Research on the clustering of marine traffic status based on AIS data using DBSCAN

Master's Thesis

RESEARCH ON THE CLUSTERING OF MARINE TRAFFIC STATUS BASED ON AIS DATA USING DBSCAN

September 2022

Graduate School of Marine Science and Technology

Tokyo University of Marine Science and Technology

Master's Course of Marine Technology and Logistics

Wang Xiya

Master's Thesis

RESEARCH ON THE CLUSTERING OF MARINE TRAFFIC STATUS BASED ON AIS DATA USING DBSCAN

September 2022

Graduate School of Marine Science and Technology

Tokyo University of Marine Science and Technology

Master's Course of Marine Technology and Logistics

Wang Xiya

Contents

Abstract

Maritime traffic is increasingly crowded due to the continuous development of economy and technology, and the development brings opportunities as well as challenges. The heavier maritime traffic and more complex ship operations all lead to increased maritime traffic risks. Using spatial and temporal data of ship navigation, we can extract and analyze the characteristics of maritime traffic flow by computer data mining technology and cluster analysis, which can help us understand the spatial and temporal distribution of traffic flow in the target sea area. This provides a reference basis for traffic management and the formulation of relevant policies and regulations. It also provides guidance for ship route planning and reduces the labor cost of maritime navigation.

Based on AIS and data clustering, the paper focused on the macro features of maritime traffic flow, maritime traffic flow division and maritime traffic spatial and temporal patterns. Based on this research, the following work has been done on the marine traffic situation:

The AIS data of vessels (mainly energy vessels) in the first half of 2016 worldwide were screened and processed, and the Arabian sea area was selected for the study to analyze the data quality. We also analyze the basic situation of traffic within this sea area, including the composition ratio of ship types regarding space and time, the spatial composition of destination, the distribution of SOG regarding geographic space and time, the spatial distribution of COG, the distribution of draught regarding geographic space and time, and the relationship between draught and destination and COG respectively.

Based on the DBSCAN algorithm, we included longitude, dimension, SOG and COG together in the calculation of distance, and added the attribute constraints of SOG and COG to perform constrained clustering analysis on AIS data to classify different maritime traffic flows and analyze the macroscopic characteristics of maritime traffic flows. The results proved that improved DBSCAN could cluster the AIS data of ships in the sea well and get the realistic traffic flow. It was shown that the distribution of traffic flow was influenced by the weather environment. The average SOG of energy tankers during April was significantly greater than that of June, with an average of 2.055kn higher. And the draught was larger in June than in April in overall.

The ST-DBSCAN algorithm is used to cluster the spatial and temporal dimensions of AIS data by introducing the temporal dimension based on the spatial distance. By this method, the busy and congested level of watershed can be judged simultaneously in space and time. The study found that the offshore areas within the Arabian Sea and the area leading to the Gulf of Aden had a high density of vessels and were prone to vessel encounters. Vessel encounters occurred in Arabian waters at night more often and less often during the day. There was a decreasing trend from 0:00 to 12:00 and an increasing trend from 12:00 to 24:00. And there is a higher chance of more than two ship encounters in the Indian offshore waters.

Keywords: Automatic Identification System, Characteristics of Marine Traffic, DBSCAN, ST-DBSCAN, Arabian Sea

1. Introduction

1.1 Research Background

Maritime transport is developing rapidly due to its advantages of large freight volume, low cost, low energy consumption, and a small investment. Globalization of world trade also provides a strong driving force for the development of maritime transport. From Figure 1.1, we can find that driven by macroeconomic growth, the global seaborne trade volume grew year by year from 2009 to 2019; in 2020, affected by the covid-19 epidemic, the global seaborne trade volume decreased by about 4% compared with 2019. And in 2021, the international trade demand gradually recovers to the 2019 level.

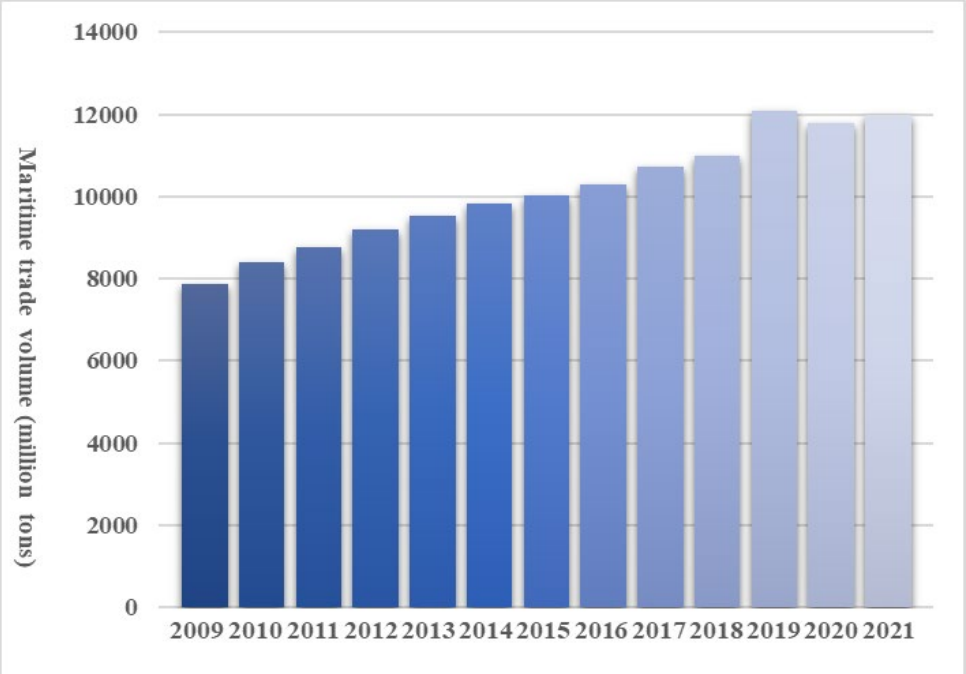


Figure 1.1 Global Seaborne Trade Volume

Data source：UNCTAD

The rapid development of the shipping industry has promoted the innovation and progress of shipbuilding technology, and the trend of vessel specialization, enlargement, high speed, and intelligence is remarkable. At the same time, the number, type, scale, and speed of vessels are also changing. While these developments bring opportunities, they also lead to the higher complexity of maritime traffic flow. The heavy marine traffic and complex ship operation significantly increase the risk of maritime traffic and bring challenges to marine traffic. The contradiction between maritime traffic development and navigation safety and navigation order is becoming more and more prominent. Marine traffic accidents not only bring a great threat to human life and property but also cause serious pollution to the ecological environment, and its loss is incalculable. How to balance navigational safety and navigational efficiency and maximize the use of navigable water resources is a crucial issue for the sustainable development of maritime traffic and is also the core issue for maritime transportation to dovetail with national strategies and achieve sustainable development of the transportation industry.

With the advancement of technology, the application of technologies such as Automatic Identification System (AIS) provides the possibility to solve the contradiction between the development of the shipping industry and safety orders and further improve the level of maritime traffic management. Through the establishment of shore-based data exchange network, the navigation information of ships is accepted and recorded around the clock, forming a global database of ship movement trajectories. Using

advanced data mining technology and data analysis technology provides a powerful tool for mining, analysis, and knowledge discovery of massive trajectory data based on spatial-temporal information. It provides effective technical support for water activities and maritime traffic management.

Unlike land traffic operations, there is no fixed road for maritime traffic due to external and internal factors. Relying on the massive AIS trajectory data accumulated over a period of time to analyze the traffic flow distribution characteristics of ships and extract the effective route. It can not only discern whether the ship navigation follows the marine route traffic rules but also provide the basis for analyzing the behavior of ships at sea, abnormality detection, and marine supervision, as well as provide theoretical positioning reference in the arrangement of marine engineering facilities and marine area planning layout.

1.2 Arabian Sea

The Arabian Sea is a part of the Indian Ocean, located between the Arabian Peninsula, the Indian Peninsula, and the African continent. It is connected to the Persian Gulf in the north through the Strait of Hormuz and to the Red Sea in the northwest through the Strait of Bab el-Mandeb. With an area (including the Gulf of Aden and the Gulf of Oman) of 3.86 million square kilometers, it is the second-largest sea in the world. Its average depth is 2734 meters, and the maximum depth is also 5203 meters of Arabian Sea resources. The Arabian Sea is rich in natural resources. The coastal shelf is rich in oil and gas resources. Seafood, such as pearls, fish, sardines, halibut, and tuna, are also abundant. The location of the Arabian Sea is crucial. It is the outlet for external oil transportation from the Persian Gulf and the shipping waterway connecting the Indian Ocean, the Persian Gulf, the Mediterranean Sea, and the Atlantic Ocean. Giant oil tankers traveling to and from the Persian Gulf must pass through the Arabian Sea, which is also the world's most immense marine oil transportation sea. Along its coast, there are famous ports such as Mumbai, Karachi, and Aden Djibouti. A map of the Arabian Sea was shown in Figure 1.2.

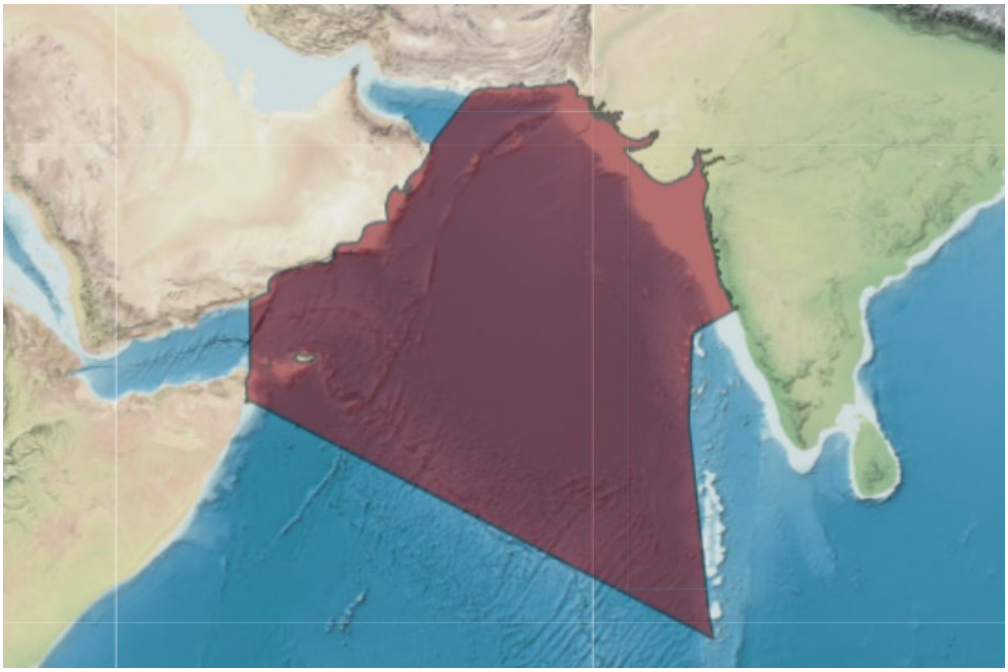


Figure 1.2 Arabian Sea

The Arabian Sea is located in the tropical and subtropical regions, with high topography in the north, and the monsoon features are obvious due to the thermal differences between land and sea. Since it is controlled by the blast circulation, the direction of ocean currents has noticeable seasonal changes during the year, roughly clockwise in summer and counterclockwise in winter. From November to April

every year, the northeast monsoon period in the North Indian Ocean is not very windy, generally, 3-4 levels, which is called the "golden season of the North Indian Ocean" in navigation. The southwest monsoon prevails in summer, starting in late May, from the southeast of the Bay of Bengal, gradually extending northwest, and the wind gradually increases. July and August, the monsoon enter its peak period, with an average wind speed of 7, east of Somalia and Socotra Island near the surface of 8, 9 gale days up to 50%. This peculiarity makes the shipping in the Arabian Sea show prominent seasonal characteristics.

The North Indian Ocean is characterized by complex currents, prevalent monsoons, large wind currents and currents, severe weather conditions in summer, and so on. Moreover, the sea area is large, and the sailing time is extended. To reduce navigation safety accidents, improve navigation economy and ensure the safety of ship navigation, selecting routes and avoiding possible collisions between ships are essential.

1.3 Research Status

1.3.1 Maritime Traffic Status

The study of maritime traffic flow draws mainly on the research methods and results of road traffic flow. A ship traveling in one direction along a route or channel is called a ship flow or water traffic flow. Due to the considerable variation of ship types, large scale span, no clear physical boundaries and traffic signs, and a great degree of freedom and inertia in navigation, the maritime traffic flow is more complex than the road traffic flow. The planning of a ship's navigational route depends mainly on the experience level of the crew. Through route mining research, we can provide maritime traffic supervisory departments with the near real-time maritime traffic situation and also understand the spatial and temporal distribution of traffic flow in the target sea area by studying historical data so as to provide a reference basis for traffic management and the formulation of relevant policies and regulations, as well as provide guidance for ship route planning and reduce the labor cost of maritime navigation. In the study of maritime traffic flow models, experts and scholars have proposed numerous parameters to characterize the traffic flow state, mainly macro-parameters and micro-parameters. Macro parameters include traffic flow(Lv, Zhuang, & Li, 2017), speed, density, width, and position. And micro parameters include vessel domain (Goodwin & M, 1975) and vessel spacing(Zhu & Zhang, 2009) .

For traffic flow locations, the location of traffic flows has been commonly inferred in recent years using AIS data trajectory clustering methods.Dobrkovic (Dobrkovic, Iacob, & Hillegersberg, 2016) used a Genetic Algorithm (GA) to cluster vessels position to get waypoints, and by connecting them get nodes and edges of a directed graph depicting. Ristic (Ristic, Scala, Morelande, & Gordon, 2008) used an adaptive kernel density algorithm to detect and warn of abnormal ships based on statistical analysis models to discover ship movement patterns from large amounts of AIS data and compare previously discovered ship behavior patterns with navigational information of incoming ships. vespe (Vespe, Visentini, Bryan, & Braca, 2012) explored ship navigation patterns and predicts the location of ships through unsupervised self-learning mining algorithms. Among these, spatial-based cluster analysis techniques are the most prevalent, with most studies using computational methods such as Minkowski distances, Euclidean distances, Manhattan distances, and Hausdor distances (Mou, Chen, & He, 2018) to measure the similarity of trajectories, and then using density-based spatial clustering algorithm DBSCAN(Birant & Kut, 2007) , Douglas-Peucker algorithm(Mou et al., 2018) , and augmented genetic algorithm (Dobrkovic et al., 2016) were then used to develop the clustering analysis.

1.3.2 Multi-vessel Encounter Situational Recognition

A ship encounter is a special situation in the spatial and temporal distribution of maritime traffic, where two or more ships are within the range of mutual influence simultaneously. Vessel encounter

information can reflect the behavioral patterns of vessels, and the time and place of encounter, encounter situation, encounter rate, and congestion can be explored to assess the risk of vessel collision.

Vessel encounters are a hot topic of research in maritime vessel traffic and are an important basis and prerequisite for determining a vessel's duty of avoidance. (Huang & M, 2020)。In the interchange waters with high traffic density and complex situations, if the ship encounter situation can be accurately identified, it will have important practical significance and application value for traffic safety supervision and reduction of ship collision accident rate in the interchange waters. (He et al., 2017)

The ship encounter based on AIS data refers to the AIS data of a certain period and range, and the search of the two ships' distance meeting specific requirements is considered as an encounter by using suitable algorithms, without considering whether the two ships take actions and their action effects. In addition to the traditional encounters in nautical science, such as chase, encounter, and cross encounter, it also covers the situation where multiple ships are in close parallel or close to other ships although they have already passed the give way. The research on ship encounter based on AIS data mainly revolves around three aspects: one is to analyze and determine the ship encounter situation from the perspective of the ship domain. Secondly, it focuses on computer technology, and the ship encounter status is found by multiple comparisons of temporary database tables or computer algorithms traversing the database (Pan, Jiang, & Shao, 2010). Thirdly, it emphasizes distributed decision making, where multiple ships in an encounter judge the encounter situation individually and preferably choose a breakthrough path (D, K, & OkimotoT, 2017).

When ships approach each other to cause an encounter, there are generally three encounter situations: chase over, encounter, and cross encounter (also can be divided into two cases of large angle and slight angle). Some scholars have further studied and divided the encounter into simple encounters and complex encounters. If a ship is only in a two-ship encounter situation, it is called a simple encounter ship (S). If a ship is in a multi-ship encounter situation, it is called a complex encounter ship (M). If one ship is a simple encounter ship and the other ship is a complex encounter ship, it is called (S-M). Moreover, if both ships are simple to encounter ships or complex encounter ships, it is called (S-S) or (M-M). When the total encounter rate of two seas is close to each other, the higher the proportion of complex encounters, the greater the danger, so it is practical to refine the encounter type and count them separately.

1.4 Research Framework

The main research objectives of this paper were to analyze the relationship between traffic flow parameters; optimize the DBSCAN algorithm by introducing two constraint attributes, SOG and COG, to divide traffic flows and perform macroscopic analysis of different traffic flows; and use the ST-DBSCAN algorithm to introduce the time dimension of AIS to analyze the spatial-temporal information of vessels. Analyzing the macro traffic situation of ships in the sea area is conducive to improving marine traffic planning and management.

Figure 1.3 presented the Technology Roadmap. The specific steps of the paper content were introduced as follows:

The chapter 1 introduced the research background of the thesis, the significance of the study, and a review of the research literature in the related fields. The framework of the thesis was also described.

Chapter 2 introduced the basic theory and research methods. The basic information of AIS data and the related theories of ship traffic flow were outlined. The basic concepts of classification algorithms were discussed, and the concept, algorithm description and algorithm flow of the DBSCAN algorithm used in this paper were introduced in detail.

Chapter 3 selected the AIS data of Arabian Sea as the sample and discussed the basic situation of

AIS data and traffic in the sea. Then the distribution of ship COG, SOG and draft in time and space were analyzed.

Chapter 4 points out the limitations of the basic DBSCAN algorithm in the study of maritime traffic. Based on the geographical distance of vessels, two constraint attributes, SOG and COG, are introduced to optimize the DBSCAN algorithm. Using AIS data, different traffic flows are classified and macroscopic analysis of different traffic flows is carried out.

Chapter 5 used the ST-DBSCAN algorithm to analyze the spatial-temporal co-occurrence characteristics of vessels by introducing the temporal dimension of AIS based on the geographical distance of vessels.

Chapter 6 summarizes the main research work of this paper, and points out the limitations and shortcomings.

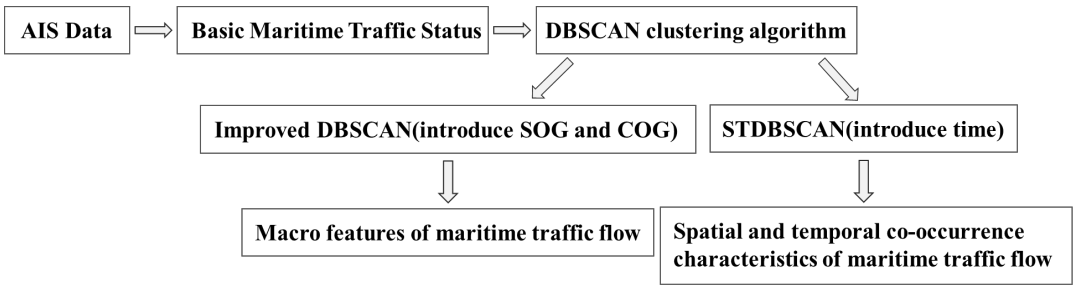


Figure 1.3 Technology Roadmap

2. Research Methodology

2.1 AIS (Automatic Identification System)

AIS is a shipboard broadcast answering system. A ship continuously sends its identity, position, heading, speed, and other data to nearby ships and shore-based authorities over a VHF public radio channel through this system. Shipboard AIS equipment made the exchange of information between ships increasingly smooth. Later, AIS shore-based network systems emerged for monitoring purposes to collect information from all AIS stations along the coast and to realize real-time forwarding of AIS data and historical data retrieval. According to SOLAS Convention: all ships of 300 GT and above engaged in international voyages, cargo ships of 500 GT and above not engaged in international voyages, and passenger ships of any size should be equipped with AIS as required. A ship equipped with AIS will record a large amount of data during the voyage.

AIS data belongs to trajectory data, which meets the characteristics of big data such as large volume, real-time, and variety. Affected by the equipment standard, sampling frequency, transmission effect, and storage method, AIS data has the following characteristics：

a) Spatial-temporal seriality

AIS data is a sampling sequence with location, time, and other information, containing the object's spatial-temporal dynamics.

b) Reporting heterogeneity

As the reporting frequency of AIS data is related to the change of ship speed and heading, the reporting interval varies significantly, which increases the difficulty of trajectory data analysis.

c) Data quality is not guaranteed

The accuracy of AIS data depends on the accuracy of the ship positioning system, the correct input of the information by crew members, the encoding and transmission of information by the sender, and the reception and decoding of information by the receiver. In addition, the self-checking and error correction function of AIS data is weak. These system design and human factors can lead to incorrect received AIS data.

As shown in Table 2.1, there are two main contents of AIS complete, which are divided into two categories: static information and dynamic information. It contains a variety of information such as the position, speed, and heading of the surrounding ships and the operation status of the ships. Therefore, when the AIS system is connected to the electronic chart, it can display the detailed situation of the surrounding ships, thus making up for the defects of the blind radar area and the unfavorable terrain in bad weather, which plays a tremendous complementary role to the safe navigation of the ships.

Table2.1 AIS Information

	Information Name	**Sample**	**Unit**
Dynamic Information	MMSI	636016273	/
	Destination	Arab Emirates	/
	Longitude	106.4822	Decimal degree
	Latitude	6.735367	Decimal degree
	Sog	9.2	Knot
	Cog	92	degree
	Rot	0	degree
	Heading	87	degree
	Base Datatime	2016/1/1 11:34	/

	Navigation Status	Under Way Using Engine	/

Static Information	IMO	9332822	/
	Vessel Name	TAURUS SUN	/
	Callsign	D5FI4	/
	Vessel Type	Tanker	/
	Vessel Class	A	/
	Length	243	Meter
	Width	43	Meter
	Flag Country	Liberia	/
	Flag Code	636	/
	Draught	12.6	Meter
	Vessel type	Crude Oil Tanker	/

2.2 Clustering Algorithms

2.2.1 Unsupervised Clustering Algorithms

Clustering is an unsupervised classification method(Zhou, 2016), which is the process of dividing data into groups that are as similar as possible for a given set of data objects(Bushra & Yi, 2021). Partitioning a data set into different classes or clusters according to some specific criterion (such as distance) makes the similarity of data objects within the same cluster as large as possible and the difference of data objects in different clusters as large as possible.

The current popular clustering algorithms can be categorized into five main types(Han & Kamber, 2001): Partitional methods(L.Huntley & Brown, 1996), Hierarchical methods(Cerri & de Carvalho, 2011), Density-based methods(Wang, Wu, Zhang, & Lu, 2019), Grid-based methods(Lee & Cho, 2016) ,and Model-based methods. Density-based methods is defined as the region with higher density than the rest of the data set as clusters, and objects in sparse regions are usually considered as noise points and boundary points(Kriegel, 2011). As shown in Figure 2.1, DBSCAN is a more representative density-based clustering algorithm, which defines a cluster as the most extensive set of densely connected points.

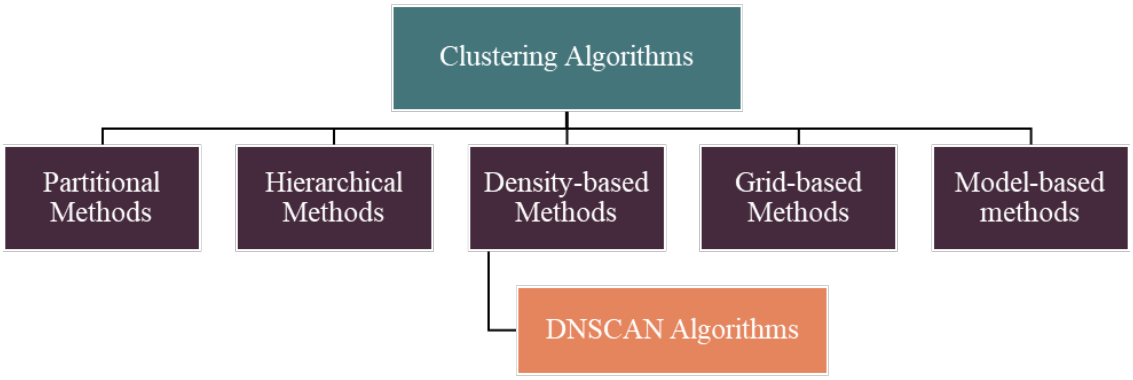


Figure 2.1 Clustering Algorithms

The similarity measure is the basis for the execution of clustering algorithms. All clustering algorithms are based on the similarity between data points for data relationship discrimination and then specific clustering algorithms for partitioning strategies to achieve the clustering requirements of high

intra-cluster similarity and low inter-cluster similarity. The choice of similarity measure directly affects the clustering results of the data. The most commonly used similarity measure for clustering is to calculate the distance between data objects, and the object distance is inversely proportional to the similarity. The smaller the distance value, the higher the similarity between data objects, and the more significant the distance value, the lower the similarity.

The distance between objects can be divided into Minkowski distance, cosine distance and Marxian distance. Among them, the more commonly used is the Minkowski distance, as shown in the equation (2-1).

$$dist(x,y) = (\sum_{i=1}^{n}|x_i - y_j|^p)^{\frac{1}{p}} \quad (2\text{-}1)$$

When p = 1, it's Manhattan distance formula; when p = 2, it's Euclidean distance; when $p \to +\infty$ it is Chebyshev distance.

2.2.2 DBSCAN

The DBSCAN (Density-based spatial clustering of applications with noise) algorithm was proposed by Martin Ester, Hans Peter Kriegel et al. in 1996(Ester, Kriegel, Sander, & Xu, 1996). Its advantages are mainly four. The first one is that the number of clusters need not be determined in advance, and the data clusters are identified based on the density reachable relationship between the core points, which overcomes the drawbacks of the division-based clustering K-means algorithm. The second is that it is insensitive to the shape of clusters, and the shape of clustering clusters can be various. The third is the ability to identify noisy points. Fourth, it is insensitive to the order of samples in the database.

1） Concepts of DBSCAN

DBSCAN is used to analyze the magnitude of the data sample distribution by two parameters (Eps, MinPts).

(Definition I) Eps: the radius of the neighborhood of p to determine the neighborhood of p as the hypersphere region with radius Eps.

(Definition II) MinPts: the minimum neighbor point support of p to determine whether p is the core point , the necessary condition for p to be a core point is to have at least M$inPts$ neighboring points within the neighborhood of p.

Suppose $N_{eps}^{D}(p)$ is the set of neighbor points of p under the data set D with Eps as the domain radius; and $\left|N_{eps}^{D}(p)\right|$ is the number of neighbor points of p.

(Definition III) Core Point (CP): If $\left|N_{eps}^{D}(p)\right| \geq MinPts$, p is a core point.

(Definition IV) Border Point (BP) If $\left|N_{eps}^{D}(p)\right| < MinPts$, $\exists CP \in N_{eps}^{D}(p)$, p is a border point.

(Definition V) Noise Point (NP): If $\left|N_{eps}^{D}(p)\right| < MinPts$, $\nexists CP \in N_{eps}^{D}(p)$, p is a noise point.

The decision tree for the division of various points can be plotted as Figure 2.2.

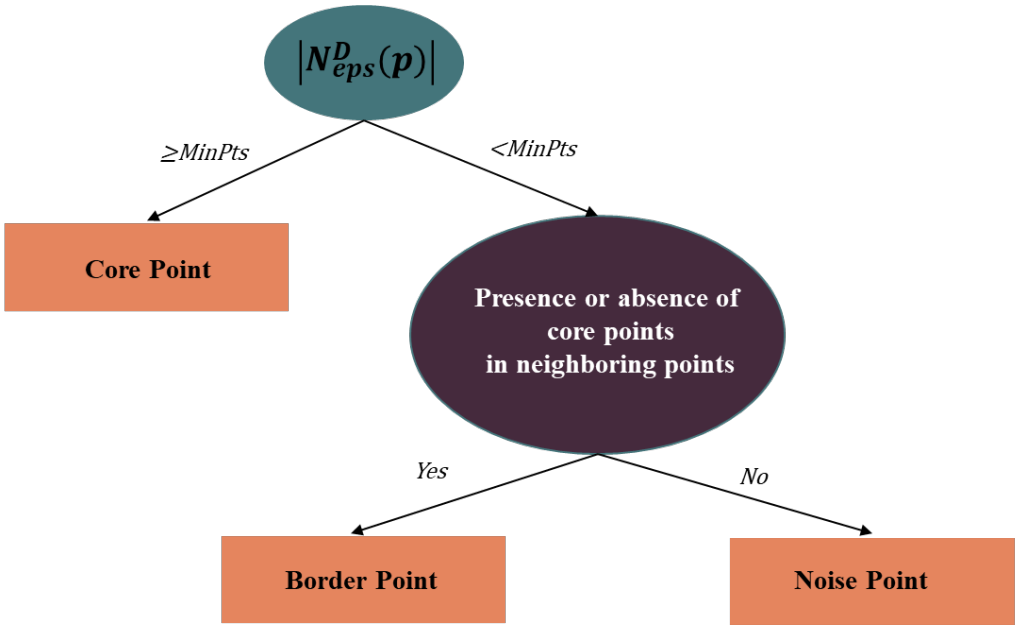


Figure 2.2 The Decision Tree for the Division

(Definition VI) Directly density reachability: assuming q is the core point and $P \in N^D_{eps}(p)$, it is considered that it is direct density reachable from q to p.

(Definition VII) Density reachability: If q is associated with p through a direct density reachable relationship of at least one data point, then q is considered to be density reachable to p.

(Definition VIII) Density connectivity: If p and q densities are reachable to the same data point, then p and q are considered to be density connected.

2）Principle of DBSCAN

The DBSCAN algorithm starts clustering by accessing a random data point p from the dataset. Neighboring points within the hypersphere volume of point p with Eps as the domain radius are retrieved. If the number of neighbor points within the neighborhood of point p (including point p itself) is lower than the minimum neighbor point support MinPts, point p is recorded as a noisy point. Here the p-point is only temporarily marked as a noise point, and subsequently the p-point may be found within the neighborhood of another point and marked as a boundary point, so it is still possible for the p-point to be clustered within a cluster. If the number of p points in the neighborhood is greater than or equal to the minimum neighbor point support MinPts, the p points are considered as core points. At this point, we start to expand outward from point p to find its surrounding neighbor points with direct density reachable in order to form a cluster, and the core points in the surrounding neighbor points also find density reachable points by continuing to expand outward until they expand to all points that can be density reachable, which forms a cluster, and all data points within this cluster are connected to each other with density(Gu et al., 2017).

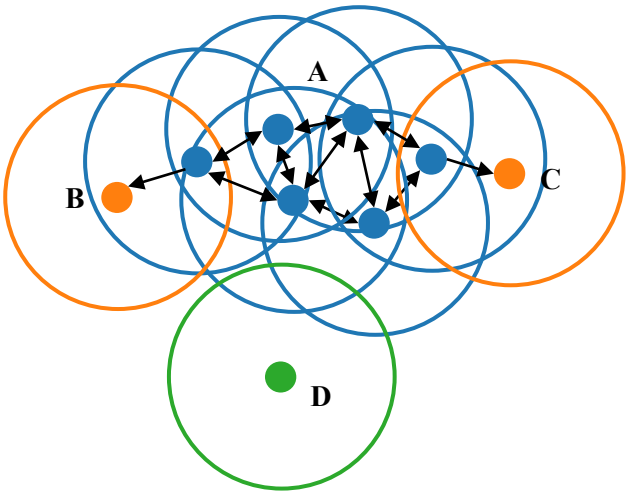


Figure 2.3 DBSCAN Algorithm Clustering Process

Figure 2.3 shows a schematic diagram of the DBSCAN clustering process with MinPts = 4. Since the Eps neighborhood radius area of A and the other red points contains at least 4 points (itself is counted), they are the core points. They are all densely accessible to each other, forming a cluster. Points B and C will initially be judged as noisy points but are reachable from A (through other core points), so they also belong to the cluster where A is located and are modified as boundary points. There is no point that is density reachable with point D, so it is a noise point.

The DBSCAN Algorithm pseudo-code is shown in Table 2.1.

Table 2.1 DBSCAN Algorithm pseudo-code

Input: Datalist, Eps，MinPts
Output：Grouplist
1: Grouplist [index ∈ Datalist] = UNVISITED
2: **for** each UNVISITED point x in X do
3: mark x as VISITED
4: Set N as Eps-neighborhood of x
5: **if** the size of |N| < MinPts
6: mark x as noise point
7: **else**
8: create a new cluster C and add x to C
9: **for** each point y in N
10: **if** y is UNVISITED then
11: mark y as VISITED
12: Set N' as Eps-neighborhood of y
13: **if** the size of N' < MinPts
14: add those point to N
15: **end if**
16: **end if**
16: **if** y does not belong to any cluster, then
17: add y to C
18: **end if**
19: **end for**
20: **end if**
21: **end for**

2.3 Vessel Traffic Flow

Vessel traffic flow refers to the overall movement of continuously running vessels and other transportation vehicles that exhibit specific fluid-like characteristics in waterway transportation. The characteristics of vessel traffic flow mainly include vessel traffic flow, vessel traffic flow structure, vessel density, vessel speed, the spatial and temporal distribution of traffic flow, and other fundamental characteristics. Vessel traffic flow refers to the number of all vessels passing through a certain location in the water within a unit of time. Its size directly reflects the scale and busyness of traffic in the water area and, to a certain extent, the congestion of ships. The structure of vessel traffic flow refers to the ratio of different types of vessels or vessels of different sizes, and the ratio of various vessels and various classes of vessels is counted to assess the navigational safety of the waters. Generally speaking, the more complex the structural characteristics are, the higher the complexity and danger level follows. Vessel density in traffic flow theory refers to the number of vessels per unit area at a particular moment in certain water, reflecting the density of vessels in the water and closely related to the flow and area of vessels in the water. Vessel speed in traffic flow theory refers to the average speed value and distribution range of speed of all ships passing through a certain water area or waterway in a water dynamic. This characteristic distribution reflects the frequency of the vessel chase over the water. The spatial-temporal distribution of traffic flow generally includes the distribution location and direction of traffic flow. The study of spatial-temporal characteristics is of great practical significance for improving the efficiency of marine passage, analyzing the danger of ship collision, and constructing marine functional area assessments.

Traffic flow characteristics are divided into static and dynamic characteristics according to their attributes, as shown in Figure 2.4. The microscopic behavioral characteristics of ship traffic flow are the characteristics and behavioral data collection and analysis of a single ship, which can be divided into static and dynamic characteristics. The microscopic static characteristics of vessel traffic flow contain type, tonnage, and size, which do not change with the change of time and space in a relatively stable period of time and belong to individual essential characteristics; the microscopic dynamic characteristics of ship traffic flow contain ship position, speed, heading, which constantly change at any time and space in the process of ship movement. The macroscopic features of vessel traffic flow are the results of calculating and analyzing the massive ship traffic flow data in certain waters, which can also be divided into static characteristics and dynamic characteristics. The macroscopic characteristics of ship traffic flow are the results of calculating and analyzing a large amount of ship traffic flow data in a particular water area. The macro static characteristics of ship traffic flow are derived from the essential characteristics of a ship and micro static characteristics of traffic flow, forming the structural characteristics of vessel traffic flow. The macro dynamic characteristics of vessel traffic flow are the spatial and temporal distribution characteristics of traffic flow, such as position distribution, speed distribution, and course change distribution, which are expressed by the behavioral characteristics data of all ships in the region within a certain period of time.

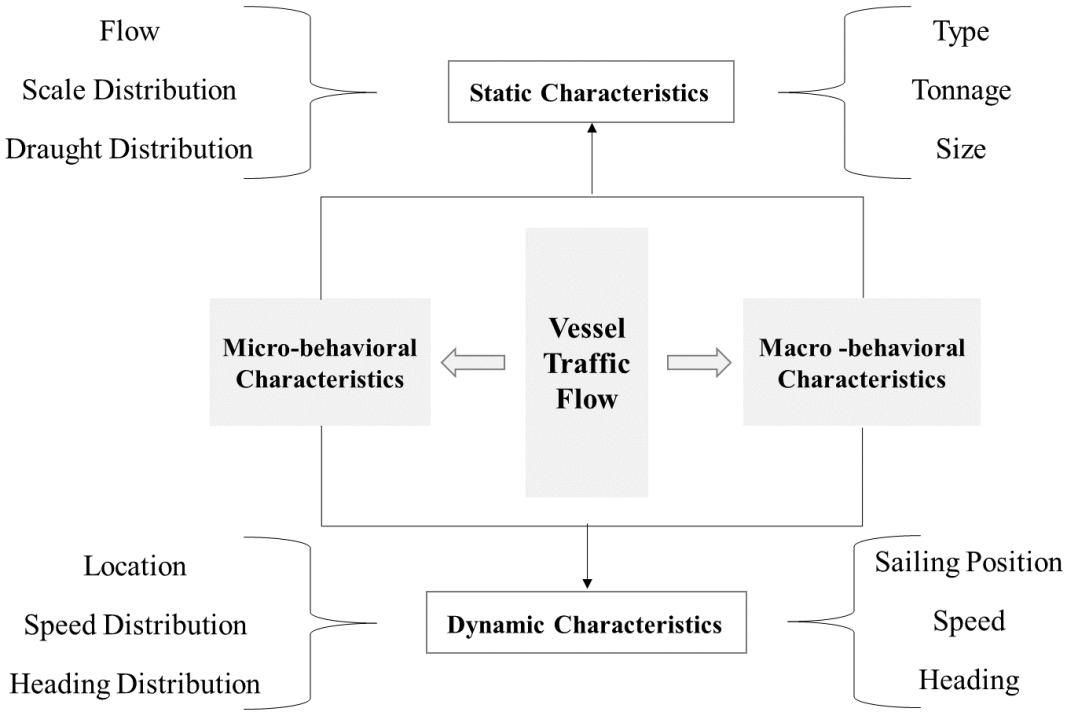


Figure 2.4 Vessel Traffic Flow Characteristics

Vessel traffic flow has multiple attributes, such as duality, limitations, variability and complexity (Liu, 2009)。Duality means that vessel traffic flow is affected by both traffic management and the ability to change the operation of navigating vessels and to adjust vessel status. Constraint refers to the fact that vessel traffic flow is generally subject to the constraints of navigation channel, water environment, meteorological conditions, etc., and ships affect each other. Variability means that the vessel traffic flow in certain waters will change with time, season and sea facilities. Complexity means that the ship is moving at the speed and direction suitable for the prevailing environment selected under the maneuvering

of human, and it is the traffic flow formed by the comprehensive movement under the influence of many factors, which has complex characteristics (L. Zhao, 2013).The characteristics of vessel traffic flow(Meng, Weng, & Li) mainly include traffic flow, traffic flow structure, vessel density, vessel speed, and spatial and temporal distribution of traffic flow. It is essential to master the elements of traffic flow and the change law of traffic flow to improve the efficiency of marine transportation, reasonably implement the marine engineering construction and optimize the marine spatial layout.

3. Marine Traffic Status Assessment

3.1 AIS Data Source Analysis

Vessel AIS data from January to June 2016 provided by exactEarth was used for this study. These data include only crude oil tankers, oil products tankers and LNG tankers on a global wide basis. And the number of messages was about 50 million. It includes vessel name, callsign, Maritime Mobile Service Identity (MMSI), vessel type, vessel type cargo, vessel class, length, width, flag country, destination, estimated Time of Arrival (ETA), draught, longitude, latitude, speed over ground (SOG), course over ground (COG), rate of turn (ROT), course, navigation (NAV) status, source, time, vessel type main and vessel type sub.

We combined 182 small datasets into one large dataset and selected AIS data for the Arabian Sea. According to Figure 3.1, more than 90% of the data had a time interval of less than 1 hour. Moreover, there was approximately 31.24% of the data interval within 1min.

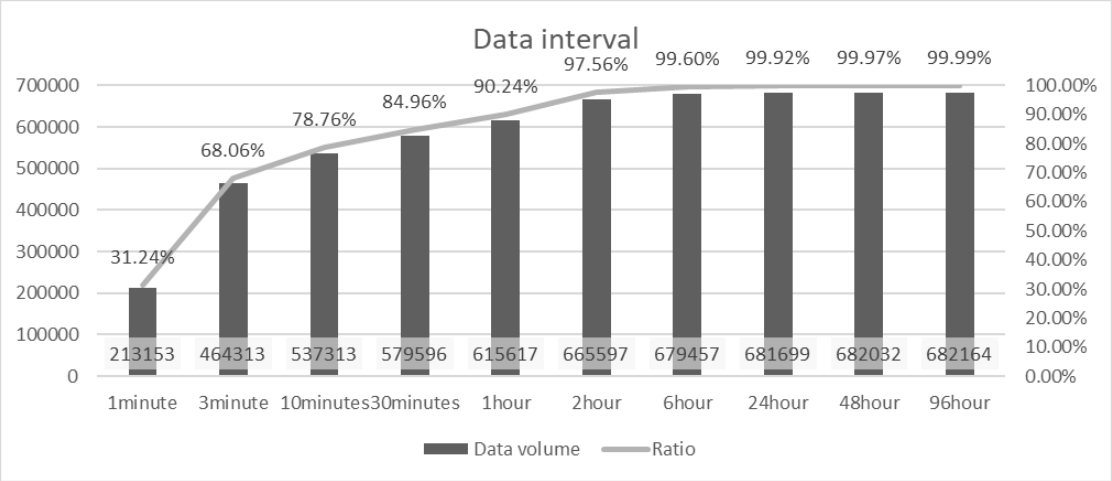


Figure 3.1 Data Interval time

3.2 Traffic Status in Arabian Sea

Global AIS data from January 1, 2016, to June 30, 2016, a total of 51,705,694 data. These data mainly included crude oil tanker, oil products tanker and LNG tanker. We filtered out the data from global data located in the Arabian Sea for a total of 1,212,922. The ship heat geographic distribution map of AIS data in the Arabian Sea was shown in Figure 3.2. In this area, four dense routes formed the main triangular-shaped route. Complex traffic conditions increased the risk of ship collisions.

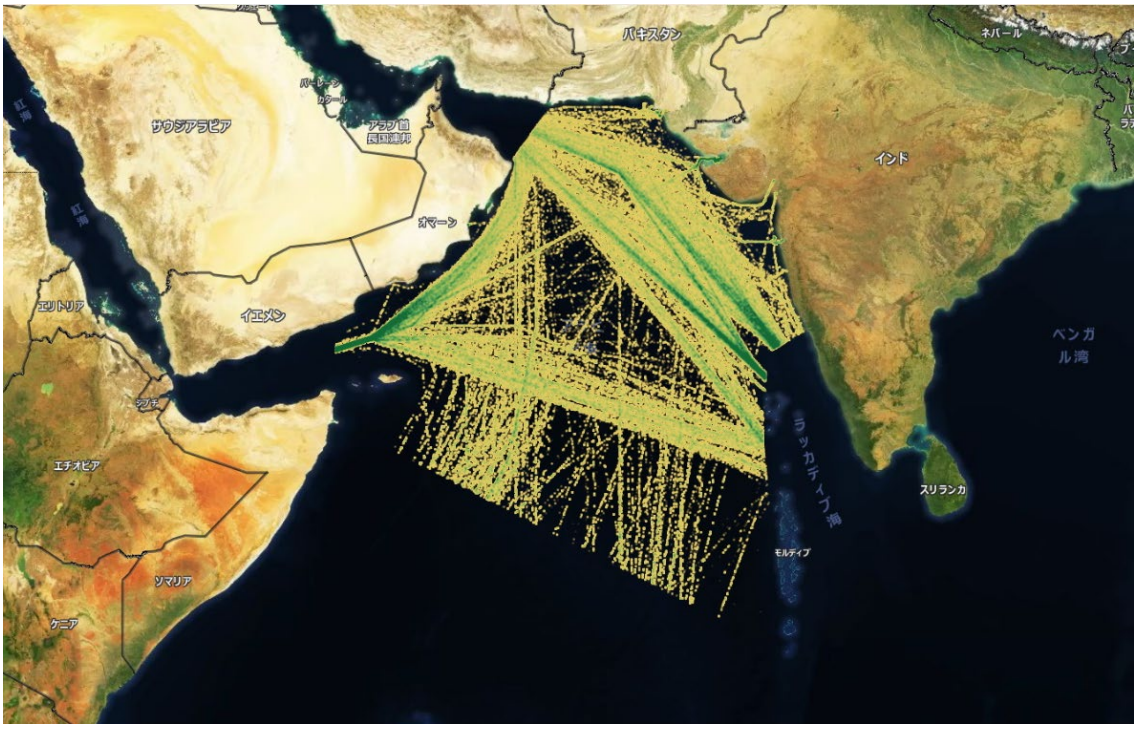


Figure 3.2 Ship Heat Distribution in Arabian Sea on January-June 2016

1）Vessel Type

As vessels generate a large amount of AIS data during their voyage, the amount of AIS data in an area does not represent the number of vessels. So, we counted the number of vessels passing through the Arabian sea areas. As demonstrated in Table 3.1 and Figure 3.3, in the first half of 2016, a total of 3659 vessels passed through the waters. Crude Oil Tankers accounted for the majority, at 73%, while LNG Tankers were the second most numerous, at 625 vessels, accounting for 17% of the total. Oil Products Tankers were only 70 vessels, accounting for 2% of the total.

Table 3.1 Amount of Tanker Types in Arabian Sea

Amount of Tanker	3659
Crude Oil Tanker	2678
Oil Products Tanker	70
LNG Tanker	625
Other Vessels	282

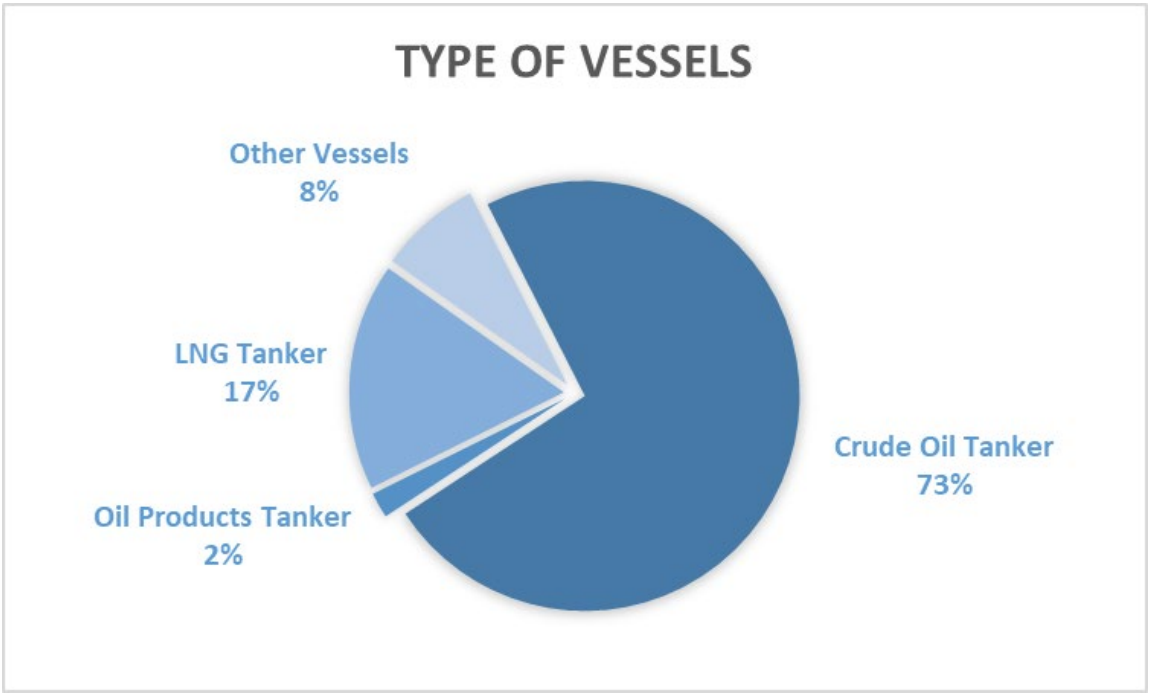


Figure 3.3 Distribution of Tanker Types in Arabian Sea

Figure 3.4 showed the number of ships passing through the Arabian Sea per month in the first half of 2016. Among them, January, March, and June were the busiest months for traffic in this sea, and April had the lowest number of vessels passing. The difference between the number of LNG Tankers and the Oil Products Tankers passing each month was insignificant. The number of Crude Oil Tankers was similar in January and March and similar in February, May, and June, respectively.

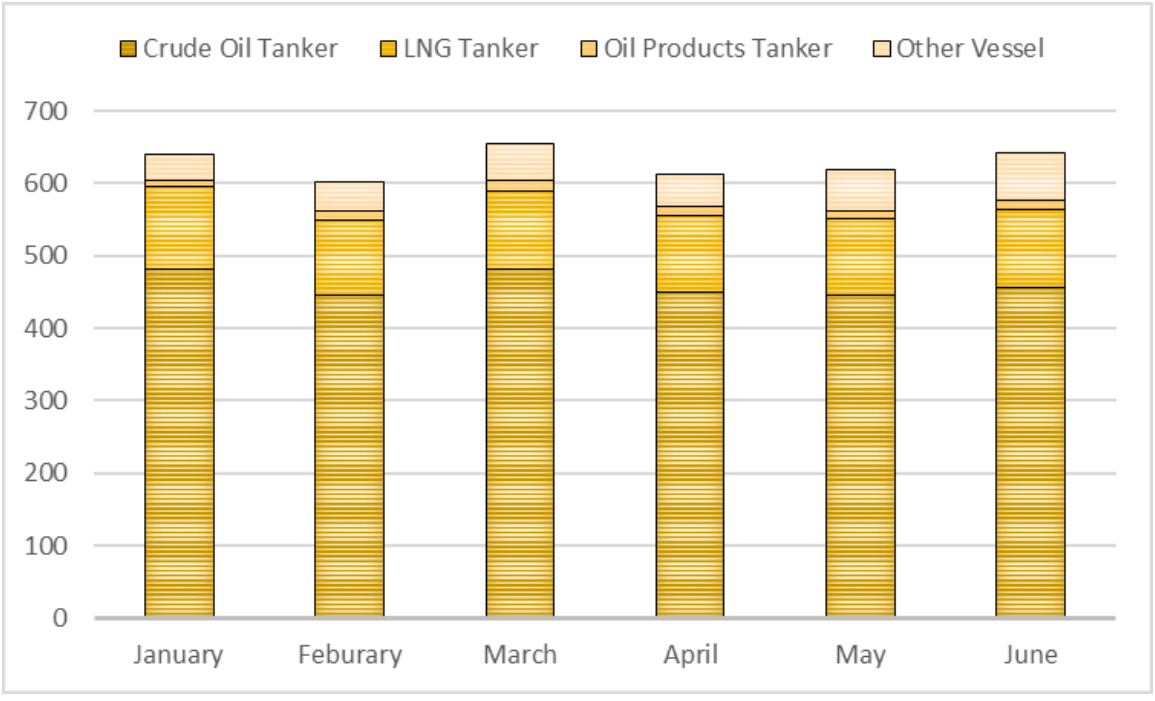


Figure 3.4 Distribution of Vessel Types in the First Half of 2016

2) Destination

With reference to Figure 3.5 and Table 3.2, it could be seen up to 50 destinations for vessels passing through the Arabian Sea in the first half of 2016 and the number of energy vessels destined for Arab Emirates, India, Saudi Arabia, Singapore, Japan and China accounted for half of all data. The most significant number of vessels destined for the Arab Emirates, 904, accounted for 25% of the total. As a major oil-producing country, its crude oil tankers accounted for 25.88% and LNG tankers for 17.28% of the sea area. There were 413 vessels destined for India, of which 303 were crude oil tankers and 63 were LNG tankers, accounting for 11.31% and 10.08%, respectively. The number of vessels in Saudi Arabia and Singapore was 244 and 139, respectively, representing 6% and 4% of the total. There were no LNG tankers going to Saudi Arabia and only two vessels going to Singapore, but there were crude oil tankers 220 and 122, respectively. There were 136 vessels destined for Japan, of which 88 were crude oil tankers and 39 were LNG tankers, accounting for 3.29% and 6.24%, respectively. And there were 135 vessels destined for China, of which 110 were crude oil tankers and 14 were LNG tankers, accounting for 4.11% and 2.24%, respectively.

Table 3.2 Type Distribution of Destinations in Arabian Sea

Destination	Crude Oil Tanker	Percentage	LNG Tanker	Percentage
Arab Emirates	693	25.88%	108	17.28%
India	303	11.31%	63	10.08%
Saudi Arabia	220	8.22%	0	0.00%
Singapore	122	4.56%	2	0.32%
Japan	88	3.29%	39	6.24%
China	110	4.11%	14	2.24%

Figure 3.5 Distribution of Destinations in Arabian Sea

3)SOG

As presented in Figure 3.6, in the Arabian Sea, the speed of ships was generally maintained between 0 and 21kn, with most ships sailing at medium speed (12kn-15kn). Vessel speeds were faster near the Indian Peninsula than near the Arabian Peninsula. We could also identify the approximate areas of anchoring, mooring, and proximity to land-based on the heat of latitude and longitude corresponding to a SOG of 0. Ships destined for India anchor mainly in the ports of the Gulf of Kutch and the port of Mumbai. Ships destined for Pakistan are mainly based in the port of Karachi. In the open sea zone, the density of ships was low, and the speed was relatively high.

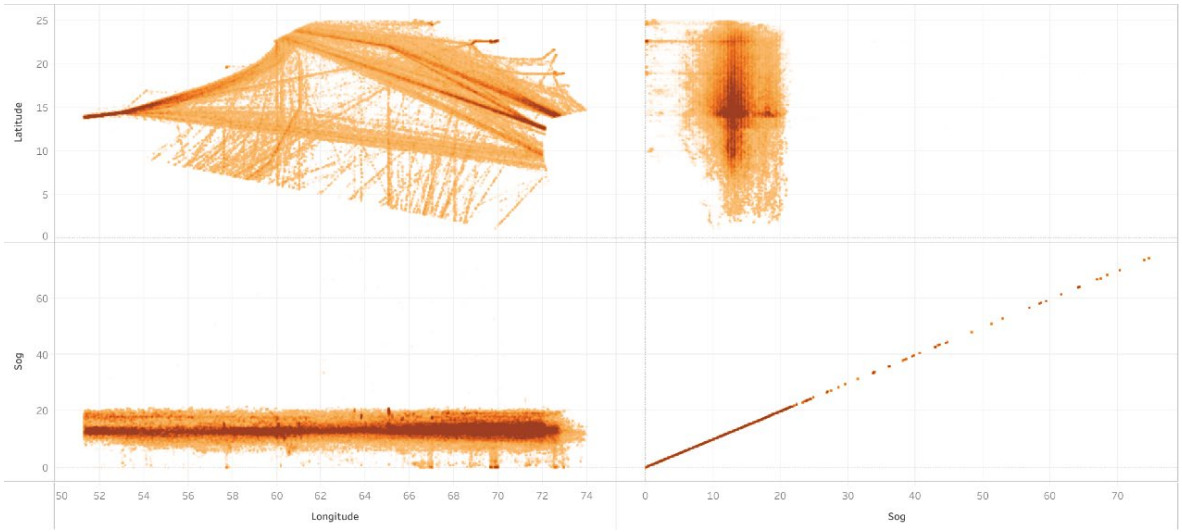


Figure 3.6 SOG distribution on geographic coordinates

4)SOG-Month

The average daily SOG of vessels in the first half of 2016 was shown in Figure 3.7. January to April belong to the northeast monsoon period, which was favorable for ship navigation, so the average daily SOG of vessels was relatively stable, concentrated between 12.5 kn and 14.5 kn. The daily average SOG distribution in May was not much different from the first four months. However, in June, the wind speed in the sea gradually increased, and the sailing conditions gradually became harsh, so the average SOG dropped steeply and showed a continuous decreasing trend, with the range of SOG between 13.644kn and 10.544kn.

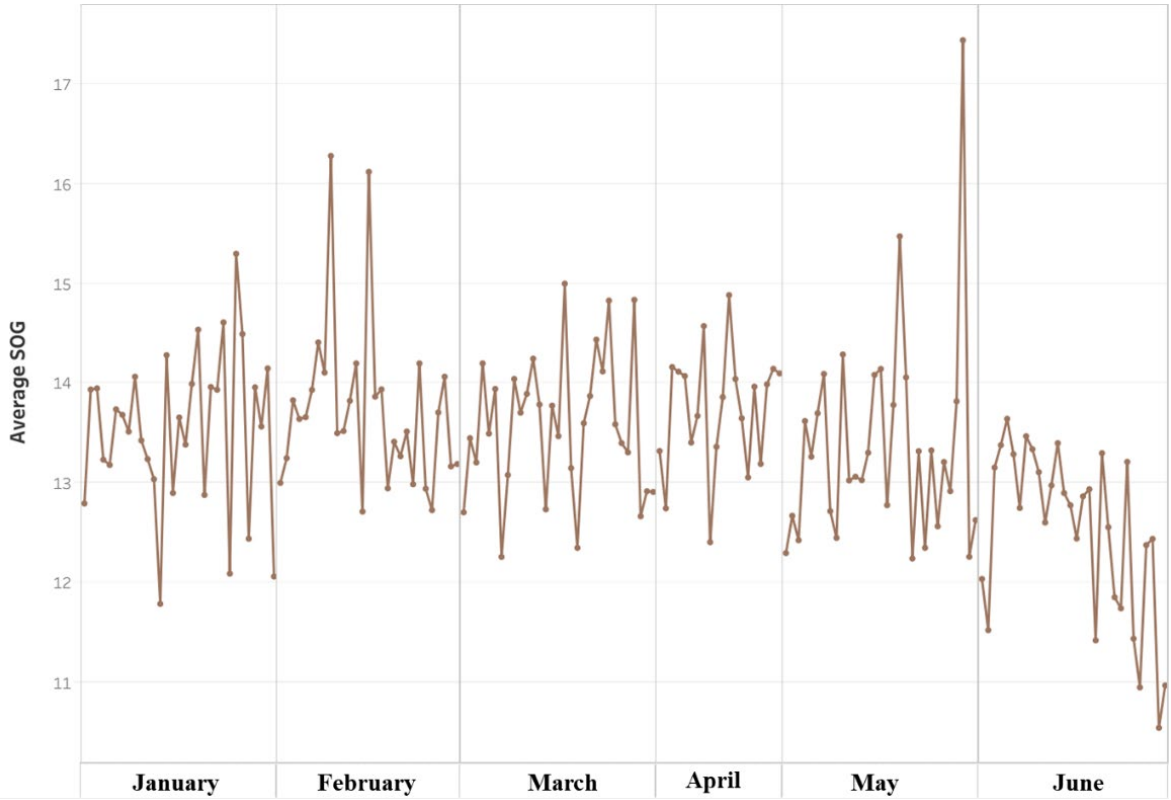


Figure 3.7 SOG distribution on month

5) COG

According to Figure 3.8, it can be seen that the course was distributed between 0°-360°, and COG showed a symmetrical distribution about longitude and latitude, respectively, which was due to the simultaneous movement of ships in both directions on the same route. In the area adjacent to the Gulf of Aden, the COG was concentrated at 70° and 250°. Vessels were heading 70° sail from the Gulf of Aden to the Arabian Sea, while Vessels were heading 250° sail from the Arabian Sea to the Gulf of Aden. In the distant areas of the Arabian Sea, the heading was more complex, spreading from 0-360°. Close to the Indian Peninsula, the ships' heading was concentrated at 110°, 140°, 290°, and 320°.

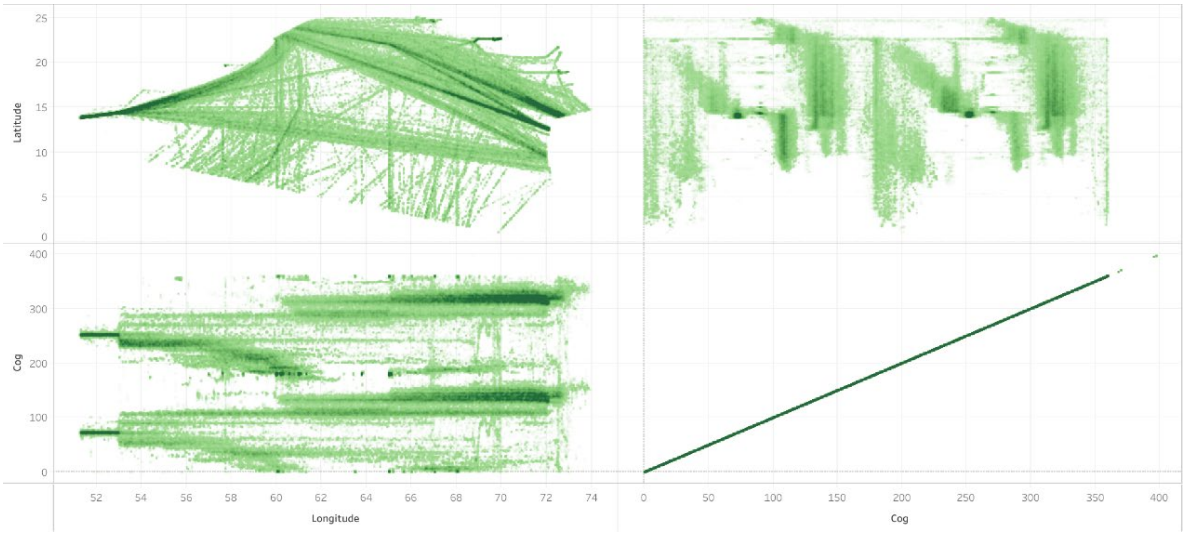


Figure 3.8 COG distribution on geographic coordinates

7) Draught

From Figure 3.9, we could observe that the draught was concentrated above 8m and under 22m. And there was a blank draught record between 17-19 m, which meant that the draught of some ships suddenly drops from above 19m to below 17m at the same position. With a longitude of 60° as the boundary, draughts were relatively greater in area less than 60° longitude than in area bigger than 60° longitude.

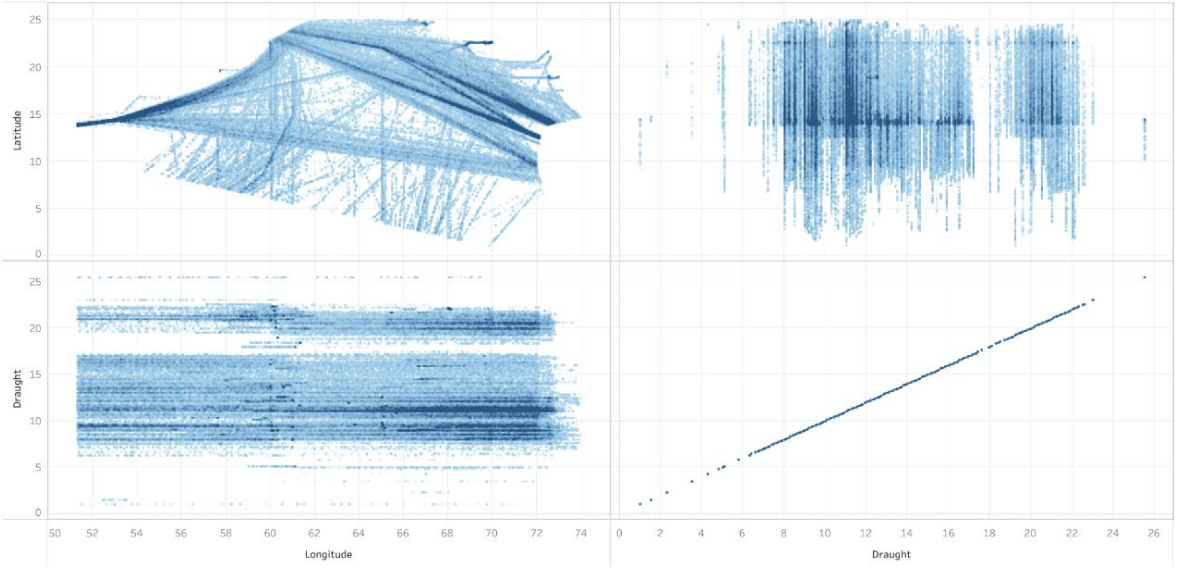


Figure 3.9 Draught distribution on geographic coordinates

8) Draught-Destination

The top 12 destination countries in terms of number of ships passing through the Arabian Sea in the first half of 2016 were selected and a Gantt chart was made in relation to their draught, as in Figure 3.10. Among them, destinations with a median draught of less than 11m included the Arab Emirates, Saudi Arabia, Iraq, Kuwait, Oman, and Qatar, all of which were located on the Arabian Peninsula or Persian Gulf coast and are large oil producing countries. China, Egypt, India, Japan, Korea, and Singapore all have median draught greater than 14m and they were all large oil importers. This was consistent with the fact that unladen tankers came to these major oil producing countries to load crude oil for onward delivery to various countries around the world.

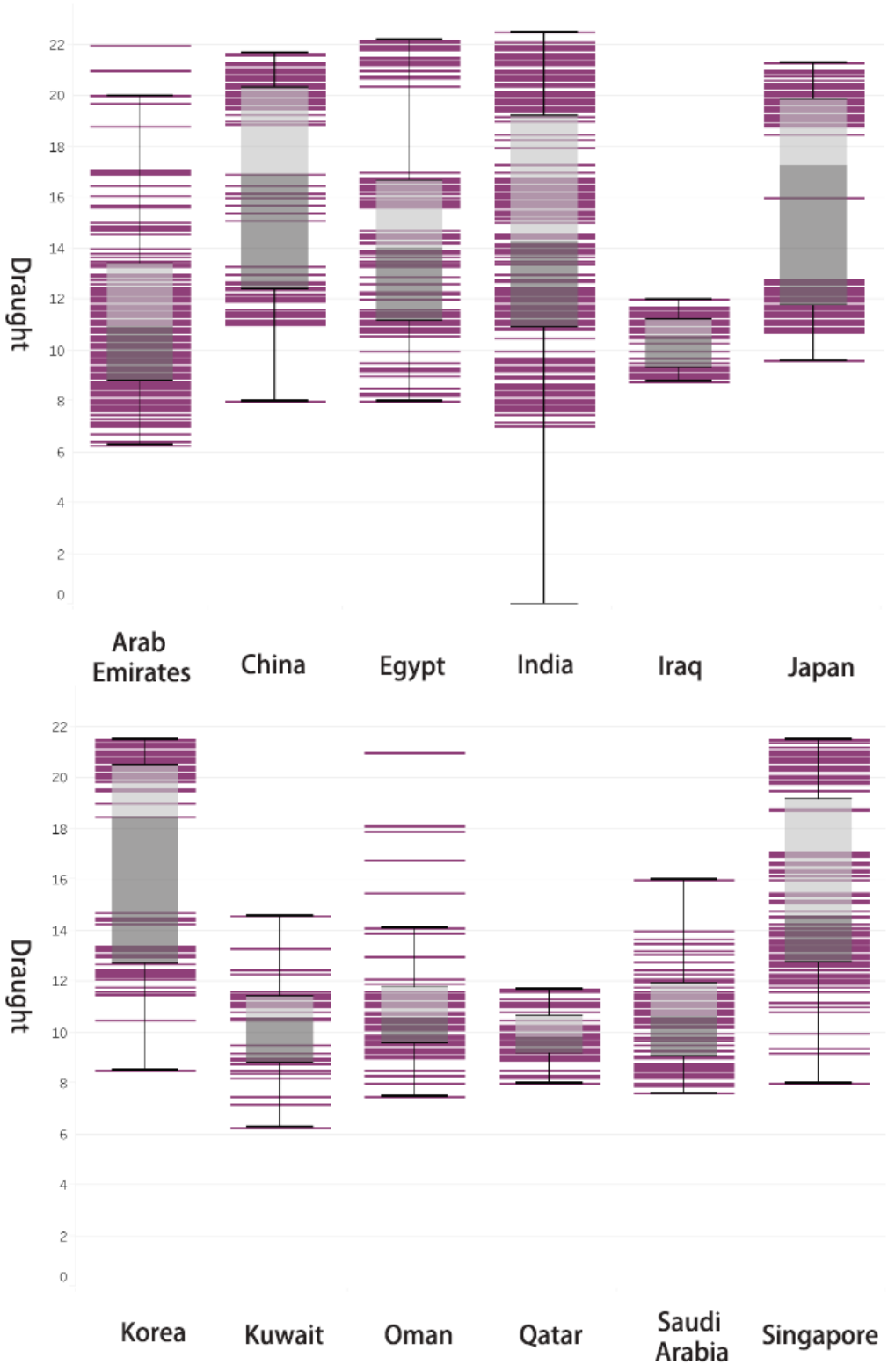


Figure 3.10 Draught distribution on destination

9) Draught-Month

As shown in Figure 3.11, overall the six months saw little difference in vessel draught levels. The median vessel's draught for each month was around 13.5. 75% of the vessels had a draft of less than 18m.

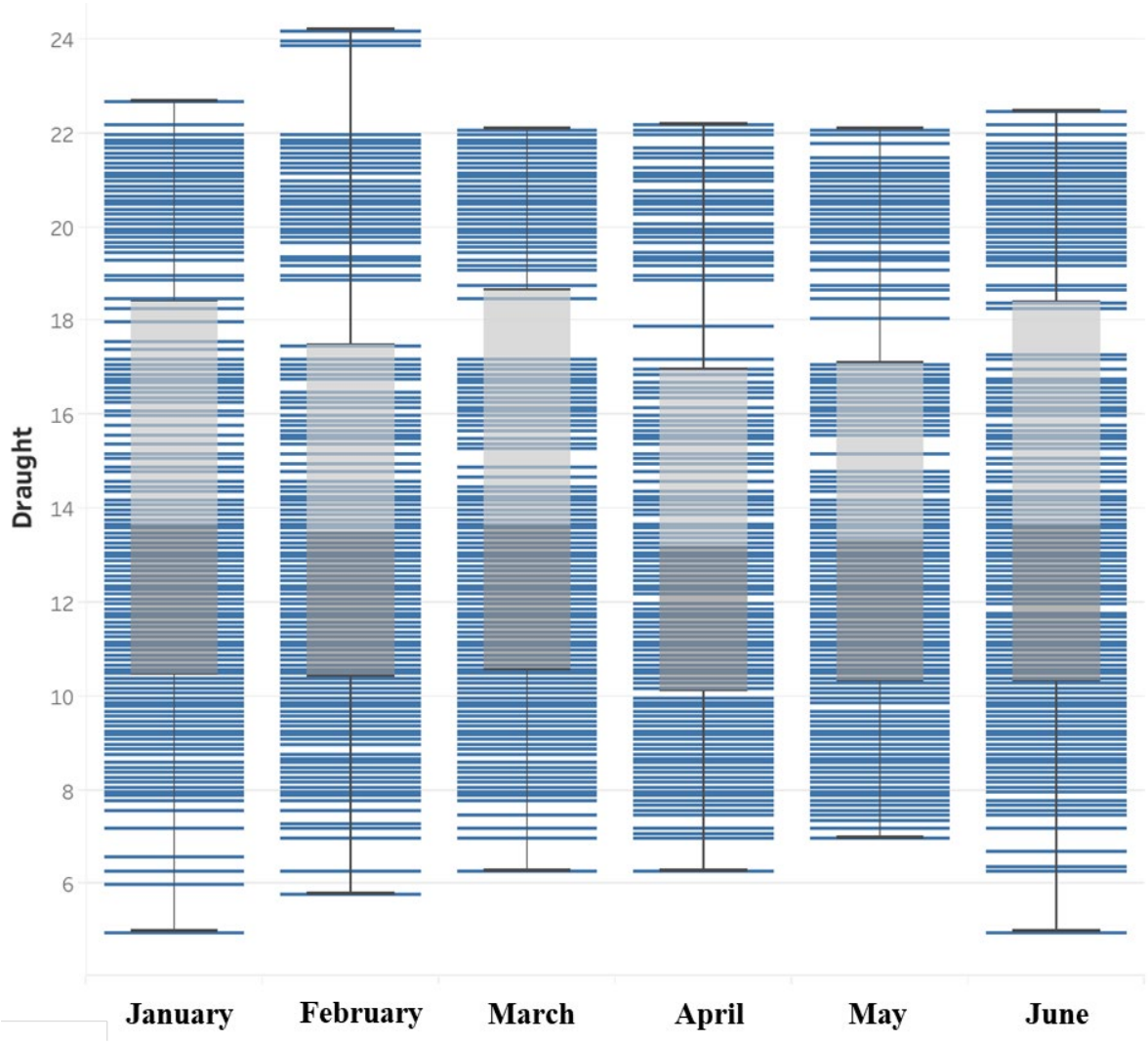


Figure 3.11 Draught distribution on month

10) Draught-COG

Figure 3.12 illustrated the relationship between the draught and COG. Vessels whose COG was between 280° and 340° generally had a smaller draught, while vessels whose COG was between 100° and 160° generally had a larger draught. This was consistent with the fact that the Arabian Peninsula was rich in oil reserves and was known as the world's oil treasure trove, with ships sailing out of the Arabian Peninsula laden with crude oil.

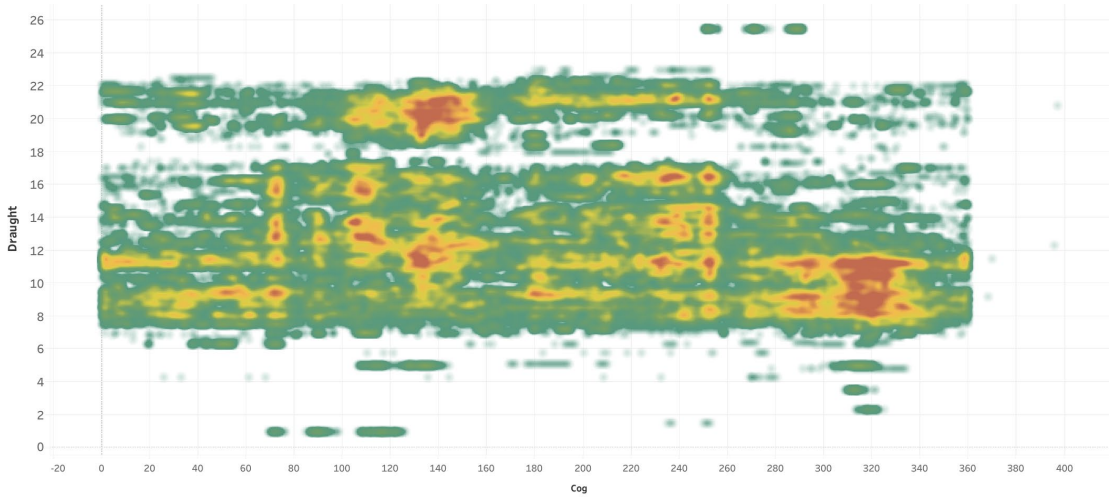


Figure 3.12 Draught distribution on COG

4. Traffic Flow Analysis Based on Improved DBSCAN

The macro features of maritime traffic flow reflect the movement patterns and characteristics of a large number of ships, which is an important reference for managing maritime traffic. Mining and analyzing the AIS data of a certain time period in a certain water area can analyze the macro features of marine traffic flow and find out the ship movement pattern.

Traditional clustering methods are mostly designed at the beginning considering only a single spatial location or thematic attribute clustering. Since AIS data objects have dual spatial, temporal and thematic characteristics, a good clustering algorithm should consider not only the spatial proximity of objects, but also the similarity of thematic attribute characteristics. That is, each AIS data object in the clustering result is continuous in the spatial domain and similar in the attribute domain.

In this chapter, the basic characteristics and parameters of maritime traffic flow are discussed. Two attributes, SOG and COG, are introduced to improve the DBSCAN algorithm. Using AIS data, constrained clustering analysis is performed on a large amount of ship AIS data over a longer period of time to delineate different traffic flows, and macroscopic analysis is performed on each traffic flow to provide a basis for the competent authorities to implement ship traffic organization.

4.1 Improved DBSCAN

Most ships navigate along the shipping lanes. Vessels navigate along recommended routes with few obstacles and well-marked navigational aids, so the risk is low. In some areas, although there are no recommended routes for the time being, customary routes have been formed as a result of long-term navigation. DBSCAN defines the area within a given object's radius ε as the object's ε-neighborhood, which is usually measured using Minkowski distances、Euclidean distances、Manhattan distances and so on..

For a large amount of AIS data, the DBSCAN algorithm can be used to carry out clustering based on the geographical distance converted from AIS latitude and longitude coordinates, however, this ignores the attributes such as heading and speed contained in the AIS data, which cannot distinguish different maritime traffic flows. Therefore, in this paper, the spatial proximity and attribute proximity are considered separately in the process of clustering, and SOG and COG are introduced in the calculation of distance, as in equation (3-1). These four attributes are normalized separately. And SOG and COG are introduced to redefine the neighborhood.

$$\mathrm{dist} = \sqrt{(lon_1 - lon_2)^2 + (lat_1 + lat_2)^2 + (sog_1 - sog_2)^2 + (cog_1 - cog_2)^2} \quad (3\text{-}1)$$

We redefine the constrained clustering ε-neighborhood as a circle centered on a particular object, with dist as the radius and satisfying a SOG difference less than MaxSog and a COG difference less than MaxCog. Given the AIS dataset D, the constrained clustering ε-neighborhood of objects m and n is given in equation (3-2).

$$\mathrm{E(m)} = \{\mathrm{n} \in \mathrm{D} \mid \mathrm{dist(m,n)} < \varepsilon \,\&\, |\mathrm{m.SOG} - \mathrm{n.SOG}| < \mathrm{MaxSog} \,\&\, |\mathrm{m.COG} - \mathrm{n.COG}| < \mathrm{MaxCog}\} \quad (3\text{-}2)$$

MaxSog and MaxCog cannot be taken arbitrarily. MaxSog with a large value, such as 5kn, then ships in the route such as 9kn and 14kn may be grouped into the same cluster and cannot distinguish ships with different characteristics in the route. If the value is taken small, such as 1kn, the differentiation is too strong and a large number of different clusters will be produced, making it difficult to identify the traffic flow. This model selected 3kn as MaxSog based on the actual data and basic information analysis.

If MaxCog takes a large value, such as 20°, after 6 cycles of calculation of AIS data points, theoretically the AIS data points with 120° heading difference in the vicinity may be grouped into the

same cluster, which obviously deviates from the original intention. if MaxCog takes a small value, such as 2°, then the normal, slight jitter operation of the rudder operator may produce more than 2° heading difference, and then the two consecutive AIS reporting points may be classified as different clusters, which also do not satisfy the requirement. This model selected 7° as MaxCog based on the actual data and basic information analysis.

4.2 Data Normalization

Longitude, latitude, SOG and COG have different magnitudes and units of magnitude, which can affect the results of data analysis. In order to remove the effects of differences between the different attributes of the input data and reduce the prediction error of the model, the input data needs to be normalized. After the normalization process, the differences between the attributes of the data can be removed and all dimensions of the data are in the same value interval, which is limited to [0,1]. The model uses the max - min normalization method to perform the normalization operation on the AIS data used in this module, as shown in the equation (3-3)

$$X' = \frac{x - \min(x)}{\max(x) - \min(x)} \qquad (3\text{-}3)$$

where max(x) is the maximum value in this attribute of the sample data, min(x) is the minimum value in this attribute of the sample data, x is the original training data; X' is the transformed data.

The normalized data still retains the relationships between the data and can reduce the complexity of the computation, and reduce the training time, as the individual data are in [0,1] between, it can also reduce bias and improve prediction accuracy.

4.3 Parameter Selection

The DBCSCAN algorithm requires the input parameters neighborhood radius Eps and the number of minimum neighbor points M$inPts$. These two parameters are not independent, they are interacting and interdependent relationships. If one of the parameters produces a change, the other parameter should follow and the DBSCAN algorithm has a high parameter sensitivity.

If a too small eps parameter is chosen, most of the data will not be clustered. And if too large a Eps value is chosen, multiple clusters will be merged and most of the objects will be located in the same cluster. In general, the smaller the eps value, the better. As a rule of thumb, only a small percentage of points should be within a distance of each other. The choice of the distance function is closely related to the choice of Eps and has a significant impact on the results.

Zhao Zhiyong (Z. Zhao, 2017) proposed a method to find the optimal radius Eps, as shown in Table 4.1. The optimal Eps can be obtained by inputting the data set and MinPts through equation.

Table 4.1 Parameter selection of DBSCAN algorithm pseudo-code

Input: datalist, MinPts
Output: Eps
1: Set m as the amount of datalist
2: Set n to the datalist dimension
3: Set xMax to the maximum value of each dimension of the datalist
4: Set xMin to the minimum value of each dimension of the datalist
5: Eps = ((np.prod(xMax-xMin) * MinPts * math.gamma(0.5 *n+1)) / (m*math.sqrt(math.pi**n))) ** (1.0/n)

In order to analyze the distribution of traffic flow in the Arabian Sea during different periods, we have selected the data for the first six days of April during the "golden season" and the last six days of June during the monsoon period. The data volume in the first six days of January was 6570, 7679, 6834, 6105, 6932, 7740, separately, and that in the last six days of June was 6475, 7157, 6954, 6312, 6287, 5270, separately. In order to avoid the influence of consecutive messages from a single ship on the clustering results, clusters with less than 0.5% of the total sample size are also set as clustering noise. So, we set MinPts for January and June at about 30. Eps obtained by the above calculation method, as shown in Table 4.2.

Table 4.2 Improved DBSCAN

Date	Volume	MinPts	Eps
1st April	6570	33	0.090242
2nd April	7679	38	0.075111
3rd April	6834	34	0.100347
4th April	6105	31	0.103412
5th April	6932	35	0.77158
6th April	7740	39	0.099199
25th June	6475	32	0.108402
26th June	7157	36	0.086310
27th June	6954	35	0.090500
28th June	6312	32	0.752550
29th June	6287	31	0.109357
30th June	5270	26	0.127136

4.4 Improved DBSCAN-based Traffic flow Characteristics Analysis

4.4.1 Major traffic flow clustering

The above algorithm was written in Python language and run in Jupyter software. The speed difference MaxSOG was set to 3kn, the heading difference MaxCOG was set to 7°, and the values of Eps and MinPts were chosen as Table 3.4. The clustering was performed for the first 6 days of April and the last 6 days of June respectively, and the clustering results were shown in Figure 4.1 and Figure 4.2. Through the figures, we could find that there were 8 main routes in the Arabian Sea, which were Gulf of Oman to Gulf of Aden traffic flow , Gulf of Aden to Gulf of Oman traffic flow , Gulf of Aden to Indian Peninsula traffic flow , Indian Peninsula to Gulf of Aden traffic flow , Indian Peninsula to Gulf of Oman traffic flow , Gulf of Oman to Indian Peninsula traffic flow , from north to south traffic flow and from south to north traffic flow .

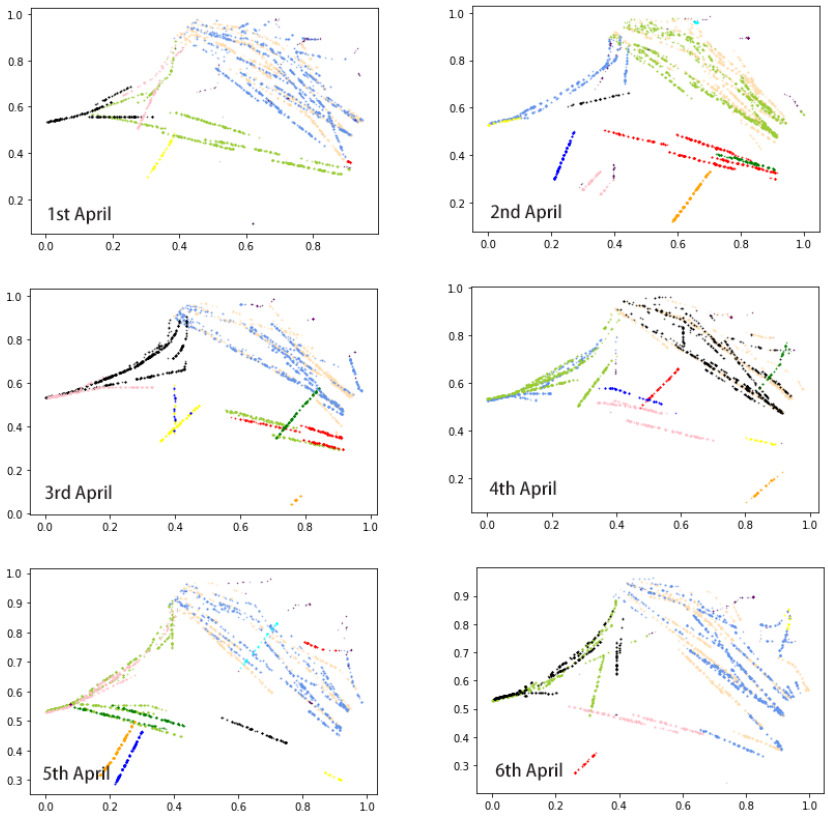


Figure 4.1 April Clustering Results

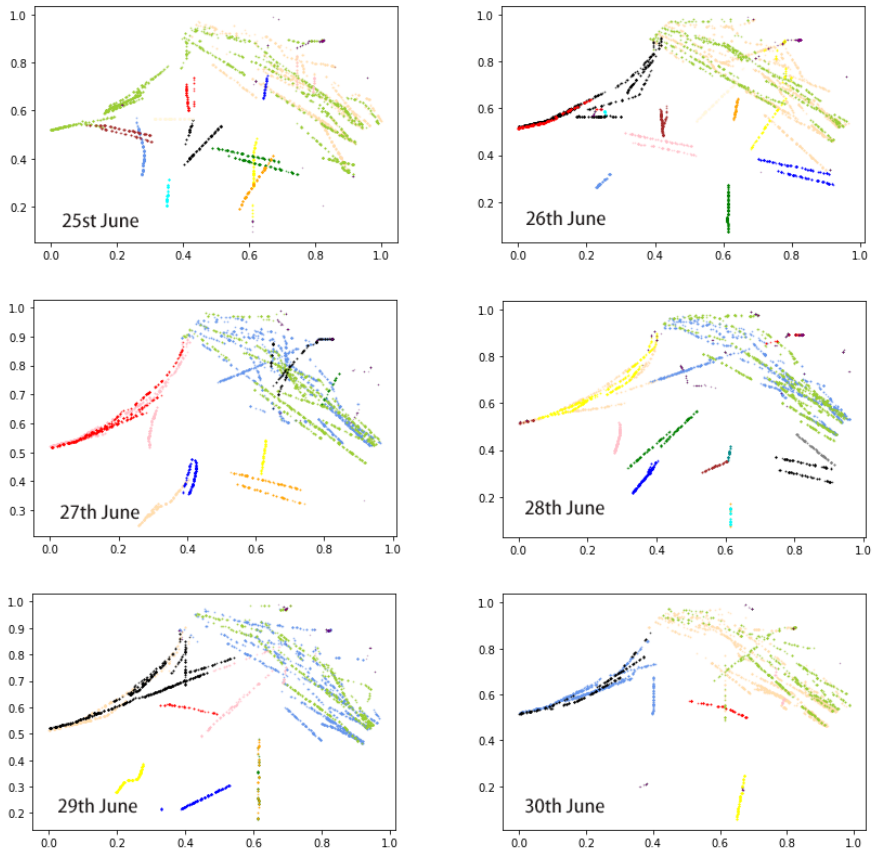


Figure 4.2 June Clustering Results

4.4.2 Cluster Analysis

By further sorting the clustering results (summarizing the clustering results for 6 days of each month; removing clusters that do not reflect the traffic flow direction; grouping clusters with the same traffic flow and the same direction into one category), we obtained 8 route clusters for April and June, as shown in Figure 4.3 and Figure 4.4.We named the Gulf of Oman direction to Gulf of Aden direction course, Gulf of Aden direction to Gulf of Oman direction course, Gulf of Aden direction to Indian Peninsula direction course, Indian Peninsula direction to Gulf of Aden direction course, Indian Peninsula direction to Gulf of Oman direction course, Gulf of Oman direction to Indian Peninsula direction, north to south course and south to north course as Cluster 1 to Cluster 8, respectively.

For the traffic flow closed to the mainland like Cluster 1, Cluster 2, Cluster 5 and Cluster 6, there was not much difference in the geographical distribution in April and June. However, for clusters 3 and 4, we found that the amount of AIS data was higher in April than in June. in April, cluster 3 includes 74 ships and cluster 4 contained 32 ships, and the number of tankers to the direction of the Indian Peninsula was more than twice as large as the number of tankers sailing in the opposite direction from the direction of the Gulf of Aden. In June, on the other hand, although the number of ships sailing from the direction of the Gulf of Aden to the direction of the Indian Peninsula was also much larger than the number of ships traveling in the reverse direction, Cluster 3 included only 23 ships and Cluster 4 contains only 5 ships. For Cluster 7 and Cluster 8, the density of ships was greater in June than in April. This may be due to the prevailing southwest monsoon in the Arabian Sea during summer.

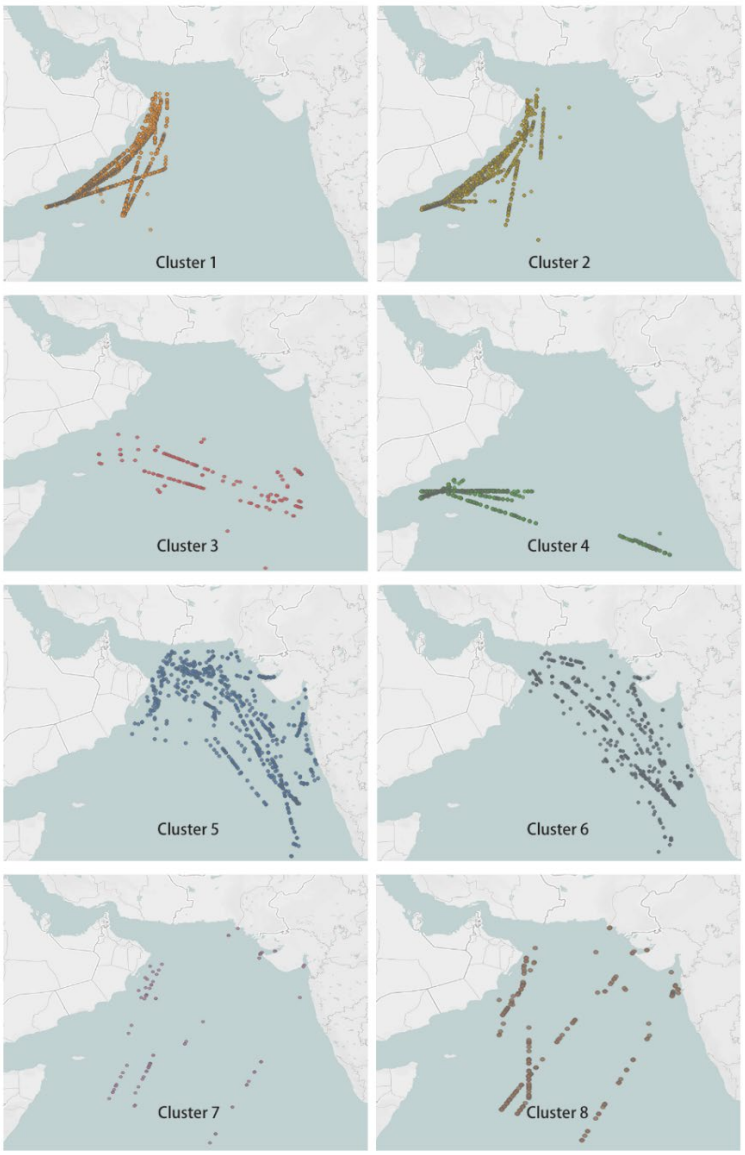


Figure 4.3 Traffic Flow in April

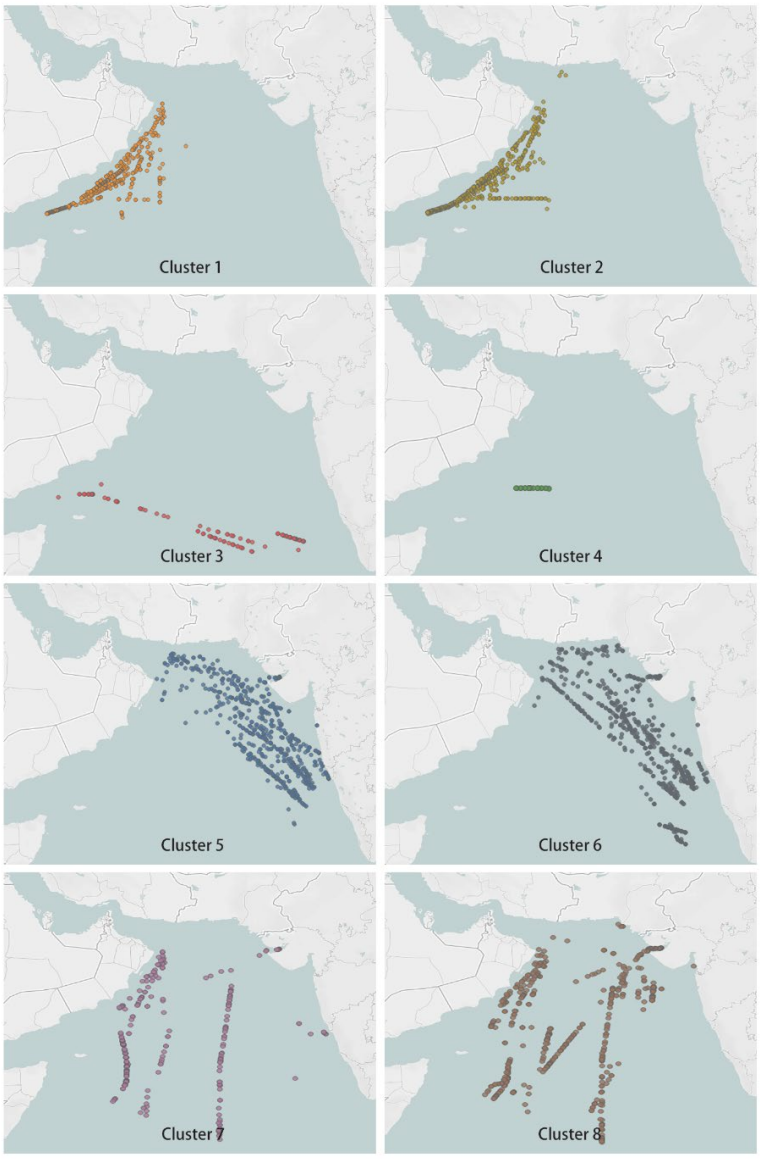


Figure 4.4 Traffic Flow in June

1) Traffic flow direction

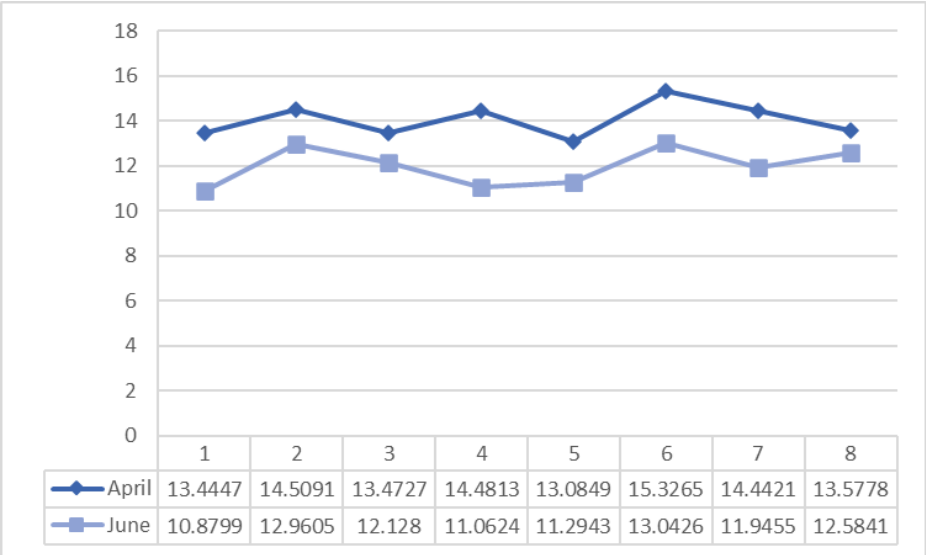
2)

In April, the average COG of sailing of energy tankers in Cluster 1 was 228.5746° and it of sailing of tankers in Cluster 2 in the opposite direction was 98.023°. The average COG of sailing of vessels in Cluster 3 was 131.9808° and it of sailing of vessels in Cluster 4 in the opposite direction was 258.8494°. The average COG of sailing for tankers in Cluster 5 is 312.4453° and the average COG of sailing for tankers in Cluster 6 in the opposite direction was 142.0153°. Cluster 7 and Cluster 8, which travel from north to south and from south to north, respectively, had an average COG of 133.9889° and 61.3264°, indicating that the vessels traveled with a certain inclination angle rather than vertical.

In June, the average COG of sailing of energy tankers in Cluster 1 was 231.4774°, while it of sailing of energy tankers in Cluster 2 was 94.7705° in the opposite direction. The average COG of sailing of vessels in Cluster 3 was 108.7232° and it of travel of vessels in Cluster 4 in the opposite direction was 270.4427°. The average direction of sailing of the energy tankers in cluster 5 was 306.593° and the average direction of travel of the ships in cluster 6, which was the opposite direction, was 171.5736°. The average COG of cluster 7 and cluster 8 were 191.7661° and 87.6246°, respectively.

3) Traffic flow SOG

As shown in Figure 4.5, the dark blue line and the light blue line represented the Average SOG of each cluster during April and June, respectively. The average SOG of energy tankers during April was significantly greater than that of June, with an average of 2.055kn higher. The average SOG of energy tankers sailing from the direction of Gulf of Aden to the direction of Gulf of Oman was about 1kn greater than that of tankers sailing from the direction of Gulf of Oman to the direction of Gulf of Aden. The Average SOG of ships sailing from the Gulf of Oman to the Indian Peninsula was about 1.3kn faster than the reverse tankers.

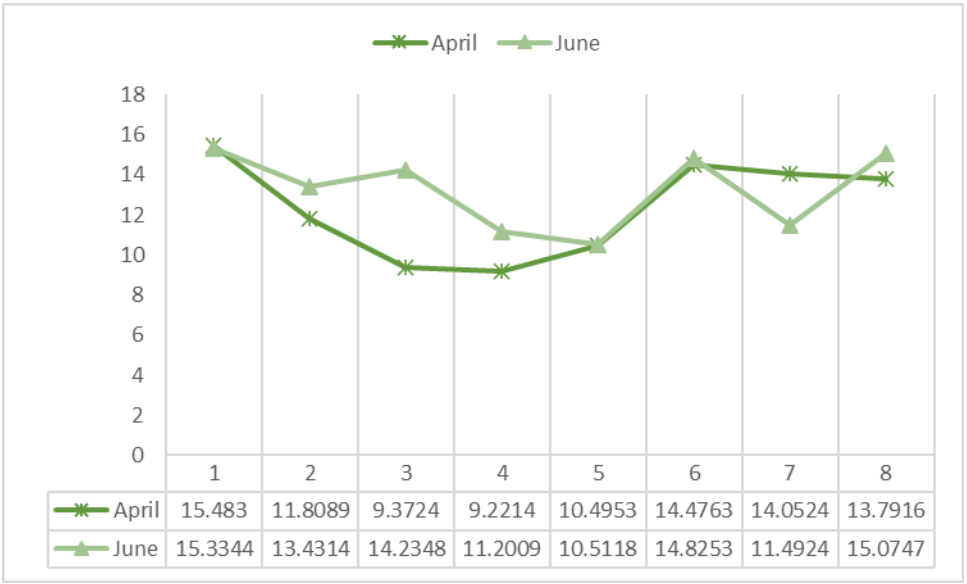


	1	2	3	4	5	6	7	8
April	13.4447	14.5091	13.4727	14.4813	13.0849	15.3265	14.4421	13.5778
June	10.8799	12.9605	12.128	11.0624	11.2943	13.0426	11.9455	12.5841

Figure 4.5 Average SOG of clusters

3) Traffic flow draught

As shown in Figure 4.6, the dark green line and the light green line represented the Average draught of each cluster within April and June, respectively. In overall, the draught was larger in June than in April. In April, the draught of ships within Cluster 1, Cluster 6, Cluster 7 and Cluster 8 was larger, 15.483, 14.4763, 14.0524 and 13.7916, respectively. In June, the draught of ships in Cluster 1, Cluster 6 and Cluster 8 was larger with 15.3344, 14.8253 and 15.0747, respectively.



	1	2	3	4	5	6	7	8
April	15.483	11.8089	9.3724	9.2214	10.4953	14.4763	14.0524	13.7916
June	15.3344	13.4314	14.2348	11.2009	10.5118	14.8253	11.4924	15.0747

Figure 4.6 Average draught of clusters

5. Vessel Encounter Recognition Based on the ST-DBSCAN

5.1 ST-DBSCAN

AIS data contains attributes such as latitude and longitude position and time, which are typical spatial-temporal data. Mining the spatial-temporal patterns of traffic in the sea through AIS data is important for predicting the future movement of ships and carrying out traffic organization. In this chapter, based on the spatial-temporal density, the spatial-temporal information contained in the trajectory is also considered in the clustering process to mine the spatial-temporal patterns of ships.

The ship density distribution directly reflects the spatial distribution of ships in the study waters, and thus indicates the busy and congested level of maritime traffic. However, ship density is often an average quantity measured many times over a long period of time, which can hardly reflect the degree of ship congestion on the spatial and temporal axes. For example, the ship density in area A is high, but the time interval between the appearance of immediately adjacent ships is long, so that no meeting situation is formed and no urgent situation is created, and the risk of marine traffic is low. However, in region B, although the ship density is not high, a large number of ships appear in a short sub-time period, resulting in a close encounter situation and an urgent situation, and the maritime traffic risk is high. In this paper, we solve the problem of reflecting ship density in time and space at the same time by adopting STDBSCAN, which is based on DBSCAN with the addition of parameters measuring the time dimension.

5.2 Parameter Selection

The DBSCAN algorithm uses only one distance parameter Eps to measure the similarity of one-dimensional spatial data. Based on this, we propose two distance measures, Eps1 and Eps2. Eps1 is used for spatial values to measure the geographical proximity of two points. Eps2 is used to measure the similarity of temporal values.

For any point X, there are at least MinPts points in a cylinder with X as the center, Eps1 as the radius, and Eps2 as the height. The cylinder travels in the 3-dimensional spatial-temporal domain until it no longer contains MinPts points. All points in a cylinder form a spatial-temporal neighborhood cluster, as presented in Figure 5.1.

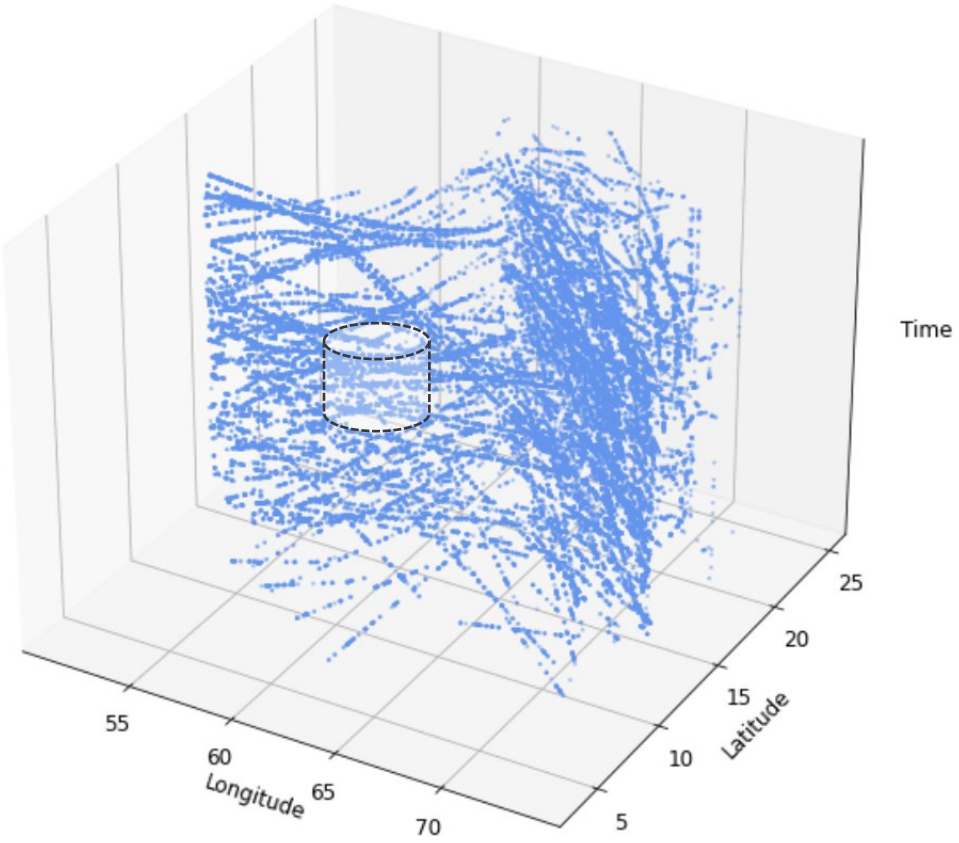


Figure 5.1 Diagram of AIS Data Spatial-temporal clustering based on STDBSCAN

The distance corresponding to space and time, dis1 and dis2 are calculated as shown in Equation (5-1) (Baskar & Xavior, 2021)and Equation (5-2), respectively.

$$dis1 = 2 * \text{asin}\left(\sqrt{\sin((lat2 - lat1)/2)^2 + \cos(lat1) * \cos(lat2) * \sin((lon2 - lon1)/2)^2}\right) * r * k \quad (5-1)$$

Where lat1 and lat2 are the latitude of the two points, respectively. lon1 and lon2 are the longitudes of the two points, respectively. r is the mean radius of the Earth, approximately 6371. k is the value of converting kilometers to nautical miles, approximately 0.5399568。The unit of dis1 is n mile.

$$dis2 = time1 - time2 \quad (5-2)$$

Where time1 and time2 are the reported times of the two points, respectively. The unit of dis2 is minute.

The selection of parameters needs to consider the navigable environment of the study waters. The average length of navigable ships in the Arabian sea is 285m, which is about 0.15nmile, and the distance between ships and ships is generally more than 2 times the length of the ship. Generally, when the distance between two vessels is more than 3nmile, they have no influence on each other. So, the value of Eps1 should be greater than 0.3nmile, less than 3 n mile.

Vessel encounter time is usually a few minutes. Although the spatial distance between two ships AIS points is similar, if the time difference is more than 10min, the probability of actually appearing at the same time and causing an urgent situation is small, and it is also difficult to find the accompanying law. Therefore, the value of A should be greater than 1min and less than 10min.

5.3 Multi-vessel Encounter Situational Recognition

5.3.1 Data preparation and algorithm running

We select the vessel data from January to June 2016 for the Arabian Sea separately for cluster analysis. The monthly AIS data are shown in Table 5.1, and the number of crude oil tankers, oil product tankers and LNG tankers passing through the sea is generally more than 600 per month. Among them, March had the highest number of tankers passing through the sea area. It can be seen that the monthly data volume and the number of passing tankers are evenly distributed. The above algorithm is written in Python language and run with the help of Jupyter software. Eps1 is set to 1nmile, Eps2 is set to 7min, MinPts is taken as 2, and different MMSI categories are selected.

Table 5.1 AIS data

Month	Amount of Vessel	Data Volume
January	641	227064
February	602	215943
March	655	228336
April	613	205878
May	618	204605
June	643	208658

The following clustering results were obtained. As shown in Table 5.2, there were more tankers encounters in the Arabian Sea in January and June 2016, with 33 and 30 encounters, respectively. The least number of encounters occurred in May, with 13.

Table 5.2 ST-DBSCAN Clustering Result

Month	Amount of cluster group
January	33
February	18
March	16
April	17
May	13
June	30

5.3.2 Cluster Analysis

1）Spatial Analysis

Figure5.2 showed the geographical distribution of the cluster results in the Arabian Sea, and Figure5.3 was a statistical bar chart of the geographical distribution of ship encounters. Except for January, the locations of encounters in other months were offshore and in the area leading to the Gulf of Aden. This indicated that the offshore areas and the area leading to the Gulf of Aden had a high density of vessels and were prone to vessel encounters. The Arabian Sea was connected to the Gulf of Aden in the west, and was a must for energy tankers to sail quickly between the Mediterranean Sea and the Indian Ocean, an important waterway for oil from the Persian Gulf to Europe and North America and an important route for trade and energy transportation for all countries in the world. The large number of energy tankers entered the narrow Gulf of Aden area from the vast Arabian Sea, had led to an increase in ship density, increased the number of encounters and the risk of collisions. Most of the encounters that occurred in the eastern side of the Arabian Sea were concentrated in the offshore area near Indian ports. The number of encounters in the offshore regions of India from January to May was about the same, but the number of encounters in June was twice as many as before. This may be due to the fact that the navigation

environment becomes progressively worse in June due to the monsoon, and ships tend to travel offshore.

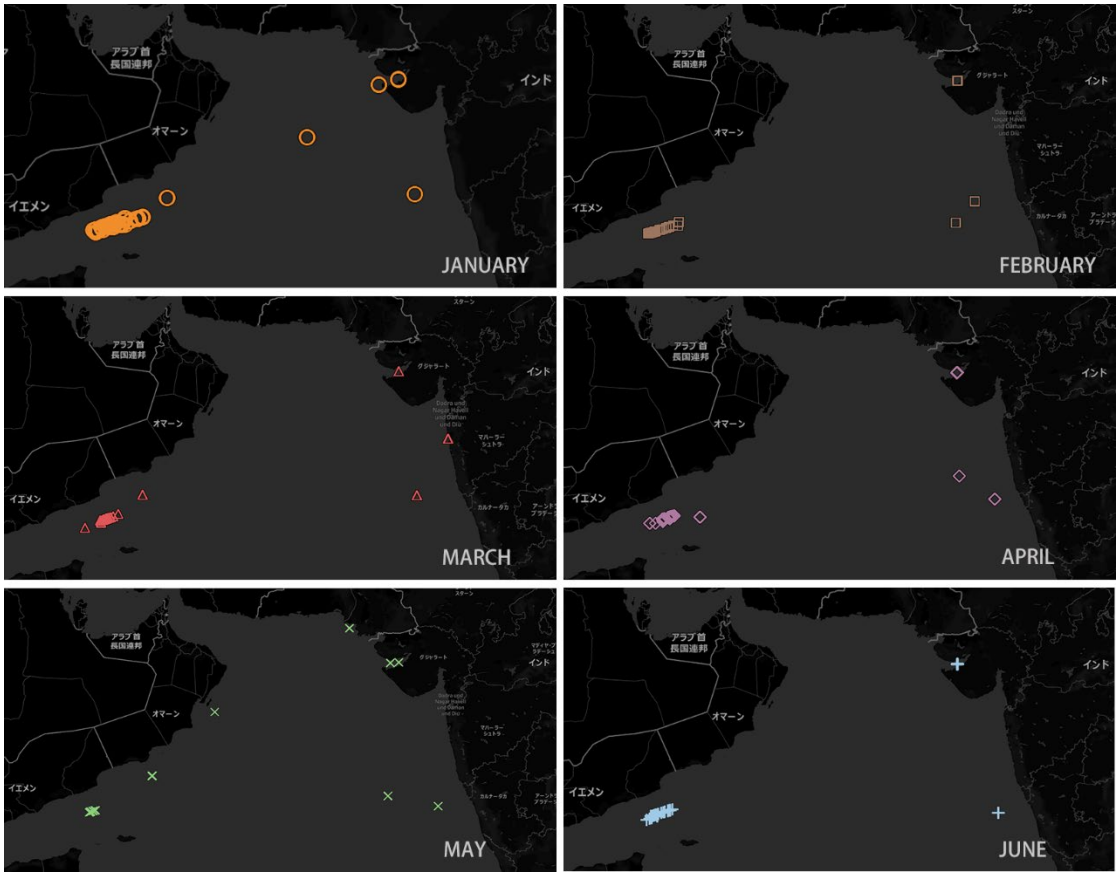


Figure 5.2 Geographical Distribution of the Cluster Results

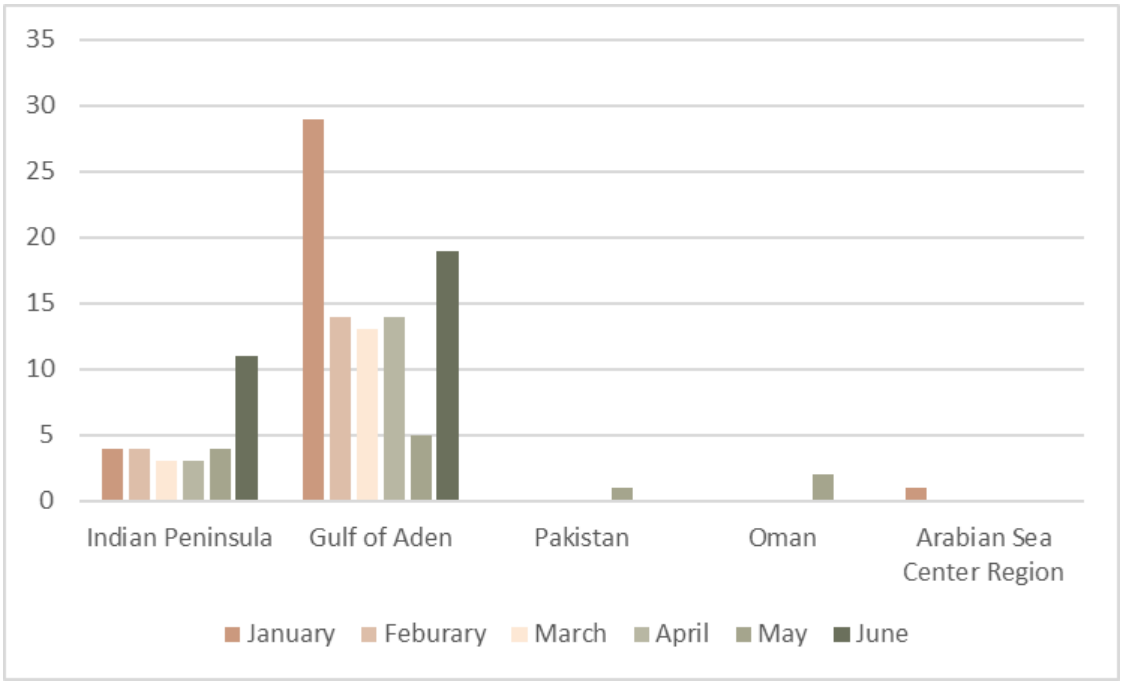


Figure 5.3 Bar chart of the Geographical Distribution of Vessel Encounters

2）Temporal Analysis

The 24-hour day was divided into 6 equal time periods according to the crew watch pattern, i.e.

00:00-04:00, 04:00-08:00, 08:00-12:00, 12:00-16:00, 16:00-20:00 and 20:00-24:00. From Figure 5.4, we found that overall, vessel encounters occur in Arabian waters at night more often and less often during the day. There was a decreasing trend from 0:00 to 12:00 and an increasing trend from 12:00 to 24:00.

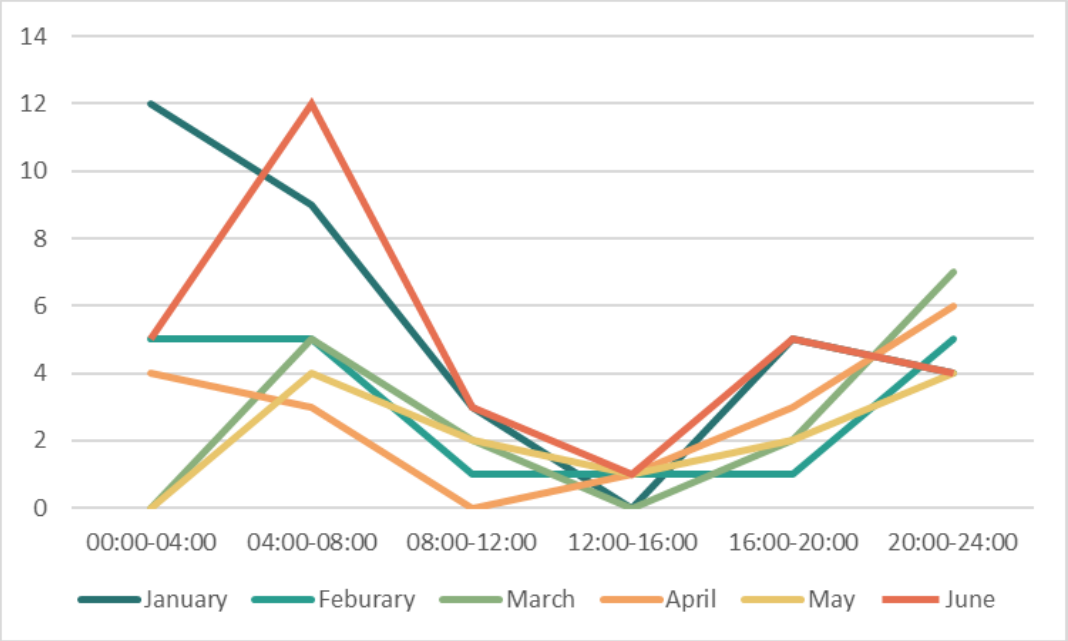


Figure 5.4 Temporal Distribution of Multi-ship Encounters

3）Amount Analysis

The encounters in Arabian waters were mainly 2-tankers encounters. In the first quarter, there were only two-vessels encounters, but in the second quarter, there were not only two-ship encounters but also three-vessels encounters. There was one 3-ship encounter in April and one in May, both of which were located in the waters off India. In June, there were four 3-ship encounters, three of which occurred in the offshore waters of India. This indicated that there was a higher density of ships in the Indian offshore waters, and there was a higher chance of more than two ship encounters.

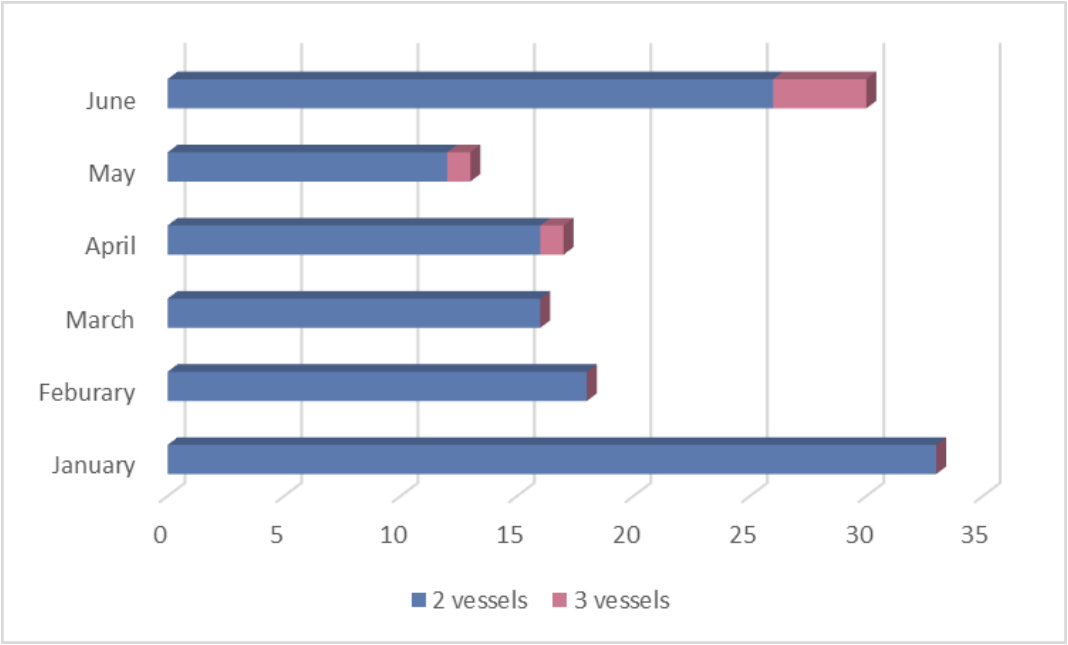


Figure 5.5 Statistics of Multi-ship Encounters

6. Conclusion

Using the spatial-temporal data of ship navigation, we extracted and analyzed the characteristics of maritime traffic flow by computer data mining technology, cluster analysis and other methods. It was conducive to understanding the spatial and temporal distribution of traffic flow in the target sea area, so as to provided a reference basis for traffic management and the formulation of relevant policies and regulations. It also could provide guidance for ship route planning and reduce the labor cost of maritime navigation. In this paper, we focused on maritime traffic flow and vessel encounters to carry out relevant research. Based on AIS data and data clustering, the macro features of maritime traffic flow, maritime traffic flow clustering and maritime traffic spatial and temporal patterns were taken as the main research contents. The main research content and innovation points of the paper included the following aspects:

1）The paper selected the Arabian seas, where the largest volume of oil is transported, for the study. The AIS data of crude oil tanker, oil products tanker and LNG tanker worldwide in the first half of 2016 were screened and processed to analyze the data quality. The basic traffic status in this sea area was also analyzed, including the composition ratio of ship types regarding space and time, the spatial composition of destinations, the distribution of SOG regarding geographic space and time, the spatial distribution of COG, the distribution of draught regarding geographic space and time, and the relationship between draught and destinations and COG respectively. In the Arabian Sea, the speed of ships was generally maintained between 0 and 21kn, with most ships sailing at medium speed (12kn-15kn). Vessel speeds were faster near the Indian Peninsula than near the Arabian Peninsula. The average daily SOG of vessels in January to May concentrated between 12.5 kn and 14.5 kn, but it dropped steeply in June and showed a continuous decreasing trend. The draught was concentrated above 8m and under 22m. And there was a blank draught record between 17-19 m. Vessels whose COG was between 280° and 340° generally had a smaller draught, while vessels whose COG was between 100° and 160° generally had a larger draught.

2）Constrained clustering analysis was performed on AIS data to classify maritime traffic flows and to analyze the macroscopic features of maritime traffic flows. Different from the traditional DBSCAN algorithm, we incorporated longitude, dimension, SOG and COG into the calculation of distance together, and added the attribute constraints of SOG and COG to distinguish different traffic flows. The results proved that improved DBSCAN could cluster the AIS data of ships in the sea well and get the realistic traffic flow. It was shown that the distribution of traffic flow was influenced by the weather environment. The average SOG of energy tankers during April was significantly greater than that of June, with an average of 2.055kn higher. The average SOG of energy tankers sailing from the direction of Gulf of Aden to the direction of Gulf of Oman was about 1kn greater than that of tankers sailing from the direction of Gulf of Oman to the direction of Gulf of Aden. The Average SOG of ships sailing from the Gulf of Oman to the Indian Peninsula was about 1.3kn faster than the reverse tankers. And the draught was larger in June than in April in overall.

3）The ST-DBSCAN algorithm was used to cluster the spatial and temporal dimensions of AIS data by introducing the temporal dimension based on the spatial distance. By this method, the busy and congested level of watershed could be judged simultaneously in space and time. The study found that the offshore areas within the Arabian Sea and the area leading to the Gulf of Aden had a high density of vessels and were prone to vessel encounters. Vessel encounters occurred in Arabian waters at night more often and less often during the day. There was a decreasing trend from 0:00 to 12:00 and an increasing trend from 12:00 to 24:00. And there was a higher chance of more than two ship encounters in the Indian offshore waters.

This thesis also has limitations.

1) The DBSCAN algorithm has weaknesses in terms of convenience and time consuming, which need further improvement.

2) The AIS data analyzed in this thesis only includes crude oil tanker, oil products tanker and LNG tanker. Other types of vessels, such as fishing vessels and container ships, can be considered in future studies of traffic flow in the Arabian Sea.

Acknowledgement

2 years have passed in a flash, and graduation is around the corner. Along with gaining knowledge, I have also gained valuable life experiences and precious friends.

First of all, I would like to express my deepest gratitude to my supervisor, Professor Watanabe Daisuke, who has done a great favor and provided Careful guidance to my thesis. I would like to thank my sensei for taking time out of his busy schedules to answer various questions and solve difficulties in my studies and life.

I owe my thanks to my labmates for helping me with my studies and life. Inoue san helped me with various procedures that I needed to go through. Nishimiya san took the initiative to chat with me when I was not familiar with everyone so that I could integrate with everyone as soon as possible. Thuta san guided me with my data, thesis, and code.

Special thanks should also go to my parents. My parents respect my decision. Because of your dedication and support, I can go forward without any worries.

I also would like to grant my cousin for staying up late with me to modify the code.

Last but certainly not least, thanks to my friends in the JCK program! May our friendship last forever! Due to the epidemic, it was only in April this year that we changed from online friends to offline friends. Because of you, I don't feel lonely, and every day is fulfilling and happy. Wang Qian, whose name is only one word away from mine, is inexplicably in tune with me in hobbies, life, etc. This is fate!

References

Baskar, A., & Xavior, M. A. (2021). A four-point direction search heuristic algorithm applied to facility location on plane, sphere, and ellipsoid surfaces. *Journal of the Operational Research Society*. doi:10.1080/01605682.2021.1984185

Birant, D., & Kut, A. (2007). ST-DBSCAN: An algorithm for clustering spatial–temporal data. *Data & Knowledge Engineering, 60*(1), 208-221. doi:10.1016/j.datak.2006.01.013

Bushra, A. A., & Yi, G. (2021). Comparative Analysis Review of Pioneering DBSCAN and Successive Density-Based Clustering Algorithms. *IEEE Access, 9*, 87918-87935. doi:10.1 109/ACCESS.202 1.3089036

Cerri, R., & de Carvalho. (2011). Adapting non-hierarchical multilabel classification methods for hierarchical multilabel classification. *Intelligent Data Analysis, 15*(6), 861-887. doi:10.3233/IDA-2011-0500

D, K., K, H., & OkimotoT. (2017). Distributed Stochastic Search Algorithm for Multi-ship Encounter Situations. *Journal of Navigation, 70*(4). doi:10.1017/S037346331700008X

Dobrkovic, A., Iacob, M.-E., & Hillegersberg, J. v. (2016). *Maritime Pattern Extraction from AIS Data Using a Genetic Algorithm*. Paper presented at the 2016 IEEE International Conference on Data Science and Advanced Analytics, Montreal, QC, Canada.

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*. Paper presented at the Proceedings of the Second International Conference on Knowledge Discovery and Data Mining.

Goodwin, & M, E. (1975). A Statistical Study of Ship Domains. *Journal of Navigation, 28*(3), 328-341. doi:10.1017/s0373463300041230

Gu, Y., Ye, X., Zhang, F., Du, Z., Liu, R., & Yu, L. (2017). A parallel varied density-based clustering algorithm with optimized data partition. *Journal of Spatial Science, 63*(1), 93-114. doi:10.1080/14498596.2017.1352542

Han, J., & Kamber, M. (2001). *Data Mining Concepts and Techniques*. San Francisco: Morgan Kaufmann Publishers.

He, Y., Jin, Y., Huang, L., Xiong, Y., Chen, P., & Mou, J. (2017). Quantitative analysis of COLREG rules and seamanship for autonomous collision avoidance at open sea. *Ocean Engineering, 140*(1), 281-291. doi:10.1016/j.oceaneng.2017.05.029

Huang, Y., & M, v. G. P. H. A. J. (2020). Time-Varying Risk Measurement for Ship Collision Prevention. *Risk Analysis, 40*(1), 24-42. doi:10.1111/risa.13293

Kriegel, H.-P. (2011). Density-based clustering. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1*(3), 231-240. doi:10.1002/widm.30

L.Huntley, C., & Brown, D. E. (1996). Parallel genetic algorithms with local search. *COMPUTERS & OPERATIONS RESEARCH, 23*(6), 559-571. doi:10.1016/0305-0548(95)00061-5

Lee, J., & Cho, N. (2016). Fast Outlier Detection Using a Grid-Based Algorithm. *Plos One, 11*(11). doi:10.1371/journal.pone.0165972

Liu, J. (2009). *Research of the Through Capacity and Traffic Organizational Model in Main Channel of Large Sea Harbor.* (Doctor), Wuhan University of Technology,

Lv, P., Zhuang, Y., & Li, Y. (2017). BP neural network combined with Markov prediction model of ship traffic flow. *Journal of Shanghai Maritime University, 38*(2). doi:10.13340 /j.jsmu.2017.02.004

Meng, Q., Weng, J., & Li, S. Analysis with Automatic Identification System Data of Vessel Traffic Characteristics in the Singapore Strait. *TRANSPORTATION RESEARCH RECORD, 2426*(1), 33-43.

doi:10.3141/2426-05

Mou, J., Chen, P., & He, Y. (2018). Fast self-tuning spectral clustering algorithm for AIS ship trajectory. *Journal of Harbin Engineering University, 39*(3). doi:10. 11990 / jheu． 201609033

Pan, J., Jiang, Q., & Shao, Z. (2010). Ship Encounter Data Mining Algorithm. *Navigation of China, 33*(4). doi:10.16411/j.cnki.issn1006-7736.2012.01.023

Ristic, B., Scala, B., Morelande, M., & Gordon, N. (2008). *Statistical analysis of motion patterns in AIS Data: Anomaly detection and motion prediction*. Paper presented at the Information Fusion, 2008 11th International Conference on.

Vespe, M., Visentini, I., Bryan, K., & Braca, P. (2012). *Unsupervised learning of maritime traffic patterns for anomaly detection*. Paper presented at the 9th IET Data Fusion & Target Tracking Conference (DF&TT 2012): Algorithms & Applications.

Wang, P., Wu, S., Zhang, H., & Lu, F. (2019). Indoor Location Prediction Method for Shopping Malls Based on Location Sequence Similarity. *ISPRS International Journal of Geo-Information, 8*(11), 517. doi:10.3390/ijgi8110517

Zhao, L. (2013). *Study on the Statistical Analysis and Prediction Model of Vessel Traffic Flow Characteristics.* (Master), Wuhan University of Technology,

Zhao, Z. (2017). *Python Machine Learning Algorithms*: Electronics Industry Press.

Zhou, Z. (2016). *Machine Learning*: Tsinghua University Press.

Zhu, J., & Zhang, W. (2009). Calculation model of inland waterway transit capacity based on ship-following theory. *Journal of Traffic and Transportation Engineering, 9*(5), 1671-1637. doi:10.19818/j.cnki.1671-1637.2009.05.015