Ocean Engineering 311 (2024) 118969

Contents lists available at ScienceDirect

Ocean Engineering

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

Research paper

From ports to routes: Extracting multi-scale shipping networks using massive AIS data

Ryan Wen Liu [a,b], Shiqi Zhou [a,b], Maohan Liang [c,*], Ruobin Gao [d], Hua Wang [e]

[a] Hubei Key Laboratory of Inland Shipping Technology, School of Navigation, Wuhan University of Technology, Wuhan, China
[b] State Key Laboratory of Maritime Technology and Safety, Wuhan University of Technology, Wuhan, China
[c] Department of Civil and Environmental Engineering, National University of Singapore, Singapore
[d] School of Civil and Environmental Engineering, Nanyang Technological University, Singapore
[e] School of Automotive and Transportation Engineering, Hefei University of Technology, Hefei, China

ARTICLE INFO

Keywords:
Shipping network
Automatic identification system
Vessel trajectory
Vessel behaviour
Feature points

ABSTRACT

Maritime transportation is a critical component of global trade and commerce. To ensure maritime safety, fixed shipping routing has been established in many complex waters. However, there is currently a lack of comprehensive digital shipping networks in wide-range maritime areas. To better understand the navigational patterns, this paper proposes a data-driven extraction framework for multi-scale shipping networks, including port-, node-, and route-level shipping networks. It is essentially a hierarchical approach, which progresses from port to route. In particular, for the extraction of port-level shipping networks, the clustering in quest (CLIQUE) and alpha-shapes algorithms are employed to accurately extract the boundaries and spatial extents of individual ports. For the node-level shipping network extraction, an adaptive Douglas-Peucker algorithm is developed to identify crucial feature points, and CLIQUE clustering is further exploited to extract the network waypoints. A novel slice-based traffic flow fitting algorithm is finally introduced to extract the route-level shipping network. To verify the performance of shipping network extraction methods, comprehensive experiments are conducted using the massive Automatic Identification System (AIS) data in different water areas. The experimental results have demonstrated that our method was capable of extracting multi-scale shipping networks, revealing traffic characteristics and vessel behaviours. Overall, the method proposed herein is useful for shipping logistic analysis and provides a foundation for several potential maritime applications, including route planning, trajectory prediction, and others.

1. Introduction

In the context of economic globalization, maritime trade demands are increasing significantly among countries worldwide with the rapid growth of the global shipping industry (Cullinane and Bergqvist, 2014; Liang et al., 2024b). Maritime transportation, one of the key components of transportation systems, accounts for approximately 90% of the global trade freight and logistics (Gu et al., 2019; Rong et al., 2020). The increasing waterborne traffic flow density can result in a complex navigational environment for vessels, with a high potential risk of traffic accidents (Yu et al., 2024; Ma et al., 2024). Especially in open water area, the navigational environment is less restricted, and the navigation safety issues are more prominent. To ensure the navigation safety, it is necessary to extract the shipping network, revealing the vessels' behaviours and traffic patterns, and provide the fundamental navigation references for vessels. Additionally, maritime shipping network research is also attracting the attention of logistics enterprises. They conduct shipping logistics planning research to improve the efficiency of cargo transportation and reduce transportation costs, making the logistics industry more competitive (Irannezhad et al., 2018; Mason et al., 2003). The information on maritime transport connectivity, routes, and others can be obtained for their decision making (Cheung et al., 2020). Furthermore, it will also help shipping companies implement Green Shipping Practices (GSP) by optimizing vessel navigation strategies to make their operations more environmentally friendly (Duan et al., 2021; Lai et al., 2011). Focusing on the aforementioned perspectives, the research on shipping network extraction has aroused the interests of many scholars, for the further aims of guaranteeing the vessel navigation safety and making maritime transportation more efficient. (Yan et al., 2020a; Arguedas et al., 2014; Yu et al., 2021).

* Corresponding author.
E-mail addresses: wenliu@whut.edu.cn (R.W. Liu), 332421@whut.edu.cn (S. Zhou), mhliang@nus.edu.sg (M. Liang), GAOR0009@e.ntu.edu.sg (R. Gao), hwang191901@hfut.edu.cn (H. Wang).

https://doi.org/10.1016/j.oceaneng.2024.118969
Received 9 May 2024; Received in revised form 6 August 2024; Accepted 7 August 2024
Available online 14 August 2024

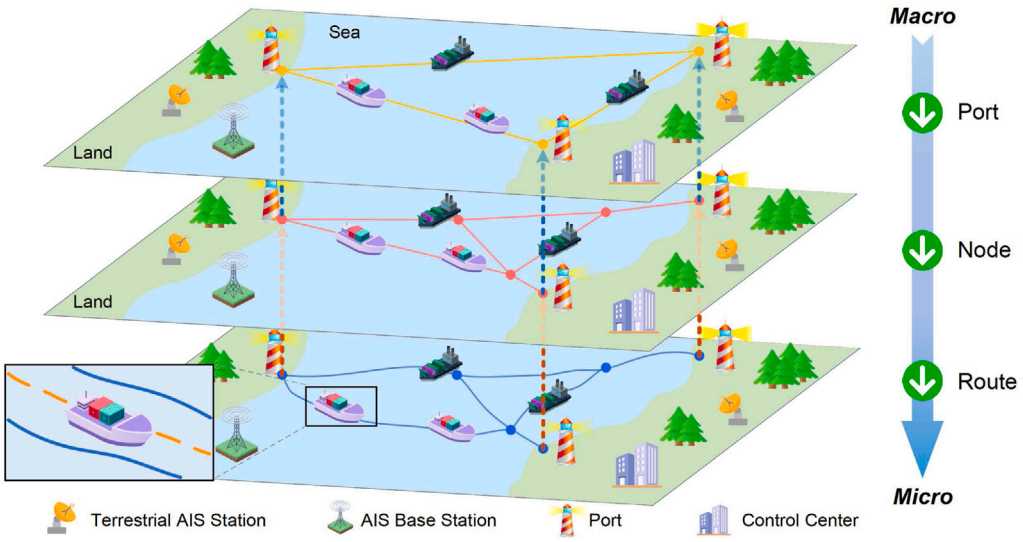


Fig. 1. Overview of maritime transportation from different perspectives. The traffic situation can be characterized from the port level, to the node level which represents the main vessel navigational characteristics, and finally to the route level which characterizes the details of the vessel's navigation.

A shipboard automatic identification system (AIS) transmitter, which records static, dynamic and voyage-related vessel information, must be installed on designated types of vessels to serve maritime management (Liang et al., 2024; Yang et al., 2021b; Yan et al., 2020b; Yang et al., 2019; Liang et al., 2024a). Currently, AIS data, which are widely used globally, are becoming a cornerstone of maritime transportation research, which is an application that is far beyond their original purpose (Liu et al., 2023b). Since AIS data contain general vessel motion patterns, AIS data mining for characterizing vessel behaviours and extracting shipping networks is feasible and meaningful (Rong et al., 2020).

Leveraging massive amounts of AIS data, prevailing studies on shipping network extraction have predominantly centred on identifying trajectory feature points (Cai et al., 2021; Zhang et al., 2018; Zhou et al., 2023; Liang et al., 2024c), port areas (Yan et al., 2022; Arguedas et al., 2014), etc., while simply linking these components to construct maritime networks. Current studies focus only on the extraction of maritime shipping networks at a single scale, which ignores the relationships that precede each network and provides limited information. There is a lack of a comprehensive research framework for shipping network extraction. In fact, maritime transportation can be viewed from different perspectives, as shown in Fig. 1. In particular, from the macro perspective, vessels sail from one port to another port. However, during the navigation, vessel may have to navigate through certain areas that can be abstractly expressed as 'nodes', which reflects the navigation characteristics of the vessels. It is from the meso perspective, an intermediate scale between macro and micro. Furthermore, to ensure safe navigation, the vessels will follow the customary route and remain within a certain lateral range, which can be viewed from the micro perspective. Therefore, taking fully into account macro and micro traffic flow characteristics, a novel approach for extracting multi-scale shipping networks from massive AIS data is proposed in this paper. To comprehensively and accurately describe vessel movement patterns, we discover motion patterns and extract shipping networks at multiple scales, including the port, node, and route levels. The primary contributions of this paper are encapsulated as follows.

- The *port-level shipping network* (PLSN) extraction method is proposed to identify ports and their respective boundaries from massive AIS data. The shipping network is then constructed using the connectivity between these identified ports.
- The *node-level shipping network* (NLSN) extraction method is proposed based on a novel trajectory compression algorithm which can automatically extract the feature points (i.e., starting points, turning points, and ending points) of vessel trajectories. The final nodes of shipping networks can then be determined by clustering the extracted feature points.
- According to the NLSN extraction results, statistical methods are used to identify the customary routes and the corresponding boundaries, leading to the establishment of a *route-level shipping network* (RLSN).
- The effectiveness and robustness of our extraction methods were demonstrated via comprehensive experiments conducted in two maritime areas with massive realistic AIS-based vessel trajectories. Each level of the extracted shipping network has significant potential for vessel behaviour analysis and intelligent maritime surveillance.

The remainder of this paper is organized as follows. Section 2 introduces related works on shipping network extraction. Section 3 introduces the main definitions and problems related to shipping network extraction. Section 4 presents our data-driven method for extracting multi-scale shipping networks. Extensive experiments are presented in Section 5 to evaluate the performance of our extraction methods. This work is concluded by summarizing our main contributions in Section 6.

2. Related works

Many efforts have been devoted to shipping network extraction in recent years. These methods can be mainly categorized into three groups, i.e., statistics-, grid- and vector-based methods (Yan et al., 2020b; Liu et al., 2023a,b).

2.1. Statistics-based shipping network extraction

Statistics-based methods are capable of constructing shipping networks by considering vessel traffic flow characteristics, e.g., traffic volume and speed (Rong et al., 2022). For example, Xiao et al. (2015) obtained the statistical distribution of the speed, course and other characteristics of different types of vessels by analysing information on vessel traffic behaviour. Li et al. (2016) simplified trajectories by applying the popular Douglas-Peucker (DP) algorithm and visualized vessel traffic density using kernel density estimation (KDE). Wen et al. (2016) extracted shipping routes from the massive AIS data using local polynomial regression. The spatial and temporal characteristics were then considered to extract the vessel movement pattern. In addition,

Zhang et al. (2019) analysed the origin-to-destination pairs and navigation routes. The spatial–temporal vessel traffic analyses, reflecting the shipping network, were implemented in the Singapore Strait. Zhang et al. (2022b) presented the vessel trajectory distribution, evaluated the traffic flow complexity, and analysed the accidents' situation of different areas of the Yangtze River. The corresponding results also indicated the shipping network.

2.2. Grid-based shipping network extraction

The grid-based methods project the vessel trajectory points onto the pre-defined grid. It is capable of generating a point map, and then extracting high-density areas from this grid map to generate a shipping network. For example, Vettor and Soares (2015) projected the trajectory point data onto a grid with a cell size 2°×2°, identified the junction points and extracted high-density routes on the basis of point density of each grid change trend. Furthermore, Silveira et al. (2019) projected the AIS data from the research area into grid cells, extracting the core nodes and forming edges on the basis of vessel movements. The most commonly used routes then were identified by applying the Dijkstra algorithm. Comprehensively considering the factors, e.g., vessel size, vessel speed, and grid size, Yang et al. (2021a) projected the vessel density of each grid unit onto a map. The high-density grid cells and the corresponding navigating routes were extracted through the non-parametric KDE method. Furthermore, Zhang et al. (2022a) divided the research region into grids, and identified the channel boundaries and safety domains of vessel sailing for anti-collision applications. Kim et al. (2023) analysed the traffic flow characteristics, including the overall traffic length (LOA), speed over ground (SOG), and frequency from 6 grid sizes. The maritime transportation network could be displayed directly from the visual results.

2.3. Vector-based shipping network extraction

The vector-based methods are used to determine how to extract vessel waypoints and connectivity from trajectory data to construct maritime shipping networks. The pioneering method, extracting the shipping network, was proposed by clustering the vessel trajectories with high similarities. However, it only considered three terminals (ports, entry, and exit points), and neglected the vessel behaviour and traffic flow features related to the vessel trajectories (Pallotta et al., 2013). Many researchers have explored effective clustering methods to extract highly-reliable shipping networks. Wang et al. (2017) defined the distance between vessel routes and clustered the trajectories using the hierarchical method. Sheng and Yin (2018) used the density-based spatial clustering of applications with noise (DBSCAN) algorithm to cluster trajectories based on structural similarity to extract representative trajectories. Huang et al. (2023), who used the DP algorithm to compress trajectories, clustered trajectories from the perspective of spatial distance and course over ground (COG). Liu et al. (2023a) clustered the trajectories via principal component analysis and the K-means algorithm and extracted the route centrelines and boundaries. However, such methods rely heavily on the accuracy of clustering methods. To improve the performance of network generation, feature point extraction-based methods have been proposed, which generated the shipping network by extracting the starting points, ending points, and waypoints. After obtaining the route via the DBSCAN algorithm, Arguedas et al. (2014) extracted the breakpoints to form a more refined maritime network. In addition, Zhang et al. (2018) considered the turning characteristics of vessel traffic flow, which exploited the DP algorithm to extract feature points and the DBSCAN algorithm to cluster them. It could obtain the turning nodes of the route and determine the connectivity of the turning nodes. The shipping network was constructed (Cai et al., 2021) by extracting the key pattern nodes via clustering methods. In particular, these pattern nodes were classified through the K-means algorithm in the open sea passage. In the local sea passage, some important nodes were obtained, using the DBSCAN algorithm, as the representative feature points. Finally, the navigational routes could be extracted for improving the vessel traffic services.

2.4. Summaries

A comprehensive overview of the reviewed literature pertaining to shipping network extraction methods is summarized in Table 1. Statistics-based methods are commonly suitable for analysing traffic flow characteristics and serve as a foundation for informed decision-making but encounter limitations when tasked with the direct extraction of shipping networks. Grid-based methods are usually feasible if the experimental area is not very large. Nevertheless, as the scope of the experimental area broadens, the computational load also increases proportionately. A pivotal consideration in employing grid-based methods is the selection of the grid cell size, which significantly impacts the outcomes of the experiment. This necessitates a deliberate discussion and choice regarding the optimal grid dimensions to ensure the accuracy of the results. In contrast, vector-based shipping network extraction methods perform well in demonstrating vessel motion patterns. Despite their strengths, the literature reveals a common limitation: the majority of existing studies focus on shipping network extraction at a singular scale, often overlooking the multifaceted aspects of the node connectivity among them. This summary underscores the need for a more nuanced approach that encompasses multiple scales, particularly to unveil the intricate micromotion patterns within maritime traffic flows.

3. Definitions and problem statements

3.1. Definitions

Definition 1 (*Vessel Trajectory*)**.** The mth trajectory T_m in vessel trajectory dataset T can be considered as time-sequence points $p_n = \{lon_n, lat_n, sog_n, cog_n, status_n, time_n\}$, where lon_n and lat_n denote the vessel's longitude and latitude at time $time_n$, sog_n and cog_n denote the vessel SOG and COG, respectively, and $status_n$ denotes the vessel's navigation status.

Definition 2 (*Shipping Network*)**.** Due to maritime traffic planning and regulations, vessel mobility and waterway characteristics, waterways, waypoints and vessel movement patterns have common characteristics for sailing vessels (Xiao et al., 2019). These elements are extracted to form a shipping network to represent a large number of real-world maritime routes. By adopting graph theory, the network can be illustrated as $G = (V, E)$, where E is the set of edges and V denotes the vertices of the graph G.

Definition 3 (*Port-Level Shipping Network (PLSN)*)**.** It refers to the shipping network composed of the identified ports (the nodes of the shipping network) and their transportation connections between them. If there are vessels navigating between 2 ports, it can be considered that there exists a connection relationship and the corresponding edge of the shipping network is formed. Similarly, the network can be illustrated as $G_P = (V_P, E_P)$, where E_P represents the connections between the ports and V_P represents the sets of the ports.

Definition 4 (*Node-Level Shipping Network (NLSN)*)**.** It refers to the shipping network composed of several nodes (namely the waypoints that the vessels pass through during their navigation) extracted from vessel trajectory feature points and their connections, and described the main movement characteristics of the vessel during navigation. Similarly, the network can be illustrated as $G_N = (V_N, E_N)$, where E_N represents the connections between the feature points and V_N represents the sets of the feature points.

Definition 5 (*Route-Level Shipping Network (RLSN)*)**.** It refers to the shipping network formed by the customary routes of vessel navigation, which describes the detailed characteristics of the vessel's navigation process. The network, displayed as $G_R = (V_R, E_R)$, is formed based

Table 1
Summary of related works reviewed.

Category	Method	Data	Element(s)	Node(s)	Reference
Statistics	Statistical distributions	Dutch & China	R	/	Xiao et al. (2015)
	KDE	China	R	/	Li et al. (2016)
	Local polynomial regression	Singapore Strait	R	/	Wen et al. (2016)
	Statistical analysis	Singapore	N & R	Origin & destination	Zhang et al. (2019)
	Basic statistic	China	R	/	Zhang et al. (2022b)
Grid	Density visualization	North Atlantic	R	/	Vettor and Soares (2015)
	Dijkstra algorithm	Portugal	R	/	Silveira et al. (2019)
	KDE	China	R	/	Yang et al. (2021a)
	KDE	China	R	/	Zhang et al. (2022a)
	Density visualization	Korea	R	/	Kim et al. (2023)
Vector	Clustering	North Adriatic Sea	N & R	Entry/exit & port	Pallotta et al. (2013)
	Clustering	Dover Strait	N & R	Entry/exit, port & breakpoint	Arguedas et al. (2014)
	Clustering	China & Germany	R	/	Wang et al. (2017)
	Clustering	China	R	/	Sheng and Yin (2018)
	Clustering	South China Sea	R	/	Huang et al. (2023)
	Clustering	Portugal	R	/	Liu et al. (2023a)
	Generate nodes & their connectivity	China	N & R	Waypoints (include turning points)	Zhang et al. (2018)
	Generate nodes & their connectivity	Pacific Ocean & Atlantic Ocean	N & R	Waypoints (include pattern nodes & connection points)	Cai et al. (2021)

Notes: The column "Element(s)" indicates the key elements forming the shipping network, where R and N represent the route and node respectively. The column "Node(s)" indicates the types of points that are extracted as nodes of the shipping network.

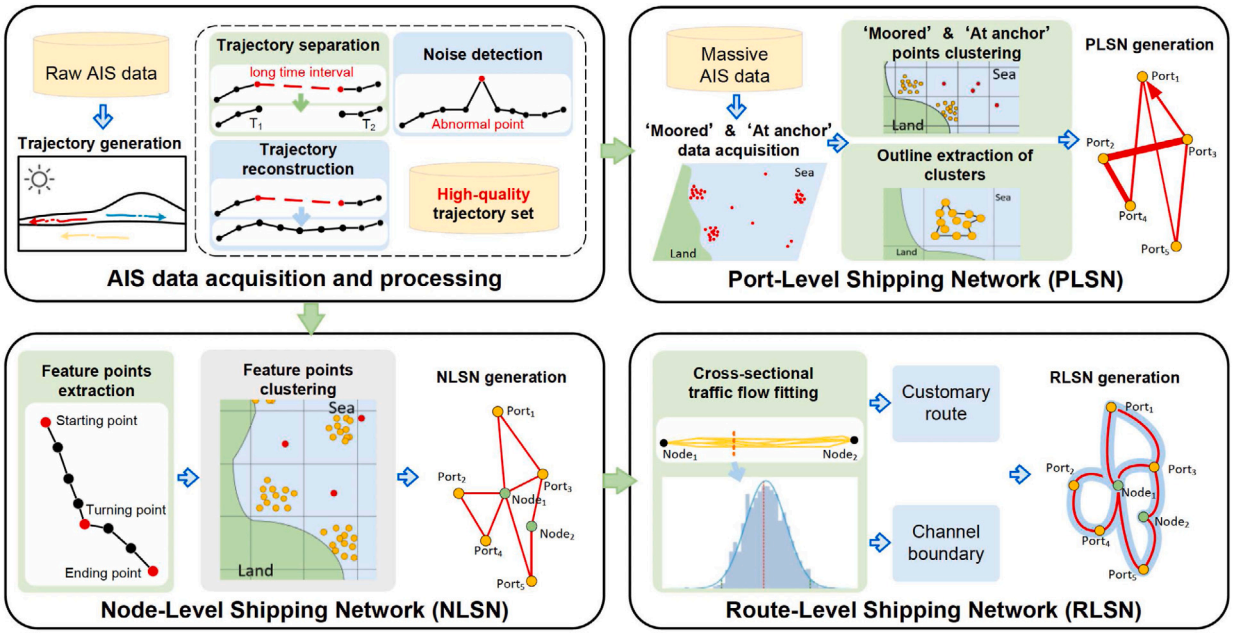


Fig. 2. Flowchart of multi-scale shipping network extraction. Both Port-level shipping network and Node-level shipping network are constructed based on the processed AIS data. Based on the Node-level shipping network, the microscopic features of vessel navigation between nodes are mined to form the Route-level network.

on NLSN, but what is different from the other two networks is that E_R is composed of extracted vessels' customary routes and channel boundaries. The differences among the multiple shipping networks are clearly demonstrated in Fig. 3.

3.2. Problem statements

Problem 1. How to extract the Port-Level Shipping Network (PLSN)?

Vessels in the port area usually are within two main areas: port anchorage and port berths (Yan et al., 2022). This vessel-specific status information can be obtained from the navigation status contained in AIS data, and significant collections of mooring and anchoring points will be identified by analysing the massive 'moored' and 'at anchor' information from AIS data (Martinčič et al., 2020). Virtual berths considered ports can be found by data mining, and a useful method is clustering. For common clustering methods, density-based methods have a high computational cost, rendering them less feasible for processing in large-scale space (Liu et al., 2013), while partition-based methods are simple but cannot find non-convex shape clusters (Xing et al., 2023). However, grid-based methods stand out for their efficiency in handling large amounts of multidimensional data. Thus, the clustering in quest (CLIQUE) clustering algorithm is applied in this paper, and clusters are obtained. Then, the outliers of each cluster are extracted using the widely recognized alpha-shapes algorithm. Finally, the connectivity among ports is determined, forming the PLSN.

Problem 2. How to extract the Node-Level Shipping Network (NLSN)?

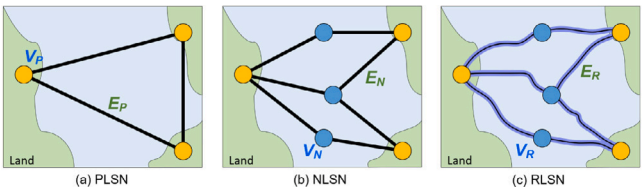


Fig. 3. The schematic diagram of the extracted shipping networks at different scales. PLSN, NLSN, and RLSN denote the port-level shipping network, node-level shipping network, and route-level shipping network, respectively.

The trajectory points where the vessel's navigation behaviour changes significantly are the feature points in the steering area corresponding to the trajectory (Li et al., 2016). The starting and ending points (the first and last points of one trajectory sequence during a vessel's voyage) also belong to the feature points. Feature points are important for constructing NLSNs. Among the existing methods for extracting feature points, the DP algorithm is a common and effective method. However, the selection of the compression threshold has a great impact on the compression effect, making it necessary to improve the algorithm (Tang et al., 2021). By combining the trajectory features, an adaptive DP algorithm without a manually set threshold for compressing vessel trajectories is designed to overcome the above issue. After obtaining the feature points, the CLIQUE algorithm is applied again for clustering, and the NLSN is generated according to the connectivity among the extracted nodes.

Problem 3. How to extract the Route-Level Shipping Network (RLSN)?

To ensure vessel navigation safety and decrease the cost of transportation, vessels usually sail on customary routes. From the perspective of mathematical statistics, massive vessel trajectories should present a normal distribution in the cross-section perpendicular to the channel. However, some abnormal trajectories need to be identified and removed, and the three-sigma rule is useful for this kind of mission. Based on this idea, the channel boundaries and the customary waterways can be successfully extracted, forming an RLSN.

4. Methodology

The ports, nodes and routes, which exist in the real world, are essential parts of the virtual shipping network. To robustly and accurately analyse the shipping information, we propose an automatic method for extracting multi-scale shipping networks from three different levels, illustrated in Fig. 2. The PLSN performs construction network operations based on so-called ports extracted from AIS data with 'At Anchor' and 'Moored' statuses. To further understand the navigation characteristics of vessels, an NLSN is formed on the basis of feature points in trajectories. In addition, the constructed RLSN considers more trajectory characteristics from a point of statistical distribution. The details on how to extract these three networks from massive AIS data will be discussed in this section.

4.1. Port-level shipping network extraction

Vessels are usually moored or anchored in and around ports and the navigation status information can be obtained from AIS data, making the port identification possible by data analysis (Yang et al., 2024). For those outer anchorages located far from the coast, they can be also viewed as 'virtual ports'. Based on this assumption, the PLSN construction is intricately grounded on the port identification by analyzing the AIS data that reflect the vessel navigation status. First, this paper employs statistical methods to distill effective vessel anchoring and mooring data from the vessel's static information. Subsequent steps include the application of the CLIQUE clustering and alpha-shapes algorithms to delineate the extent of ports, and a PLSN is formed based on the identified ports and their connectivity. The PLSN reflects the relationships and connections between macro ports and is an important means to analyse the macro behaviour of vessels. The pseudocode of the PLSN extraction is shown in Algorithm 1.

Algorithm 1 PLSN extraction

Input: Trajectory point set P; Trajectory set T; CLIQUE algorithm parameters (the number of grid divisions K and the density threshold r)
Output: Extracted PLSN

```
For p in P: // Obtain the points with navigation status 'at anchor' and 'moored'
if (p['navigation status'] !=1 or p['navigation status'] !=5) and p['SOG'] >0.5 then
    Delect p
end if
Clusters=CLIQUE(P, K, r) // Cluster the points in P
for cluster in Clusters do
    Boundary[cluster_index] = alphashape(cluster, 0.01) // Get port boundaries
end for
for t in T do
    Obtain the traffic volumes among ports
end for
Generating PLSN
return PLSN
```

4.1.1. Mooring/anchoring data clustering

The navigation status information used to obtain the mooring and anchoring data in AIS data is represented by numbers from 0 to 15. In particular, if the navigation status in the AIS data is '0', the vessel is 'underway using engine'. '1' indicates the vessel is at anchor, and '5' indicates that the vessel is moored (Sturgis et al., 2024). Therefore, extracting AIS data with navigation status values of 1 and 5 can help to extract mooring and anchoring points.

However, the static information of AIS data is manually input by operators, and operators can enter incorrect information or fail to enter information due to negligence (Yang et al., 2021b), which leads to noise in the navigation status information. To obtain high-quality mooring data and anchoring data, constraints are imposed on vessel trajectory data in this work:

$$P_{anchoring} = (lat, lng, sog, status), \quad if \quad status == 1 \quad and \quad sog < 0.5 \tag{1}$$

$$P_{mooring} = (lat, lng, sog, status), \quad if \quad status == 5 \quad and \quad sog < 0.5 \tag{2}$$

The anchoring points are obtained by Eq. (1), and the mooring points are obtained by Eq. (2). A vessel SOG less than 0.5 is used to improve the quality of the results. However, the points that are initially obtained usually include noise and the data characteristics cannot be automatically obtained. The CLIQUE clustering algorithm is used to cluster the data to remove erroneous AIS data from the processed data and to obtain each cluster.

Agrawal et al. (1998) proposed the CLIQUE algorithm to achieve clustering in high-dimensional data. As a grid-based spatial clustering algorithm combining the advantages of the dense-based clustering algorithm, the main advantages of the CLIQUE algorithm include scalability, the ability to uncover clusters in high-dimensional spaces, and its notable interpretability. Additionally, the algorithm is independent of the data input order and does not necessitate the assumption of a particular probability distribution within the dataset. Unlike other density-based clustering algorithms, the CLIQUE algorithm does not consider the entire high-dimensional space, so it can more effectively

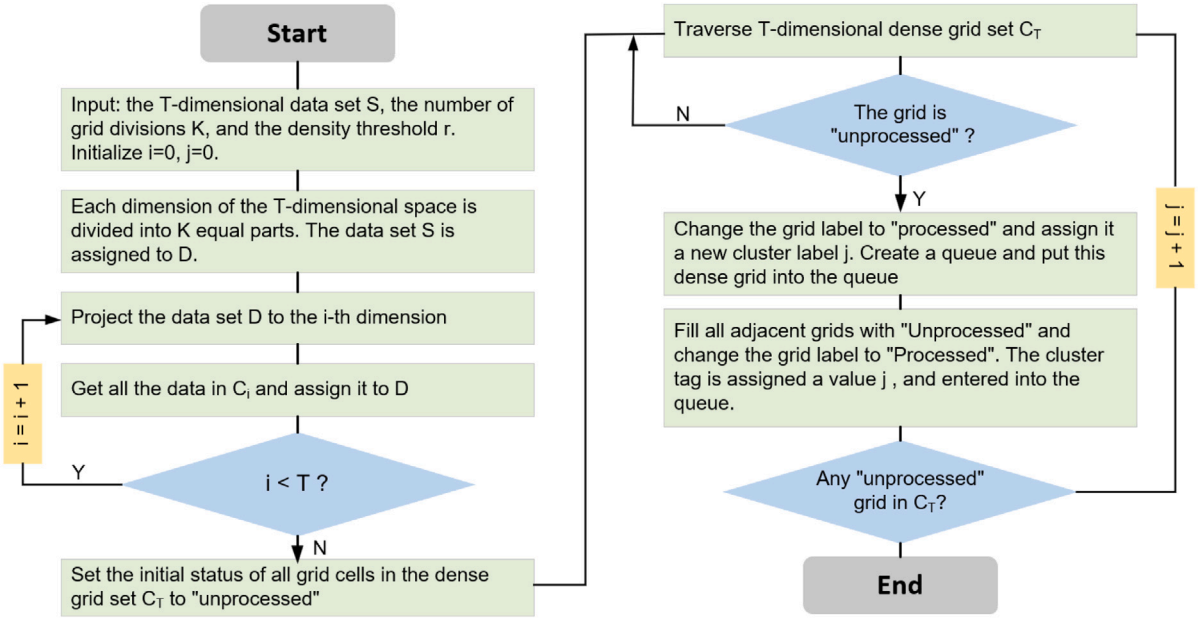


Fig. 4. Flowchart of the CLIQUE algorithm, which is employed for efficient clustering of trajectory points.

discover highly dense areas and extract subspace clusters in the original data space.

The CLIQUE algorithm divides the data space into non-overlapping rectangular cells and determines whether the space is dense based on the number of points in the cell. A cell is considered dense if the number of all points it contains exceeds a certain model parameter. The largest set of densely connected units is considered a cluster. In summary, the overall purpose of employing the CLIQUE algorithm is to identify sparse and dense regions (or units) in space in a multi-dimensional dataset, thereby revealing the overall distribution pattern of the dataset.

The flowchart of the CLIQUE algorithm is displayed in Fig. 4. The geographical position information (longitude and latitude) is considered and extracted from the above anchoring and mooring points. After this, the clustering process of the CLIQUE algorithm is divided into three steps:

(1) Identify the subspace containing clusters

To improve the calculation efficiency, the CLIQUE algorithm adopts a bottom-up calculation method to reduce the search space through the uniqueness of dense dimensions. It is noteworthy that the dense part of a point set S in k-dimensional space is also dense in $(k-1)$-dimensional space. If the units of the point set are connected in the k dimension, then these point set units are still connected in the $(k-1)$ dimension.

The CLIQUE algorithm traverses all the data to determine the dense units in one-dimensional space and then generates a candidate set of k-dimensional dense units based on the determined $(k-1)$-dimensional dense units. The algorithm terminates when no new candidate sets are generated.

An increase in the subspace dimension will cause rapid growth of dense units. An algorithm must be used to retain the dense units of interest and prune some unqualified candidate sets. This process can be achieved by pruning the minimal description length (MDL). The essence of the MDL concept is to encode the input data according to a specific pattern to make the code the shortest.

In a subspace set $\{S_1, S_2, \ldots, S_n\}$, the pruning method calculates the number of records contained in each subspace:

$$x_{s_j} = \sum_{u_i \in s_j} \mathrm{C}\left(u_i\right), u_i \in s_j \tag{3}$$

where $\mathrm{C}(u_i)$ denotes the number of points in u_i and x_{s_j} is the cover of subspace s_j.

The CLIQUE algorithm arranges the subspaces from high to low according to the degree of coverage and then divides the subspaces into two sets: the selected subspace set R and the pruned subspace set P. For each set, the average area coverage is calculated, and the difference between each subspace in the set and the average is calculated. Finally, the number of digits required for the calculation is added to obtain the objective function encoded by the CLIQUE algorithm:

$$\begin{aligned} CL(i) = & \log_2\left(\mu_I(i)\right) + \sum_{1\le j\le i} \log_2\left(\left|x_{S_j} - \mu_I(i)\right|\right) \\ & + \log_2\left(\mu_P(i)\right) + \sum_{i+1\le j\le n} \log_2\left(\left|x_{S_j} - \mu_P(i)\right|\right) \end{aligned} \tag{4}$$

where $\mu_I(i)$ and $\mu_P(i)$ can be calculated respectively as $\mu_I(i) = \left(\sum_{1\le j\le i} x_{s_j}\right)/i$, $\mu_P(i) = \sum_{i+1\le j\le n} x_{s_j}/(\mathrm{n}-i)$. One i value is determined to minimize the objective function $CL(i)$, and the i value is the dividing point that the CLIQUE algorithm needs to find.

(2) Identify these clusters

The input of the CLIQUE algorithm is a set ψ of dense units, which are all in the same k-dimensional space S. The output is a partition of ψ, and each partition contains only connected units. There is no unit connecting multiple partitions.

Clustering can be viewed as the process of finding connected subgraphs in a graph, where the vertices represent dense units. If two dense cells are coplanar, then there is an edge between their corresponding vertices in the graph. Thus, the corresponding dense units are in the same cluster if two vertices are in the same connected subgraph, while they are not in the same cluster if two vertices are not in the same connected subgraph.

To find connected subgraphs in a graph, the depth-first search (DFS) algorithm can be used. Starting from any unit u in ψ, find all units connected to u. If there is any unvisited node in ψ, select another unit and repeat the process.

(3) Generate the "minimum description"

This step accepts a set of units that should be connected together in k-dimensional space but are separated due to the operation. Each unit represents a cluster, so the purpose of this step is to generate a concise description of each cluster. To generate the minimum description of each cluster, the CLIQUE algorithm needs to find all the units that make up the cluster and contain only the minimum number of connected units. Assuming that R is a set in the same subspace S, the Z coverage

condition is that every $R \in Z$ is included in Z, and any unit in Z is included in at least one R. The algorithm consists of two parts: finding the coverage of the largest area and then finding the minimum coverage. In step one, the CLIQUE algorithm uses a greedy strategy: the rectangular bundles with the maximum number are used for coverage clustering. In step two, the CLIQUE algorithm finds the minimum coverage by removing redundant rectangles.

4.1.2. Mooring/anchoring boundary extraction

Based on the CLIQUE algorithm, clusters of anchors and moorings can be obtained, but the anchors and berths represent a physical space range. In this paper, the alpha-shapes algorithm is applied to extract the outline of the berth and anchorage area. Specifically, the outline is formed by the outermost trajectory points in each cluster, and the method of boundary extraction is fully driven by data. The alpha-shapes algorithm is a simple and effective algorithm for quickly extracting boundary points proposed by Edelsbrunner and Mücke (1994). This approach overcomes the influence of the shape of point cloud boundary points and can quickly and accurately extract boundary points. Its principle is shown in Fig. 5. For a two-dimensional point cloud of any shape, if a circle with a radius α rolls on its surface, the points formed by its rolling trajectory are boundary points. There is no point within the rolling circle (Liao et al., 2021). Identifying the contour points is essential, and the steps are summarized as follows:

(1) For a point $P(x, y)$ to be judged in the dataset, search for other points whose distance from P is less than 2α; these points are gathered and recorded as a point set S;

(2) Randomly select one point $P_1(x_i, y_i)$ from S. Form two circles with radii α through P_1 and P and calculate the two circle centre coordinates $o_1(x_{o1}, y_{o1})$ and $o_2(x_{o2}, y_{o2})$, as shown in Fig. 5(b). The calculation process of the circle centre coordinates is as follows:

$$\begin{cases} x_{o1} = x + \frac{1}{2}(x_i - x) - \beta * (y_i - y) \\ y_{o1} = y + \frac{1}{2}(y_i - y) - \beta * (x - x_i) \\ x_{o2} = x + \frac{1}{2}(x_i - x) + \beta * (y_i - y) \\ y_{o2} = y + \frac{1}{2}(y_i - y) + \beta * (x - x_i) \end{cases} \tag{5}$$

with $\beta = \sqrt{\frac{\alpha^2}{d^2} - \frac{1}{4}}$, where $d^2 = (x - x_1)^2 + (y - y_1)^2$.

(3) For points in the point set S other than point P_1, calculate the distances to the centres o_1 and o_2. If the distance from all points to o_1 or o_2 is greater than α, P is a contour point, and the judgement process of this point is terminated;

(4) If the distances from points other than P_1 to the circle centres o_1 and o_2 are not all greater than α, traverse all points of the point set S and recalculate the circle centre coordinates using them as point P_1. Then, follow step 3 to determine point P. If there is a point such that P is marked as a contour point, then P is a contour point; otherwise, P is determined as a non-contour point.

4.1.3. PLSN generation

Previous work has extracted the anchorages and berths of ports based on vessel anchoring and mooring data. To further construct a port-level shipping network, this article determines the connections between identified ports by drawing on historical vessel trajectory data. If there are many vessels having frequently sailed between two identified ports, the corresponding edge of the shipping network is formed. Notably, this research considers a vessel to have stopped at a port if it stays anchored for a prolonged duration, with this paper setting the threshold at over 5 h.

4.2. Node-level shipping network extraction

Due to geographical, navigational and other environmental constraints, in order to ensure the safety of navigation, vessels usually navigate through certain specific areas that can be abstractly expressed as nodes. To detailedly consider the vessels' navigational characteristics, this section introduces a method for the construction of a NLSN. Specifically, an adaptive DP algorithm is proposed, which can iteratively select the optimal compression evaluation index, facilitating the extraction of trajectory feature points of vessels. The CLIQUE clustering algorithm is also introduced to cluster these extracted feature points and determine waypoints (nodes of NLSN). These processes collectively lay the groundwork for NLSN formation. The pseudocode detailing the NLSN extraction process is provided in Algorithm 2.

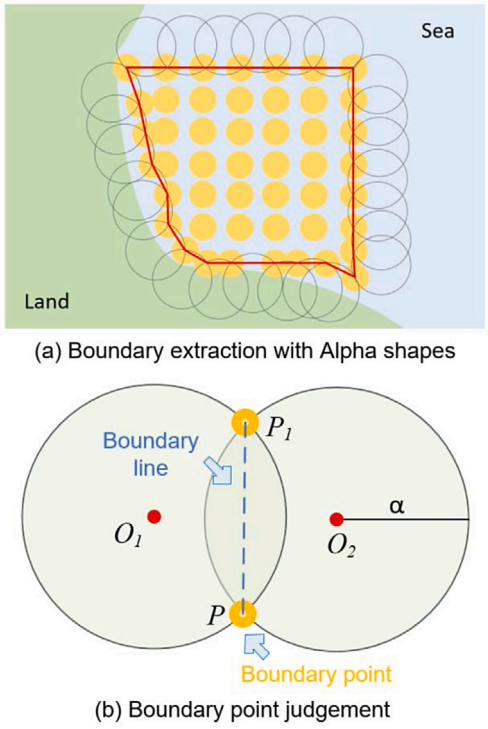


Fig. 5. The principle of the alpha-shapes algorithm. From up to down: (a) schematic diagram of using alpha-shapes algorithm, where yellow dots represent trajectory points, and black circles represent the rolling circles. (b) method of determining boundary points.

Algorithm 2 NLSN extraction

Input: Trajectory set T; CLIQUE algorithm parameters (the number of grid divisions K and the density threshold r)
Output: Extracted NLSN

```
Feature_pointset = list()
for t in T do
    Feature_points = Adaptive_DP(t) //Obtain the feature points
    Feature_pointset.extend(Feature_points)
end for
Clusters = CLIQUE(Feature_pointset, K, r) // Cluster the feature points
for t in T do
    Obtain the connectivity among the clusters (nodes) in Clusters
end for
Generating NLSN
return NLSN
```

4.2.1. Adaptive DP algorithm for the extraction of feature points

The essence of the traditional DP compression algorithm is to approximate the traditional trajectory using the line cut method (Douglas and Peucker, 1973; Liang et al., 2022). The compressed trajectory is topologically consistent with the traditional trajectory and retains the neighbourhood characteristics in the trajectory. Generally speaking,

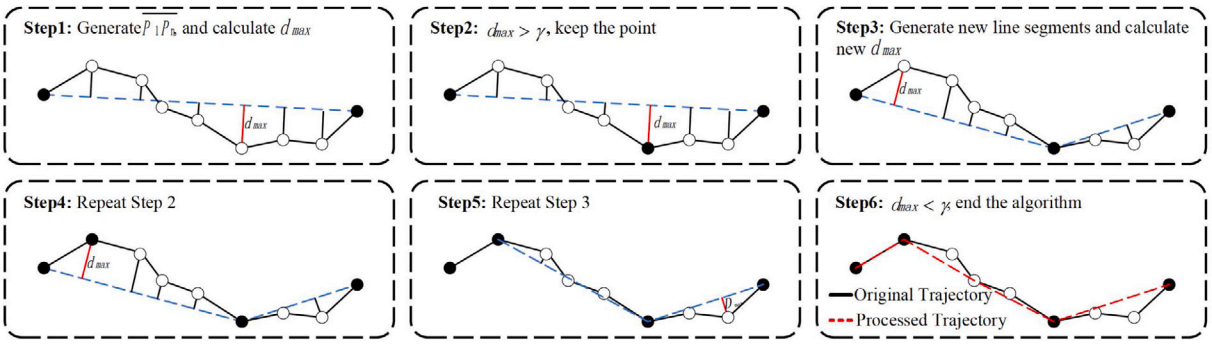


Fig. 6. An example of the traditional DP algorithm. The hollow circles represent the original trajectory points, the solid circles represent the kept trajectory points, the black line represents the original trajectory, and the red line represents the compressed trajectory.

the remaining points defined as feature points of a trajectory are the starting point and the ending point (i.e. the first and last one points), and the turning points. The feature points extracted from massive vessel trajectories indirectly reflect vessels' sailing habits, such as where most vessels depart, where they turn and where they arrive. The feature point extraction process of this method involves translation and rotation invariance, and the compression result is certain under the condition of a given trajectory and threshold. The process of the traditional DP compression algorithm can be summarized as follows, and an example of the algorithm's application is depicted in Fig. 6.

Assume that the traditional trajectory T is expressed as $T = (P_1, P_2, \ldots, P_i, \ldots, P_n)$, including multiple line segments $\overline{P_1P_2}$, $\overline{P_2P_3}$, $\ldots$, and $\overline{P_{n-1}P_n}$. Thus, the traditional DP algorithm flow is as follows:

(1) Mark the start point P_1 and the end point P_n, and connect the 2 points into line segment $\overline{P_1P_n}$. In addition, the compression threshold γ is set.

(2) All trajectory points in the original trajectory T are traversed, the Euclidean distance between each point and the line segment is calculated, and the maximum value d_{max} is obtained. Keep the point P_m corresponding to the maximum distance d_{max} as the segmentation point if $d_{max} > \gamma$. Otherwise, delete all points between P_1 and P_n.

(3) When a new point P_m is retained, divide the line segment $\overline{P_1P_n}$ into $\overline{P_1P_m}$ and $\overline{P_mP_n}$, and calculate the distance from all points between P_1 and P_m to $\overline{P_1P_m}$. If the maximum distance is $d_{max} > \gamma$, retain the point corresponding to the maximum distance; otherwise, delete all points between P_1 and P_m. The same treatment method is used for line segments $\overline{P_mP_n}$.

(4) Repeat step 3 until all the points are processed.

However, the threshold γ must be predefined by the user to simplify the trajectory. Selecting a threshold for each vessel trajectory is also complex and difficult since it is directly related to different vessel trajectories and navigation areas. Different thresholds should be selected for different trajectories to obtain the optimal compression rate and ensure the minimum distance loss. To solve this problem, a method for selecting the trajectory compression rate through iterative adaptation is proposed in this paper. The pseudocode of the improved adaptive DP algorithm is shown in Algorithm 3. First, a trajectory compression evaluation index is designed, and then the optimal compression evaluation index is obtained based on the iterations. The compression rate D_r and distance similarity rate D_l of the vessel trajectory can be expressed as follows:

$$D_r = 1 - \frac{n}{N} \tag{6}$$

$$D_l = 1 - \frac{DL}{dist} \tag{7}$$

$$DL = \|dist - dist_{ori}\| \tag{8}$$

where n is the number of trajectory points after compression, N is the number of trajectory points before compression, dist represents the total length of the trajectory after compression, and the distance loss DL is calculated by $dist$ and the length of the original trajectory $dist_{ori}$. An evaluation index L_D to evaluate the performance of DP compression is proposed and expressed as:

$$L_D = w_1 \times D_r + w_2 \times D_l \tag{9}$$

where w_1 and w_2 are two weight parameters, which are set to 1 in many experiments. As the threshold γ of the DP compression algorithm increases, the compression rate of the vessel trajectory gradually increases, and the distance loss gradually increases. The maximum D_l and D_r are expected to be obtained in this paper.

4.2.2. Feature point clustering for node extraction and NLSN generation

During a vessel's voyage, vessels navigate routes adhering to local legal mandates, aiming for the shortest sailing distance while accounting for traffic environmental constraints. Consequently, despite the independent nature of vessel trajectories, their behavioural patterns exhibit similarity and regularity in their routing. Accordingly, the feature points within vessel trajectories are likely to be spatially clustered. The clustering analysis method enables the extraction of traffic nodes critical for the construction of a shipping network. To overcome the influences of bias and noise in feature points on the construction of shipping networks, the CLIQUE clustering algorithm is also introduced to distinguish outlier points and normal feature points, and the centroid of normal feature points within the same cluster is designated as a waypoint after clustering.

Similarly, the NLSN is constructed by connecting the identified waypoints and high-frequency route segments.

Algorithm 3 Adaptive DP algorithm

Input: Trajectory T, composed of points P_1, P_2, $\ldots$, P_i, $\ldots$, P_n
Output: Point set T', after compression
1: Initialize compression threshold γ
2: Iterative step length s
3: Initialize evaluation indicators $L_D^1 = 0, \quad L_D^2 = 0$
4: $i = 0$
5: **while** $L_D^2 - L_D^1 < 0$ **do**
6: $\quad L_D^1 = L_D^2$
7: $\quad T'[i] = DP(T, \quad \gamma)$ // Traditional DP algorithm is applied.
8: $\quad L_D$ is calculated and gotten
9: $\quad \gamma = \gamma - s$
10: $\quad L_D^2 = L_D$
11: $\quad i+ = 1$
12: **end while**
13: **return** $T'[i-1]$

4.3. Route-level shipping network extraction

The RLSN refers to the microscopic vessel route and reflects the vessel path that vessels are accustomed to choosing when sailing.

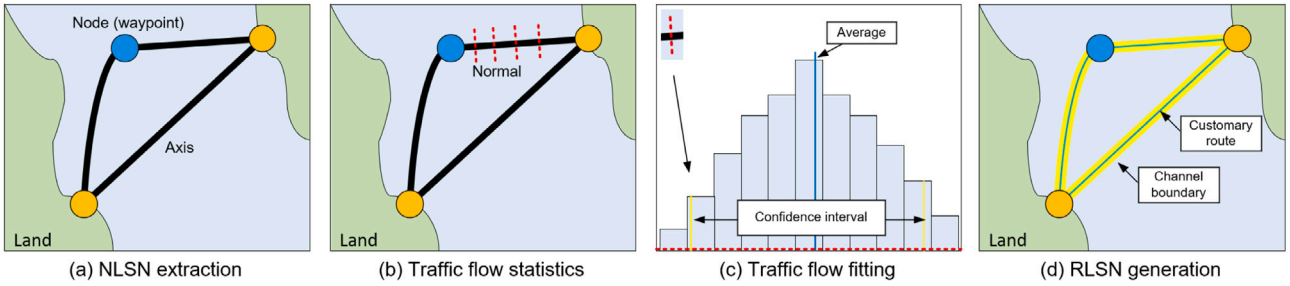


Fig. 7. The flowchart of route-level shipping network generation. From left to right: (a) the extracted local Node-level shipping network, (b) making n normals on a line segment as cross-sections for traffic flow statistics, (c) using the Gaussian function for fitting the traffic flow of each cross-section, and (d) Route-level shipping network generation.

This path is affected by vessel navigation rules and environmental restrictions. In this paper, the extracted NLSN is used to obtain the detailed RLSN, and the process is shown in Fig. 7.

Specifically, each edge within the NLSN is selected and designated as an axis, with n normals to this axis chosen to serve as the cross-sections for traffic flow analysis. For each of these cross-sections, the cross-section itself is treated as the X-axis, and the frequency of traffic flow is represented on the Y-axis. The number of vessels passing through this section within a specific time interval is counted to obtain an accurate distribution of the channel traffic flow. The steps for conducting the cross-sectional traffic flow analyses are detailed as follows.

Two nodes $A(x_1, y_1)$ and $B(x_2, y_2)$ corresponding to the edge in NLSN are selected to obtain the axis between $\overline{AB}$:

$$f(x) = \frac{y_1 - y_2}{x_1 - x_2}(x - x_1) + y_1, \quad x_1 < x < x_2 \tag{10}$$

The channel section is defined as the normal to the node axis. A certain channel section is randomly selected, and the mathematical representation of this section can be succinctly described by its abstract equation as follows:

$$y = -\frac{1}{f'(\omega)}(x - \omega) + f(\omega) \tag{11}$$

with $\omega = x_1 + \frac{|x_1 - x_2|c}{n}, c \in (1, N)$.

By substituting the vessel AIS point (lat, lng) into the channel section abstract equation, we can obtain:

$$\mathrm{Tmp} = -\frac{1}{f'(\omega)}(\mathrm{lat} - \omega) + f(\omega) - lng \tag{12}$$

If $Tmp > 0$, the point is on the left side of the section; otherwise, it is on the right side of the section; calculate $I = Tmp_i \times Tmp_{i+1}$, $i \in (1, N-1)$, where N is the number of vessel AIS points. If $I < 0$, the trajectory passes through this traffic flow section. The intersection point of each vessel is obtained with the solution of the section equation to obtain the cross-sectional vessel traffic flow. The coordinates $(x_n, y_n)(n = 1, 2, 3 \ldots)$ of the obtained cross-section traffic flow statistical histogram are described by the Gaussian function as:

$$y_n = y_{\max} \times \exp\left[-\frac{(x_n - x_{\max})^2}{S}\right] \tag{13}$$

In the formula, the parameters to be estimated x_{max}, y_{max} and S are the peak value, peak position and half-width information of the Gaussian curve, respectively. Taking the natural logarithm of both sides of the formula, we can obtain:

$$\ln y_n = \left(\ln y_{\max} - \frac{{x_{\max}}^2}{S}\right) + \frac{2x_n x_{\max}}{S} - \frac{{x_n}^2}{S} \tag{14}$$

If we set $\ln y_n = z_i$, $\ln y_{\max} - \frac{{x_{\max}}^2}{s} = b_0$, $\frac{2x_{\max}}{s} = b_1$, and $-\frac{1}{s} = b_2$ and consider all the fitted data, the above formula can be expressed in matrix form as:

$$\begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 & {x_1}^2 \\ 1 & x_2 & {x_2}^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & {x_n}^2 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} \tag{15}$$

$$Z = XB \tag{16}$$

The least squares solution to construct the matrix is given by:

$$B = \left(X^T X\right)^{-1} X^T Z \tag{17}$$

The fitting result of the traffic flow section can be obtained as follows:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right) \tag{18}$$

where μ is the mean, σ is the variance and $f(x)$ is the Gaussian function of fitting traffic flow.

Upon deriving n traffic flow cross-section fitting outcomes, this paper uses the average positions of each traffic flow section fitting as the route points and connects them in sequence to form a customary route. The boundaries of these routes are delineated using the 3σ boundary derived from the fitting results. To enhance the precision of the extracted customary routes and channel boundaries, this paper uses mean filtering to smooth the extracted customary routes and boundaries so that the extracted results are more consistent with actual navigation habits.

5. Experimental results and analysis

In this paper, the Bohai Sea and Zhoushan waters in China are selected as typical research areas. The original vessel trajectories we collected were sourced from the global AIS data for March 2018, and only specific information (navigation status, longitude, latitude, etc.) was extracted from the raw AIS data for the experiment. The statistical information is summarized and presented in Table 2. Additionally, the results of overlaying the AIS data on a geographical map are depicted in Fig. 8. The traffic flows in all the study areas exhibit distinctive characteristics that warrant further analysis.

Due to vessel equipment failure, crew negligence, environmental disturbances, and other factors, there are many errors, such as noise data and missing data. To improve the trajectory quality and the final results, the original trajectories are preprocessed, which includes identifying noise data and missing data and reconstructing the trajectories. Noise data are pinpointed and removed according to the principles of vessel manoeuvrability. For the segments of missing data, polynomial interpolation is utilized to reconstruct the vessel trajectories, thereby ensuring the integrity and reliability of the data for subsequent analysis.

Table 2
Statistics of the two research areas in experiments.

Area	Period	Longitude range	Latitude range	Trajectory quantity	Point quantity
Bohai Sea	March, 2018	117° E–125° E	37° N–41° N	12,997	10,818,603
Zhoushan waters	March, 2018	122° E–122.3° E	29.45° N–30° N	3,694	1,796,404

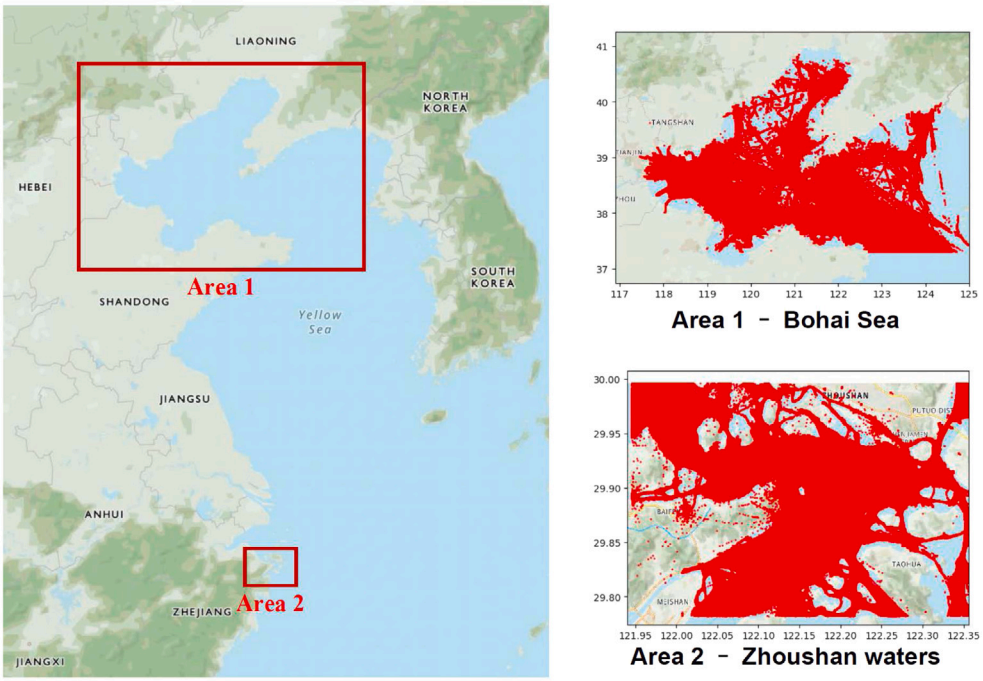


Fig. 8. Results of AIS data visualization for each research area, i.e., (a) Bohai Sea, (b) Zhoushan waters.

5.1. Extraction results of PLSN

The foundation of constructing a PLSN lies in effectively extracting AIS data labelled with navigation statuses 'At Anchor' and 'Moored'. After data filtering, 1,832,508 points exhibiting these statuses were obtained from the Bohai Sea, and 1,469,945 points were obtained from the Zhoushan waters. Given the computational challenges posed by the sheer volume of extracted points for subsequent analysis, the grid division strategy is applied to divide the research area into 60*60 grid units, and 10% of the points in each grid unit are extracted as representative points for subsequent work.

5.1.1. Parameter setting

The CLIQUE clustering algorithm is the core of this task. To determine the correct parameter settings for the CLIQUE algorithm, various combinations of the number of grid divisions K and the density threshold r are set appropriately based on previous experience. K is set to 20, 30 and 40, and the other parameter r is set to 100, 200 and 300. Each combination of the two parameters is applied in the CLIQUE algorithm. To intuitively display the clustering results, the alpha-shapes algorithm, in which the parameter α is set to 0.01, is applied to directly extract the boundaries of each cluster. Here, the set α is not discussed due to its minimal impact on the experimental results. The clustering results are shown in Fig. 9. If K is set too small, multiple ports will be identified as one port, while if it is set too large, one port will be classified as multiple ports. For r, if it is set too small, valid ports will be identified, while if it is set too large, some ports cannot be identified. Based on the above analysis and the results, the number of grid divisions K is set to 30, and the density threshold r is set to 100.

5.1.2. Results analysis

The appropriate parameters are determined, and the extraction processes for the ports (berths and anchorages) and the results for the Bohai Sea area are shown in Fig. 10. The AIS data with navigation statuses 'At Anchor' and 'Moored' are projected into each divided grid diagram, as shown in Fig. 10(a). The points are basically distributed on the shore, and the distribution of points conforms to general rules. However, several unreasonable noises can be identified by applying the CLIQUE algorithm. Fig. 10(b) displays the clustering result. 'x' in the figure represents noise. Then, the alpha-shapes algorithm is applied, the boundaries of 15 identified ports are extracted, and the results are shown in Fig. 10(c).

In addition, the port extraction process and identification results for Zhoushan waters are shown in Fig. 11. The experimental process and the parameter settings of the two algorithms are the same as above. The identified ports are almost on the coastline. Some are slightly further from the coastline, and these may be the outer anchorages for vessels to stay. Ultimately, 34 ports of this area are extracted.

By analysing the historical trajectories, the transportation relationships between ports in the Bohai Sea area can be obtained. Specifically, for each trajectory point in each vessel trajectory, if it is within the range of a port, the vessel is considered to be in this port. If a vessel stays at the port for more than 5 h, it is considered to have stayed at the port. After analysing all trajectory data, the final constructed PLSN is shown in Fig. 12. An undirected line segment indicates that there are vessels frequently travelling between two ports. Notably, some vessels sail directly out of the Bohai Sea from the port and do not stop at other ports in the Bohai Sea. Therefore, these vessel trajectories are not represented in this scale network. In addition, since there are too many ports identified in Zhoushan waters and too many connections between ports, the generated PLSN is confusing. Thus, the PLSN of this area is not displayed.

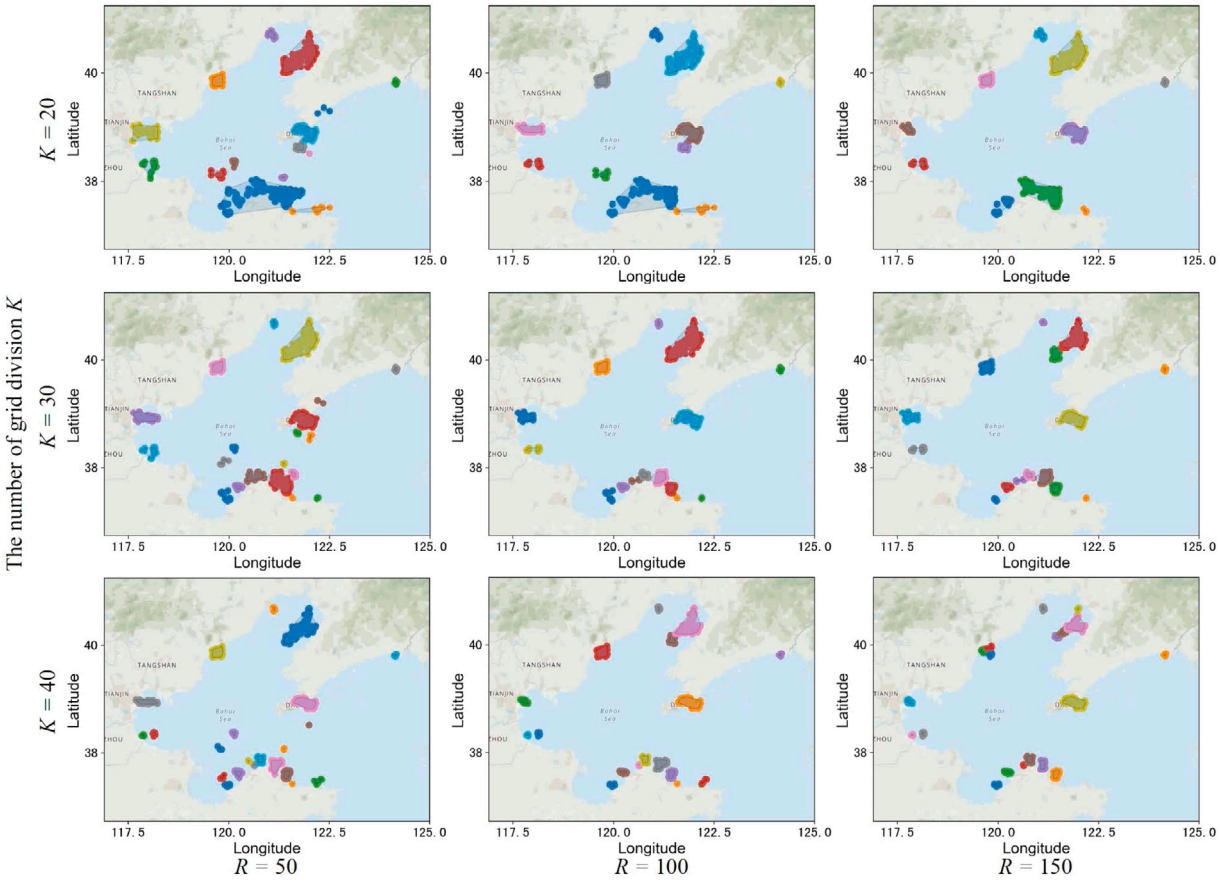


Fig. 9. Port identification results under different settings of the two parameters. R and K denote the density threshold and the number of grid division of the CLIQUE algorithm respectively.

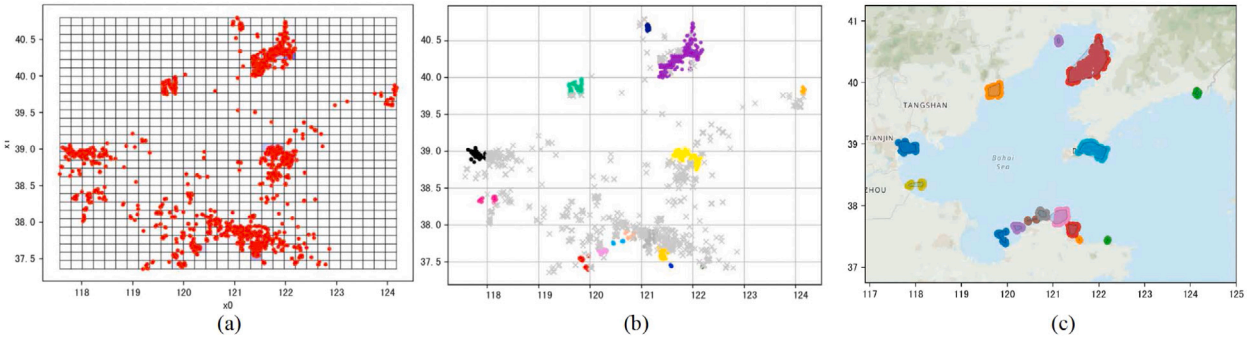


Fig. 10. The results of port identification in the Bohai Sea area. From left to right: (a) the projection of AIS data with navigation statuses 'At Anchor' and 'Moored', (b) extracted clusters, and (c) identified ports and their boundaries.

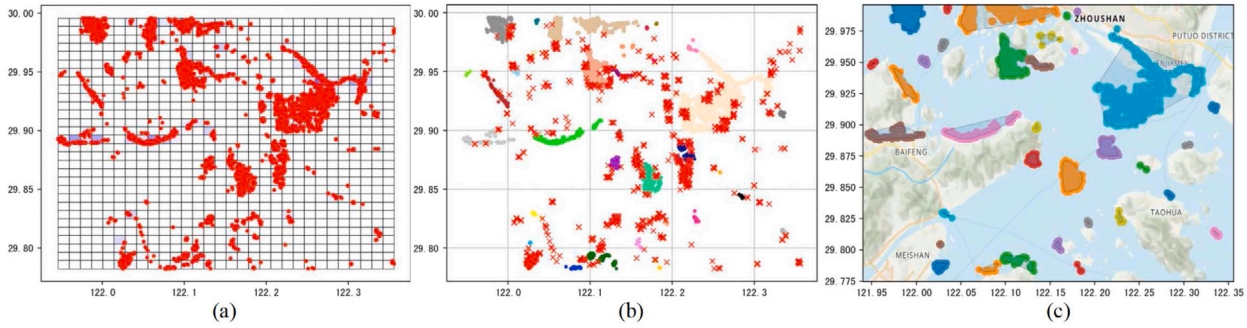


Fig. 11. The results of port identification in Zhoushan waters. From left to right: (a) the projection of AIS data with navigation statuses 'At Anchor' and 'Moored', (b) extracted clusters, and (c) identified ports and their boundaries.

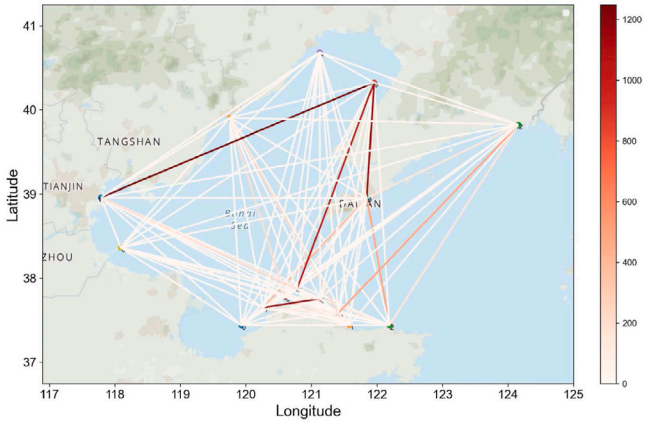


Fig. 12. The results of port-level shipping network generation for the Bohai Sea research area. Line segments denote the connectivity between ports, and the colour indicates the traffic flow value.

5.2. Extraction results of NLSN

To extract the feature points of trajectories from a vast array of vessel trajectories, an adaptive DP compression algorithm is proposed. Subsequently, the CLIQUE algorithm is utilized to cluster these extracted feature points, enabling the identification of waypoints (the nodes of the NLSN).

5.2.1. Parameter verification

To evaluate the efficacy of the compression algorithm, this paper randomly selected two vessel trajectories from the two research areas. The performance of the adaptive DP algorithm, including a comparison of the original trajectories and the compressed trajectories, is shown in Figs. 13 and 14. Fig. 13(a) and Fig. 14(a) detail the calculation process in which the evaluation index gradually obtains the optimal value as the compression threshold changes. The results indicate that, following the adaptive selection of the threshold by the algorithm, both the trajectory compression rate and the distance similarity rate attain high levels. Specifically, for the trajectory from the Bohai Sea area, the compression rate D_r and distance similarity rate D_l are 0.9926 and 0.9609, respectively. Similarly, for the trajectory from Zhoushan waters, these rates are 0.9925 and 0.9851, respectively, showing consistent, exceptional performance across different regions. The experimental results show the same excellent performance. Fig. 13(b) and Fig. 14(b) provide visual comparisons of the vessel trajectories before and after compression, clearly demonstrating that the fundamental shapes of the trajectories are well preserved.

5.2.2. Results analysis

By applying the trajectory compression algorithm, vessel trajectory feature points are extracted in both the Bohai Sea area and the Zhoushan waters. Subsequently, the CLIQUE algorithm is also applied to cluster the feature points to obtain the nodes of the NLSN, maintaining consistency in the density threshold r and the number of grid divisions K, as outlined in the prior section.

For the Bohai Sea area, the procedure for extracting NLSN and its outcomes are illustrated in Fig. 15. The extracted feature points indicate the vessel navigation routes, as shown in Fig. 15(a). Most noise was identified by the CLIQUE algorithm, and several clusters were identified, as shown in Fig. 15(b). Then these feature points in each cluster are recorded, and the average position of these feature points is abstractly expressed as the node of NLSN. NLSN is extracted based on the traffic flow analysis between nodes. Specifically, for one trajectory, the position of each trajectory point is compared to these kept feature points. If the vessel trajectory point is in some cluster, it is considered that the vessel passes through this node. Since the number of extracted trajectory feature points is extremely smaller than the number of identified anchoring and mooring points, and these feature points in one node range are extracted from these historical trajectories. Only when the vessels pass through two nodes continuously at high frequencies, it is considered that there is traffic flow between the nodes. Eventually, compared to the extracted PLSN, the main ports are also recognized within the NLSN.

In the Zhoushan waters area, the proposed method achieved the comparable performance, with the results shown in Fig. 16. By extracting the feature points and subsequently clustering them, the nodes are extracted and NLSN for this area is successfully generated.

5.3. Extraction results of RLSN

The RLSN is constructed based on the NLSN in this section, and the channel boundaries and customary routes are determined by statistically fitting the traffic flow sections. However, when sailing, vessels in each study area are easily affected by many factors, such as weather and tides. The navigation of some vessels presents its own particularities and may be inconsistent with the overall behavioural habits of vessels in that area. In addition, NLSNs extracted for both areas display numerous intersections among connecting line segments, which complicates the portrayal of a coherent RLSN. To address these challenges and to accurately represent the RLSN, selected fractional traffic flows were subjected to both statistical and visual analyses. The results of the RLSN construction are analysed below.

5.3.1. Method verification

This section mainly verifies the fitting performance of the traffic flow cross-section. Vessel trajectories spanning one week are extracted for analysis. Six traffic flow sections are randomly chosen, and the traffic flow statistics of the sections are shown in Fig. 17. The figure features a red dotted line representing the average traffic flow frequency, while the green dotted lines denote the lower specification limit (LSL) and upper specification limit (USL). The distribution of traffic flow across each selected section aligns closely with a normal distribution pattern. It indicates that the vessels unusually sail on the certain routes and the trajectories show the common navigation pattern.

To further validate the model fitting effect, this paper employs the Shapiro–Wilk Test (W test) and Kolmogorov–Smirnov Test (KS test) to verify the traffic flow fitting results (Belhadi et al., 2020; Lilliefors, 1967). The W test and the KS test are both testing methods used to determine whether a sample conforms to a normal distribution. The difference is that the W test is more suitable for data with a smaller sample size, while the KS test is suitable for data with a larger sample size. Here, a W statistic closer to 1 indicates a closer fit to a normal distribution, and a W test P value close to 0 indicates that the data do not follow a normal distribution. In addition, a lower KS statistic and a higher P value suggest a good fit between the sample data and the theoretical distribution, with no significant difference. According to the test results of channel section fitting, presented in Table 3, the values of both W Statistic and KS Test P-value are close to 1 and the values of the other two items are small. This indicates that the traffic flow in the channel section predominantly conforms to the normal distribution of a large sample.

5.3.2. Results analysis

To intuitively analyse the performance of the RLSN extraction method, this paper projects the extracted results onto an electronic chart (EC) for visualization, and the results are shown in Figs. 18 and 19. Based on the visualized local RLSN results for the Zhoushan waters, it can be clearly found that the vessel trajectory distribution is centralized, the channel boundaries and customary routes are clearly extracted, and they are all in line with navigation habits and navigation rules, especially as shown in Fig. 18.

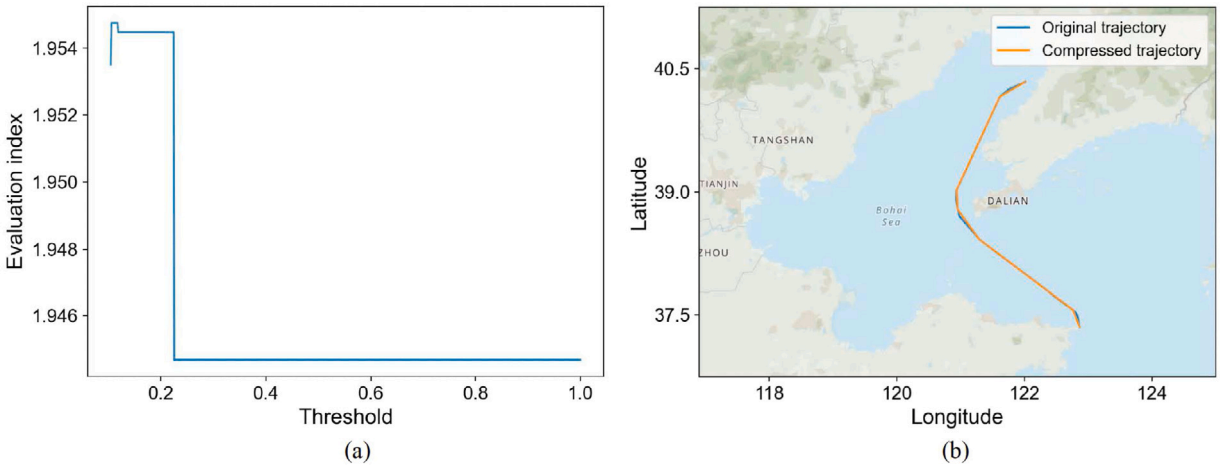


Fig. 13. The performance of the adaptive DP algorithm in the Bohai Sea area. From left to right: (a) the calculation process in which the evaluation index gradually obtains the optimal value as the compression threshold changes. (b) the comparison between the original trajectory and the compressed trajectory.

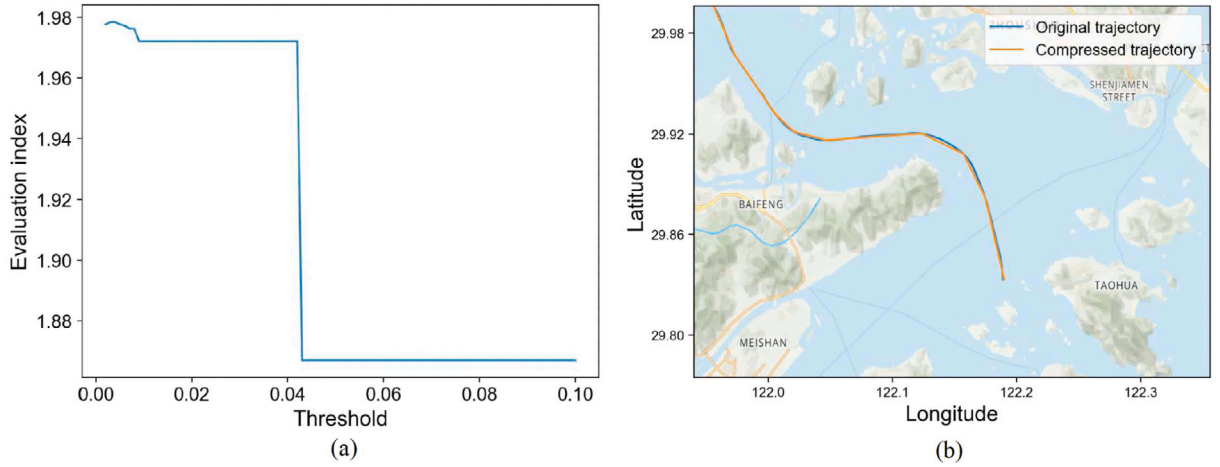


Fig. 14. The performance of the adaptive DP algorithm for Zhoushan waters. From left to right: (a) the calculation process in which the evaluation index gradually obtains the optimal value as the compression threshold changes. (b) the comparison between the original trajectory and the compressed trajectory.

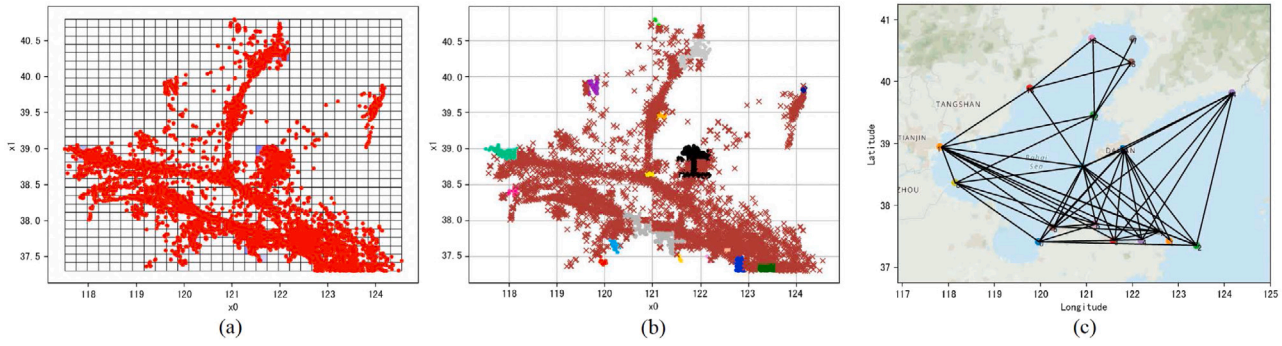


Fig. 15. Node-level shipping network generation for Bohai Sea area. From left to right: (a) the projection of extracted trajectory feature points, (b) extracted clusters, and (c) identified nodes and corresponding Node-level shipping network.

Table 3
Traffic flow fitting verification results of 6 cross-section.

	CS 1	CS 2	CS 3	CS 4	CS 5	CS 6
W Statistic	0.9891	0.9873	0.9865	0.9877	0.9819	0.9805
W Test *P*-value	0.5701	0.4252	0.4019	0.4811	0.1804	0.1402
KS Statistic	0.0495	0.0492	0.0504	0.0523	0.0504	0.0536
KS Test *P*-value	0.9513	0.9502	0.9504	0.9319	0.9480	0.9182

Notes: The column name CS indicates the selected cross-sectional traffic flow.

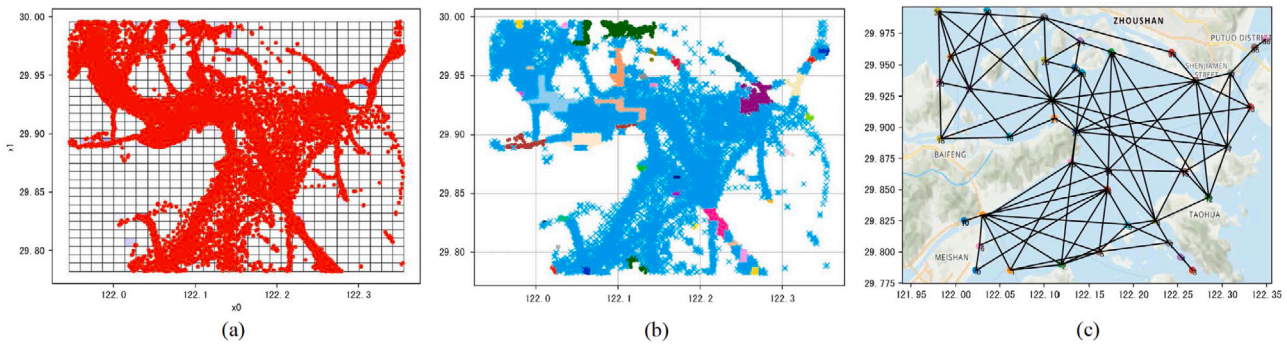


Fig. 16. Node-level shipping network generation for the Zhoushan waters. From left to right: (a) the projection of extracted trajectory feature points, (b) extracted clusters, and (c) identified nodes and corresponding Node-level shipping network.

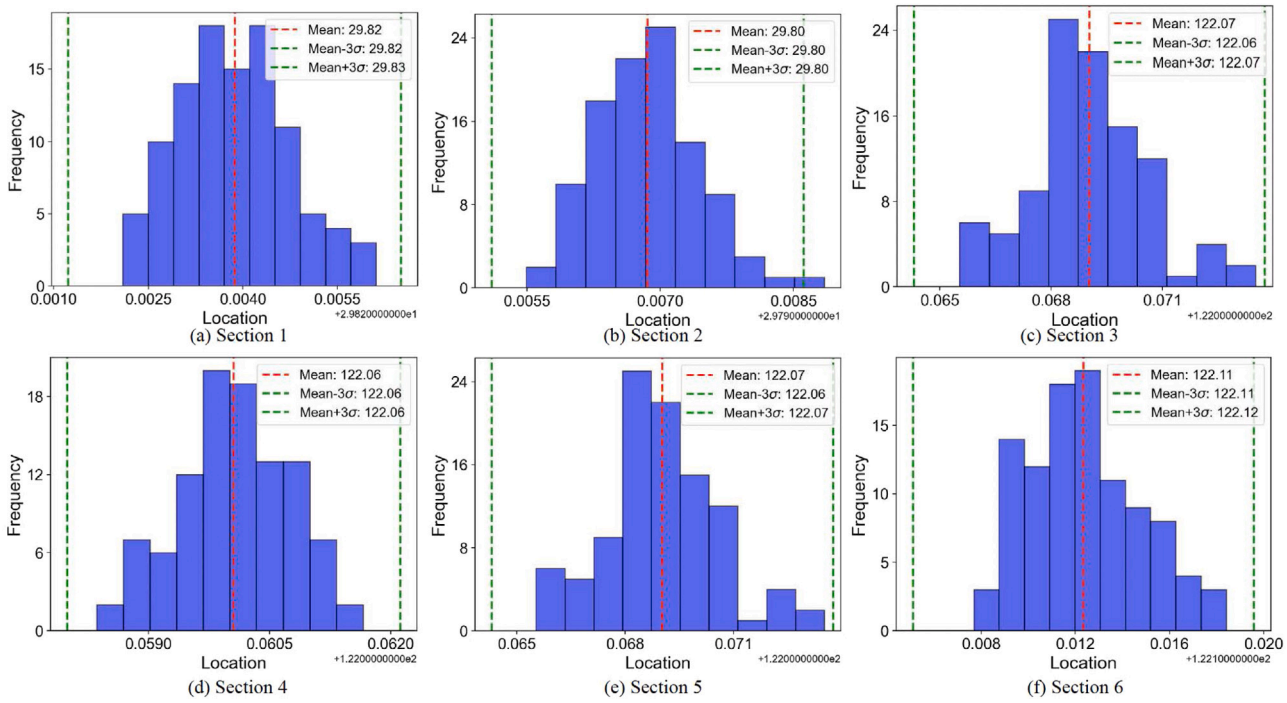


Fig. 17. Traffic flow statistics of six traffic flow sections. Red line means the average position and the green line means the route boundary.

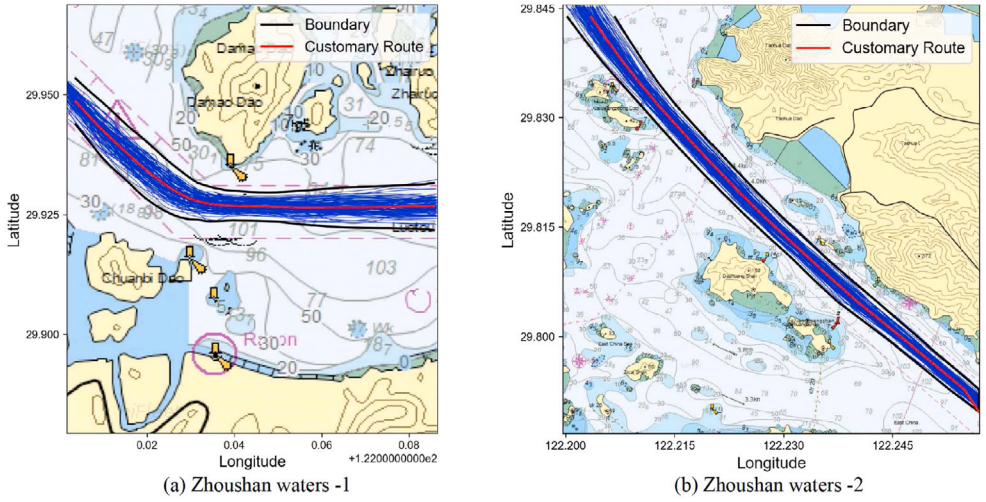
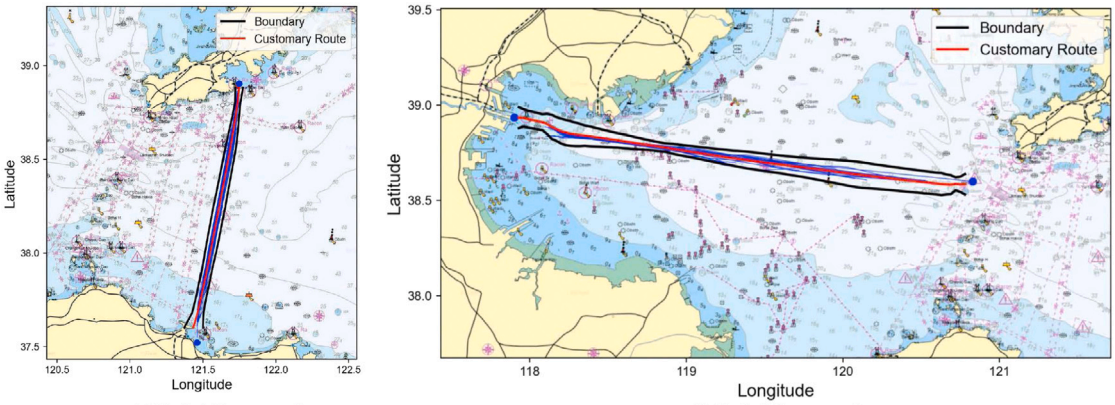


Fig. 18. The local route-level shipping network generation for the Zhoushan waters.



(a) Bohai Sea area -1 (b) Bohai Sea area -2

Fig. 19. The local route-level shipping network generation for the Bohai Sea area.

For the Bohai Sea area, boundary and customary route extraction also perform well in this area, regardless of the distance between the nodes. Moreover, even though there is a large open area for vessels to navigate, vessels are accustomed to sailing in a straight line, as revealed by the results shown in Fig. 19.

6. Conclusion

This paper proposes a massive AIS data-driven computational framework to automatically extract multi-scale shipping networks, encompassing port-level shipping networks, node-level shipping networks and route-level shipping networks. The massive AIS data of two research areas, the Bohai Sea and Zhoushan waters, were obtained from global AIS data for March 2018 and used to construct multi-scale shipping networks.

For PLSN extraction, effective vessel mooring and mooring data are mined from a massive AIS dataset, and then a port-port shipping network is constructed by extracting identified port boundaries through application of the CLIQUE clustering and alpha-shapes algorithms. In detail, the traffic flows between two ports are also calculated and shown.

For NLSN extraction, an adaptive DP algorithm is proposed and used to obtain the trajectory feature points. The CLIQUE clustering algorithm is used to obtain clusters again, and the identified waypoints are then connected to form the NLSN. Based on the extracted NLSN, the normal line of n axes, connecting line segments between nodes, is selected as the section to count the number of vessels passing through the section in a specific time interval to obtain an accurate channel traffic flow distribution. The customary routes are formed based on the mean traffic flow statistics, with Gaussian fitting confidence intervals defining the boundaries, culminating in the RLSN formation.

The experimental results show that the proposed method is feasible and effective and can provide a valuable reference for shipping logistic analysis and lay the groundwork for future endeavours in trajectory clustering, anomaly detection and trajectory prediction. However, different types of vessels have their own customary sailing habits. For example, passenger vessels travel on fixed routes, and cargo vessels usually do not change their behavioural habits to ensure economic and environmental protection. However, fishing vessels may act in a disorderly manner, frequently changing course and speed, which may cause relatively poor accuracy of network construction. In this paper, the impact of these differences on shipping network extraction has not been considered. Additionally, the navigation patterns of vessels may also vary in different seasons. Some passenger vessels only perform the navigation work in the tourist season, which may also affect the final results of shipping network extraction. Furthermore, the efficiency of the adopted clustering algorithm affects the accuracy of the network extraction results. In conclusion, future studies will not only examine the impact of such these variances on shipping network extraction but also explore higher performance multi-scale shipping network extraction methods to enhance practical application and relevance.

CRediT authorship contribution statement

Ryan Wen Liu: Software, Methodology, Conceptualization. **Shiqi Zhou:** Writing – original draft, Visualization, Validation, Methodology. **Maohan Liang:** Writing – original draft, Visualization, Validation, Methodology. **Ruobin Gao:** Methodology, Investigation, Data curation, Conceptualization. **Hua Wang:** Investigation, Data curation, Conceptualization.

Declaration of competing interest

This manuscript has not been published or presented elsewhere in part or in entirety and is not under consideration by another journal. We have read and understood your journal's policies, and we believe that neither the manuscript nor the study violates any of these. There are no conflicts of interest to declare.

Funding

This research has been supported by the National Key Research and Development Program of China (No.2022YFC3302702).

References

Agrawal, R., Gehrke, J., Gunopulos, D., Raghavan, P., 1998. Automatic subspace clustering of high dimensional data for data mining applications. In: Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data. pp. 94–105.

Arguedas, V.F., Pallotta, G., Vespe, M., 2014. Automatic generation of geographical networks for maritime traffic surveillance. In: 17th International Conference on Information Fusion. FUSION, IEEE, pp. 1–8.

Belhadi, A., Djenouri, Y., Srivastava, G., Djenouri, D., Cano, A., Lin, J.C.-W., 2020. A two-phase anomaly detection model for secure intelligent transportation ride-hailing trajectories. IEEE Trans. Intell. Transp. Syst. 22 (7), 4496–4506.

Cai, J., Chen, G., Lützen, M., Rytter, N.G.M., 2021. A practical AIS-based route library for voyage planning at the pre-fixture stage. Ocean Eng. 236, 109478.

Cheung, K.F., Bell, M.G., Pan, J.-J., Perera, S., 2020. An eigenvector centrality analysis of world container shipping network connectivity. Transp. Res. Part E Logist. Transp. Rev. 140, 101991.

Cullinane, K., Bergqvist, R., 2014. Emission control areas and their impact on maritime transport. Transp. Res. Part D Transp. Environ. 28, 1–5.

Douglas, D.H., Peucker, T.K., 1973. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. Cartographica 10 (2), 112–122.

Duan, G., Fan, T., Chen, L., Ma, J., 2021. Floating marine debris mitigation by vessel routing modeling and optimization considering carbon emission and travel time. Transport. Res. Part C: Emerging Technol. 133, 103449.

Edelsbrunner, H., Mücke, E.P., 1994. Three-dimensional alpha shapes. ACM Trans. Graph. 13 (1), 43–72.
Gu, Y., Wallace, S.W., Wang, X., 2019. Can an emission trading scheme really reduce CO2 emissions in the short term? Evidence from a maritime fleet composition and deployment model. Transp. Research Part D Transp. Environ. 74, 318–338.
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.
Irannezhad, E., Prato, C.G., Hickman, M., 2018. The effect of cooperation among shipping lines on transport costs and pollutant emissions. Transp. Res. Part D Transp. Environ. 65, 312–323.
Kim, H.-S., Lee, E., Lee, E.-J., Hyun, J.-W., Gong, I.-Y., Kim, K., Lee, Y.-S., 2023. A study on grid-cell-type maritime traffic distribution analysis based on AIS data for establishing a coastal maritime transportation network. J. Mar. Sci. Eng. 11 (2), 354.
Lai, K.-H., Lun, V.Y., Wong, C.W., Cheng, T.C.E., 2011. Green shipping practices in the shipping industry: Conceptualization, adoption, and implications. Resour. Conserv. Recycl. 55 (6), 631–638.
Li, Y., Liu, R.W., Liu, J., Huang, Y., Hu, B., Wang, K., 2016. Trajectory compression-guided visualization of spatio-temporal AIS vessel density. In: 2016 8th International Conference on Wireless Communications & Signal Processing. WCSP, IEEE, pp. 1–5.
Liang, M., Li, H., Liu, R.W., Lam, J.S.L., Yang, Z., 2024a. PiracyAnalyzer: Spatial temporal patterns analysis of global piracy incidents. Reliab. Eng. Syst. Saf. 243, 109877.
Liang, M., Liu, R.W., Gao, R., Xiao, Z., Zhang, X., Wang, H., 2024. A survey of distance-based vessel trajectory clustering: Data pre-processing, methodologies, applications, and experimental evaluation. arXiv preprint arXiv:2407.11084.
Liang, M., Liu, R.W., Zhan, Y., Li, H., Zhu, F., Wang, F.-Y., 2022. Fine-grained vessel traffic flow prediction with a spatio-temporal multigraph convolutional network. IEEE Trans. Intell. Transp. Syst. 23 (12), 23694–23707.
Liang, M., Su, J., Liu, R.W., Lam, J.S.L., 2024b. AISClean: AIS data-driven vessel trajectory reconstruction under uncertain conditions. Ocean Eng. 306, 117987.
Liang, M., Weng, L., Gao, R., Li, Y., Du, L., 2024c. Unsupervised maritime anomaly detection for intelligent situational awareness using AIS data. Knowl.-Based Syst. 284, 111313.
Liao, Z., Liu, J., Shi, G., Meng, J., 2021. Grid partition variable step alpha shapes algorithm. Math. Probl. Eng. 2021, 1–8.
Lilliefors, H.W., 1967. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. J. Amer. Statist. Assoc. 62 (318), 399–402.
Liu, H., Oyama, S., Kurihara, M., Sato, H., 2013. Landmark FN-DBSCAN: an efficient density-based clustering algorithm with fuzzy neighborhood. J. Adv. Comput. Intell. Inform. 17 (1).
Liu, D., Rong, H., Soares, C.G., 2023a. Shipping route modelling of AIS maritime traffic data at the approach to ports. Ocean Eng. 289, 115868.
Liu, L., Shibasaki, R., Zhang, Y., Kosuge, N., Zhang, M., Hu, Y., 2023b. Data-driven framework for extracting global maritime shipping networks by machine learning. Ocean Eng. 269, 113494.
Ma, Q., Du, X., Liu, C., Jiang, Y., Liu, Z., Xiao, Z., Zhang, M., 2024. A hybrid deep learning method for the prediction of ship time headway using automatic identification system data. Eng. Appl. Artif. Intell. 133, 108172.
Martinčič, T., Štepec, D., Costa, J.P., Čagran, K., Chaldeakis, A., 2020. Vessel and port efficiency metrics through validated AIS data. In: Global Oceans 2020: Singapore–US Gulf Coast. IEEE, pp. 1–6.
Mason, S.J., Ribera, P.M., Farris, J.A., Kirk, R.G., 2003. Integrating the warehousing and transportation functions of the supply chain. Transp. Res. Part E Logist. Transp. Rev. 39 (2), 141–159.
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., Soares, C.G., 2020. Data mining approach to shipping route characterization and anomaly detection based on AIS data. Ocean Eng. 198, 106936.
Rong, H., Teixeira, A., Soares, C.G., 2022. Maritime traffic probabilistic prediction based on ship motion pattern extraction. Reliab. Eng. Syst. Saf. 217, 108061.
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., Teixeira, A., Guedes-Soares, C., 2019. AIS based shipping routes using the Dijkstra algorithm. TransNav 13 (3).
Sturgis, R., Emiya, V., Couëtoux, B., Garreau, P., 2024. Beyond geofencing: Behavior detection using AIS. Ocean Eng. 293, 116630.
Tang, C., Wang, H., Zhao, J., Tang, Y., Yan, H., Xiao, Y., 2021. A method for compressing AIS trajectory data based on the adaptive-threshold Douglas-Peucker algorithm. Ocean Eng. 232, 109041.
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.
Wang, S., Gao, S., Yang, W., 2017. Ship route extraction and clustering analysis based on automatic identification system data. In: 2017 Eighth International Conference on Intelligent Control and Information Processing. ICICIP, IEEE, pp. 33–38.
Wen, R., Yan, W., Zhang, A.N., Chinh, N.Q., Akcan, O., 2016. Spatio-temporal route mining and visualization for busy waterways. In: 2016 IEEE International Conference on Systems, Man, and Cybernetics. SMC, IEEE, pp. 000849–000854.
Xiao, Z., Fu, X., Zhang, L., Goh, R.S.M., 2019. Traffic pattern mining and forecasting technologies in maritime traffic service networks: A comprehensive survey. IEEE Trans. Intell. Transp. Syst. 21 (5), 1796–1825.
Xiao, F., Ligteringen, H., Van Gulijk, C., Ale, B., 2015. Comparison study on AIS data of ship traffic behavior. Ocean Eng. 95, 84–93.
Xing, S., Su, Q.-M., Xiong, Y.-J., Xia, C.-M., 2023. PDCSN: A partition density clustering with self-adaptive neighborhoods. Expert Syst. Appl. 227, 120195.
Yan, Z., Cheng, L., He, R., Yang, H., 2022. Extracting ship stopping information from AIS data. Ocean Eng. 250, 111004.
Yan, Z., Xiao, Y., Cheng, L., Chen, S., Zhou, X., Ruan, X., Li, M., He, R., Ran, B., 2020a. Analysis of global marine oil trade based on automatic identification system (AIS) data. J. Transp. Geogr. 83, 102637.
Yan, Z., Xiao, Y., Cheng, L., He, R., Ruan, X., Zhou, X., Li, M., Bin, R., 2020b. Exploring AIS data for intelligent maritime routes extraction. Appl. Ocean Res. 101, 102271.
Yang, Y., Liu, Y., Li, G., Zhang, Z., Liu, Y., 2024. Harnessing the power of machine learning for AIS data-driven maritime research: A comprehensive review. Transp. Res. Part E Logist. Transp. Rev. 183, 103426.
Yang, J., Ma, L., Liu, J., 2021a. Modeling and application of ship density based on ship scale conversion and grid. Ocean Eng. 237, 109557.
Yang, D., Wu, L., Wang, S., 2021b. Can we trust the AIS destination port information for bulk ships?–Implications for shipping policy and practice. Transp. Res. Part E Logist. Transp. Rev. 149, 102308.
Yang, D., Wu, L., Wang, S., Jia, H., Li, K.X., 2019. How big data enriches maritime research–a critical review of automatic identification system (AIS) data applications. Transp. Rev. 39 (6), 755–773.
Yu, H., Fang, Z., Fu, X., Liu, J., Chen, J., 2021. Literature review on emission control-based ship voyage optimization. Transp. Res. Part D Transp. Environ. 93, 102768.
Yu, Y., Liu, K., Fu, S., Chen, J., 2024. Framework for process risk analysis of maritime accidents based on resilience theory: A case study of grounding accidents in arctic waters. Reliab. Eng. Syst. Saf. 110202.
Zhang, L., Chen, P., Li, M., Chen, L., Mou, J., 2022a. A data-driven approach for ship-bridge collision candidate detection in bridge waterway. Ocean Eng. 266, 113137.
Zhang, L., Meng, Q., Fwa, T.F., 2019. Big AIS data based spatial-temporal analyses of ship traffic in Singapore port waters. Transp. Res. Part E Logist. Transp. Rev. 129, 287–304.
Zhang, S.-k., Shi, G.-y., Liu, Z.-j., Zhao, Z.-w., Wu, Z.-l., 2018. Data-driven based automatic maritime routing from massive AIS trajectories in the face of disparity. Ocean Eng. 155, 240–250.
Zhang, M., Zhang, D., Fu, S., Kujala, P., Hirdaris, S., 2022b. A predictive analytics method for maritime traffic flow complexity estimation in inland waterways. Reliab. Eng. Syst. Saf. 220, 108317.
Zhou, C., Xiang, J., Huang, H., Yan, Y., Huang, L., Wen, Y., Xiao, C., 2023. TTMRN: A topological-geometric two-layer maritime route network modeling for ship intelligent navigation. Ocean Eng. 287, 115884.