Ocean and Coastal Management 239 (2023) 106622

Contents lists available at ScienceDirect

Ocean and Coastal Management

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

A data mining method to extract traffic network for maritime transport management

Zhao Liu [a,b], Hairuo Gao [a,b], Mingyang Zhang [c,*], Ran Yan [d], Jingxian Liu [a,b]

[a] School of Navigation, Wuhan University of Technology, Wuhan, 430063, China
[b] Hubei Key Laboratory of Inland Shipping Technology, Wuhan, 430063, China
[c] School of Engineering, Department of Mechanical Engineering, Aalto University, Espoo, 20110, Finland
[d] School of Civil and Environmental Engineering, Nanyang Technological University, 50 Nanyang Avenue, Singapore

ARTICLE INFO

Keywords:
Maritime transport management
Maritime traffic network
AIS
Big data analytics
Machine learning

ABSTRACT

Maritime traffic network is essential for navigation efficiency and safety of the maritime transport system. This study proposes a framework for extracting maritime traffic network based on Automatic Identification System (AIS) data. The framework consists of maritime traffic pattern recognition, semantic routes extraction, route decomposition, and network generation. Firstly, a data-driven method is introduced to recognize ship behavior patterns and extends the single ship behaviors to regional characteristics to determine the departure-arrival areas. Then, based on the different combination of departure-arrival areas, the ship trajectories are classified to traffic groups. Subsequently, the grid-system is used to rasterize each traffic group, which realizes the fusion of trajectory data and geographic location information. Finally, to obtain the main routes and navigation channels, the extraction method is introduced by establishing the cumulative grid importance function. The main routes, together with the navigation channels, compose the maritime traffic network. The method is applied to AIS data in the Beibu Gulf, and the results show that the traffic network contains 12 stop areas, 4 entry/exit locations, 13 main routes as well as their corresponding navigation channels. It is therefore concluded that the proposed method helps (1) provide a theoretical framework to obtain and analyze the maritime traffic network and (2) enrich navigation channel identification methods for maritime transport management.

1. Introduction

With the development of the economy, maritime traffic is becoming heavier, and the huge traffic volume makes traffic management more difficult. The great mobility of ships has allowed more goods to be transported to remote ports (Xu et al., 2021a,b,c; Aregall et al., 2018; Lin and Chang, 2018), and shipping is also the most environmentally friendly mode of transportation if the goods' value is considered (Zhang et al., 2021). So, expanding ports and shipping are key driving forces in growing the world economy (Andersson and Ivehammar, 2017; Xu et al., 2021a,b,c; Wang and Meng, 2012). However, there also exist potential dangers when sea transportation booms. For example, the total losses of marine accidents make them the most serious accidents in the world (Chen et al., 2019), especially for grounding and fire/explosion accidents (Fu et al., 2022a,b). Nuclear leakage of the nuclear-powered ships may occur to threat the maritime safety (Fu et al., 2022a,b). And the ecological environment pollution is becoming more and more serious, which raises the concern for effective governance of shipping pollution (Chen et al., 2022a,b,c; Xu et al., 2021a,b,c). As is known to all, the increase of shipping activity will give a rise to the carbon emissions (Xu et al., 2023). However, such damages can be controlled through specific research and technologies (Zhang et al., 2023). Nowadays, the widely used shipboard Automatic Identification System (AIS) provides a large amount of AIS data, which can be used for traffic flow analysis. And such analysis can reflect the traffic information and develop many applications for transport optimization and efficiency/safety management (Fan et al., 2010; Wei et al., 2020).

Generally, with fuel efficiency and safety issues being considered, vessels often navigate under international regulations. For example, liner shipping relies on route planning (Chen et al., 2021), and appropriate routes can save time and energy (Wan et al., 2021). In the meanwhile, the distribution of routes can reflect the navigation rules, and the shipping routes can be identified from historical AIS trajectories. Such traffic data provide effective information for decision-makers, and

* Corresponding author. Otakaari 4, 02150, Koneteknikka 1, Espoo, Finland.
E-mail address: mingyang.0.zhang@aalto.fi (M. Zhang).

https://doi.org/10.1016/j.ocecoaman.2023.106622
Received 12 February 2023; Received in revised form 9 April 2023; Accepted 20 April 2023
Available online 3 May 2023

they can help optimize navigation decisions. In addition, analyzing these traffic data can reveal ships' historical navigation patterns, which can be used in anomaly detection, route planning, etc. As a result, extracting shipping network patterns is of great necessity for maritime traffic planning and management.

However, the extraction methods of network have yet to be studied. While the traditional manual measurement methods and image recognition methods all have the drawbacks of high cost and slow update speed, AIS data-driven methods are more suitable with the characteristics of low cost and real-time. As an ideal information source for ship behavior studies, the data-driven methods are mainly divided into three aspects, respectively statistics-based, grid-based, and vector-based methods. In general, the statistic-based methods are easy to operate, but the model will be difficult to construct when processing large-scale data. Comparatively, the vector-based methods can effectively relieve the computational burden by identifying the waypoints in ship trajectories. But the methods also do not perform well in high ship density areas, as the single parameter set from experience is difficult to deal with the area having uneven ship density. Moreover, the problem will be trickier when the ship behaviors are complex, especially in uncontrolled waters (Liu et al., 2023; Zhang et al., 2021a, 2021b).

Based on this analysis, we develop a grid-system method based on ship behavior patterns to extract the maritime traffic network. On one hand, relying on the recognition of maritime traffic patterns, the regional characteristics can be identified, such as the departure-arrival areas. And with the ship trajectories classified by the different combination of departure-arrival areas, we apply the method to each traffic group separately. In this method, the shortcoming of many density-based clustering approaches, that is, ignoring local information of complex trajectories, can be overcome. On the other hand, the grid system is characterized in the area attribute of grid and storage ability of AIS information, which promotes the identification of navigation channels. In this method, we propose the grid importance function considering the grid value as well as the distance between the grid and the main route. It can effectively deal with the density differences among traffic groups to identify the important area of the group. To verify the effectiveness of the method, we apply it to the historical AIS data covering 6 months of 2019 in the Beibu Gulf.

This paper is organized as follows: in Section 2, the related research of maritime network extraction using historical navigation data is reviewed; in Section 3, the maritime traffic network extraction method is introduced in detail, including maritime traffic pattern extraction, semantic route extraction, route decomposition and traffic network generation; in Section 4, the results of the experiment that uses historical AIS data to prove the effectiveness of the method are presented; in Section 5, the conclusions are drawn, and the future work is described.

2. Literature review

Unlike other tools of transportation, ships navigate in wide waters, which means maritime traffic is characterized by the spatial freedom. However, a high degree of freedom will increase uncertainty, which may threaten ships' navigation safety. Therefore, the need to ensure maritime traffic safety has stimulated the development of research on maritime traffic control (Wang et al., 2014) and safety management (Rong et al., 2019). The complexity of maritime traffic flow is considered as one of the main influencing factors on maritime safety (Zhang et al., 2022). To effectively deal with the complexity, grasping traffic information is of great importance for both ship management and navigational safety. Wang et al. (2022) proposed a framework to analyze the characteristics of ship traffic flow and found that ships have special behavior patterns. In general, the crew often rely on navigation-related materials (i.e., sailing directions, guide to port entry, notice to mariners, admiralty list of lights, and fog signals) and experience from previous similar routes (Zhang et al., 2018). Besides, with the consideration of fuel efficiency and safety issues, ships generally follow fixed routes. Consequently, a maritime shipping network is essential for ship routing, scheduling, and flexibility analysis of the shipping system (Liu et al., 2023). The methods for constructing the maritime traffic network can be classified as statistics-based, grid-based, and vector-based.

To be more specific, statistics-based methods perform a statistical analysis to obtain traffic flow characteristics (i.e., traffic flow, traffic speed, traffic density, and traffic lane width). Early research is centered on specific waters. Silveira et al. (2013) analyzed the Portuguese coast statistically. And based on existing traffic separation schemes (TSS), the results can help route planning studies. Additionally, Wu et al. (2016) conducted a statistical analysis of the navigation behaviors of ships in hot regions in the Sabine-Neches Waterway. The study can provide accurate collision risk warnings by analyzing behavior patterns of ships entering or leaving the port. However, these methods have the limitation of lack in systematic study. To fill the gap, Kang et al. (2018) first developed a research framework to analyze big data and explored the speed-density relationship of maritime traffic. What's more, Lee et al. (2020) applied kernel density estimation (KDE) to generate the centerline of a maritime route.

The grid-based methods construct a grid reference system to discretize AIS data to realize the fusion of multiple features. By means of using grids to split the area, the ship position is applied to the grid (Wang et al., 2019). Bomberger et al. (2007) placed a uniform square grid over the area of interest surrounding the port of Miami so as to discretize vessel locations. Dobrkovic et al. (2015) subdivided the area into cells, and tracked its pheromone density. The grid-based methods can adequately solve the problem of DBSCAN (Density-Based Spatial Clustering of Applications with Noise) when deal with different traffic densities, and it contributes to the research of main route identification. Vettor and Soares (2015) proposed a grid-system method of tracking the areas where the reports were relatively dense between the previously defined junction points to distinguish the routes. However, the method is rather empirical, as the grid-based methods are effective for small-area surveillance applications but have the problem of heavy computation when the scale increases. To solve the problem, Xiao et al. (2017) proposed a novel information-assisted methodology using a grid-based DBSCAN algorithm to extract the waterway patterns. It used the kernel density estimation method for the first time to model the ship motion behaviors quantitatively.

Vector-based methods build maritime traffic networks by extracting network nodes (waypoints) and edges (trajectory segments) to characterize ship navigation routes. Specifically, routes are viewed as a set of straight lanes connecting waypoints. Kaluza et al. (2010) constructed a navigation network, viewing the ports as nodes connected by ship journeys. They found that the differences in the movement patterns of different ship types are an important characteristic of the network. The TREAD, that is, Traffic Route Extraction and Anomaly Detection, was developed by Pallotta et al. (2013). In this methodology, the stream of AIS messages was processed to show maritime motion patterns, which leads to the discovery of waypoints. Arguedas et al. (2017) established two hierarchical layers on the traffic network: the external and internal layers, which can reflect more real-world situations to represent the maritime traffic in the monitored area more accurately. Sheng and Yin (2018) comprehensively considered the geospatial information as well as the contextual features of ship trajectories and proposed a method that can automatically classify different shipping routes without prior information. Yan et al. (2020) transformed the rich ship-position information into a ship trip semantic object (STSO) to define ship behavior patterns. To process the big data, relevant clustering algorithms were introduced. Wen et al. (2020) applied the DBSCAN algorithm to recognize the key regions and connected them through cluster similarity measuring, and then generated the routes. Murray and Perera (2022) applied machine learning techniques to extrapolate commonalities in relevant trajectory segments. Huang et al. (2023) clustered the trajectories through the multi-dimensional density-based spatial clustering applications with the noise (MD-DBSCAN) algorithm.

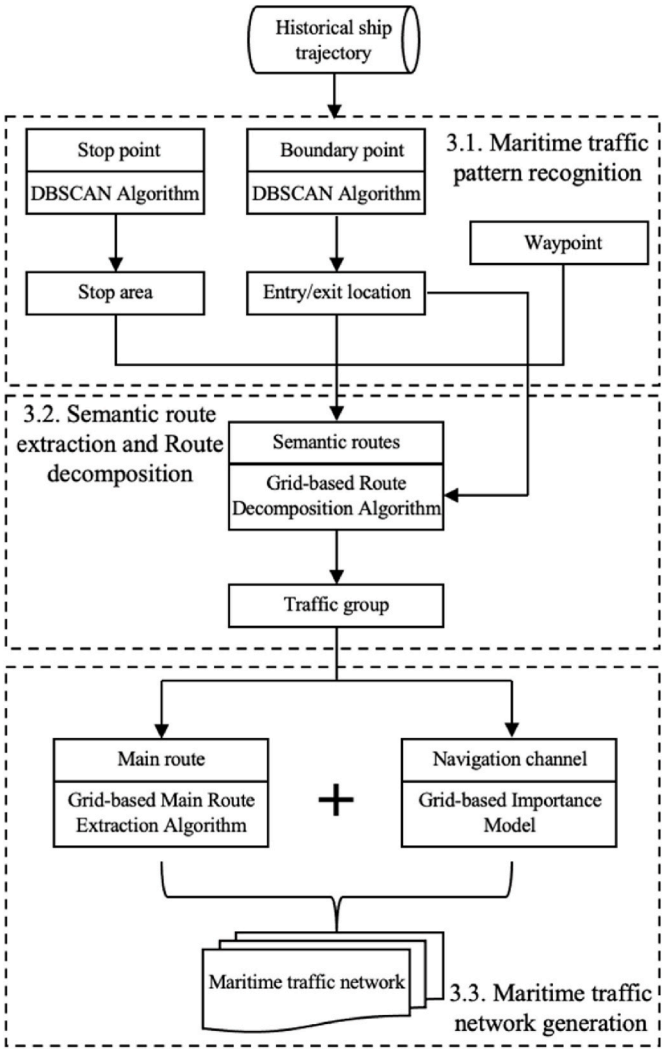

Fig. 1. The proposed framework for maritime traffic network generation.

Overall, the statistic-based methods are easy to operate, but the model will be difficult to construct when processing large-scale data. Comparatively, the vector-based methods can effectively relieve the computational burden by identifying the waypoints in ship trajectories. However, the methods also do not perform well in high ship density areas, as the single parameter set from experience is difficult to deal with the area having uneven ship density. Moreover, the problem will be trickier when the ship behaviors are complex, especially in uncontrolled waters. Therefore, in this paper we adopt a grid-system extraction method based on ship behavior patterns. Applying the method to each traffic group that matches the different ship behavior pattern separately can overcome the shortcoming of most density-based clustering algorithms. Besides, the grid system is characterized in the area attribute of grid and storage of AIS information, which promotes the identification of navigation channels.

3. Methodology

Ships navigate at sea, where they have great spatial freedom. The flexibility of ship navigation increases the difficulty of ship motion investigations. However, the rational distribution of fleet resources is an important way to improve the productivity of the maritime transport industry (Zhang et al., 2021; Chen et al., 2021, 2022). So, extracting the traffic network patterns is of great importance for ship management and safety.

Fig. 1 shows the framework for maritime traffic network generation proposed. And the framework includes three stages:

- Stage I: Maritime traffic pattern recognition.

The identification of ship behavior patterns can give trajectories well-defined semantic meaning, with the structure of 'stop point-waypoint-boundary point'. Besides, by extending the single ship behaviors to regional characteristics with the DBSCAN algorithm, the departure-arrival areas, which are composed of stop areas and entry/exit locations, can be identified. And they provide the support for subsequent research.

- Stage II: Semantic route extraction and route decomposition.

With the application of departure-arrival areas, the ship trajectories can be classified to obtain the semantic routes. And in order to separate the mixed traffic groups caused by the simplification of entry/exit locations, the route decomposition method is proposed. Through this method, the traffic groups representing different ship behavior patterns are obtained.

- Stage III: Maritime traffic network generation.

According to the knowledge that the feature line of the traffic group has relatively high density, the main route of each traffic group is extracted with the grid system. Besides, based on the experience that traffic flow often follows the normal distribution, the cumulative grid importance function is established to identify the navigation channels around the main routes. The main routes and navigation channels constitute the maritime traffic network. The results are evaluated by comparing the official navigation channels.

3.1. Maritime traffic pattern recognition

Although the restrictions on maritime transport are not as tight as those on road traffic, ship navigation also follows certain rules. For example, most ships navigate along channels, and ships of the same group have similar navigational characteristics. And these navigational characteristics can be discovered from a large amount of AIS data. There are generally three types of ship behavior patterns: static, normal navigation, and maneuvering (Chen et al., 2020). As ships rarely maneuver during navigation, compressing AIS data on the basis of the three behavior patterns can be well performed. It can preserve the original ship navigation features effectively. Specifically, the data-driven approach recognizes stop points matching the static pattern, and waypoints matching the maneuvering pattern, forming the ship trip semantic object (STSO). The method for recognizing stop points and waypoints is described in the following section.

3.1.1. Departure-arrival area recognition

As one typical maritime traffic pattern, the ship's static pattern reflects the stop points in the ship's trajectory. Generally, stop points are aggregated, so it is easy for us to recognize a ship's static status to identify its stop points. Rather than official areas (i.e., anchorages and ports), stop points are applied by data-driven approaches, resulting in higher accuracy and a higher matching degree. Additionally, as the area being studied has a boundary, the endpoints of a ship's trajectory where the ship enters or exits the area should also be considered. And such entering or exiting locations are described as entry/exit locations. The steps to recognize a ship's departure-arrival area are described below.

Step 1 stop point recognition

Stop areas of a ship include ports and anchorages. There are generally two types of stop areas, specifically, the mooring and anchoring

areas. When a ship is moored, it is relatively stable. In contrast, it will shift randomly when it is anchored. Meanwhile, due to ocean currents and positioning accuracy, ships generally do not remain stationary when they stop. On the basis of the ship stop-behavior just mentioned, it can be summarized that if the period during which a ship maintains a relatively static status around a location point exceeds a certain threshold, this point can be viewed as a possible stop point.

Based on the analysis above, we set a distinguishing principle to identify the possible stop points. Through voyage identification and data preprocessing (i.e., trajectory partition, data cleaning and data interpolation) (Liu et al., 2020), the AIS data are processed into a set of trajectory lines of different voyages of different vessels. The trajectory set is expressed as $tra_{all} = \{tra_1, tra_2, \cdots, tra_n\}$, and each trajectory consists of multiple points, which are expressed as $tra_i = \{tp_1, tp_2, \cdots, tp_n\}$. Start from the first point in the time series, and determine whether the following conditions are satisfied at adjacent moments tp_j and tp_{j+1}:

$$v_{tp_j} < v_T, t_T \le t_{tp_{j+1}} - t_{tp_j} < t_T{}', dist_{tp_j,tp_{j+1}} < dist_T \tag{1}$$

where v_{tp_j} represents the instantaneous velocity at the trajectory point tp_j in trajectory tra_i, t_{tp_j} represents the timestamp at point tp_j, $dist_{tp_j,tp_{j+1}}$ represents the distance between tp_j and tp_{j+1}, and v_T, t_T, $t_T{}'$ and $dist_T$ represent the speed threshold, time threshold and distance threshold, respectively. With reference to existing literature and real practices, the thresholds set in the case study are as follows: $v_T = 0.5kots, t_T = 300s$, $t_T{}' = 1800s, dist_T = 1km$ (Yan et al., 2020). If conditions (1) are satisfied at point tp_j, it can be viewed as a candidate stop point and then added to the set of candidate stop points, $ps_{all} = \{ps_1, ps_2, ps_3, \cdots, ps_n\}$. If conditions (1) are not met, then keep checking the following points in tra_i until a new point satisfies the conditions. Accordingly, single out all candidates stop points to complete the recognition process.

Step 2 stop area recognition

In the previous step, many candidates of stop points are singled out. However, the thresholds cannot guarantee the absolute accuracy of these selected stop points, which means these points may be misjudged stop points. To solve the problem, stop area recognition is introduced. Stop areas are defined as the ports and anchorages important for ship navigation. Many ships tend to stop in these areas. And from a systematic perspective, many stop points cluster in the stop areas, where the ships stopping there all exhibit the ship static pattern. Therefore, we introduce a clustering algorithm to recognize the stop areas.

The DBSCAN algorithm is a density-based clustering algorithm that can effectively remove noise data. It is efficient in mining the data of high-density areas. The DBSCAN algorithm is applied to the set of candidates of stop points to obtain the stop areas within the water being studied through a clustering method. The obtained stop areas can be compared with the planned anchoring areas on the electronic charts to evaluate the accuracy of the clustering parameters selected.

Step 3 entry/exit location recognition

In addition, as the area being studied has a boundary, the endpoints of a ship's trajectory exist not only at the ports or anchorages but also at the boundary of the area. Thus, the locations where the ship enters and exits the area should also be considered. These locations exhibit the navigation trend, which has a feature of aggregation. So, entry/exit locations can be obtained from the clustered navigation trajectory points at the area boundary.

With the consideration that the entry/exit locations are influenced by the situation of water being studied, a clustering method is used to identify the high-density locations at the area boundary based on historical AIS data. Firstly, the set of boundary points of the water being studied is obtained by identifying the intersection points of trajectories and the water boundary. The set is expressed as $pb_{all} = \{pb_1, pb_2, \cdots pb_n\}$. Secondly, the DBSCAN algorithm is applied to cluster the set of intersection points pb_{all} to acquire the set of clusters containing boundary points, which is expressed as $pb_{all}^{clu} = \{pb_1^{clu}, pb_2^{clu}, \cdots, pb_N^{clu}\}$. Then, we single out the longest side parallel to the direction of the boundary in each cluster, and obtain two end points of the side, which are expressed as $[P_1(pb_i^{clu}), P_2(pb_i^{clu})]$. It is described as the entry/exit location loc_i.

Furthermore, identify the longest side of all clusters to obtain the entry/exit locations of the study water, which are expressed as loc_{all}. It presents the ship's navigation tendency outside the study water.

3.1.2. Waypoint recognition

Ships are large and generally not flexible in maneuvering. So, they always need a relatively long time to maneuver. Large ships, in particular, do not frequently maneuver during their voyages and usually take the optimal routes to minimize total fuel consumption. Most of the time, ships change their courses slowly in the turning areas. Thus, the ship trajectory can be simplified as a line segment formed by waypoints.

A large amount of redundant data is effectively removed from the compressed data, which greatly improves the operational efficiency of the algorithm. Besides, the compressed data can retain the shape and time information of the original trajectory. The method of waypoint recognition is described below.

The idea of a sliding window is introduced to identify a turning section. Starting from the first point of the trajectory tra_i, we slide the window in sequence to calculate each point's vectorial angle. We assume the sliding window is composed of a sequence of trajectory points with the number p. And the number is determined according to the distribution of trajectory points within the study area. Then the vectorial angle of two vectors in the sequence is calculated to represent the steering extent of this sequence. One vector is made up of the first point and the middle point in the sequence, and the other vector is made up of the middle point and the last point in the sequence. The vectorial angle calculation is defined as follows:

$$\cos\theta_j = \frac{\overrightarrow{tp_k tp_{k-a}} \cdot \overrightarrow{tp_{k+b} tp_k}}{\left|\overrightarrow{tp_k tp_{k-a}}\right| \cdot \left|\overrightarrow{tp_{k+b} tp_k}\right|}, a + b + 1 = p \tag{2}$$

where $\cos\theta_j$ represents the vector cosine angle of the jth trajectory sequence in the trajectory tra_i, $\overrightarrow{tp_k tp_{k-a}}$ represents the vector from point tp_k to point tp_{k-a}, and $\overrightarrow{tp_{k+b} tp_k}$ represents the vector from point tp_{k+b} to point tp_k. The point tp_{k-a} and tp_{k+b} are relatively the first point and last point of the sequence, and the point tp_k is the middle point. Then we can see whether the value calculated above can satisfy the following condition:

$$\left|\cos\theta_j\right| \le v_T \tag{3}$$

where v_T represents the threshold of vector cosine angle, which is set by experiment and experience. If the above condition is satisfied, the trajectory sequence is viewed as a turning section. In particular, the trajectory sequence, $tra_m^{turn} = \{tp_{k-a}, tp_{k-a+1}, \cdots, tp_{k+b}\}$, has a bending feature, which means it can be regarded as a turning section. However, we need the key turning point in the turning section rather than a trajectory section. Thus, we introduce the secondary recognition method. The same idea of the sliding window is applied, but we reduce the length of the window to 3 trajectory points. Then, we can calculate each vector cosine angle of the secondary trajectory section in the turning section tra_m^{turn}, with the following formula:

$$\cos\theta_n^{turn} = \frac{\overrightarrow{tp_{k-a+r} tp_{k-a-1+r}} \cdot \overrightarrow{tp_{k-a+1+r} tp_{k-a+r}}}{\left|\overrightarrow{tp_{k-a+r} tp_{k-a-1+r}}\right| \cdot \left|\overrightarrow{tp_{k-a+1+r} tp_{k-a+r}}\right|} \tag{4}$$

where $\cos\theta_n^{turn}$ represents the vector cosine angle of the nth secondary trajectory section within the identified turning section tra_m^{turn}, and r is an integer varying from 1 to $p-2$. As the turning section identified is

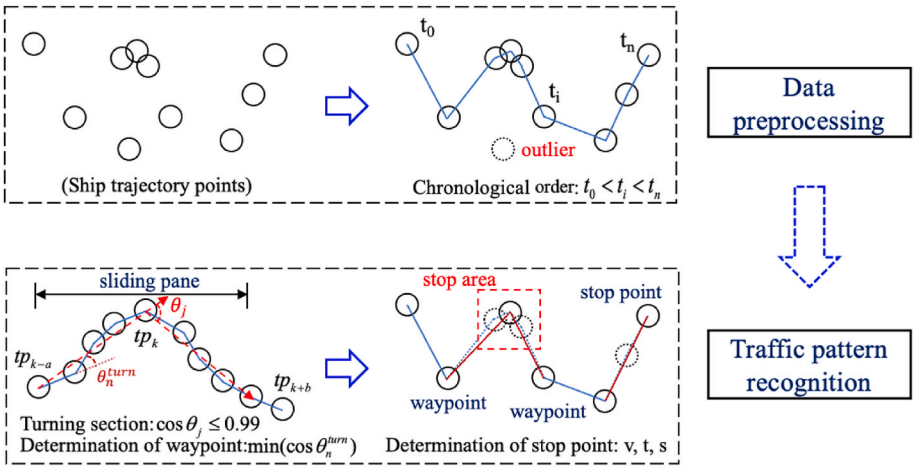

Fig. 2. Maritime traffic pattern recognition.

composed of points with the number p, the number of secondary trajectory sections obtained by cutting it into 3 points in turn should be $p-2$. Thus, a turning section identified has $p-2$ angle values, with $cos\ \theta_n^{turn} = \{cos\ \theta_n^{turn}(1), cos\ \theta_n^{turn}(2), \cdots, cos\ \theta_n^{turn}(p-2)\}$. Then we take the smallest value $\min(\cos\theta_n^{turn})$, which indicates the largest turning amplitude in the trajectory turning section tra_m^{turn}. And the middle point of the chosen secondary section, tp_{k-a+r}, is the waypoint, which is denoted by pw_t^i. After all trajectories are checked, each trajectory can be simplified as a set of ordered waypoints, tra_i^{cpr}. It should be noted that the subsequent points in the section should be skipped if the section has been identified as a turning section, which means the next trajectory section to be checked consists of points with the number of p from tp_{k+b+1} to tp_{k+b+p}.

In addition, if a trajectory section does not satisfy condition (3), the section starting from the next point in tra_i should be checked, which means the section to be checked consists of the points from tp_{k-a+1} to tp_{k+b+1}. Check the sections until a new one satisfies the condition, and then make the corresponding calculation. After all trajectory points in the area being studied are checked through the waypoint recognition method as described, a set of waypoints corresponding to all trajectories can be obtained, which is $tra_{all}^{cpr} = \{tra_1^{cpr}, tra_2^{cpr}, \cdots, tra_N^{cpr}\}$.

Fig. 2 shows the processes of stop point and waypoint recognition. In the stage of data preprocessing, this study filters the outliers from three aspects: duplicate data, data with speed exceeding normal range, and data on trajectory position offset. The position offset is judged by comparing the average speed between trajectory points with the maximum speed, which is obtained through the overall distribution of the experimental data. And the maximum speed is also used to filter the point with data with abnormal speed. After the data preprocessing process, the stop points and waypoint recognition method above is applied to the data.

3.2. Semantic route extraction and route decomposition

Ships always follow certain navigation rules, meaning that ships taking the same route have similar behavior and spatial distribution patterns during their navigation. The previous analysis shows that the ship's behavior patterns are divided into three types, which are static, normal navigation, and maneuvering. In the preceding section, we have identified the stop areas corresponding to the ship's static pattern and the entry/exist locations at the boundary of the water, which form the departure-arrival areas. Furthermore, the compressed data are obtained by recognizing waypoints corresponding to the maneuvering pattern. Thus, according to the navigation rules, we can classify ship trajectories based on the departure-arrival areas recognized. Specifically, the trajectories that arrive at the same departure-arrival areas should belong to one cluster. It should be noted that the stop areas identified are polygons in size, and the identified anchorages cover a large area in the water being studied. So in order to ease the research, only anchorages in the stop areas are analyzed in the following section. With the application of departure-arrival areas, the acquirement of traffic groups can be obtained more specifically and comprehensively.

3.2.1. Semantic route extraction

In the water, a ship heading to a port enters its boundary at an entry location, then sails at low speed within the stop area before arriving at the port.

And the ship will reverse the process when it departs from the port. Therefore, vessels passing through a same combination of entry/exit locations and stop areas should have similar behavior pattern and spatial distribution, which are called the semantic route. Specifically, the cluster that passes through both stop area $area_i$ and entry/exit location loc_i is defined as Tra_i^{group}. The set of all semantic routes is denoted as $Tra_i^{group} = \{Tra_1^{group}, Tra_2^{group}, \cdots, Tra_n^{group}\}$, which corresponds to a certain combination of the departure-arrival areas.

In addition, the stop areas identified are polygons in size, so the sematic routes of some trajectory groups may overlap. Among these trajectory groups, the groups having broader ranges or more trajectories should be retained while the remaining groups should be eliminated from the set of semantic routes. Then, a set of extracted semantic routes retaining the main features of study objects can be obtained.

3.2.2. Route decomposition

As mentioned above, the semantic routes are obtained according to the different combination of departure-arrival areas, which are composed of the entry/exit locations and stop areas. However, based on the idea that the intersection points at boundaries are covered with the locations as few as possible, the results of entry/exit locations are simplified by adjusting parameters in the proposed method in Section 3.1.1 (see Step 3). Therefore, the semantic route extraction process is also simplified. And this may lead to a mix of semantic routes of trajectory groups with different spatial distributions. To address this problem, we propose a grid-based semantic route decomposition algorithm to separate the mixed semantic routes.

Considering the spatial and temporal continuity of the traffic flow, we introduce a grid-system method. It can convert vector data into matrix data, utilizing the continuity characteristic to differentiate the data effectively. What's more, the grid-based method can remove noise data and greatly reduce the order of operation magnitude. Each grid stores the number of trajectories entering or leaving the grid, and then we can differentiate the trajectory groups based on the continuity of the

grids.

Based on the continuity of traffic flow, we can recognize whether the semantic route needs to be decomposed and the splitting location. First, we iterate through each row to acquire the discontinuity location within the trajectory group. Specifically, according to the continuous feature, the grids with values should distribute continuously vertically and laterally in the grid system. Then, we determine whether the group needs to be separated by comparing the vertical discontinuous length with the threshold. If there exist enough empty grids, we can suppose the trajectory group has a gap, and it needs to be decomposed. Based on the above analysis, a grid-based semantic route decomposition algorithm is proposed. Firstly, we perform the de-noising procedure with the grid-based model. Specifically, the values of grids with few trajectories are set to zero to retain the more important data. Then, we identify the separation sections. Finally, we extract the continuous grids to form the interval where the separation sections of the trajectory group are located. Algorithm 1 shows the pseudocode for the grid-based semantic route decomposition algorithm.

Algorithm 1. Grid-based semantic route decomposition

Input: Anchorages identified $area_{all} = \{area_1, area_2, \cdots, area_I\}$, entry/exit locations $loc_{all} = \{loc_1, loc_2, \cdots, loc_J\}$ and compressed trajectory data $tra_{all}^{cpr} = \{tra_1^{cpr}, tra_2^{cpr}, \cdots, tra_N^{cpr}\}$
Output: Grid-based traffic groups $mat_{all} = \{mat_1, mat_2, \cdots, mat_m\}$
Process:
/*Annotations for line 2-9: Semantic route extraction algorithm*/
For $i = 1,2, \cdots, I$:
  For $j = 1,2, \cdots, J$:
    For each ship trajectory tra_k^{cpr} in tra_{all}^{cpr} do:
    If tra_k^{cpr} passes through the anchorage $area_i$ and the entry/exit location loc_j at the same time:
      Add tra_k^{cpr} to the semantic route group corresponding to the combination of $area_i$ and loc_j
    End if
    End for
  End for
End for
/*Annotations for line 12-24: Grid-based route decomposition algorithm*/
The extracted semantic routes are modeled by grids to form matrixes, $mat_{all} = \{mat_1, mat_2, \cdots, mat_n\}$
For $r = 1,2, \cdots, n$:
  Count the frequency of all grid values to obtain the frequency distribution line graph. The graph reveals the trend of monotonically decreasing, and identify the turning point as threshold
  Locate the first row and last row of grids with value in mat_r, respectively recorded as t_1, t_2
  For $t = t_1, \cdots, t_2$:
    Locate the minimum index and maximum index of columns with value in row t within mat_r, respectively stored in s_1, s_2
  End for
  Calculate the mean value of s_1, s_2, and expand or reduce the value according to the migretion direction, respectively recorded as $\overline{s_1}, \overline{s_2}$
  Then identify the vertical interval where there exists at least 5 consecutive values are less than $\overline{s_1}$ in s_1, and the index of endpoints are respectively m_1, m_2, that is the longitudinal separation interval of s_1 is marked as $sepa = [[m_1, m_2], \cdots]$. The same goes for s_2
  For $m \in sepa$:
    Identify the horizontal continuous segments with grid values, and determine the left and right index, respectively recorded as o', o'', that is the separation section in mat_r is obtained, $sepa_{left} = \{[m_1, o'], [o'', m_2], \cdots\}$. The same goes for s_2
  End for
  Get the position of separation section of mat_r, respectively recorded as $sepa_{left}$ and $sepa_{right}$, which mark the coverage area of new route after route decomposition, recorded as mat_r', mat_r'', $\cdots$, and update mat_r
End for
Output $mat_{all} = \{mat_1, mat_2, \cdots, mat_m\}$.

3.3. Maritime traffic network generation

With a large amount of historical AIS data mined, ship behavior patterns are recognized. Then, ship trajectories can be classified according to different behavior patterns. And according to the knowledge that the traffic flow often follows the normal distribution, we can extract the main routes of traffic groups, as well as identify their navigation channels with the grid system. On one hand, the grid system provides a good environment for locating the area of navigation channels. On the other hand, the proposed grid importance function considering the grid value as well as the distance between the grid and the main route can effectively deal with the density differences among traffic groups, to identify the important area of the group. The routes and channels constitute the maritime traffic network. And the traffic network is an important part of the decision support system, especially in the high-traffic-density areas.

3.3.1. Main route extraction

The main routes are the primary lines of the trajectory clusters in the water, with a feature of relatively high traffic density. Gridded clusters of navigational trajectories in different departure-arrival areas are obtained from the above grid-based semantic route decomposition algorithm, with each grid storing the number of trajectories passing through the area grid.

In general, the traffic flow along the main route follows a normal distribution, which means its probability distribution curve has a mean value with a variation statistically. Therefore, it can be assumed that the cluster's centre line basically reflects where the main route is located. However, the trajectory lines do not follow a standard normal distribution. So, there will be some errors if the centre line of trajectory lines is viewed as the main route directly.

Based on the above analysis, we propose a grid-based main route extraction algorithm to locate the main route of each trajectory cluster. For the trajectory clusters whose entry/exit locations are located at the boundary of the water being studied, their main routes should have the highest traffic density at the boundary. Therefore, it is possible to use a boundary traffic density threshold to locate the main route. If the number stored in a grid is less than a specific threshold value, the grid value is set as 0, otherwise, its value is set as the trajectory number that is stored. With the increase in the threshold value, the number of grids at the boundary with a value other than zero gradually decreases until all their values become zero. Then, the threshold searching process is completed.

With this threshold method, the main route of a trajectory cluster can be singled out. With the above method, it can avoid errors and minimize the interference of abnormal trajectories in precisely locating the main routes passing through the boundary of the water. Algorithm 2 shows the pseudocode for the grid-based main route extraction algorithm.

Algorithm 2. Grid-based main route extraction algorithm

Input: Grid-based traffic groups $mat_{all} = \{mat_1, mat_2, \cdots, mat_m\}$
Output: Grid position of main routes $mat_{all}^{main} = \{mat_1^{main}, mat_2^{main}, \cdots, mat_m^{main}\}$ and longitude and latitude position $tra_{all}^{main} = \{tra_1^{main}, tra_2^{main}, \cdots, tra_m^{main}\}$
Process:
/*Annotations for line 2-15: Critical value searching algorithm*/
For $r = 1,2, \cdots, m$:
  Extract all grid values in mat_r and arrange them from large to small, forming the matrix mat_value_r
  For $\epsilon = 0.01, 0.02, 0.03, \cdots 0.5$:
    For $t = t_1, \cdots, t_2$:
      For $s = s_1, \cdots, s_2$:
        If $mat_r[t,s] \le mat\_value_r[len(mat_{value_r}) * \epsilon]$:
          set $mat_r[t,s]$ to zero // set $mat_r[t,s]$ to 0
        End if
      End for
    End for
    If $mat_r[seleted\ items] = 0$: // For north-south routes, $seleted\ items$ is the boundary row of the matrix; For east-west routes, $seleted\ items$ is the boundary column
      $\varepsilon = \epsilon$
      $mat_r^* = mat_r$
  End for
/*Annotations for line 17-21: Main route extraction algorithm*/
For $r = 1,2, \cdots, m$:
  Count the number of grids with value of the new matrix whose designated grid values adjust to 0, and obtain the list mat_num_r
  Find the position where the rate increases sharply, recorded as mat_node_r, which is considered as the key point of the main route (the boundary between port channnel and open water)
  Get the first row, key row and last row where grid of value exists in mat_r^*, and then extract the middle point of the three lines, connecting sequentially to form the distribution of main route, recorded as tra_r^{main}
End for
Output $mat_{all}^{main} = \{mat_1^{main}, mat_2^{main}, \cdots, mat_m^{main}\}$ and $tra_{all}^{main} = \{tra_1^{main}, tra_2^{main}, \cdots, tra_m^{main}\}$.

3.3.2. Identification of navigation channels

According to the statistical theory, the traffic flow follows a normal distribution. So according to the symmetry and uniform variation characteristics of a normal distribution, the distribution curve of each traffic group presents a pattern of high in the middle and low on both sides. All cluster lines show this distribution pattern, so the peak lines in the middle of their distribution curves can be viewed as their main

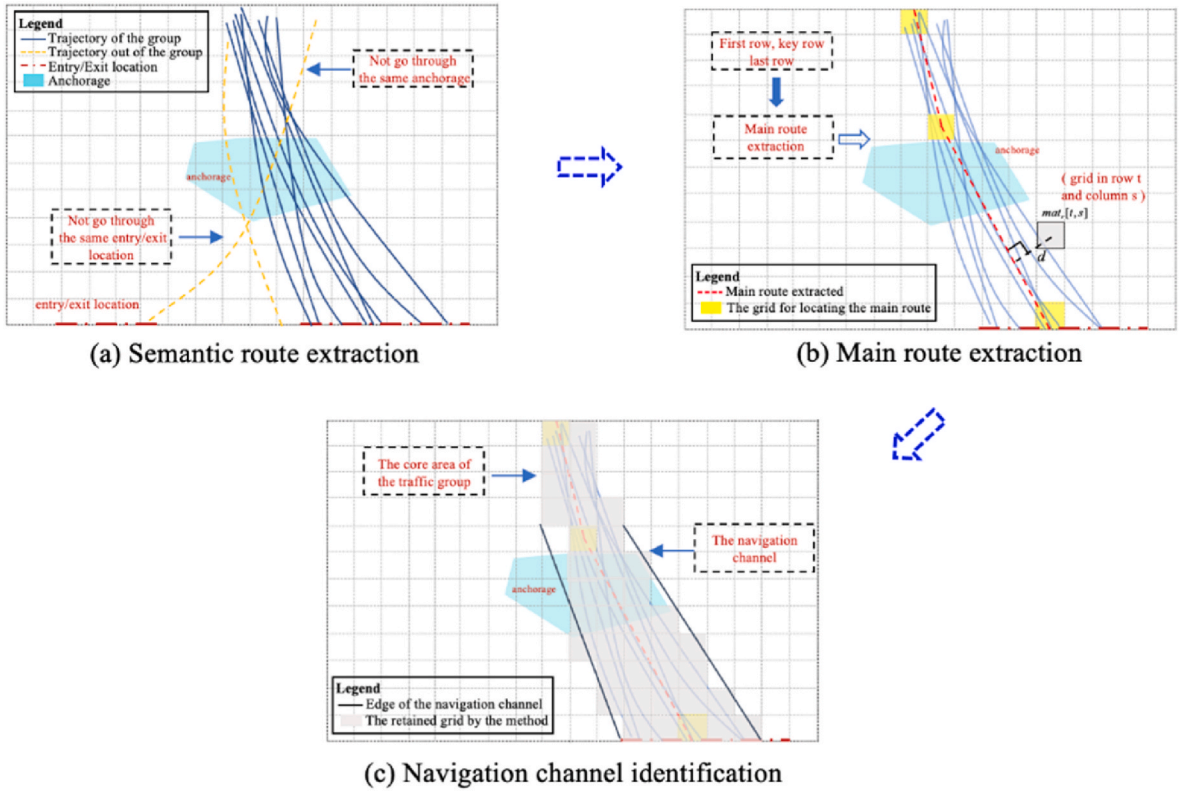

Fig. 3. Maritime traffic network generation.

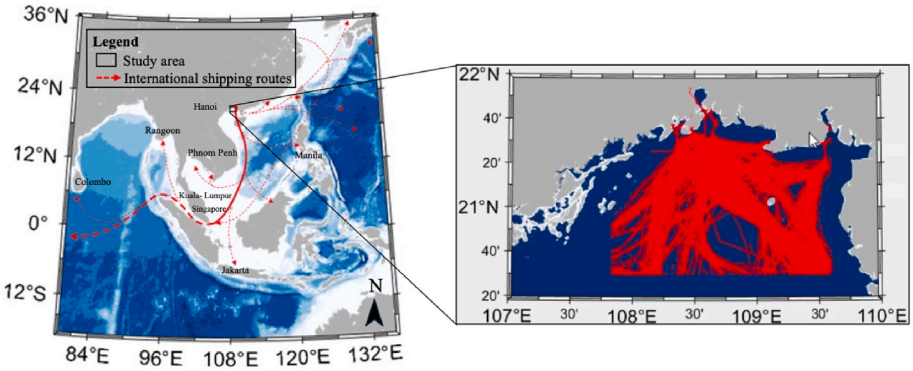

Fig. 4. The international shipping routes linked to the Beibu Gulf and visualization of AIS messages within the monitored area.

routes. Therefore, it is possible to identify the navigation channels on the basis of this symmetry characteristic.

Based on the grid-systems of traffic groups, we can roughly locate the blurred edges of the areas where the grids with values are distributed. However, because the data are inadequately preprocessed, there may exist a small number of trajectories that interfere with the identification of the navigation channels. To solve the problem, a grid-based scarification model is proposed, with the importance function established to eliminate the grids outside the core area of the traffic group. In the normal distribution curve of traffic flow, the probability decreases uniformly from the center to both sides. Therefore, we establish a grid-based importance function, using grid values and the distances between the grid lines and peak lines as metric factors. The function is as follows:

$$grid\_import(i,j) = mat_k[i,j] \times \frac{1}{d(mat_k^{main}[i], mat_r[i,j])} \tag{5}$$

where $grid\_import(i,j)$ represents the importance of the grid in the traffic group, $mat_k[i,j]$ represents the stored value of the grid located in row i and column j in the grid matrix, and $d(mat_k^{main}[i], mat_r[i,j])$ represents the distance between the grid and the main route in the same matrix row.

With the function above, the importance of each grid to the traffic group can be measured. To deal with the density differences among traffic groups, a cumulative importance function is applied to each traffic group. It can evaluate the overall importance of the retained grids in each group, and retain the important area according to the function results. The cumulative importance function is as follows:

$$\text{Important}_\rho^k = \sum grid\_import(i,j) \tag{6}$$

where Important_ρ^k represents the cumulative importance of the reserved areas under a threshold value of ρ, ρ represents the threshold value relating to the number of trajectories stored in the grid, and $grid\_import(i,j)$ represents the importance of the reserved grid in the *tth*

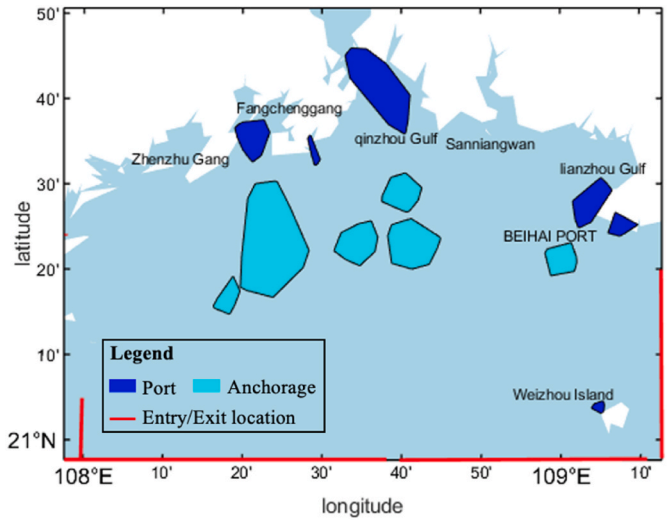

Fig. 5. Stop areas and entry/exit locations identified.

row and *sth* column, calculated by equation (5). With increasing of the threshold , the number of calculated grids will grow, and the cumulative importance of the retained area will raise.

A line chart is created according to the cumulative importance function. It clearly shows that as the threshold value increases, which means the area of interest expands, the cumulative importance tends to increase monotonically, and its curve slope gradually decreases until it tends to level off. This indicates that under this state, the importance of the retained grids has risen to a limit, and continuing to expand the retained area will result in an increase in redundancy far greater than the raise in effective information. Therefore, it can be considered that when the curve tends to be horizontal, the area of interest is gradually covering the core area of the traffic group. And the corresponding area of interest is the navigation channel of the group.

Based on the function curve above, we obtain the inflection point of the cumulative importance curve and view the reserved area corresponding to the threshold at this point as the navigation channel of the traffic group. The inflection point can be obtained by calculating the second-order derivative to zero, and the formula is as follows:

$$f''(x) = \lim_{\Delta x \to 0} \frac{f'(x_0 + \Delta x) - f'(x_0)}{\Delta x} \tag{7}$$

where $f''(x)$ is the second-order derivative of $\text{Imptant}_{\rho}^{k}$, $f'(x)$ is the first-order derivative of $\text{Imptant}_{\rho}^{k}$. A σ making $f''(\sigma)$ equal to 0 is the threshold value corresponding to the inflection point.

Following the above steps, we obtain the distribution of the navigation channels. However, the outline of these channels needs to be clarified. With the cumulative importance model, the noise grids in the traffic group are eliminated, and the edges of the channels are clear to a certain extent. Then, according to the distribution of the valued grids, the points of each edge are connected to form the outer frame. The closed loop polygon has the characteristic of long strips and approximate symmetry. In addition, the edges identified are mostly segmental, so a smoothing process is required. Finally, we retain the key points of the edges and discard the points between key points. With this method, the navigation channels of each traffic group are obtained. And the main routes and navigation channels constitute the network. Fig. 3 shows the process of maritime traffic network generation. Algorithm 3 shows the pseudocode for the grid-based navigation channel identification algorithm.

Algorithm 3. Grid-based navigation channel identification algorithm

Input: Grid-based traffic groups $mat_{all} = \{mat_1, mat_2, \cdots, mat_m\}$, Grid position of main routes $mat_{all}^{main} = \{mat_1^{main}, mat_2^{main}, \cdots, mat_m^{main}\}$, longitude and latitude position $tra_{all}^{main} = \{tra_1^{main}, tra_2^{main}, \cdots, tra_m^{main}\}$, minimun longitude and latitude of the research area lon_{min}、lat_{min}

Output: Channel position coordinates $lane_{all} = \{lane_1, lane_2, \cdots, lane_m\}$

Process:

/*Annotations for line 2-20: Grid importance sparse algorithm*/
function $getDistance(point, line)$
function $getExtremepoint(x_1, y_1, x_2, y_2)$
for τ *in* $np.arange(0.1,0.1,15,0.01)$: // The threshold interval will affect the function result
  for ρ *in* $np.arange(0.1,0.1 + (50 // (\tau \times 100)) \times \tau, \tau)$: // Define the threshold
    $mat\_point_{i,j}^{\rho} = \left[\left[lon_{min} + (j \times 0.01), lat_{min} + ((88 - i) \times 0.01)\right] \text{for i in range}(mat_r) \text{ for j in range}(mat_r[i])\right]$ // Get the coordinates of this grid
    $mat\_dist_{i,j}^{\rho} = \text{getDistance}(mat\_point, \text{mat\_line})$ // Calculate the distance between the grid and main route
    $mat\_value_{i,j}^{\rho} = mat_r[i,j] \times \frac{1}{mat\_dist}$ // Get the importance of the grid
    $mat\_value_{all}^{\rho} = \sum_{[i,j]} mat\_value$ // Accumulate the importance of all grids with value under this threshold to get the overall importance
    $mat\_value_{all}.append(mat\_value_{all}^{\rho})$
    $\rho_{all}.append(\rho)$
    $mat\_value_{all}^{\rho} = mat\_value_{all}^{\rho} - np.mean(mat\_value_{all})$ // Data standardization
    $plt.plot(\rho_{all}, mat\_value_{all})$ // Draw the function image of threshold and overall importance, and find it has convexity
    $mat\_value_{all}^{\rho}{}' = \frac{\Delta mat\_value_{all}}{\Delta \rho_{all}}$ // First derivative of overall importance
    $mat\_value_{all}^{\rho}{}'' = \frac{\Delta mat\_value_{all}^{\rho}{}'}{\Delta \rho_{all}'}$ // Second derivative of overall importance
    $mat\_value_{all}''.append(mat\_value_{all}^{\rho}{}'')$
    $\rho_{all}''.append(\Delta \rho_{all}')$
    $plt.plot(\rho_{all}'', mat\_value_{all}'')$ // Draw the second derivative image, and the point intersecting the zero axis is the possible inflection point
    $extremepoint\_num.append(num^{\rho})$ // Count the number of inflection points under different threshold
$\sigma = getExtremepoint(x_1, y_1, x_2, y_2)$ // Calculate the inflection point for the image corresponding to the minimum and stable inflection points, and σ is the threshold
/*Annotations for line 22-42: Navigation channel identification algorithm*/
function $getMinlat(mat)$
function $getMinlon(mat)$
function $getUpperlon(mat)$
function $getLowerlon(mat)$
for i *in* $range(len(mat_r{}^*))$: // Identify the left and right boundary of the traffic group after sparing
  $minlat_i = getMinlat(mat_r{}^*[i])$
  $minlon_i = getMinlon(mat_r{}^*[i])$
  $leftedge\_lat.append(lat_{min} + (matrix\_num - minlat_i) \times 0.01)$ // Calculate the latitude of upper left vertex
  $leftedge\_lat.append(lat_{min} + (matrix\_num - minlat_i - 1) \times 0.01)$ // Calculate the latitude of lower left vertex
  $leftedge\_lat.append(lon_{min} + minlon_i \times 0.01)$ // Calculate the longitude of upper and lower left vertex. The same goes for the right
  for j *in* $range(len(mat_r{}^*[0]))$: // Identify the upper boundary of the traffic group after sparing
    $upperlon_j = getUpperlon(mat_r{}^*[j])$
    $upperedge\_lat.append(lat_{min} + (matrix\_num - minlat_0) \times 0.01)$
    $upperedge\_lon.append(lon_{min} + j \times 0.01)$
  for k *in* $range(len(mat_r{}^*[-1]))$: // Identify the lower boundary of the traffic group after sparing
    $lowerlon_k = getUpperlon(mat_r{}^*[k])$
    $loweredge\_lat.append(lat_{min} + (matrix_{num} - minlat_{-1} - 1) \times 0.01)$
    $loweredge\_lon.append(lon_{min} + k \times 0.01)$
  $rightedge\_lat = list(reversed(rightedge\_lat))$ // Reverse the order of the right and upper boundary to connect the edges in a closed loop
  $lane_r.extend([leftedge\_lon, loweredge\_lon, rightedge\_lon, \cdots])$ // Integrate the coordinates to obtain the channel of traffic group
  return $lane_{all} = \{lane_1, lane_2, \cdots, lane_m\}$.

3.3.3. Performance evaluation

The effectiveness of the maritime traffic network extraction method can be judged both qualitatively and quantitatively. From a qualitative point of view, high-quality results should be generally consistent with high-density areas of ship traffic flow and reflect the routes taken by the majority of ships. And from a quantitative point of view, we can compare the extraction results with officially published routes to quantify the effectiveness of the proposed method.

The officially published information is commonly found in Navigation Guides or published marine charts for the waters in question. On the basis of the practical needs of mariners and full reference to international and domestic rules, these guides provide the informative and important reference for the majority of navigating ships.

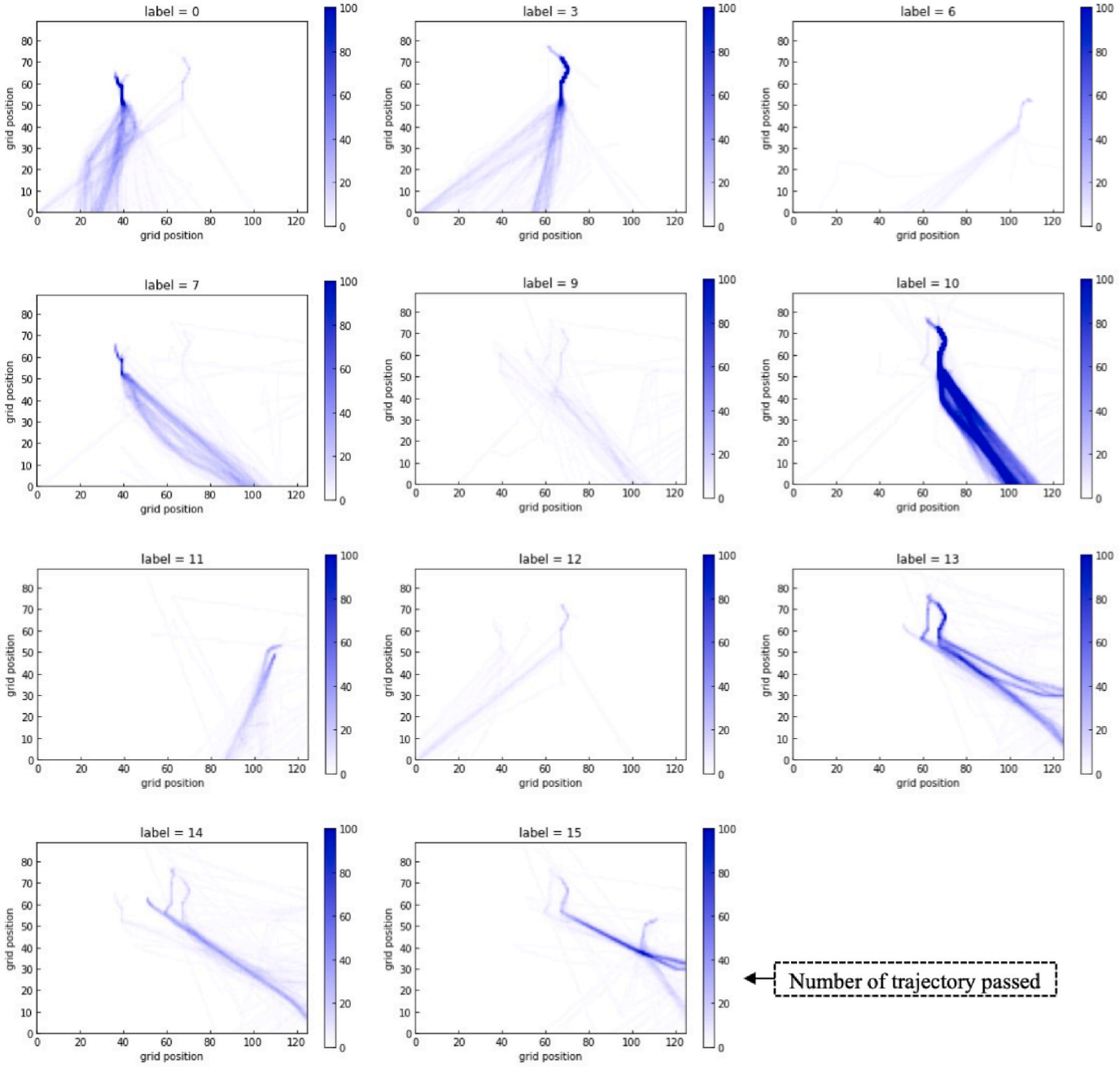

Fig. 6. Results of semantic routes.

According to the information provided in the Navigation Guides, the officially published routes consist of multiple coordinates and are polygons in shape, which means they have area properties. Therefore, the validity of the model proposed is analyzed by calculating the coverage rate of the extracted results with the corresponding official routes:

$$\text{Coverage Rate} = \sum_{i=1}^{m} \frac{S_i^{\text{extra}}}{S_i^{\text{offici}}} \tag{8}$$

where Coverage Rate is the coverage rate of extraction results with the corresponding official routes in the study water, S_i^{extra} is the area of the *ith* extraction navigation channel, and S_i^{offici} is the area of the official published route corresponding to the *ith* extraction route.

In summary, the quality of the extraction results can be measured not only from a qualitative point of view by observing the matching degree, but also from a quantitative point of view by calculating its coverage rate with the official published routes.

The above approaches can comprehensively assess the effectiveness of the extraction results in reflecting the traffic flow status in the study water.

4. Experimental results and analysis

4.1. Case study area

To verify the effectiveness of the maritime traffic network extraction method proposed in this study, real AIS historical data were used to evaluate the method. The Beibu Gulf has a coastline of 1628 km and covers a total sea area of 129,000 square kilometers, and it is convenient

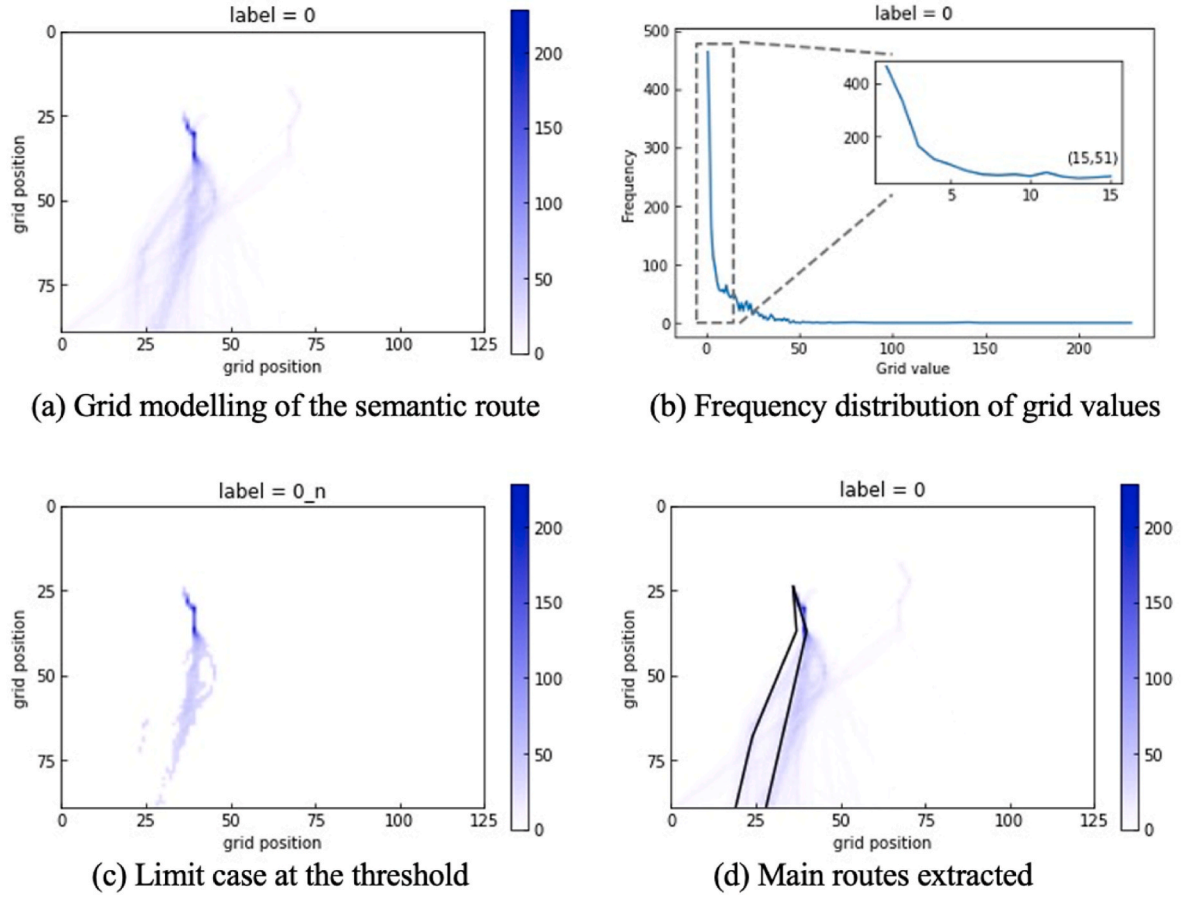

Fig. 7. A case of main route extraction
(This semantic route needs to be decomposed because of the anchorage).

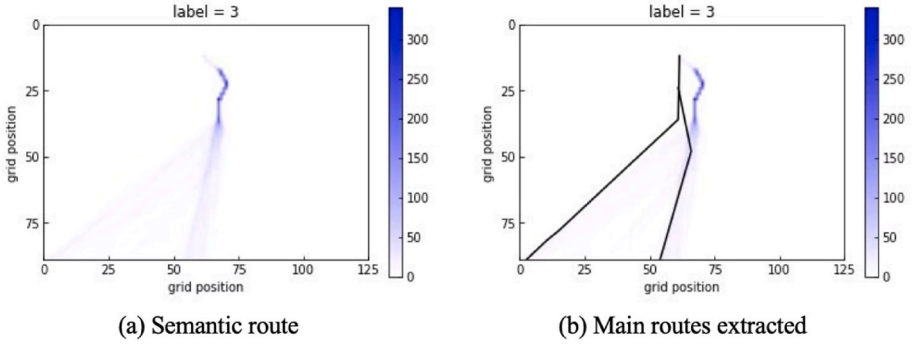

Fig. 8. A case of main route extraction
(This semantic route needs to be decomposed because the entry/exit area identified is too large).

to access the sea in the southwest of China. As the key node of maritime transportation channels, many international shipping routes meet here. And the monitored area in this study includes Fangcheng Port, Qinzhou Bay, Beihai Port, Weizhou Island and some other parts of the sea, covering an area of more than 12,000 square kilometers and spanning from 107.96 °E to 109.21 °E in longitude. The AIS historical data used are the data broadcast from January 1 to June 30, 2019, consisting of more than 2.27 million pieces of AIS information broadcast. Fig. 4 shows the international shipping routes linked to the Beibu Gulf and the spatial distribution of AIS data. The study area is a typical complicated water, for the shipping routes run through the vertical and horizontal area, as well as the ship traffic density is relatively high.

4.2. Ship trajectory waypoints extraction

The maritime traffic pattern recognition method was used to identify the stop points and waypoints within the data. It should be noted that the size of sliding window in the waypoint recognition method was 10 trajectory points and the threshold of vector cosine angle was set as 0.99 by experiment and experience in this study. As the upper limit of ships' stop time was 1800 s in the stop point recognition method and ships' time intervals were primarily about 300 s by analyzing the timestamp distribution, we defined the length of window as 10 points considering the fluctuation of data. And then the DBSCAN-based stop area recognition method was applied to identify the departure-arrival areas.

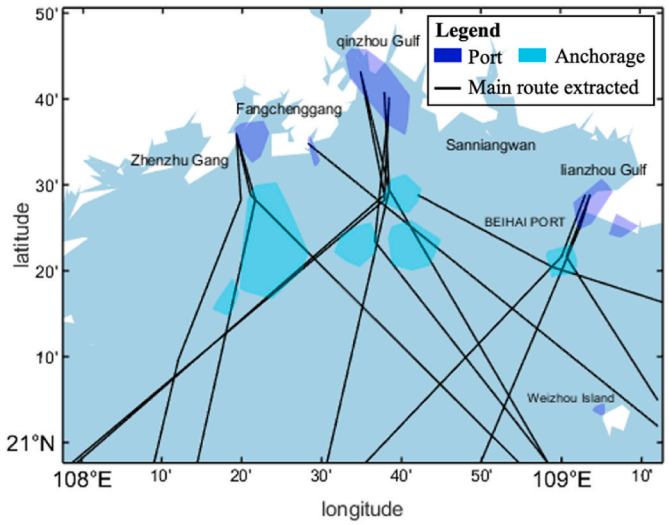

Fig. 9. Result of main route extraction.

Meanwhile, because the study area has a boundary, the DBSCAN algorithm was also used to identify the entry/exit locations. The DBSCAN algorithm contains two threshold parameters, namely, *MinLns* and ε, where ε denotes a spatial distance threshold delimiting the neighborhood of the point and *MinLns* denotes the minimum number of the points required to form a dense cluster (Zhang et al., 2021). For the stop area recognition method, the clustering parameters were determined according to the match of results and relevant materials containing the actual distribution of anchorages as well as ports in the study area. In this paper, several groups of *MinLns* (100–1000) were compared with ε between 0.01 and 0.05. The experiences show that when the *MinLns* and ε were determined as 500 and 0.03, the results of stop areas were most compatible with the materials. Similarly, the clustering parameters of the entry/exit location identification method were determined based on the idea that the intersection points were covered with the locations as few as possible. And the results performed best when the *MinLns* and ε were defined as 300 and 0.2. As shown in Fig. 5, twelve stop areas were identified with the algorithm, with dark blue polygons indicating ports and light blue polygons indicating anchorages. And four entry/exit locations were identified. In Fig. 5, these locations are represented by red line segments. The anchorages and entry/exit locations formed the departure-arrival areas in this study.

In addition, to ease the computation without affecting data quality, a waypoint recognition algorithm was used to compress the trajectory data with the ship behavior patterns identified. After compression, the number of data points was reduced from 2,389,749 to 181,808, and the compression rate was 7.6%. A large amount of ship trajectory data was redundant, making it necessary to compress the data.

4.3. Ship routes extraction

Based on the departure-arrival areas identified through the methods mentioned, we generated semantic routes with different combinations of departure-arrival areas. Basically, the semantic routes reflect the behavioral characteristics of ships which follow the same behavior pattern and similar spatial distribution during the course. And then, to realize the fusion of trajectory data and geographic location information, the grid-system was applied in this paper. Constructing the grid model with each grid, which stored the number of trajectories passed by, can convert the vector data into matrix data. Such data structure contributes to the quantitative analysis of trajectory data, for example, the traffic density at different positions. And with the statistical character-

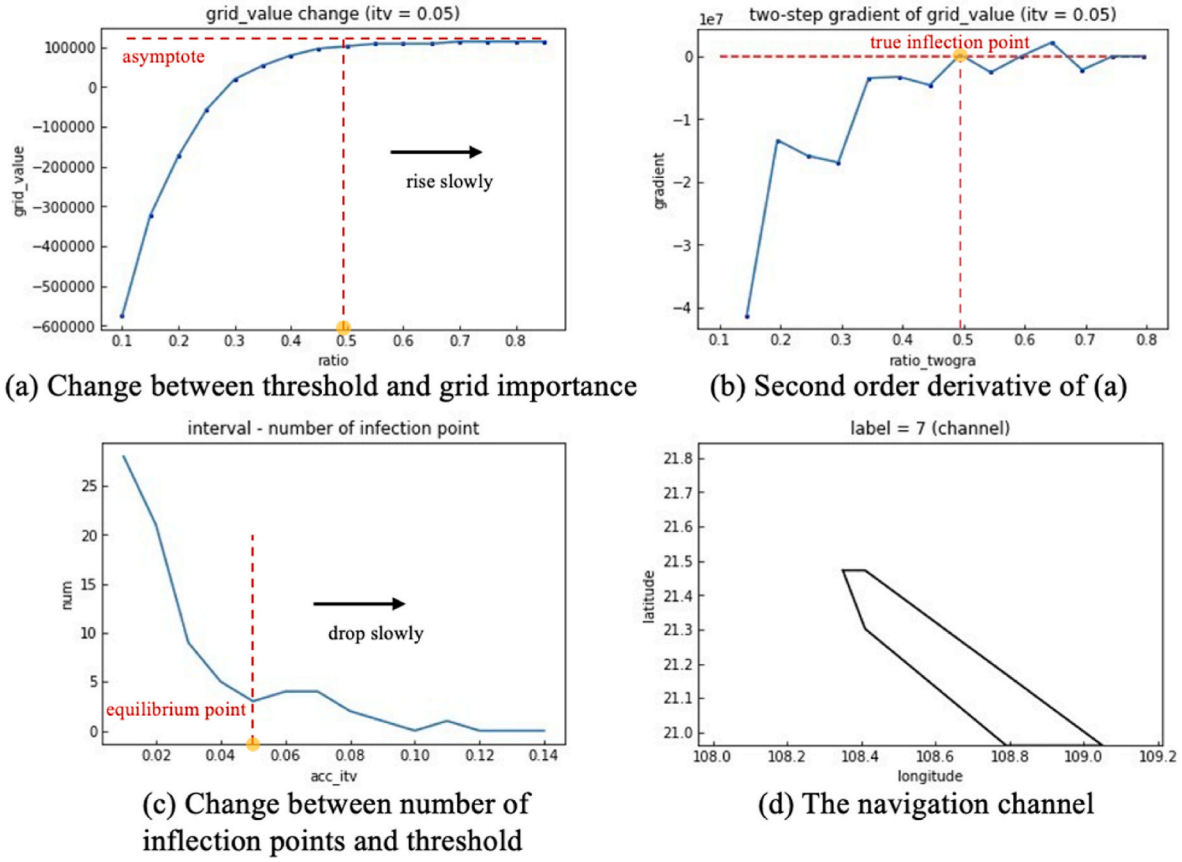

Fig. 10. A case of navigation channel identification.

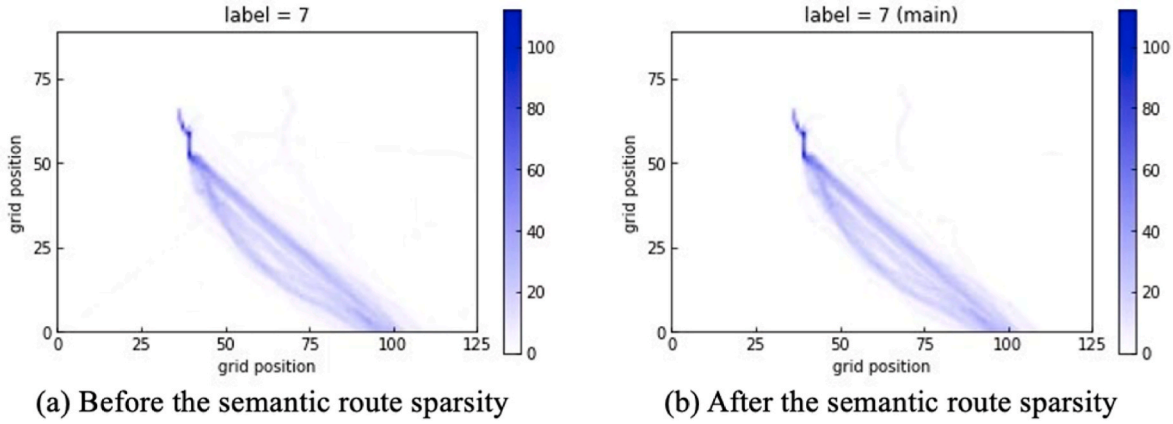

Fig. 11. A case of semantic route scarification.

istics of ship traffic flow introduced, the density-oriented route extraction method can be achieved. With a division unit of 0.01° in latitude and longitude, the area was converted into an 89 × 125 grid matrix. Then, the semantic routes were transformed in the grid system, with different label numbers. However, some stop areas in the water may overlap, so it is necessary to eliminate those semantic routes that overlap with each other. As shown in Fig. 6, eleven final semantic routes were identified. Each route passed through a different combination of entry/exit locations and stop areas, characterizing a different ship behavior pattern. And it is clear that the ship traffic density of the groups changes in a large range, which further illustrates the complexity of the study area.

In addition, simplifying the entry/exit locations during the clustering process may result in semantic routes containing more than one ship behavior pattern. A grid-based semantic route decomposition algorithm was used in this study to solve the problem. The semantic routes extracted in the grid-system are also called traffic groups. The feature lines characterized by the maximum relative densities of the traffic groups can be extracted to present their features as a whole. With the consideration that trajectories of a traffic group do not always follow a standard normal distribution, it is necessary to single out high-density lines according to the distribution of the data. Based on the above analysis, a grid-based main route extraction algorithm was proposed to extract the main routes in the study area. The grid-system that transforms vectors to grids storing the number of trajectories passing through can easily reflect the density distribution of ship traffic flow. What's more, this method can avoid interference from wrong or abnormal trajectories and precisely locate the main routes at the boundary. For example, suppose there was an anchorage shown in Fig. 7(a), which was resulted from the mixing of two different trajectory clusters that pass through the same port and entry/exit location. By applying the grid-based semantic route decomposition algorithm, we first counted the frequencies of all grid values and obtained a monotonically decreasing line graph of the frequency distribution of grid values, as shown in Fig. 7 (b). The horizontal coordinate of the inflection point was 15. Then, the grid values not greater than the threshold in the traffic group were set to zero for the initial denoising. The final result is shown in Fig. 7(c). Then the spatially separated segments were identified, and finally, the grid-based main route extraction algorithm was used to extract the main routes after route decomposition. The results are shown in Fig. 7(d). Similarly, the traffic groups labelled as 3 were apparently different. The reason for this misjudgment was that the entry/exit location was too large. Therefore, the decomposition algorithm identified and separated the traffic groups, as shown in Fig. 8. Finally, the results of the main route extraction in the water are shown in Fig. 9, with a total of thirteen main routes extracted.

4.4. Maritime traffic network extraction and results evaluation

In addition to the main routes, the cross-section of a ship traffic flow needed to be explored to obtain more distribution information of the trajectory data to analyze the ship behavior patterns in the water fully. So, a grid-based scarification model was established, which used the grid importance function to identify the core region of a traffic group. Then, we obtained the smoothed boundary points in the region to identify the navigation channels. Fig. 10 shows the identified channels of each traffic group. Fig. 10(a) shows the convex relationship curve between the threshold and the overall grid importance. The inflection point has the practical significance that when the threshold is reached, and the overall importance of the traffic group will increase extremely slowly, with the grid area retained under that threshold close to the core area of the traffic group. This core area is also the navigation channel to be identified. Fig. 10(b) shows the second order derivative curve of the overall grid importance, and the point with a horizontal coordinate of zero was the possible the inflection point. Fig. 10(c) shows the relationship curve between the number of inflection points and the threshold. The number of inflection points decreases with the increase in the threshold. The point from which the curve is shown in Fig. 10(c) becomes leveling off corresponds to the point of the second-order derivative curve shown in Fig. 10(b) from which the oscillation effect becomes less significant under the same threshold. Therefore, this threshold was selected to identify the inflection points in Fig. 10(a) and (b). After that, this threshold was substituted into the channel recognition algorithm, with the recognition result shown in Fig. 10(d). Fig. 11 shows the heat map results before and after the application of the grid importance sparse algorithm. It can be seen that the grid-based scarification algorithm can effectively remove the noise data and highlight the core region.

Finally, the navigation channels extracted were analyzed qualitatively and quantitatively. From a qualitative perspective, the routes extracted in the water should be basically consistent with the high-density areas of ship traffic flow, which means they should reflect the spatial distribution of most ships. From a quantitative perspective, the results of routes can be compared with the corresponding sailing materials. For the safety of navigation, there exist officially published Sailing Direction materials, which record the regulations about ships' navigation, communication, power systems, environmental protection and personnel job requirements (Han and Song, 2016). Based on this, the officially published "Guide to Vessel Navigation in Guangxi Waters of Beibu Gulf" was introduced in this study. According to the routes recorded in this guideline, the matching degrees between the extracted results and the official planned routes can be compared. In Fig. 12, the extracted navigation channels are shown in the left column, while the distributions of the official routes are shown in the right column.

Furthermore, the coverage rates between the extracted channels and

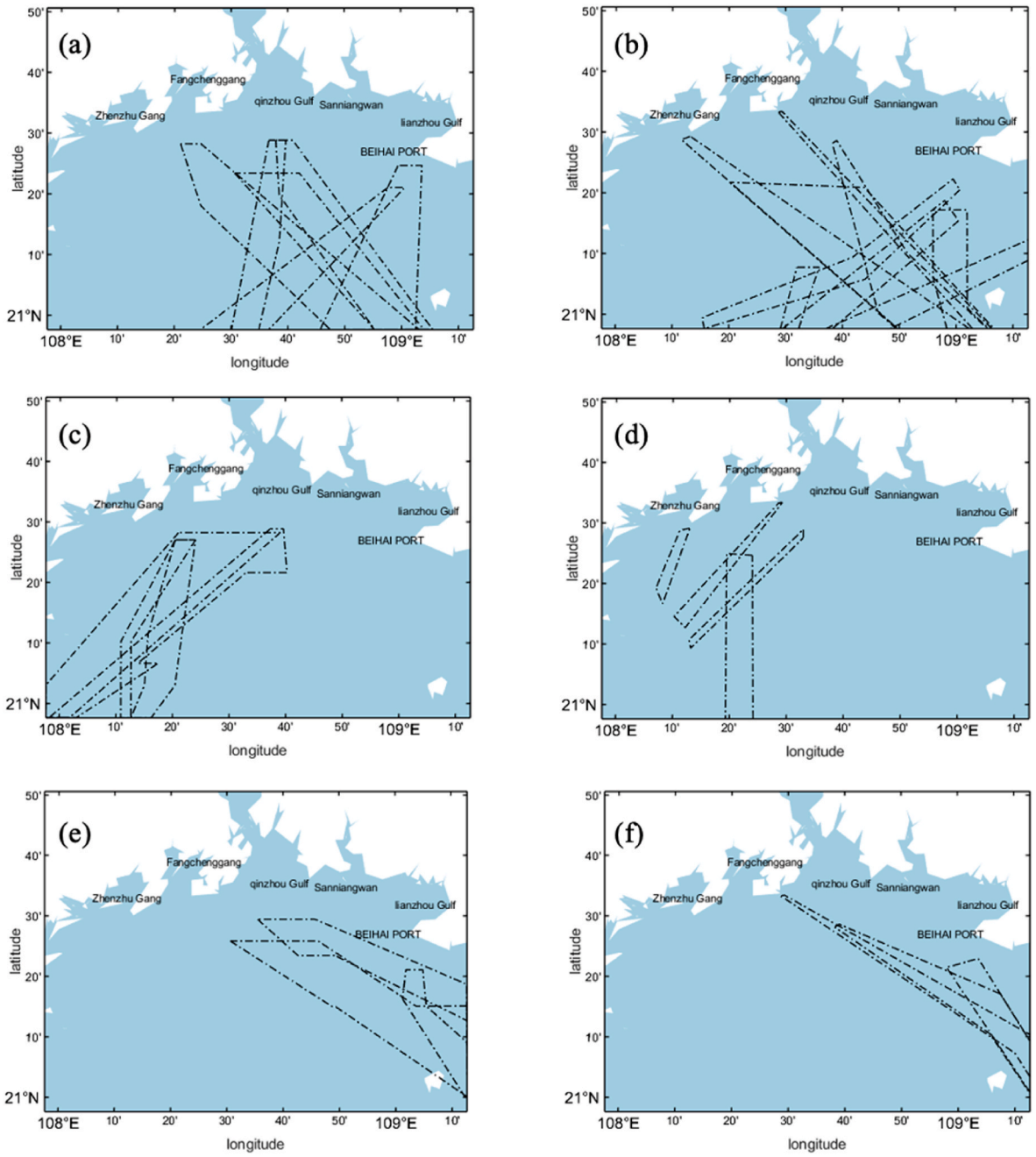

Fig. 12. Result comparison (Figures (a, c, and e) show the navigation channels extracted, and Figures (b, d, and f) show the official routes).

the official routes were calculated as presented in the equation (8). Analyzing the area coverage between both can verify the effectiveness of the method. And the results are shown in Table 1. Furthermore, the obtained findings indicate that a proportion of ships are unable to operate entirely within the official navigation channels within the actual navigational setting, consequently increasing the probability of maritime incidents. Thus, it is imperative to reinforce the regulation of ships to curtail ships that traverse outside the designated channel, ultimately enhancing the safety of ships and advancing maritime traffic management.

The results show that among the total eighteen official routes published, seventeen routes were located in the study water. The routes extracted in this study were all consistent with these seventeen official routes. Besides, nearly half of the results covered more than 80% of the corresponding official published routes, and the average coverage rate of the results was 75.2%, indicating that the model proposed in this paper is valid and has practical significance.

5. Conclusion and future work

In this paper, a maritime traffic network extraction method is proposed based on ship behavior patterns. According to the principle that trajectories of a same route should follow similar ship behavior and spatial distribution patterns, a data-driven method is applied to recognize ship behavior patterns to identify the departure-arrival areas in the study water and compress the trajectory data. After that, the ship trajectories can be classified with different combinations of departure-arrival areas. Then, a grid-system is introduced to rasterize each traffic group. With the fusion of trajectory data and geographic location information, the grid-based semantic route decomposition algorithm and

Table 1
Quantitative analysis of the results.

ID (Navigation channel)	Corresponding official route	Coverage rate
1	West Route from Fangchengggang to Southeast Asian Countries	45.15%
2	West Route from Fangchenggang to Southeast Asian Countries	67.42%
3	Route from Tieshangang to Southeast Asian Countries	83.34%
4	Route from Beihai Port to Southeast Asian Countries	51.08%
5	Route from Bailong Port to Qiongzhou Strait	86.94%
6	Route from Fangchenggang to Qiongzhou Strait	49.05%
7	Route from Qinzhou Strait to Qiongzhou Strait	71.61%
8	Route from Beihai Port to Qiongzhou Strait	96.11%
9	Route from Bailong Port to Vietnam	45.15%
10	Route from Qisha Port to Vietnam	88.17%
11	Route from Qisha Port to Qiongzhou Strait	87.42%
12	Route from Qinzhou Port to Qiongzhou Strait	27.87%
13	Route from Beihai Port to Qiongzhou Strait	67.04%
14	Route from Qisha Port to Qiongzhou Strait via Weizhou Island	98.87%
15	Route from Qinzhou Port to Qiongzhou Strait via Weizhou Island	87.52%

main route extraction algorithm are proposed to extract the main routes. Finally, with the consideration of statistical characteristics of traffic flow distribution, a grid cumulative importance function has been constructed. Based on this, the grid-based channel identification algorithm is applied to identify the channels around the main routes. And the real-world AIS data in the first half of the year 2019 are used to extract the maritime traffic network within the framework. By the waypoint recognition method of identifying the ship maneuvering pattern, the data are compressed with a compression rate of 7.6%. Furthermore, a total of twelve stop areas corresponding to the static ship behavior pattern are identified, which include six ports and six anchorages. As the study area has a boundary, the four entry/exit locations are identified. And the stop areas and entry/exit locations form the departure-arrival areas. Subsequently, based on the departure-arrival areas recognized, we extract the maritime traffic network, obtaining thirteen main routes and their corresponding navigation channels. The routes extracted in this study are analyzed quantitively, which shows that they are all consistent with the seventeen official routes published. Besides, nearly half of the results covered more than 80% of the corresponding official published routes, and the average coverage rate of the results is 75.2%.

In general, the ship behavior patterns are fully considered in the model, which can be applied to analyze the complicated water. On one hand, the grid-based route extraction and channel identification methods applied in this paper utilize the advantages of grid-system in removing noise data and greatly relieving the computational burden. In addition, the channel identification process can be well performed under a grid-cell environment with area properties. On the other hand, it can reveal the distribution of historical traffic flow in the water by mining the information of ports, anchorages, and routes there, thus providing the reference for the management departments on route planning. Meanwhile, with the extracted main routes and channels, crew members can make more appropriate navigation plans or pay more attention to high-risk areas during navigation to avoid traffic accidents. Thus, the results can enhance the safety of the ship navigation environment.

Ship behavior patterns are applied in the maritime traffic network extraction method proposed in this study. The patterns can reveal the correlation between AIS data and ship behavior features, as well as improve the accuracy of network extraction. However, for those ships with incorrect input of ship type in static information, such as those that are fishing vessels but display as cargo ships in AIS information, this study has not yet conducted research on identifying these ships. As the fishing vessels that need to change navigation behaviors frequently, the effectiveness of the maritime traffic pattern identification could be compromised because many false stop points and waypoints are generated. Therefore, the future research should consider the influencing factor of vessel type on the traffic network extraction along with the factors of speed, distance and time, thus improving the robustness of the method. Furthermore, this method will be further tested in other complicated water to strengthen the universality.

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.

Data availability

Data will be made available on request.

Acknowledgments

This study was supported by the National Natural Science Foundation of China (Grant No. 52171351).

Appendix. Nomenclature

Variable	Definition
tra_i	Each ship trajectory
tp_j	The trajectory point
v_{tp_j}	The instantaneous velocity at point t_{tp_j}
t_{tp_j}	The timestamp
$dist_{tp_j,tp_{j+1}}$	The distance
v_T, t_T, $t_T{}'$, $dist_T$	The threshold
loc_{all}	The set of entry/exit locations
p	The size of the sliding window
$\cos\theta_j$	The vector cosine angle
$\overrightarrow{tp_k tp_{k-a}}$	The vector from point tp_k to tp_{k-a}
v_T	The threshold of vector cosine angle
tra_m^{turn}	The *mth* turning section identified
$\cos\theta_n^{turn}$	The vector cosine angle of the section
r	An integer varying from 1 to $p-2$
pw_t^i	The waypoint in the turning section
tra_i^{cpr}	The compressed trajectory

(*continued on next page*)

(*continued*)

Variable	Definition
Tra_i^{group}	The set of semantic routes
$grid\_import(i,j)$	The importance of the grid in the group
$mat_k[i,j]$	The stored value of the grid
$d(mat_k^{main}[i], mat_r[i,j])$	The distance between grid and main route
Important_ρ^k	The cumulative importance under ρ
ρ	The threshold of number in grid
$grid\_import(i,j)$	The importance of the reserved grid
$f''(x)$	The second-order derivative
$f'(x)$	The first-order derivative
Coverage Rate	The coverage rate
S_i^{extra}	The area of the navigation channel
S_i^{offici}	The area of the official published route

References

Andersson, P., Ivehammar, P., 2017. Green approaches at sea–The benefits of adjusting speed instead of anchoring. Transport. Res. Transport Environ. 51, 240–249.

Aregall, M.G., Bergqvist, R., Monios, J., 2018. A global review of the hinterland dimension of green port strategies. Transport. Res. Transport Environ. 59, 23–34.

Arguedas, V.F., Pallotta, G., Vespe, M., 2017. Maritime traffic networks: from historical positioning data to unsupervised maritime traffic monitoring. IEEE Trans. Intell. Transport. Syst. 19 (3), 722–732.

Bomberger, N.A., Waxman, A.M., Rhodes, B.J., Sheldon, N.A., 2007. A new approach to higher-level information fusion using associative learning in semantic networks of spiking neurons. Inf. Fusion 8 (3), 227–251.

Chen, J., Bian, W., Wan, Z., Yang, Z., Zheng, H., Wang, P., 2019. Identifying factors influencing total-loss marine accidents in the world: analysis and evaluation based on ship types and sea regions. Ocean Eng. 191, 106495.

Chen, J., Ye, J., Zhuang, C., Qin, Q., Shu, Y., 2022a. Liner shipping alliance management: overview and future research directions. Ocean Coast Manag. 219, 106039.

Chen, J., Zhang, W., Song, L., Wang, Y., 2022b. The coupling effect between economic development and the urban ecological environment in Shanghai port. Sci. Total Environ. 841, 156734.

Chen, J., Zhuang, C., Xu, H., Xu, L., Ye, S., Rangel-Buitrago, N., 2022c. Collaborative management evaluation of container shipping alliance in maritime logistics industry: CKYHE case analysis. Ocean Coast Manag. 225, 106176.

Chen, J., Zhuang, C., Yang, C., Wan, Z., Zeng, X., Yao, J., 2021. Fleet co-deployment for liner shipping alliance: vessel pool operation with uncertain demand. Ocean Coast Manag. 214, 105923.

Chen, X., Liu, Y., Achuthan, K., Zhang, X., 2020. A ship movement classification based on Automatic Identification System (AIS) data using Convolutional Neural Network. Ocean Eng. 218, 108182.

Dobrkovic, A., Iacob, M.E., Van Hillegersberg, J., 2015. Using machine learning for unsupervised maritime waypoint discovery from streaming AIS data. In: Proceedings of the 15th International Conference on Knowledge Technologies and Data-Driven Business, pp. 1–8.

Fan, L., Wilson, W.W., Tolliver, D., 2010. Optimal network flows for containerized imports to the United States. Transport. Res. E Logist. Transport. Rev. 46 (5), 735–749.

Fu, S., Yu, Y., Chen, J., Han, B., Wu, Z., 2022a. Towards a probabilistic approach for risk analysis of nuclear-powered icebreakers using FMEA and FRAM. Ocean Eng. 260, 112041.

Fu, S., Yu, Y., Chen, J., Xi, Y., Zhang, M., 2022b. A framework for quantitative analysis of the causation of grounding accidents in arctic shipping. Reliab. Eng. Syst. Saf. 226, 108706.

Han, L., Song, S., 2016. The regulation of international law in the arctic sea area and its impact on China's use of polar route. Research on Chinese Maritime Law 27 (3), 56–63.

Huang, C., Qi, X., Zheng, J., Zhu, R., Shen, J., 2023. A maritime traffic route extraction method based on density-based spatial clustering of applications with noise for multi-dimensional data. Ocean Eng. 268, 113036.

Kaluza, P., Kölzsch, A., Gastner, M.T., Blasius, B., 2010. The complex network of global cargo ship movements. J. R. Soc. Interface 7 (48), 1093–1103.

Kang, L., Meng, Q., Liu, Q., 2018. Fundamental diagram of ship traffic in the Singapore Strait. Ocean Eng. 147, 340–354.

Lee, J.S., Son, W.J., Lee, H.T., Cho, I.S., 2020. Verification of novel maritime route extraction using kernel density estimation analysis with automatic identification system data. J. Mar. Sci. Eng. 8 (5), 375.

Lin, D.Y., Chang, Y.T., 2018. Ship routing and freight assignment problem for liner shipping: application to the Northern Sea Route planning problem. Transport. Res. E Logist. Transport. Rev. 110, 47–70.

Liu, C., Liu, J., Zhou, X., Zhao, Z., Wan, C., Liu, Z., 2020. AIS data-driven approach to estimate navigable capacity of busy waterways focusing on ships entering and leaving port. Ocean Eng. 218, 108215.

Liu, L., Shibasaki, R., Zhang, Y., Kosuge, N., Zhang, M., Hu, Y., 2023. Data-driven framework for extracting global maritime shipping networks by machine learning. Ocean Eng. 269, 113494.

Liu, Z., Zhang, B., Zhang, M., Wang, H., Fu, X., 2023. A quantitative method for the analysis of ship collision risk using AIS data. Ocean Eng. 272, 113906.

Murray, B., Perera, L.P., 2022. Ship behavior prediction via trajectory extraction-based clustering for maritime situation awareness. J. Ocean Eng. Sci. 7 (1), 1–13.

Pallotta, G., Vespe, M., Bryan, K., 2013. Vessel pattern knowledge discovery from AIS data: a framework for anomaly detection and route prediction. Entropy 15 (6), 2218–2245.

Rong, H., Teixeira, A.P., Guedes Soares, C., 2019. Ship trajectory uncertainty prediction based on a Gaussian Process model. Ocean Eng. 182, 499–511.

Sheng, P., Yin, J., 2018. Extracting shipping route patterns by trajectory clustering model based on automatic identification system data. Sustainability 10 (7), 2327.

Silveira, P.A.M., Teixeira, A.P., Soares, C.G., 2013. Use of AIS data to characterise marine traffic patterns and ship collision risk off the coast of Portugal. J. Navig. 66 (6), 879–898.

Vettor, R., Soares, C.G., 2015. Detection and analysis of the main routes of voluntary observing ships in the North Atlantic. J. Navig. 68 (2), 397–410.

Wan, Z., Nie, A., Chen, J., Ge, J., Zhang, C., Zhang, Q., 2021. Key barriers to the commercial use of the Northern Sea Route: view from China with a fuzzy DEMATEL approach. Ocean Coast Manag. 208, 105630.

Wang, H., Wang, S., Meng, Q., 2014. Simultaneous optimization of schedule coordination and cargo allocation for liner container shipping networks. Transport. Res. E Logist. Transport. Rev. 70, 261–273.

Wang, S., Meng, Q., 2012. Sailing speed optimization for container ships in a liner shipping network. Transport. Res. E Logist. Transport. Rev. 48 (3), 701–714.

Wang, X., Li, J., Zhang, T., 2019. A machine-learning model for zonal ship flow prediction using AIS data: a case study in the south atlantic states region. J. Mar. Sci. Eng. 7 (12), 463.

Wang, X., Liu, Z., Yan, R., et al., 2022. Quantitative analysis of the impact of COVID-19 on ship visiting behaviors to ports- A framework and a case study. Ocean Coast Manag. 230, 106377.

Wei, X., Jia, S., Meng, Q., Tan, K.C., 2020. Tugboat scheduling for container ports. Transport. Res. E Logist. Transport. Rev. 142, 102071.

Wen, Y., Sui, Z., Zhou, C., Xiao, C., Chen, Q., Han, D., Zhang, Y., 2020. Automatic ship route design between two ports: a data-driven method. Appl. Ocean Res. 96, 102049.

Wu, X., Mehta, A.L., Zaloom, V.A., Craig, B.N., 2016. Analysis of waterway transportation in Southeast Texas waterway based on AIS data. Ocean Eng. 121, 196–209.

Xiao, Z., Ponnambalam, L., Fu, X., Zhang, W., 2017. Maritime traffic probabilistic forecasting based on vessels' waterway patterns and motion behaviors. IEEE Trans. Intell. Transport. Syst. 18 (11), 3122–3134.

Xu, L., Di, Z., Chen, J., Shi, J., Yang, C., 2021a. Evolutionary game analysis on behavior strategies of multiple stakeholders in maritime shore power system. Ocean Coast Manag. 202, 105508.

Xu, L., Yang, S., Chen, J., Shi, J., 2021b. The effect of COVID-19 pandemic on port performance: evidence from China. Ocean Coast Manag. 209, 105660.
Xu, L., Di, Z., Chen, J., 2021c. Evolutionary game of inland shipping pollution control under government co-supervision. Mar. Pollut. Bull. 171, 112730.
Xu, L., Yang, Z., Chen, J., Zou, Z., 2023. Impacts of the COVID-19 epidemic on carbon emissions from international shipping. Mar. Pollut. Bull. 189, 114730.
Yan, Z., Xiao, Y., Cheng, L., He, R., Ruan, X., Zhou, X., et al., 2020. Exploring AIS data for intelligent maritime routes extraction. Appl. Ocean Res. 101, 102271.
Zhang, M., Montewka, J., Manderbacka, T., Kujala, P., Hirdaris, S., 2021. A big data analytics method for the evaluation of ship - ship collision risk reflecting hydrometeorological conditions. Reliab. Eng. Syst. Saf. 213, 107674.
Zhang, M., Taimuri, G., Zhang, J., Hirdaris, S., 2023. A deep learning method for the prediction of 6-DoF ship motions in real conditions. In: Proceedings of the Institution of Mechanical Engineers, Part M. Journal of Engineering for the Maritime Environment, 14750902231157852.
Zhang, M., Zhang, D., Fu, S., Kujala, P., Hirdaris, S., 2022. A predictive analytics method for maritime traffic flow complexity estimation in inland waterways. Reliab. Eng. Syst. Saf. 220, 108317.
Zhang, S., Chen, J., Wan, Z., Yu, M., Shu, Y., Tan, Z., Liu, J., 2021. Challenges and countermeasures for international ship waste management: IMO, China, United States, and EU. Ocean Coast Manag. 213, 105836.
Zhang, M., Montewka, J., Manderbacka, T., Kujala, P., Hirdaris, S., 2021a. A big data analytics method for the evaluation of ship-ship collision risk reflecting hydrometeorological conditions. Reliab. Eng. Syst. Saf. 213, 107674.
Zhang, M., Montewka, J., Manderbacka, T., Kujala, P., Hirdaris, S., 2021b. A big data analytics method for the evaluation of ship-ship collision risk reflecting hydrometeorological conditions. Reliab. Eng. Syst. Saf. 213, 107674.
Zhang, S., Shi, G., Liu, Z., Zhao, Z., Wu, Z., 2018. Data-driven based automatic maritime routing from massive AIS trajectories in the face of disparity. Ocean Eng. 155, 240–250.