sustainability

MDPI

Article

Extracting Shipping Route Patterns by Trajectory Clustering Model Based on Automatic Identification System Data

Pan Sheng and Jingbo Yin *

State Key Laboratory of Ocean Engineering, Department of International Shipping, School of Naval Architecture, Ocean and Civil Engineering, Shanghai Jiao Tong University, Shanghai 200240, China; shengv5@sjtu.edu.cn
* Correspondence: jingboyin@sjtu.edu.cn

Received: 28 May 2018; Accepted: 3 July 2018; Published: 5 July 2018

Abstract: Shipping route analysis is essential for vessel traffic management and relies on professional technical facilities for collecting and recording specific information about vessel behaviors. The recent Automatic Identification System (AIS) onboard has been made available to provide ship-related information for the research. However, the complexity and large quantity of AIS data overload traditional surveillance operations and increase the difficulty of vessel traffic analysis. An unsupervised approach is urgently desired to effectively convert the raw AIS data to regular shipping route patterns. In this paper, we proposed a trajectory clustering model based on AIS data to analyze the shipping routes. The whole model consists of four parts: Data preprocessing, structure similarity measurement, clustering, and representative trajectory extraction. Our model comprehensively considered the geospatial information and contextual features of ship trajectory. The revised density-based clustering algorithm could automatically classify different shipping routes with trajectory features without prior knowledge. The experimental evaluation showed the effectiveness of the proposed model by real AIS data from Port of Tianjin. The results contribute to the further understanding of shipping route patterns and assists maritime authorities and the officers in stable and sustainable vessel traffic management.

Keywords: shipping route analysis; ship trajectory clustering model; AIS data

1. Introduction

With the rapid development of the global economy, maritime transportation becomes increasingly important and represents approximately 90% of global trade by volume and 70% by value [1]. However, maritime transportation is recognized as a relatively risky mode. In addition to market risk [2], the large volume and potential damage to a vessel, which could bring loss of life and property and destruction of the marine environment, must be considered. According to the Global Integrated Shipping Information System (GISIS) database, hundreds of reported maritime accidents occurred in 2017 [3]. Therefore, implementation of procedures to decrease maritime damage and ensure a safe and stable maritime traffic environment would be of great significance. As described in the 2012 to 2017 strategic plan for the International Maritime Organization (IMO), maritime safety and security have become a priority in vessel traffic management [4].

In numerous sailing events, acquiring knowledge of the sailing area in advance and selecting a reliable shipping route is essential for ships' navigational safety. Traditionally, the officers mainly acquire voyage-related information through the electronic chart display system or other navigation materials and take their personal, limited sailing experience in evaluating the performance of sailing. It concentrates more on HydroMet information and lacks the concrete sailing practice from other vessels.

With the development of technology, more and more surveillance systems have been installed in the harbor and on board to improve the level of maritime situational awareness and strengthen maritime safety. Automatic Identification System (AIS) is one kind of surveillance system. The International Maritime Organization's International Convention for the Safety of Life at Sea requires AIS to be fitted aboard all ships of 300 and more gross tonnage, cargo ships of 500 and more tonnage and all passenger ships regardless of size since 2004 [5]. The self-reporting system could achieve global coverage and send a ship's location-related information including identity number, speed, course and heading direction to coastal maritime authorities and other ships every few minutes which makes it possible to track the ship's movement and monitor the ship's behavior in detail. The collected sequences of AIS logs could include vessel trajectory. Analysis of these vessel trajectories reveals typical movement patterns and provides an overview of the maritime traffic. This shipping route knowledge could provide mariners with a more practical reference than other sailing experience. Further, these discovered patterns from AIS data could support decisions for the sustainable development of maritime traffic.

As mentioned above, the real-time and historical AIS data seem to be an ideal source to reconstruct ships' movement trajectories and extract shipping route knowledge. However, there are a number of challenges in analyzing maritime trajectory data. First, unlike the constrained movement of vehicles on road networks, vessels are moving relatively freely in the maritime environment. There are main recommended shipping channels for vessels to follow, but it is difficult to define the normal movement of the vessel. Second, the refresh rate of AIS is every few seconds or minutes according to different motion mode. The volume of AIS data to be processed is very large and complex. Traditional analysis and evaluation methods are overloaded with the dramatically increasing quantity of AIS data. An automated and effective solution for exploring ships' movement patterns and extracting shipping route knowledge is needed urgently for maritime traffic surveillance and sustainability management.

The purpose of this paper is to present a method that makes it possible to extract relevant shipping knowledge from AIS data through the use of unsupervised learning techniques. In this work, we propose a density-based clustering model for mining AIS data of various sailing features. The remainder of the paper is organized as follows: Section 2 reviews the various methods in mining trajectory data as well as the limitations of existing algorithms; Section 3 presents the procedure of the shipping trajectory clustering model; Section 4 evaluates the effectiveness of the clustering model with empirical AIS data taken from Port of Tianjin, and finally, conclusions and recommendations for future work are provided in Section 5.

2. Literature Review

Recent improvements in location acquisition technologies and tracking facilities have made it possible to collect trajectory data of moving objects. There is increasing interest in performing analysis to discover knowledge from these trajectory data. Many researches have put much effort into trajectory data analysis and proposed some algorithms to mine movement pattern. To find common movement trends from complicated trajectory data, we usually need to group similar trajectories into clusters. A general clustering approach is used to represent a trajectory with similarity of feature vectors. However, it is not easy to generate a uniform criterion for different trajectories as different trajectories contain various and complex attributes. Generally, clustering algorithms could be classified into four categories: partitioning methods (e.g., K-means [6]), hierarchical methods (e.g., the BIRCH [balances iterative and clustering using hierarchies] algorithm [7]), density-based methods (e.g., the DBSCAN [density-based spatial clustering of applications with noise] algorithm [8], the OPTICS [ordering points to identify the clustering structure] algorithm [9]), and grid-based methods (e.g., the sting algorithm [10]). Gaffney and Smyth [11] proposed grouping similar trajectories into clusters by using a regression mixture model and the Expectation–Maximization (EM) algorithm. This algorithm clusters trajectories with respect to the overall distance between two entire trajectories. However, dealing with the whole trajectories might miss some common characteristics in sub-trajectories. Thus, Lee et al. [12] proposed a framework to partition trajectories into small segments and then built groups of

close trajectory segments into clusters using the Trajectory-Hausdorff distance. A representative path was later found for each cluster of segments. Since trajectory data were often received incrementally, Li et al. [13] further proposed an incremental clustering algorithm, aiming to reduce the computational cost and storage of received trajectory data.

In the maritime domain, Knorr et al. [14] applied the similarity measurement to compare the features of vessel trajectories such as starting and ending points, direction and velocity. The similarity distance between two trajectories was measured by the summed score of these feature vectors. Bomberger et al. [15] divided the entire region of interest area into different grid locations and applied the grid-based method to explore the vessel motion patterns for maritime situational awareness. Dahlbom and Niklasson [16] introduced a density-based clustering method to establish the representation of normal vessel behavior for coastal surveillance. The traffic was investigated, and trajectories of detected vessels were further processed to maintain knowledge and conceive the representation of the normal course of movements. Auslander et al. [17] supposed that different kinds of maritime traffic were characterized by different levels of complexity and presented two global and two local anomaly detection algorithms, whose performance varied depending on the maritime traffic type. Vespe et al. [18] proposed a waypoint density algorithm to compress a large volume of AIS data into a list of adequate waypoints and ultimately delineate sea lanes and subsequently routes through a set of lines connecting waypoints. A typical density-base algorithm was conducted by Pallotta et al. [19] to cluster the vessel trajectory into the route object, waypoint object and stationary object and construct a framework to detect the anomalous behavior and predict ship movement. Bo Liu et al. [20] developed the method with speed and direction information. The method could detect moving and stopping areas and extract the representative center of the clustered region. Lei [21] modeled vessel sailing patterns by considering the spatial, sequential and dynamic features of vessels, and defined a combined index of outlying scores to detect anomalous vessel behavior. To effectively detect the frequent region, the author developed a clustering method based on a grid-based approach and set two parameters to define the cell size and mitigate the loss problem. Zhen et al. [22] formulated a new approach to detect maritime anomalies based on a combination of vessel trajectory clustering and Naïve Bayes Classifier. In the process of vessel trajectory clustering, the structure similarity measurement was introduced to solve the difficulty in spatial and directional similarity, while the optimal combination parameters were obtained quantitatively.

Among all these studies, the approaches in References [18,20] seem to be the most suitable for exploring AIS trajectory data. However, the existing approaches did not work well in high-density vessel area because they did not preprocess the trajectory data such as segmentation or compression to select the most important data. In addition, the approaches just set local clustering factor and could not cluster the whole data with different attributes. In our work, we first selected the most important characteristics from the original data according to the trajectory shape and other attributes. Then, we modified DBSCAN algorithm with structure similarity measurement to generate the most significant clusters and to find vessel traffic patterns from AIS dataset.

3. Ship Trajectory Clustering Model

In this section, we describe a method to build the ship trajectory clustering model. The model assumes that the majority of the ship behaviors correspond to normal activities. The method consists of four main parts: Data preprocessing, structure similarity distance measurement, clustering, and representative route extraction.

As mentioned in the introduction, AIS transmits ship-related data including static data, such as ship name, call sign and MMSI (Maritime Mobile Service Identity), and dynamic information such as speed, heading and course. The types and content of AIS data are numerous, but our research only needs the most remarkable features. For example, from our perspective, ship name, call sign, IMO and MMSI are similar in meaning to identify and recognize the ship. Therefore, we could use only MMSI, the unique 9-digit identification number, to represent the ship and arranged it as an

index of each ship. In addition, feature selection is an important aspect of the model and depends on the expected performance of the method. The minimal set of information for shipping route discovery is the location (longitude and latitude) of the ship. However, if we want to explore more information about the ship routes, we need to enrich the model with contextual knowledge which leads to better clustering and classification results. Here, we decided to distinguish the lanes oriented in opposite directions according to the Traffic Separation Scheme, and eventually introduced directional information, the course over ground (COG), into the model. Since we also wanted to identify speed, we incorporated another attribute, the speed over ground (SOG), in our trajectory representation. Consequently, the vessel trajectory could be expressed by the following formula:

$$T_i = (TS_{P_{t_1}P_{t_2}}, \cdots, TS_{P_{t_{m-1}}P_{t_m}})^i = (P_{t_1}, P_{t_2}, \cdots, P_{t_m})^i \tag{1}$$

where T_i represents the whole vessel trajectory, i is MMSI of the vessel and $TS_{P_{t_i}P_{t_j}}$ represents the trajectory segment. P_{t_m} is location point at timestamp t_m which consists of the feature vector $P_{t_m} = [x, y, v, \omega]_{t_m}$. (x, y) are the longitude and latitude, v is the SOG and ω is the COG. Therefore, the trajectory formula can also be written as a matrix expression:

$$T_i = (TS_{P_{t_1}P_{t_2}}, \cdots, TS_{P_{t_{m-1}}P_{t_m}})^i = (P_{t_1}, P_{t_2}, \cdots, P_{t_m})^i = [x, y, v, \omega]^i_{t_m} \tag{2}$$

3.1. AIS Data Preprocessing

To reduce the computation time and improve model efficiency, we first needed to process the trajectory data and selected the most characteristic points to represent the original trajectory.

According to Lee et al. [11], there are two basic criteria on which to pick out trajectory representative points: Conciseness and preciseness. Conciseness means that the number of selected trajectory points should be as small as possible. Preciseness means that the difference between the original trajectory and its representative set should be as small as possible. In other words, the processed trajectory should maintain, as far as possible, the same shape and features as the original one. Here, we applied these two principles to select the AIS trajectory data to meet the above requirements.

First, we selected the most characteristic points of a ship's trajectory. The shipborne AIS sends a message when the ship's COG or SOG changes in short time cycles. According to this regular pattern, our selection criteria could be defined as:

$$CRC = \frac{\left|\omega_{P_{t_m}} - \omega_{P_{t_n}}\right|}{t_m - t_n} \tag{3}$$

where CRC is the change rate of the COG; $\omega_{P_{t_m}}$ is the COG at the location P_{t_m}; t_m is the time timestamp.

$$CRS = \frac{\left|v_{P_{t_m}} - v_{P_{t_n}}\right|}{t_m - t_n} \tag{4}$$

where CRS is the change rate of the SOG; $v_{P_{t_m}}$ is the SOG at the location P_{t_m}.

If the change rates of trajectory points are bigger than the thresholds- α for CRC and β for CRS, it means that the ship's course over ground and speed changes greatly at these locations, making a significant impact on the ship's sailing trajectory. Therefore, we select these points to present the semantic meaning of a trajectory.

However, only using these characteristic points may not satisfy the requirements of preciseness because the trajectory shape criterion is not considered above. Therefore, we applied another approach, the minimum descriptionlength (MDL) principle [23], to maintain the trajectory shape as far as possible. We used a parameter λ to measure the distance between unselected points and characteristic trajectory segments [24]. If the distance exceeds the threshold, the unselected points will be chosen into the characteristic points set.

3.2. Ship's Trajectory Structural Similarity

The similarity measurement is the fundamental basis to implement the following clustering and classification. For our trajectory formula, every location point, P, is a multi-dimensional point with feature vectors. Considering these features of every trajectory, we applied the structure similarity (SSIM) [25] method to measure the similarity between different ship trajectories.

(1) Spatial distance measurement

The spatial distance refers to the physical distance between two trajectories and is calculated by longitude and latitude data. Among different distance functions, Hausdorff distance is the more appropriate to measure the spatial similarity of vessel trajectory because it does not require the same number of points between different trajectories. The Hausdorff distance of vessel trajectories can be written as:

$$D_S(TS^i_{P_{t_{m-1}}P_{t_m}}, TS^j_{P_{t_{n-1}}P_{t_n}}) = h(TS^i_{P_{t_{m-1}}P_{t_m}}, TS^j_{P_{t_{n-1}}P_{t_n}}) \tag{5}$$

$$h(TS^i_{P_{t_{m-1}}P_{t_m}}, TS^j_{P_{t_{n-1}}P_{t_n}}) = \max \min \|(P_{t_{m-1}}, P_{t_m}) - (P_{t_{n-1}}, P_{t_n})\| \tag{6}$$

where $h(TS^i_{P_{t_{m-1}}P_{t_m}}, TS^j_{P_{t_{n-1}}P_{t_n}})$ is the direct Hausdorff distance from trajectory segment $TS^i_{P_{t_{m-1}}P_{t_m}}$ to $TS^j_{P_{t_{n-1}}P_{t_n}}$, which identifies the point as the farthest from its nearest neighbors; $\|(P_{t_{m-1}}, P_{t_m}) - (P_{t_{n-1}}, P_{t_n})\|$ is the Euclidean distance for comparing location points of trajectory segment $TS^i_{P_{t_{m-1}}P_{t_m}}$ and $TS^j_{P_{t_{n-1}}P_{t_n}}$.

(2) Directional distance measurement

In addition to spatial differences, moving direction has a great impact on vessel behavior as well. According to Traffic Separation Scheme (TSS) rules, ships within the main traffic-lane should sail in the general direction of that lane [26]. The traffic pattern will be totally different due to different navigation directions. We have introduced COG in the trajectory formula to distinguish the trajectory direction. Thus, the directional distance function can be written as:

$$D_D(TS^i_{P_{t_{m-1}}P_{t_m}}, TS^j_{P_{t_{n-1}}P_{t_n}}) = \begin{cases} \min(\|TS^i_{P_{t_{m-1}}P_{t_m}}\|, \|TS^j_{P_{t_{n-1}}P_{t_n}}\|) * sin(\theta),\ 0^\circ \le \theta \le 90^\circ \\ \min(\|TS^i_{P_{t_{m-1}}P_{t_m}}\|, \|TS^j_{P_{t_{n-1}}P_{t_n}}\|),\ 90^\circ \le \theta \le 180^\circ \end{cases} \tag{7}$$

where $\|TS^i_{P_{t_{m-1}}P_{t_m}}\|$ is the Euclidean length of trajectory segment $TS^i_{P_{t_{m-1}}P_{t_m}}$, and θ $(0^\circ \le \theta \le 180^\circ)$ is the intersecting angle between $TS^i_{P_{t_{m-1}}P_{t_m}}$ and $TS^j_{P_{t_{n-1}}P_{t_n}}$.

An illustration of spatial and directional distance between trajectory segments was given in Figure 1.

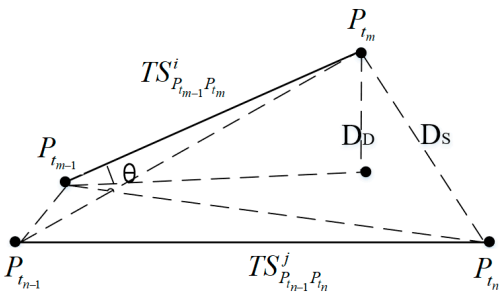


Figure 1. Illustration of trajectories' spatial and directional similarity distance.

(3) Speed distance measurement

Speed is another critical feature vector in maritime traffic pattern analysis, and we have taken it into our trajectory expression. Here, we used average speed of trajectory to express the similarity distance.

$$D_V(TS^i_{P_{t_{m-1}}P_{t_m}}, TS^j_{P_{t_{n-1}}P_{t_n}}) = \left| V_{avg}(TS^i_{P_{t_{m-1}}P_{t_m}}) - V_{avg}(TS^j_{P_{t_{n-1}}P_{t_n}}) \right| \tag{8}$$

$$V_{avg}(TS^i_{P_{t_{m-1}}P_{t_m}}) = \frac{\left| v_{P_{t_m}} - v_{P_{t_{m-1}}} \right|}{t_m - t_{m-1}} \tag{9}$$

where $V_{avg}(TS^i_{P_{t_{m-1}}P_{t_m}})$ is the average speed of trajectory segment $TS^i_{P_{t_{m-1}}P_{t_m}}$.

To ensure the same numerical magnitude, we used the liner transformation to normalize the spatial, directional, and speed similarity distance. The general formula of liner transformation is shown as:

$$D_{norm} = (D - D_{min})/(D_{\max} - D_{min}) \tag{10}$$

where D stands for the distance to be normalized, the $D_{\max}$ and D_{min} are the maximum and minimum value of each spatial, directional and speed similarity distance.

Finally, based on the above normalized spatial and directional distance, we propose a synthetic distance function to calculate the trajectories' structure similarity (SSIM) distance. The mathematical expression was written as:

$$SDIS(TS^i_{P_{t_{m-1}}P_{t_m}}, TS^j_{P_{t_{n-1}}P_{t_n}}) = W_S * D_S + W_D * D_D + W_V * D_V \tag{11}$$

where W_S, W_D, W_V are weights and their values are nonnegative and $W_S + W_D + W_V = 1$.

In this way, we obtain a pairwise distance matrix to measure the similarity between each pair of trajectory segments.

3.3. Ship Trajectory Clustering

We revised the DBSCAN algorithm with the above synthetic similarity distance, $SDIS(TS^i_{P_{t_{m-1}}P_{t_m}}, TS^j_{P_{t_{n-1}}P_{t_n}})$, to create the cluster. Set T denote all processed trajectories. A trajectory segment $TS^i_{P_{t_{m-1}}P_{t_m}} \in T$ is defined as the core trajectory segment if $\left| N_\varepsilon(TS^i_{P_{t_{m-1}}P_{t_m}}) \right| \geq MinLns$ where $N_\varepsilon(TS^i_{P_{t_{m-1}}P_{t_m}}) = \left\{ TS^j_{P_{t_{n-1}}P_{t_n}} \in T \middle| SDIS(TS^i_{P_{t_{m-1}}P_{t_m}}, TS^j_{P_{t_{n-1}}P_{t_n}}) \leq \varepsilon \right\}$. Besides the core trajectory segment related to similarity distance, other definitions are similar to the concepts in DBSCAN algorithm [8]. We just linked pairs of vessel trajectories that were close together into the clusters according to their distance from each other.

To automatically recognize clusters of different vessel density, we also introduced some concepts in hierarchical clustering algorithm into our algorithm. We used core distance to express the distance from core trajectory distance to its *MinLns*-nearest neighbor. We redefined the distance metric between points as mutual reachability distance [27].

The clustering algorithm computed a membership score of each point which indicates the saturation level and centrality degree. We transited such attributes into the outlier score [28]. After setting the clustered data with outlier scores, we could label and detect the abnormal vessel location and activities.

3.4. Extract Representative Trajectory

After clustering the ship trajectories, it is necessary to extract a representative trajectory to describe the overall movement features and present the typical traffic patterns.

The representative trajectory also consisted of a sequence of location points. These points were determined by the sweep line approach [11]. Perpendicular to the clustered trajectory direction,

the sweeping line scanned the trajectories and counted the number of intersections. If the number exceeded the threshold, we would calculate coordinates of these trajectories. Otherwise, we skipped the intersection. The average coordinates of those intersections created the new representative trajectory, and the value of their average speed was used to represent the clustered trajectories. An example of the approach was given in Figure 2 and the threshold value of intersection was set to be 3 in this example.

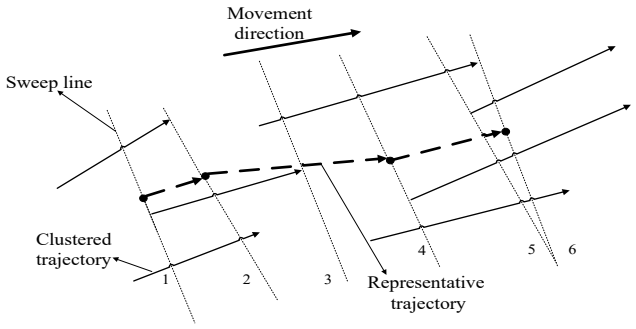


Figure 2. An example of the sweep line approach.

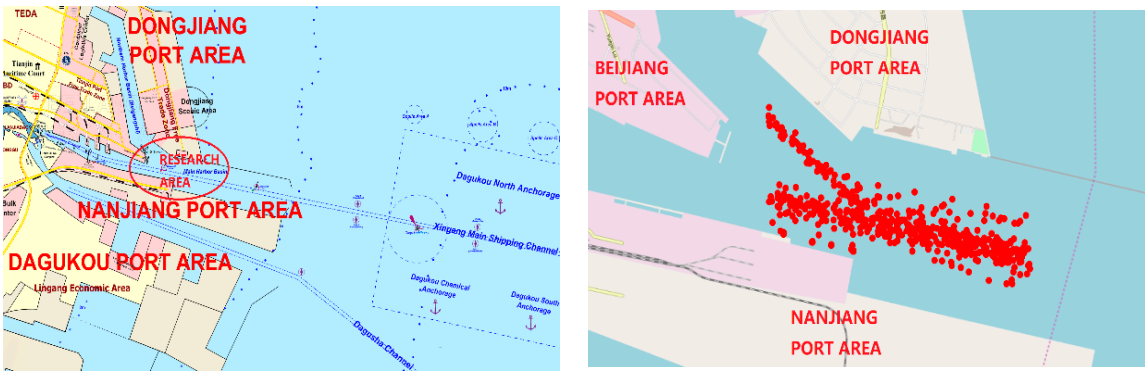
4. Case Study

In this section, we carried out an experiment to verify the effectiveness of the proposed method.

4.1. Data Description

This study collected the historical AIS data of cargo ships in Port of Tianjin from 1 October 2017 to 10 October 2017. The data were collected from http://www.shipxy.com/.

The traffic condition is complicated in Port of Tianjin, especially the central zone surrounded by three core areas of Dongjiang, Beijiang and Nanjiang around the Xingang fairway. This zone includes the original port areas and new facilities and contains mostly container and general cargo terminals. Therefore, we chose this area as our research area. Figure 3a shows the research area and red points in Figure 3b represent the locations of ships.



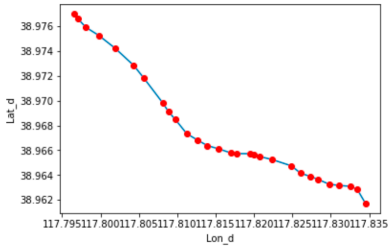
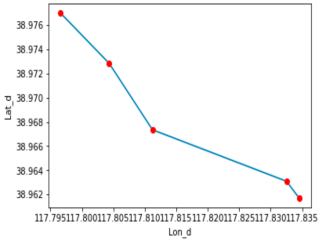
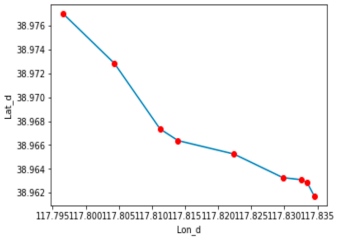
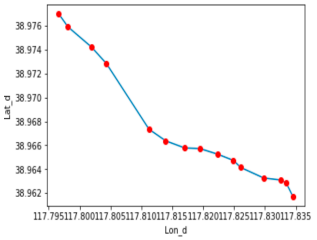
(**a**) (**b**)

Figure 3. The research area in (**a**) and collected Automatic Identification System (AIS) data in (**b**).

4.2. Parameter Setting

We set the thresholds α to be 2°/min for CRC and β to be 2 knots for CRS according to the average change rate of the course over ground and speed over ground of our collected AIS data. A relationship between λ and trajectory shape of one vessel is given in Table 1. We finally set λ equals to 3×10^{-4} to get the expected trajectory shape.

Table 1. The relationship between parameters and trajectory shape.

	Original Trajectory	$\alpha = 2,\ \beta = 2$ $\lambda = 0$	$\alpha = 2,\ \beta = 2$ $\lambda = 3 \times 10^{-4}$	$\alpha = 2,\ \beta = 2$ $\lambda = 1 \times 10^{-4}$
Number of data	30	5	9	15
Trajectory shape				

To comprehensively consider the influence of location, direction and speed, the values of W_S, W_D, W_V were set to be 0.4, 0.3, 0.3 respectively. The clustering algorithm connected the AIS data within mutual reachability distance and then increased the distance to create a newly merged cluster. If the density around a point was greater than *MinLns*, the point would be clustered [29]. We could decide the threshold value of *MinLns* according to the cluster hierarchy and the number of clusters. In our study, we set the threshold value of *MinLns* to be 15 according to the cluster hierarchy in Figure 4. The threshold value of intersections was set to be four according to the range of course over ground.

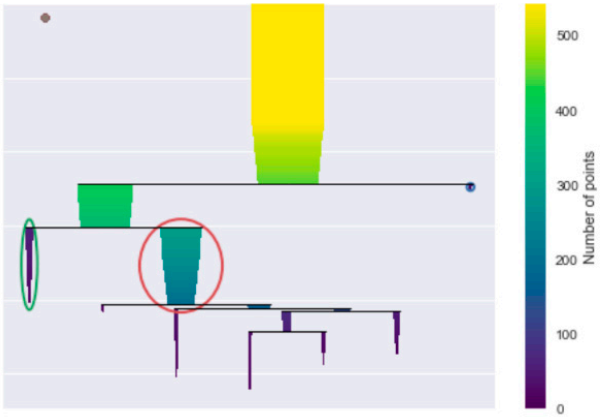
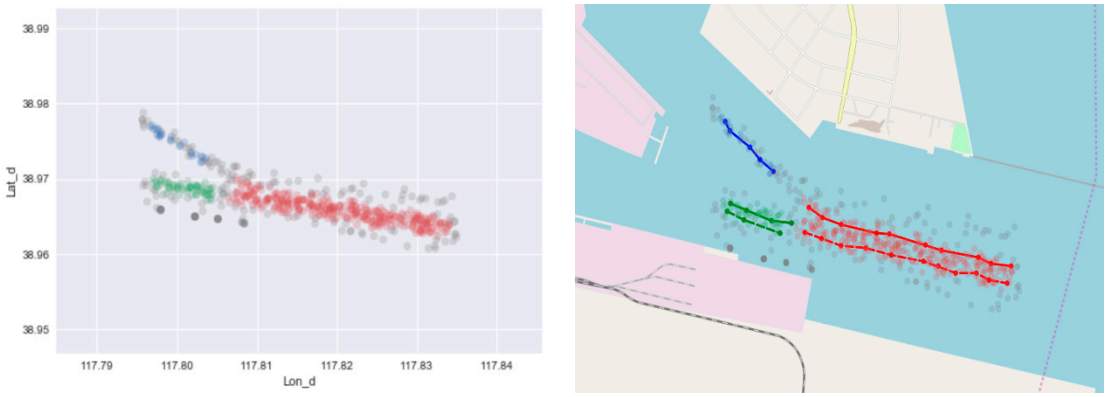


Figure 4. The cluster hierarchy.

4.3. Clustering Result

According to our previous parameters setting, we visualized the clustering result with different colors which indicated the centrality degree in Figure 5a. The processing was implemented using Python 3 and the computational time was 32 s with Inter(R) Core(TM) i5-6300 CPU 2.50 GHz. The representative trajectories were remarked in Figure 5b by two kinds of lines in corresponding colors. The solid line represented entrance direction of the vessel and the dashed line represented departure direction.



(**a**) (**b**)

Figure 5. The clustering result. (**a**) The clusters in different colors; (**b**) The representative trajectories.

From the result in Figure 5, we found that the whole AIS data were clustered into three parts. The red part conformed to vessel traffic in Xingang main shipping channel. The result gave a recommended entrance shipping lane with speed changing from 10.3 knots to 8.4 knots and the departure speed changing from 6.6 knots to 9.4 knots. The green part represented the characteristic traffic follow in Chuanzhadong channel. The average speed of departure lane was 4.6 knots, which was

possibly influenced by the speed limit near the wharf. The blue part indicated the direction towards to Northern Harbor Basin. Due to the narrow condition and mixed use of the channel, the clustering result just showed one entrance direction. The result showed a significant agreement with the current traffic schemes in this area. It could provide scientific guidance to maritime authorities and the officers to understand the vessel traffic patterns in the ports better.

In Figure 6, we showed the abnormal data out of the clustered AIS data. The general reason for these outliers was that they located too far from the most regular points as the situation in area 1. Some location points were labeled as outliers because their speed differed widely from other points like area 2. Some points were excluded due to abnormal course over ground like outliers in area 3. The result could be used to inform coastal authorities to pay attention to these abnormal behaviors of vessels and enhance the level of maritime surveillance.

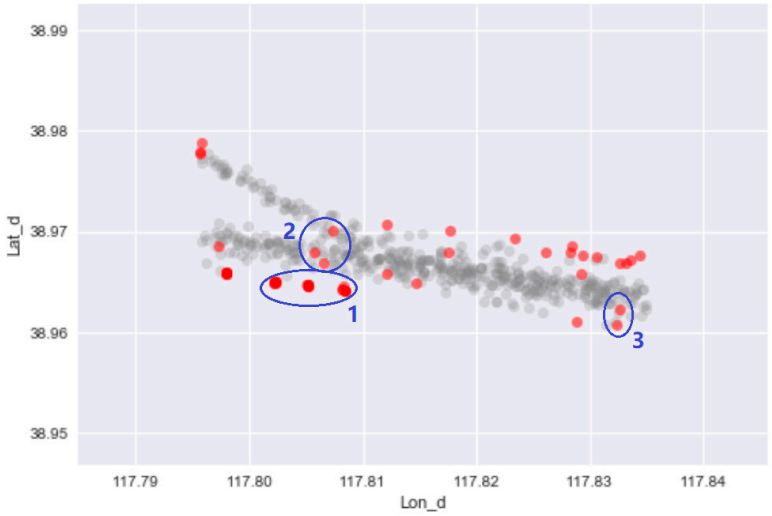


Figure 6. The abnormal AIS data.

In addition, as shown in Figure 7, if we input new AIS data into the research area, the proposed algorithm could classify and label them automatically according to the previous clustering result. With this application, we could verify the rationality and distribution of planned vessel shipping route and re-plan the route based on the clustering result.

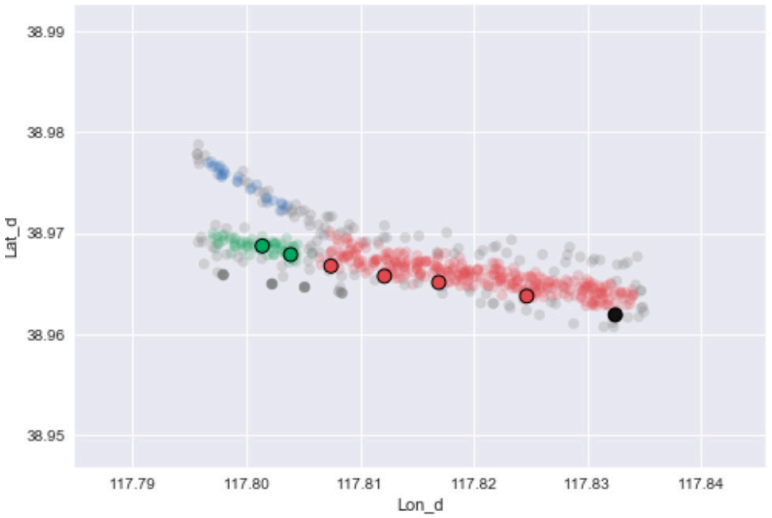


Figure 7. The clusters of new AIS data.

5. Conclusions and Discussions

In this paper, we proposed a ship trajectory clustering model to extract shipping route knowledge based on Automated Identification System (AIS) data. The important conclusions can be summarized as follows:

- The feature points and MDL principle were used to reduce the complexity and amount of AIS data while maintaining consistency with the original trajectory data.
- The revised DBSCAN algorithm was well suited to exploring AIS data. Structural similarity measurement and hierarchical density estimates were built to automatically cluster the AIS data in different trajectory features and overcome the limitations of vessel high-density.
- The experimental results demonstrate the effectiveness of this ship trajectory clustering model, which has much lower computer time, and expected result.

The results could be used to enhance security and safety in the maritime environment, which is a basis for the stable and sustainable development of the port. First, the model could convert the mass and complex data into reliable information about possible situations and help port authorities to make corresponding decisions in advance. This will help in reducing the maritime accidents. The cause of many maritime accidents is that the operators could not predict the possible situations in the area and could not react immediately according to the current traffic condition. But the model in this paper could prejudge such issues. For example, the outliers in the clustering result could notify decision makers in port of abnormal vessel behaviors and give an early warning to vessels, which increases the capability of monitoring and improve the level of maritime situational awareness. Second, the result could be applied in navigation and provide scientific guidance in shipping route planning. For example, the representative trajectory which was extracted from practical situations could help seafarers to check whether the vessel obeys the recommended navigation lanes or not. It enables seafarers to be aware of the recommended route used by the majority of vessels and avoid incidents caused by an unfamiliar navigational environment.

The future work will consist of automatically extracting shipping route patterns to enhance the level of maritime surveillance and create a stable and sustainable port environment. We will incorporate more AIS information into the model to improve the completeness of the model. For instance, only location (longitude and latitude), direction and speed attributes were considered in this model. If we were to take more contextual attributes of ship trajectory into account, the clustering result would be more accurate. Another objective would be to create a proper method to preprocess the AIS data without expert knowledge in determining the parameters.

Author Contributions: Study conception and design, analysis and interpretation of results, and manuscript preparation: P.S.; data collection and correspondence: J.Y.

Acknowledgments: This research was funded by National Social Science Fund of China, grant number [17BGL259].

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Rodrigue, J.-P. *The Geography of Transport Systems*, 4th ed.; Routledge: New York, NY, USA, 2017; p. 200.
2. Yin, J.; Luo, M.; Fan, L. Dynamics and interactions between spot and forward freights in the dry bulk shipping market. *Marit. Policy Manag.* **2016**, *44*, 1–18. [CrossRef]
3. Global Integrated Shipping Information System. Available online: https://gisis.imo.org/ (accessed on 4 April 2018).
4. International Maritime Organization. *Strategic Plan for the Organization (for Six-Year Period 2012 to 2017)*; Resolution A 1037(27); International Maritime Organization: London, UK, 2011; pp. 3–4.
5. Safety of Life at Sea (SOLAS) Convention Chapter V, Regulation 19. Available online: http://www.imo.org/en/OurWork/safety/navigation/pages/ais.aspx (accessed on 4 April 2018).

6. Lloyd, S. Least Squares Quantization in PCM. *IEEE Trans. Inf. Theory* **1982**, *28*, 129–137. [CrossRef]
7. Zhang, T.; Ramakrishnan, R.; Livny, M. BIRCH: An efficient data clustering method for very large databases. In Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data, Montreal, QC, Canada, 4–6 June 1996; pp. 103–114.
8. Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X. A Density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, Portland, OR, USA, 2–4 August 1996; pp. 226–231.
9. Ankerst, M.; Breunig, M.M.; Kriegel, H.-P.; Sander, J. OPTICS: Ordering points to identify the clustering structure. In Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data, Philadelphia, PA, USA, 31 May–3 June 1999; pp. 49–60.
10. Wisdom, M.J.; Cimon, N.J.; Johnson, B.K.; Garton, E.O.; Thomas, J.W. Spatial partitioning by mule deer and Elk in relation to traffic. In *Transactions of the North American Wildlife and Natural Resources Conference*; U.S. Forest Service: Washington, DC, USA, 2004; pp. 509–530.
11. Gaffney, S.; Smyth, P. Trajectory clustering with mixtures of regression models. In Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, 15–18 August 1999; pp. 63–72.
12. Lee, J.G.; Han, J.W.; Whang, K.Y. Trajectory clustering: A partition-and-group framework. In Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data, Beijing, China, 11–14 June 2007; pp. 593–604.
13. Li, Z.; Lee, J.-G.; Li, X.; Han, J. Incremental clustering for trajectories. In *DASFAA 2010: Database Systems for Advanced Applications*; Springer: Berlin/Heidelberg, Germany, 2010; Volume 5982, pp. 32–46. [CrossRef]
14. Knorr, E.M.; Ng, R.T.; Tucakov, V. Distance-based Outliers: Algorithms and Applications. *VLDB J.* **2000**, *8*, 237–253. [CrossRef]
15. Bomberger, N.A.; Rhodes, B.J.; Seibert, M.; Waxman, A.M. Associative learning of vessel motion patterns for maritime situation awareness. In Proceedings of the 9th Conference on Information Fusion, Florence, Italy, 10–13 July 2006; pp. 1–8.
16. Dahlbom, A.; Niklasson, L. Trajectory clustering for coastal surveillance. In Proceedings of the 10th International Conference on Information Fusion, Quebec City, QC, Canada, 9–12 July 2007; pp. 1–8.
17. Gupta, K.M.; Auslander, B.; Aha, D.W. A Comparative evaluation of anomaly detection algorithms for maritime video surveillance. In Proceedings of the SPIE 8019, Sensors, and Command, Control, Communications, and Intelligence (C3I) Technologies for Homeland Security and Homeland Defense X, Orlando, FL, USA, 2 June 2011; pp. 1684–1687. [CrossRef]
18. Vespe, M.; Visentini, I.; Bryan, K.; Braca, P. Unsupervised learning of maritime traffic patterns for anomaly detection. In Proceedings of the 9th IET Data Fusion & Target Tracking Conference: Algorithms & Applications, London, UK, 16–17 May 2012; pp. 1–5.
19. Pallotta, G.; Vespe, M.; Bryan, K. Vessel Pattern Knowledge Discovery from AIS Data: A Framework for Anomaly Detection and Route Prediction. *Entropy* **2013**, *15*, 2288–2315. [CrossRef]
20. Liu, B.; de Souza, E.N.; Matwin, S.; Sydow, M. Knowledge-based clustering of ship trajectories using density-based approach. In Proceedings of the IEEE International Conference on Big Data, Washington, DC, USA, 27–30 October 2014; pp. 603–608.
21. Lei, P.R. A Framework for Anomaly Detection in Maritime Trajectory Behavior. *Knowl. Inf. Syst.* **2016**, *47*, 189–214. [CrossRef]
22. Zhen, R.; Jin, Y.; Hu, Q.; Shao, Z.; Nikitakos, N. Maritime Anomaly Detection within Coastal Waters Based on Vessel Trajectory Clustering and Naïve Bayes Classifier. *J. Navig.* **2017**, *70*, 648–670. [CrossRef]
23. Han, J. *Data Mining: Concepts and Techniques*; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 2005; pp. 323–333.
24. Barron, A.; Rissanen, J.; Yu, B. The Minimum Description Length Principle in Coding and Modeling. *IEEE Trans. Inf. Theory* **1998**, *44*, 2743–2760. [CrossRef]
25. Yuan, G.; Xia, S.X.; Zhang, L.; Zhou, Y. Trajectory Clustering Algorithm Based on Structural Similarity. *J. Commun.* **2011**, *32*, 103–110. (In Chinese)
26. Ships' Routeing. Available online: http://www.imo.org/en/OurWork/Safety/Navigation/Pages/ShipsRouteing.aspx (accessed on 30 April 2018).

27. Campello, R.J.G.B.; Moulavi, D.; Sander, J. Density-based clustering based on hierarchical density estimates. In *PAKDD 2013: Advances in Knowledge Discovery and Data Mining*; Springer: Berlin/Heidelberg, Germany, 2013; Volume 7819, pp. 160–172.
28. Campello, R.J.G.B.; Moulavi, D.; Zimek, A.; Sander, J. Hierarchical Density Estimates for Data Clustering, Visualization, and Outlier Detection. *ACM Trans. Knowl. Discov. Data* **2015**, *10*, 1–51. [CrossRef]
29. McInnes, L.; Healy, J.; Astels, S. HDBSCAN: Hierarchical Density Based Clustering. *J. Open Source Softw.* **2017**, *2*, 205. [CrossRef]