

# Unsupervised Hierarchical Methodology of Maritime Traffic Pattern Extraction for Knowledge Discovery

Huanhuan Li<sup>a,b,c</sup>, Jasmine Siu Lee Lam<sup>a,\*</sup>, Zaili Yang<sup>b</sup>, Jingxian Liu<sup>c</sup>, Ryan Wen Liu<sup>c,\*\*</sup>, Maohan Liang<sup>c</sup> and Yan Li<sup>c</sup>

<sup>a</sup>*School of Civil and Environmental Engineering, Nanyang Technological University, Singapore*

<sup>b</sup>*School of Engineering, Technology and Maritime Operations, Liverpool John Moores University, Liverpool L3 3AF, UK*

<sup>c</sup>*Hubei Key Laboratory of Inland Shipping Technology, School of Navigation, Wuhan University of Technology, Wuhan 430063, China*

## ARTICLE INFO

### Keywords:

Pattern Extraction  
Knowledge discovery  
Trajectory compression  
Trajectory clustering  
Maritime traffic safety management

## ABSTRACT

Owing to the space-air-ground integrated networks (SAGIN), seaborne shipping has attracted increasing interest in the research on the motion behavior knowledge extraction and navigation pattern mining problems in the era of maritime big data for improving maritime traffic safety management. This study aims to develop a novel unsupervised methodology for feature extraction and knowledge discovery based on automatic identification system (AIS) data, allowing for seamless knowledge transfer to support trajectory data mining. The unsupervised hierarchical methodology is constructed from three parts: trajectory compression, trajectory similarity measure, and trajectory clustering. In the first part, an adaptive Douglas-Peucker with speed (ADPS) algorithm is created to preserve critical features, obtain useful information, and simplify trajectory information. Then, dynamic time warping (DTW) is utilized to measure the similarity between trajectories as the critical indicator in trajectory clustering. Finally, the improved spectral clustering with mapping (ISCM) is presented to extract vessel traffic behavior characteristics and mine movement patterns for enhancing marine safety and situational awareness. Comprehensive experiments are conducted and implemented in the Chengshan Jiao Promontory in China to verify the feasibility and effectiveness of the novel methodology. Experimental results show that the proposed methodology can effectively compress the trajectory, determine the number of clusters in advance, guarantee the clustering accuracy, and extract useful navigation knowledge while significantly reducing the computational cost. The clustering results are further explored and follow the Gaussian mixture distribution, which can further help provide new discriminant criteria for trajectory clustering.


## 1. Introduction

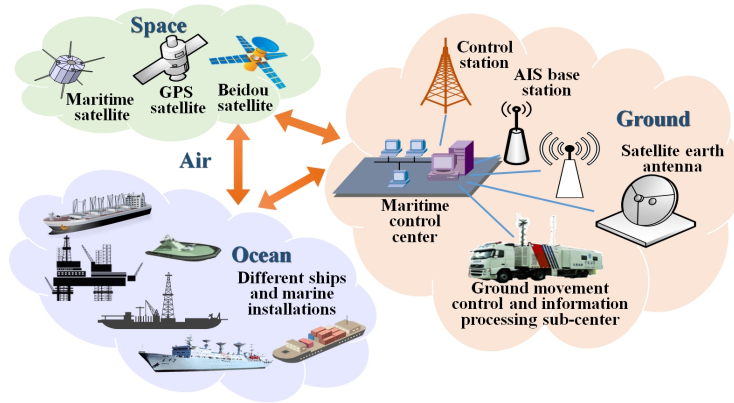
The development of space-air-ground integrated networks (SAGIN) (Yang et al., 2020), such as wireless communication technology, monitoring system, radar system, sensor network, and satellite constellation, makes it possible to store, analyze and apply explosive data with wide applications to all walks of life (Yang et al., 2019), including maritime transport (Huo et al., 2020). SAGIN is promising to provide and improve seamless communication and traffic services (Wei et al., 2021), as demonstrated in Fig. 1, where the comprehensive transportation system based on SAGIN is presented. Maritime vessels based on SAGIN are key parts of intelligent traffic and also play important roles in the global shipping network (Liu et al., 2020). Shipping has undertaken more than 80% global trading freight and logistics as the most efficient and low-cost transport means, placing maritime safety and security as a high priority for each country (Li and Lam, 2017; Magirou et al., 2015; Millefiori et al., 2016).

Following the extensive use of automatic identification system (AIS) equipment due to the mandatory installation requirements of the International Maritime Organization (IMO), the vast amount of near-real AIS-based spatiotemporal vessel trajectories information from different types of maritime communications has significantly grown in recent years (Tu et al., 2020), including maritime mobile service identity (MMSI), time, longitude, latitude, and speed over ground (SOG), and course over ground (COG) (Xiao et al., 2019). AIS-based vessel trajectories are increasingly used to extract the moving knowledge and marine traffic patterns, which are critical in trajectory data mining and

\*Principal corresponding author

\*\*Corresponding author

 huanhuanlisun@gmail.com (H. Li); SLLam@ntu.edu.sg (J.S.L. Lam); Z.Yang@ljmu.ac.uk (Z. Yang); liujingxian@whut.edu.cn (J. Liu); wenliu@whut.edu.cn (R.W. Liu); mliang@whut.edu.cn (M. Liang); li\_yan@whut.edu.cn (Y. Li)  
ORCID(s):



**Fig. 1:** The comprehensive transportation system based on SAGIN for maritime vehicles.

knowledge discovery. The potential and hidden knowledge of the trajectory helps to solve the problems in intelligent traffic management and assist in maritime traffic surveillance (Yan et al., 2020). Therefore, trajectory data mining has become a research hotspot and has broad applications in vessel behavior modeling (Zhou et al., 2019), maritime situational awareness (Coscia et al., 2018; Graziano et al., 2019; Pitsikalis et al., 2020; Murray and Perera, 2020), and intelligent transportation management (Rong et al., 2020). Furthermore, the development of trajectory data mining technologies, such as storing, processing, analysis, and application, arise as an emerging research topic lately (Yap and Lam, 2020; Aslam et al., 2020; Wang et al., 2020b, 2019).

Vessel trajectory data mining can discover navigational knowledge and perceive maritime situations from AIS-based spatiotemporal vessel trajectories (Vespe et al., 2012; Pan et al., 2014; Li et al., 2017). The flowchart of trajectory data mining is presented in Fig. 2, involving AIS data collection, data preprocessing, trajectory retrieval and indexing, trajectory pattern mining, and application. In particular, it follows three steps of trajectory compression, trajectory similarity measurement, and trajectory clustering analysis in a hierarchical order. A strong relationship among the steps to stimulate knowledge discovery has been well documented in the literature (Andrienko et al., 2018; Atev et al., 2010; Wang et al., 2020a; Pallotta et al., 2013a; Li et al., 2016; Zhang et al., 2018; De Mulder, 2014; Ding et al., 2018; Elhamifar and Vidal, 2013; Hong et al., 2017).

Trajectory compression is among the most important contents in AIS-based vessel trajectory preprocessing. Trajectory compression technologies are widely used in maritime trajectory simplification, route extraction, trajectory clustering, and data mining (Huang et al., 2020; Zhang et al., 2016). Trajectory compression is the key foundation of data mining and knowledge discovery, and it also determines the accuracy of subsequent trajectory similarity measurement and trajectory clustering. However, it is difficult to determine the trajectory compression threshold.

There is a large volume of published studies on trajectory similarity measurement methods (Li et al., 2020; Zhao and Shi, 2019b; Zheng and Zhou, 2011). Trajectory similarity measurement is regarded as a crucial factor in calculating the distance between trajectories, and hence it is one of the critical indicators in trajectory clustering (Tu et al., 2017; Talat et al., 2020; Lei, 2020).

Clustering analysis is a classical method in data mining techniques (Han et al., 2018; Shi et al., 2019). It can group the data sets into different clusters while ensuring that the data points in the same cluster are more similar to each other than to those in other clusters (Lee et al., 2007). Trajectory clustering is applied for mining vessel customary routes and hidden movement patterns. Therefore, it aids in realizing knowledge discovery and situational awareness (Pallotta et al., 2013b; Chen et al., 2014). However, it is still challenging to optimally cluster vessel trajectories because of the volume and spatio-temporal characteristics. Furthermore, the choice of a rational trajectory clustering method is complex. Moreover, it is difficult to choose the parameters used in different clustering methods, such as the appropriate number of clusters, density threshold, and radius threshold. Furthermore, the issues about trajectory data with time, position, and speed remain to be solved in trajectory data mining.

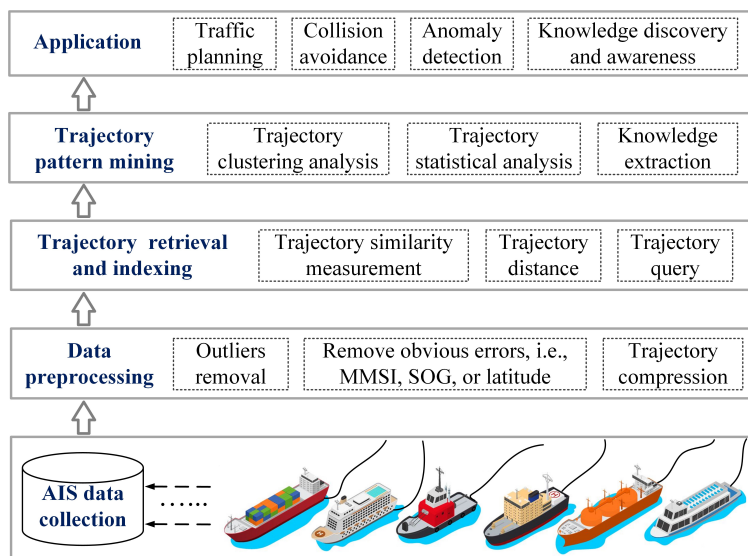
In light of the above, the key research problems to be addressed in our paper are summarized as follows:

**Question 1: How to conduct unsupervised trajectory data mining for maritime knowledge discovery systematically?**

**Question 2: How to effectively simplify the trajectories without human intervention while retaining critical features?**

**Question 3: How to mine moving patterns and extract hidden knowledge accurately without presetting parameter values?**

To address the research questions, this work aims to develop a novel unsupervised hierarchical methodology to improve the effectiveness and applicability of vessel trajectory data mining. The main contributions of this work include automatic trajectory compression, improved trajectory clustering, and the finding of within-cluster similarity distribution fitting analysis. Specifically, automatic trajectory compression, an adaptive Douglas-Peucker with speed (ADPS) algorithm, can automatically compress and simplify the trajectories. Then, the similarity between trajectories is measured by dynamic time warping (DTW). Furthermore, it conducts trajectory clustering analysis based on the improved spectral clustering with mapping (ISCM) and clustering internal evaluation indexes. Finally, the similarity distribution of each cluster is further fitted to validate the clustering performance.



**Fig. 2:** The flowchart of trajectory data mining.

The previous related work in the literature is reviewed in Section 2, followed by the revealed state of the art and the description of the novelties of this work. Section 3 describes the details of the proposed unsupervised hierarchical methodology for AIS-based vessel trajectories. Comprehensive experiments and evaluation analysis on realistic trajectories in the Chengshan Jiao Promontory (CJP) are carried out in Section 4. Section 5 concludes this paper with the implications of the findings and limitations for future work.

## 2. Related Work

Trajectory data mining is an important research focus for knowledge discovery in maritime traffic management. Most of the studies on trajectory data mining based on AIS data only focus on trajectory compression, trajectory similarity, or trajectory clustering methods, and few integrate the three with novel algorithms in a holistic framework. Moreover, the trajectory clustering results need to be further explored and studied to verify the accuracy and the validity of trajectory clustering methods based on the within-cluster similarity fitting. To our knowledge, no previous studies have been performed on the systematic framework from unsupervised and hierarchical perspectives without presetting parameters values, including trajectory compression, trajectory similarity, and trajectory clustering methods. The critical analysis of the trajectory data mining research in maritime knowledge discovery is conducted from three perspectives, including traditional methods, newly-developed machine learning-based methods, and deep learning-based methods.

## 2.1. Traditional Methods in Maritime Knowledge Discovery

A large and growing body of research has investigated trajectory mining methods for maritime knowledge discovery to extract vessel behaviors and motion patterns. Ristic et al. (2008) proposed a simple framework of adaptive kernel density estimation to extract the motion patterns, detect anomalies, and predict the vessel motion. Mazzearella et al. (2014) introduced a knowledge extraction method based on the SOG, COG, and density-based spatial clustering of applications with noise (DBSCAN) algorithm to mine vessel moving patterns and behavior characteristics on AIS data, thus aiding the detection of fishing areas. The parameters in DBSCAN are determined by a heuristic approach. Vespe et al. (2016) firstly introduced a map for EU fishing based on the analysis and investigation of fishing vessel patterns from AIS data to identify fishing activities, then to track and manage vessels' fishing footprint.

To further mine hidden and useful navigational characteristics and patterns, many scholars proposed new frameworks for trajectory clustering from different angles. Bomberger et al. (2006) and Rhodes et al. (2007) proposed a new system to learn normal behavior patterns and predict vessel motion based on an artificial neural network method on real AIS data. The traditional methods in maritime knowledge discovery do not consider new algorithms and techniques. Up to now, different trajectory data mining algorithms and frameworks have been created to improve trajectory clustering. Lee et al. (2007) proposed a trajectory clustering (TRACCLUS) framework based on the DBSCAN and a partition framework to conduct trajectory clustering and mine vessel patterns. Lee et al. (2007) further developed the partition framework and trajectory partition clustering based on the sub-trajectories to detect trajectory outliers. Based on the previous research, Lee et al. (2008) proposed a feature generation framework "TraClass" to cluster the sub-trajectories and class the different patterns. Mascaro et al. (2014) developed a Bayesian network model to learn vessel behaviors and motion patterns from typical AIS data, then to detect ship anomaly behaviors.

Classical methods of trajectory compression and trajectory clustering are widely used in trajectory extraction and analysis. However, it is difficult to determine the parameter values, such as trajectory compression thresholds, the appropriate number of clusters, density threshold, and radius threshold. Previous works have not taken into account the innovative solution that a multi-layer algorithm can bring. Instead, most of them focus on clustering methods. In the meantime, the answer to how to set parameter values in different frameworks remains unclear.

## 2.2. Newly-Developed Machine Learning Methods in Maritime Knowledge Discovery

It is crucial to take advantage of trajectory data mining and knowledge discovery based on AIS data are crucial to exploit the full ship behavior characteristics and moving patterns and facilitate their applications in traffic management. Gaffney and Smyth (1999) proposed a new trajectory clustering method based on the maximum likelihood principle, an expectation-maximization algorithm, and a mixed regression model to extract the hidden information in trajectories. Nanni and Pedreschi (2006) proposed a spatio-temporal trajectory clustering method based on density, time semantics, and the average distance to discover time features and intricate clustering patterns. The clustering method based on the regression mixture model was introduced in Gaffney et al. (2007) to find and analyze the moving characteristics of the extratropical cyclones. Pallotta et al. (2013b) developed a useful learning framework based on traffic route extraction and anomaly detection (TREAD) and the popular Ornstein-Uhlenbeck stochastic process to extract traffic routes and model vessel behavior. Yu et al. (2013) proposed an online clustering method "CtraStream" based on density to group the trajectory with spatial-temporal information for trajectory data from moving objects. Liu et al. (2014) proposed a new clustering method, density-based spatial clustering of applications with noise considering Speed and direction (DBSCANSD), to extract the traffic patterns and discern abnormalities. However, it has five parameters in DBSCANSD. Li et al. (2017) proposed a multi-step clustering method based on dynamic time warping (DTW), principal component analysis (PCA), and improved center clustering algorithm to ensure robust AIS trajectory clustering, find the customary vessel routes, and detect abnormal trajectories. Zhao et al. (2017) combined the decision graph and data field to conduct the trajectory clustering method and discover dynamic patterns and hotspots to support transportation planning and management. Zhen et al. (2017) designed a new trajectory similarity measurement method and applied it into hierarchical clustering to learn the typical patterns, and then detect anomalies based on the Naïve Bayes classifier. Li et al. (2018a) introduced merge distance (MD) and multidimensional scaling (MDS) into a spatio-temporal trajectory clustering method to analyze the vessel behavior characteristics and navigation patterns. Lehmann et al. (2019) developed the trajectory similarity measurement method "SMSM" to detect stops and moves, and extract semantic information in urban transportation, then verified the performance of SMSM in three different trajectory datasets. Zhao and Shi (2019a) put forward a new trajectory similarity method to measure the distance between the vessel trajectories to mine different routes and patterns in data from the Zhoushan Islands waters. Zhao and Shi (2019b) proposed a new trajectory clustering method based on the Douglas and Peucker (DP) algorithm with an

appropriate threshold and the improved DBSCAN algorithm with better parameters  $\epsilon$  and MinLns to analyze vessel behavior and extract traffic patterns. Liu et al. (2019) developed the traditional DP algorithm to realize the adaptive compression and conduct AIS-based vessel trajectory clustering. However, the improved method failed to address the speed factor. Li et al. (2020) considered the corresponding relationship between trajectories and proposed an adaptive constrained dynamic time warping (ACDTW) method to measure the similarity between trajectories. Then the trajectory clustering and classification are conducted in time series and AIS-based trajectories. Kontopoulos et al. (2021) proposed a distributed framework, including waypoint extraction, route segment, and DBSCAN clustering, to construct the trajectory network, identify the trajectory lanes, and mine the traffic patterns.

The aforementioned works reveal that the existing trajectory compression and clustering methods have not yet provided an effective solution on how to automatically compress trajectory and select appropriate thresholds. It also reveals other critical research problems in trajectory clustering, including eliminating the influence of different parameters, determining the number of clustering centers, and choosing an appropriate clustering method.

### 2.3. Deep Learning Methods in Maritime Knowledge Discovery

In recent years, deep learning methods, such as Recurrent Neural Network (RNN) and Convolutional Neural Network (CNN), have also attracted significant attention in maritime data mining. They are applied to encode trajectories and decode them into feature vectors. Then, different point clustering methods can be conducted to cluster the trajectories and mine useful information. A RNN-based Seq2Seq autoencoder model is developed to extract the features and improve the accuracy of similarity measurement (Yao et al., 2017). Li et al. (2018b) proposed a Seq2Seq learning method (t2vec) to compute the similarity and enhance its robustness based on a new spatial proximity loss function. Taghizadeh et al. (2019) further verified the effectiveness of t2vec in achieving good similarity measurement results. To reduce the unexpected influence of noise points in the process of similarity measurement, a novel auto-encoder model is proposed by introducing an attention mechanism to realize the feature representations of noisy vessel trajectories in a low dimensional space (Zhang et al., 2019a). The three kinds of semantic information of active trajectories are taken into account in the Seq2Seq auto-encoder model to generate a new At2vec model to acquire the robust feature representation and discover implicit features (Zhang et al., 2019b). Liang et al. (2021) proposed an unsupervised learning method to efficiently calculate the similarities between vessel trajectories based on convolutional auto-encoder (CAE).

Although deep learning methods have strong feature learning ability, they fail to provide an effective solution to adjust and find optimal parameter values such as learning rate, batch size, gride size, epoch size, iteration number, and loss function. Moreover, deep learning models often keep the inference content in black boxes, leading to unexplainable and invisible processes. Therefore, it is highly demanded but challenging to develop a systematical and unsupervised framework based on deep learning to learn and mine traffic patterns.

Following the above critical analysis of the previous studies in the field, we summarize the state of the art against the four novelties (i.e. N1-N4) of this study in the ensuing section.

#### **N1. The Unsupervised Hierarchical Methodology.**

**State of the art:** In maritime transportation trajectory studies, analyzing navigation direction is the main innovation in trajectory compression, while optimizing the clustering parameters is the primary improvement in trajectory clustering methods (Lee et al., 2007; Pallotta et al., 2013b). The use of different parameters will lead to the existence of biases in the results. While the current frameworks focus on the improvement of trajectory compression and clustering methods (Arguedas et al., 2017; Zhao and Shi, 2019b), they seldomly tackle the speed factor and the influence of different parameters simultaneously. Therefore, a new systematic and unsupervised methodology is developed from the perspective of the hierarchical model (Allen, 2018) in this manuscript, and it has made significant value for trajectory data mining.

**Our solution:** The proposed unsupervised hierarchical methodology is constructed to mine the typical behavioral characteristics and moving patterns of ships. It relies on three new developments described below. The methodology can help realize the extraction of navigational behavior patterns and hidden knowledge more quickly, precisely, and robustly.

#### **N2. The New Adaptive Compression Method.**

**State of the art:** The original Douglas and Peucker (DP) algorithm has translation and rotation invariance to simplify the time series. However, the threshold must be defined by the users in advance. On the other hand, the same threshold for all trajectories is also arguably not optimal. The improvement trajectory compression methods take into account the navigation course and segmentation, but still set the same threshold for the whole data (Tang et al., 2021; Ji

et al., 2021). Meanwhile, speed is presented as an important factor in trajectory compression. Therefore, the questions on automating the threshold setting and selecting appropriate threshold values for different trajectories have not yet been well addressed.

**Our solution:** To effectively simplify ship trajectories and extract features, we propose the Adaptive Douglas and Peucker with Speed (ADPS) algorithm to compress the trajectories and retain effective features. It can automatically calculate and set an appropriate threshold for each of the investigated trajectories without subjective intervention.

#### **N3. The Novel Trajectory Clustering Method.**

**State of the art:** Due to the volume and spatio-temporal characteristics, traditional point clustering methods cannot be directly applied to trajectory clustering. Moreover, it is difficult to select the parameters in different clustering methods, such as the appropriate number of clusters, density threshold, and radius threshold. Density-based clustering methods are the common improved methods dealing with maritime data processing and mining. However, the big data volume and parameter optimization often lead to memory overflow and the presentation of the local optimal solutions (Li et al., 2018a; Xu et al., 2021). Therefore, the question of how to eliminate the influence of the parameters in trajectory clustering remains a research challenge, wanting an effective solution to be found.

**Our solution:** The Improved Spectral Clustering with Mapping (ISCM) is put forward to map the trajectories into points and mine trajectory patterns based on the graph theory and the normalized cut. It can determine the number of clustering centers and extract the hidden knowledge, hence presenting an effective solution to the aforementioned question.

#### **N4. The Within-cluster Similarity Distribution Fitting.**

**State of the art:** The Gaussian mixture model is commonly used for trajectory clustering based on an iterative algorithm. The probability is taken as the clustering criterion (Wang et al., 2021; Fu et al., 2021). From the perspective of good clustering results, its distribution of within-cluster similarity should also obey the Gaussian mixture function. Researchers use the Gaussian mixture function to model the density distribution and fit the result to show the centrality of clustering results with the fitting graph (Lee, 2005; Hu et al., 2017). However, they fail to use the different evaluation indexes to measure the degree of the fitting. Therefore, the finding of the fitting function of clustering results requires further verification based on the multiple evaluation indexes.

**Our solution:** We further explore the within-cluster similarity distribution based on the DTW and the Gaussian mixture function to conclude and verify the finding. The experimental results and fitting analysis indicate that this finding provides a reliable standard and index for the performance evaluation of future clustering methods.

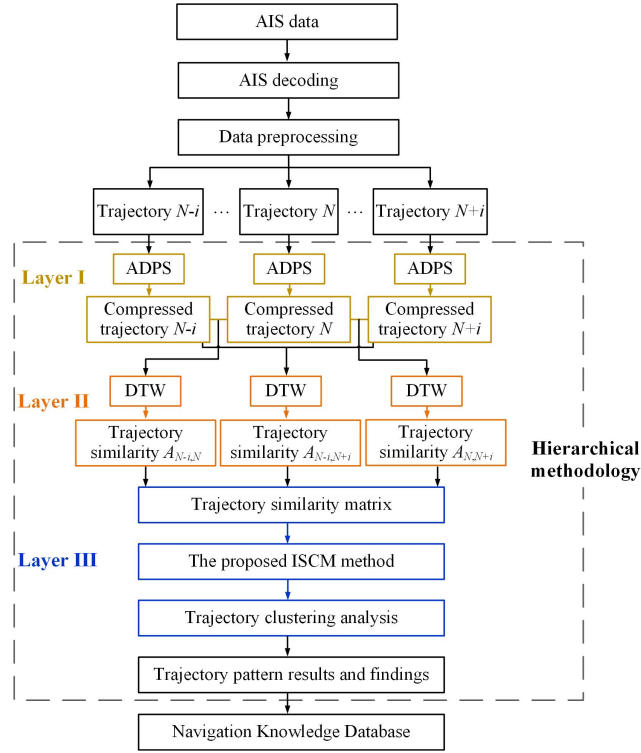
### **3. The Proposed Unsupervised Hierarchical Methodology**

In Fig. 3, it is seen that the unsupervised hierarchical methodology consists of three main steps, including trajectory compression, trajectory similarity measure, and trajectory clustering analysis. Trajectory compression and trajectory clustering analysis are two essential parts of tackling large-scale data volume and data mining. The trajectory compression method (i.e. the proposed ADPS algorithm) can preserve the key features, remove some redundant information, and simplify information for further trajectory similarity measurement. The trajectory similarity measurement method (i.e. the DTW algorithm) can be used to calculate the distance between trajectories based on dynamic warping, which provides a distance criterion for trajectory clustering. The trajectory clustering method (i.e. the proposed ISCM method) is able to extract hidden patterns and discover navigational knowledge, thus correspondingly making route planning, mining vessel movement patterns, and improving maritime safety. The hierarchical methodology is illustrated in Fig. 3, while the three important methods are described in detail in Sections 3.1, 3.2, and 3.3.

As shown in Fig. 3, the original AIS data should be decoded and processed to remove noise and delete obvious errors. In layer I, the trajectories are compressed by the ADPS algorithm to retain their critical features and simplify data. The distance between trajectories  $N - i$  and  $N + i$  are measured based on the DTW method in layer II. Finally, the proposed ISCM method can help mine the traffic patterns and discover the knowledge according to the distance, mapping transformation, and the new spectral clustering method in layer III.

#### **3.1. Adaptive Douglas-Peucker with Speed Algorithm**

Douglas and Peucker (1973) initially proposed the classical DP algorithm to simplify the time series with the line segments. The compressed trajectories are topologically consistent with the original ones, especially for the neighborhood characteristics. Many scholars have shown that the DP algorithm could compress the trajectories effectively while preserving the main geometrical structures (Saalfeld, 1999; Tienah et al., 2015).



**Fig. 3:** The flowchart of our hierarchical methodology.

The traditional DP algorithm has two disadvantages: one is that it needs to pre-define the thresholds artificially; the other is that different trajectories have the same threshold. There are few studies on the automatic selection of the optimal thresholds in the literature. The issue of automatically selecting an appropriate compression threshold for each trajectory is the focus of current research. To address this issue, we propose a novel trajectory compression algorithm, an adaptive Douglas-Peucker with speed (ADPS) algorithm, to set a different threshold for each trajectory based on their geometric features, speed variation rate, and the distance of feature points. This improvement will lay a solid foundation for the subsequent trajectory similarity measurement and clustering analysis.

For time series trajectory with speed, the thresholds of different trajectories are calculated based on the slope of the baseline of the trajectory starting and ending points, and the distance between all trajectory points and the baseline. The threshold model is shown below.

$$\theta = \begin{cases} \frac{1}{|k|} (a_i^j + \sum_{i=2}^{n-1} d_i / (n-2)), |k| > 1 \text{ and } a_i^j > \overline{a^j}, \\ |k| (a_i^j + \sum_{i=2}^{n-1} d_i / (n-2)), |k| \leq 1 \text{ and } a_i^j \leq \overline{a^j}. \end{cases} \quad (1)$$

where  $\theta$  denotes the threshold,  $k$  indicates the slope of baseline,  $d_i$  expresses the distance from the different points in the trajectory to the baseline,  $a_i^j$  is the rate of SOG at the  $i^{th}$  coordinate point of the  $j^{th}$  trajectory,  $\overline{a^j}$  indicates the average rate of SOG in the  $j^{th}$  trajectory. The pseudo-code of the ADPS algorithm is described in Algorithm 1.

The trajectory compression schematic diagram based on the original DP and the proposed ADPS algorithm is displayed in Fig. 4. The compressed results of the DP algorithm can be seen in Figs. 4 (a) and (b), while the results of the ADPS algorithm are presented in Figs. 4 (c), (d), and (e). The trajectory feature points in Figs. 4 (a) and (b) are selected by the same threshold based on the DP algorithm. However, the moving characteristics of the two trajectories are extremely different. Finally, a straight line will approximately replace the trajectory with unobvious features in Fig. 4 (b). Therefore, it is not reasonable and effective to compress all the trajectories with the same threshold. The useful

---

**Algorithm 1** ADPS algorithm

---

**Input:**  $T_i^j = (x_i^j, y_i^j, t_i^j, v_i^j) \in D$ ,  $T_1^j = (x_1^j, y_1^j, t_1^j, v_1^j)$ ,  $T_n^j = (x_n^j, y_n^j, t_n^j, v_n^j)$ ,  $i = 1, \dots, n$ ,  $j = 1, \dots, m$   
▷ /\*  $(x_i^j, y_i^j)$  indicates the  $i^{th}$  coordinate point in the  $j^{th}$  trajectory,  $t_i^j$  and  $v_i^j$  express the time and speed of point  $(x_i^j, y_i^j)$  . \*/

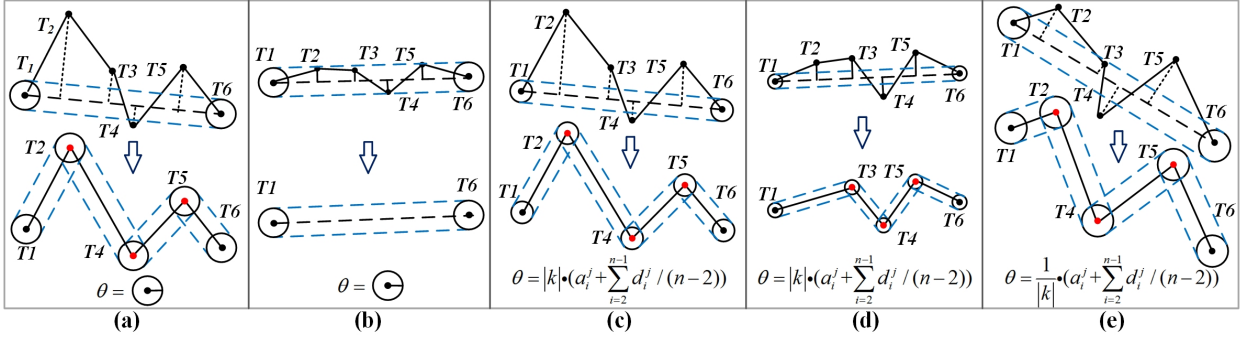
**Output:**  $\theta^j$ ,  $d_i^j$ ,  $TT_i^j$ .  
▷ /\*  $\theta^j$  is the threshold,  $d_i^j$  is Euclidean distance from the point  $(x_i^j, y_i^j)$  to the base line,  $TT_i^j$  is the compressed trajectories . \*/

```
1: //A segmentation framework for annular and semi-annular trajectories based on waterway characteristics and trajectory features.//
2: Set research waterway and analyze the characteristics;
3: if there is annular and semi-annular trajectories, then
4:   split them;
5: else
6:   next step;
7: end if
8: //Calculate  $\theta^j$ //
9: for j = 1:m do
10:   for i = 1:n do
11:      $a_i^j = \left| \frac{v_{i+1}^j - v_i^j}{t_{i+1}^j - t_i^j} \right|$ ;
12:      $\bar{a}^j = \sum_{i=1}^n a_i^j / (n - 1)$ ;
13:      $d_i^j = \frac{|k(x - x_i^j) + y_i^j - y|}{\sqrt{1 + k^2}}$ ;
14:     if  $a_i^j > \bar{a}^j$  and  $|k| > 1$  then
15:        $\theta_v^j = \frac{1}{|k|} a_i^j$ ;
16:        $\theta_d^j = \frac{1}{|k|} \sum_{i=2}^{n-1} d_i^j / (n - 2)$ ;
17:        $\theta^j = \theta_v^j + \theta_d^j$ ;
18:     else
19:        $\theta_v^j = |k| a_i^j$ ;
20:        $\theta_d^j = |k| \sum_{i=2}^{n-1} d_i^j / (n - 2)$ ;
21:        $\theta^j = \theta_v^j + \theta_d^j$ ;
22:     end if
23:   end for
24: end for
25: // Generate  $TT_i^j$  //
26: for j = 1:m do
27:   for i = 1:n do
28:     if  $d_i^j > \theta^j$  then
29:        $TT_i^j \leftarrow i$ ;
30:     else
31:       delete  $i$ ;
32:     end if
33:   end for
34: end for
```

---

and critical feature points can be retained in Figs. 4 (c), (d), and (e). It can be observed that the ADPS algorithm can set appropriate thresholds for different trajectories based on their features. Comparing Fig. 4 (a) and Fig. 4 (d), the same trajectory is not approximately replaced by a straight line. It also presents the reliability of the proposed ADPS method. The comparison in Fig. 4 further verifies the effectiveness and superiority of the proposed ADPS method. The original DP has the same threshold for all trajectories, and it is difficult to determine the optimal value directly. The proposed ADPS algorithm can automatically calculate and set rational thresholds for different trajectories.

The proposed ADPS method can significantly compress vessel trajectories by automatic calculation of different thresholds for each trajectory while maintaining the main geometric structure and features. It is important to ensure the structural characteristics and improve the compression quality in trajectory clustering and classification. Therefore, in practical applications, trajectory compression based on ADPS can significantly improve the accuracy of similarity measurement and clustering.



**Fig. 4:** Schematic diagram of trajectory compression based on different algorithms, (a)-(b) the trajectory compression results based on the DP algorithm with the same threshold value, (c)-(e) the trajectory compression results based on the ADPS algorithm with the adaptive threshold values.

### 3.2. Dynamic Time Warping Method

The trajectory is a kind of time series with consecutive locations and time-stamps. Trajectory similarity measurement has been thought of as a crucial factor in calculating the distance between trajectories, and hence it is one of the critical indicators in trajectory clustering. Trajectory similarity is calculated based on the correspondence and distance between points. DTW is chosen as the similarity measurement method in this work because it is easy to find similarity patterns and the optimal path based on dynamic programming (Morel et al., 2018). It can measure the similarity and receive similar patterns based on the path warping from feature to feature, also does not limit the length of the trajectory (Loh et al., 2011). DTW can minimize the cumulative distance between two trajectories with local optimization. The theory is described within the context of ship trajectory similarity measurement as follows.

Let  $Q = \{q_1, \dots, q_i, \dots, q_m\}$  and  $C = \{c_1, \dots, c_j, \dots, c_n\}$  are the two ship trajectories. Sorting all points according to the time, then we can construct a matrix  $A_{m \times n}$  to store distance, and  $a_{ij} = d(q_i, c_j) = \sqrt{(q_i - c_j)^2} \in A_{m \times n}$ .  $d(q_i, c_j)$  denotes the Euclidean distance between the  $i^{th}$  point in series  $Q$  and the  $j^{th}$  point in series  $C$ .

The essence of DTW is to calculate the distance matrix between two trajectories, then find the optimal warping path. The warping path  $W = \{w_1, w_2, \dots, w_t, \dots, w_M\}$ ,  $w_t = (a_{ij})_t$  consists of a set of adjacent matrix elements in  $A_{m \times n}$ ,  $\max\{m, n\} < M \leq m + n - 1$ . The warping path must meet the following conditions:

- (1) Boundary condition:  $w_1 = a_{11}, w_t = a_{mn}$ ;
- (2) Continuity: if  $w_{t-1} = a_{i't'}, w_t = a_{ij}$ , then  $i - i' \leq 1, j - j' \leq 1$ ;
- (3) Monotonicity: if  $w_{t-1} = a_{i't'}, w_t = a_{ij}$ , then  $i - i' \geq 0, j - j' \geq 0$ , the time at each point is also monotonic in  $W$ .

The path with the lowest warping cost can be calculated as follows:

$$DTW(Q, C) = \min \left\{ \frac{1}{M} \sum_{t=1}^M w_t \right\}. \quad (2)$$

The DTW distance is described as follows:

$$D(1, 1) = d_{11},$$

$$D(i, j) = d_{ij} + \min \begin{cases} D(i, j-1) \\ D(i-1, j-1) \\ D(i-1, j) \end{cases}. \quad (3)$$

### 3.3. The Improved Spectral Clustering with Mapping Method

The spectral clustering algorithm is a classical method, which can find the optimal clustering results according to the graph theory and classify the feature vectors from the feature decomposition (Li et al., 2017). Its essence is the optimal graph partition problem based on the spectral graph partition theory. Spectral clustering can transform

the clustering problem into graph space, identify the sample space with arbitrary shape, and converge to an optimal global solution (Li et al., 2018a). However, it is also sensitive to the input parameter: the number of clustering centers  $k$ . Therefore, we propose the ISCM method to conduct the trajectory clustering, integrating mapping transformation, graph structure, the normalized cut, the number of clustering centers, and the selection of clustering centers

From the viewpoint of the whole trajectory research, the transformation of trajectories into points can not only reduce the computation time but also can select more available clustering algorithms in the future. The proposed ISCM method includes trajectory mapping, internal evaluation indexes, and improved spectral clustering methods. Firstly, MDS is used to map the trajectories to points in the two-dimensional (2D) plane. Then the number of clustering centers  $k$  is determined by the clustering internal evaluation index functions. Finally, the improved spectral clustering method is applied to mine the trajectory patterns and hidden features.

MDS is a classical nonlinear data mapping method, and it can visualize the relationships between the trajectories in a 2D and three-dimensional (3D) space. The similarity matrix between trajectories can be represented as the distance of points in the lower dimension space. The essence of MDS is to find the space representation of the point based on the similarity between trajectories.

Suppose  $X$  is a set of points that has the same distance as  $D$  based on the Euclidean constraints,  $x_i, x_j \in X, d_{ij} \in D$ . The matrix  $T$  is introduced to decompose  $X$  and  $T = XX^T, t_{ij} \in T$ .

The distance between any two points after dimension reduction is still the same as the original distance between trajectories, then

$$d_{ij}^2 = (x_i - x_j)^2 = x_i^2 + x_j^2 - 2x_i x_j, \quad (4)$$

with

$$t_{ij} = x_i x_j \Rightarrow t_{ij} = -\frac{1}{2}(d_{ij}^2 - x_i^2 - x_j^2), \quad (5)$$

Suppose the samples after dimension reduction are centered, namely  $\sum_j x_j = 0$  and  $\sum_i x_i = 0$ , then

$$\begin{aligned} \sum_j d_{ij}^2 &= nx_i^2 + \sum_j x_j^2 - 2x_i \sum_j x_j = nx_i^2 + \sum_j x_j^2, \\ \sum_i d_{ij}^2 &= nx_j^2 + \sum_i x_i^2 - 2x_i \sum_i x_i = nx_j^2 + \sum_i x_i^2, \\ \sum_{ij} d_{ij}^2 &= n \sum_i x_i^2 + n \sum_j x_j^2, \end{aligned} \quad (6)$$

Based on Eqs. (5) and (6), the matrix  $T$  is presented, as shown in Eq. (7).

$$\begin{aligned} t_{ij} &= -\frac{1}{2}(d_{ij}^2 - \frac{1}{n} \sum_k d_{ik}^2 - \frac{1}{n} \sum_k d_{kj}^2 + \frac{1}{n^2} \sum_{k,l} d_{kl}^2), \\ T &= U\Lambda U^T = U\Lambda^{1/2}\Lambda^{1/2}U^T = XX^T. \end{aligned} \quad (7)$$

The running time of trajectory clustering and the volume of data are significantly reduced after the trajectories are mapped into points. For example, the 3904 trajectories being investigated in this study that consist 207267 points are mapped into 3904 points with MDS. The similarity matrix is calculated by DTW in this paper and then is transformed into the relative distance representation of spatial points based on MDS.

The trajectories are mapped into points by the MDS method. Meanwhile, the normalized cut is introduced to avoid the small cluster problem in the proposed ISCM method. Meanwhile, the normalized cut is introduced to avoid the small cluster problem in the proposed ISCM method. Furthermore, the new k-means clustering with the cluster centers  $k$  is proposed and applied to cluster the mapping points based on the similarity between the trajectories from the DTW method. The similarity distribution is divided into  $k$  parts, and the original centers are selected to reduce the iterations. However, since the number of clusters  $k$  is unknown, it is necessary to determine the optimal number of cluster centers in advance.

In this paper, we select three internal evaluation indexes to determine the optimal number of cluster centers in the ISCM algorithm. Clustering evaluation can measure the performance of clustering analysis based on internal evaluation indexes. The comprehensive evaluation is carried out mainly from two aspects: the density of points within the same cluster and the separation degree between the clusters. Internal evaluation indexes include Silhouette Coefficient

(*SC*) index, Calinski-Harabasz Score (*CHS*) index, and Davies-Bouldin Index (*DBI*) index. *SC* can compare the similarity of the sample in the same cluster with that in other clusters. It can also measure the degree of compactness and separation for different clusters. The *SC* index is the average *SC* value of all samples. It can show how similar the point is within a group compared to other groups. The value range is  $[-1, 1]$ , and the larger the value is, the better the clustering performance. The essence of the *CHS* index is to compare the between-cluster and within-cluster scatters. The larger the value is, the closer the same cluster, and the more dispersed the different group. It means that a better clustering performance could be obtained. The *DBI* index can compare the difference between the sum of the mean distance of all samples in different clusters and the center point of different groups. The smaller the value is, the better the clustering result.

After the number of clusters  $k$  is determined by the optimal value of the *SC* index, *CHS* index, and *DBI* index. The global solution and proof of the proposed ISCM method based on the normalized cut is presented as follows.

**Problem modeling.** Given a data set  $X = \{x_1, \dots, x_m\}$  that is divided into  $k$  clusters  $C_1, C_2, \dots, C_k$  by the undirected graph  $G(V, E)$ , where  $C_1 \cup C_2 \cup \dots \cup C_k = V$ ,  $C_i \cap C_j = \emptyset$ .

The edge weight is  $w_{ij}$  and

$$Cut(C_l, \overline{C_l}) = \frac{1}{2} \sum_{i \in C_l, j \notin C_l} w_{ij}. \quad (8)$$

The goal of clustering is

$$Ncut(C_1, C_2, \dots, C_k) = \sum_{l=1}^k \frac{Cut(C_l, \overline{C_l})}{vol(C_l)} = \sum_{l=1}^k \frac{y_l^T L y_l}{y_l^T D y_l}. \quad (9)$$

with

$$vol(C_l) = \sum_{v_i \in C_l} d_{ii}.$$

**Proof and problem solving.** Introduce the indicator matrix  $Y$  and indicator vector  $y_{il}$ .

$$y_{il} = \begin{cases} 0, & v_i \notin C_l \\ \sqrt{\frac{1}{vol(C_l)}}, & v_i \in C_l \end{cases}, \quad Y^T D Y = I. \quad (10)$$

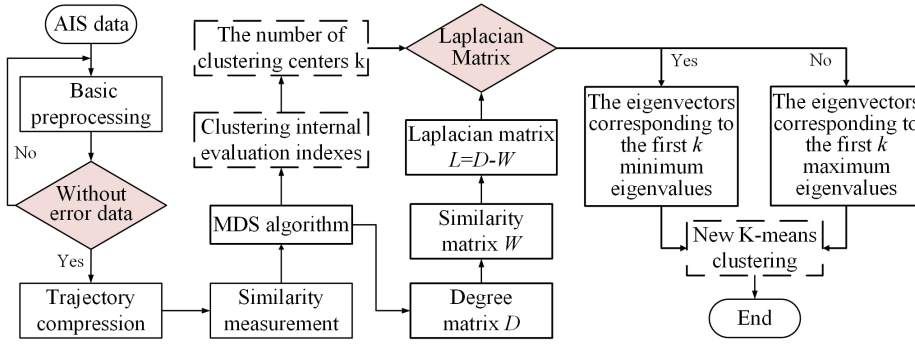
then

$$\begin{aligned} y_l^T L y_l &= \frac{1}{2} \sum_{i,j} w_{ij} (y_{il} - y_{jl})^2, \\ &= \frac{1}{2} \left[ \sum_{i \in C_l, j \notin C_l} w_{ij} \left( \sqrt{\frac{1}{vol(C_l)}} \right)^2 + \sum_{i \notin C_l, j \in C_l} w_{ij} \left( \sqrt{\frac{1}{vol(C_l)}} \right)^2 \right], \\ &= \frac{1}{2} \left[ \sum_{i \in C_l, j \notin C_l} w_{ij} \frac{1}{vol(C_l)} + \sum_{i \notin C_l, j \in C_l} w_{ij} \frac{1}{vol(C_l)} \right], \\ &= \frac{1}{2} \left[ Cut(C_l, \overline{C_l}) \frac{1}{vol(C_l)} + Cut(\overline{C_l}, C_l) \frac{1}{vol(C_l)} \right], \\ &= \frac{Cut(C_l, \overline{C_l})}{vol(C_l)}, \\ &= Ncut(C_l, \overline{C_l}). \end{aligned} \quad (11)$$

with

$$Y_l^T L Y_l = (Y^T L Y)_{ll},$$

$$Ncut(C_1, C_2, \dots, C_k) = \sum_{l=1}^k Y_l^T L Y_l = \sum_{l=1}^k (Y^T L Y)_{ll},$$



**Fig. 5:** The flowchart of the proposed ISCM method.

$$= Tr(Y^T LY).$$

The objective function can be converted to

$$\min_{Y^T DY=I} Tr(Y^T LY). \quad (12)$$

Let us normalize the Laplacian matrix. The property of the Laplacian matrix is as follows

$$L_{norm} = D^{-1/2} L D^{-1/2}, \quad (13)$$

with

$$H = D^{1/2} Y, \quad Y = D^{-1/2} H,$$

The problem can be rewritten based on the graph Laplacian.

$$\begin{aligned} \min_{H \in \mathbb{R}^{m \times k}} Tr(H^T L_{norm} H), \\ s.t. \quad H^T H = I_k. \end{aligned} \quad (14)$$

then

$$\begin{aligned} D^{-1/2} L D^{-1/2} H_l &= \lambda H_l, \\ D^{-1/2} L D^{-1/2} D^{1/2} Y_l &= D^{1/2} Y_l, \\ D^{-1} L &= Y_l. \end{aligned} \quad (15)$$

$Y$  has the  $k$  eigenvectors of  $D^{-1} L$  corresponding to its  $k$  smallest eigenvalues. Finally, the new k-means clustering is applied to mine the hidden patterns. The  $k$  centers are determined by the similarity distribution from the DTW. The proposed ISCM method can better mine the behavior characteristics and pattern information of moving vessels. The flowchart of the proposed ISCM method is illustrated in Fig. 5. The algorithm flow is shown in Algorithm 2.

## 4. Experimental Results and Discussion

### 4.1. Experimental Setting

To verify the accuracy and effectiveness of the proposed method, this paper selects the Chengshan Jiao Promontory (CJP) as the experimental research area. The CJP is one of the busiest coastal waters in China, involving complicated traffic flow and diverse natural environments (wind, wave, current, fog, etc.). The high vessel density and numerous intersection areas have caused the increased risks of ship collision and grounding accidents. The waters of the CJP have many vessel routing systems and navigation rules. The navigation routes and directions are complex and complicated in this area because there are different types of vessels from different graphical regions, such as South Korea, Japan, and some parts of Taiwan, Bohai Bay, Dalian, Dandong Port, and the Shandong Peninsula. Therefore, the number of clustering centers is set within the range of [15, 35] based on the routing scheme.

---

**Algorithm 2** Pseudocode of the proposed ISCM algorithm

---

```
1: Input: the trajectory dataset  $TD$ , the trajectories  $Q, C \in TD_C$ .
2: Output: cluster  $C_1, C_2, \dots, C_k$ 
3: Initialize: the compressed trajectory dataset  $TD_C = \emptyset$ , similarity matrix  $W_T, W_D, W = \emptyset$ , degree matrix  $D = \emptyset$ .
4:  $TD_C \leftarrow ADPS(TD)$ 
5:  $W_T \leftarrow DTW(Q, C)$ 
6:  $X \leftarrow MDS(TD_C)$ 
7: for each point in  $X$  do
8:    $W_D \leftarrow d(x_i, x_j)$ 
9: end for
10:  $D, W \leftarrow G(X, W_D, W_T)$ 
11:  $L = D - W$ 
12:  $L_{norm} \leftarrow D^{-1/2} L D^{-1/2}$ 
13: Eigenvalue decomposition of  $L_{norm}$ 
14:  $\Lambda \leftarrow k$  smallest eigenvalues
15:  $U \leftarrow$  the corresponding eigenvectors
16:  $Y \leftarrow$  row vector of  $U$ 
17:  $k \leftarrow \max(SC, CHS) \& \min(DBI)$ 
18: for  $i = 1:m$  do
19:    $\mu_r$  select the cluster center based on the similarity distribution
20:   repeat
21:      $d(y_i) = \arg \min \|y_i - \mu_r\|_2^2, r = 1, \dots, k$ 
22:      $r \leftarrow \{i, \max d(y_i)\}$ 
23:     update  $\mu_r$ 
24:   Until  $\mu_1, \dots, \mu_k$  are received
25: end for
26: for  $i = 1:m$  do
27:   for  $r = 1:k$  do
28:      $d_{ir} = \min \|y_i - \mu_r\|_2^2$ 
29:      $C_{\mu_r} = C_{\mu_r} \cup \{y_i\}, r = 1, \dots, k$ 
30:   end for
31: end for
```

---

The experiment flowchart is illustrated in Fig. 6. The visualization and discussion of trajectory compression is presented in Section 4.2. Section 4.3 describes the number of clustering centers in the following clustering method. The clustering results using the four different clustering methods (i.e. the proposed ISCM method, original spectral clustering (OSC) method (Von Luxburg, 2007), affinity propagation (AP) method (Wang et al., 2018), and fast affinity propagation (FAP) method (Shang et al., 2012)) are compared and discussed in Section 4.4. Furthermore, the clustering analysis using the four methods is carried out on the data sets before and after compression. The finding based on the clustering results of the proposed ISCM method is shown in Section 4.5. Section 4.6 presents the time complexity analysis of the whole experiment.

The experiments are conducted based on the AIS data set from the CJP in January 2018, and the scope of the research area is  $122^\circ 18' - 123^\circ 17' E$ ,  $37^\circ 16' - 38^\circ 16' N$ . All numerical experiments are performed using 64-bit Windows 10 on a 3.60 GHz Intel Core i7-7700U CPU, 1080 Ti GPU with 16 GB memory. The proposed algorithms are programmed in the MATLAB R2016a and Python. The original data set has 4257 trajectories with 3153535 timestamped points. Data cleaning is a necessary step for trajectory mining. The 3904 trajectories with 2908685 timestamped points are preserved after data cleaning.

## 4.2. Visualization and Analysis of Trajectory Compression

The data set after data cleaning is compressed based on the proposed ADPS algorithm. The trajectories before and after compression are presented in Fig. 7. The research area is shown in Fig. 7 (a), clearly illustrating the vessel routing system and different routes. The original and compressed trajectories are displayed in Fig. 7 (b) and Fig. 7 (d),

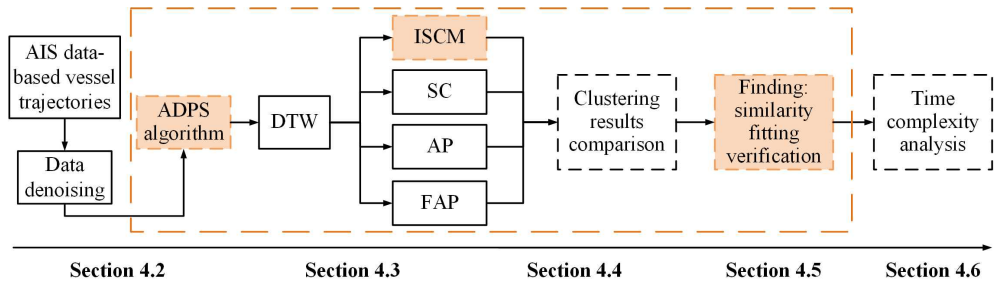


Fig. 6: The experiment flowchart.

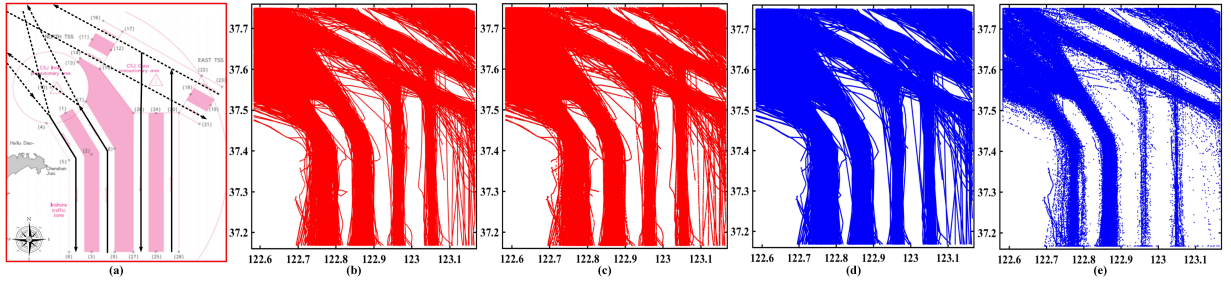


Fig. 7: The original vessel trajectories and compressed trajectories in the CJP, (a) the research area, (b) the original trajectory data, (c) the original point data, (d) the compressed trajectories, (e) the compressed point data.

respectively. The comparison of Fig. 7 (c) and Fig. 7 (e) shows that the data volume is significantly reduced, while the critical features are effectively retained after trajectory compression. The original number of points on all trajectories is 2908685, and it is 207267 after the trajectory compression. It can be seen that the data volume of trajectories after the ADPS method has reduced significantly from Fig. 7.

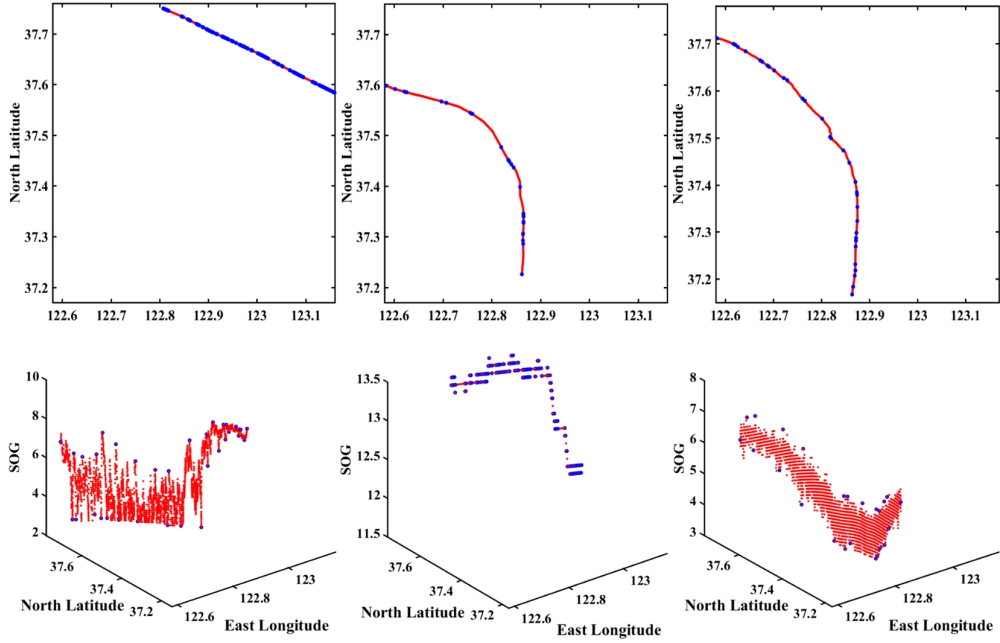
To further verify the performance of the proposed ADPS algorithm, the trajectories are randomly selected to visually compare and analyze the trajectory information before and after the compression. The 2D and 3D image visualization of vessel trajectories before and after trajectory compression are shown in Fig. 8. Fig. 8 (top and bottom) represents the comparison of 2D and 3D visualization of vessel trajectories before and after trajectory compression, respectively. Red points represent the original trajectory information, while blue ones indicate the compressed trajectory information. The comparative results can clearly show that the compressed trajectory retains the essential structure information and the speed information of the feature points in the original trajectory. Therefore, the performance of the ADPS algorithm is further verified.

Visualization of the number of points in the trajectory data set and the threshold is shown in Fig. 9. The red line represents the number of points before compression, while the blue line expresses the number of points after compression. Fig. 9 (c) displays the compression threshold of each trajectory, and the range is  $[0, 0.18]$ .

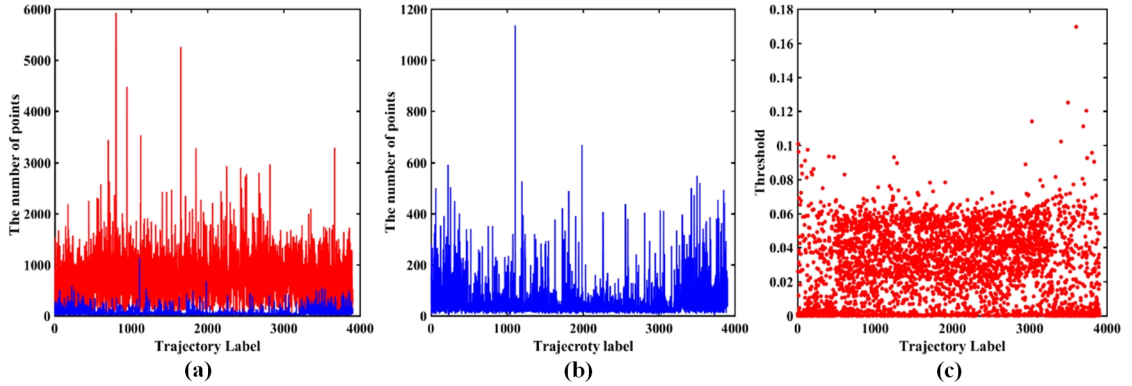
The findings reveal that the proposed ADPS algorithm significantly compresses the AIS trajectories while maintaining their main geometrical structures and key information, and also automatically calculates appropriate thresholds for different trajectories.

### 4.3. The Number of Clustering Centers

The comparison of the *SC* index, *CHS* index, and *DBI* index based on the number of different clusters is shown in Fig. 10. The larger *SC* and *CHS* indexes are better, and the smaller *DBI* index is better. Therefore, the higher values of the *SC* index and *CHS* index in Fig. 10 (top) are better, and the number of common cluster centers corresponding to the larger values in the two lines is the best cluster center number. Similarly, after the analysis of the *CHS* and *DBI* indexes of all the figures in Fig. 10, the different indexes of 23 and 25 are good, especially when the number of cluster centers is 23. The *SC* index is relatively the largest, and the *CHS* index is relatively the smallest, while the *DBI* index is also relatively small. Therefore, 23 is selected as the number of clustering centers. In the clustering performance evaluation, the clustering results of 23 and 25 will be further compared to prove the validity of the chosen



**Fig. 8:** 2D (Top) and 3D (Bottom) trajectory comparisons before and after compression. Red and blue labels denote the original point data and the compressed point, respectively.

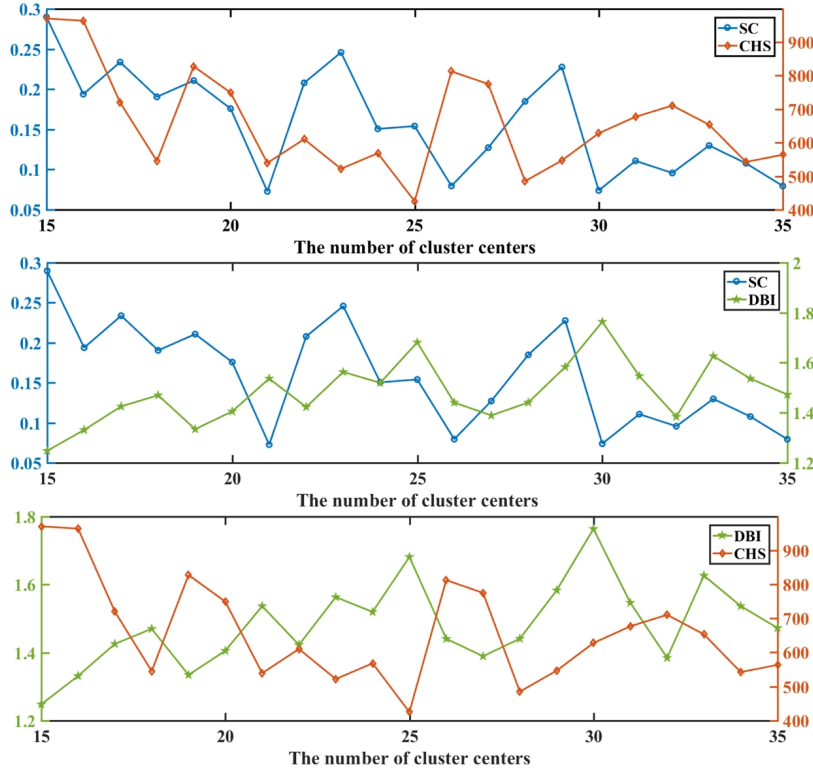


**Fig. 9:** Comparison of trajectory information before and after compression, (a) the number of points before and after compression, (b) the number of points based on the ADPS method, (c) the threshold of each trajectory.

number of cluster centers.

#### 4.4. Clustering Results and Comparative Analysis of Four Different Methods

Clustering analysis is carried out on the data sets before and after trajectory compression. The clustering performance of different clusters is further compared based on the proposed ISCM method, the OSC method, the AP method, and the FAP method. The clustering results corresponding to the different methods and the different number of clustering centers ( $k=23$  and  $k=25$ ) are shown in Fig. 11. As displayed in Fig. 11, we can see that the clustering results of four methods after compression (columns 3 and 4) are better than those before compression (columns 1 and 2). These results show that trajectory compression is critical in maritime traffic pattern extraction. Moreover, the visualization results of the proposed ISCM method (the first row in Fig. 11) are superior to those of the OSC method (the second row in Fig. 11) from Figs. 11 (a)-(h). It reveals that the clustering result based on the proposed ISCM method is better than those based on the OSC method. Similarly, comparing the results of Figs. 11(a)-(d), (i)-(m), and (n)-(q), the



**Fig. 10:** The determination of the number of clustering centers based on the results of clustering evaluation indexes, the comparison of *SC* and *CHS* index (top), the comparison of *SC* and *DBI* index (middle), and the comparison of *DBI* and *CHS* index (bottom).

results of the proposed ISCM method are better than the FAP and AP methods.

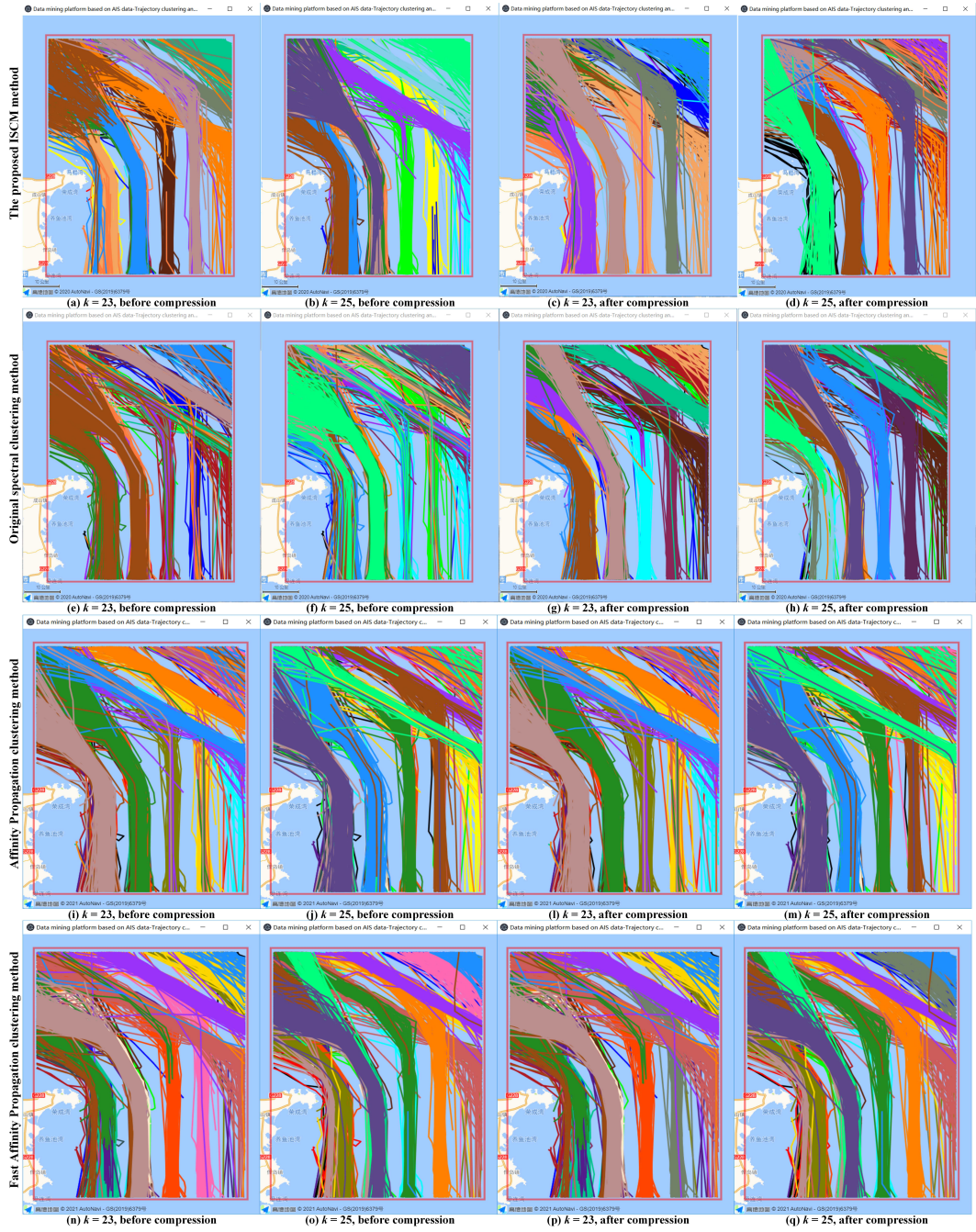
Finally, we analyze the clustering results of different clustering numbers based on the proposed ISCM method. Fig. 11 (a) and Fig. 11 (c) are the clustering result of clustering centers  $k=23$  before and after trajectory compression, respectively, while Fig. 11 (b) and Fig. 11 (d) are the results with  $k=25$ . It visually shows that the clustering results after trajectory compression based on the proposed ISCM method are better than those before trajectory compression.

From the comparative analysis in Fig. 11, the clustering performance after compression is better than before, and the clustering results of clustering centers  $k=23$  are better than that of  $k=25$  based on the proposed ISCM method. The superiority of the proposed ISCM method is also verified and shown in Fig. 12 ( $k=23$ ) and Fig. 13 ( $k=25$ ), respectively. To further highlight the effectiveness of the proposed method, we present the results of  $k=23$  and  $k=25$  based on the AP (Fig. 14 and Fig. 15) and FAP (Fig. 16, and Fig. 17) clustering methods in the Appendix.

As shown in Fig. 12, different clusters clearly show the routes and moving patterns of vessels. In contrast, clusters 1, 2, 3, 4, 5, 7, 9, 12, 14, 17, 21, and 23 are the primary movement patterns of different vessel types corresponding to various navigation directions in Fig. 12. Some of the trajectories are misclassified in clusters 5, 6, 9, 10, 13, 16, 18, 21, and 22 in Fig. 12. The clustering result with  $k=25$  is shown in Fig. 13, and some trajectories are misclassified in clusters 2, 3, 4, 5, 6, 7, 11, 12, 17, 21, 22, 23, and 25. The comparative analysis of different clusters can further verify the clustering performance of  $k=23$  is better than that of  $k=25$ . The vessel movement patterns are extracted based on the clustering results. They are conducive to planning navigation routes and detecting abnormal trajectories.

#### 4.5. Statistical Analysis based on Gaussian Mixture Model

Statistics-based methods are extensively applied to provide quantitative analysis in maritime traffic research. The DTW algorithm measures the similarity in the same cluster, and then the distance matrix will be obtained. The similarity distribution is fitted based on the distance matrix by the normal fitted curve in each cluster. The statistical analysis results of different clusters in Section 4.4 are listed in Table 1. The number of trajectories in 23 groups is clearly shown in Table 1.

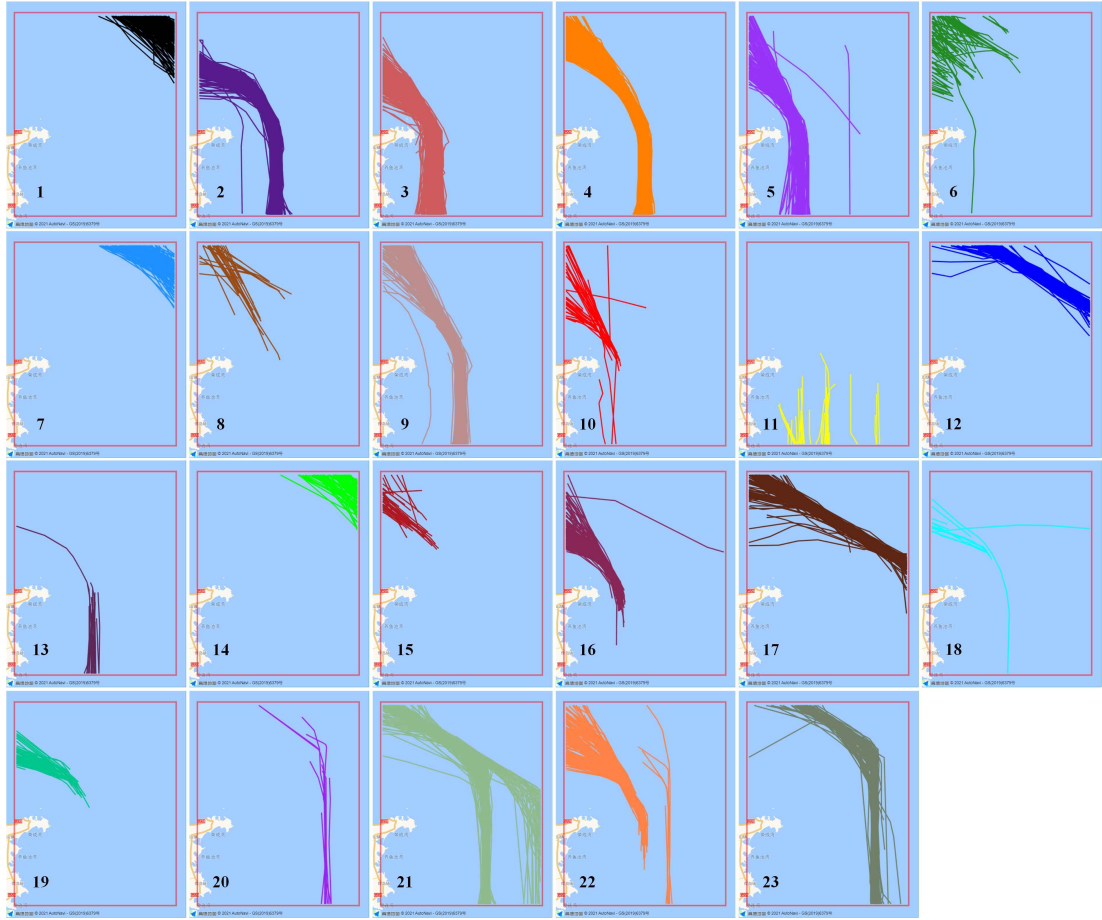


**Fig. 11:** Clustering results based on different methods and different number of clustering centers in trajectory data set before and after trajectory compression.

To directly and effectively compare the results of the fitting results, the simplified formula is used to express the Gaussian mixture model. The coefficients  $a_i$ ,  $b_i$ , and  $c_i$ ,  $i = 1, 2, \dots, 23$  are visualized in Table 1.

$$f(x) = \sum_{i=1}^n a_i \exp\left(-\frac{1}{2c_i^2} (x - b_i)^2\right). \quad (16)$$

The evaluation indexes of the fitting performance mainly include  $R^2$ , *Adjusted  $R^2$* , and Root Mean Square Error



**Fig. 12:** The clustering results with 23 clusters ( $k=23$ ) based on the proposed ISCM clustering method

(*RMSE*). From Table 1, the distribution of clusters 10, 11, 13, and 18 follow Gaussian single model. The trajectory similarity of cluster 1, 2, 3, 4, 5, 6, 7, 8, 12, 14, 15, 19, 20, 22 and 23 obey two-dimensional Gaussian distribution. The distribution of clusters 9, 16, 17, and 21 follow a Gaussian mixture distribution.

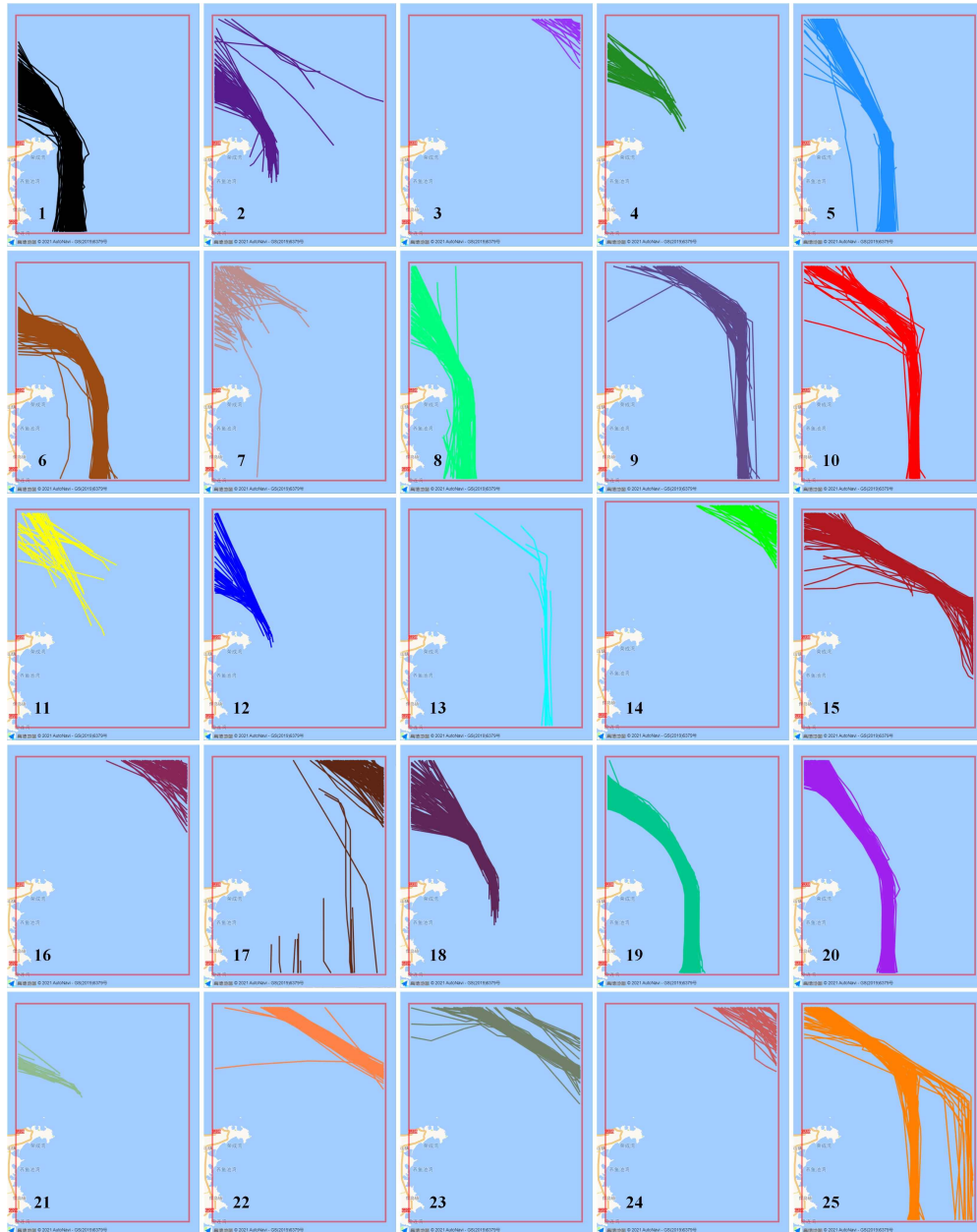
The number of trajectories, the different evaluation indexes, and the fitting results in different clusters are presented in Table 1. It can be seen that the range of the  $R^2$  and *Adjusted*  $R^2$  is [0.9281, 0.9989] and [0.9132, 0.9988], respectively. Both  $R^2$  and *Adjusted*  $R^2$  are close to 1; meanwhile, *RMSE* is also small. The statical analysis results from Table 1 reveal that the proposed ISCM method is valid and appropriate. The fitting results of similarity distribution in different clusters further verify the validity and effectiveness of our proposed methodology.

#### 4.6. The Time Complexity Analysis

The time complexity of the proposed ADPS method, the DTW method, the proposed ISCM method, the OSC method, the AP method, and the FAP method is  $O(n^2)$ , where  $n$  indicates the number of ship trajectories. The comparison results are shown in Table 2.

The data set in the CJP includes 3904 trajectories with 2908685 points after the trajectory preprocessing. After the trajectory compression, 3904 trajectories are represented by 207267 points. The time spent on the trajectory compression is 2318.863s. By comparison, data storage has been decreased by 14 times. The running time of the DTW algorithm in the data set before and after compression was 151053.265s and 14296.23s, respectively, which represents a 10.5-fold reduction in distance computing time.

The running time of four different clustering methods before and after compression is listed in Table 2, with increasing order of the FAP method, the AP method, the ISCM method, and the OSC method. Furthermore, the clustering



**Fig. 13:** The clustering results with 25 clusters ( $k=25$ ) based on the proposed ISCM clustering method

performance is evaluated based on three clustering internal evaluation indicators: *SC*, *CHS*, and *DBI*. As can be seen from Table 2, the proposed ISCM method is better than the other methods. The clustering performance further verifies the effectiveness of the proposed unsupervised hierarchical methodology.

The running time of different steps after trajectory compression is lower than that before trajectory compression, and the performance of distance calculation and clustering analysis in the data set after trajectory compression is improved. Through the comparative analysis of experiments, the proposed ISCM method can extract similar and useful behaviors and patterns better. The comparison results before and after trajectory compression further proved the feasibility and effectiveness of the proposed methodology. The unsupervised hierarchical methodology can extract and retain key trajectory characteristics, realize automatic compression, reduce the distance calculation time, improve the

**Table 1**

The fitting results of different clusters based on Gaussian mixture distribution.

Cluster ID	Number	$R^2$	Adjusted $R^2$	RMSE	a1	b1	c1	a2	b2	c2	a3	b3	c3
1	161	0.9596	0.9583	24.00	285.6	-0.730	1.084	113.2	-1.345	0.289	\	\	\
2	627	0.9962	0.9962	126.4	7496	-1.442	0.113	3436	-1.294	0.191	\	\	\
3	535	0.9886	0.9885	162.9	5981	-1.539	0.087	3113	-1.405	0.167	\	\	\
4	739	0.9931	0.993	142.4	4153	0.604	0.192	2897	0.895	0.359	\	\	\
5	144	0.9984	0.9983	24.49	2847	1.349	1.523	1298	1.792	3.816	\	\	\
6	100	0.9763	0.975	14.77	168.1	2.451	2.272	157.8	5.95	5.841	\	\	\
7	118	0.9794	0.9787	16.41	163.6	2.274	1.640	271.5	7.916	8.151	\	\	\
8	029	0.9281	0.9132	6.402	34.95	3.925	1.645	48.36	11.35	19.92	\	\	\
9	242	0.9886	0.9882	82.16	3700	-1.663	0.043	2138	-1.591	0.091	208.1	-1.34	0.630
10	038	0.9462	0.9451	6.728	4313	-4.055	1.265	\	\	\	\	\	\
11	033	0.9887	0.9879	4.497	115.7	-1.714	1.209	\	\	\	\	\	\
12	196	0.9989	0.9988	73.46	4853	-1.613	0.090	9250	-2.008	0.512	\	\	\
13	011	0.9501	0.9418	2.716	47.16	-2.101	0.795	\	\	\	\	\	\
14	086	0.9604	0.9568	16.09	106.9	-1.345	0.272	187	-0.656	1.258	\	\	\
15	040	0.9507	0.9435	7.620	77.96	-0.947	1.085	37.01	-1.471	0.087	\	\	\
16	133	0.9833	0.9823	32.44	552.1	-1.635	0.049	719.5	-1.525	0.126	282.5	-1.287	0.428
17	149	0.9937	0.9932	42.93	1335	-1.608	0.055	1456	-1.519	0.117	607.2	-1.342	0.278
18	020	0.9966	0.9953	4.038	5865	-6.838	2.873	\	\	\	\	\	\
19	047	0.9403	0.9348	9.445	35.26	-1.229	0.274	73.72	-1.359	1.207	\	\	\
20	010	0.9965	0.9922	1.249	161.2	-1.347	0.1234	7.654	-0.518	0.691	\	\	\
21	220	0.9846	0.9839	50.18	1453	-1.552	0.075	810.6	-1.424	0.156	282.2	-0.861	0.918
22	157	0.9877	0.9871	99.57	4527	-1.668	0.120	9.338e+15	-38.49	6.551	\	\	\
23	069	0.9757	0.9721	39.46	803.3	-1.514	0.134	339.6	-1.295	0.304	\	\	\

**Table 2**

The time complexity comparison results of different algorithms.

Symbol	Time complexity	Raw dataset	Trajectory dataset after preprocessing	Trajectory dataset after compression	$SC$	$CHS$	$DBI$
$N_{Tra}$	\	4257	3904	3904	\	\	\
$N_{Point}$	\	3153535	2908685	207267	\	\	\
$T_{ADPS}(s)$	$O(n^2)$	\	\	2318.863	\	\	\
$T_{DTW}(s)$	$O(n^2)$	\	151053.265	14296.23	\	\	\
$T_{ISCM}(s)$	$O(n^2)$	\	287.917	260.681	0.2457	521.8965	1.5630
$T_{OSC}(s)$	$O(n^2)$	\	285.510	259.317	0.1531	217.0527	3.5760
$T_{AP}(s)$	$O(n^2)$	\	290.230	265.350	0.1873	359.7881	2.9231
$T_{FAP}(s)$	$O(n^2)$	\	282.326	256.827	0.2265	490.7683	1.8235

\*  $N_{Tra}$  and  $N_{Point}$  represent the number of trajectories and points in the dataset, respectively.  $T_{ADPS}(s)$ ,  $T_{DTW}(s)$ ,  $T_{ISCM}(s)$ ,  $T_{OSC}(s)$ ,  $T_{AP}(s)$ , and  $T_{FAP}(s)$  indicate the running time based on the ADPS, DTW, ISCM, original spectral clustering method, affinity propagation clustering method, and fast affinity propagation clustering method, respectively.  $SC$ ,  $CHS$ , and  $DBI$  respectively represent three internal evaluation indexes.

accuracy of clustering, and mine the hidden knowledge. It is of great significance for maritime traffic pattern extraction for knowledge discovery, as demonstrated by maritime AIS data. The similarity fitting analysis after trajectory clustering further verifies the validity and rationality of the proposed methodology. It is feasible to extract and mine the hidden patterns and critical knowledge based on the novel unsupervised hierarchical methodology.

## 5. Conclusions and Future Research

This paper proposed an unsupervised hierarchical methodology to extract knowledge and navigation characteristics from hidden pattern information and better understand maritime situational awareness. It consists of three inherent elements (i.e. trajectory compression, trajectory similarity measurement, and trajectory clustering) in a sequence. The findings make contributions to the maritime pattern extraction from three parts: the ADPS algorithm, the ISCM method, and the new clustering evaluation criterion. The experimental results also show that the ADPS algorithm can automatically set appropriate thresholds for different trajectories and reduce the cost of data storage. Furthermore, the

proposed ISCM method can convert the trajectories into graph space with points, determine the number of clustering centers based on the inner evaluation indexes in advance, and avoid local solution. Finally, the deep statistical analysis of within-cluster trajectory similarity shows that all the similarity distribution between trajectories in each cluster follows a Gaussian mixture model. The finding hence provides a reliable standard for the clustering performance evaluation.

To generalize the proposed methodology beyond the scope of ship trajectory analysis, the adaptive selection of the number of clustering centers and the new similarity measurement methods can be further studied. Lastly, the application of trajectory mining to realize the operational practicality, reliability, and robustness in maritime transportation industries can be further enhanced.

## CRedit authorship contribution statement

**Huanhuan Li:** Investigation, Data curation, Conceptualization, Methodology, Software, Experiments, Writing - review & editing. **Jasmine Siu lee Lam:** Supervision, Writing - review & editing, Validation. **Zaili Yang:** Methodology, Writing - review & editing. **Jingxian Liu:** Supervision, Writing - original draft. **Ryan Wen Liu:** Writing - original draft & review, Validation. **Maohan Liang:** Investigation, Visualization. **Yan Li:** Investigation, Visualization, Validation.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

This work was supported by the grants from the Nanyang Technological University Project (04SBS000097C120), the National Key R&D Program of China (2018YFC1407400), and the EU project GOLF (H2020-MSCA-RISE-2017-777742).

## References

- Allen, M., 2018. Hierarchical model, in: The sage encyclopedia of communication research methods. Thousand Oaks, pp. 661–663.
- Andrienko, G., Andrienko, N., Fuchs, G., Garcia, J.M.C., 2018. Clustering trajectories by relevant parts for air traffic analysis. *IEEE Transactions on Visualization and Computer Graphics* 24, 34–44. doi:10.1109/tvcg.2017.2744322.
- Arguedas, V.F., Pallotta, G., Vespe, M., 2017. Maritime traffic networks: From historical positioning data to unsupervised maritime traffic monitoring. *IEEE Transactions on Intelligent Transportation Systems* 19, 722–732.
- Aslam, S., Michaelides, M.P., Herodotou, H., 2020. Internet of ships: A survey on architectures, emerging applications, and challenges. *IEEE Internet of Things journal* 7, 9714–9727.
- Atev, S., Miller, G., Papanikolopoulos, N.P., 2010. Clustering of vehicle trajectories. *IEEE transactions on intelligent transportation systems* 11, 647–657.
- Bomberger, N., Rhodes, B., Seibert, M., Waxman, A., 2006. Associative learning of vessel motion patterns for maritime situation awareness, in: *proc. 9th Int. Conf. Inf. Fusion (FUSION)*, Florence, Italy. pp. 1–8.
- Chen, C.H., Khoo, L.P., Chong, Y.T., Yin, X.F., 2014. Knowledge discovery using genetic algorithm for maritime situational awareness. *Expert Systems with Applications* 41, 2742–2753.
- Coscia, P., Braca, P., Millefiori, L.M., Palmieri, F.A., Willett, P., 2018. Multiple ornstein–uhlenbeck processes for maritime traffic graph representation. *IEEE Transactions on Aerospace and Electronic Systems* 54, 2158–2170.
- De Mulder, W., 2014. Instability and cluster stability variance for real clusterings. *Information Sciences* 260, 51–63.
- Ding, F., Wang, J., Ge, J., Li, W., 2018. Anomaly detection in large-scale trajectories using hybrid grid-based hierarchical clustering. *Int. J. Robot. Autom* 33, 5–206.
- Douglas, D.H., Peucker, T.K., 1973. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Cartographica: the international journal for geographic information and geovisualization* 10, 112–122.
- Elhamifar, E., Vidal, R., 2013. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE transactions on pattern analysis and machine intelligence* 35, 2765–2781.
- Fu, Y., Liu, X., Sarkar, S., Wu, T., 2021. Gaussian mixture model with feature selection: An embedded approach. *Computers & Industrial Engineering* 152, 107000.
- Gaffney, S., Smyth, P., 1999. Trajectory clustering with mixtures of regression models, in: *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 63–72.
- Gaffney, S.J., Robertson, A.W., Smyth, P., Camargo, S.J., Ghil, M., 2007. Probabilistic clustering of extratropical cyclones using regression mixture models. *Climate dynamics* 29, 423–440.

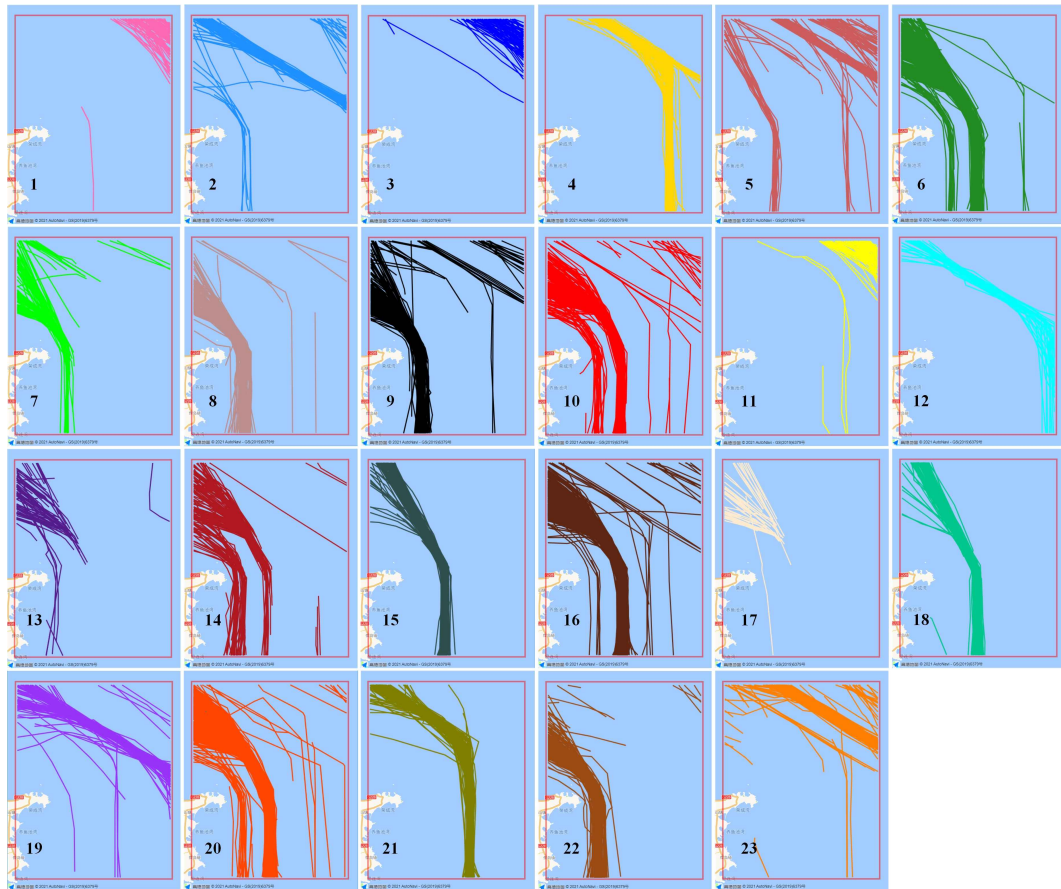
- Graziano, M.D., Renga, A., Moccia, A., 2019. Integration of automatic identification system (ais) data and single-channel synthetic aperture radar (sar) images by sar-based ship velocity estimation for maritime situational awareness. *Remote Sensing* 11, 2196.
- Han, Y., Zhu, L., Cheng, Z., Li, J., Liu, X., 2018. Discrete optimal graph clustering. *IEEE transactions on cybernetics* 50, 1697–1710.
- Hong, Z., Chen, Y., Mahmassani, H.S., 2017. Recognizing network trip patterns using a spatio-temporal vehicle trajectory clustering algorithm. *IEEE Transactions on Intelligent Transportation Systems* 19, 2548–2557.
- Hu, B., Liu, R.W., Wang, K., Li, Y., Liang, M., Li, H., Liu, J., 2017. Statistical analysis of massive ais trajectories using gaussian mixture models, in: 2017 2nd International Conference on Multimedia and Image Processing (ICMIP), IEEE. pp. 113–117.
- Huang, Y., Li, Y., Zhang, Z., Liu, R.W., 2020. Gpu-accelerated compression and visualization of large-scale vessel trajectories in maritime iot industries. *IEEE Internet of Things Journal* 7, 10794–10812.
- Huo, Y., Dong, X., Beatty, S., 2020. Cellular communications in ocean waves for maritime internet of things. *IEEE Internet of Things Journal* 7, 9965–9979.
- Ji, Y., Qi, L., Balling, R., 2021. A dynamic adaptive grating algorithm for ais-based ship trajectory compression. *The Journal of Navigation* , 1–17.
- Kontopoulos, I., Varlamis, I., Tserpes, K., 2021. A distributed framework for extracting maritime traffic patterns. *International Journal of Geographical Information Science* 35, 767–792.
- Lee, D.S., 2005. Effective gaussian mixture learning for video background subtraction. *IEEE transactions on pattern analysis and machine intelligence* 27, 827–832.
- Lee, J.G., Han, J., Li, X., 2008. Trajectory outlier detection: A partition-and-detect framework, in: 2008 IEEE 24th International Conference on Data Engineering, IEEE. pp. 140–149.
- Lee, J.G., Han, J., Whang, K.Y., 2007. Trajectory clustering: a partition-and-group framework, in: Proceedings of the 2007 ACM SIGMOD international conference on Management of data, pp. 593–604.
- Lehmann, A.L., Alvares, L.O., Bogorny, V., 2019. Smsm: a similarity measure for trajectory stops and moves. *International Journal of Geographical Information Science* 33, 1847–1872.
- Lei, P.R., 2020. Mining maritime traffic conflict trajectories from a massive ais data. *Knowledge and Information Systems* 62, 259–285.
- Li, H., Liu, J., Liu, R.W., Xiong, N., Wu, K., Kim, T.h., 2017. A dimensionality reduction-based multi-step clustering method for robust vessel trajectory analysis. *Sensors* 17, 1792.
- Li, H., Liu, J., Wu, K., Yang, Z., Liu, R.W., Xiong, N., 2018a. Spatio-temporal vessel trajectory clustering based on data mapping and density. *IEEE Access* 6, 58939–58954.
- Li, H., Liu, J., Yang, Z., Liu, R.W., Wu, K., Wan, Y., 2020. Adaptively constrained dynamic time warping for time series classification and clustering. *Information Sciences* 534, 97–116.
- Li, Q., Lam, J.S.L., 2017. Conflict resolution for enhancing shipping safety and improving navigational traffic within a seaport: vessel arrival scheduling. *Transportmetrica A: Transport Science* 13, 727–741.
- Li, X., Zhao, K., Cong, G., Jensen, C.S., Wei, W., 2018b. Deep representation learning for trajectory similarity computation, in: 2018 IEEE 34th international conference on data engineering (ICDE), IEEE. pp. 617–628.
- Li, Y., Liu, R.W., Liu, J., Huang, Y., Hu, B., Wang, K., 2016. Trajectory compression-guided visualization of spatio-temporal ais vessel density, in: 2016 8th International Conference on Wireless Communications & Signal Processing (WCSP), IEEE. pp. 1–5.
- Liang, M., 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 Engineering* 225, 108803.
- Liu, B., de Souza, E.N., Matwin, S., Sydow, M., 2014. Knowledge-based clustering of ship trajectories using density-based approach, in: 2014 IEEE International Conference on Big Data (Big Data), IEEE. pp. 603–608.
- Liu, J., Li, H., Yang, Z., Wu, K., Liu, Y., Liu, R.W., 2019. Adaptive douglas-peucker algorithm with automatic thresholding for ais-based vessel trajectory compression. *IEEE Access* 7, 150677–150692.
- Liu, R.W., Nie, J., Garg, S., Xiong, Z., Zhang, Y., Hossain, M.S., 2020. Data-driven trajectory quality improvement for promoting intelligent vessel traffic services in 6g-enabled maritime iot systems. *IEEE Internet of Things Journal* 8, 5374–5385.
- Loh, W.K., Mane, S., Srivastava, J., 2011. Mining temporal patterns in popularity of web items. *Information Sciences* 181, 5010–5028.
- Magirou, E.F., Psaraftis, H.N., Bouritas, T., 2015. The economic speed of an oceangoing vessel in a dynamic setting. *Transportation Research Part B: Methodological* 76, 48–67.
- Mascaro, S., Nicholso, A.E., Korb, K.B., 2014. Anomaly detection in vessel tracks using bayesian networks. *International Journal of Approximate Reasoning* 55, 84–98.
- Mazzarella, F., Vespe, M., Damalas, D., Osio, G., 2014. Discovering vessel activities at sea using ais data: Mapping of fishing footprints, in: Proc. 17th Int. Conf. Inf. Fusion (FUSION), Salamanca, Spain. pp. 1–7.
- Millefiori, L.M., Braca, P., Bryan, K., Willett, P., 2016. Modeling vessel kinematics using a stochastic mean-reverting process for long-term prediction. *IEEE Transactions on Aerospace and Electronic Systems* 52, 2313–2330.
- Morel, M., Achard, C., Kulpa, R., Dubuisson, S., 2018. Time-series averaging using constrained dynamic time warping with tolerance. *Pattern Recognition* 74, 77–89.
- Murray, B., Perera, L.P., 2020. A dual linear autoencoder approach for vessel trajectory prediction using historical ais data. *Ocean Engineering* 209, 107478.
- Nanni, M., Pedreschi, D., 2006. Time-focused clustering of trajectories of moving objects. *Journal of Intelligent Information Systems* 27, 267–289.
- Pallotta, G., Vespe, M., Bryan, K., 2013a. Traffic knowledge discovery from ais data, in: Proceedings of the 16th International Conference on Information Fusion, IEEE. pp. 1996–2003.
- Pallotta, G., Vespe, M., Bryan, K., 2013b. Vessel pattern knowledge discovery from ais data: A framework for anomaly detection and route prediction. *Entropy* 15, 2218–2245.
- Pan, J., Jiang, Q., Shao, Z., 2014. Trajectory clustering by sampling and density. *Marine Technology Society Journal* 48, 74–85.
- Pitsikalis, M., Bereta, K., Vodas, M., Zissis, D., Artikis, A., 2020. Event processing for maritime situational awareness, in: Big Data Analytics for

- Time-Critical Mobility Forecasting. Springer, pp. 255–274.
- Rhodes, B.J., Bomberger, N.A., Zandipour, M., 2007. Probabilistic associative learning of vessel motion patterns at multiple spatial scales for maritime situation awareness, in: 2007 10th International Conference on Information Fusion, IEEE. pp. 1–8.
- 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., Soares, C.G., 2020. Data mining approach to shipping route characterization and anomaly detection based on ais data. *Ocean Engineering* 198, 106936.
- Saalfeld, A., 1999. Topologically consistent line simplification with the douglas-peucker algorithm. *Cartography and Geographic Information Science* 26, 7–18.
- Shang, F., Jiao, L., Shi, J., Wang, F., Gong, M., 2012. Fast affinity propagation clustering: A multilevel approach. *Pattern recognition* 45, 474–486.
- Shi, D., Zhu, L., Li, Y., Li, J., Nie, X., 2019. Robust structured graph clustering. *IEEE transactions on neural networks and learning systems* 31, 4424–4436.
- Taghizadeh, S., Elekes, A., Schäler, M., Böhm, K., 2019. How meaningful are similarities in deep trajectory representations? *Information Systems* , 101452.
- Talat, R., Obaidat, M.S., Muzammal, M., Sodhro, A.H., Luo, Z., Pirbhulal, S., 2020. A decentralised approach to privacy preserving trajectory mining. *Future Generation Computer Systems* 102, 382–392.
- Tang, C., Wang, H., Zhao, J., Tang, Y., Yan, H., Xiao, Y., 2021. A method for compressing ais trajectory data based on the adaptive-threshold douglas-peucker algorithm. *Ocean Engineering* 232, 109041.
- Tienaaah, T., Stefanakis, E., Coleman, D., 2015. Contextual douglas-peucker simplification. *Geomatica* 69, 327–338.
- Tu, E., Zhang, G., Mao, S., Rachmawati, L., Huang, G.B., 2020. Modeling historical ais data for vessel path prediction: A comprehensive treatment. *arXiv preprint arXiv:2001.01592* .
- Tu, E., Zhang, G., Rachmawati, L., Rajabally, E., Huang, G.B., 2017. Exploiting ais data for intelligent maritime navigation: A comprehensive survey from data to methodology. *IEEE Transactions on Intelligent Transportation Systems* 19, 1559–1582.
- Vespe, M., Gibin, M., Alessandrini, A., Natale, F., Mazzarella, F., Osio, G.C., 2016. Mapping eu fishing activities using ship tracking data. *Journal of Maps* 12, 520–525.
- Vespe, M., Visentini, I., Bryan, K., Braca, P., 2012. Unsupervised learning of maritime traffic patterns for anomaly detection , 1–5.
- Von Luxburg, U., 2007. A tutorial on spectral clustering. *Statistics and computing* 17, 395–416.
- Wang, F., Zhu, L., Liang, C., Li, J., Chang, X., Lu, K., 2020a. Robust optimal graph clustering. *Neurocomputing* 378, 153–165.
- Wang, L., Zheng, K., Tao, X., Han, X., 2018. Affinity propagation clustering algorithm based on large-scale data-set. *International Journal of Computers and Applications* 40, 1–6.
- Wang, R., Zhou, J., Jiang, H., Han, S., Wang, L., Wang, D., Chen, Y., 2021. A general transfer learning-based gaussian mixture model for clustering. *International Journal of Fuzzy Systems* 23, 776–793.
- Wang, S., Yan, R., Qu, X., 2019. Development of a non-parametric classifier: Effective identification, algorithm, and applications in port state control for maritime transportation. *Transportation Research Part B: Methodological* 128, 129–157. URL: <https://www.sciencedirect.com/science/article/pii/S0191261519301390>, doi:<https://doi.org/10.1016/j.trb.2019.07.017>.
- Wang, W., Xia, F., Nie, H., Chen, Z., Gong, Z., Kong, X., Wei, W., 2020b. Vehicle trajectory clustering based on dynamic representation learning of internet of vehicles. *IEEE Transactions on Intelligent Transportation Systems* , 1–10.
- Wei, T., Feng, W., Chen, Y., Wang, C.X., Ge, N., Lu, J., 2021. Hybrid satellite-terrestrial communication networks for the maritime internet of things: key technologies, opportunities, and challenges. *IEEE Internet of Things Journal* , 1–28.
- Xiao, Z., Fu, X., Zhang, L., Goh, R.S.M., 2019. Traffic pattern mining and forecasting technologies in maritime traffic service networks: A comprehensive survey. *IEEE Transactions on Intelligent Transportation Systems* 21, 1796–1825.
- Xu, X., Cui, D., Li, Y., Xiao, Y., 2021. Research on ship trajectory extraction based on multi-attribute dbscan optimisation algorithm. *Polish Maritime Research* 28, 136–148.
- Yan, Z., Xiao, Y., Cheng, L., He, R., Ruan, X., Zhou, X., Li, M., Bin, R., 2020. Exploring ais data for intelligent maritime routes extraction. *Applied Ocean Research* 101, 102271.
- Yang, D., Wu, L., Wang, S., Jia, H., Li, K.X., 2019. How big data enriches maritime research—a critical review of automatic identification system (ais) data applications. *Transport Reviews* 39, 755–773.
- Yang, T., Chen, J., Zhang, N., 2020. Ai-empowered maritime internet of things: A parallel-network-driven approach. *IEEE Network* 34, 54–59.
- Yao, D., Zhang, C., Zhu, Z., Huang, J., Bi, J., 2017. Trajectory clustering via deep representation learning, in: 2017 international joint conference on neural networks (IJCNN), IEEE. pp. 3880–3887.
- Yap, W.Y., Lam, J., 2020. Data analytics for international transportation management. *Res. Transp. Bus. Manag.* 34, 100470.
- Yu, Y., Wang, Q., Wang, X., Wang, H., He, J., 2013. Online clustering for trajectory data stream of moving objects. *Computer science and information systems* 10, 1293–1317.
- Zhang, B., Zhu, L., Sun, J., Zhang, H., 2018. Cross-media retrieval with collective deep semantic learning. *Multimedia Tools and Applications* 77, 22247–22266.
- Zhang, R., Xie, P., Jiang, H., Xiao, Z., Wang, C., Liu, L., 2019a. Clustering noisy trajectories via robust deep attention auto-encoders, in: 2019 20th IEEE International Conference on Mobile Data Management (MDM), IEEE. pp. 63–71.
- Zhang, S., Liu, Z., Cai, Y., Wu, Z., Shi, G., 2016. Ais trajectories simplification and threshold determination. *The Journal of Navigation* 69, 729–744.
- Zhang, Y., Liu, A., Liu, G., Li, Z., Li, Q., 2019b. Deep representation learning of activity trajectory similarity computation, in: 2019 IEEE International Conference on Web Services (ICWS), IEEE. pp. 312–319.
- Zhao, L., Shi, G., 2019a. A novel similarity measure for clustering vessel trajectories based on dynamic time warping. *The Journal of Navigation* 72, 290–306.

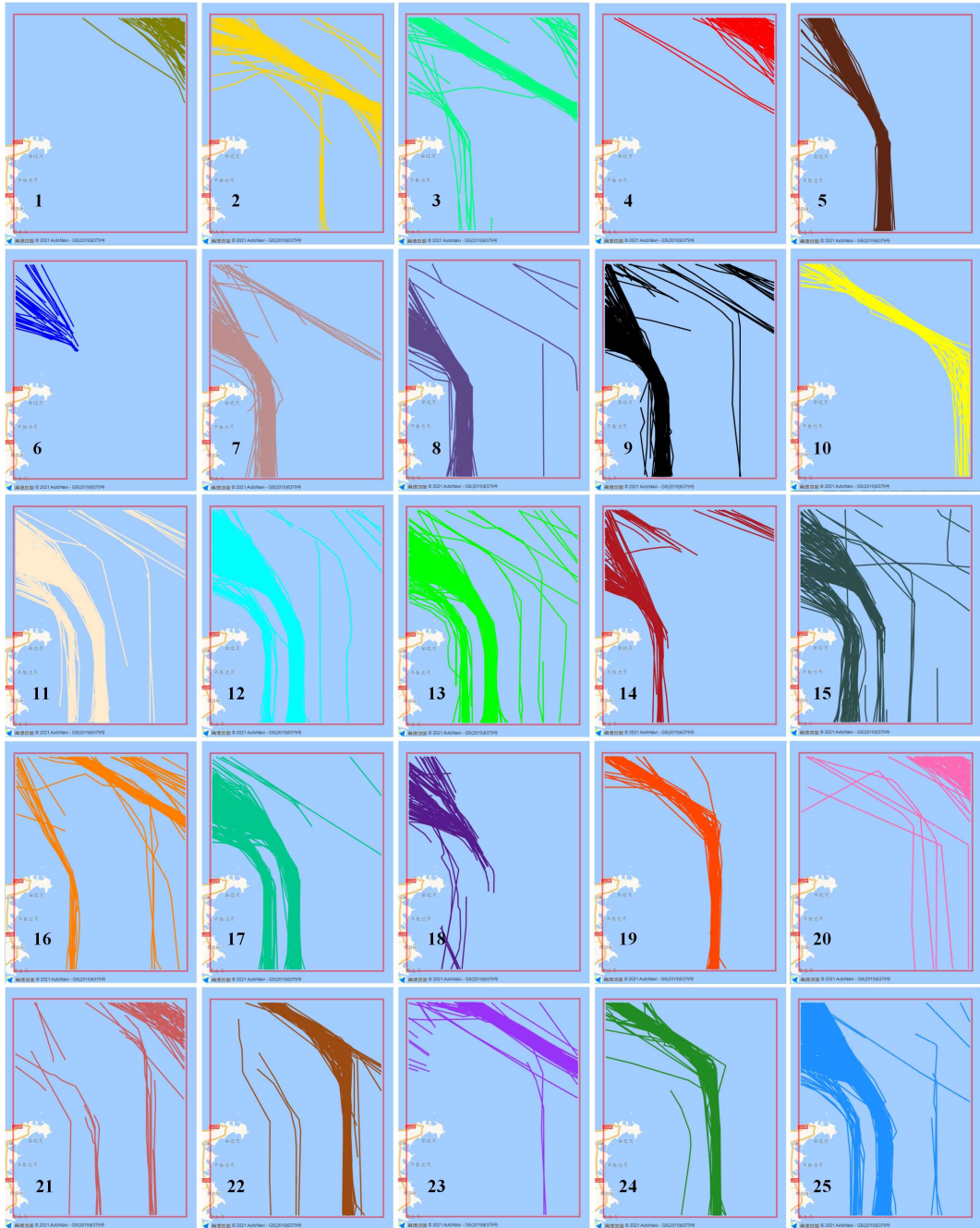
- Zhao, L., Shi, G., 2019b. A trajectory clustering method based on douglas-peucker compression and density for marine traffic pattern recognition. *Ocean Engineering* 172, 456–467.
- Zhao, P., Qin, K., Ye, X., Wang, Y., Chen, Y., 2017. A trajectory clustering approach based on decision graph and data field for detecting hotspots. *International Journal of Geographical Information Science* 31, 1101–1127.
- Zhen, R., Jin, Y., Hu, Q., Shao, Z., Nikitakos, N., 2017. Maritime anomaly detection within coastal waters based on vessel trajectory clustering and naïve bayes classifier. *The Journal of Navigation* 70, 648.
- Zheng, Y., Zhou, X., 2011. *Computing with spatial trajectories*. Springer Science & Business Media.
- Zhou, Y., Daamen, W., Vellinga, T., Hoogendoorn, S., 2019. Review of maritime traffic models from vessel behavior modeling perspective. *Transportation Research Part C: Emerging Technologies* 105, 323–345.

## Appendix

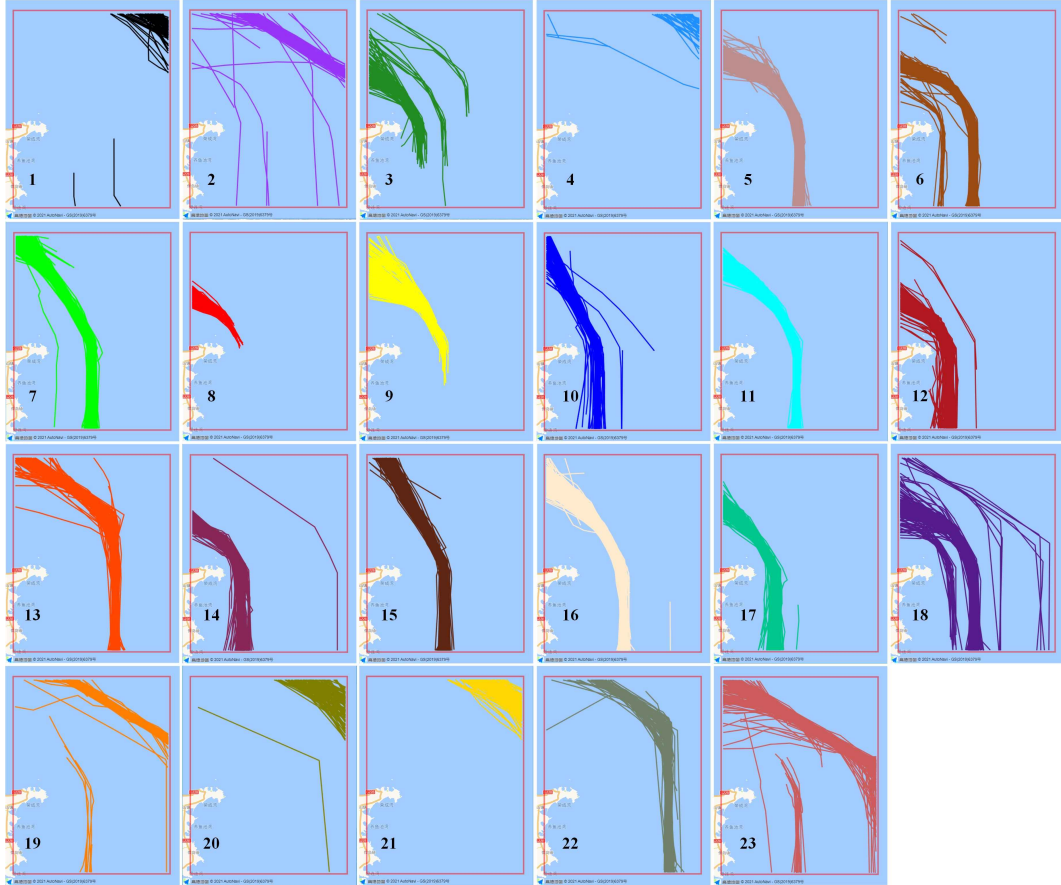
The clustering results based on the affinity propagation (AP) clustering method and fast affinity propagation (FAP) clustering method are displayed in Fig. 14 and Fig. 15, Fig. 16, and Fig. 17, respectively. The comparison results further verify that the FAP method has better clustering performance than the AP method. From Fig. 12, Fig. 14, and Fig. 16, it can be clearly seen that the performance of the proposed methodology is superior to that of other methods. Furthermore, the results in Fig. 13, Fig. 15, and Fig. 17 also show that the performance of the proposed methodology is better than that of other methods. All the comparison experiments verify that the performance of the proposed methodology is the best in the maritime pattern extraction.



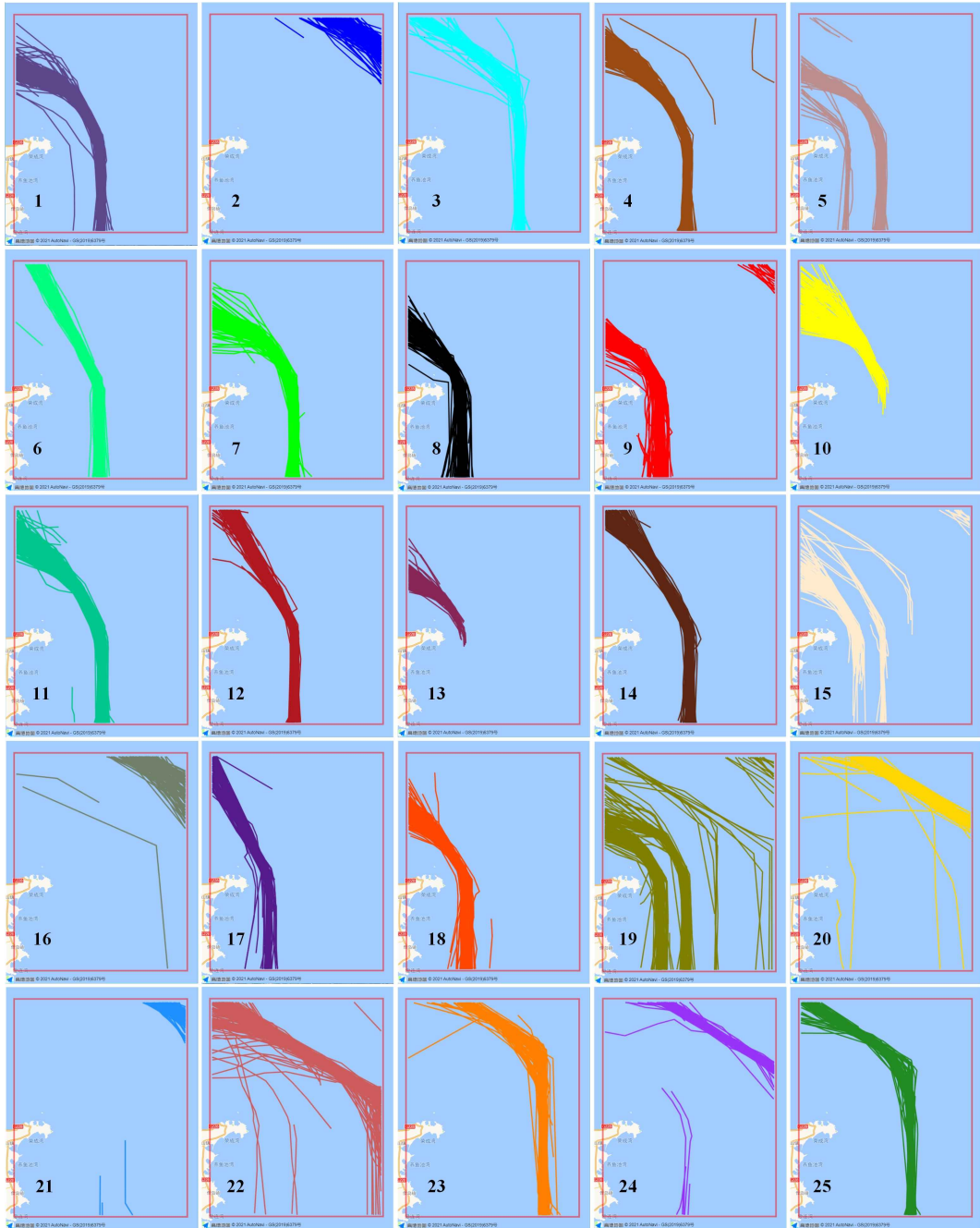
**Fig. 14:** The clustering results with 23 clusters ( $k=23$ ) based on the AP clustering method.



**Fig. 15:** The clustering results with 25 clusters ( $k=25$ ) based on the AP clustering method.



**Fig. 16:** The clustering results with 23 clusters ( $k=23$ ) based on the FAP clustering method.



**Fig. 17:** The clustering results with 25 clusters ( $k=25$ ) based on the FAP clustering method.