic clusters.

3. This study ranks and captures the critical traffic clusters based on the sum of conflict criticality and the average density. Therefore, a traffic cluster associated with a larger number of ships normally has a higher conflict criticality. Additionally, the relationships between the traffic cluster risk and the traffic topological properties influencing the resolving difficulty of collisions should be further considered. Some advanced techniques such as complex network theory are promising to effectively measure the interactions between multiple ships to assist in the identification of key traffic clusters.

6 Conclusion and future work

Detecting the potential high-risk ship traffic clusters in complex waters is critical for MSA and safety management. This study proposes a systematic maritime traffic clustering approach to adaptively capture high-risk multi-ship encounters. It effectively integrates a composite distance measure, a constrained SNN clustering, a fine-tuning strategy, and a hierarchical bi-objective optimization, enabling the whole ship traffic to be partitioned into multiple conflict-connected and spatially compact clusters. The developed methodology has been integrated with the following new features: 1) being capable of adaptively discovering the traffic clusters with strong conflict connectivity at high density; 2) allowing for robust handling for the traffic scenarios with varying densities; and 3) being capable of helping determine the conflict resolution constraints and assist in the design and deployment of anti-collision risk mitigation solutions.

A case study using the historical AIS data from the Ningbo-Zhoushan Port is conducted to evaluate and check the proposed clustering approach. Experimental results reveal that the clustering methodology has a reliable and rational performance in capturing the high-risk/density clusters in complex traffic scenarios and offers valuable insights into where to enhance maritime monitoring and offer hazard warnings. Also, the functionality and utility of the key modules in the proposed approach are examined and demonstrated through comparison with other clustering techniques. Therefore, the developed methodology shows excellent potential to enhance the operators' intelligence surveillance capabilities and facilitate the implementation of risk mitigation measures.

Future explorations and development will focus on the following research directions. First, more vessel motion features (e.g., converging/diverging trend of ship pairs, ship movement behaviour patterns, and ship traffic evolutionary characteristics) could be factored into the traffic partitioning approach to help better reveal the hidden information during traffic propagation. Second, it would be insightful to develop a multi-ship collision risk evaluation model to adapt to the risk comparison between the traffic clusters with the different numbers of ships so that the critical traffic clusters can be more reasonably determined and monitored.

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.

Disclaimer

This paper is the opinion of the authors and does not represent the belief and policies of their employers.

Acknowledgments

This research was funded by the National Natural Science Foundation of China (Grant No. 52031009), and China Scholarship Council (No. 202006950039). This work is also financially supported by an EU H2020 ERC Consolidator Grant program (TRUST Grant No. 864724).

References

[1] Yu H, Fang Z, Murray AT, Peng G. A Direction-Constrained Space-Time Prism-Based Approach for Quantifying Possible Multi-Ship Collision Risks. IEEE Trans Intell Transp Syst 2019;22:131–41.

[2] Fang Z, Yu H, Ke R, Shaw S-L, Peng G. Automatic identification system-based approach for assessing

the near-miss collision risk dynamics of ships in ports. IEEE Trans Intell Transp Syst 2018;20:534–43.

[3] Xin X, Yang Z, Liu K, Zhang J, Wu X. Multi-stage and multi-topology analysis of ship traffic complexity for probabilistic collision detection. Expert Syst Appl 2022:118890.

[4] Liu W, Liu Y, Bucknall R. Filtering based multi-sensor data fusion algorithm for a reliable unmanned surface vehicle navigation. J Mar Eng Technol 2022:1–17.

[5] Liu RW, Guo Y, Lu Y, Chui KT, Gupta BB. Deep Network-Enabled Haze Visibility Enhancement for Visual IoT-Driven Intelligent Transportation Systems. IEEE Trans Ind Informatics 2022.

[6] Liu RW, Guo Y, Nie J, Hu Q, Xiong Z, Yu H, et al. Intelligent Edge-Enabled Efficient Multi-Source Data Fusion for Autonomous Surface Vehicles in Maritime Internet of Things. IEEE Trans Green Commun Netw 2022.

[7] Zhang Y, Sun X, Chen J, Cheng C. Spatial patterns and characteristics of global maritime accidents. Reliab Eng Syst Saf 2021;206:107310.

[8] Montewka J, Manderbacka T, Ruponen P, Tompuri M, Gil M, Hirdaris S. Accident susceptibility index for a passenger ship-a framework and case study. Reliab Eng Syst Saf 2022;218:108145.

[9] Yu Q, Teixeira ÂP, Liu K, Rong H, Soares CG. An integrated dynamic ship risk model based on Bayesian Networks and Evidential Reasoning. Reliab Eng Syst Saf 2021;216:107993.

[10] Zhang M, Zhang D, Fu S, Kujala P, Hirdaris S. A Predictive Analytics Method for Maritime Traffic Flow Complexity Estimation in Inland Waterways. Reliab Eng Syst Saf 2022:108317.

[11] Zhang M, Kujala P, Hirdaris S. A machine learning method for the evaluation of ship grounding risk in real operational conditions. Reliab Eng Syst Saf 2022:108697.

[12] Bye RJ, Aalberg AL. Maritime navigation accidents and risk indicators: An exploratory statistical analysis using AIS data and accident reports. Reliab Eng Syst Saf 2018;176:174–86.

[13] Liu Z, Wu Z, Zheng Z. A novel framework for regional collision risk identification based on AIS data. Appl Ocean Res 2019;89:261–72.

[14] Zhen R, Riveiro M, Jin Y. A novel analytic framework of real-time multi-vessel collision risk assessment for maritime traffic surveillance. Ocean Eng 2017;145:492–501.

[15] Zhen R, Shi Z, Shao Z, Liu J. A novel regional collision risk assessment method considering aggregation density under multi-ship encounter situations. J Navig 2021:1–19.

[16] Chen P, Huang Y, Mou J, van Gelder P. Probabilistic risk analysis for ship-ship collision: state-of-the-art. Saf Sci 2019;117:108–22.

[17] 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. Reliab Eng Syst Saf 2020:106933.

[18] Huang Y, Chen L, Chen P, Negenborn RR, van Gelder P. Ship collision avoidance methods: State-of-the-art. Saf Sci 2020;121:451–73.

[19] Lei P-R. Mining maritime traffic conflict trajectories from a massive AIS data. Knowl Inf Syst 2020;62:259–85.

[20] Zhang W, Goerlandt F, Montewka J, Kujala P. A method for detecting possible near miss ship collisions from AIS data. Ocean Eng 2015;107:60–9.

[21] Szlapczynski R, Szlapczynska J. Review of ship safety domains: Models and applications. Ocean Eng 2017;145C:277–89.

[22] Zhang L, Meng Q. Probabilistic ship domain with applications to ship collision risk assessment. Ocean Eng 2019;186:106130.

[23] Liu K, Yuan Z, Xin X, Zhang J, Wang W. Conflict detection method based on dynamic ship domain model for visualization of collision risk Hot-Spots. Ocean Eng 2021;242:110143.

[24] Liu J, Feng Z, Li Z, Wang M, Wen LR. Dynamic Ship Domain Models for Capacity Analysis of Restricted Water Channels. J Navig 2016;69:481–503.

[25] Weng J, Meng Q, Qu X. Vessel Collision Frequency Estimation in the Singapore Strait. J Navig 2012;65:207–21.

[26] Zhang M, Montewka J, Manderbacka T, Kujala P, Hirdaris S. A big data analytics method for the evaluation of ship-ship collision risk reflecting hydrometeorological conditions. Reliab Eng Syst Saf 2021;213:107674.

[27] Zhao Y, Li W, Shi P. A real-time collision avoidance learning system for Unmanned Surface Vessels. Neurocomputing 2016;182:255–66.

[28] Gil M. A concept of critical safety area applicable for an obstacle-avoidance process for manned and autonomous ships. Reliab Eng Syst Saf 2021;214:107806.

[29] 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 Eng 2020;213:107709.

[30] Li B, Pang F-W. An approach of vessel collision risk assessment based on the D–S evidence theory. Ocean Eng 2013;74:16–21.

[31] Gang L, Wang Y, Sun Y, Zhou L, Zhang M. Estimation of vessel collision risk index based on support vector machine. Adv Mech Eng 2016;8:1687814016671250.

[32] Goerlandt F, Montewka J, Kuzmin V, Kujala P. A risk-informed ship collision alert system: framework and application. Saf Sci 2015;77:182–204.

[33] Hao S, Zhang Y, Cheng S, Liu R, Xing Z. Probabilistic multi-aircraft conflict detection approach for trajectory-based operation. Transp Res Part C Emerg Technol 2018;95:698–712.

[34] Zhang N, Liu H, Ng BF, Low KH. Collision probability between intruding drone and commercial aircraft in airport restricted area based on collision-course trajectory planning. Transp Res Part C Emerg Technol 2020;120:102736.

[35] Cho Y, Han J, Kim J. Efficient COLREG-compliant collision avoidance in multi-ship encounter situations. IEEE Trans Intell Transp Syst 2020;23:1899–910.

[36] Liu J, Zhang J, Yan X, Soares CG. Multi-ship collision avoidance decision-making and coordination mechanism in Mixed Navigation Scenarios. Ocean Eng 2022;257:111666.

[37] Rong H, Teixeira AP, Soares CG. Data mining approach to shipping route characterization and anomaly detection based on AIS data. Ocean Eng 2020;198:106936.

[38] Rong H, Teixeira AP, Soares CG. Ship trajectory uncertainty prediction based on a Gaussian Process model. Ocean Eng 2019;182:499–511.

[39] Zhang L, Meng Q, Fwa TF. Big AIS data based spatial-temporal analyses of ship traffic in Singapore port waters. Transp Res Part E Logist Transp Rev 2019;129:287–304.

[40] Rong H, Teixeira AP, Soares CG. Spatial correlation analysis of near ship collision hotspots with local maritime traffic characteristics. Reliab Eng Syst Saf 2021;209:107463.

[41] Li H, Lam JSL, Yang Z, Liu J, Liu RW, Liang M, et al. Unsupervised hierarchical methodology of maritime traffic pattern extraction for knowledge discovery. Transp Res Part C Emerg Technol 2022;143:103856.

[42] Li H, Liu J, Yang Z, Liu RW, Wu K, Wan Y. Adaptively constrained dynamic time warping for time series classification and clustering. Inf Sci (Ny) 2020;534:97–116.

[43] Liang M, Liu RW, Li S, Xiao Z, Liu X, Lu F. An unsupervised learning method with convolutional auto-encoder for vessel trajectory similarity computation. Ocean Eng 2021;225:108803.

[44] Gil M, Kozioł P, Wróbel K, Montewka J. Know your safety indicator–A determination of merchant vessels Bow Crossing Range based on big data analytics. Reliab Eng Syst Saf 2022;220:108311.

[45] Yu H, Murray AT, Fang Z, Liu J, Peng G, Solgi M, et al. Ship Path Optimization That Accounts for Geographical Traffic Characteristics to Increase Maritime Port Safety. IEEE Trans Intell Transp Syst 2021;23:5765–76.

[46] Svanberg M, Santén V, Hörteborn A, Holm H, Finnsgård C. AIS in maritime research. Mar Policy 2019;106:103520.

[47] Yang D, Wu L, Wang S, Jia H, Li KX. How big data enriches maritime research–a critical review of automatic identification system (AIS) data applications. Transp Rev 2019;39:755–73.

[48] Xiao Z, Fu X, Zhang L, Goh RSM. Traffic pattern mining and forecasting technologies in maritime traffic service networks: A comprehensive survey. IEEE Trans Intell Transp Syst 2019;21:1796–825.

[49] Rong H, Teixeira AP, Soares CG. Maritime traffic probabilistic prediction based on ship motion pattern extraction. Reliab Eng Syst Saf 2022;217:108061.

[50] Arguedas VF, Pallotta G, Vespe M. Maritime traffic networks: From historical positioning data to unsupervised maritime traffic monitoring. IEEE Trans Intell Transp Syst 2017;19:722–32.

[51] Yan Z, Xiao Y, Cheng L, He R, Ruan X, Zhou X, et al. Exploring AIS data for intelligent maritime routes extraction. Appl Ocean Res 2020;101:102271.

[52] Arguedas VF, Pallotta G, Vespe M. Automatic generation of geographical networks for maritime traffic surveillance. 17th Int. Conf. Inf. fusion, IEEE; 2014, p. 1–8.

[53] Pallotta G, Vespe M, Bryan K. Vessel pattern knowledge discovery from AIS data: A framework for anomaly detection and route prediction. Entropy 2013;15:2218–45.

[54] Fiorini M, Capata A, Bloisi DD. AIS data visualization for maritime spatial planning (MSP). Int J e-

Navigation Marit Econ 2016;5:45–60.

[55] Ristic B. Detecting anomalies from a multitarget tracking output. IEEE Trans Aerosp Electron Syst 2014;50:798–803.

[56] Osekowska E, Johnson H, Carlsson B. Grid size optimization for potential field based maritime anomaly detection. Transp Res Procedia 2014;3:720–9.

[57] Tsou M. Discovering knowledge from AIS database for application in VTS. J Navig 2010;63:449–69. https://doi.org/10.1017/S0373463310000135.

[58] Xiao Z, Ponnambalam L, Fu X, Zhang W. Maritime traffic probabilistic forecasting based on vessels' waterway patterns and motion behaviors. IEEE Trans Intell Transp Syst 2017;18:3122–34.

[59] George J, Crassidis JL, Singh T, Fosbury AM. Anomaly Detection using Context-Aided Target Tracking. J Adv Inf Fusion 2011;6:39–56.

[60] Yu Q, Liu K, Teixeira AP, Soares CG. Assessment of the Influence of Offshore Wind Farms on Ship Traffic Flow Based on AIS Data. J Navig 2020;73:131–48.

[61] Xin X, Liu K, Yang X, Yuan Z, Zhang J. A simulation model for ship navigation in the "Xiazhimen" waterway based on statistical analysis of AIS data. Ocean Eng 2019;180:279–89.

[62] Wu X, Mehta AL, Zaloom VA, Craig BN. Analysis of waterway transportation in Southeast Texas waterway based on AIS data. Ocean Eng 2016;121:196–209.

[63] Shelmerdine RL. Teasing out the detail: How our understanding of marine AIS data can better inform industries, developments, and planning. Mar Policy 2015;54:17–25.

[64] Kang L, Meng Q, Liu Q. Fundamental diagram of ship traffic in the Singapore Strait. Ocean Eng 2018;147:340–54.

[65] Greidanus H, Alvarez M, Eriksen T, Gammieri V. Completeness and accuracy of a wide-area maritime

situational picture based on automatic ship reporting systems. J Navig 2016;69:156–68.

[66] Mazzarella F, Alessandrini A, Greidanus H, Alvarez M, Argentieri P, Nappo D, et al. Data fusion for wide-area maritime surveillance. Work. Mov. objects Sea, 2013.

[67] Simsir U, Amasyalı MF, Bal M, Çelebi UB, Ertugrul S. Decision support system for collision avoidance of vessels. Appl Soft Comput 2014;25:369–78.

[68] Qi L, Zheng Z. Trajectory prediction of vessels based on data mining and machine learning. J Digit Inf Manag 2016;14:33–40.

[69] Young BL. Predicting vessel trajectories from AIS data using R. Naval Postgraduate School Monterey United States; 2017.

[70] Perera LP, Oliveira P, Soares CG. Maritime traffic monitoring based on vessel detection, tracking, state estimation, and trajectory prediction. IEEE Trans Intell Transp Syst 2012;13:1188–200.

[71] Mazzarella F, Vespe M, Santamaria C. SAR ship detection and self-reporting data fusion based on traffic knowledge. IEEE Geosci Remote Sens Lett 2015;12:1685–9.

[72] Stergiopoulos G, Valvis E, Mitrodimas D, Lekkas D, Gritzalis D. Analyzing congestion interdependencies of ports and container ship routes in the maritime network infrastructure. IEEE Access 2018;6:63823–32.

[73] Stergiopoulos G, Valvis E, Anagnou-Misyris F, Bozovic N, Gritzalis D. Interdependency analysis of junctions for congestion mitigation in transportation infrastructures. ACM SIGMETRICS Perform Eval Rev 2017;45:119–24.

[74] Lykou G, Dedousis P, Stergiopoulos G, Gritzalis D. Assessing interdependencies and congestion delays in the aviation network. IEEE Access 2020;8:223234–54.

[75] Saeedmanesh M, Geroliminis N. Clustering of heterogeneous networks with directional flows based on

"Snake" similarities. Transp Res Part B Methodol 2016;91:250–69.

[76] Saeedmanesh M, Geroliminis N. Dynamic clustering and propagation of congestion in heterogeneously congested urban traffic networks. Transp Res Procedia 2017;23:962–79.

[77] Ji Y, Geroliminis N. On the spatial partitioning of urban transportation networks. Transp Res Part B Methodol 2012;46:1639–56.

[78] Gu Z, Saberi M. A bi-partitioning approach to congestion pattern recognition in a congested monocentric city. Transp Res Part C Emerg Technol 2019;109:305–20.

[79] Fujii Y, Tanaka K. Traffic Capacity. J Navig 1971;24:543–52. https://doi.org/10.1017/S0373463300022384.

[80] Hu L, Naeem W, Rajabally E, Watson G, Mills T, Bhuiyan Z, et al. A multiobjective optimization approach for COLREGs-compliant path planning of autonomous surface vehicles verified on networked bridge simulators. IEEE Trans Intell Transp Syst 2019;21:1167–79.

[81] Wang H, Song Z, Wen R. Modeling air traffic situation complexity with a dynamic weighted network approach. J Adv Transp 2018;2018.

[82] Xin X, Liu K, Yang Z, Zhang J, Wu X. A probabilistic risk approach for the collision detection of multi-ships under spatiotemporal movement uncertainty. Reliab Eng Syst Saf 2021:107772.

[83] 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. Reliab Eng Syst Saf 2013;118:93–105. https://doi.org/10.1016/j.ress.2013.04.006.

[84] Tu E, Zhang Y, Zhu L, Yang J, Kasabov N. A graph-based semi-supervised k nearest-neighbor method for nonlinear manifold distributed data classification. Inf Sci (Ny) 2016;367:673–88.

[85] Zhao L, Shi G. Maritime anomaly detection using density-based clustering and recurrent neural

network. J Navig 2019;72:894–916.

[86] Jiang H, Li J, Yi S, Wang X, Hu X. A new hybrid method based on partitioning-based DBSCAN and ant clustering. Expert Syst Appl 2011;38:9373–81.

[87] Jahirabadkar S, Kulkarni P. Algorithm to determine ε-distance parameter in density based clustering. Expert Syst Appl 2014;41:2939–46.

[88] Ertöz L, Steinbach M, Kumar V. Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data. Proc. 2003 SIAM Int. Conf. data Min., SIAM; 2003, p. 47–58.

[89] Zhao W, He Q, Ma H, Shi Z. Effective semi-supervised document clustering via active learning with instance-level constraints. Knowl Inf Syst 2012;30:569–87.

[90] De Amorim RC, Hennig C. Recovering the number of clusters in data sets with noise features using feature rescaling factors. Inf Sci (Ny) 2015;324:126–45.

[91] Zhao L, Shi G, Yang J. Ship trajectories pre-processing based on AIS data. J Navig 2018;71:1210–30.

[92] Chai T, Weng J, Xiong DQ. Development of a quantitative risk assessment model for ship collisions in fairways. Saf Sci 2017;91:71–83.

[93] van Westrenen F, Ellerbroek J. The effect of traffic complexity on the development of near misses on the North Sea. IEEE Trans Syst Man, Cybern Syst 2015;47:432–40.

[94] Tan P-N, Steinbach M, Kumar V. Introduction to data mining. Pearson Education India; 2016.