Liu, Zhao; Chen, Wanli; Liu, Cong; Yan, Ran; Zhang, Mingyang

A data mining-then-predict method for proactive maritime traffic management by machine learning

Published in:
Engineering Applications of Artificial Intelligence

DOI:
10.1016/j.engappai.2024.108696

Published: 01/09/2024

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

Engineering Applications of Artificial Intelligence 135 (2024) 108696

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

A data mining-then-predict method for proactive maritime traffic management by machine learning

Zhao Liu [a,b], Wanli Chen [a,b], Cong Liu [c,*], Ran Yan [d], Mingyang Zhang [c,**]

[a] School of Navigation, Wuhan University of Technology, Wuhan, China
[b] Hubei Key Laboratory of Inland Shipping Technology, School of Navigation, Wuhan University of Technology, Wuhan, 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 traffic management
Traffic pattern extraction and prediction
Machine learning
Automatic identification system data

ABSTRACT

Proactive traffic management is increasingly critical in maritime intelligent transportation systems. Central to this is maritime traffic forecasting, which leverages specific structures and properties of the problem. This study focuses on the traffic dynamics within convergent areas of inland waterways and proposes a method based on data mining followed by prediction using Automatic Identification System (AIS) data. This approach addresses uncertainties in ship voyage destinations and optimizes predictions for temporary stops in inland waterways. AIS data is processed to depict complete ship motion trajectories, grouping them into trajectory sets based on shared origin, destination, and route. These groups help represent maritime traffic patterns using the entrance and exit points of channels and the boundaries of the study area. Additionally, a stop detection model is applied to these trajectories to identify nodes within maritime traffic networks. A decision tree algorithm is then employed to train a classifier for predicting traffic patterns. The method was validated in the convergent area of the Yangtze River and the Hanjiang River, demonstrating effective pattern extraction from inland maritime traffic and high accuracy in predicting single ship trajectories, achieving a 96.7% accuracy rate and 80.9% precision. The findings suggest that the proposed method (1) effectively extracts and predicts traffic patterns, (2) identifies congestion in convergent waters, and (3) supports traffic management strategies.

1. Introduction

With the rapid growth in maritime transportation demand, the number and diversity of ships have increased significantly, making maritime traffic more complex and increasing the associated risks (Li et al., 2023; Zhang et al., 2022a; Chen et al., 2024). Ship traffic safety management is a critical concern in maritime activities, particularly in restricted waterways and busy waters where accidents can result in substantial direct and indirect human and property losses (Zhang et al., 2023a; Cheng et al., 2024). In response to complex traffic conditions and significant safety challenges, intelligent maritime equipment and systems, such as Vessel Traffic Services (VTS) and Geographical Information Systems (GIS), have been employed to enhance maritime traffic monitoring (Xin et al., 2023; Liang et al., 2021, 2024). While these systems effectively monitor and regulate maritime traffic behaviors, yielding significant benefits, they predominantly depend on human decision-making and judgment. Consequently, they often lack advanced intelligence and automation for risk warnings, which hampers their ability to provide timely and accurate alerts.

Due to safety and efficiency concerns, ships in inland waterways typically adhere to designated navigational waterways, constrained by maritime regulations and traffic separation schemes. These spatial patterns reflect ship routes shaped by factors such as water activities, traffic planning, ship maneuvering, and hydrological characteristics (Rong et al., 2022). At intersections within inland waterways, ships often experience queuing, leading to significant time and space resource wastage. These intersections are conflict hotspots, primarily due to the frequent crossing and converging of vessels, which can lead to ship conflicts (Cheng et al., 2023), as illustrated in Fig. 1(a). Furthermore, collision risks increase when ships deviate from and attempt to return to their routes, particularly during traffic congestion. This may necessitate actions such as slowing down, altering course, or temporary docking to

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

https://doi.org/10.1016/j.engappai.2024.108696
Received 12 January 2024; Received in revised form 18 April 2024; Accepted 26 May 2024
Available online 7 June 2024

avoid collisions, as depicted in Fig. 1(b). Consequently, effective ship dispatching and intelligent traffic control at these critical points are both necessary and urgent.

Ship trajectories, influenced by hydrological, meteorological, and navigational maneuvers, encapsulate the behavior and routes of ships during voyages (Zhang et al., 2021a). As maritime traffic increases, many trajectories within the same traffic environment exhibit similar characteristics such as route, speed, and course. These patterns provide insights into broader traffic conditions, including scenarios of ships entering or leaving ports. In traffic research, maritime traffic patterns are derived from data mining techniques like ship trajectory clustering, allowing the aggregated historical trajectory clusters to represent specific traffic patterns. These clusters not only encapsulate past ship behavior but also guide future navigation, as illustrated in Fig. 1(c).

Therefore, by identifying historical traffic patterns and predicting future ship behavior based on these patterns and the latest traffic data, maritime supervisory authorities can effectively perceive, analyze, and manage traffic scenarios, and prepare dispatching strategies. This forms the technical foundation for collaborative ship control and intelligent dispatching, which is crucial for efficient maritime traffic management. The widespread implementation of AIS supports this goal by enhancing traffic data availability and accuracy (Zhang et al., 2021b; Liu et al., 2022).

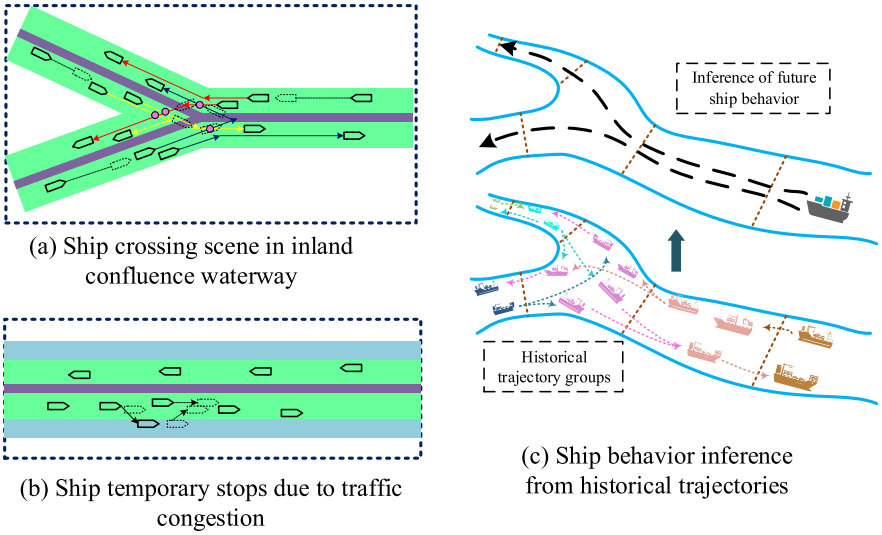
With the vast amount of AIS data available, new opportunities arise for maritime traffic research, which can be enriched by integrating other maritime sources like radar and video (Guo et al., 2023; Lu et al., 2022; Li et al., 2024; Liu et al., 2021). Various technologies—including computer vision, statistical analysis, and machine learning—are employed to analyze AIS data, enabling researchers to reconstruct historical ship behaviors and assess regional traffic conditions (Rong et al., 2022; Liu et al., 2023a,b,c). However, due to the complexity of ship behaviors and route distributions, extracting and modeling maritime traffic pattern networks from extensive AIS data remains a significant challenge (Zhao and Shi, 2019). Addressing these complexities, this paper introduces a method that combines data mining and prediction to identify and categorize traffic patterns within inland waterways, uncover hotspot areas for stops, visualize maritime traffic with Origin-Destination (OD) graphs, and predict the traffic pattern of individual ships using a decision tree model. The main contributions of this study are as follows:

- A method for traffic management in inland waterways based on AIS data within "data processing and trajectory reconstruction-traffic pattern extraction-feature mining-traffic pattern prediction" is proposed, which provide a framework reference for inland maritime traffic situation awareness and prediction research.
- The method for extracting traffic patterns and maritime network nodes takes into account the dynamic traffic conditions. Unlike conventional trajectory clustering, this method uses the trends of the channels and the origin and destination of ship voyages to determine patterns, thereby reducing the computational burden associated with measuring trajectory similarity. Furthermore, the trajectories of ships that have made temporary stops are also recognized as indicative of traffic patterns. This understanding aids in labeling comprehensive traffic patterns for subsequent prediction tasks, addressing uncertainties in ship voyage destinations and temporary stops within proactive traffic management.
- A decision tree model is employed for predicting traffic patterns, chosen for its high interpretability as a machine learning classification method. The classifier inputs include both static traffic features (such as the origin area and stopping behavior) and dynamic traffic features (such as speed and course). This selection considers the differential traffic flow characteristics of various traffic patterns, effectively reflecting the specific traffic conditions of interest.

The rest of this paper is organized as follows: in Section 2, the related research of maritime traffic pattern extraction and ship behavior prediction is reviewed; in Section 3, the method of data mining and traffic pattern prediction are introduced, which includes three parts: data processing, data mining, and traffic pattern prediction; 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 supplementary experimental part is discussed; Section 6, the conclusions are drawn, and the future work is described.

2. Literature review

Maritime traffic pattern extraction refers to the use of AIS and other data sources to mine representative and regular maritime traffic behaviors and navigation routes from massive ship trajectories to facilitate applications such as maritime supervision (Xiao et al., 2023), route planning (Liu et al., 2024a,b,c), and position prediction (Zhang et al., 2022b; Liang et al., 2022). Generally, when the ship voyage origin and destination are determined and connected by specific routes, a type of traffic pattern can be determined accordingly (Zhang et al., 2023b; Fuentes, 2021; Li et al., 2023; Ma et al., 2024). There are three main methods for pattern extraction: (1) grid-based methods, (2) vector-based



(a) Ship crossing scene in inland confluence waterway

(b) Ship temporary stops due to traffic congestion

(c) Ship behavior inference from historical trajectories

Fig. 1. Ship behavior inference and traffic conflict scenarios in inland waterway.

methods, and (3) statistics-based methods (Xiao et al., 2020). The grid-based method is mainly suitable for fractional traffic areas. Each grid is accompanied by basic attribute statistics, such as traffic density, route and speed, and the construction of a grid database can reduce the size of the data and facilitate efficient maritime knowledge retrieval (Xiao et al., 2017; Vettor and Guedes Soares, 2015; Zhu, 2011; Liu et al., 2024a,b,c). Vector-based technology can compactly represent traffic patterns, usually connected by extracted nodes and lanes, but simplified route modeling may lose necessary information (Murray and Perera, 2022; Yan et al., 2020a; Huang et al., 2023; Dobrkovic et al., 2018); Statistical techniques can only provide basic traffic statistics, with the aim of revealing the distribution profile of traffic attributes (Han and Yang, 2020; Lee and Cho, 2022; Ristic et al., 2008). For instance, it describes the distribution characteristics of ship traffic, and captures traffic hotspots (Silveira et al., 2013; Wu et al., 2016). These studies have great significance in determining important traffic parameter values or thresholds (Kang et al., 2018).

Ship voyages typically adhere to established waterway patterns, resulting in predictable ship motion patterns within specific areas over long periods. Consequently, ships follow traffic patterns that are extracted from historical trajectories, making these patterns significantly influential for guiding future navigation. The ability to mine the distinct traffic flow characteristics of different patterns and to classify ships according to the features of their current trajectories is crucial. The importance of this approach is highlighted in several key aspects (Arguedas et al., 2017):

- Maritime traffic patterns are the regular and representative expressions of ship behavior. Mining traffic patterns is to analyze the traffic characteristics and operational conditions, for the supplement to the regional maritime knowledge base (Liu et al., 2023b; Ma et al., 2024).
- Maritime traffic patterns provide prediction targets for ship behavior prediction. Ship behavior prediction often involves classification algorithms, which generally require labeled data (Duan et al., 2022). Therefore, the extracted traffic patterns provide the labeled categories and the expected results of classification.
- Maritime traffic patterns mining, and prediction provides support for maritime traffic management and ship dispatching and control. After classifying ships into specific patterns, it provides specific management objectives and the expected effects for maritime management departments. For example, it helps to perceive the sailing intentions and sailing obstacles in advance, to optimize the sailing routes and speeds of ships, etc.

Maritime traffic pattern prediction is a key research area in maritime studies. The ability to anticipate ship behavior and navigation status is crucial for evolving maritime supervision from passive to active (Xiao et al., 2023). Both machine learning and deep learning have become dominant technologies for predicting ship behaviors. Prominent machine learning techniques include Support Vector Machines (SVM), Gaussian Process Regression (GPR), Random Forests (RF), and Artificial Neural Networks (ANN). Qi and Zheng (2016) trained an SVM classifier using grouped trajectory data to predict new trajectories based on the central trajectory of each group. Chen et al. (2021) enhanced this approach with the Dimension Learning Grey Wolf Optimizer and Support Vector Regression for better trajectory prediction. Liu et al. (2019) combined the Adaptive Chaotic Differential Evolution algorithm with SVR for similar purposes. SVM is effective with small datasets and avoids overfitting but struggles with large datasets. Rong et al. (2019) introduced a GPR model to predict ship motion in lateral and longitudinal components, and later (Rong et al., 2022), they developed a joint MLR and GPR model to predict ship destinations and behavior probabilities. GPR works well for short-term predictions but is less efficient with extensive data. Zhang et al. (2020) used a Random Forest model to predict ship destinations by comparing trajectories, while Abebe et al. (2020) employed the same model to predict ship speeds. However, Random Forest can be overfit in data-rich, noisy environments. Volkova et al. (2021) utilized ANN for trajectory predictions, though this method requires extensive parameter tuning and faces challenges in gradient management and computational demand.

With technological advancements, deep learning has gained prominence, particularly methods like LSTM, Seq2seq, and GRU (Gao et al., 2023; Park et al., 2021; Capobianco et al., 2021). These models, capable of handling complex data scenarios, require large datasets and careful parameter tuning but offer superior performance over traditional machine learning (Zhang et al., 2024; Liu et al., 2024a,b,c). The interpretability of these models, however, remains a challenge, necessitating a balance between efficiency and clarity in model behavior.

Current research often assumes that a ship present state continues into the near future, a premise not always valid in maritime operations where immediate changes are frequent. By utilizing historical data, predictions about future maritime traffic patterns can be improved, thereby enhancing the accuracy and utility of predictive models. This paper focuses on leveraging historical trajectories for predicting future maritime traffic patterns, highlighting the importance of model transparency and interpretability in enhancing prediction accuracy.

The simultaneous consideration of ship destination uncertainty and the unpredictability of temporary berthing during navigation is rare, especially in convergent areas of inland waterways where traffic congestion and spatial-temporal resource waste are significant concerns. Timely and strategic resource allocation and scheduling are crucial to pre-emptively addressing these issues and minimizing traffic conflicts, which are key to intelligent maritime traffic management. Addressing these gaps, this paper introduces a method that applies data mining followed by prediction to analyze traffic patterns in inland waters, uniquely accounting for both the uncertainties of ship destination and temporary berthing behaviors during navigation. The paper also extend our research to include a prediction experiment that examines the impact of ship sailing time on the accuracy of traffic pattern prediction. Our approach begins by mining regional traffic characteristics, which are then fed into a prediction model that integrates distinct traffic features and stop detection results, considering the historical maritime network traits of the area concerned. The proposed machine learning prediction model is not only highly interpretable and precise but also yields results and processes that are easily adaptable and user-friendly for maritime regulators. This study enriches the knowledge base of regional transportation and supports proactive management strategies.

3. Methods

The data mining-then-predict method for maritime traffic management in inland waterways is illustrated in Fig. 2. This method is divided into three stages, each briefly described below and detailed further in Sections 3.1 to 3.3.

- Step (I): Data processing: AIS data undergo several processing steps, including data cleaning, matching static and dynamic data, and trajectory reconstruction. The processed AIS data are used to extract traffic characteristics, reflect the historical motion features of ships, and provide support for feature mining in subsequent sections.
- Step (II): Data mining for various traffic groups: Traffic groups are delineated based on the origin and destination zones to represent distinct traffic patterns. The process begins by detecting and matching the start and end points of historical ship trajectories with predefined origin and destination zones, leveraging prior knowledge. If multiple historical trajectories share the same pair of origin and destination zones, they are classified into the same traffic group. Additionally, the detection of temporary stops in ship trajectories helps to identify a range of stop patterns, further enriching the understanding of maritime traffic networks.

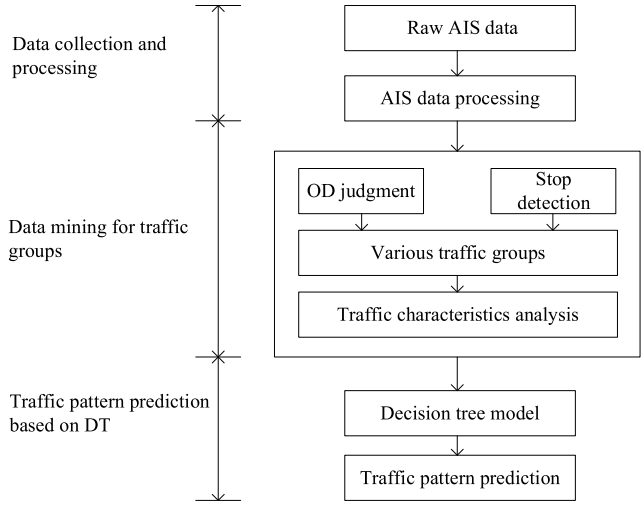


Fig. 2. Framework of a data mining-then-predict method for maritime traffic management.

- Step (III): Traffic pattern prediction: A decision tree model is developed to predict the traffic patterns of new ship trajectories. This model integrates static traffic features, such as origin area and stop behavior, with dynamic features like speed and heading. These selected features are crucial, as they exhibit distinct differences that are essential for effective feature mining and accurately reflect the unique traffic flow characteristics of the targeted area. The prediction process tackles uncertainties related to the ship's sailing destination and potential temporary stops. It begins by labeling previously extracted traffic patterns, creating a feature dataset, and training a decision tree classifier on this dataset. The classifier is then applied to new trajectories, assigning each to a specific traffic pattern based on the learned distinctions. Additionally, a supplementary experiment assesses the impact of the duration a ship spends navigating the waters on the accuracy of predictions. This experiment involves decision tree models trained on datasets with varying trajectory lengths to evaluate how time influences prediction outcomes.

3.1. AIS data and data processing

AIS data is categorized into two primary types: ship static data and ship dynamic data (Liu et al., 2022). Static data encompasses details like ship length, width, and type, while dynamic data records variables such as timestamp, longitude, latitude, speed, and course. Initially, the integration of dynamic and static data is achieved by utilizing the Maritime Mobile Service Identity (MMSI) to match related records(Liu et al., 2024). Subsequently, the AIS data is transformed into ship trajectories, each representing a series of time-sequenced data points enriched with multidimensional information, as $\{p_1, \ldots p_i, \ldots, p_n\}$. Each data point p_i consists of traffic information characteristics as Eq. (1).

$$p_i = \{MMSI, Timestamp, Lon, Lat, Sog, Cog, L, W, Type\} \tag{1}$$

where p_i is the trajectory point, *MMSI* is the unique identifier of the ship, *Timestamp* is the AIS timestamp, *Lon*, *Lat*, *Sog*, *Cog*, *L*, *W*, *Type* are the ship longitude, latitude, speed, course, ship length, ship width and ship type information respectively.

The subsequent step involves trajectory segmentation, designed to distinguish between different voyages. According to Zhang et al. (2022a), the criteria for segmentation is that the time interval between adjacent trajectory points exceeds 900 s. Consequently, the first point after each segmentation threshold marks the origin, and the last point indicates the destination. This process allows ship trajectories to accurately represent the real operational conditions.

Due to potential issues such as equipment failures, signal delays, and obstacle interference during the collection and transmission of AIS data, there may be anomalies in AIS data such as missing, drifting, and errors (Wang et al., 2022). Therefore, in the study of ship transportation using AIS, it is necessary to clean AIS data. Mainly, the paper considers the filtering of abnormal speed and position data. In addition, the cleaning of *MMSI* and static data mistakes are also conducted. The last task is interpolation, which is to fill the missing data that are generated by cleaning task and data collection. The linear interpolation is applied for the interpolation, while cubic spline interpolation is discarded because of its overfitting. The time interval is set as the mode of time interval in each trajectory. The flowchart of AIS data processing is depicted in Fig. 3.

3.2. Data mining for various traffic groups

3.2.1. Traffic patterns in inland waterways

The inland waterway consists of enclosed geographic boundaries, channels and wharves and anchorages within the water. Ships enter the waterway through the channel entrance at a certain velocity and depart from the waterway through the channel exit. Ships may berth in this area, which is determined by the existence of wharves and anchorages in the waterway and the berthing requirements of the ship. As a result, the velocity of ships with berthing behavior will first decrease and then gradually increase, and the concentration of ships in the stop area will increase. Fig. 4 illustrates the velocity and location changes of ships. Trajectory 1 shows that the ship sails normally and the velocity does not fluctuate greatly. Trajectory 2 shows that the ship stopped at a wharf or anchorage, and the velocity first decreased and then increased.

For traffic pattern analysis in inland waterways, the paper classifies ship trajectories that follow identical routes from the same origin to the same destination as one traffic pattern. As previously noted, inland ships typically navigate along the channel's direction, making it crucial to accurately determine the origin and destination areas. These areas are identified by setting judgment areas at the waterway boundaries, based on prior knowledge. Judgment areas serve as potential origins and destinations within the area of concern. The paper then align these judgment areas with the start and end points of ship trajectories. A trajectory is assigned to a traffic group if its start and end points correspond to a specific pair of judgment areas.

The area judgment process, detailed as Algorithm I, involves two key components within the departure and arrival zones: (1) the intersection between the channel and the waterway, which acts as the entrance and exit, and (2) the wharves or anchorages, if present. While the intersections always exist, wharves or anchorages might not and are considered as OD (Origin-Destination) child nodes within the area. Trajectories that pass through these child nodes are identified as new traffic patterns, yet they still intersect with the main entrances and exits. The pattern extraction process, therefore, begins with determining the entrances and exits, followed by stop detection on the identified patterns, as illustrated in Fig. 5.

Algorithm 1. Area Judgment

Input: Coordinates of area,ship trajectory
Output: Trajectory cross results
$\boldsymbol{P_1}$: The first point of trajectory, that is origin; $\boldsymbol{P_n}$: The last point of trajectory, which is the destination
1: **Begin**
2: **If min**(area A.longitude)< $\boldsymbol{P_1}$.longitude < **max**(area A.longitude)
3: and **min**(area A.latitude)< $\boldsymbol{P_1}$. latitude < **max**(area A. latitude), **then**
4: $\boldsymbol{P_1}$ in area A = = True
5: **End if**
6: **If min**(area B.longitude)< $\boldsymbol{P_n}$.longitude < **max**(area B.longitude)
7: and **min**(area B.latitude)< $\boldsymbol{P_n}$ latitude < **max**(area B. latitude), **then**
8: $\boldsymbol{P_n}$ in area B = = True

(*continued on next page*)

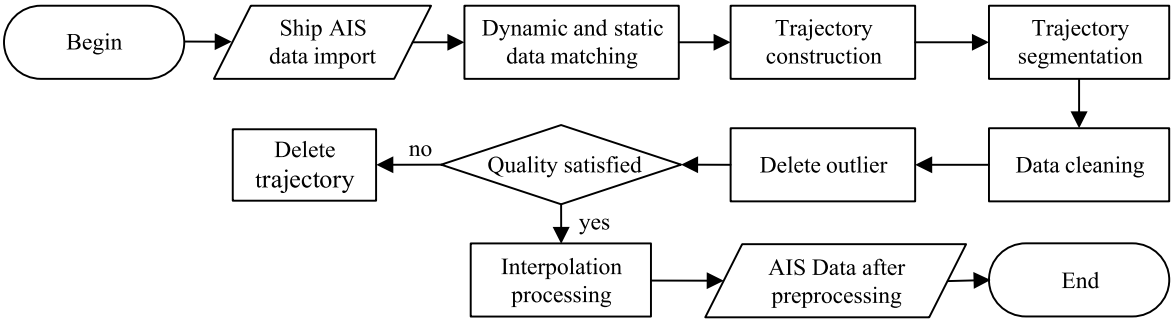


Fig. 3. The flowchart of AIS data processing.

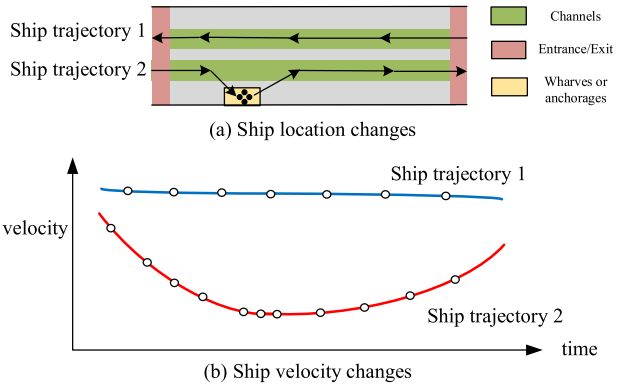


Fig. 4. Schematic diagram of ship velocity and location changes in inland waters.

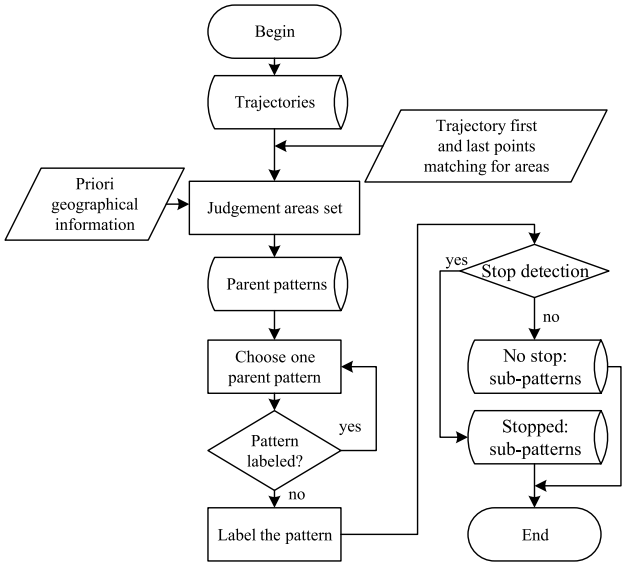


Fig. 5. The flow chart of traffic pattern division.

(*continued*)

9: **End if**
10: **If** P_1 in area A = = True, P_n in area B = = True, **then**
11: trajectory cross area A, area B
12: **END**

Fig. 6 illustrates an example of extracting traffic patterns based on trajectory analysis. Trajectory 1 and Trajectory 2 both pass through Areas A and B, but in different orders, resulting in their separation into distinct traffic patterns. Similarly, Trajectory 4 and Trajectory 6 are treated as separate patterns. Trajectory 3, and Trajectories 5 and 7 also traverse through Areas A and C. However, the sequence of areas differs between Trajectory 3 and Trajectories 5 and 7, leading to further differentiation. Additionally, Trajectory 7 includes a stop, which distinguishes it from Trajectory 5, even though both share the same route through Areas A and C.

3.2.2. *Detection of stop behavior*

To explore the stop behavior and potential stop areas in the waterway, stop detection is necessary (Guo et al., 2020), and its schematic diagram is shown in Fig. 7. First, the sliding window is applied to extract the speed sequence, then low-speed sailing trajectory without stopping is removed, and finally the sequence of acceleration and deceleration segments are removed.

3.2.2.1. Sliding window to extract the speed sequence. The sliding window algorithm is applied to initially extract the stop trajectory sequence by velocity. N is the window length threshold of the sliding window. If ship trajectory repeatedly shows more than certain number of points N with a velocity of less than m knots, it is considered that the ship has stop behavior in the area. Generally, N can be taken as 40, and m is 2. The schematic diagram of the sliding window is shown in Fig. 8. Using sliding window, a series of stop trajectory segments of the ship are obtained. However, these trajectory segments may include the case where the ship is moving at low velocity without stopping, for which the paper can restrict it based on the farthest opposite point distance of the trajectory segment.

3.2.2.2. True stop sequence extraction. Before berthing or anchoring, ships navigate at low speed and decelerates for a short distance, then gradually increase speed and depart from the stop area after completing the berthing operation. During this period, the distance between adjacent trajectory points of the ship will first decrease and then increase. To obtain the trajectory of the stop part, the trajectory points of the acceleration and deceleration segments (before and after the stop) need to be eliminated. Fig. 9(a) shows the plane coordinate curve of the cumulative value of the distance between adjacent trajectory points of the ship varying with time. When the cumulative distance changes with time, the growth rate will first magnify, then reduce, and then magnify again. A piecewise linear function is proposed to fit the plane coordinate curve to calculate the function breakpoint position. After getting breakpoint positions, the paper can retain the segment with the smallest derivative as the ship's mooring or anchoring, as shown in Fig. 9(b).

The general equation system of piecewise linear function is as Eq. (2).

$$y(x)=\begin{cases}\eta_1+\mu_1(x-b_1) & b_1<x\le b_2\\ \eta_2+\mu_2(x-b_2) & b_2<x\le b_3\\ \vdots & \vdots\\ \eta_n+\mu_n(x-b_n) & b_n<x\le b_{n+1}\end{cases} \tag{2}$$

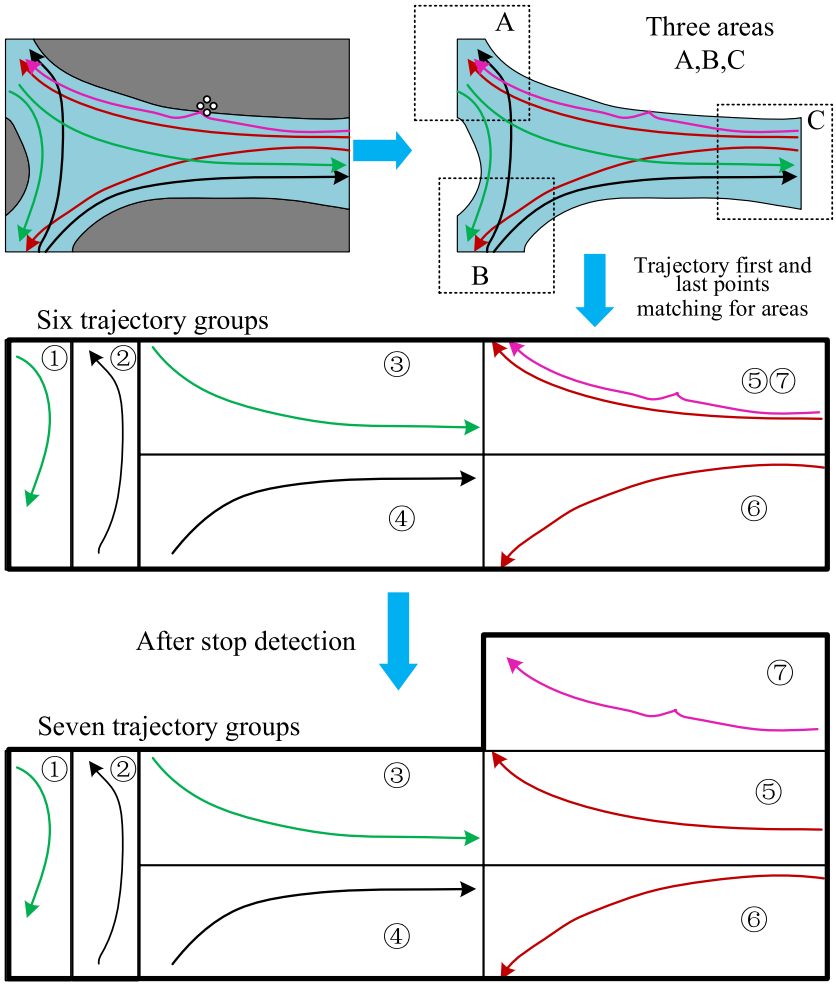


Fig. 6. Schematic diagram of pattern division by setting judgment area and stop detection.

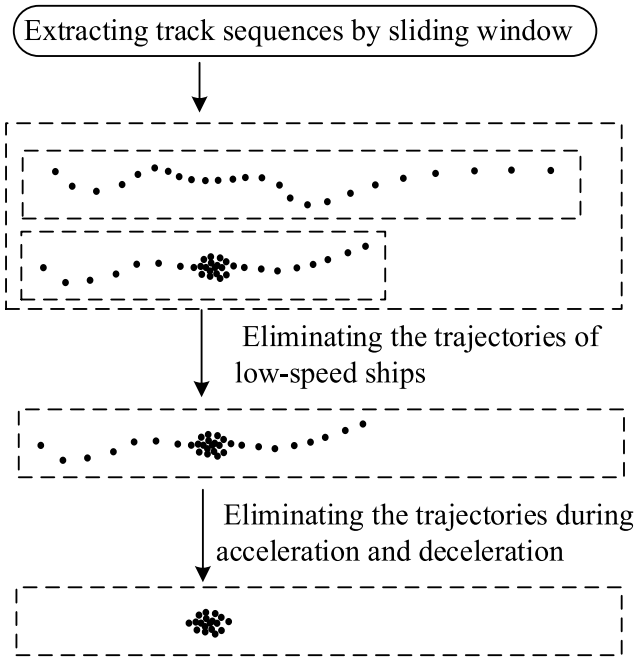


Fig. 7. Schematic diagram of extraction of stop points.

where n is the number of segments, and $b_1 \ldots b_{n+1}$ is the abscissa of the break points.

Generally, the paper can divide the accumulative plane curve into three sections, and there are four breakpoints, that is, n is 3. Eq. (3) can be represented by Eq. (2).

$$y(x) = \begin{cases} \beta_1 + \beta_2(x - b_1) & b_1 < x \le b_2 \\ \beta_1 + \beta_2(x - b_1) + \beta_3(x - b_2) & b_2 < x \le b_3 \\ \beta_1 + \beta_2(x - b_1) + \beta_3(x - b_2) + \beta_4(x - b_3) & b_3 < x \le b_4 \end{cases} \tag{3}$$

In matrix form, it can be expressed as Eq. (4).

$$\begin{bmatrix} 1 & x_1 - b_1 & (x_1 - b_2) \bullet D_{x_1 > b_2} & (x_1 - b_3) \bullet D_{x_1 > b_3} \\ 1 & x_2 - b_1 & (x_2 - b_2) \bullet D_{x_2 > b_2} & (x_2 - b_3) \bullet D_{x_2 > b_3} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_m - b_1 & (x_m - b_2) \bullet D_{x_m > b_2} & (x_m - b_3) \bullet D_{x_m > b_3} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} \tag{4}$$

where m is the number of sample points, and β is an unknown parameter. $D_{x_j > b_k}$ is a dummy variable, as shown in Eq. (5).

$$D_{x_j > b_k} = \begin{cases} 1 & x_j \ge b_k \\ 0 & x_j < b_k \end{cases} \quad 0 \le j \le m, 2 \le k \le 3 \tag{5}$$

The purpose of piecewise fitting is to know the position of the breakpoint, then the problem is transformed into how to obtain the piecewise fitting situation when the breakpoint position is unknown. For any given breakpoint position b, a least squares fitting can be performed to solve for the β parameters that minimize the sum of squared errors of

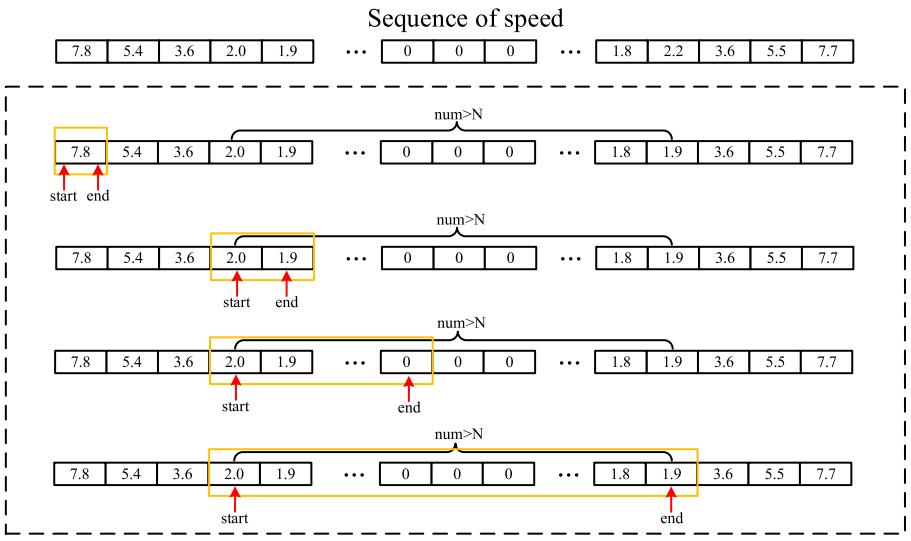
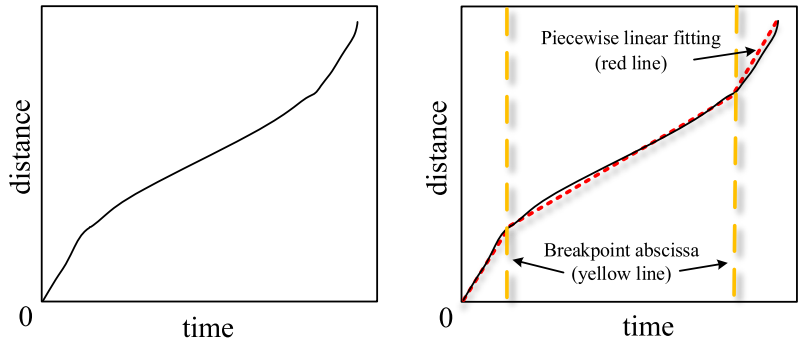


Fig. 8. Extraction of speed sequence by sliding window.



(a) Cumulative distance (b) Piecewise linear fitting for breakpoints

Fig. 9. Schematic diagram of segmented linear fitting.

residuals. The sum of squared residuals can be expressed as a function of the breakpoint position $SSR(\boldsymbol{b})$. Knowing n is 3, the first breakpoint is $b_1 = x_1$ (the smallest x), the last breakpoint is $b_4 = x_m$ (the largest x), then there are 2 unknown breakpoints. The formulation of the optimization problem is to find the breakpoint position so that the overall residual sum of squares is minimized. The optimization problem is summarized as shown in Eq. (6).

$$\begin{aligned} &\min\ SSR(\boldsymbol{b}), \boldsymbol{b} = [b_2, b_3]^{\mathrm{T}} \\ &s.t.\quad x_1 \le b_k \le x_m, k = 1, 2, 3, 4 \end{aligned} \tag{6}$$

When the number of segments is known, differential evolution algorithm (DE) and global optimization is utilized to find the breakpoint location that minimizes the sum of square errors. Differential evolution is an optimization algorithm based on swarm intelligence theory, which is an intelligent optimization search algorithm generated by cooperation and competition among individuals within a population. The principle of differential evolution algorithm is as Algorithm II.

Algorithm 2. DE

Input: Population: N_p; Dimension: D; Max generation times: G_{max};Mutation factor F; Crossover rate: CR
Output: Optimal vector
1: **Begin**
2: generate initial population/*initialization*/
3: **while** $G < G_{max}$ **do**

(*continued on next column*)

(*continued*)

4: **for** $i = 1$ to N_p **do**
5: Randomly select three indexes r_1, r_2, r_3 and $r_1 \neq r_2 \neq r_3 \neq i$
6: $v_i^G = x_{r_1}^G + F\left(x_{r_2}^G - x_{r_3}^G\right)$/*mutation*/
7 j_{rand} = randint(1, D)/*crossover*/
8: **for** j = 1 to D **do**
9: **if** rand$[0, 1] \le CR$ or $j == j_{rand}$
10: $u_{i,j}^G = v_{i,j}^G$
11: **else**
12: $u_{i,j}^G = x_{i,j}^G$
13: **end if**
14: **end for**
15: **if** $f(u_i^G) \le f(x_i^G)$/*selection*/
16: $x_i^{G+1} = u_i^G$
17: **else**
18: $x_i^{G+1} = x_i^G$
19: **end if**
20: **end for**
21: $G = G + 1$
22: **end while**
23: **END**

3.3. Traffic pattern prediction based on decision tree

3.3.1. Introduction of decision tree model

Currently, many artificial intelligence algorithms possess an issue of being uninterpretable, leading humans knowing little about their decision-making processes and reasons. Such opaque AI systems fail to gain the understanding and trust of users, hindering deep integration of AI across various fields. Explainable machine learning refers to machine learning methods that present the model's decision-making process and outcomes in a form comprehensible to humans. The primary aspects of explainability include model transparency, the importance assessment of features, and the interpretability of results. Based on this, this study proposes a traffic pattern prediction model based on decision tree, which is highly explainable and easy to understand.

Decision tree is a supervised learning classification algorithm that is widely used in machine learning (Abreu et al., 2023; Yan et al., 2020b). It builds a tree-like structure of classification rules by inductive reasoning. Each path from the root node to the leaf node of the decision tree forms a rule. The features of the internal nodes on the path correspond to the conditions of the rule, and the classes of the leaf nodes correspond to the conclusions of the rule. The decision tree model divides the feature space into a limited number of disjoint sub-regions through a set of such a series of decision rules. For samples falling in the same sub-region, the decision tree model gives the same prediction value. Compared with other supervised learning classifiers, such as neural networks, decision trees have the advantages of fast classification speed, high efficiency, and no need for standardization of input data. At the same time, decision trees are easy to understand and explain, suitable for visualization, and can be converted into If-Then rules. The phenomena observed in the model can be easily explained by logical analysis.

The interpretability advantages of decision trees are:

(1) The rules are clear. The decision tree makes decisions from top to bottom through a series of judgment rules. These rules form an easy-to-understand tree structure. Each decision node divides the data into two parts based on a threshold of the feature. The threshold of the feature can be ship speed, course etc.
(2) Intuitive expression. Tree structures can be drawn intuitively, and decisions at each node can be clearly expressed as "If-Then" rules. An intuitive visual tree structure can display the decision-making process of each rule, that is, the precise reason for improving a single prediction.
(3) Feature importance. When training a decision tree, you can calculate how important each feature is to the model's decision. Based on this, the feature weight ranking in the prediction task can be obtained, and then which ship traffic characteristics are more important for traffic mode prediction decisions can be obtained, which can provide reference for maritime traffic control.

The process of generating a decision tree is to determine what attribute should be selected for each layer (node) on the tree for judgment, that is, how to select the optimal partition attribute (of each node). As the division process continues, the samples contained in the branch nodes of the decision tree are expected to belong to the same category as much as possible, that is, the "purity" of the nodes is getting higher. Generally, entropy and gini index are used to measure purity. The usage of entropy mainly includes information gain and information gain ratio, and the gini index mainly corresponds to gini impurity gain (Zhang et al., 2024).

Relevant definitions and descriptions are as follows:

- **Entropy**. Entropy is a measure of the uncertainty of a discrete random variable. When the random variable has only one value, the entropy is 0. When the random variable has more possibilities, the probability distribution among the possibilities is more average, and the entropy is greater. Note that, entropy can only measure the uncertainty of discrete random variables. X is a discrete random variable that takes on a finite number of values. Its probability distribution formula is shown as Eq. (7).

$$P(X=x_i)=p_i, i=1,\ldots,n \tag{7}$$

Entropy of X is calculated as Eq. (8). In the application of the decision tree, the paper actually use the empirical entropy to measure the "purity" of the label value distribution, that is, use the frequency distribution instead of the probability distribution for calculation, as Eq. (9).

$$H(X)=-\sum_{i=1}^{n} p_i \bullet \log p_i \tag{8}$$

$$H(X)=-\sum_{i=1}^{n} \frac{m_i}{M} \bullet \log\left(\frac{m_i}{M}\right) \tag{9}$$

- **Information gain.** Y is a discrete random variable that takes on a finite number of values. Similar to X, the probability distribution of Y is presented as Eq. (10).

$$P\left(Y=y_j\right)=p_j, j=1,\ldots,m \tag{10}$$

The information gain of a random variable X with respect to a random variable Y is defined as the difference between the entropy of Y and the conditional entropy of Y to X. Conditional entropy refers to a measure of the uncertainty of a random event Y given the value of a random variable X, as Eq. (11). In the application of the decision tree, the meaning of conditional entropy is clearer, that is, the sample space is divided into multiple leaf nodes according to the value of the discrete feature X, and the weighted average of the entropy impurity of the sample label Y value on each leaf node.

$$H(Y|X)=-\sum_{i=1}^{n} p(x_i)\sum_{j=1}^{m} p\left(y_j|x_i\right) \bullet \log\left(p\left(y_j|x_i\right)\right) \tag{11}$$

where $p\left(y_j|x_i\right)$ is the conditional probability that the random variable Y takes the value j under a given X.

Therefore, the information gain calculation formula is as Eq. (12).

$$InfoGain(Y,X)=H(Y)-H(Y|X) \tag{12}$$

- **Information gain ratio**. The information gain ratio of X to Y is the ratio of the information gain of X to Y to the entropy of X, as Eq. (13).

$$InfoGainRatio(Y,X)=\frac{H(Y)-H(Y|X)}{H(X)} \tag{13}$$

- **Gini index**. Gini index have similar effects with entropy, and can measure the uncertainty or "impurity" of a random variable's value. When the random variable has only one possible value, the gini impurity is 0. When the number of possible values of the random variable is larger, the value probability distribution is more average, and the gini impurity is greater. Given a data set D, K is the number of categories in D, p_k is the probability of category k in D, Gini impurity is calculated as Eq. (14).

$$Gini(D)=\sum_{k=1}^{K} p_k \bullet (1-p_k) \tag{14}$$

Gini impurity gain is the change in gini impurity after splitting the data set. Gini impurity gain is calculated as Eq. (15).

$$GiniGain(D,A)=Gini(D)-\sum_{v=1}^{u}\frac{|D_v|}{|D|} \bullet Gini(D_v) \tag{15}$$

where A is the attribute feature used for segmentation, and u is the number of different subsets under attribute A. D_v is the sub-dataset v under attribute A after segmentation. $|D_v|$ is the size of the sub-data set D_v, $|D|$ is the size of the total set D.

Three decision tree algorithms are widely used: ID3, C4.5 and CART (Classification and Regression Tree) (Breiman, 2017). Among them, ID3 uses information gain as a division standard and cannot handle continuous values; C4.5 uses information gain ratio as a division standard and can handle continuous values; CART uses gini index as a division standard and can handle continuous values. CART can be used for both classification and regression, while the former two can only be used for classification. When C4.5 and CART classification trees deal with continuous values, they both discretize continuous values. CART is the most versatile due to their ability to handle continuous values and be used for regression forecasting. See Table 1 for details and their comparisons about several typical algorithms of decision trees.

3.3.2. Traffic pattern prediction based on decision tree

Ship behavior is influenced by multiple factors including the navigation purpose, maritime traffic conditions, and the ship's captain. To address the uncertainties associated with ship behavior, particularly the uncertainties of the ship's sailing destination and temporary stopping areas, predictive techniques are employed, as depicted in Fig. 10. While the departure area of a ship entering a zone of interest is known, the destination might not be immediately clear to the shore-based command center.

Data mining techniques such as clustering analysis of historical trajectory data are used to discern intended destinations and specific navigation routes. However, these methods can be limited by temporal delays. By learning from the characteristic features of historical trajectories and aligning them with the current target ship's information, this approach mitigates the uncertainty surrounding the ship's navigational intent. This process is conducted before the actual maneuvering of the ship, providing maritime supervisors with a buffer to strategically dispatch and command, while also ensuring that the crew receives timely feedback from shore-based controls.

Fig. 10 illustrates that the target ship could be headed towards various destinations, with stop behavior also being uncertain. To effectively manage these uncertainties, a traffic pattern prediction model based on a decision tree is proposed.

The paper maps individual ship behaviors onto regional traffic patterns and uses predictions of the traffic patterns to which a ship's trajectory belongs to infer potential future ship dynamics. The traffic pattern prediction model, based on a decision tree, operates through the following mechanism: it extracts traffic patterns from historical trajectories, labels them according to the extracted traffic patterns, constructs a feature dataset, and trains a decision tree classification model. This model is then applied to new trajectories in the water, classifying each new trajectory into a specific category, as shown in Fig. 11. Since the target variable is a discrete label, the research aims to develop a multi-class decision tree model.

This paper uses the decision tree method in the python scikit-learn package to build a traffic pattern classification prediction model, which requires training samples for model construction and evaluation. Scikit-learn uses the optimized CART algorithm to build a decision tree.

Our idea is to analyze regional ship traffic flow to find traffic characteristics that are significantly different in various traffic patterns. The characteristics considered include not only the dynamic traffic characteristics of the ship trajectory (speed, course distribution, etc.), but also the static characteristics of the ship trajectory (such as the origin area and the previous stop detection results). The characteristic form should be specifically related to the area of concern, so the detailed characteristic form is not given here, but is referred to as *Feature* 1, *Feature* 2… *Feature n*. Specific feature names and explanations can be seen in the case study. For ship trajectory segments, an array is constructed as shown in Eq. (16).

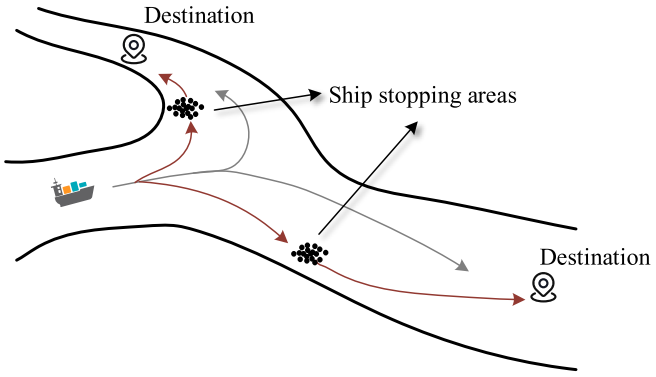


Fig. 10. Schematic diagram of the uncertainty of the ship's sailing destination and temporary stop areas.

$$\{Feature\ 1, Feature\ 2, Feature\ 3, Feature\ 4, \cdots, Feature\ n, \boldsymbol{R}\} \tag{16}$$

where *Feature* 1, *Feature* 2… *Feature n* are the features chosen for decision tree classifier. $\boldsymbol{R}$ is a label vector, indicating the trajectory segment categories. In actual feature selection, the research should focus on features that are significantly different between different label categories in the data set. Table 2 is the normalized sample data format. As indicated in the table, the dataset encompasses n feature parameters and a label vector $\boldsymbol{R}$. The n feature parameters serve as independent variables, whereas the label vector $\boldsymbol{R}$ is the dependent variable. Within the training set, the decision tree autonomously identifies the relationship between the feature parameters and labels, essentially learning the correlation between static and dynamic parameters of ship navigation and traffic pattern categories. In the test set, by inputting a target ship's feature parameters into the trained model, a predictive label can be obtained. This predicted label is then contrasted with the actual label to evaluate the model's effectiveness.

3.3.3. Metrics for model validation

To demonstrate model performance on the test set, four typical classifier performance metrics are used: accuracy, precision, recall, and F1 score. Accuracy is the ratio of the number of correct predictions divided by the total number of predictions. Precision represents the ratio of the number of correct class predictions to the number of class predictions. Recall indicates the number of correctly predicted classes divided by the amount of data. The metric f1 score can be defined as the adjusted accuracy based on the precision and recall metrics, which is the harmonic mean of the two. In general, the calculation formulas of the four metrics are shown in Table 3.

The main parameters for calculating the metrics are described as follow:

Table 1
Details and comparisons for typical decision tree model.

Algorithms	Model use	Tree structure	Classification criteria	Continuous value handling	Missing value handling	Pruning
ID3	classification	Multi-fork tree	Information gain	no	no	no
C4.5	classification	Multi-fork tree	Information gain ratio	yes	yes	yes
CART	classification\regression	binary tree	Gini\MSE(mean square error)	yes	yes	yes

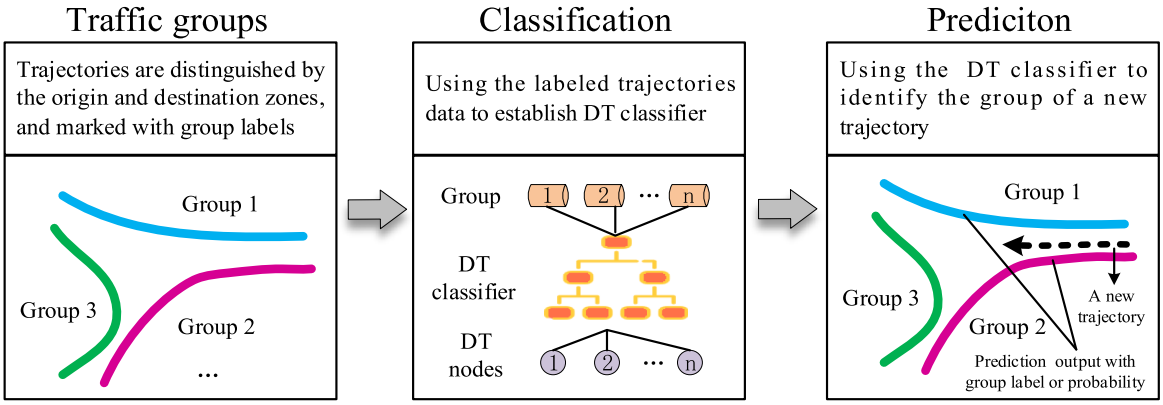


Fig. 11. The entire process of the proposed DT based traffic pattern prediction approach.

Table 2
Sample data format.

Features Trajectories	*Feature* 1 (discrete)	*Feature* 2 (continuous)	*Feature* 3 (continuous)	*Feature* 4 (continuous)	...	*Feature n* (continuous)	***R***
$Traj_1$	1	0.32	0.78	0.56	...	0.65	1
$Traj_2$	2	0.31	0.45	0.53	...	0.72	2
$Traj_3$	2	0.93	0.43	0.63	...	0.23	3

Table 3
Classification metrics.

Accuracy	Precision	Recall	F1 score
$\frac{TP+TN}{TP+TN+FP+FN}$	$\frac{TP}{TP+FP}$	$\frac{TP}{TP+FN}$	$\frac{2}{\frac{1}{Pricision}+\frac{1}{Recall}}$

- True Positive (TP): The true label is positive, and the predicted label is positive.
- True Negative (TN): The true label is negative, and the predicted label is negative.
- False Positive (FP): The true label is negative, and the predicted label is positive.
- False Negative (FN): The true label is positive, and the predicted label is negative.

Since traffic pattern classification is a multi-classification problem, the performance of the overall evaluation classification needs to be considered from the classification of all categories, so the macro-average and micro-average need to be considered when calculating the last three metrics. If the distribution of the data set is unbalanced and the category of small samples is concerned, macro-average can be used; if the category of large samples is concerned, micro-average can be used. Table 4 illustrates the equation for each of the indicators.

where the metrics of the macro average are the arithmetic mean of the metrics of each category. In the equation of micro-average metrics, l_i represents the number of samples predicted by the model as class i and actually belongs to class i, m_i indicates the number of samples predicted by the model as class i, and n_i is the number of samples actually belonging to class i.

Table 4
Classification metrics for macro-average and micro-average.

Metrics	Accuracy	Precision	Recall	F1 score
Macro-average	$\frac{TP+TN}{TP+TN+FP+FN}$	$\frac{1}{k}\sum_{i=1}^{k}P_i$	$\frac{1}{k}\sum_{i=1}^{k}R_i$	$\frac{1}{k}\sum_{i=1}^{k}F1_i$
micro-average		$\frac{\sum_{i=1}^{k}l_i}{\sum_{i=1}^{k}m_i}$	$\frac{\sum_{i=1}^{k}l_i}{\sum_{i=1}^{k}n_i}$	$\frac{2}{\frac{1}{P_{micro}}+\frac{1}{R_{micro}}}$

In addition, the confusion matrix is also a widely used metrics which indicates the result of the classifier in the classification algorithm. Quantity of correctly classified items are on the main diagonal, while values outside of this represents wrong ratings. An example of a confusion matrix for a three-traffic-pattern classification is shown in Fig. 12. From the figure, it is evident that out of the total of 143 data in the test set, 120 have been correctly categorized, appearing on the diagonal, with a few data points misclassified. It is possible to calculate metrics for each pattern separately, which can then be used to evaluate the classification performance of this multi-class model example.

4. Case study

In this section, the method outlined in Section 3 is applied to an AIS dataset from Wuhan. The processed trajectories are visualized and discussed in Section 4.1. Sections 4.2 and 4.3 explore the traffic patterns in the study waters and analyze the spatial distribution of traffic flow

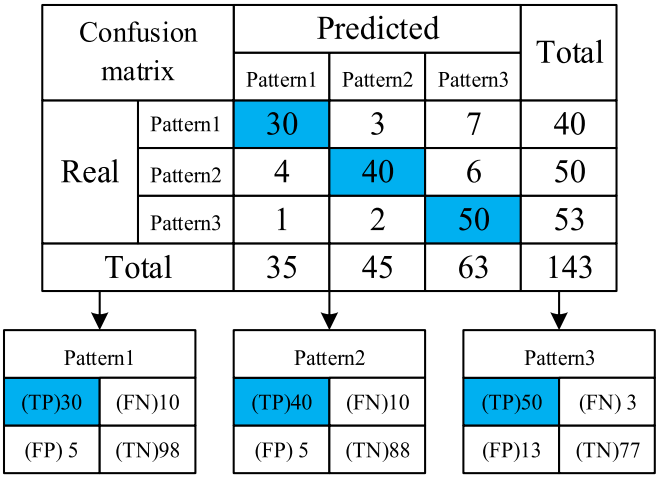


Fig. 12. Illustration of confusion matrix and parameters for metrics in classification problems.

parameters, respectively. Finally, in Section 4.4, a ship pattern prediction model based on a decision classification tree is developed.

4.1. Data visualization

The areas of concern are from 30.511°N to 30.611°N, 114.22°E to 114.34°E, which includes part of the waters of the Hanjiang River. Time span from June to July in 2021. After data preprocessing, a total of 4297 high-quality ship trajectory data were performed, as shown in Fig. 13. Fig. 13 reveals that the ship tracks in the main channel of the Yangtze River are significantly wider than those in the Hanjiang River, indicating a clear navigational advantage of the main channel of the Yangtze River over the Hanjiang River. As ships transition from the Yangtze into the Hanjiang River, they encounter constricted navigational waters. In this confluence area, ship maneuverability is somewhat limited, resulting in increased navigation risks. Moreover, near 114°18′E, 30°34′N, the presence of serrated, irregular protrusions in the ship tracks suggests that the maritime status of these ships might differ from others, possibly indicating stopping behavior.

4.2. Traffic pattern division

The study area is typical convergent area of inland rivers. Judgment areas are in the Hanjiang River, the upper reaches of the Yangtze River, and the lower reaches of the Yangtze River. The judgment area in the Hanjiang River is from 114.215°E to 114.260°E,30.55°N to 30.60°N, and from 114.227°E to 114.308°E, 30.496°N to 30.54° N in upper reaches of the Yangtze River, and 114.282°E to 114.350°E, 30.582°N to 30.620° N in lower reaches of the Yangtze River, as shown in Fig. 14.

The research classified the trajectories of ships crossing the judgment area, and obtained three types of ship trajectories, which navigate between the upper reaches and the lower reaches of the Yangtze River, the upper reaches of the Yangtze River and the Hanjiang River, and the lower reaches of the Yangtze River and the Hanjiang River. The number and proportion of trajectories of each traffic flow are shown in Table 5. The table indicates that the number of ship trajectories between the upper and lower reaches of the Yangtze River is the highest by a significant margin, far exceeding the number of trajectories between the upper Yangtze River and the Hanjiang River, as well as those between the lower Yangtze River and the Hanjiang River. There is also a notable difference between the latter two categories.

In the further division of the three types of trajectories, the paper subdivided them according to the order of crossing areas and whether there is a stop behavior. In theory, 12 types of trajectory clusters will be obtained. In order to avoid repeating the directions of different trajectory clusters, the label of the 12 types of trajectory clusters is explained, see Table 6.

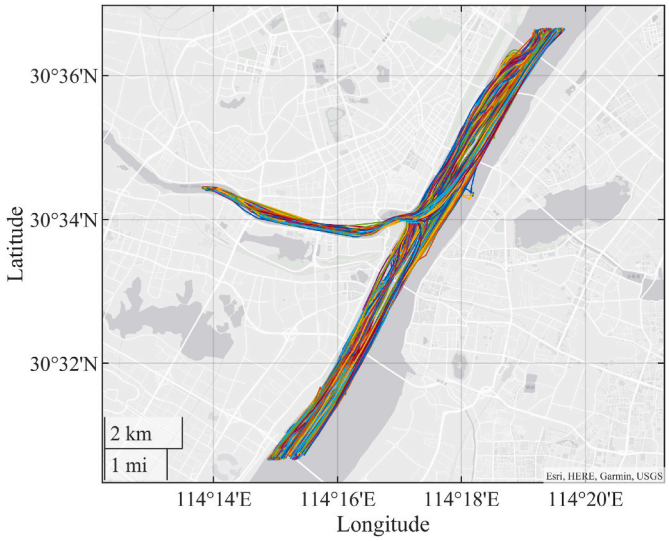


Fig. 13. Trajectories from June to July in 2021.

As the traffic groups division result, total of 8 subsets were obtained, as shown in Table 7. In other words, in the area of concern from June to July in 2021, there are only 8 types of maritime traffic patterns, instead of the theoretical 12 types. Ship trajectories with stop behavior only appear in ship navigating between the upper and lower reaches of the Yangtze River. The statistics of the number and time of ship stops are shown in Fig. 15. The figure indicates that ships typically make a single stop, with one ship having made two stops. The duration of each stop is generally less than 5,000 s, with only a few ships having a layover time between 25,000 and 50,000 s. There is even one ship that has remained stationary for over 100,000 s. The stop duration shows that most ships only make temporary stops rather than loading and unloading cargo.

The visualization of trajectories in each subset are shown in Figs. 16 and 17. Among the ship trajectories with stop behavior, the ship trajectories sailing at low speed instead of stopping are removed according to the length constraint of the stopping trajectories, and they are divided into non-stopping subsets. According to the statistical analysis of the length of the stopping trajectories, the threshold is 800m, as shown in Fig. 18.

After updating the trajectories according to the constraints, the number and proportion of trajectories of each subset are shown in Table 8. The stop points and the major stop areas from the trajectories are identified and the common clusters of points crossing by both traffic flows are highlighted with red circles, as shown in Fig. 19. Taking it as the node of OD, the matching and connection of nodes are completed, and the OD display is shown in Fig. 20.

In addition, the number of trajectories in the waterway and the stop in July 2015 are compared. There are more ship stops in 2015 than in 2021 (up to 4 in 2015 and up to 2 in 2021). The statistics of the number of stops and the stop time are shown in Fig. 21.

The number and flow of traffic streams in 2015 are shown in Table 9. In 2015, the same stop areas for each pair of traffic flows are marked with red circles, green circles, and blue circles, as shown in Fig. 22. Compared with 2021, the distribution of stop areas in 2015 is wider and the number of areas is also larger. The paper speculates that in 2021, compared with 2015, a certain number of wharves were dismantled in the waters, so there will be fewer gathering areas. It is worth noting that there are trajectory points in cluster 12 that are not on the normal route. After restoring the ship's trajectory, it was found that the ship departed from the Hanjiang River and headed towards the upper reaches of the Yangtze River, stopping in the red rectangular marked area, and then heading towards the lower reaches of the Yangtze River, as shown in Fig. 22 (f).

4.3. Parameter analysis

Due to the limited number of ship trajectories traveling between the upper reaches of the Yangtze River and the Hanjiang River, the main analysis focuses on cluster 1, cluster 2, cluster 5, and cluster 6 of ship trajectories in 2021. First, the statistics of their average velocity distribution are performed, as shown in Fig. 23. There is a significant difference in the numerical size between cluster1 and cluster 2. In other words, the average speed holds difference between these two types of trajectories, but it is not obvious between cluster 5 and cluster6. Cluster 5 and cluster 6 have more significant differences in spatial distribution. Then, after gridding the research area with a 500 * 500 grid, statistics are conducted on the trajectory density, average course, course standard deviation, average speed, maximum speed, and minimum speed within the grid.

(1) Visualization of the parameters of cluster 1

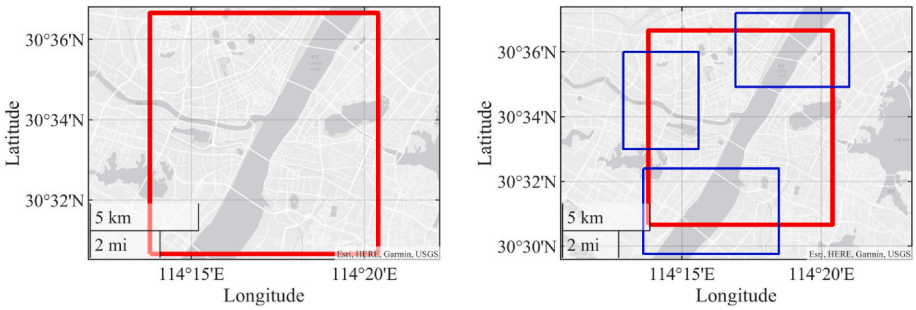


Fig. 14. Schematic diagram of research area (red) and judgment area (bule).

Table 5
Statistics of trajectories for each pair.

Trajectory pairs	number	proportion (%)
The upper and lower reaches of the Yangtze River	4140	96.26
The upper reaches of the Yangtze River and the Hanjiang River	9	0.21
The lower reaches of the Yangtze River and the Hanjiang River	152	3.53

Table 6
Trajectory clusters directions and label description.

Directions		Stop or not	Cluster label
from	to		
The upper reaches of the Yangtze River	The lower reaches of the Yangtze River	no	1
The lower reaches of the Yangtze River	The upper reaches of the Yangtze River	no	2
The upper reaches of the Yangtze River	The Hanjiang River	no	3
The Hanjiang River	The upper reaches of the Yangtze River	no	4
The lower reaches of the Yangtze River	The Hanjiang River	no	5
The Hanjiang River	The lower reaches of the Yangtze River	no	6
The upper reaches of the Yangtze River	The lower reaches of the Yangtze River	yes	7
The lower reaches of the Yangtze River	The upper reaches of the Yangtze River	yes	8
The upper reaches of the Yangtze River	The Hanjiang River	yes	9
The Hanjiang River	The upper reaches of the Yangtze River	yes	10
The lower reaches of the Yangtze River	The Hanjiang River	yes	11
The Hanjiang River	The lower reaches of the Yangtze River	yes	12

Table 7
Statistics of trajectories for each subset in 2021.

Cluster label	number	proportion (%)	Cluster label	number	proportion (%)
Cluster 1	2233	51.97	Cluster 7	13	0.30
Cluster 2	1866	43.43	Cluster 8	28	0.65
Cluster 3	5	0.12	Cluster 9	0	0
Cluster 4	4	0.09	Cluster 10	0	0
Cluster 5	77	1.79	Cluster 11	0	0
Cluster 6	75	1.75	Cluster 12	0	0
			Total	4297	100

The trajectory density map of cluster 1 shows that the density will be higher near the bridge area, which is consistent with our experience and knowledge base, because the width of the ship traffic flow in the bridge area is limited and it will show aggregation. The spatial distribution of course and speed of cluster 1 is relatively uniform. The course is at a small angle (30–50°), the average speed is about 8–10 knots, and the maximum speed is about 14 knots. As shown in Fig. 24.

(2) Visualization of the parameters of cluster 2

The trajectory density map of cluster 2 still shows that the density will be higher near the bridge area, showing aggregation, and even void areas. The spatial distribution of course and speed of cluster 2 is not uniform enough, and the course is at a large angle (210–260°). When the traffic flow passes through the first bridge area (the top bridge), the track band becomes wider, and the ships turn a lot here, temporarily lifting the restrictions on the bridge area. The average speed is around 6–8 knots, and the maximum speed is around 10 knots. The high-speed area is mainly distributed on the right side, suggesting that there may be a distinction between high-speed and low-speed routes. As shown in Fig. 25.

(3) Visualization of the parameters of cluster 5

There is a lot of turning behavior in cluster 5. After entering different waters, the course changes from 220-240° to 260–280°. The speed distribution shows that after entering different waters, the overall speed will increase. This is because the water flow of the Yangtze River has a greater restriction on the speed of the upward ship, while the speed restriction of the Hanjiang River is smaller. As shown in Fig. 26.

(4) Visualization of the parameters of cluster 6

There is also a lot of turning behavior in cluster 6. After entering different waters, the course changes from 100-120° to 30–50°. The speed distribution shows that when entering different waters, the speed will first decrease and then increase. This is because ships from the Hanjiang River to the Yangtze River have deceleration and collision avoidance behaviors when passing through the convergent area. After leaving the intersection area, the restriction is lifted and the speed increases. As shown in Fig. 27.

Figs. 24–27 illustrate that the spatial distribution of the speed and course of cluster1, cluster2, cluster5, and cluster6 show obvious differences. More details in Table 10. Therefore, when constructing the data set of decision tree, speed and course indicators (mean, range, standard deviation, median) were considered.

4.4. Traffic pattern prediction

When building a trajectory classification model based on decision

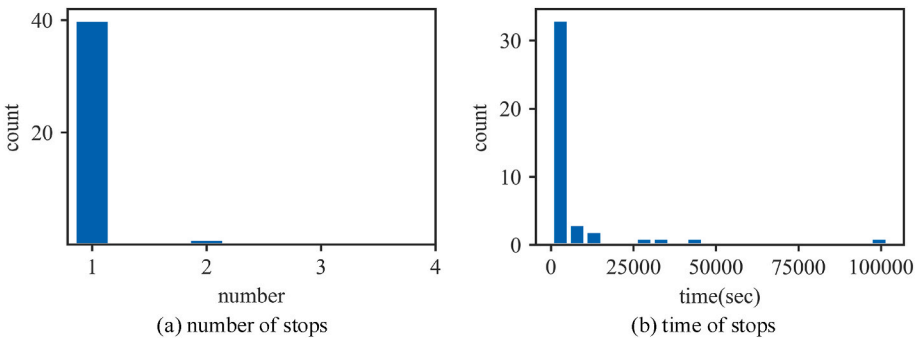


Fig. 15. Statistics of the number and duration of stops.

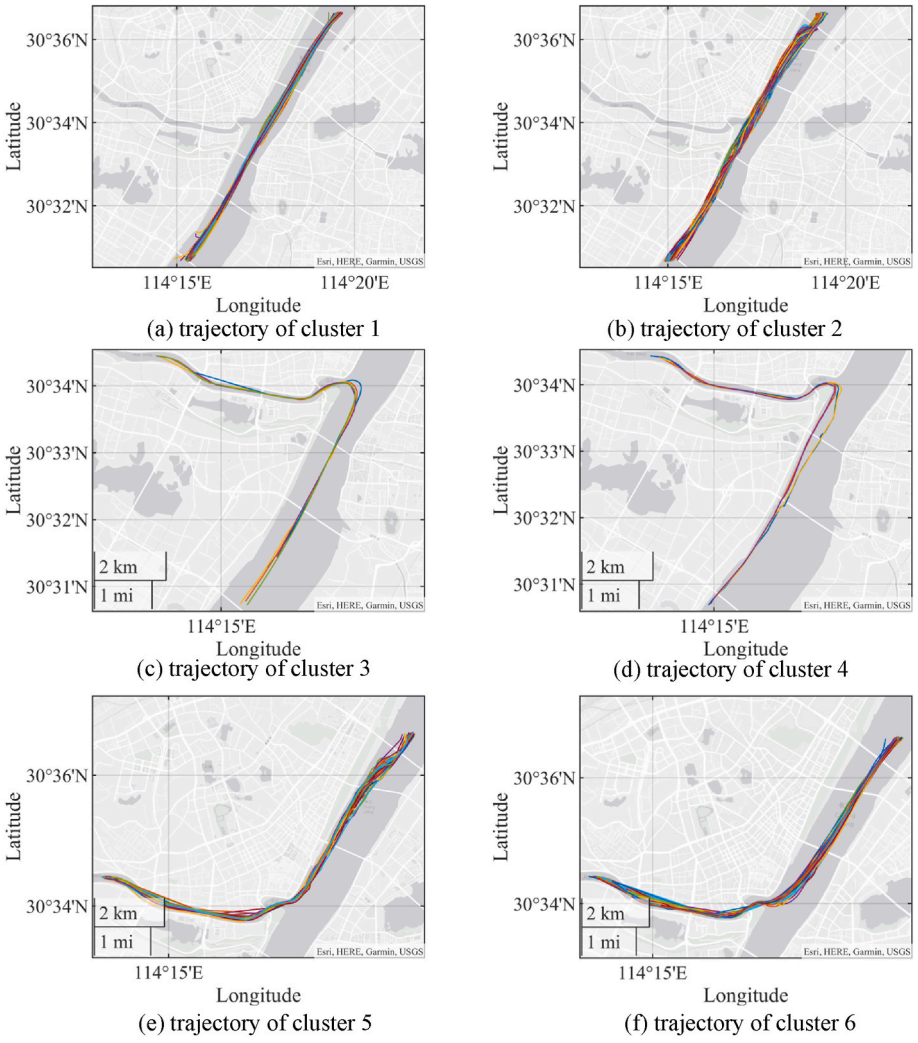
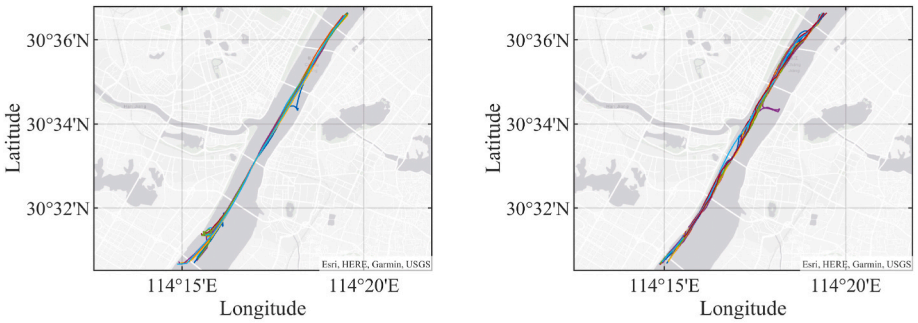


Fig. 16. No stop trajectories of cluster 1–6.

trees, a ship trajectory classification dataset is constructed by manual annotation according to the extracted traffic patterns. The historical trajectory set from June to July 2021 in the area of concern only has 8 categories (There are few or even 0 trajectories in stops). In order to obtain the trajectories with stops between the Yangtze River and the Hanjiang River, the trajectory data in 2015 and other periods in 2021 were extracted by traffic groups to expand the data set of the decision tree classifier to 12 categories. Ultimately 1267 trajectories were extracted from historical trajectories and the data distribution is shown in Table 11. After normalizing the sample data, it was divided into training set and test set according to the ratio of 80% and 20%.

In the previous parameter analysis, different traffic groups showed huge differences in ship course and speed, so ships' course and speed are considered in the dataset features. In addition to these dynamic features, static features of the ship trajectory are also taken into account, that is trajectory departure area and whether to stop during voyages. So, the dataset format is depicted as Table 12.

where *Origin* is the origin of trajectories, labeled by 1,2,3 … etc. Cog_{mean}, Cog_{median}, Cog_{range}, Cog_{std} are the mean, median, range, standard deviation of course respectively, and Sog_{mean}, Sog_{median}, Sog_{range}, Sog_{std} are



(a) trajectory of cluster 7 (b) trajectory of cluster 8

Fig. 17. Stopped trajectories of cluster 7 and 8.

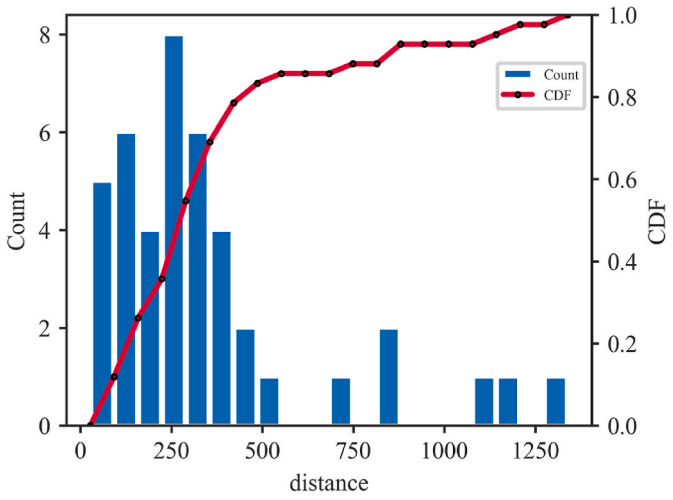


Fig. 18. Statistics of the length of stop trajectory intervals.

Table 8
Statistics of trajectories for each subset after updating.

Cluster label	number	proportion (%)	Cluster label	number	proportion (%)
Cluster 1	2233	51.97	Cluster 7	13	0.30
Cluster 2	1871	43.54	Cluster 8	23	0.54
Cluster 3	5	0.12	Cluster 9	0	0
Cluster 4	4	0.09	Cluster 10	0	0
Cluster 5	77	1.79	Cluster 11	0	0
Cluster 6	75	1.75	Cluster 12	0	0
			Total	4297	100

the mean, median, range, standard deviation of speed respectively. *Stop* is the result of stop detection, as $\{0|1\}$,where 0 represents that the ship stopped, and 1 is the opposite. The value of feature *Origin* is $\{1, 2, 3\}$, where 1 represents the departure area is the upper reaches of the Yangtze River, 2 represents the departure area is the lower reaches of the Yangtze River, and 3 represents the departure area is the Hanjiang River.

Since the distribution of the data set is not uniform enough, when generating a decision tree, it is necessary to balance the data set and calculate the weight. When pruning the decision tree, the grid search is used first to determine the parameters for the high accuracy. The search parameters are as follows:

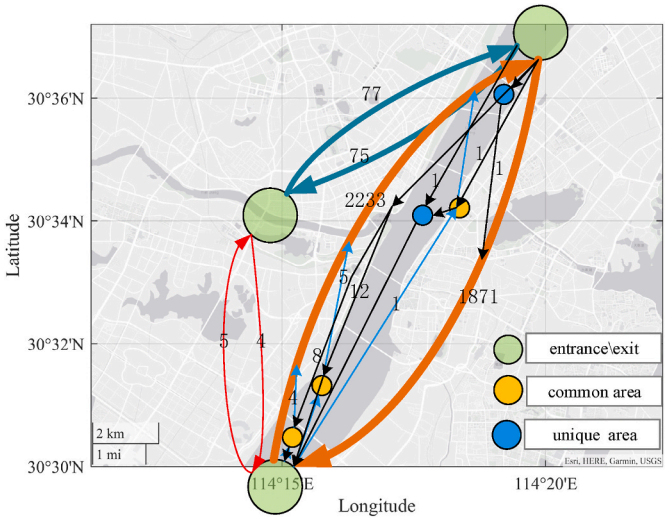
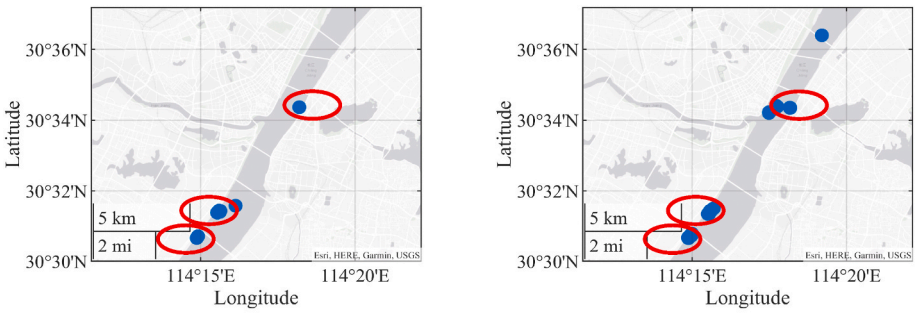


Fig. 20. Traffic patterns presentation in Wuhan waters from June to July in 2021.



(a) stop trajectory point of cluster 7 (b) stop trajectory point of cluster 8

Fig. 19. Stop trajectory points visualization of cluster 7 and cluster 8.

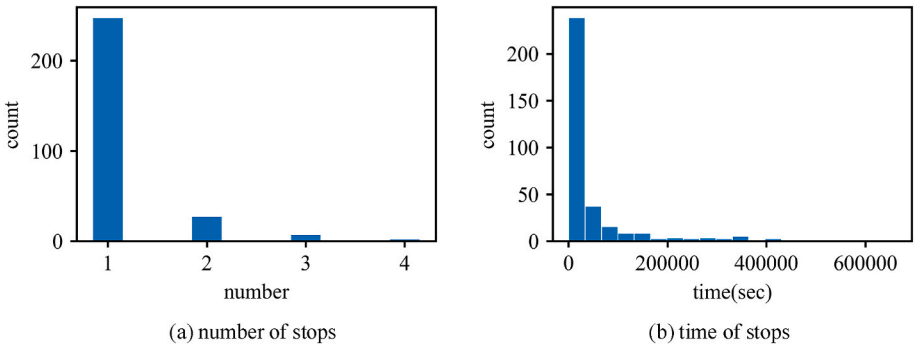


Fig. 21. Statistics of the number and duration of stops in 2015.

Table 9
Statistics of trajectories for each subset in 2015.

Cluster label	number	proportion (%)	Cluster label	number	proportion (%)
Cluster 1	1688	49.21	Cluster 7	90	2.62
Cluster 2	1164	33.94	Cluster 8	157	4.58
Cluster 3	132	3.85	Cluster 9	7	0.21
Cluster 4	42	1.22	Cluster 10	11	0.32
Cluster 5	90	2.62	Cluster 11	13	0.38
Cluster 6	31	0.90	Cluster 12	5	0.15
			Total	3430	100

- **Criterion.** Split criteria. The criterion determines the method used to measure the split quality of the tree. When the criterion value is "gini", the Gini impurity is used to construct the decision tree; when the criterion value is "entropy", the information gain is used instead.
- **Max_depth.** Maximum tree depth. max_depth determines the maximum depth of the tree. When the maximum depth is reached, no further branches can be made.
- **Min_samples_leaf.** The minimum number of samples required at a leaf node. The continuation branch is only considered if a split point of any depth leaves at least min_samples_leaf training samples on each branch.
- **Min_samples_split.** The minimum number of samples required to split an internal node. A node can branch further only if it contains no less than min_samples_split.

The grid search range and final value are shown in Table 13.

Then the cost complexity pruning is performed for post-pruning. Post-pruning is to examine the non-leaf nodes from bottom to top after generating a complete decision tree. If replacing the subtree corresponding to the node with a leaf node contributes the improvement of the generalization performance of the decision tree, then replace the subtree with a leaf node. The main idea of cost complexity pruning is to construct a loss function that takes into account both the impurity and the complexity of the tree, and then use this loss function to guide the decision tree to prune. Scikit-learn achieves pruning effect by directly setting the parameter *ccp_alpha* and calculating the cost complexity path. It cannot directly specify nodes, but to retrain the model to get the pruning model. Therefore, a distribution of model accuracy with different *ccp_alpha* changes was calculated as shown in Fig. 28. As the value of *ccp_alpha* increases, both the test set and the training set exhibit a stepped decline in accuracy. This indicates that with the intensification of post-pruning, the structure of the decision tree becomes simplified, leading to a deterioration in model performance. Consequently, there is a necessity to strike a balance between ensuring high model performance and avoiding overfitting. When *ccp_alpha* is 0.01, the post-pruning effect is better, and it retains a high accuracy while the complexity of the tree is not high. The final decision tree model is shown in Fig. 29.

The importance of features can be calculated based on the reduction in impurity at the node splits of the decision tree. The order of importance of the ten feature variables is in Fig. 30. From the importance of features and their ranking, the feature *Origin* is the most important, followed by *Stop*, then Cog_{std}, and then Cog_{median}, Cog_{mean}, and Sog_{std}. The feature importance of Cog_{range}, Sog_{mean}, Sog_{range}, and Sog_{median} is 0, indicating that these four items have no effect on the classifier. Feature importance reveals which features are more important to the classifier. This is also instructive for maritime practice and situational analysis of ship trajectories.

Therefore, for ships entering Wuhan waters, it is more meaningful to pay attention to their origin area, stop behavior, and the standard deviation and median of their course. Speed has limited effect on distinguishing patterns, although it shows differences in different trajectory clusters, but compared with the previous features, it is not obvious enough.

The confusion matrix of the traffic prediction decision tree model is shown in Fig. 31. To evaluate the performance of the decision tree model, accuracy, precision, recall, and F1 scores were calculated. They are 0.967, 0.809, 0.824, 0.815 respectively. From the classification results of the decision tree, it can be seen that only 11 of the expected 12 categories of results appeared, and category 9 was not included in the classification results. The confusion matrix shows that two trajectories that should have been assigned to category 9 were misallocated to category 7; three trajectories that should have been assigned to category 11 were misallocated to category 8; and one trajectory that should have been assigned to category 8 was assigned to category 11. The visual representation of the decision tree furnishes 12 decision rules, encompassing 11 categorization outcomes. Each rule corresponds to a discrete decision path, with explicit rationale for each decision made. The visualization of the CART tree and the feature importance scores provided reflect interpretability in the context of traffic pattern classification.

The decision tree feature set had continuous variables such as speed and course normalized using min-max scaling. To clearly investigate the impact of real values in the model, the research has retrained using the data that was not normalized. The outcome revealed that models using non-normalized data yielded grid search parameterization results identical to those harnessing normalized data, with no changes in the model's tree structure or its division paths. Nonetheless, when it came to feature splits involving speed and course, the threshold values altered. These thresholds have been transformed into intuitive actual values, more effectively reflecting the influence of real maritime traffic characteristics on traffic model predictions in maritime practices. The model trained on unstandardized data is shown in Fig. 32.

By comparing Figs. 29 and 32, the research found that whether the data is standardized or not does not affect the rule decision-making and classification results of the decision tree. On the contrary, the model obtained by training without standardized data can directly display real values such as speed and course. This also illustrates an advantage of the

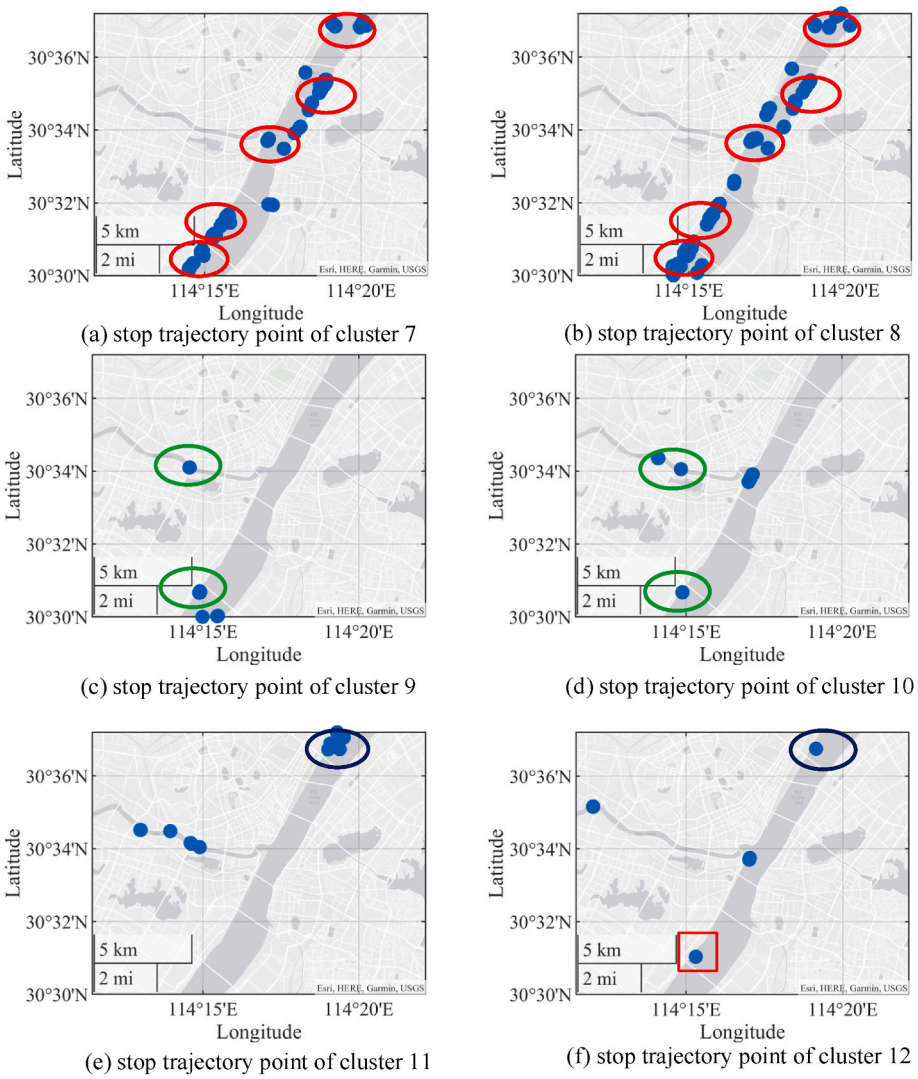


Fig. 22. Stop points and areas in 2015.

decision tree, that is, it does not require strict processing of data and is easy to visualize.

The practical application of this trained algorithm could assist VTS operators in decision-making by predicting the future navigational intentions of ships. Although the real-time nature is somewhat diminished for the complete trajectories used in training, the model can still classify and predict based on the AIS data obtained. Integrating the model into maritime systems provides authorities with essential information on ship routes and stopping behavior. This information can be combined with other methods to further refine predictions of ship behavior at future time points, thereby supporting proactive maritime traffic management. The interpretability of this method is robust, as it facilitates the visualization of decision rules and the model, while also enabling the output of thresholds and importance for decision criteria. This means that managers can directly compare the actual navigational status of ships with the model's outputs and then take further actions, such as controlling ships or assisting in optimizing routes or speed and course adjustments, to achieve coordinated control of regional fleets. The application of this method and its inherent features greatly enhance the potential for system use within the maritime industry.

This method is not intended to replace human intelligence but to serve as a powerful tool to aid in decision-making. It should be noted that to augment the capability for real-time prediction, supplementary experiments were conducted in discussions to explore the predictive power for potential traffic patterns of target ships based solely on current, incomplete trajectory data. The supplementary experiment treated the initial portion of historical trajectories as known information, while the latter part was considered as not having occurred yet. Furthermore, understanding the demands for model reliability and stability in industrial engineering applications, the paper has also addressed improvements in these aspects in our discussions.

5. Discussion

The classification and prediction of complete trajectories is the identification of traffic patterns of ship trajectories when the ship navigation is completed. It has limited practical significance in real-time in maritime traffic management. In order to explore the influence of ships appearing in different periods in the waterway on the prediction of traffic pattern, the historical trajectories fragments of ships with different percentage time lengths are selected as training data, as shown in Fig. 33, see more in Appendix A. Finally, the corresponding decision tree prediction model is performed, and the parameters of each model are shown in Table 14. Appendix A shows the visualization of each model. The different evaluation indicators are shown in Table 15 and Fig. 34.

From the metrics with different percentage time lengths, it can be seen that as the time length increases, the accuracy, precision, recall,

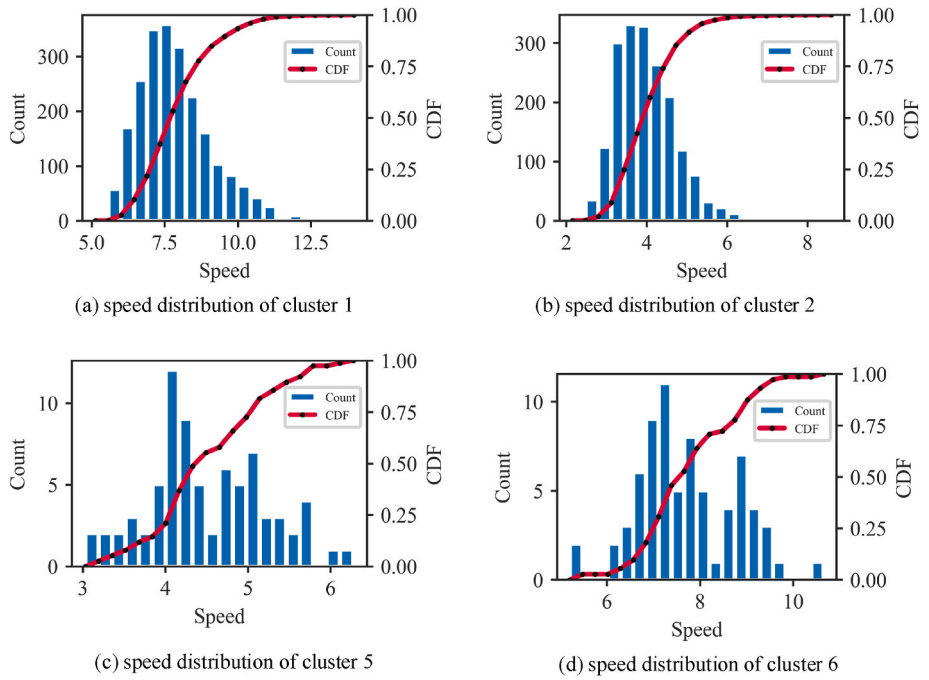


Fig. 23. Speed distribution of cluster 1,2,5, and 6.

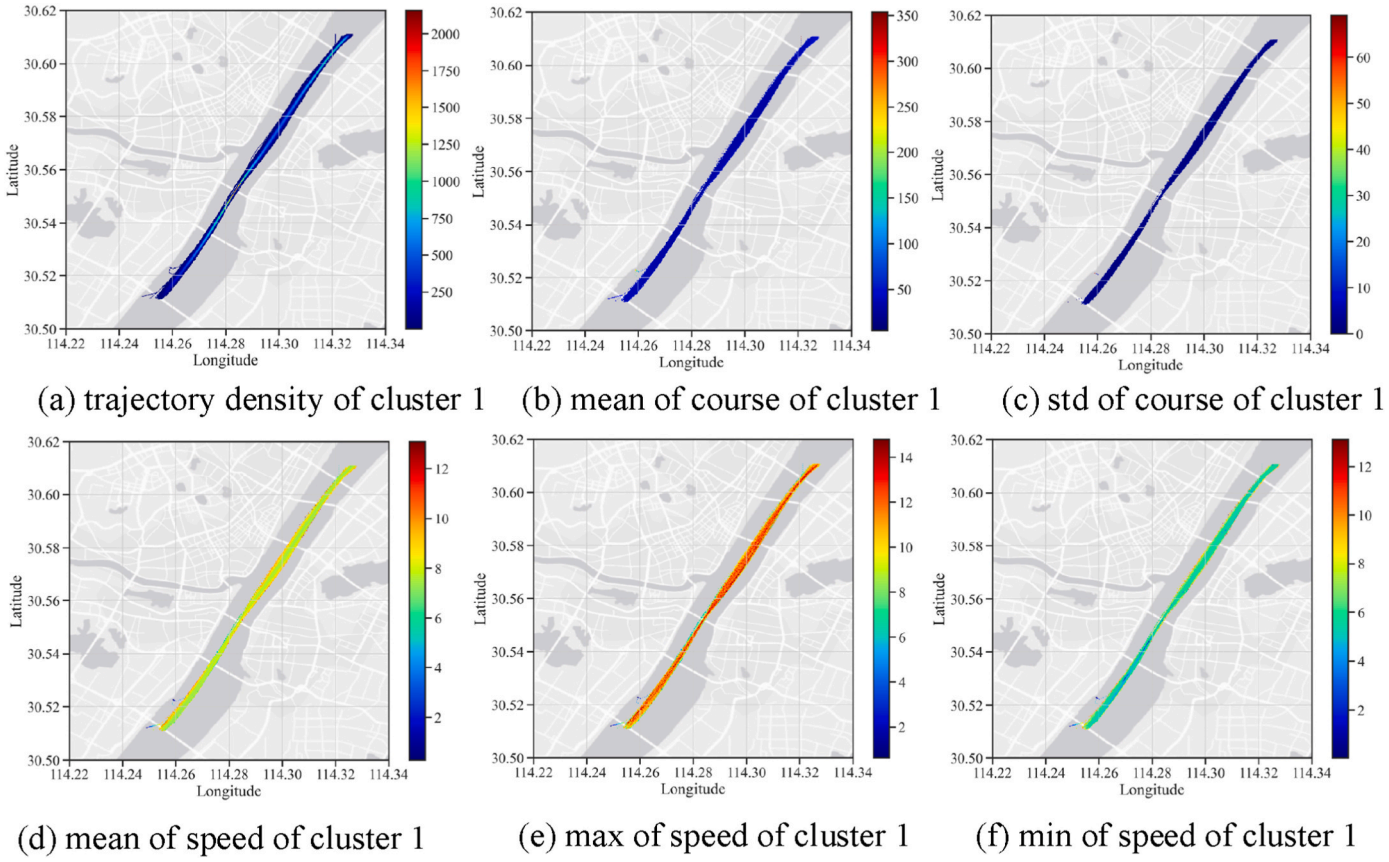


Fig. 24. Traffic parametric statistics of cluster 1.

and F1 score gradually increase. In maritime practice, as the ship sails in the waterway for a longer time, the prediction accuracy of the voyage intention is higher. Specifically, the accuracy rate gradually increased from 0.827 to 0.967, with the precision from 0.541 to 0.809, the recall from 0.597 to 0.824, and the F1 score from 0.554 to 0.815. When the trajectory length reaches 80% of the time percentage, a good prediction accuracy is achieved. Numerically speaking, the accuracy rate is higher than the precision, recall and F1 score, which shows that the imbalance

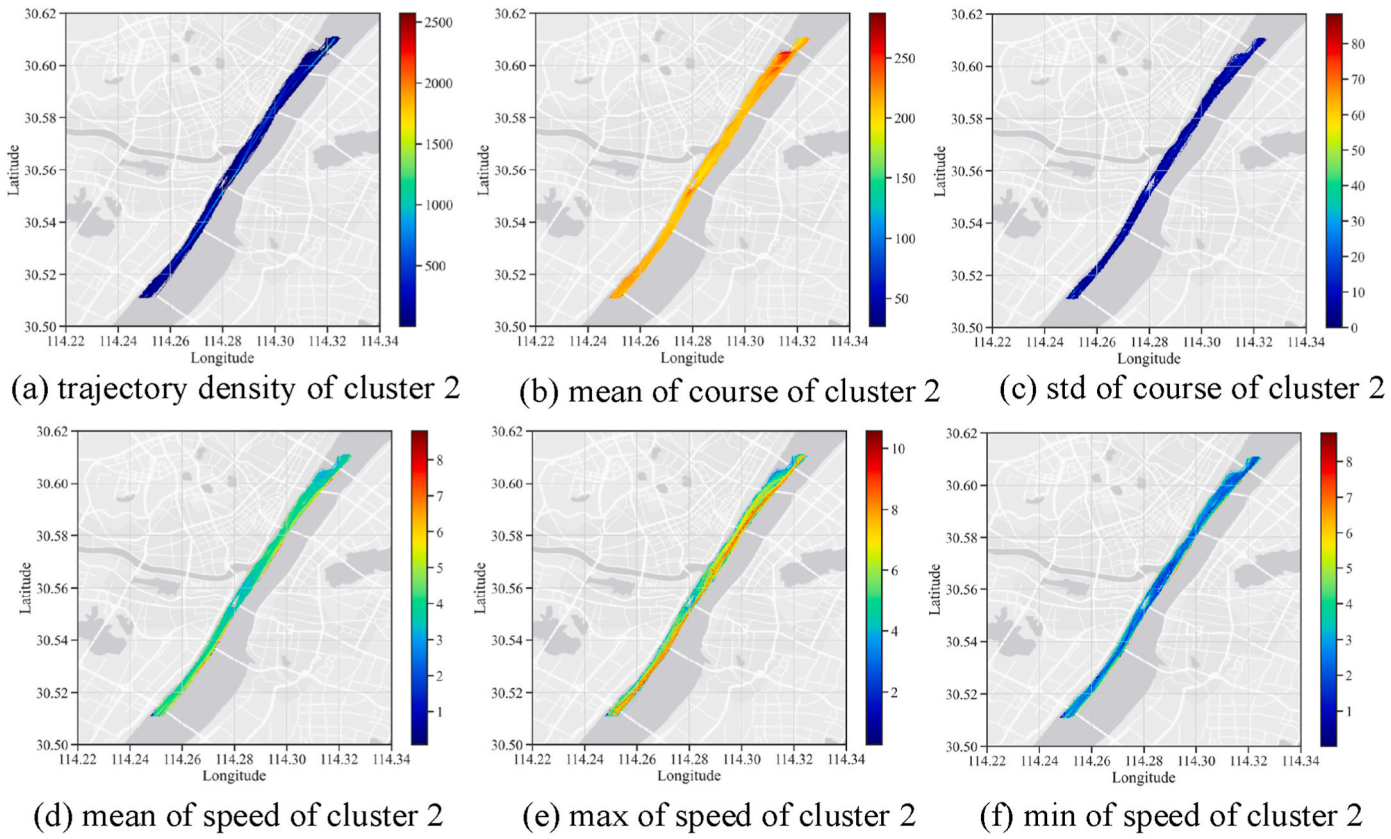


Fig. 25. Traffic parametric statistics of cluster 2.

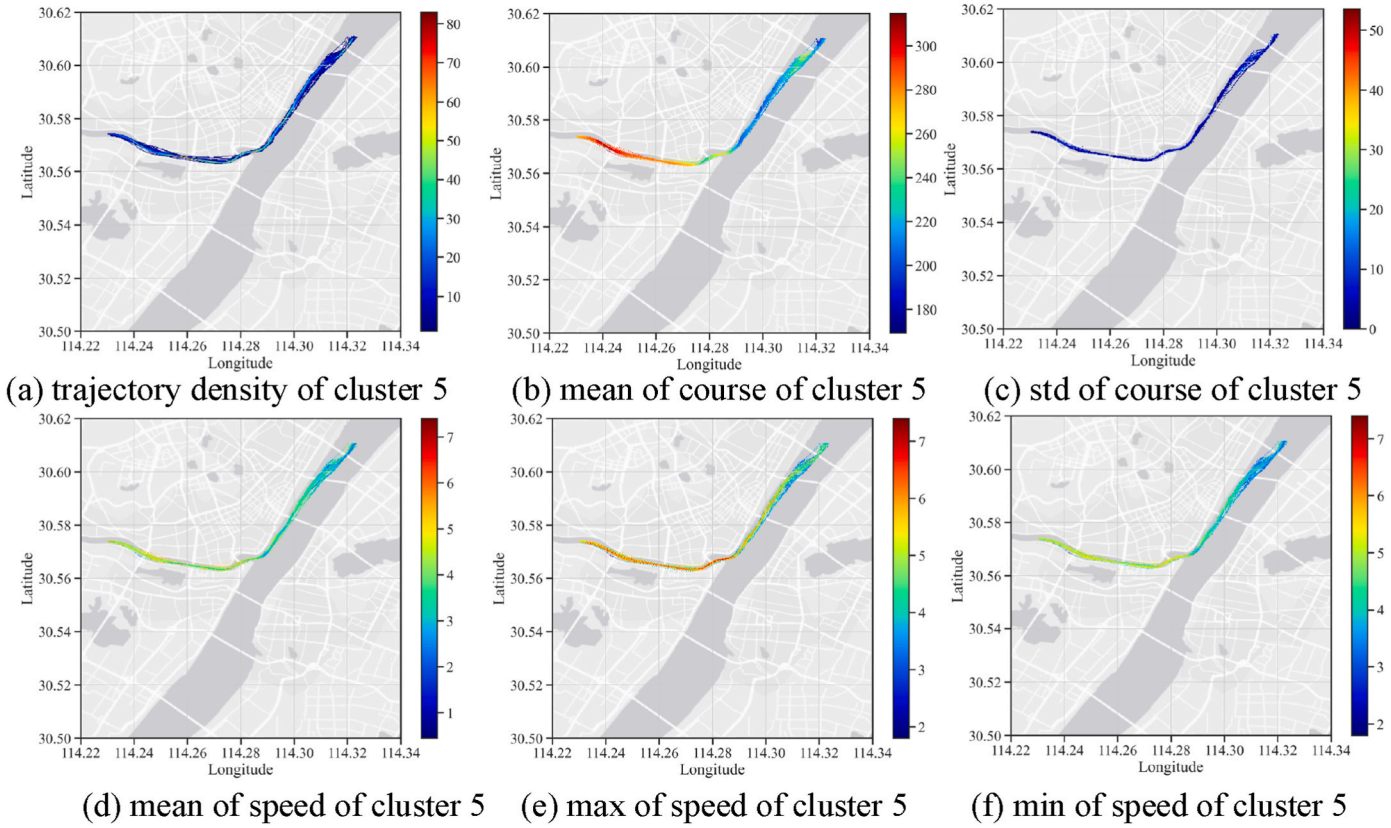


Fig. 26. Traffic parametric statistics of cluster 5.

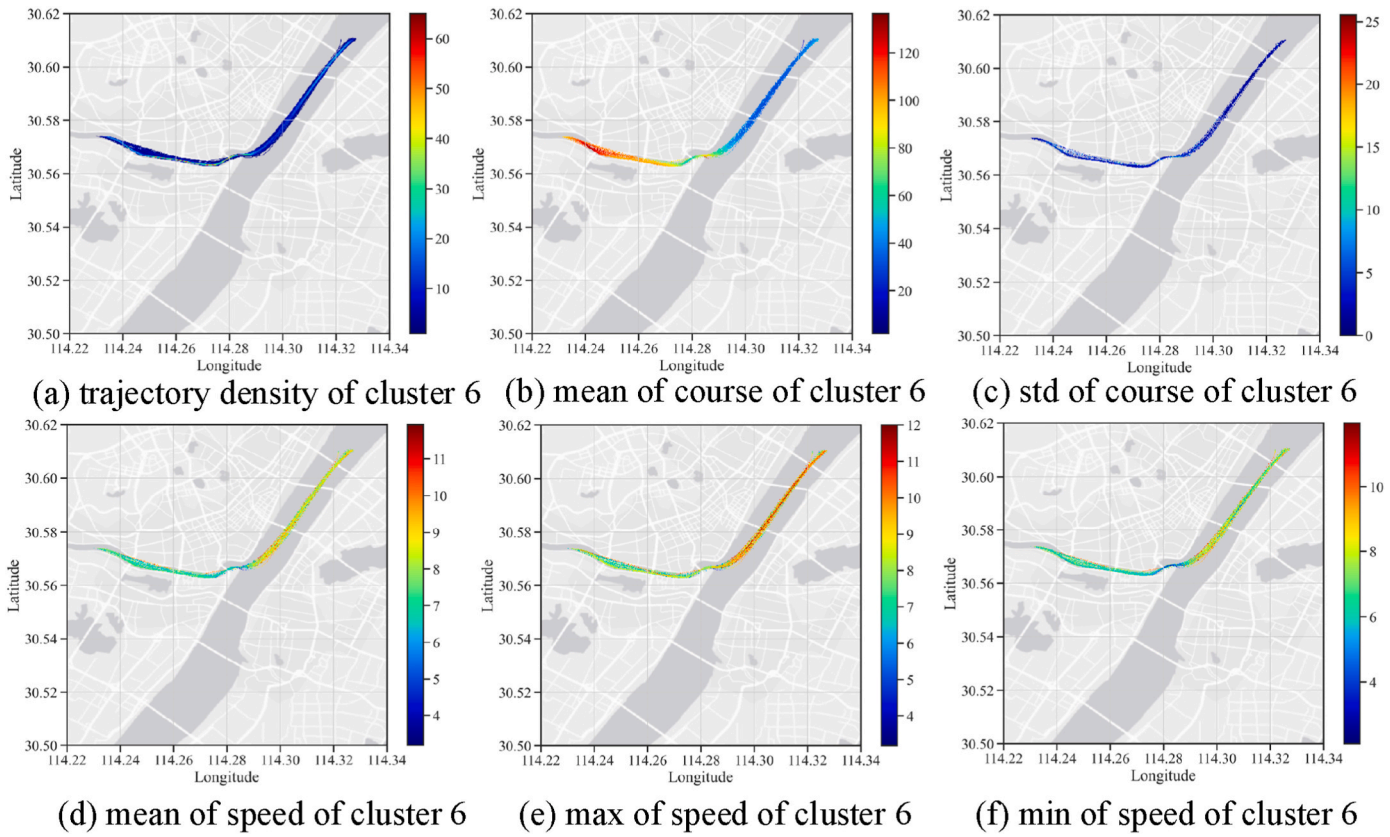


Fig. 27. Traffic parametric statistics of cluster 6.

Table 10
Comparison of speed and course distribution.

	Cluster1	Cluster2	Cluster5	Cluster6
Course	30°–50°	210°–260°	from 220°-240° to 260°–280°	from 100°-120° to 30°–50°
Speed	8kn –10kn	6kn -8kn	–	–

Table 11
Data distribution.

	cluster1	cluster 2	cluster 3	cluster 4	cluster 5	cluster 6
Number	200	200	137	46	166	103
Proportion	0.158	0.158	0.108	0.036	0.131	0.081
	cluster 7	cluster 8	cluster 9	cluster 10	cluster 11	cluster 12
Number	125	247	8	15	14	6
Proportion	0.099	0.195	0.006	0.012	0.011	0.005

of the data set still affects the model results. In Wuhan waters, the number of ships with stopping behaviors or those traversing between the Hanjiang River, and the Yangtze River is relatively small. This demonstrates that the model has indeed achieved a satisfactory classification outcome for the more numerous traffic patterns of ships traveling between the upstream and downstream areas of the Yangtze River. However, ships exhibiting stopping behaviors, or navigating the confluence of the Hanjiang River and the Yangtze River are of particular concern for maritime monitoring. Therefore, in future work, operations such as oversampling on these data will be conducted.

It is important to note that while the research has studied the impact of different percentages of initial trajectory lengths on traffic pattern prediction, there is a significant caveat to consider. When truncating the dataset, the departure area feature of the ships remained unchanged, and dynamic traffic characteristics such as speed and course were recalculated based on the current data. However, the feature indicating whether a ship has stopped was also left unchanged. In fact, the paper retained the initial portion of the complete historical trajectory to simulate the new trajectory of the target ship before completing its navigation. Yet, the determination of whether a stop has occurred is made through a posteriori detection based on the complete trajectory. This creates a temporal inconsistency in the dataset. If a stop is detected

Table 13
Pre-pruning parameters.

parameters	criterion	max_depth	min_samples_leaf	min_samples_split
range	gini/ entropy	[1,10]	[1,20,3]	[2,20,3]
value	entropy	5	1	14

Table 12
Dataset format.

	Origin	Cog_{mean}	Cog_{median}	Cog_{range}	Cog_{std}	Sog_{mean}	Sog_{median}	Sog_{range}	Sog_{std}	*Stop*	***R***
$Traj_1$	1	0.32	0.78	0.56	0.65	0.56	0.42	0.19	0.77	0	1
$Traj_2$	2	0.31	0.45	0.53	0.72	0.78	0.13	0.37	0.30	1	2
$Traj_3$	2	0.93	0.43	0.63	0.23	0.12	0.91	0.34	0.55	0	3

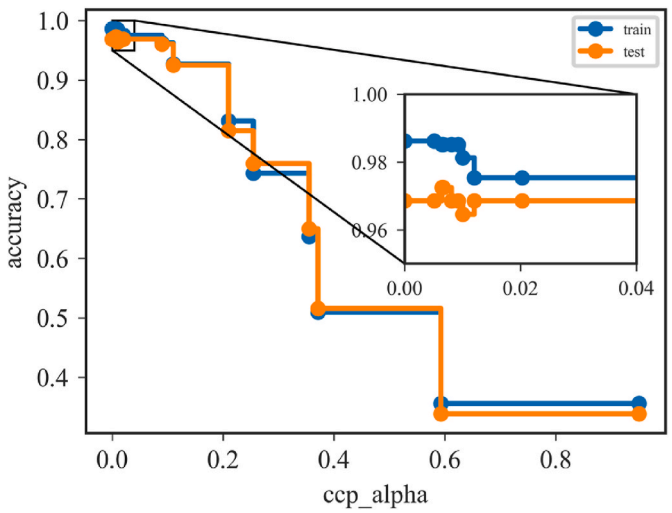


Fig. 28. The relationship between accuracy and *ccp_alpha*.

in the current trajectory and the trajectory is correctly classified into a stopping traffic pattern, it is consistent. However, for stopping behaviors that occur after the truncated time segment, the feature is marked as a

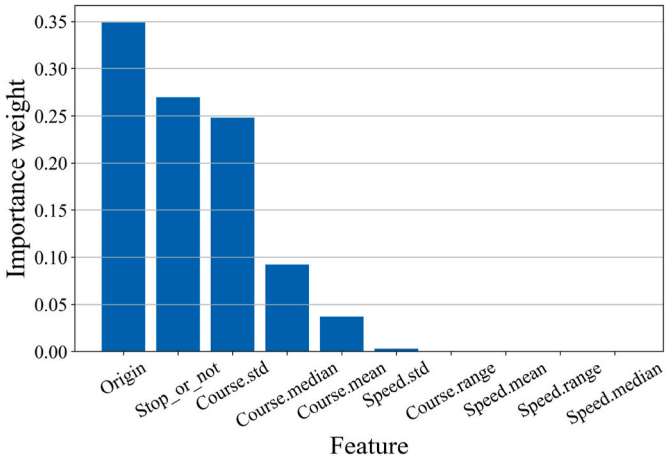


Fig. 30. Feature importance weight ranking of decision tree model.

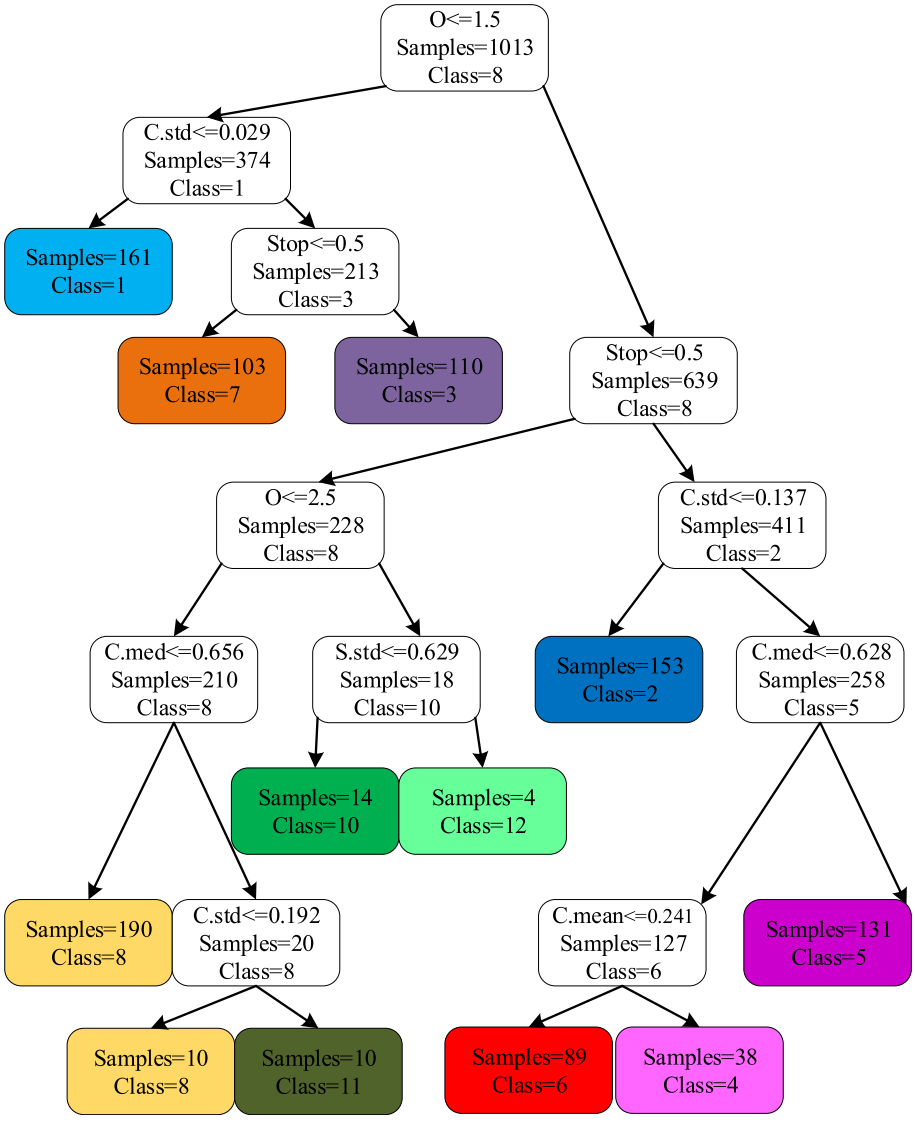


Fig. 29. Decision tree model for traffic pattern prediction.

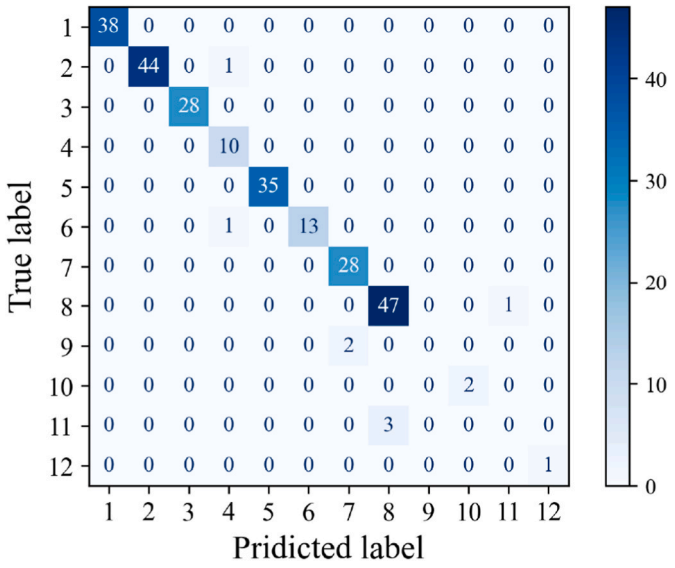


Fig. 31. The confusion matrix of decision tree model.

stop even though the stop has not yet occurred, meaning this feature should be considered indeterminate. In other words, directly truncating without detecting changes for different initial percentages of trajectory lengths presents a potential issue. Unless the model can precisely capture changes in speed and thus predict stopping behavior based solely on velocity, a more accurate model should be developed to predict stops.

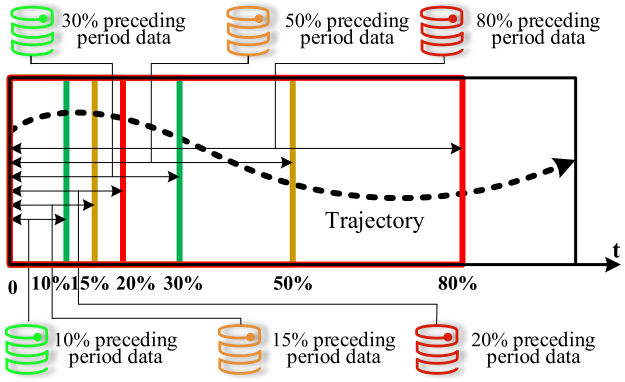


Fig. 33. Trajectory segments with different percentage time lengths.

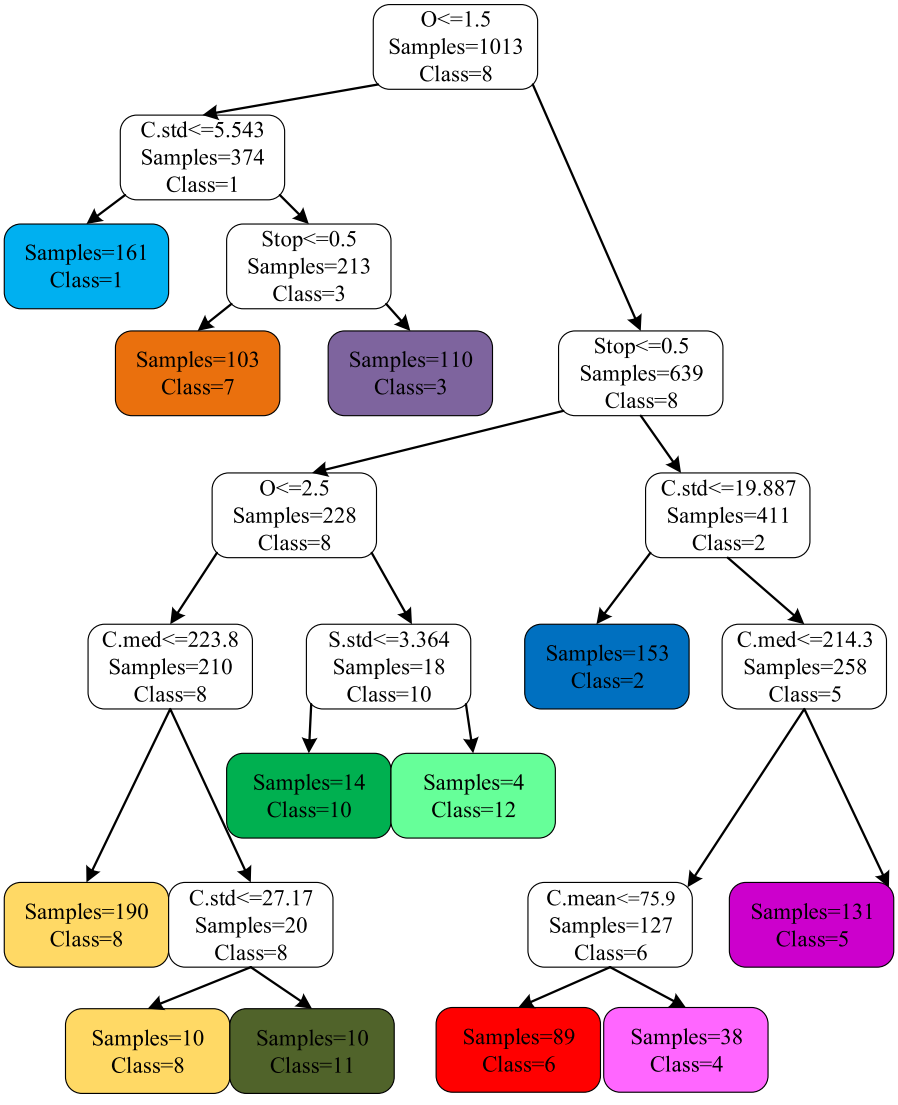


Fig. 32. Decision tree model trained on unstandardized data.

Table 14
Parameters for each decision tree model.

	10%	15%	20%	30%	50%	80%	100%
criterion	entropy	gini	entropy	gini	gini	entropy	entropy
max_depth	7	9	6	6	7	6	5
min_samples_leaf	13	13	13	13	13	1	1
min_samples_split	2	2	2	2	2	8	14
ccp_alpha	0.03	0.02	0.02	0.02	0.005	0.02	0.01

Table 15
Experiment results of evaluation metrics.

	10%	15%	20%	30%	50%	80%	100%
Accuracy	0.827	0.839	0.854	0.902	0.899	0.972	0.967
Precision	0.541	0.543	0.552	0.660	0.669	0.709	0.809
Recall	0.597	0.607	0.617	0.692	0.709	0.748	0.824
F1 score	0.554	0.564	0.574	0.671	0.684	0.725	0.815

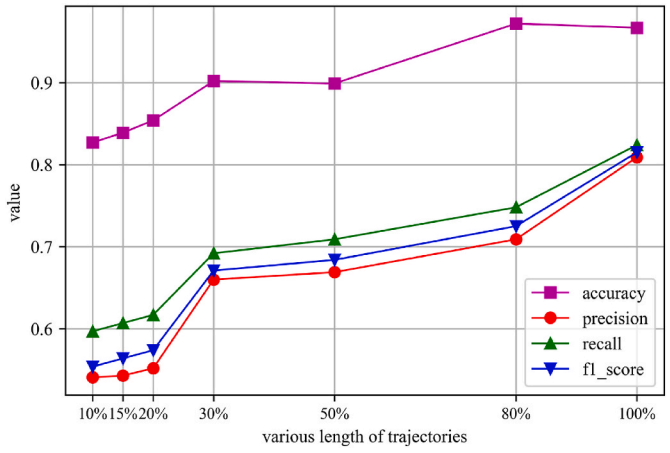


Fig. 34. Results for different time lengths.

The imbalanced metrics indicate a level of instability in the performance of a single CART tree. To enhance the model's generalization capabilities and its reliability for maritime traffic management, there is a need to develop robust models that can effectively guide maritime practices. Both the academic and industrial communities have put forth ensemble learning decision tree methods to address the shortcomings of single trees. In future research, incorporating methods such as RF and GBDT could be promising for the study of traffic patterns. When applying this model for universal verification, an anomaly detection mechanism can be added in the future to help the model identify key patterns and further strengthen management.

In actuality, following the prediction of traffic patterns, the subsequent prediction of ship behaviors stands as an inevitable and challenging subject of research. The connection between traffic pattern prediction and ship behavior prediction is established as follows: once the traffic pattern of the target ship is identified, the historical trajectory cluster that possesses the closest feature parameters and bears the same label is determined, known as the reference set. It is posited that the future individual behavior of the target ship can be deduced in conjunction with the current navigational state, based on the behaviors manifest in the reference set. Correspondingly, a machine learning model can be constructed to automatically discern the coupling relationships of latitude, longitude, speed, and course within the behavior of the reference set's ships. Upon training the model and inputting the current navigational state, a prediction of the future behavior of the ship can be obtained.

6. Conclusion and future work

This paper presents a data mining-then-predict method for maritime traffic management. The proposed method encompasses three primary facets: (1) Mining traffic patterns within the convergent waters of inland rivers. (2) Utilizing a decision tree-based classifier to predict traffic pattern, leveraging the existing traffic clusters. (3) Depicting the impacts of different time-length trajectories on the performance of the classifier. The method validated at the convergence of the Yangtze and Hanjiang Rivers effectively extracts maritime traffic patterns and predicts ship trajectories with high accuracy and precision, enhancing traffic management and congestion identification. Key findings and conclusions are as follows:

- The regional judgment technique for extracting traffic patterns in inland waterways holds substantial promise and utility.
- The identification of stop behaviors aids in delineating high-traffic zones, bearing significant implications for practical maritime supervision.
- The accuracy of the constructed decision tree classifier on complete trajectories attains a notable 96.7% while the precision is 80.9%. Among the indicative parameters, the ship's origin area stands out as the most influential and contributes significantly.
- The extended experiment shows that as the time of the ship sailing in the waterway increases, the accuracy of the ship's behavior prediction increases, from 82.7% of the 10% time-length to 96.7% of the 100% time -length and the precision is from 54.1% to 80.9%.

The limitations of the study are following: (1) Developing automatic methods to collect maritime traffic patterns while considering both dynamic and static characteristics of ship trajectories. (2) Employing stable ensemble learning techniques to enhance the usability and reliability of this study in maritime supervision. (3) Given the certain limitations in the real-time capabilities of AIS, considering integration with radar, video, and other data sources to form a high-quality, real-time, multi-source dataset.

CRediT authorship contribution statement

Zhao Liu: Writing – review & editing, Validation, Supervision, Resources, Methodology, Investigation, Data curation, Conceptualization. **Wanli Chen:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Cong Liu:** Writing – review & editing, Writing – original draft, Validation, Supervision, Software, Resources, Data curation, Conceptualization. **Ran Yan:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Mingyang Zhang:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

We declare that we have no financial and personal relationships with other people or organizations that can inappropriately influence our

work, there is no professional or other personal interest of any nature or kind in any product, service and/or company that could be construed as influencing the position presented in, or the review of, the manuscript entitled, '*A Data Mining-then-predict Method for Proactive Maritime Traffic Management in Inland Waterways by Machine Learning*'.

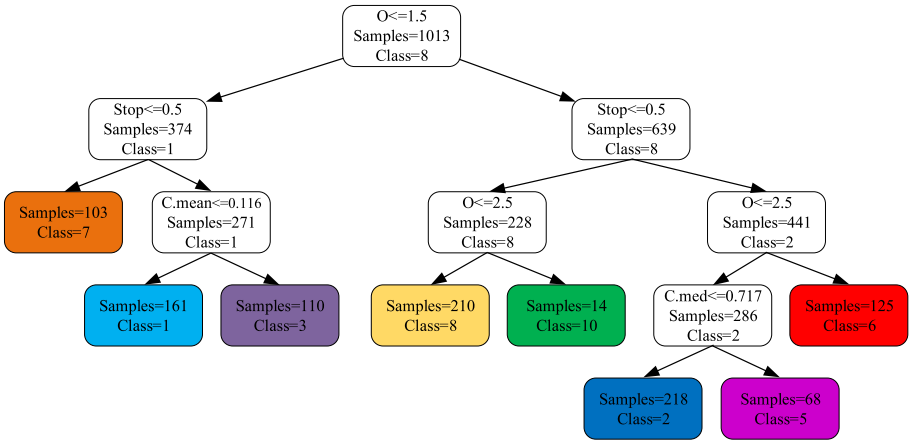
Data availability

Data will be made available on request.

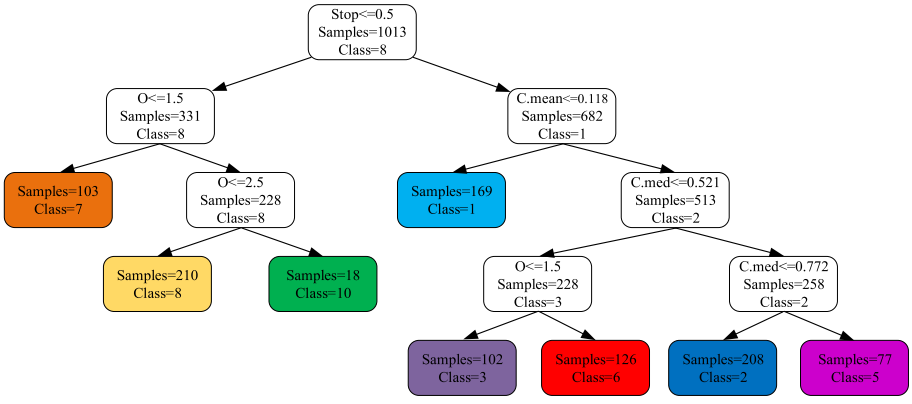
Acknowledgments

This study was supported by the National Natural Science Foundation of China (No.52171351) and the project supported by Sanya Science and Education Innovation Park of Wuhan University of Technology (No. 2022KF0018).

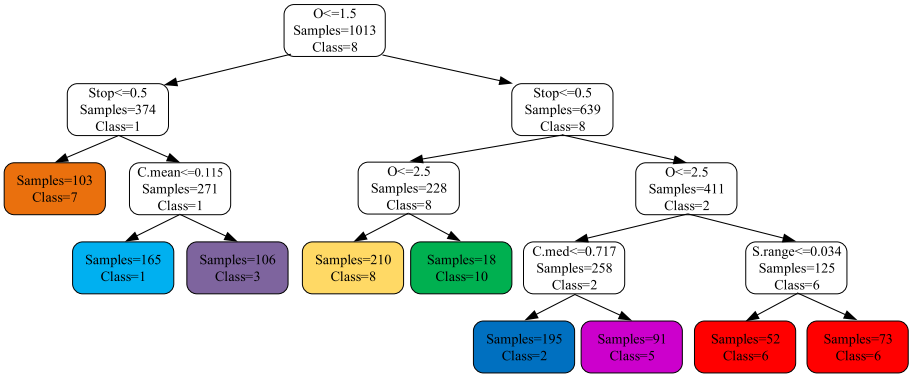
Appendix A. The comparison studies on traffic pattern prediction are based on the different preceding period of navigation



(a) DT prediction model for 10% preceding period



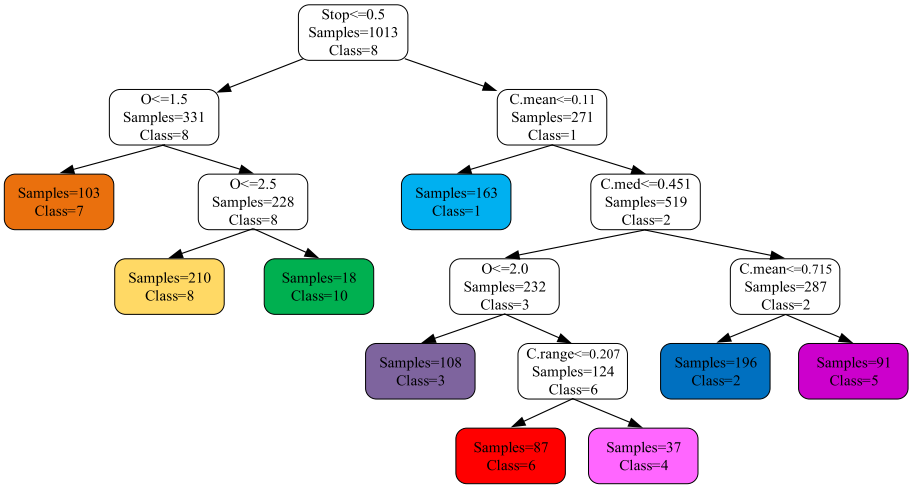
(b) DT prediction model for 15% preceding period



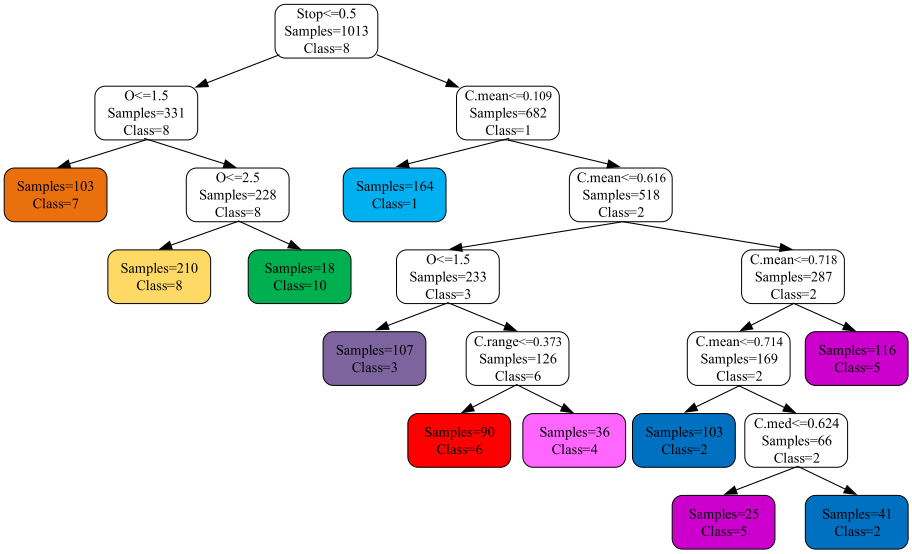
(c) DT prediction model for 20% preceding period

(*continued on next page*)

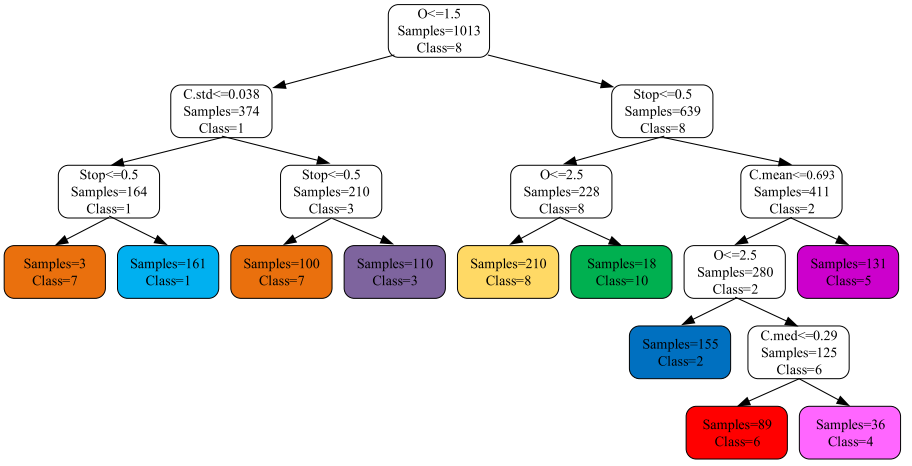
(*continued*)



(d) DT prediction model for 30% preceding period



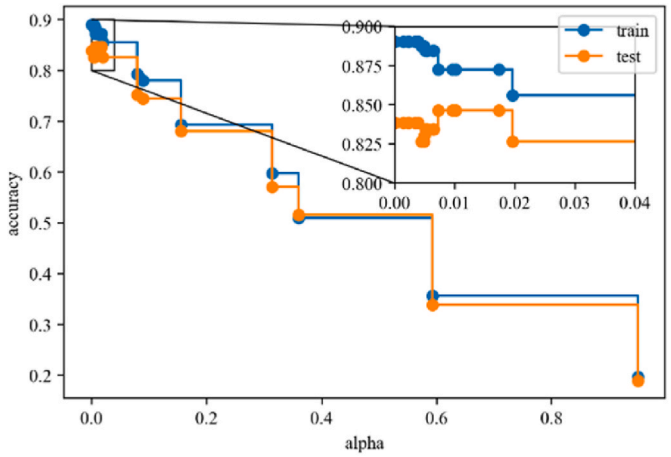
(e) DT prediction model for 50% preceding period



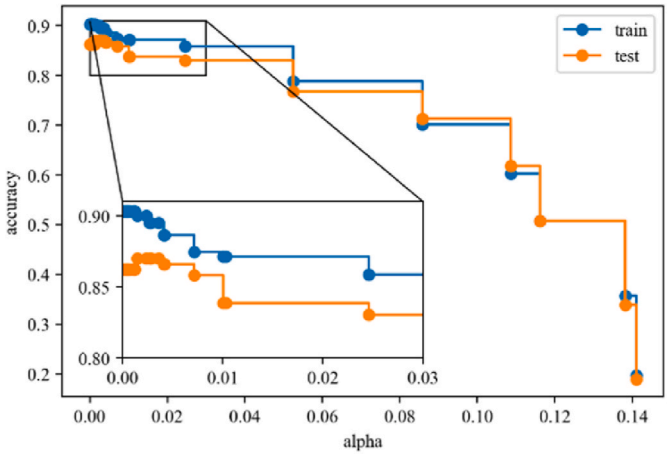
(f) DT prediction model for 80% preceding period

(*continued on next page*)

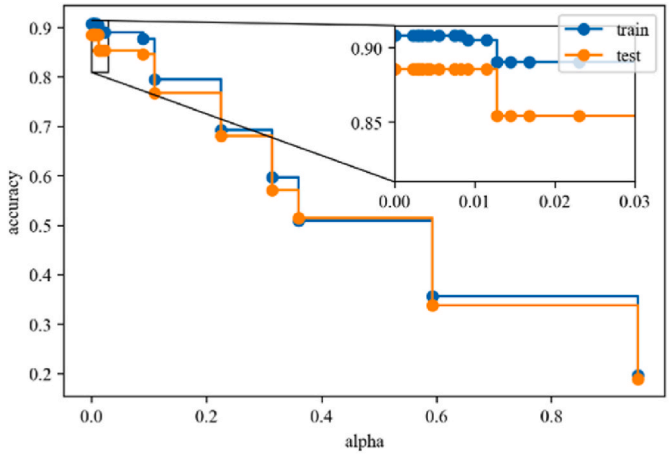
(*continued*)



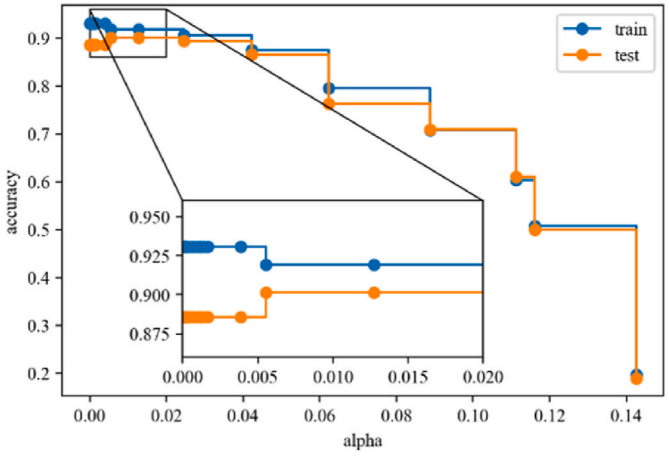
(g) The relationship between accuracy and *ccp_alpha* of 10% preceding period



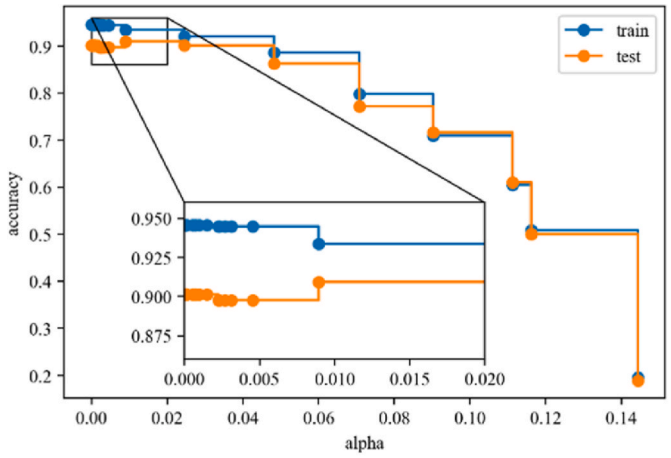
(h) The relationship between accuracy and *ccp_alpha* of 15% preceding period



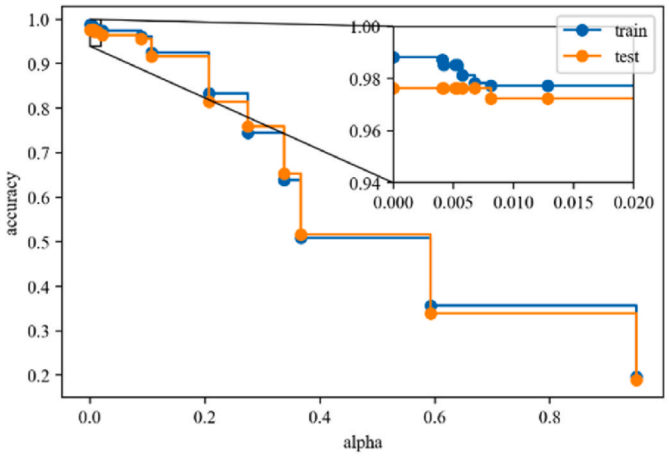
(i) The relationship between accuracy and *ccp_alpha* of 20% preceding period



(j) The relationship between accuracy and *ccp_alpha* of 30% preceding period



(k) The relationship between accuracy and *ccp_alpha* of 50% preceding period



(l) The relationship between accuracy and *ccp_alpha* of 80% preceding period

References

Abebe, M., Shin, Y., Noh, Y., Lee, S., Lee, I., 2020. Machine learning approaches for ship speed prediction towards energy efficient shipping. Appl. Sci. 10 (7), 2325.

Abreu, L.R., Maciel, I.S.F., Alves, J.S., Braga, L.C., Pontes, H.L.J., 2023. A decision tree model for the prediction of the stay time of ships in Brazilian ports. Eng. Appl. Artif. Intell. 117, 105634.

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.

Breiman, L., 2017. Classification and Regression Trees. Routledge.
Capobianco, S., Millefiori, L.M., Forti, N., Braca, P., Willett, P., 2021. Deep learning methods for vessel trajectory prediction based on recurrent neural networks. IEEE Trans. Aero. Electron. Syst. 57 (6), 4329–4346.
Chen, Y., Liu, Z., Zhang, M., Yu, H., Fu, X., Xiao, Z., 2024. Dynamics collision risk evaluation and early alert in busy waters: a spatial-temporal coupling approach. Ocean Eng. 300, 117315.
Chen, Y., Yang, S., Suo, Y., Zheng, M., 2021. Ship track prediction based on DLGWO-SVR. In: Scientific Programming, pp. 1–14, 2021.
Cheng, T., Veitch, E.A., Utne, I.B., Ramos, M.A., Mosleh, A., Alsos, O.A., Wu, B., 2024. Analysis of human errors in human-autonomy collaboration in autonomous ships operations through shore control experimental data. Reliab. Eng. Syst. Saf. 246, 110080.
Cheng, Z., Zhang, Y., Wu, B., Guedes Soares, C., 2023. Traffic-conflict and fuzzy-logic-based collision risk assessment for constrained crossing scenarios of a ship. Ocean Eng. 274, 114004.
Dobrkovic, A., Iacob, M.E., van Hillegersberg, J., 2018. Maritime pattern extraction and route reconstruction from incomplete AIS data. International journal of Data science and Analytics 5, 111–136.
Duan, H., Ma, F., Miao, L., Zhang, C., 2022. A semi-supervised deep learning approach for vessel trajectory classification based on AIS data. Ocean Coast Manag. 218, 106015.
Fuentes, G., 2021. Generating bunkering statistics from AIS data: a machine learning approach. Transport. Res. E Logist. Transport. Rev. 155, 102495.
Gao, D., Zhu, Y., Guedes Soares, C., 2023. Uncertainty modelling and dynamic risk assessment for long-sequence AIS trajectory based on multivariate Gaussian Process. Reliab. Eng. Syst. Saf. 230, 108963.
Guo, W., Zhao, Z., Zheng, Z., Xu, Y., 2020. A cloud-based approach for ship stay behavior classification using massive trajectory data. In: 2020 International Conference on Service Science (ICSS), pp. 82–89.
Guo, Y., Liu, R.W., Qu, J., Lu, Y., Zhu, F., Lv, Y., 2023. Asynchronous trajectory matching-based multimodal maritime data fusion for vessel traffic surveillance in inland waterways. IEEE Trans. Intell. Transport. 24 (11), 12779–12792.
Han, P., Yang, X., 2020. Big data-driven automatic generation of ship route planning in complex maritime environments. Acta Oceanol. Sin. 39 (8), 113–120.
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.
Kang, L., Meng, Q., Liu, Q., 2018. Fundamental diagram of ship traffic in the Singapore Strait. Ocean Eng. 147, 340–354.
Lee, J., Cho, I., 2022. Extracting the maritime traffic route in Korea based on probabilistic approach using automatic identification system big data. Appl. Sci. 12 (2), 635.
Li, H., Jiao, H., Yang, Z., 2023. Ship trajectory prediction based on machine learning and deep learning: a systematic review and methods analysis. Eng. Appl. Artif. Intell. 126, 107062.
Li, L., Lu, Y., Yang, D., 2024. Aerial visual data-driven approach for berthing capacity estimation in restricted waters. Ocean Coast Manag. 248, 106961.
Liang, M., Liu, R.W., Li, S., Xiao, Z., Liu, X., Lu, F., 2021. An unsupervised learning method with convolutional auto-encoder for vessel trajectory similarity computation. Ocean Eng. 225, 108803.
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. Transport. Syst. 23 (12), 23694–23707.
Liang, M., Weng, L., Gao, R., Li, Y., Du, L., 2024. Unsupervised maritime anomaly detection for intelligent situational awareness using AIS data. Knowl. Base Syst. 284, 111313.
Liu, C., Kulkarni, K., Suominen, M., Kujala, P., Musharraf, M., 2024a. On the data-driven investigation of factors affecting the need for icebreaker assistance in ice-covered waters. Cold Reg. Sci. Technol. 221, 104173.
Liu, C., Musharraf, M., Li, F., Kujala, P., 2022. A data mining method for automatic identification and analysis of icebreaker assistance operation in ice-covered waters. Ocean Eng. 266, 112914.
Liu, J., Shi, G., Zhu, K., 2019. Vessel trajectory prediction model based on AIS sensor data and adaptive chaos differential evolution support vector regression (ACDE-SVR). Appl. Sci. 9 (15), 2983.
Liu, L., Liu, K., Shibasaki, R., Zhang, Y., Zhang, M., 2024b. Assessment of the feasibility of vessel trains in the ocean shipping sector. Transport. Res. Transport Environ. 130, 104188.
Liu, R.W., Lu, Y., Guo, Y., Ren, W., Zhu, F., Lv, Y., 2023a. AiOENet: all-in-one low-visibility enhancement to improve visual perception for intelligent marine vehicles under severe weather conditions. IEEE Transactions on Intelligent Vehicles.
Liu, R.W., Yuan, W., Chen, X., Lu, Y., 2021. An enhanced CNN-enabled learning method for promoting ship detection in maritime surveillance system. Ocean Eng. 235, 109435.
Liu, R.W., Zheng, W., Liang, M., 2024c. Spatio-temporal multi-graph transformer network for joint prediction of multiple vessel trajectories. Eng. Appl. Artif. Intell. 129, 107625.
Liu, Z., Gao, H., Zhang, M., Yan, R., Liu, J., 2023b. A data mining method to extract traffic network for maritime transport management. Ocean Coast Manag. 239, 106622.
n.d. Liu, Z, Yuan, W, Liang, M, et al., 2024. An online method for ship trajectory compression using AIS data Journal of Navigation 1–22. https://doi.org/10.1017/S0373463324000171. Published online.
Liu, Z., Zhang, B., Zhang, M., Wang, H., Fu, X., 2023c. A quantitative method for the analysis of ship collision risk using AIS data. Ocean Eng. 272, 113906.
Lu, Y., Guo, Y., Liu, R.W., Chui, K.T., Gupta, B.B., 2022. GradDT: gradient-guided despeckling transformer for industrial imaging sensors. IEEE Trans. Ind. Inf. 19 (2), 2238–2248.
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.
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.
Park, J., Jeong, J., Park, Y., 2021. Ship trajectory prediction based on Bi-LSTM using spectral-clustered AIS data. J. Mar. Sci. Eng. 9 (9), 1037.
Qi, L., Zheng, Z., 2016. Trajectory prediction of vessels based on data mining and machine learning. J. Digit. Inf. Manag. 14 (1), 33–40.
Ristic, B., La Scala, B., Morelande, M., Gordon, N., 2008. Statistical analysis of motion patterns in AIS data: anomaly detection and motion prediction. In: 2008 11th International Conference on Information Fusion. IEEE, pp. 1–7.
Rong, H., Teixeira, A.P., Guedes Soares, C., 2022. Maritime traffic probabilistic prediction based on ship motion pattern extraction. Reliab. Eng. Syst. Saf. 217, 108061.
Rong, H., Teixeira, A.P., Soares, C.G., 2019. Ship trajectory uncertainty prediction based on a Gaussian Process model. Ocean Eng. 182, 499–511.
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., Guedes Soares, C., 2015. Detection and analysis of the main routes of voluntary observing ships in the North Atlantic. J. Navig. 68 (2), 397–410.
Volkova, T.A., Balykina, Y.E., Bespalov, A., 2021. Predicting ship trajectory based on neural networks using AIS data. J. Mar. Sci. Eng. 9 (3), 254.
Wang, X., Liu, Z., Yan, R., Wang, H., Zhang, M., 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.
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., Fu, X., Zhang, L., Goh, R.S.M., 2020. Traffic pattern mining and forecasting technologies in maritime traffic service networks: a comprehensive survey. IEEE Trans. Intell. Transport. Syst. 21 (5), 1796–1825.
Xiao, Z., Fu, X., Zhao, L., Zhang, L., Teo, T.K., Li, N., Zhang, W., Qin, Z., 2023. Next-Generation vessel traffic services systems—from "passive" to "proactive". IEEE Intelligent Transportation Systems Magazine 15 (1), 363–377.
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.
Xin, X., Liu, K., Loughney, S., Wang, J., Yang, Z., 2023. Maritime traffic clustering to capture high-risk multi-ship encounters in complex waters. Reliab. Eng. Syst. Saf. 230, 108936.
Yan, R., Wang, S., Du, Y., 2020b. Development of a two-stage ship fuel consumption prediction and reduction model for a dry bulk ship. Transport. Res. E Logist. Transport. Rev. 138, 101930.
Yan, Z., Xiao, Y., Cheng, L., He, R., Ruan, X., Zhou, X., Li, M., Bin, R., 2020a. Exploring AIS data for intelligent maritime routes extraction. Appl. Ocean Res. 101, 102271.
Zhang, C., Bin, J., Wang, W., Peng, X., Wang, R., Halldearn, R., Liu, Z., 2020. AIS data driven general vessel destination prediction: a random forest-based approach. Transport. Res. C Emerg. Technol. 118, 102729.
Zhang, L., Yang, D., Bai, X., Lai, K.H., 2023b. How liner shipping heals schedule disruption: a data-driven framework to uncover the strategic behavior of port-skipping. Transport. Res. E Logist. Transport. Rev. 176, 103229.
Zhang, M., Conti, F., Le Sourne, H., Vassalos, D., Kujala, P., Lindroth, D., Hirdaris, S., 2021a. A method for the direct assessment of ship collision damage and flooding risk in real conditions. Ocean Eng. 237, 109605.
Zhang, M., Kujala, P., Musharraf, M., Zhang, J., Hirdaris, S., 2023a. A machine learning method for the prediction of ship motion trajectories in real operational conditions. Ocean Eng. 283, 114905.
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, M., Tsoulakos, N., Kujala, P., Hirdaris, S., 2024. A deep learning method for the prediction of ship fuel consumption in real operational conditions. Eng. Appl. Artif. Intell. 130, 107425.
Zhang, M., Zhang, D., Fu, S., Kujala, P., Hirdaris, S., 2022a. A predictive analytics method for maritime traffic flow complexity estimation in inland waterways. Reliab. Eng. Syst. Saf. 220, 108317.
Zhang, X., Fu, X., Xiao, Z., Xu, H., Qin, Z., 2022b. Vessel trajectory prediction in maritime transportation: current approaches and beyond. IEEE Trans. Intell. Transport. Syst. 23 (11), 19980–19998.
Zhao, L., Shi, G., 2019. A trajectory clustering method based on Douglas-Peucker compression and density for marine traffic pattern recognition. Ocean Eng. 172, 456–467.
Zhu, F., 2011. Mining ship spatial trajectory patterns from AIS database for maritime surveillance. In: 2011 2nd IEEE International Conference on Emergency Management and Management Sciences, pp. 772–775.