Research on assessment and prediction of maritime traffic status based on AIS data using Deep learning

Master’s Thesis

RESEARCH ON ASSESSMENT AND PREDICTION OF MARITIME TRAFFIC STATUS BASED ON AIS DATA USING DEEP LEARNING

September 2020

Graduate School of Marine Science and Technology
Tokyo University of Marine Science and Technology
Master’s Course of Marine Technology and Logistics

Wang Yongpeng

Master’s Thesis

RESEARCH ON ASSESSMENT AND PREDICTION OF MARITIME TRAFFIC STATUS BASED ON AIS DATA USING DEEP LEARNING

September 2020

Graduate School of Marine Science and Technology
Tokyo University of Marine Science and Technology
Master’s Course of Marine Technology and Logistics

Wang Yongpeng

Abstract

The supervision of marine traffic conditions is more difficult than that of land traffic. At the same time, the cost of traffic accident losses and environmental pollution caused by ships is very high. Therefore, it is very important to strengthen the assessment of marine traffic conditions and monitor the navigation status of ships. Based on this research, the following work has been done on the marine traffic situation:

First, this study used real-time data recorded by AIS, including IMO number, dynamic time, speed of ground (SOG), course of ground (COG), draught and latitude and longitude, as well as K-means clustering and traffic flow methods, to analyze the sea of the Malacca Strait from January to June 2016 Traffic conditions (mainly energy vessels). The course is mainly divided into the southeast course and the northwest course. The southeast heading is close to the Indonesian side, the northwest course is close to the Malaysian side, the traffic flow of the ship is the largest in January and March, the northwest course's traffic flow and speed are faster than the southeast course, the draft of the southeast course It is greater than the northwest course, indicating that the supply speed of the oil and gas supply route passing through the Strait of Malacca in the first half of 2016 is lower than the demand speed, and the supply volume is less than the demand volume.

Secondly, this study selected the AIS data of LNG ships in the Strait of Malacca as a sample and used cubic spline interpolation and linear interpolation to repair the trajectory of 140 LNG ships in the Strait of Malacca with one AIS data per minute. Using the ITTC formula to estimate the GHG emissions of LNG ships, the GHG emission inventory of LNG ships in the Straits of Malacca was obtained. The results were as follows: The GHG emissions of LNG ships in the Straits of Malacca in the first half of 2016 were 607,719.690MT. The most greenhouse gas emissions were 235,732.699MT. In addition, QGIS was also used to divide the Malacca Strait according to a 5 km*5km grid evolution, to obtain the spatial distribution of greenhouse gas emissions. Since the speed of the Malacca Strait energy ship changes steadily, most of them were distributed on the southeast route and the northwest route. The difference was small, so the total greenhouse gas emissions in the ship-dense area were large, and they were mostly distributed in the narrow part of the southeast end of the Strait of Malacca.

Finally, this study proposed to use LSTM to predict the ship's navigation status and predict the ship's navigation dynamics (SOG, COG, position, CO2) in advance, in order to achieve the purpose of monitoring marine traffic. In this study, 2 deep learning methods, RNN and LSTM were used to predict the ship's navigation status. The results showed that LSTM had the best prediction effect.

Keyword: AIS, Traffic flow, K-means clustering, interpolation, ITTC, Deep learning

Contents

1. Introduction

1.1 Research Background

(1) Maritime traffic management is a complex and systematic project, which covers a wide area from port to sea, air to sea, including maritime investigation, pollution management, traffic management, navigation management, beacon management, etc., which requires Invest a lot of manpower and material resources. With the rapid growth of maritime traffic, this has increased the burden of water bearing, congested waterways and serious air pollution. In this case, the problems of the ship itself and the increase in accidents caused by human factors will cause huge economic losses. Based on the above-mentioned problems, the maritime traffic authority has established a corresponding Vessel Traffic Service (VTS) in the port, and the information such as the ship name, position, and SOG is displayed on the screen and passed through the Very High Frequency (Very High Frequency) , VHF) and other means to intervene in real time, refer to Figure 1.1. Although modern monitoring equipment improves navigation safety and navigation efficiency, it still faces many problems.

Figure 1.1 VTS Center

Source: SAAB BRAZIL

First, the macroscopic characteristics of traffic flow in the VTS area for a long period of time are difficult to accurately grasp. Although the nautical book materials indicate the relevant routes, channels and turning points, the nature of the port terminals of the territories, the operating characteristics and the daily supervision experience also allow the VTS duty personnel to have a general understanding of the traffic flow profile, but the actual traffic flow parameters and the traffic flow status of the key ships under supervision , it is still difficult to grasp. Secondly, the increasing number of ships and types of ships make VTS duty personnel have to deal with more and more information, which makes duty personnel easily confused by the ship data on the screen, increasing the cognitive load and relying on the intuitive judgment of the duty personnel And analysis can no longer meet the needs of marine traffic management, so how to mine marine traffic characteristics and calculate the complexity of marine traffic from a large number of ship data is a problem that the marine traffic authorities need to solve urgently.

AIS technology is more and more widely used in marine traffic management with its huge data and information advantages, which improves management efficiency. Because the application of AIS is obviously more reliable than pure radar signals. It can realize the exchange of information between the ship and the ship's shore, realize the real-time online monitoring of the ship, and has great advantages in ship identification and ship tracking. In actual operations, the application of AIS technology can reduce the high-frequency talk time between VTS operators and crew. This is due to the intelligence and automation of AIS. VTS operators can grasp the traffic situation of ships in real time, which greatly Improve the efficiency of traffic control.

(2) In addition, the issue of air pollution from ships has received great attention both at home and abroad in recent years. According to Third IMO GHG Study (2014), International shipping CO2estimates range between approximately 596 million and 649 million tonnes calculated from top-down fuel statistics, and between approximately 771 million and 921 million tonnes according to bottom-up results. International shipping is the dominant source of the total shipping emissions of other GHGs: nitrous oxide (N2O) emissions from international shipping account for the majority (approximately 85%) of total shipping N2O emissions, and methane (CH4) emissions from international ships account for nearly all (approximately 99%) of total shipping emissions of CH4. Maritime CO2 emissions are projected to increase significantly in the coming decades. Depending on future economic and energy developments, this study's BAU scenarios project an increase by 50% to 250% in the period to 2050. Further action on efficiency and emissions can mitigate the emissions growth, although all scenarios but one project emissions in 2050 to be higher than in 2012.

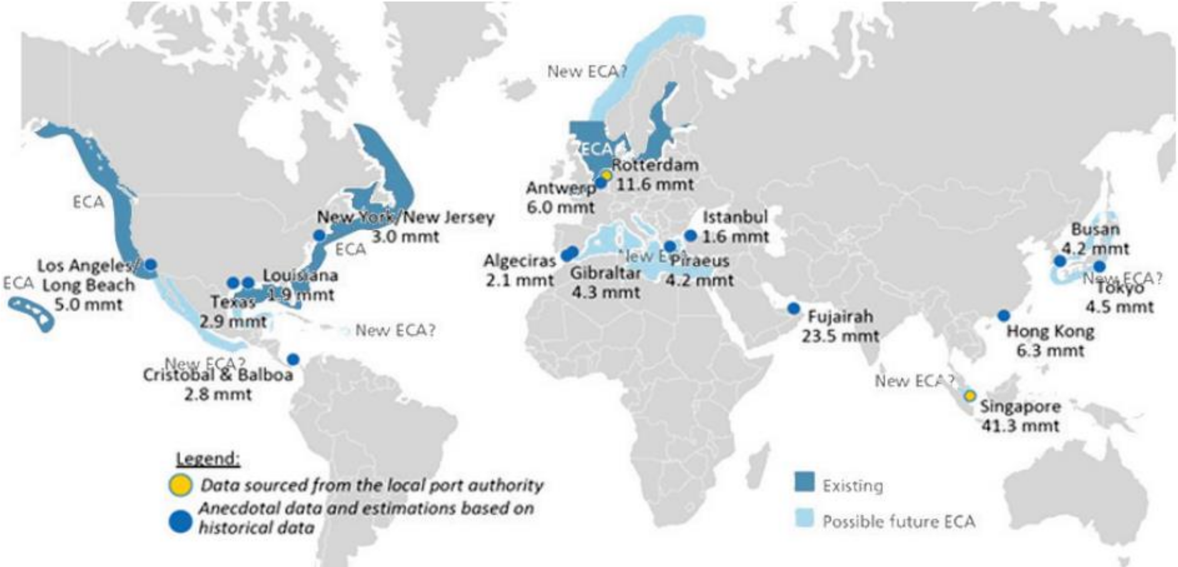


Figure 1.2 ECA Area in 2020

Source: Poten & Partners, "Marine Fuel Regulations 2010-2025"

Therefore, the International Maritime Organization (IMO) and major shipping countries are formulating corresponding policies and measures to reduce emissions from shipping vessels. For example, the 70th meeting of the IMO Marine Environmental Protection Committee (MEPC) passed a resolution requiring that the sulphur content of fuel oil for ships in the global sea area not exceed 0.5% from 2020. At the same time, various countries have successively established ECA (Emission Control Area), refer to Figure 1.2. Therefore, accurate estimation of ship energy consumption and emissions is particularly important for marine traffic management departments. For example, various emission reduction policies can be formulated and evaluated based on emission data.

At the same time, the AIS data can also find the abnormal trajectory of the ship in time, reducing the risk of marine traffic accidents. By predicting the trajectory of the ship, the trajectory of the ship and the change of the navigation status can be found in time, which is conducive to the effective monitoring of the ship. Real-time, accurate and reliable ship trajectory prediction can ensure the safety of ship navigation, effectively improve the efficiency of marine traffic, and improve the monitoring ability of ship's greenhouse gas emissions.

1.2 Research Status

1.2.1 Maritime Traffic Status

Fujii (1971) believes that maritime traffic research is mainly based on the results of traffic surveys to make a quantitative statement of the overall behavior of ships and maritime traffic. Tasseda et al. (2014) divided the TAZ of the eight port areas in Tokyo Bay, and used the information of the ship's terminal port in the AIS data to analyze which port area is the most attractive. Minami et al. (2014) discussed the relationship between ship density and ship accidents in Tokyo Bay. Shirai et al. (2016) used the AIS data in Tokyo Bay from March 5 to March 9, 2013, the characteristics of marine traffic in Tokyo Bay are analyzed, the relationship between ship size and ship speed is discussed, and DCPA and TCPA are used to identify ships in Tokyo Bay Possible locations. Altan et al. (2017) selected the Istanbul Strait as a sample for the study of the maritime traffic situation, and 13 regions within the Strait were divided for comparative study. It counts the number of southbound and northbound ships, the number of ships passing through the strait each month, the density of ships in each zone, the type and size of ships passing through the strait, speed distribution and course distribution. Huang et al. (2019) established a traffic model to explore the relationship between traffic flow and traffic density.

1.2.2 Greenhouse Gases (GHG) of Ship

With global warming gets worse, every government emissions management of greenhouse gases (GHG) is becoming more meticulous. GHG emissions from the vessel are receiving increasing attention, especially in port cities, vessels have become a major source of pollution emissions. Carbon dioxide emissions are the most important greenhouse gas species. Aiming at this fact, most of the government have introduced carbon dioxide emission tax rates to reduce GHG emission of vessels.

AIS data stores detailed real-time trajectory data of the vessel, we can use longitude, latitude, SOG (speed over ground), COG (course over ground) to accurately and real-time estimate the carbon dioxide emissions during the vessel sailing. Many studies have been carried out to estimate vessel inventories using AIS data. Sérgiomabunda et al. (2014) estimated ship emission inventory near the strait of Gibraltar, Winther et al. (2014) implemented emission inventory estimation in the artic though S-AIS (Satellite Automatic Identification System), Smith et al. (2015) implemented full-scale vessel emission inventory analysis using AIS data, Coello et al. (2015) estimated emission inventory from UK fishing fleet, and Yao et al. (2016) estimated vessel emission inventories in estuary of the Yangtze River. Wang et al. (2019a) used STEAM2 to estimate fuel consumption of the vessel, but wave and wind resistance were not considered and used the average speed that insufficient accuracy. Kim et al. (2019) considered that accuracy of calculation can be improved by solving the problem that unstable AIS data interval.

1.2.3 Ship Trajectory Prediction in Deep Learning

The traditional method like that Laxhammar et al. (2009) proposed a Gaussian mixture model that combines multiple probability distributions to model vessel trajectories, its disadvantage was to determine the number and quantity of Gaussian components, few quantities of Gaussian components were difficult to describe the vessel trajectory, and too many were prone to overfitting, Zhao et al. (2012) proposed an improved Kalman filter algorithm with system noise estimation to predict vessel trajectory, this method of setting parameters was complicated and difficult to reproduce. Traditional trajectory prediction methods were linear prediction methods, and some methods require expert knowledge to construct the kinematics equations of the vessel. In addition, the marine environment had a great influence on the trajectory of vessels, which made research difficult.

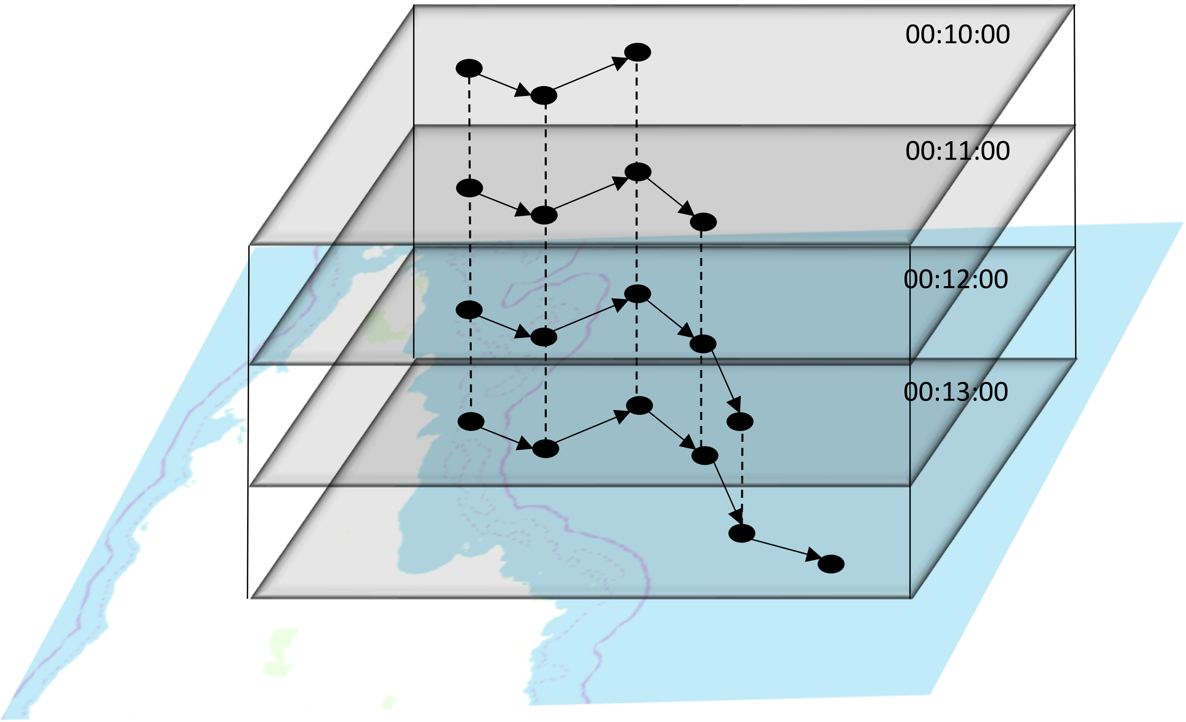


Figure 1.3 Ship Trajectory Prediction Model

Generally, vessel trajectory data can be represented as a set of multi-dimensional spatial-temporal sequences $\{(p_1, a_1, t_1)$,(p_2, a_2, t_2),(p_3, a_3, t_3),…,$(p_n, a_n, t_n)\}$, where p_i is the position (longitude, latitude), a_i is the data attribute (SOG, COG), and t_i is the recording time of the AIS data. Refer to Figure 1.3.

There are complex linear relationships for different spatial-temporal data, so we can use the neural network method to fit the relationship between the historical trajectory data and the next time data. Deep learning is an algorithm that uses artificial neural networks as a framework to perform characterization learning on data, it can automatically learn the information in the data, and can learn the essential characteristics of the data set from the small sample set. Compared to machine learning, which requires manual extraction of data features, but deep learning uses machines to automatically extract data features, which reduces the manual extraction workload and improves prediction efficiency.

Some Scholars have applied deep learning to AIS data prediction. Quan et.al. (2018) established the recurrent neural network-long short-term memory(RNN-LSTM) model to predict vessel trajectory. Nguyen et al. (2018) used the Sequence-to-Sequence model to predict vessel trajectory, encoder and decoder used LSTM model. However, vessel trajectory data was Spatial-temporal sequence data that had both temporal and spatial correlation. SOG of the vessel will affect the change in the distance of the vessel, and the COG of the vessel would affect the position of the vessel.

Therefore, some scholars proposed to use the convolutional layer to extract the spatial characteristics of the trajectory data, and then used the LSTM model to analyze the temporal characteristics. This also achieved precise prediction results.For example, Ljunggren et al. (2018) used convolutional networks of different sizes to learn Spatial-temporal sequences in trajectory features, separate the different features in the data, and finally synthesize a feature to perform trajectory prediction. Wang et al. (2019b) proposed a model of the convolutional neural network combined with LSTM, studying the spatial structure and temporal characteristics of vessel trajectory data simultaneously, the prediction effect was better than using the convolutional neural network alone, but the prediction error was not overwhelming compared to the LSTM model. So, the LSTM model is currently popular in the vessel trajectory prediction applications.

1.3 Research Framework

Chapter 1 mainly discusses the background and significance of this study from the three aspects of marine traffic management, greenhouse gas emissions, ship trajectory and status prediction, and at the same time sorts out and discusses relevant research literature.

Chapter 2 mainly discusses the basic knowledge of AIS working mechanism, ITTC algorithm, interpolation algorithm and deep learning algorithm.

Chapter 3 selects ships in the Straits of Malacca (mostly oil and gas ships) as samples, first discusses the basic conditions of the Straits of Malacca, and secondly discusses the COG, SOG, draught distribution and ship traffic flow in the sea, Finally, discussed the GHG emission status of LNG ships in the sea.

Chapter 4 selects single ship samples to analyze the SOG, COG, draught and CO2 emission status. Secondly, it uses RNN and LSTM to predict the ship trajectory and ship status.

Chapter 5, Conclusion.

The research subject of this study follows Figure 1.4, the algorithm flow diagram used in this study is shown in Figure 1.5.

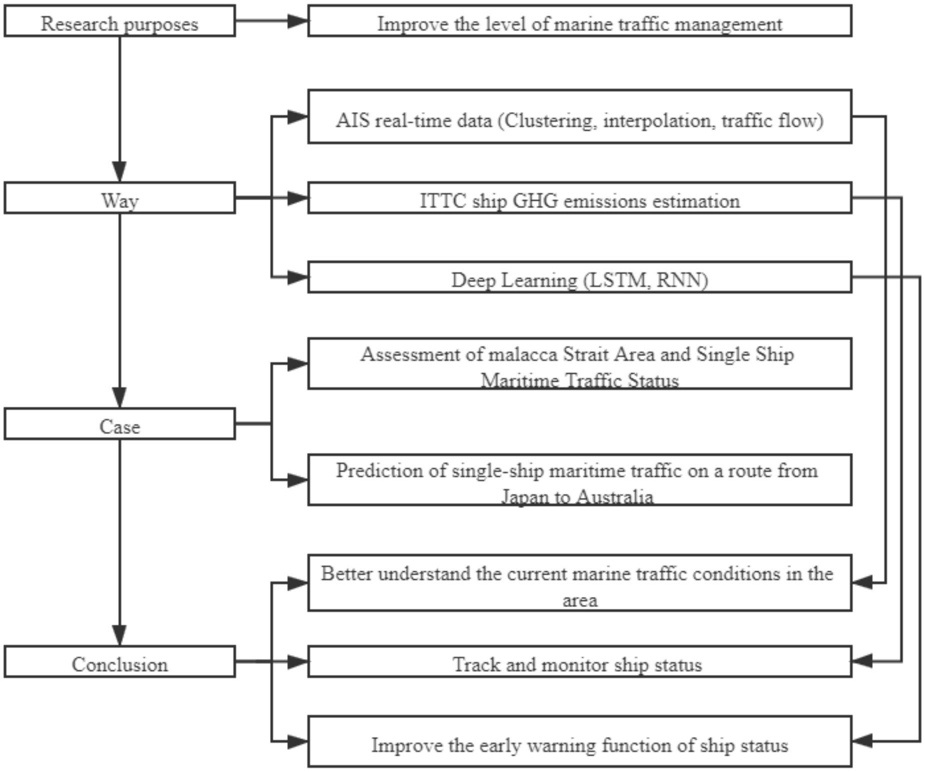

Research purposes
Improve the level of marine traffic management
AIS real-time data (Clustering, interpolation, traffic flow)
Way
ITTC ship GHG emissions estimation
Deep Learning (LSTM, RNN)
Assessment of malacca Strait Area and Single Ship Maritime Traffic Status
Case
Prediction of single-ship maritime traffic on a route from Japan to Australia
Better understand the current marine traffic conditions in the area
Conclusion
Track and monitor ship status
Improve the early warning function of ship status

Figure 1.4 Research Subject

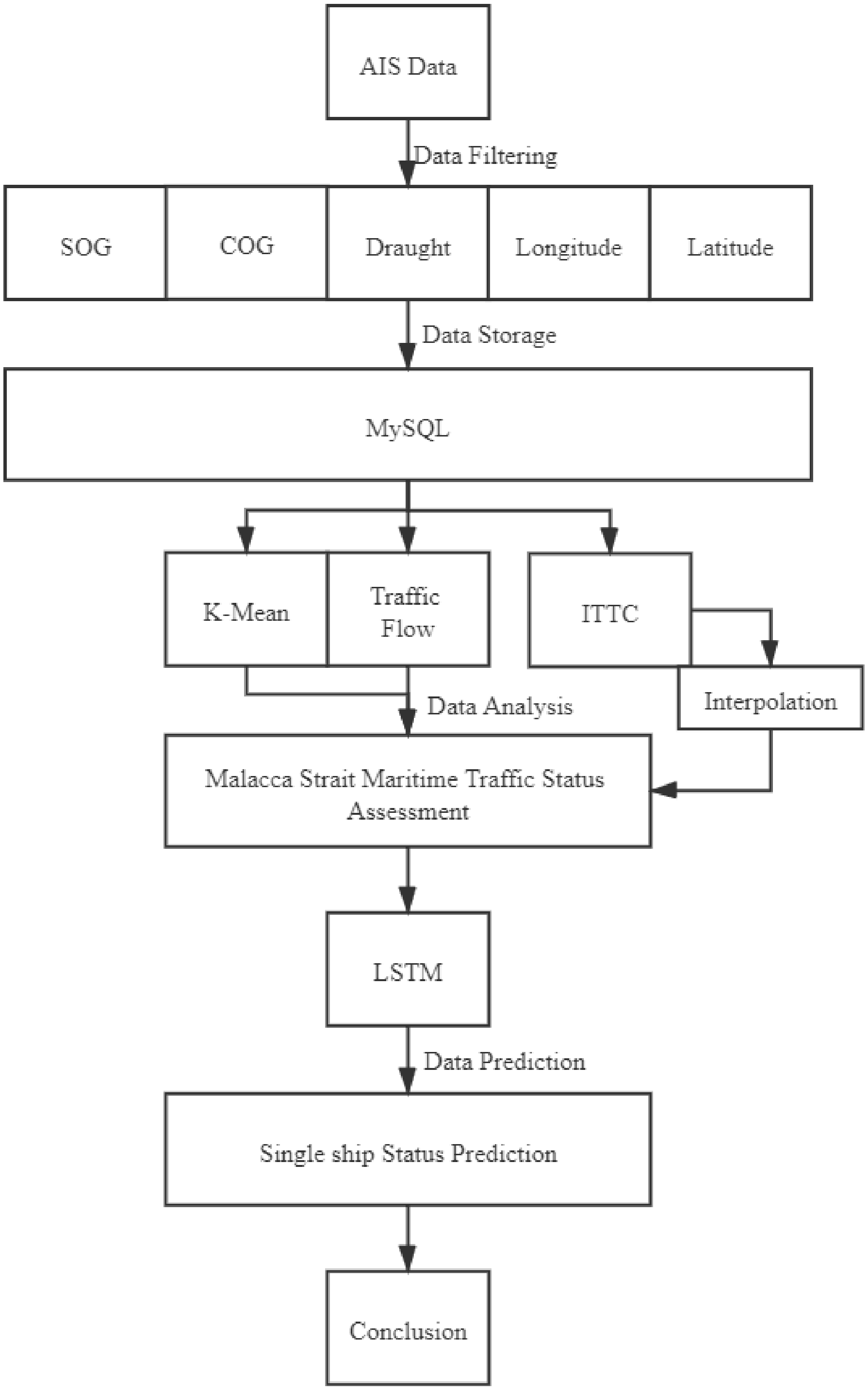


Figure 1.5 Algorithm Roadmap

2. Research theoretical basis

2.1 AIS ((Automatic Identification System)

The General Automatic Identification System (AIS) is a new type of ship collision avoidance system. The shipborne AIS transceiver is a signal transcribing device equipped on the ship. On the one hand, the device dynamically broadcasts the dynamic and static information of the ship collected and manually placed by the sensor, and on the other hand, captures the dynamic and static information of other ships around. In order to realize the real-time grasp of the surrounding marine environment by ships. The spaceborne AIS signal reconnaissance system, through the assembly of low-orbit satellites, can receive the AIS signals of ships within hundreds of nautical miles or even thousands of nautical miles and download them to the ground station receiving system, which can realize the tracking of ship information in the country's surrounding waters and even the global waters. Based on the above advantages, the spaceborne AIS system is currently receiving great attention from various countries.

AIS data characteristics. AIS data belongs to trajectory data. Refer to Figure 2.1, which has the characteristics of big data: large number, real-time performance, and diversity. AIS data has the following characteristics due to factors such as equipment standards, sampling frequency, transmission effects, and storage methods:

1) Space-time sequence. AIS data has a sequence of position, time and other information, including the spatiotemporal dynamics of the object;
2) The frequency of reporting is different. Because the frequency of AIS data reporting is related to changes in ship speed and course, the difference in reporting interval is significant, which increases the difficulty of trajectory data analysis;
3) Poor data quality. Due to the impact of AIS equipment failure rate, reporting accuracy, communication methods and data processing methods, there is no absolute guarantee for AIS data quality;

The accuracy of AIS data depends on the accuracy of the ship's positioning system, the correct input of information by the crew, the encoding and transmission of information by the sending end, and the reception and decoding of information by the receiving end. In addition, the self-checking and error correction capabilities of AIS data are weak. These system design and human factors will cause the received AIS data to be incorrect and brought into the maritime management system. This erroneous information will interfere with maritime supervision, lead to misjudgment of the maritime traffic situation, and cause trouble to the analysis of maritime traffic characteristics based on AIS data. Therefore, the AIS data needs to be cleaned and then analyzed.

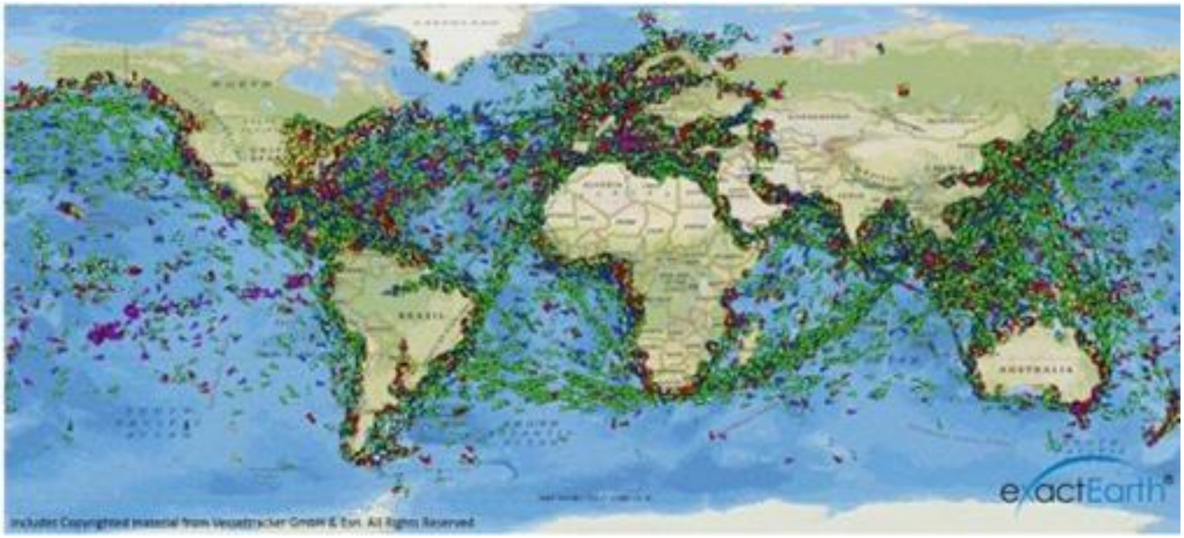

exactEarth
Includes Copyrighted material from Vesseltracker GmbH & Esri. All Rights Reserved

Figure 2.1 exactAIS Global View

Source: exactEarth

There are many sources of AIS data sources. The data sources in this study come from exactEarth and are presented in CSV file format, including attributes such as sailing time, ship name, longitude, latitude, speed, course, etc. AIS data is divided into two categories, one is static Information, one is dynamic information. Static information usually refers to information that is input during installation or relocation of AIS equipment and does not need to be changed frequently. It usually includes maritime mobile service identification code, ship name and paging number, IMO code, captain and ship width, etc. The specific content is shown in the table 2.1 shown.

Table 2.1 Static Information

Information name	Type	Input method	Input time	Update time
MMSI	String	Manual input	Device installation	When the sale of the ship is transferred
Ship name and call sign	String	Manual input	Device installation	When the ship is renamed
IMO	String	Manual input	Device installation	Never change
Length and width	String	Manual input	Device installation	When changing
Ship type	String	Manual input	Device installation	When changing

Dynamic information refers to the ship motion parameters that are automatically updated by sensors. Such information is automatically updated at a certain period, such as ship position, time, COG, and SOG. However, dynamic information such as the ship's draft, destination and estimated time of arrival need to be manually entered before the ship sails.

With the rapid development of information technology and the wide application of AIS equipment, AIS data shows a geometric growth trend. Faced with such large-scale AIS data, traditional data storage and analysis methods are far from meeting the requirements, and the emergence of distributed technology provides a good solution. In current academic research and enterprise applications, Hadoop is used as a widely distributed computing framework, which provides a good platform for distributed storage and calculation of massive data. Hadoop can run on a cluster of cheap computers to achieve distributed processing of large-scale data. With the widespread application of Hadoop, it has also exposed high latency and does not support iterative calculations. Spark is a rising star. With its in-memory computing, low latency, iterative support, and effective inheritance of Hadoop, Spark has developed rapidly since its introduction and has been widely used in various industries.

This study selects MySQL database and Python programming statements to process AIS data, and the performance can also meet the research needs. The AIS data preprocessing process is shown in Table 2.2.

Table 2.2 AIS Data Preprocessing Pseudo Program Code

```
def data_clean(y):                                        1.Enter the target AIS data
    for x in y:                                           into the program
        if y.x = y.x, y.x = null values then              2. Determine if data has null
            y.x.duplicated()                              and duplicate values
            y.x.dropna()
        else
            If y.x.sog < 0 kn, y.x.sog > 30 kn            3. Determine whether there
              y.x.cog <=0°  , y.x.cog >=360°              are abnormal values in sog,
              y.x.draught <0 m, y.x.draught >=30 m then   cog, draught
                y.x.drop()
             else
                Set up IMO and time as the main key       4. Establish ship trajectory
                Store x to MySQL                          database
     return y
```

2.2 Interpolation Model

At the same time, the observation of AIS data can be found that there are still missing AIS data. This is actually a very common phenomenon. There are many reasons for the lack of AIS data, such as AIS equipment aging, AIS data transmission system failure, equipment network Connection failure, etc. The lack of data will affect the accuracy of GHG emissions estimation and trajectory prediction later, so it is necessary to deal with the missing values. The interpolation process can refer to Figure 2.2.

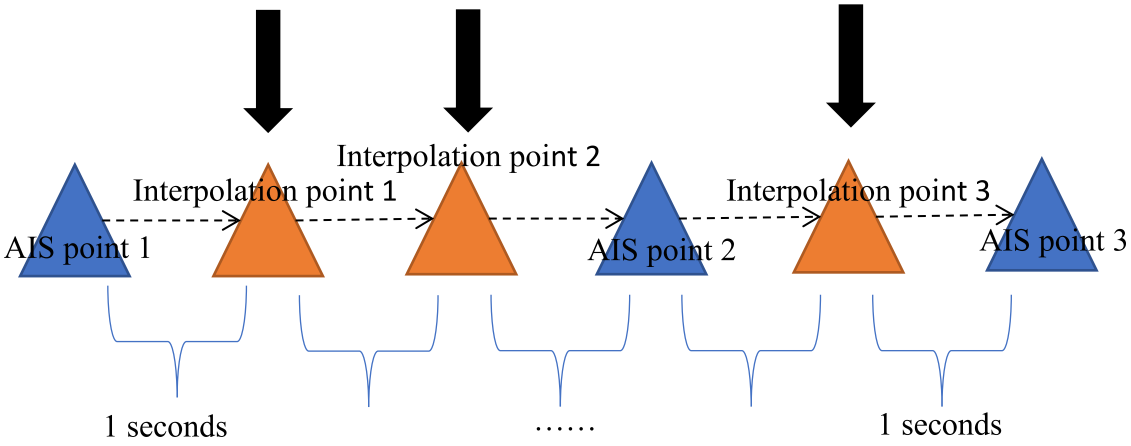


Figure 2.2 Interpolation Diagram

2.2.1 Interpolation Definition

Let the function $y = f(x)$ be defined in the interval $[a, b]$ and be known the values at the points $(a \leq x_0 \leq x_1 \leq \cdots \leq x_n \leq b)$ are known to be $\mathrm{y}_0, y_1, \cdots y_{n-1}, y_n$. If there is a simple function $p(x)$, make $p(x_i) = y_i$, $(\mathrm{i} = 0, 1, \cdots, \mathrm{n})$, then call $p(x)$ as the interpolation function of $f(x)$, the method of finding the interpolation function $p(x)$ is called the interpolation method.

2.2.2 Interpolation Types

The interpolation method is divided into one-dimensional interpolation, two-dimensional interpolation and three-dimensional interpolation, one-dimensional interpolation is used for data repair, two-dimensional interpolation is used for image repair, and three-dimensional interpolation is used for spatial repair. This study only needs to repair the data, so only introduced the 1-dimensional interpolation method.

1) LaGrange interpolation method
   LaGrange interpolation can give a polynomial function that just passes through several known points on the two-dimensional plane. If the actual ship trajectory data is processed, there are several known points. Lagrange interpolation method can obtain a polynomial, which can contain all known points, and finally obtain the interpolation point data from the polynomial. This method is concise and simple in the implementation of the algorithm, but due to the discreteness of the AIS data and the irregularity of continuous changes, the accuracy of this interpolation method is not high; and the amount of AIS data is large, according to Lagrange interpolation When constructing an interpolation polynomial, the more interpolation points, the higher the degree of interpolation polynomial. When the interpolation degree is too high, the Runge phenomenon occurs (it can be approximated within a certain range, but the closer to the interpolation endpoint, the error is bigger the phenomenon), the interpolation result obtained in this way will be very error-prone, and will also produce false fluctuations, which does not have the shape-preserving effect.

2) Cubic spline interpolation method
   Cubic spline interpolation (Cubic interpolation) is the process of obtaining a curve function group by solving the three bending moment equations. The result is a smooth curve through a series of shape points. The cubic interpolation spline curve makes a reasonable compromise between flexibility and calculation speed. Compared with higher-order splines, cubic interpolation splines require less calculation and storage, and are more stable. Compared with quadratic interpolation splines, cubic interpolation splines are more flexible when simulating arbitrary shapes. This avoids the Runge phenomenon that occurs when using higher-order polynomials, so spline interpolation has become popular. In addition to the third-order spline interpolation method, there are first-order spline interpolation method (Slinear) and second-order spline interpolation method (Quadratic).

3) Special value interpolation method
   The zero interpolation method is to make the missing value 0, the nearest value interpolation method is to make the missing value equal to the value closest to it, the future interpolation method is to make the missing value equal to its next value, and the past interpolation method is to make the missing value The value is equal to the previous value. This type of method has poor smoothness and is suitable for small-scale data interpolation operations.

4) Machine learning interpolation method
   We can fill in missing values through machine learning models. The main algorithms include support vector machine, Newton search algorithm, KNN algorithm, neural network, etc.

The effect of different interpolation methods (1--dimensional) is shown in Figure 2.3.

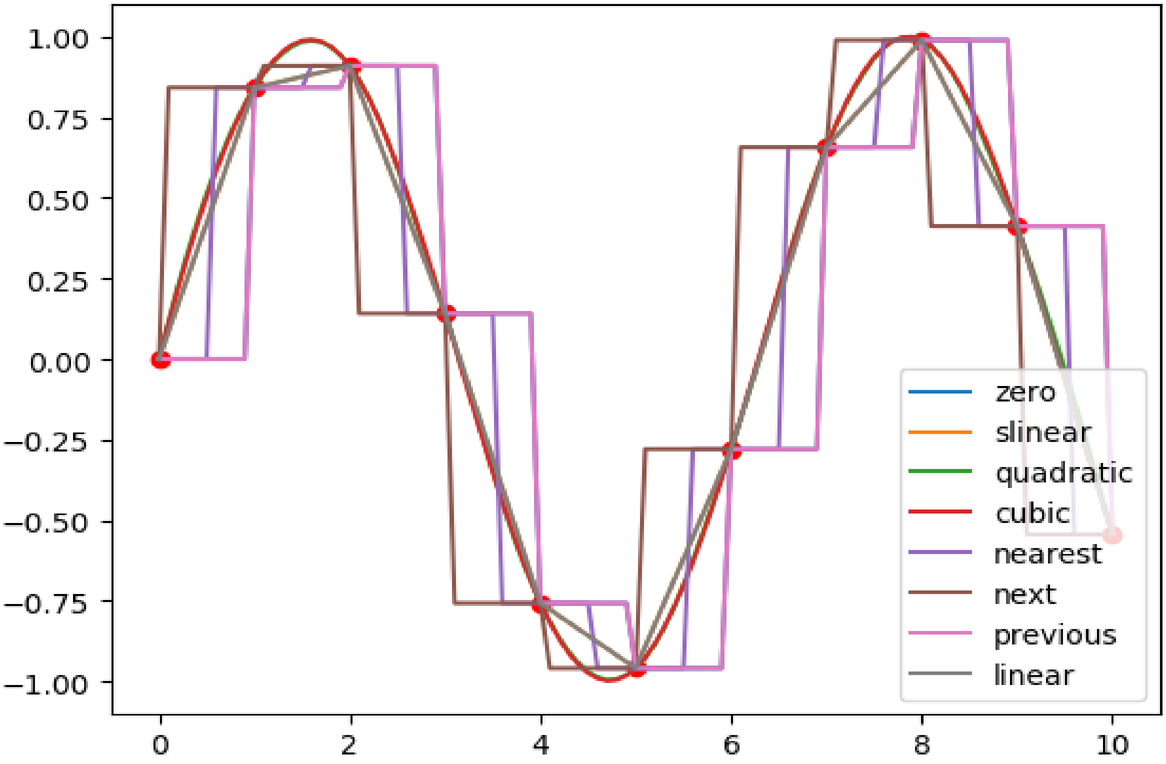


Figure 2.3 1-Dimensional Interpolation Effects

2.2.3 Cubic Spline Interpolation Method

Due to incorrect operation of the AIS system by shore and vessel personnel, information transmission failure between AIS and shore base, subjective and objective factors such as the random failure of the AIS system itself or the problem of artificial improper maintenance. Therefore, we select method of Cubic spline interpolation, smooth the vessel trajectory and repair the trajectory to obtain more accurate and complete vessel trajectory information.

We suppose that interval of trajectory data is $[a, b]$, divide$[a, b]$ into n intervals, like$[(x_0, x_1), (x_1, x_2), \cdots, (x_{n-1}, x_n)]$, $x_0 = a, x_n = b$, the function expression for each interval is $S(x)$. Cubic spline means that the curve of each interval is a cubic equation $S_i(x)$ and meet interpolation conditions, $S(x_i) = y_i$. Meet the condition of smooth curve that $S_i(x)$, $S_i{}'(x)$, $S_i{}''(x)$ are continuous function. Solved equation (Bartels et al. (1998)) is as follows:

$$S_i(x) = a_i + b_i(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^3 \tag{1}$$

$$S_i{}'(x) = b_i + 2c_i(x - x_i) + 3d_i(x - x_i)^2 \tag{2}$$

$$S_i{}''(x) = 2c_i + 6d_i(x - x_i) \tag{3}$$

Where $S_i(x)$: Cubic spline model expression, a_i, b_i, c_i, d_i: Parameters to be solved.

According to $S_i(x)$ must meet interpolation conditions, $S(x_i) = y_i$, and equations (1), (2) and (3), we can get the equation as follows:

$$a_i = y_i \tag{4}$$

$$h_i = x_{i+1} - x_i \tag{5}$$

$$a_i + b_i h_i + c_i h_i{}^2 + d_i h_i{}^3 = y_{i+1} \tag{6}$$

According to continuous function conditions,

$$\mathrm{S}_i''(x_{i+1}) = \mathrm{S}_{i+1}''(x_{i+1}) \tag{7}$$

$$\mathrm{S}_i'''(x_{i+1}) = \mathrm{S}_{i+1}'''(x_{i+1}) \tag{8}$$

we can get the equation as follows:

$$b_i + 2h_i c_i + 3h_i^2 d_i = b_{i+1} \tag{9}$$

$$2c_i + 6h_i d_i = 2c_{i+1} \tag{10}$$

$$d_i = \frac{m_{i+1} - m_i}{6h_i}(m_i = 2c_i) \tag{11}$$

Use equation (4), (10), (11) to input equation (6), we can get the equation as follows:

$$b_i = \frac{y_{i+1} - y_i}{h_i} - \frac{h_i}{2} m_i - \frac{h_i}{6}(m_{i+1} - m_i) \tag{12}$$

Use equation (4), (10), (11), (12) to input equation (9), we can get the equation as follows:

$$h_i m_i + 2(h_i + h_{i+1})m_{i+1} + h_{i+1}m_{i+2} = 6\left[\frac{y_{i+2}-y_{i+1}}{h_{i+1}} - \frac{y_{i+1}-y_i}{h_i}\right] \tag{13}$$

We build linear equations with m as the unknown($m_0 = 0, m_n = 0$):

$$\begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & 0 \\ h_0 & 2(h_0+h_1) & h_1 & 0 & \cdots & 0 \\ 0 & h_1 & 2(h_1+h_2) & h_2 & \cdots & 0 \\ 0 & 0 & h_2 & 2(h_1+h_2) & h_3 & \vdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & 0 & 0 & h_{n-2} & 2(h_{n-2}+h_{n-1}) & h_{n-1} \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} m_0 \\ m_1 \\ m_2 \\ m_3 \\ \vdots \\ m_n \end{bmatrix}$$

$$= 6\begin{bmatrix} 0 \\ \frac{y_2-y_1}{h_1} - \frac{y_1-y_0}{h_0} \\ \frac{y_3-y_2}{h_2} - \frac{y_2-y_1}{h_1} \\ \vdots \\ \frac{y_n-y_{n-1}}{h_{n-1}} - \frac{y_{n-1}-y_{n-2}}{h_{n-2}} \\ 0 \end{bmatrix} \tag{14}$$

We can calculate $m_0, m_1, \cdots, m_n$ from equation (14) and use it to calculate a_i, b_i, c_i, d_i and know the function expression for each interval to repair the vessel trajectory data.

The advantages of cubic spline interpolation: this method can not only achieve the convergence of the function, but also maintain the smoothness and continuity of the data and reduce the loss of information.

2.3 Deep Learning

2.3.1 Trajectory Data Preprocess

Before using the deep learning model, you need to convert AIS data into trajectory data with time and space attributes and effective.

1) Trajectory Drift
   Trajectory drift refers to the situation where two AIS data points with a small interval time have a large deviation, as shown in Figure 2.4. The deviation of trajectory data affects route analysis and may mislead the results of trajectory prediction. If we do not remove the trajectory drift point, when doing interpolation repair work, it will make the data appear Runge phenomenon, the value of the data will be abnormally high or low somewhere, affecting the quality of the data. This is not conducive to our estimation of ships' greenhouse gas emissions and trajectory prediction operations. If the time interval between points is approximately equal, but the distance between points is too large, it is determined that this point belongs to the trajectory drift point, and we need to remove it.

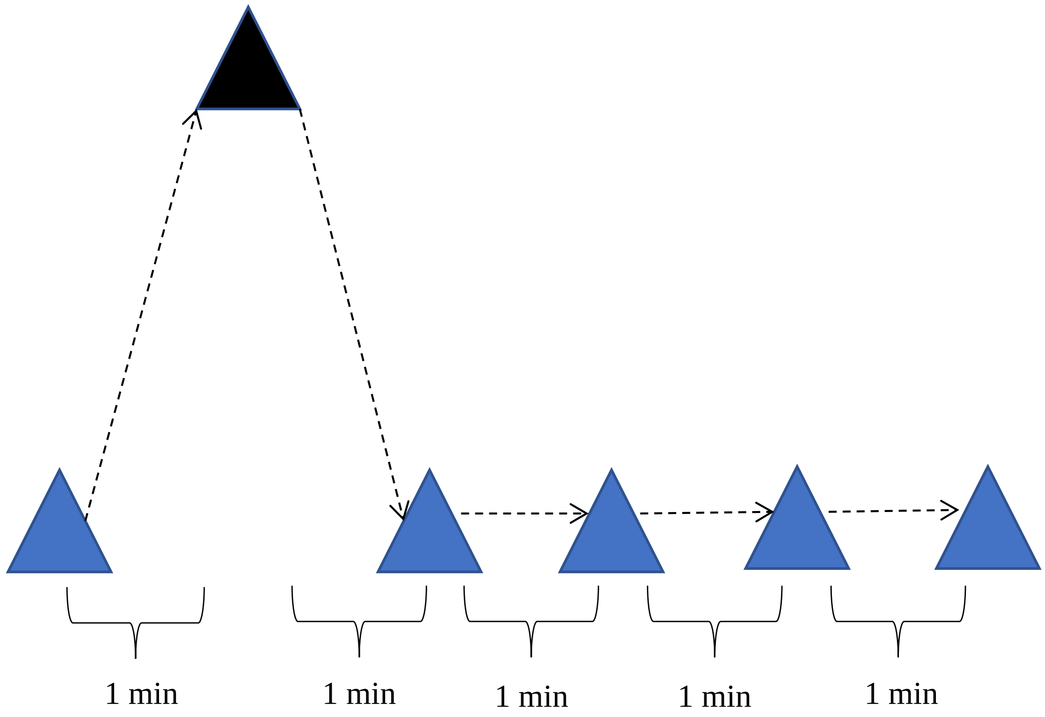


Figure 2.4Trajectory Drift

2) Trajectory Sparse

When dealing with ship trajectories, there is a difference in the amount of data between single ships in a certain area. Some ships have one day of data, and some ships may have only a few points, although the distribution of these points covers the area to be studied, we cannot determine its true trajectory, as shown in Figure 2.5. It is impossible to restore the trajectory using interpolation. Even if the trajectory is restored, the trajectory point will cross the land, refer to Figure 2.6, it needs to be removed.

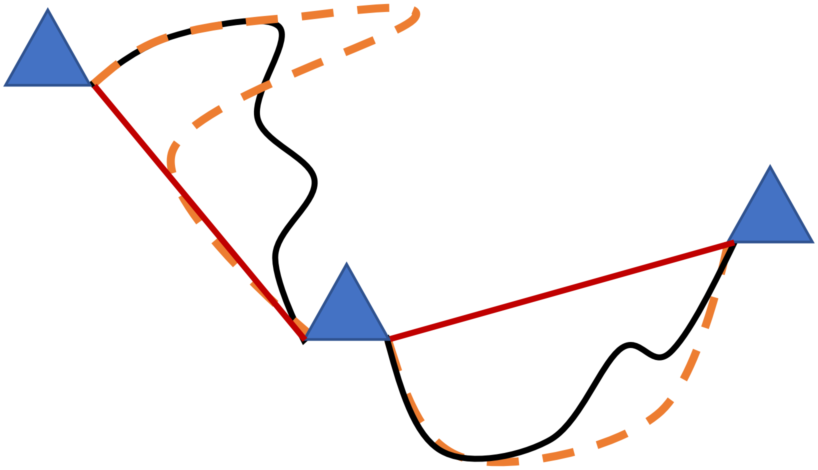


Figure 2.5 Trajectory Sparse

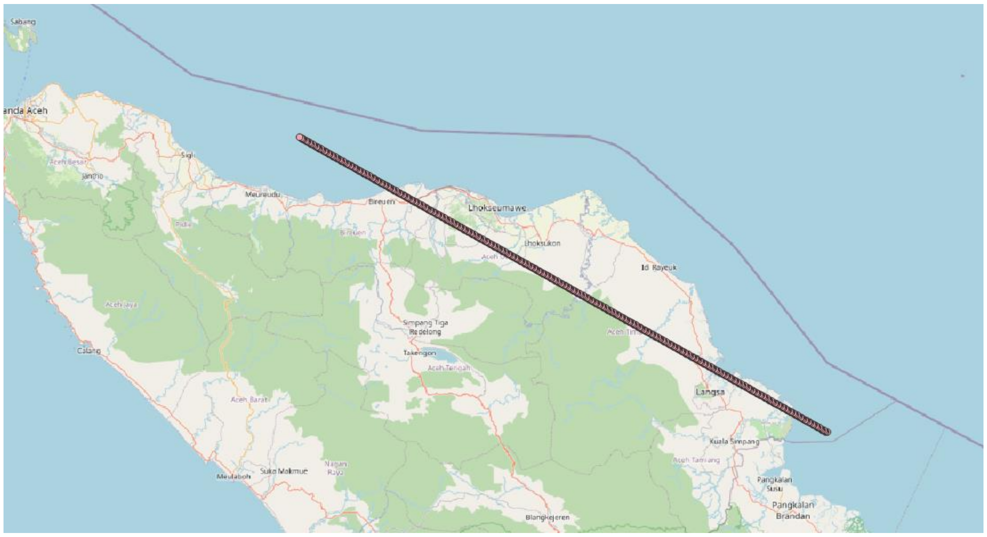


Figure 2.6 The Phenomenon of Trajectory Crossing the Land

2.3.2 Deep Learning Model

BP neural network is the basic neural network model. Rumelhart et al. (1986) founded Error Back Propagation Training (BP), which can effectively train the connection weights and neuron thresholds of multi-layer feedforward neural networks and gave rigorous mathematical derivation and demonstration. People call the multi-layer feedforward network using this algorithm for error correction to be called BP neural network. The BP neural network contains a three-layer perceptron of the hidden layer. It consists of three parts, from left to right are the input layer, hidden layer and output layer.

With the development of technology, the effect of deep learning in the field of neural networks is better than that of BP neural networks. As one of the popular emerging technologies, deep learning has been welcomed and applied by many research scholars. Many excellent models have been proposed one after another. There are currently three types of deep learning models commonly used:

1) Automatic Encoder
   Rumelhart et al. (1986) proposed the concept of Autoencoder in 1986 and used it to analyze high-dimensional complex functions, which promoted the rapid development of neural networks. Figure 2.7 is the structure of the prototype automatic encoder. As shown in the Figure, the automatic encoder is an unsupervised learning algorithm, which can reproduce the network input signal as much as possible through the encoder and decoder. In 2006, Hinton et al. (2006) made some improvements based on the original structure and proposed a deep automatic encoder (DAE). The network pre-trains the unsupervised layer-by-layer greedy training algorithm to extract the characteristics of high-dimensional complex input data, and then optimizes and adjusts the parameters of the network model through the back-propagation algorithm. This model largely avoids the problem that the traditional BP algorithm is prone to fall into local small values. The autoencoder can improve the prediction accuracy to a great extent by combining the features learned by it based on the original features, especially in the classification problem.

2) Convolutional Neural Network
   Hubel et al (1962) proposed Convolutional Neural Network (CNN). It was based on the concepts proposed in biological research on the visual cortex of cat brain. The main advantages of convolutional neural networks are parameter sharing and sparse connections. It can effectively reduce the complexity of the network mode and the number of network parameters, so that we can train it with a smaller training set to prevent overfitting. At present, it is the most widely used in the image field, because

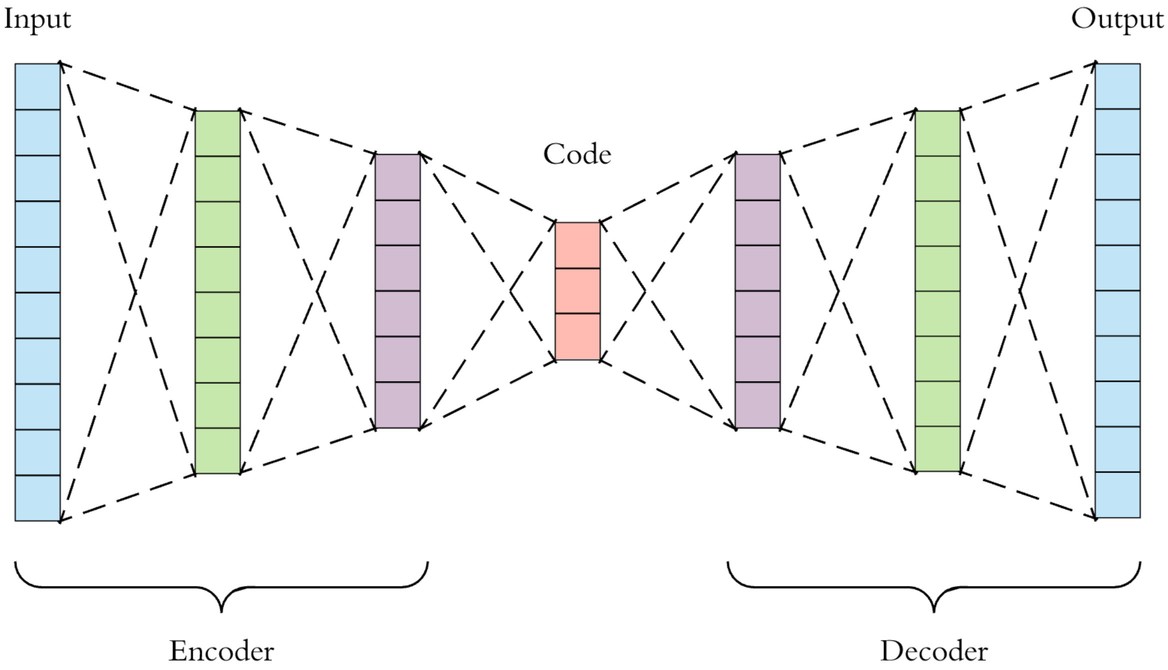


Figure 2.7 Automatic Encoder Structure (Arden Dertat, 2017)

CNN can directly use the image as the input of the network, without the very complicated feature extraction process like the traditional recognition algorithm. Figure 2.7 is a classic CNN network structure. As shown in the Figure, the convolutional neural network includes a convolution layer (Convolution layer) and a sampling layer (Sampling layer). Among them, the main function of the convolutional layer is to extract the convolutional features. The pooling layer immediately follows the convolutional layer to reduce the dimension of the feature map, and to a certain extent, the invariance of the convolutional feature scale can be guaranteed. This can greatly reduce training parameters and reduce the degree of model overfitting. In recent years, convolutional neural networks (CNN) have achieved rich research results in various fields such as image processing, target detection, and semantic segmentation with their powerful feature learning and classification capabilities. In addition, one-dimensional data can also use CNN, usually used in combination with other models, popular models are Conv1D-lstm and Conv1D-rnn.

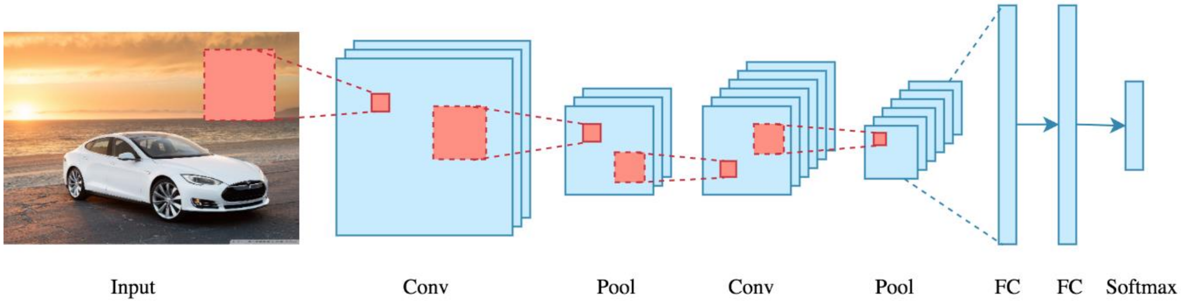


Figure 2.8 CNN Network Structure (Arden Dertat, 2017)

3) Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM)
   Recurrent Neural Networks (RNN) are derived from the Hopfield network proposed by John J Hopfield et al. (1982). Recurrent neural network is a neural network for modeling sequence data, which is mainly used to analyze and predict sequence data. The special feature of the recurrent neural network is that it can quickly process input sequences of arbitrary timing with its internal memory. With its unique advantages, recurrent neural networks have achieved outstanding results in areas such as speech recognition, machine translation, and timing analysis. Figure 2.9 is a typical RNN network structure. It can be seen from the Figure that, unlike the structure of the traditional neural network, the nodes of the RNN hidden layer are connected.

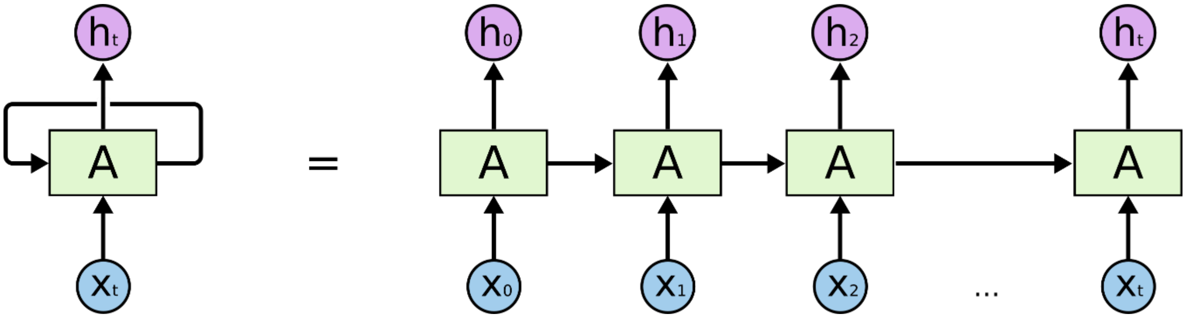


Figure 2.9 RNN Network Structure (Christopher Olah, 2015)

However, with the research of RNN, it is found that its performance decline over a long time series is obvious. The main reason for this phenomenon is that gradient explosion and gradient disappearance are easy to occur during model training, so that when the prediction time span is long, the "memory" of RNN cannot be maintained all the time. Therefore, in practical applications, it is difficult to use RNN to deal with long-distance dependence.

Long Short-Term Memory (LSTM) is a special RNN network structure. It relies on three gate structures to achieve long-term storage of memory and solves the problem of long-distance dependence of RNN networks. With its characteristics, it has made great breakthroughs in natural language processing, machine translation and other fields. Based on the above analysis and comparison, the three commonly used neural network models have their own areas of expertise. Because the ship trajectory data belongs to a longer time series, the position of the ship at the previous moment and the position at the next moment are related, so when analyzing the trajectory sequence, we need to analyze the entire sequence. Therefore, this study chooses the LSTM suitable for processing and forecasting longer time series as the model of trajectory prediction.

The LSTM model (Hochreiter et.al. (1997)) is a variant of RNN. The RNN cannot learn longer histories data, resulting in a gradient decline or even disappearance at further time steps. To solve this problem, LSTM model introduces storage units and unit states to control information transfer based on RNN.

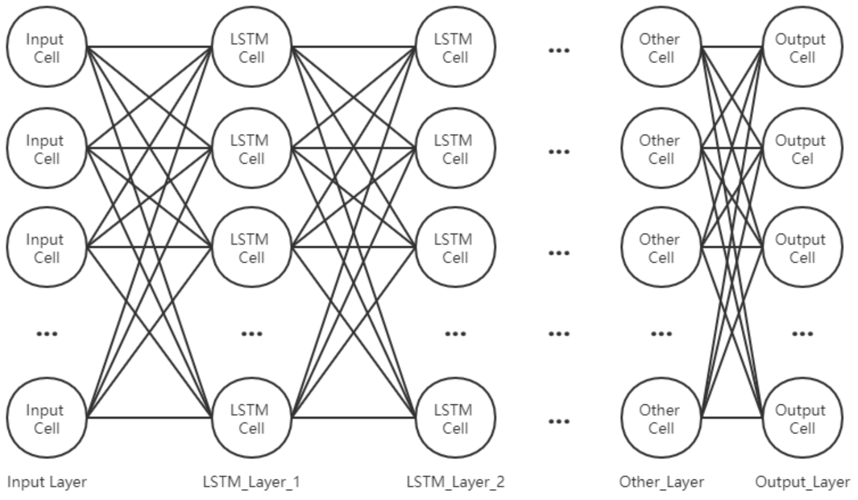


Figure 2.10 LSTM Model Structure

There are four gates (Forget Gate, Input Gate, Update Gate and Output Gate) in the storage unit of LSTM model. The input gate controls the addition of new information. The forget gate can forget the information that needs to be discarded and retain the useful information of the past. The update gate can update data. The output gate causes the storage unit to output only information related to the current time step. These four gate structures perform matrix multiplication and non-linear summing in the memory cells so that the memory does not decay in constant iterations.

As shown in Figure 2.10 is structure of LSTM neural network, it consists of three layers: input layer, hidden layer (LSTM_Layer_1 and Other Layer) and output layer .As shown in Figure 2.11 (a) is the structure of RNN. Its cell contains only one activation function, and the cells are only linked in order. Figure 2.11 (b) is the structure of LSTM. Its cell is more complex than RNN. It needs four gate calculations to output to the next cell. Figure 2.11 (c) is an enlarged view of Figure 2.11 (b).

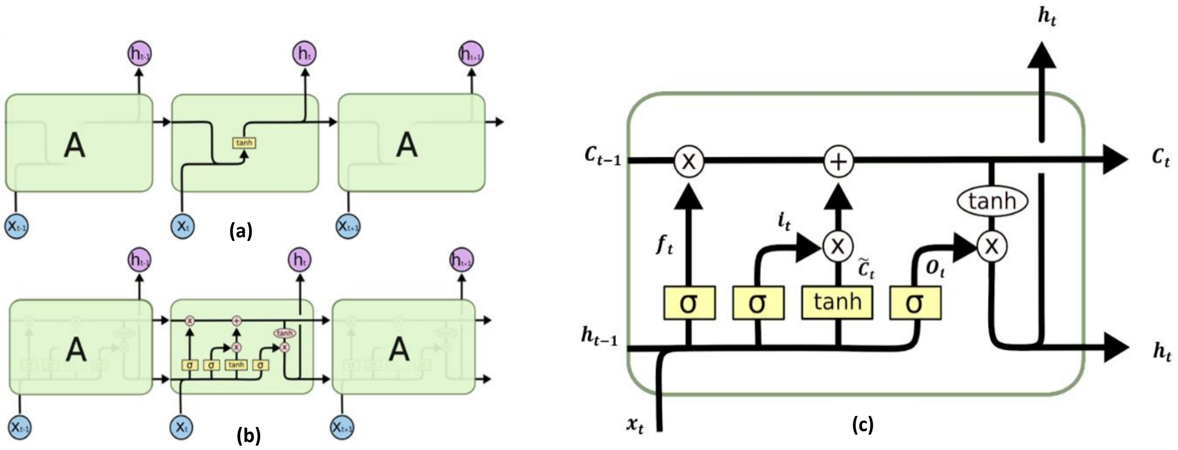


Figure 2.11 LSTM cell Model Structure (Christopher Olah, 2015)

1) Forget Gate: For time t, the state h_{t-1} at the previous time and the current training data x_t can get f_t through the forget gate, the structure is shown in Figure 2.12, the formula is as follows:

$$f_t = \sigma(W_f \cdot [x_t, h_{t-1}] + b_f) \tag{15}$$

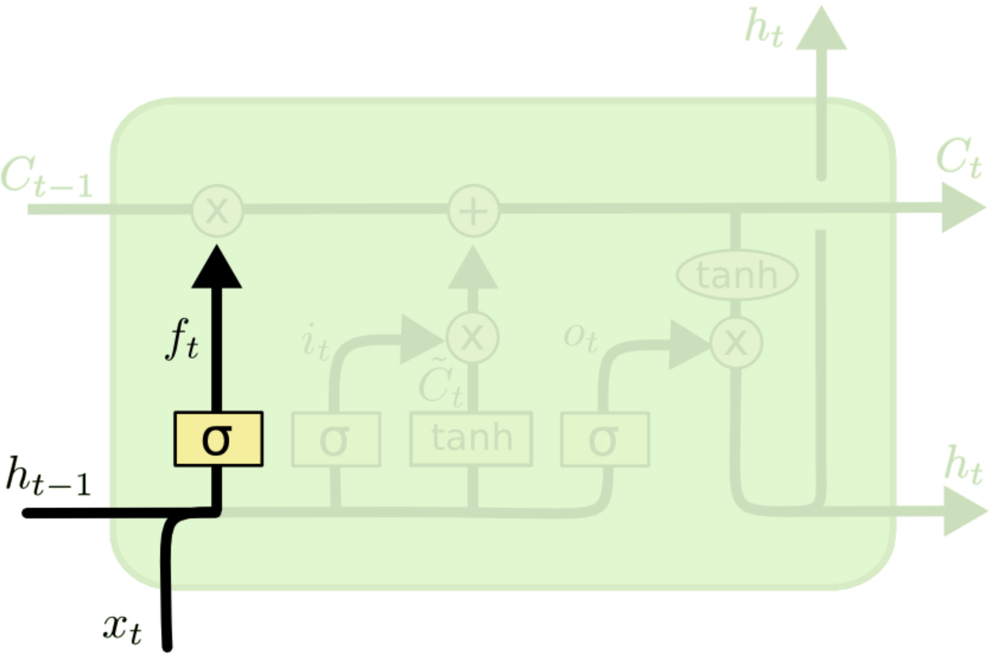


Figure 2.12 Forget Gate Structure (Christopher Olah, 2015)

Input Gate: $\tilde{C}_t$ decides the amount that can be added to the Cell state in the tanh network layer, i_t outputs a number between 0 and 1 through the Sigmoid network layer to decide which status values to update, the structure is shown in Figure 2.13, the formula is as follows:

$$\widetilde{C_t} = tanh(W_c \cdot [x_t, h_{t-1}] + b_c) \tag{16}$$

$$i_t = \sigma(W_i \cdot [x_t, h_{t-1}] + b_i) \tag{17}$$

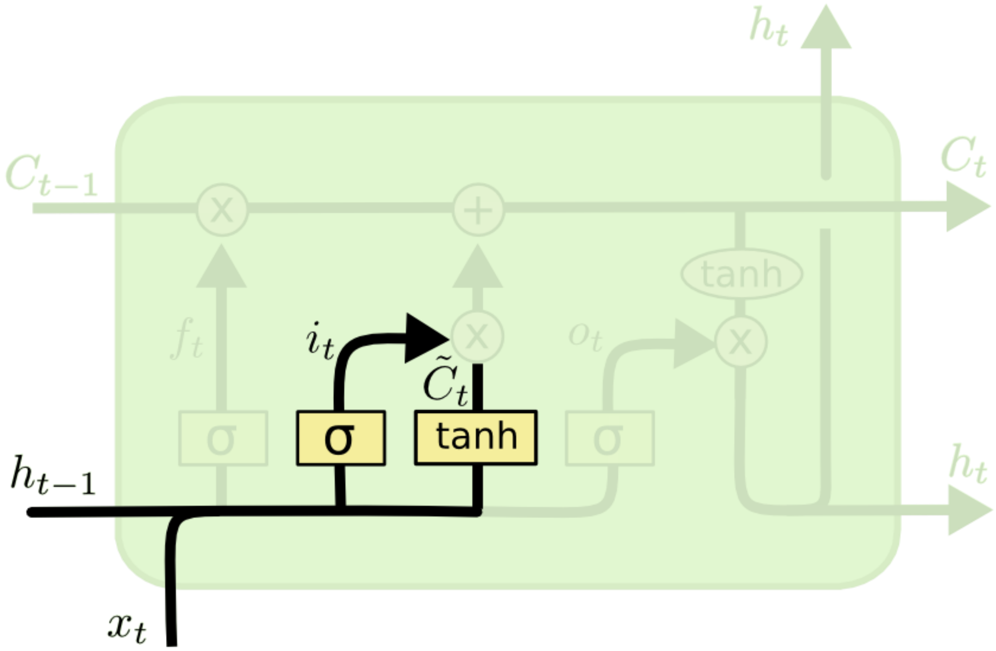


Figure 2.13 Input Gate Structure (Christopher Olah, 2015)

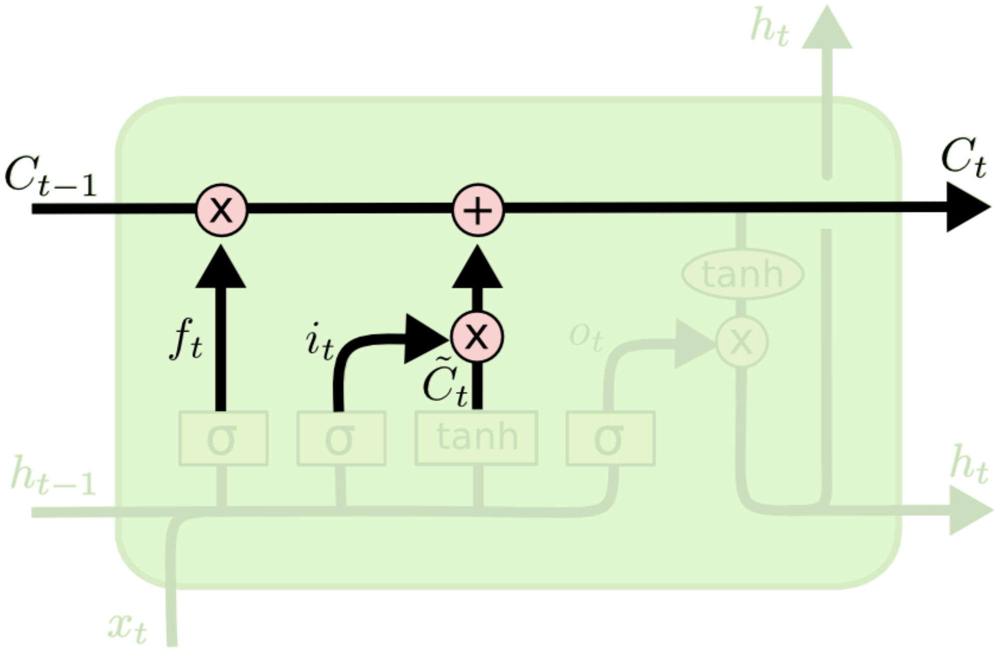


Figure 2.14 Update Gate Structure (Christopher Olah, 2015)

Update Gate: Update old state C_{t-1} to new state C_t, This gate retains long-term and short-term memory in different proportions of the cell, the structure is shown in Figure 2.14, the formula is as follows:

$$C_t = i_t * \widetilde{C_t} + f_t * C_{t-1} \tag{18}$$

Output Gate: The third sigmoid network layer determines which parts of the output cell state, combined with Equation (20) to get the output value of the cell, the structure is shown in Figure 2.15, the formula is as follows:

$$o_t = \sigma(W_o \cdot [x_t, h_{t-1}] + b_o) \tag{19}$$

$$h_t = o_t * \theta_h(C_t) \tag{20}$$

Where:

f_t: Forget gate,

$\widetilde{C_t}$: Input cell,

i_t:Input gate,

C_t: Update gate,

o_t:Output Gate,

σ: sigmoid, Activation function,

W: Weights for different gates,

x_t: Input value at time *t*,

h_{t-1}: Output value at the previous moment,

b: Bias term for different gates.

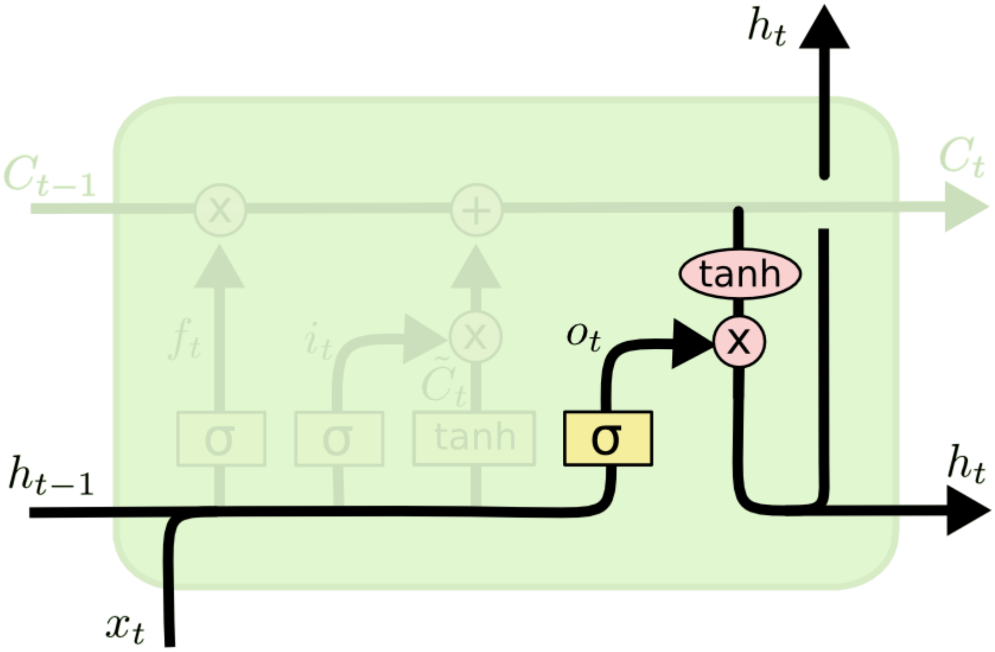


Figure 2.15 Output Gate Structure (Christopher Olah, 2015)

2.4 GHG Emission Estimation Model

In this study, we adopted the ITTC recommended procedure (2017) to estimate vessel resistance. This formula calculates the resistance during navigation based on the ship's real-time SOG and ship dimensions (summer draught, length, width) recorded by AIS, thereby estimating the ship's main engine power at different time and space, and then estimating the main engine's fuel consumption rate based on the ship's manufacturing year. The fuel consumption formula calculates the fuel consumption of the ship, and then combines the emission factors of different fuel types to estimate the ship's greenhouse gas emissions. The calculation process is shown in Figure 2.16.

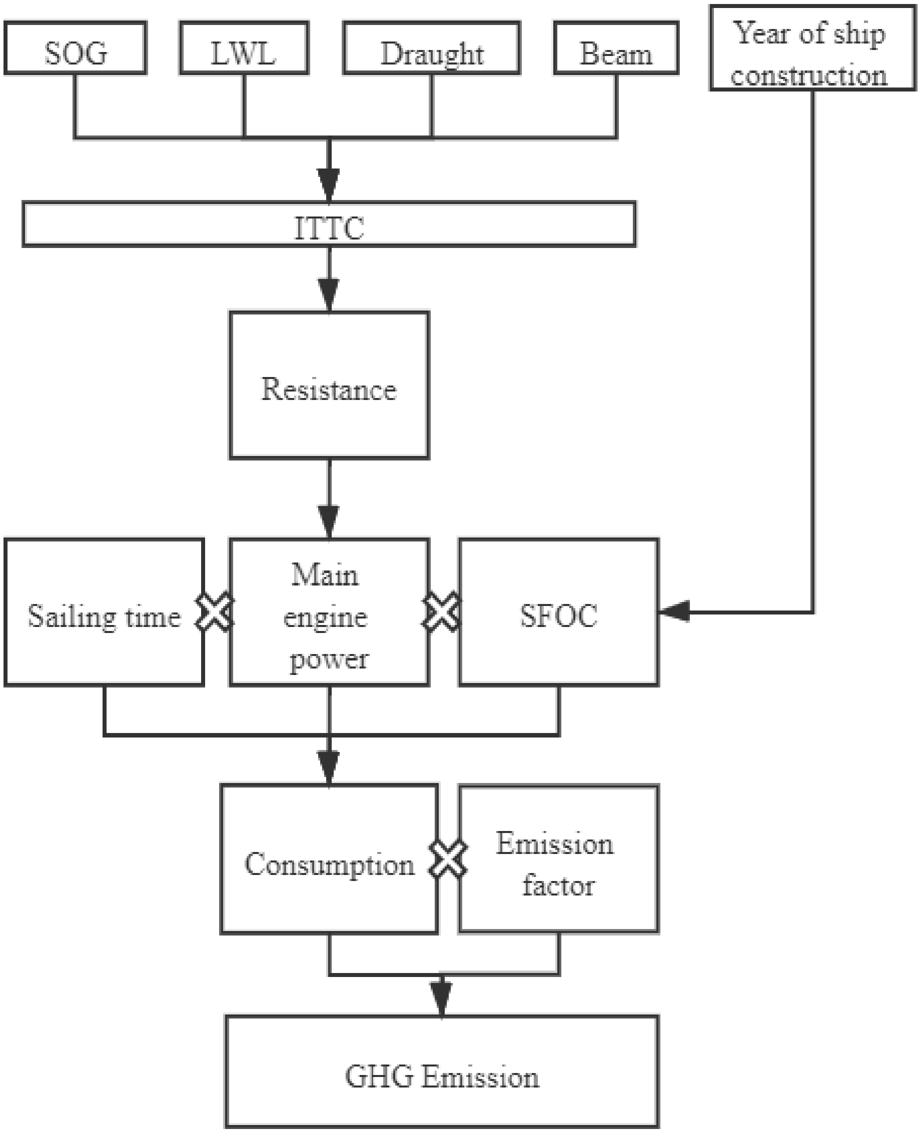


Figure 2.16 Emissions Estimation process

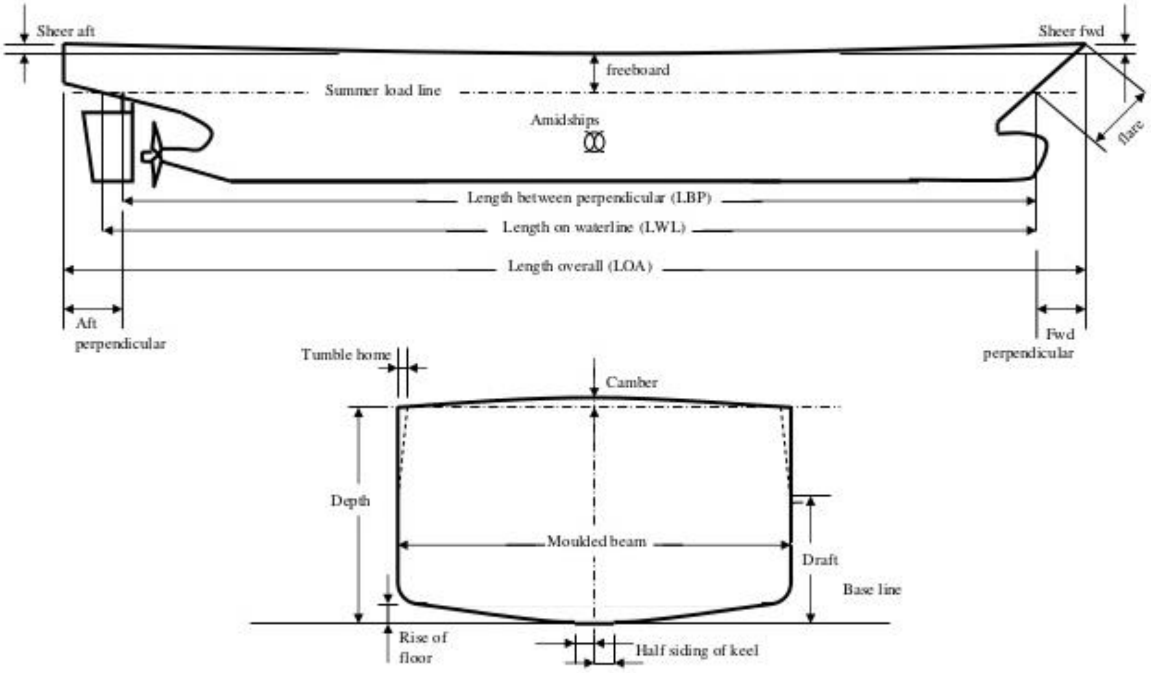


Figure 2.17 Ship Dimensions (Dewan, 2014)

First, we need to confirm the ship dimensions, Length of the waterline, LWL, and the length between perpendiculars, LPP, are one of the most frequently shown up values when calculating the hull resistance of the vessel. The length between perpendiculars is the length between foremost perpendicular and furthest perpendicular. Figure 2.17 illustrate it. LPP is, generally, slightly shorter than the waterline length. Relation between LPP and LWL could be expressed as (MAN Diesel & Turbo, 2011);

$$L_{PP} = 0.97 * \ L_{WL} \qquad (21)$$

Next, it is necessary to derive total resistance first. Total resistance can be denoted as (Molland, A. F. et al. (2016)):

$$R_T = \frac{1}{2} C_T \rho S V^2 \tag{22}$$

Where:

R_T: total resistance,

C_T: total resistance coefficient,

ρ: density of water,

S: wetted surface of the hull,

V: SOG.

C_T, total resistance coefficient, can denoted as:

$$C_T = C_F + C_A + C_{AA} + C_R \tag{23}$$

Where

C_F: frictional resistance coefficient,

C_A: incremental resistance coefficient,

C_{AA}: air resistance coefficient,

C_R: Residual resistance coefficient.

C_F, frictional resistance, of the hull is often occupying some 70-90% of the vessel total resistance for low-speed vessel (Bulk carriers and tankers), and sometimes less than 40% of vessel total resistance for high-speed vessel (MAN Diesel & Turbo, 2011). C_F can be described as (International Towing Tank Conference, 2017);

$$C_F = \frac{0.075}{(log_{10} R_n - 2)^2} \tag{24}$$

Where R_n is the Reynolds number, described as;

$$R_n = \frac{V * L_{WL}}{\varphi} \tag{25}$$

φ is Kinematic viscosity of water. It is defined as;

$$\varphi = \left((43.4233 - 31.38 * \rho) * (t + 20)^{1.72*\rho - 2.202} + 4.7478 - 5.779 * \rho\right) * 10^{-6} \qquad (26)$$

t is temperature of water in degrees Celsius.

C_A, C_{AA}, C_R, for their solution formulas, please refer to Molland, A. F. (2016).

Based on calculated total resistance of vessel, estimating required power when the vessel sailing at speed V in calm sea condition can be calculated by considering the components of propulsion efficiencies. Installed power is the power required to tow vessel with speed V in a calm sea. Service power can be derived from (Molland, A. F. et al. (2016)):

$$P_I = \frac{R_T V}{(\eta_0 \eta_R)} + m \qquad (27)$$

Where:
P_I: Installed power,
η_T: Transmission efficiency,
η_D: Quasi-Propulsive Coefficient,
m: Sea margin.

Fuel oil consumption is calculated by using Specific Fuel Oil Consumption (SFOC) on table 2.3 released in third IMO study. As the vessel engine gets older, an efficiency of the engine goes down and advent of technology make a newer engine more efficient.

Table 2.3 SFOC

Engine age	SSD (IMO)	MSD (IMO)	HSD (IMO)
before 1983	205	215	225
1984-2000	185	195	205
post 2001	175	185	195

Source: Smith et al. (2015)

Table 2.4 Emissions Factors (International Maritime Organization, 2015)

Emissions substance	Marine HFO emissions factor	Marine MDO emissions factor	Marine LNG emissions factor
CO2	3.114	3.206	2.75
CH4	0.00006	0.00006	0.0512
N2O	0.00016	0.00015	0.00011
NOx	0.093	0.08725	0.00783
CO	0.00277	0.00277	0.00783
NMVOC	0.00308	0.00308	0.00301

Combining the data in Tables 2.3 and 2.4, emission estimation model can denote as:

$$E_i = \sum P_I \times SFOC \times emission\ factor \times T \tag{28}$$

Where:

E_i: emission, it is calculated by summing emissions at each trajectory point,

T: time interval of each trajectory point.

3. Marine Traffic Status Assessment

3.1 AIS Data Source Analysis

This study used AIS data of ship provided by exactEarth, as shown in Figure 3.1. Time period was from 1st January 2016 to 30th June 2016, and number of messages was about 50 million, It included 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, SOG, COG, Rate of Turn (ROT), course, navigation(nav) status, source, time, vessel type main, and vessel type sub.

The AIS data used this research was comma-separated values (CSV) form. Every data was divided by day based on Greenwich Mean Time (GMT). A total of 182 days, so the complete data of a vessel was divided into 182 small data sets. This was very inconvenient, so we used MySQL to build a trajectory database, combining 182 small data sets into a large data set. This Trajectory data set mainly included: MMSI, Longitude, Latitude, SOG, COG, draught, Time.

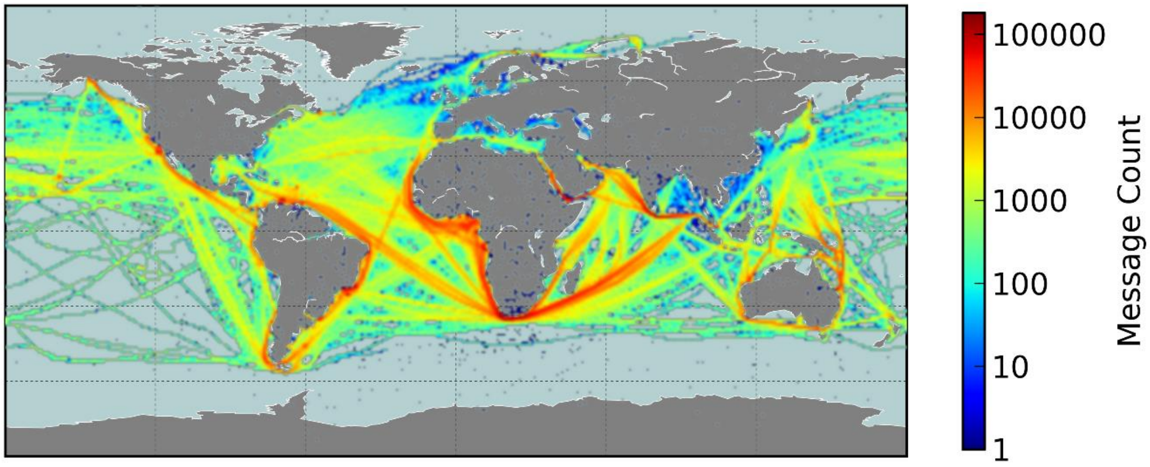


Figure 3.1 Heat Map of AIS Data

Source: exactEarth

In the original data set, we founded that the average time interval of each piece of data in table 3.1 was 520.52 seconds, 25% of the data was 6 seconds, 50% was 17 seconds, 75% was 42 seconds. According to Figure 3.2, we also founded that more than 90% of the data had a time interval of more than 30 minutes. Only about 8% of the data had a time interval of 2 seconds. It should be noted that the total amount of raw data was 51,705,694. Looking at this ratio on a single vessel, the accuracy of the estimated vessel carbon dioxide emissions was not enough. It may because AIS data collected through satellite shows longer data collecting interval when the vessel was sailing areas with high traffic compare to areas with less traffic.

Table 3.1 Summary Statistics of AIS Data interval

	Hour	Minutes	Seconds
Count (number)	9,072,291	9,072,291	9,072,291
Mean	0.14	8.68	520.52
25%	0.00	0.10	6.00
50%	0.00	0.28	17.00
75%	0.01	0.70	42.00

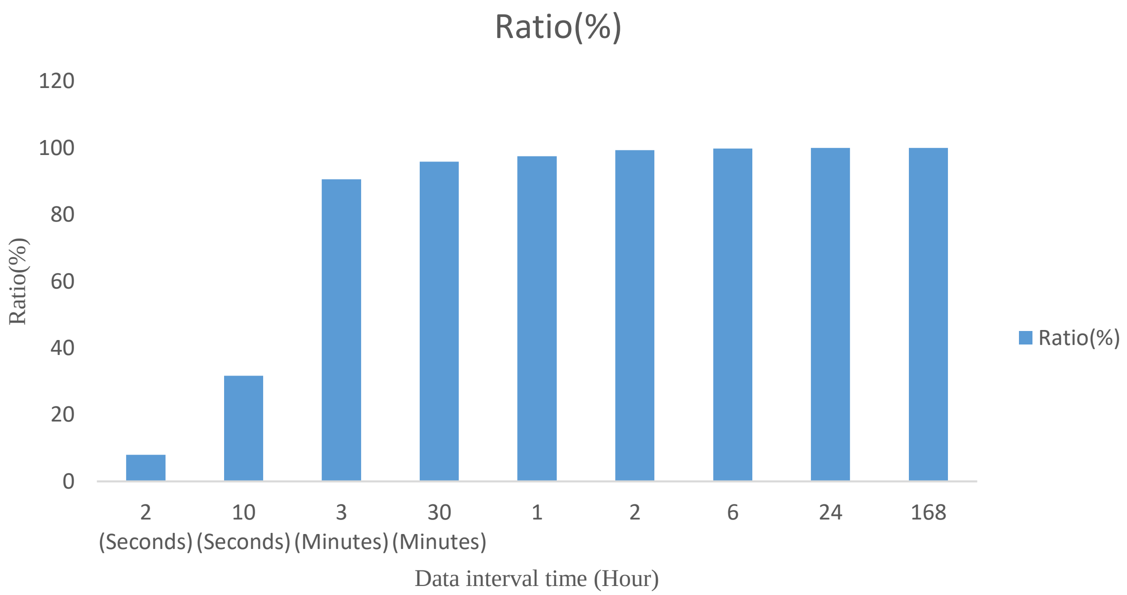


Figure 3.2 Distribution of Data Interval time

3.2 Introduction of the Strait of Malacca

The Strait of Malacca is located between Southeast Asia's Malay Peninsula and Sumatra Island. It is about 1080 kilometers long. It is 1,185 kilometers long with the Singapore Strait. It is northwest-southeast, with the Andaman Sea in the northwest and the South China Sea in the southeast. The strait is wide at the west and narrow at the east. The northwest is 370 kilometers wide and the southeast is the narrowest at 37 kilometers. The narrowest part of the strait is less than 20 kilometers. The water depth of the strait is 25-115 meters, the main channel is close to the side of the Malay Peninsula, 1.5-2 nautical miles wide.

Accidents in the Malacca Strait occurred frequently. The marine damage accident in this strait is more than 3 times that of Suez Canal and more than 5 times that of Panama Canal. The main reasons are as follows: First, the channel is narrow. Second, the water flow is gentle, the bottom of the gorge is flat and sandy, and it is easy to accumulate to form islands, reefs and shoals, causing ships to hit the reef or run aground. Thirdly, there are tropical rainstorms throughout the year, especially during the monsoon transition period, which is very frequent. Fourth, forest fires often occur in parts of the strait to which Indonesia belongs. Some Indonesians have a tradition of burning trees for fire farming. As a result, smoke often appears in the strait and visibility is insufficient. Fifthly, the number of ships passing through the strait has increased rapidly in recent years. In addition, the large-scale ships have caused congested waterways. Sixth, the traffic management in the Straits is chaotic, especially the local small fishing boats are arbitrarily passing through, which affects the normal navigation of merchant ships. Accidents such as ship collision, reef collision, and grounding are very likely to cause straight blockage and serious pollution.

For example, on June 19, 2016, a total of 315,814 AIS data were collected globally, and there were 17,239 AIS data on the Strait of Malacca to Hormuz Strait, accounting for 5.5% of the world. The geographical distribution is shown in the Figure 3.3 . The number of ships involved is 265. At the same time, East Asia's economy is developing rapidly, and the energy demand is large. A large amount of energy is collected and passes through the Strait of Malacca. It can be seen from Figures 3.4and 3.5 that the ship heat distribution of the Strait of Malacca (10km radius) is wider than that of the Strait of Hormuz, and there are more high-heat areas. Therefore, it is necessary to assess the maritime traffic situation in the Straits of Malacca

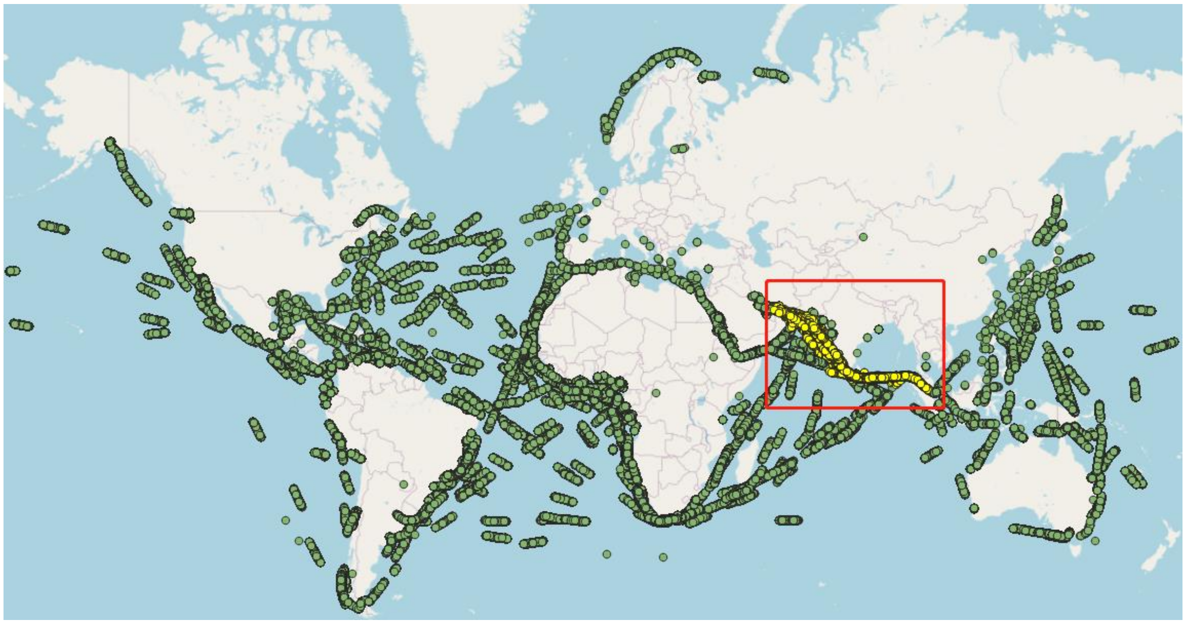


Figure 3.3 AIS data for the Strait of Hormuz-Malacca on June 19, 2016

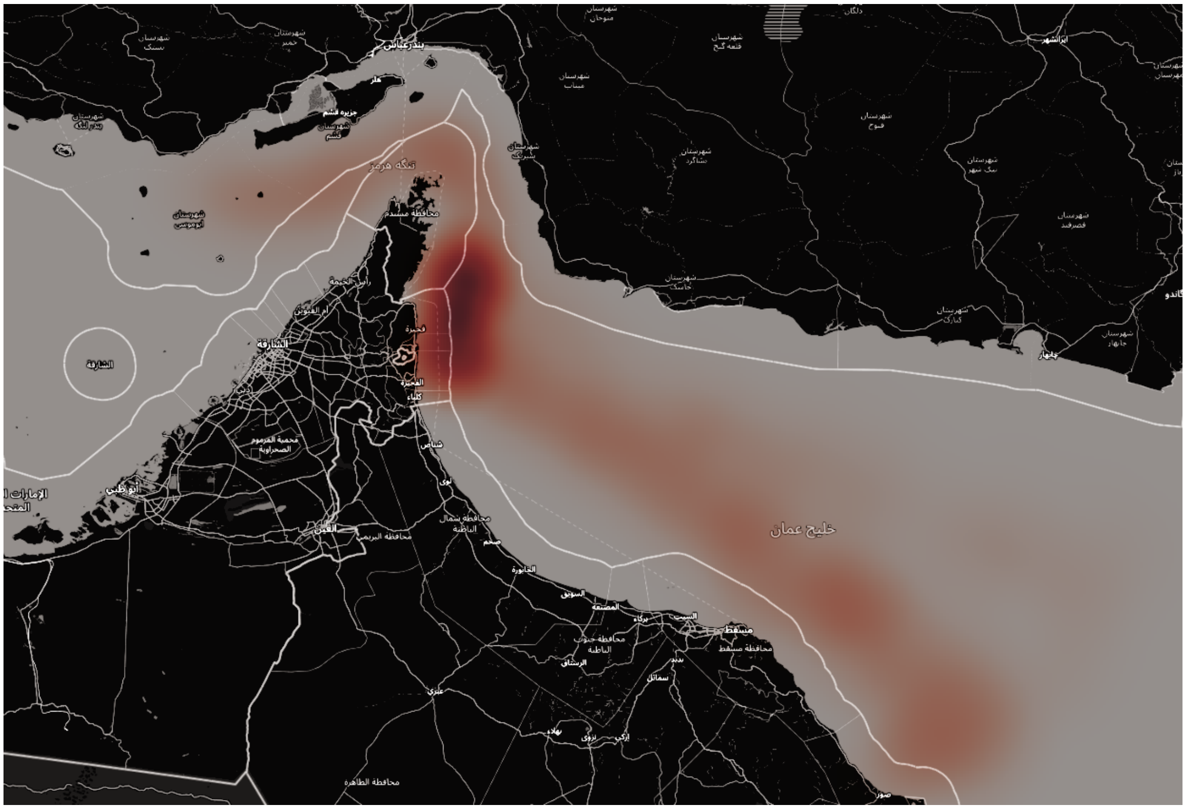


Figure 3.4 Ship Heat Distribution in Strait of Hormuz on January-June 2016

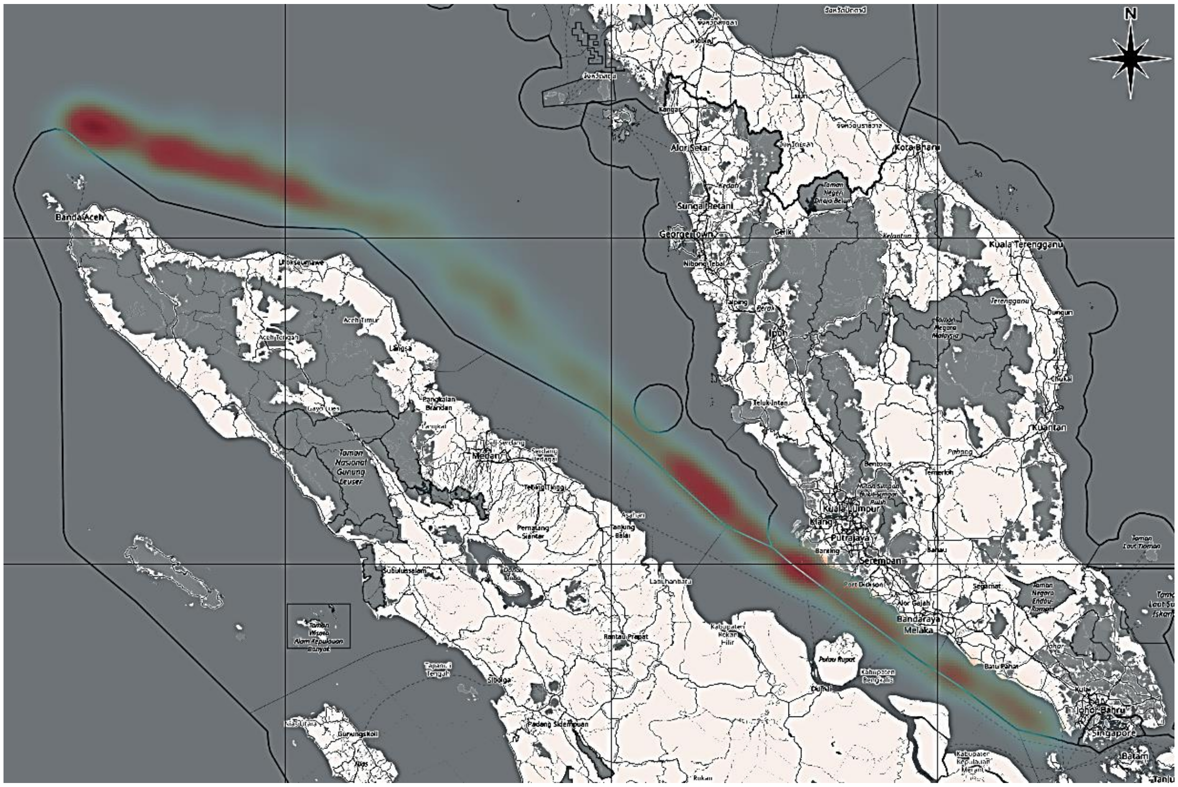


Figure 3.5 Ship Heat Distribution in Strait of Malacca on January-June 2016

3.3 Malacca Strait Maritime Traffic Status

The data source used in this study was provided by eaxctEarth, and the time span was the global AIS data from January 1, 2016 to June 30, 2016, a total of 51,705,694 data. First of all, we will add the data to the MySQL database for management, set the MMSI and dynamic time as the primary key, thus merging 182 CSV files, this is mainly to quickly filter out the AIS data located in the Strait of Malacca from each csv file. After processing, we selected 4,058,314 data from the database. The geographic distribution map of AIS data in the Straits of Malacca is shown in Figure 3.6.

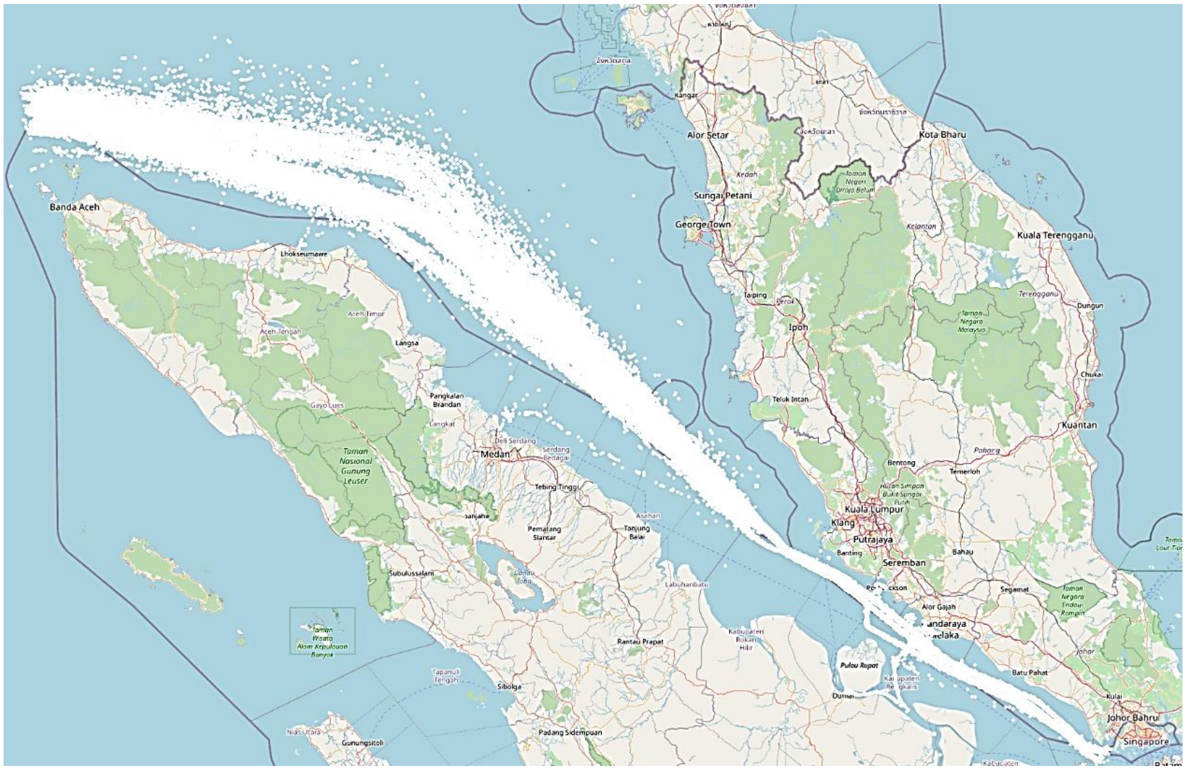


Figure 3.6 Strait of Malacca AIS Data Distribution

From Figure 3.6, we can observe the distribution of AIS data in the Strait of Malacca from January 1, 2016 to June 30, 2016. The overall distribution is that the amount of AIS data decreases from northwest to southeast, which also shows that the Strait of Malacca. The waterway is wide in the northwest and narrow in the southeast. The southeast channel is more complicated than the northwest channel, and there are routes across the strait in the southeast channel, which increases the risk of ship collision.

The amount of AIS data in an area does not represent the number of ships. Because a ship can generate a large amount of AIS data, in order to count the number of ships in the sea, we designed an algorithm to solve this problem, as shown in table 3.2.

Table 3.2 Count the Number of Ships

```
start
    read csv file, then                               Read data from the DB
    set list[ ], then                                 Create a new list
    a = list(csv file.imo.unique()), then             Add the unique imo value to the list
    print(len(a))                                     Print imo numbers
end
```

The Strait of Malacca is mainly a channel for energy transportation. After analyzing the AIS data and table 3.2 of algorithm provided by exactEarth, the data source collected 2,434 ships from January to June in the world, most of which are energy ships. As shown in table 3.3, there are 1,091 ships passing through the Strait of Malacca from January to June, accounting for 31% of the global energy ships, of which 823 are crude oil tankers, accounting for 75% of the number of ships in the Strait of Malacca. The site contains 925 ports, and the destination is Singapore with 282 ships, accounting for 34% of the Malacca Strait crude oil tankers, China with 102 ships, and 12% Malacca Strait crude oil tankers, and Korea with 81 ships, accounting for There are 9.8% of Malacca Strait crude oil tankers, Japan has 75 vessels, accounting for 9% of Malacca Strait crude oil tankers, and 130 LNG tankers, accounting for 12% of the Malacca Strait ships. The destination includes 231 ports. The number of LNG ships to Singapore is 9, accounting for 7% of the Malacca Strait LNG ships, the number of LNG ships going to China is 7, accounting for 5% of the Malacca Strait LNG ships, and the number of LNG ships going to Japan is 28, accounting for the Malacca Strait LNG The number of ships is 9%, and the number of LNG ships course to Korea is 11, accounting for 8% of the number of LNG ships in the Straits of Malacca. As shown in table 3.4, the final rankings of crude oil tankers in the Straits of Malacca from January to June 2016 are Singapore-China-Korea-Japan, and LNG ships are Japan-Korea-Singapore-China.

Table 3.3 Distribution of Ship Types in Straits of Malacca

	Crude oil tankers	LNG tankers
Strait of Malacca	75%	12%

Table 3.4 Distribution of Destination Types in Straits of Malacca

Destination	Crude oil tankers	LNG tankers
Singapore	34%	7%
China	12%	5%
Japan	9%	9%
Korea	9.8%	8%

We set the radius of the Malacca Strait Heat Map to 5 kilometers and merge the layers into Figure 3.6. The effect is shown in Figure 3.7. We can find that ships in the Straits of Malacca use the hottest routes.

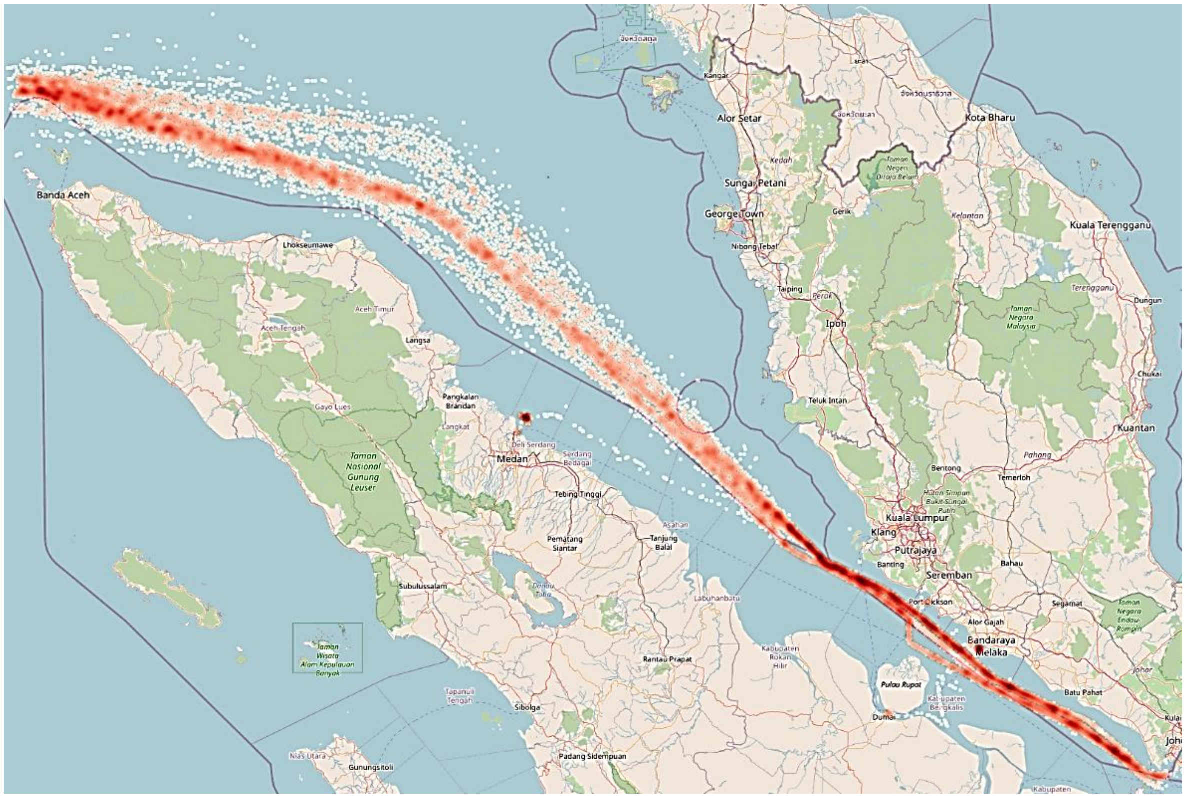


Figure 3.7 The Hottest Route in Strait of Malacca

Although the area of the central and northwestern Straits of Malacca is large, ships will still concentrate on one channel, which is close to the Indonesian side, and this channel is also a two-way channel, mainly northwest to southeast and southeast to northwest. course. The waterway with high utilization rate in the southeast sea is close to the Malaysian side.

In addition, we conducted a statistical analysis of the COG, SOG, and draught of ships in the Malacca Strait. From Figure 3.8, we can see the course distribution of the Malacca Strait from northwest to southeast from January to June. According to statistics, the course is distributed between 0°-360°, but most of them are concentrated around 300° and 120°, which is in line with the geographical trend of the northwest-southeast of the Strait of Malacca. On the main course (about 120° and about 300°), the ship's speed is mainly maintained at medium speed (12 kn - 17 kn), while the speed of other courses is mostly medium and low speed (0 kn - 11 kn) and part of the high speed (20 kn - 23 kn). We can also identify the areas where ships are anchored, moored, and close to land, the black points in the Figure 3.8, it has a speed (0 kn-4 kn).

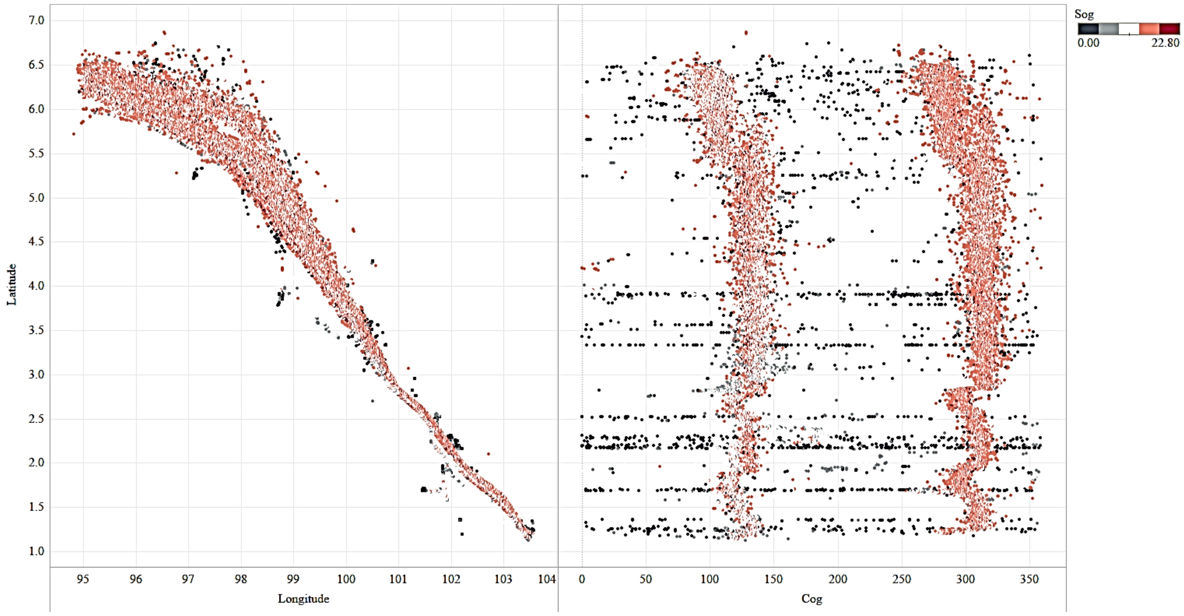


Figure 3.8 SOG - COG Distribution

According to Figure 3.9, we find that the ship's COG crossing degree in the northwestern Malacca Strait is higher. The closer it is to the southeast sea area, the clearer the COG, and the SOG decreases from northwest to southeast. This is related to the geographical features of the Strait of Malacca, the northwestern part of the Strait of Malacca is wide, and the southeast is narrow. If the course is changed frequently, the risk of collision will increase.

From Figure 3.9 we can observe that orange and yellow represent the northwest course, dark green and light green represent the southeast course, the downward trend of SOG is from northwest to southeast, the orange and yellow on the trend line are more obvious, the dark green and the light green is more obvious, which also shows that the SOG of the northwest course is faster than that of the southeast course. Overall, the SOG of the Strait of Malacca is concentrated at 12 kn-18 kn.

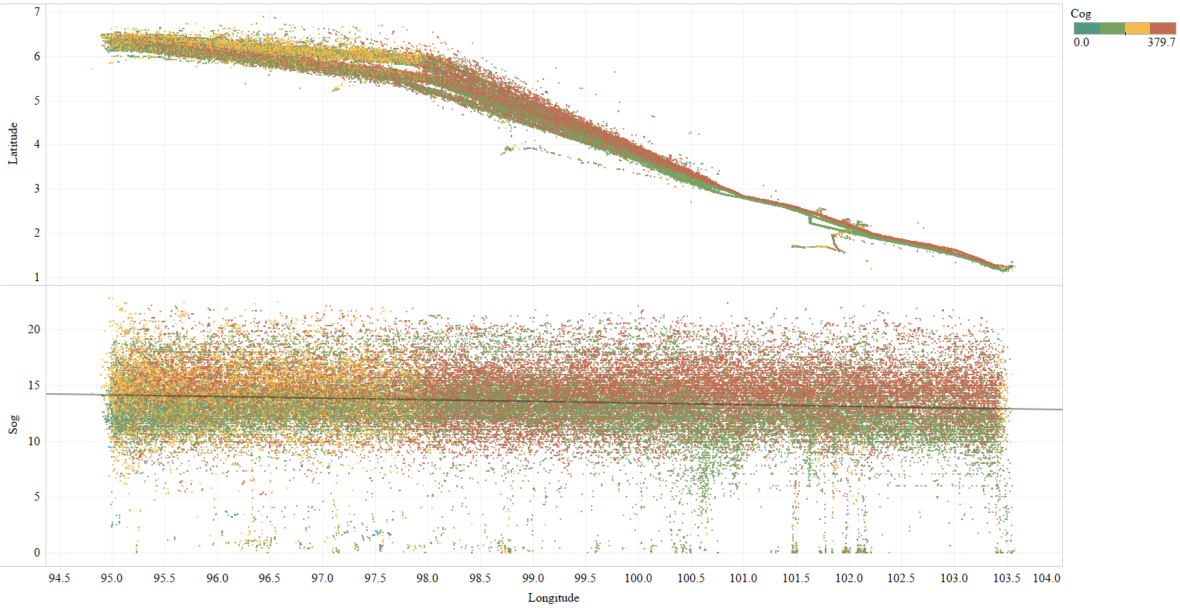


Figure 3.9 COG - SOG Distribution

In addition, we observe from Figure 3.10 that in areas where the course changes sharply, the ship's heat is also very high, the course distribution of this area is also relatively wide, while the courses of the low-medium heat area are concentrated around 300° and 120°.

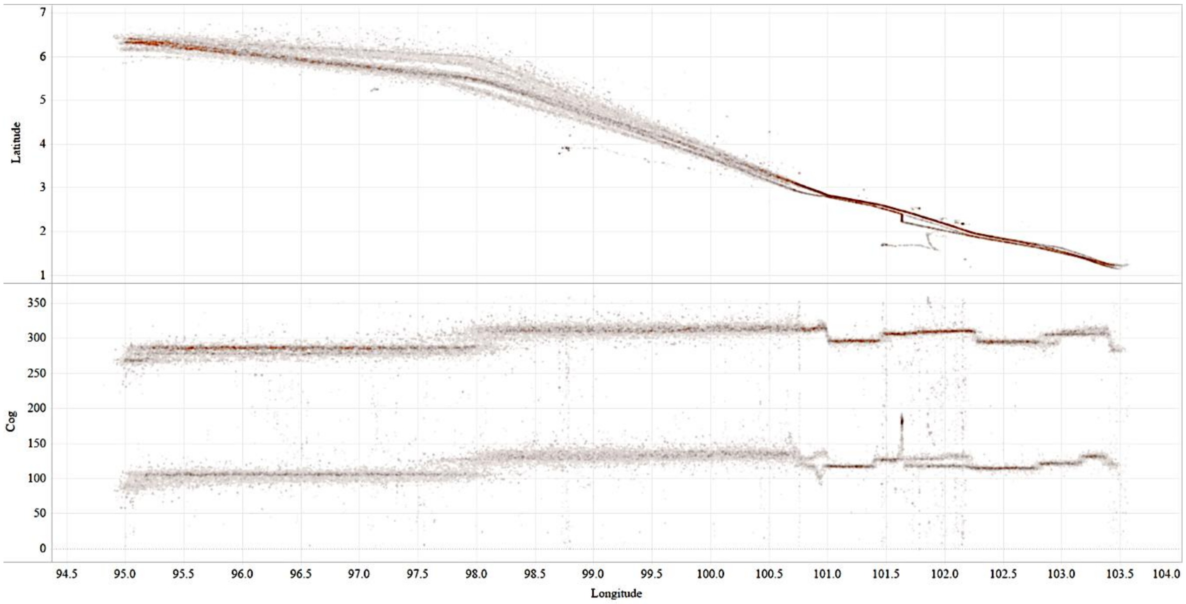


Figure 3.10 COG - Heat Distribution

From Figure 3.11, you can observe the relationship between the course and the draught of the ship. Dark green and light green represent the southeast course. The draught is concentrated above 11m. And there is a blank draught record between 18-19 m, which means that the draught of some ships suddenly drops from above 19m to below 18m at the same position. This may involve unloading goods by port. Orange and yellow represent the northwest course, the draft is concentrated below 11m, and the ship may sail without load and go to the Middle East to load oil and gas.

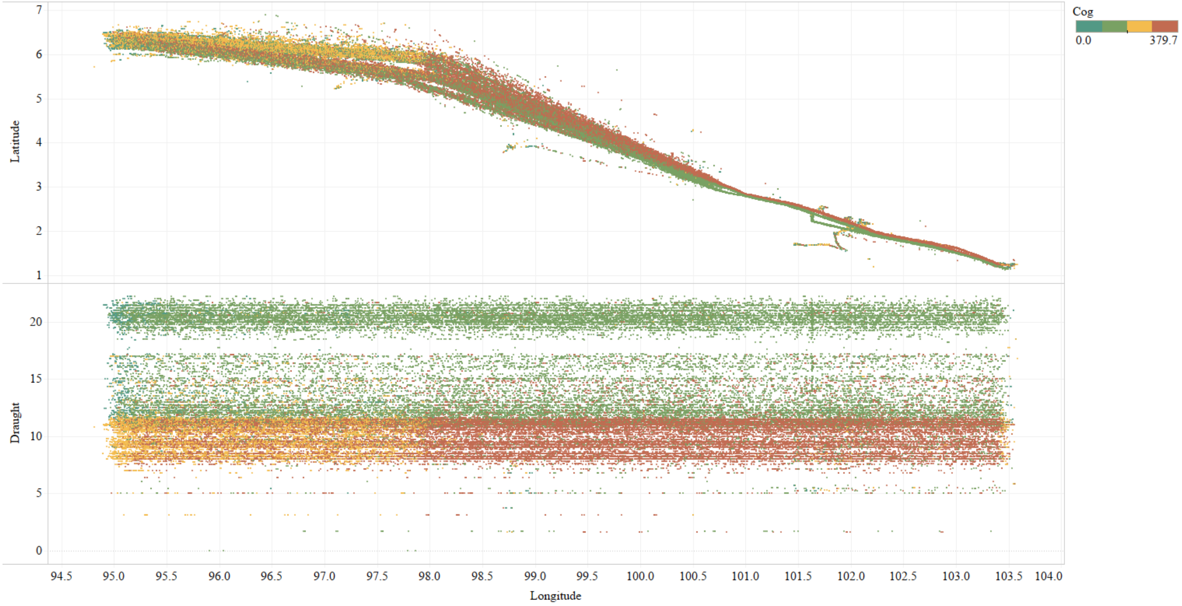


Figure 3.11 Draught - COG Distribution

It can be observed from Figure 3.9 that the original course data has been clearly classified. One type is concentrated at about 120°, which belongs to the southeast direction, and the other type is concentrated at about 300°, which belongs to the northwest direction. Based on the ship's course characteristics of the Strait of Malacca, we use the K-means method to classify the COG, SOG, and Draught.

The k-means algorithm uses distance as the standard for measuring the similarity between data objects, and usually uses Euclidean distance to calculate the distance between data objects. The calculation formula of Euclidean distance is given below:

$$dist(x_i, x_j) = \sqrt{\sum_{d=1}^{D}(x_{i,d} - x_{j,d})^2} \tag{29}$$

Where:

D: represents the number of data object attributes.

In the clustering process of the K-means algorithm, at each iteration, the corresponding cluster center needs to be recalculated (updated): the mean of all data objects in the corresponding cluster is the cluster center of the cluster after the update. Define the cluster center of the kth cluster as $Center_k$, then the cluster center update method is as follows:

$$Center_{\mathrm{k}} = \frac{1}{|C_k|} \sum_{x_i \in C_k} x_i \tag{30}$$

Where:

C_k: represents the k-th cluster,

$|C_k|$: represents the number of data objects in the kth cluster,

The K-means algorithm needs to iterate continuously to reclassify the clusters and update the cluster center. The conditions for termination are as follows:

$$J = \sum_{k=1}^{K} \sum_{x_i \in C_k} dist(x_i, Center_{\mathrm{k}}) \tag{31}$$

Where:

K: represents the number of clusters.

When the difference between the two iterations *J* is less than a certain threshold, $\Delta J < \delta$, δ is the iteration termination threshold, the iteration is terminated, and the resulting cluster is the final clustering result.

When *K*=2, the calculation results are shown in table 3.5.

Table 3.5 K-means Clustering Results

Direction	Clusters	Number of AIS	Cog	Sog	Draught
Northwest	1	72,538	296.83	14.1	10.3
Southeast	2	50,011	119.73	12.9	16.3

The northwest course has 72,538 AIS data, the course center point is 296.83°, the speed center point is 14.1kn, the draught center point is 10.3 meters, the southeast course has 50,011 AIS data, the course center point is 119.73°, and the speed center point is 12.9kn, the draught center is 16.3 meters. This shows that when the energy ships (mainly crude oil ships and LNG ships) of the Straits of Malacca sail to the northwest, most ships go to the Middle East to load oil and gas, and the speed is faster, and the empty load rate is high. When sailing to the southeast, most ships Loaded with oil and gas, the draft is deeper, and the speed is slower. Most ships go to East Asia to unload oil and gas. Therefore, the course of the energy ship in the Strait of Malacca is shown in Figure 3.12.

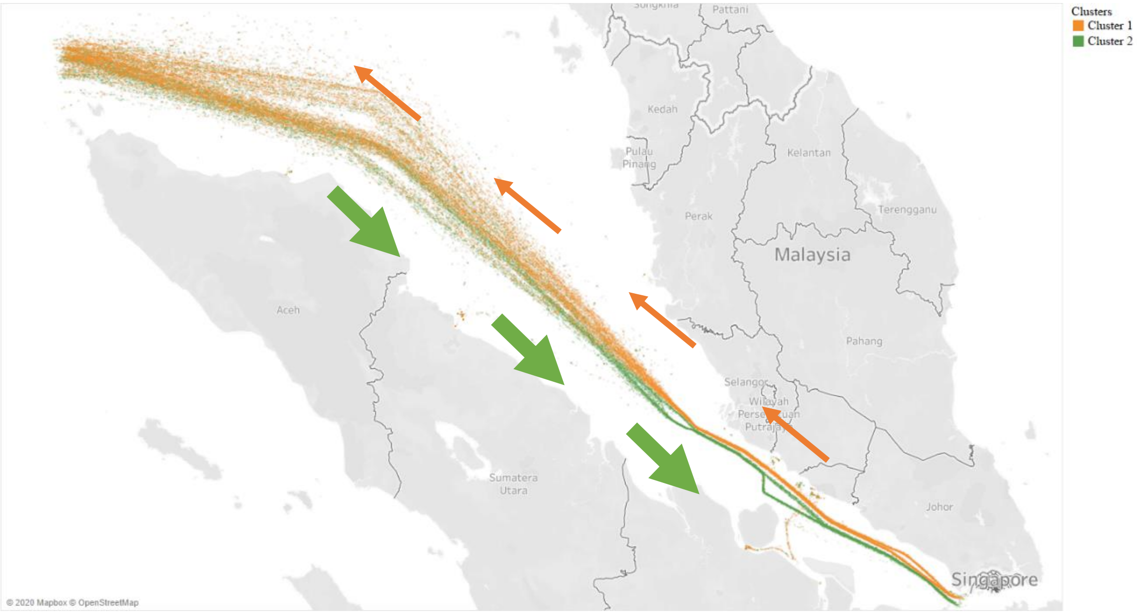


Figure 3.12 Malacca Strait Energy Ship Course, Speed, Draught Distribution

In addition, we introduce the concept of maritime traffic flow, and statistics of ship traffic flow in the Strait of Malacca. Li et al. (2019) proposed a method of calculating maritime traffic flow, the formula was as follows:

$$STF = \rho \cdot v \cdot W = \frac{N}{L \cdot W} \cdot v \cdot W = \frac{N}{L} \cdot v = \rho^{L} \cdot v \qquad (32)$$

Where:

STF: Ship traffic flow (*ship/hour*);

ρ: Traffic flow density based on area (*ship/square nautical mile*);

v: Traffic flow speed (*kn*);

W: Traffic flow width (*n mile*);

N: Number of ships (*ship*);

L: Traffic flow length (*n mile*);

ρ^{L}: Traffic flow density based on length (*ship/nautical mile*).

We calculated the traffic flow of the northwest course and the southeast course of the Straits of Malacca:

1) The half-year traffic flow is 11.68 ships/hour and 10.16 ships/hour. As shown in Figure 3.13;
2) The traffic flow during the day is 11.34 ships/hour and 9.95 ships/hour, and the night traffic flow is 11.35 ships/hour and 9.94 ships/hour. As shown in Figure 3.14;
3) The traffic flow in January was 3.90 ships/hour and 3.08 ships/hour, the traffic flow in February was 3.59 ships/hour and 2.96 ships/hour, and the traffic flow in March was 3.64 ships/hour. And 3.35 ships/hour, traffic flow in April is 3.28 ships/hour and 3.12 ships/hour, in May traffic flow is 3.74 ships/hour and 3.06 ships/hour, in June traffic flow 2.59/h and 3.45/h. As shown in Figure 3.15;
4) Therefore, each month's northwest traffic flow is greater than southeast traffic. We found that the total number of ships in the southeast course in March and April were 293 and 275, respectively, which were larger than the total number of ships in the northwest course of 291 and 266.

   However, the traffic flow in the southeast direction is lower than that in the northwest direction, because the ship speed in the southeast direction is lower than the ship speed in the northwest direction.

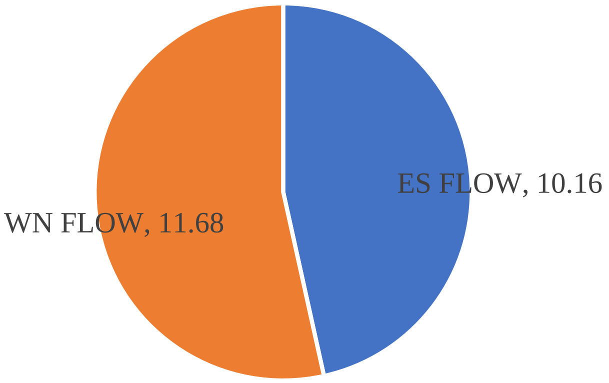


Figure 3.13 Malacca Strait Traffic Flow Distribution

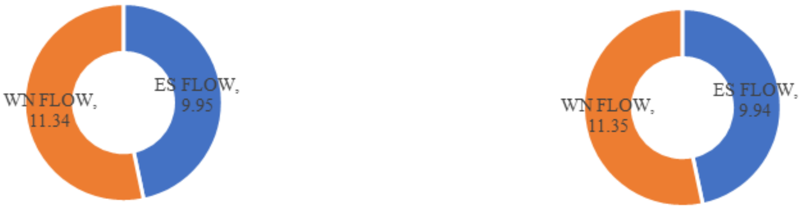


Figure 3.14 Malacca Strait Traffic Flow Distribution Day and Night

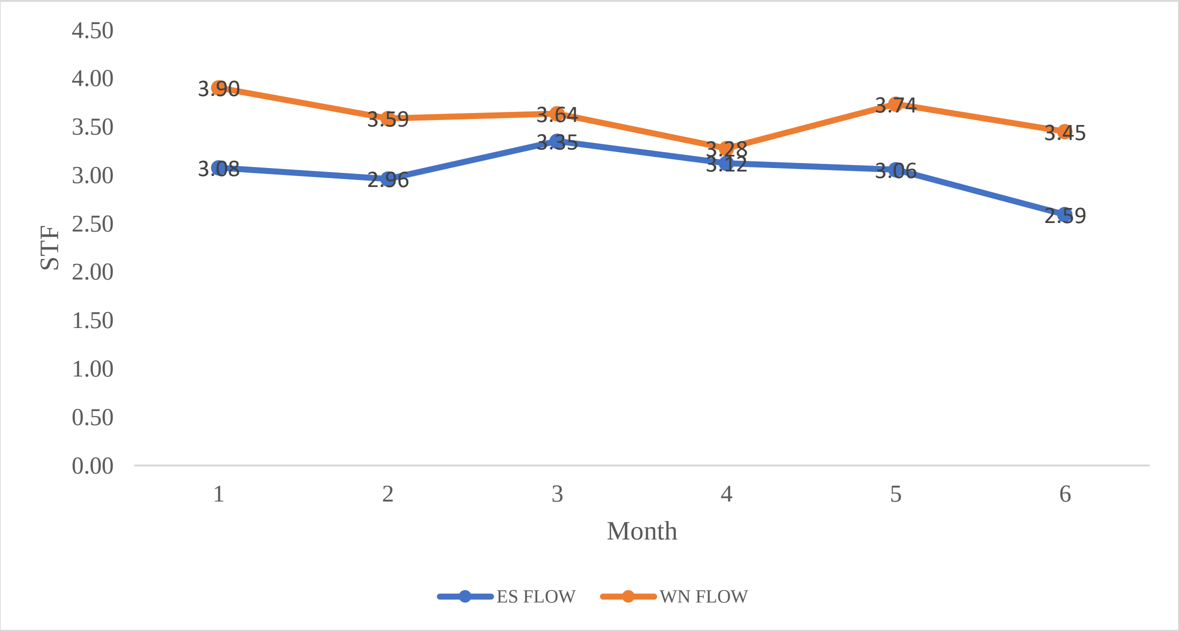


Figure 3.15 Malacca Strait Traffic Flow Distribution January to June

According to the above Figure, the Northwest ship traffic flow in Malacca Strait is greater than that of the southeast ship during half a year and during the day and night. The northwest ship traffic flow is the largest in January, the northwest ship traffic flow is the smallest in April, and the southeast ship traffic flow is the largest in March.6 Ship traffic in southeast China is the smallest in April, the difference between ship traffic in northwest and southeast in April is the smallest, and that in June is the largest.

Finally, we estimated greenhouse gas emissions based on ITTC combined with AIS data. The greenhouse gas estimation in this chapter uses LNG vessels in the Strait of Malacca as samples. The estimated results of the overall greenhouse gas emissions of LNG vessels in the Straits of Malacca are shown in table 3.6.

Table 3.6 Total emissions

Category	Emission amount (MT)
GHG	607,719.690
CO_2	592,638.651
CH_4	11,033.854
N_2O	23.705
NO_X	1,687.404
CO	1,687.404
$NMVOC$	648.669
Total consumption	222,636.604

Table 3.7 Month emissions

Category	January Emission amount (MT)
GHG	72,049.426
CO_2	70,261.463
CH_4	1,308.141
N_2O	2.810
NO_X	200.053
CO	200.053
$NMVOC$	76.904
Total consumption	25,549.623
	February Emission amount (MT)
GHG	161,871.842
CO_2	157,954.866
CH_4	2,938.971
N_2O	6.314
NO_X	449.455
CO	449.455
$NMVOC$	172.779
Total consumption	57,401.769
	March Emission amount (MT)
GHG	235,732.699
CO_2	229,882.807
CH_4	4,280.000
N_2O	9.195

NO_X	654.539
CO	654.539
$NMVOC$	251.617
Total consumption	85,721.490
	April Emission amount (MT)
GHG	41,648.181
CO_2	40,614.649
CH_4	756.171
N_2O	1.624
NO_X	115.641
CO	115.641
$NMVOC$	44.455
Total consumption	16,239.307
	May Emission amount(MT)
GHG	54,792.150
CO_2	53,432.440
CH_4	994.1814
N_2O	2.137
NO_X	152.136
CO	152.136
$NMVOC$	58.484
Total consumption	19,429.979
	June Emission amount(MT)
GHG	41,625.391
CO_2	40,592.425
CH_4	755.757
N_2O	1.623
NO_X	115.577
CO	44.430
$NMVOC$	44.430
Total consumption	18,294.437

It can be seen from Table 3.7 that the greenhouse gas emissions from LNG vessels in the Straits of Malacca from January to June 2016 were 607,719.690 MT, of which January GHG was 72,049.426 MT, February GHG was 161,871.842 MT, March GHG was 235,732.699 MT, and April GHG was 41,648.181 MT, May GHG is 54,792.150MT, June GHG is 41,625.391MT, the visualization image is shown in Figure 3.16.

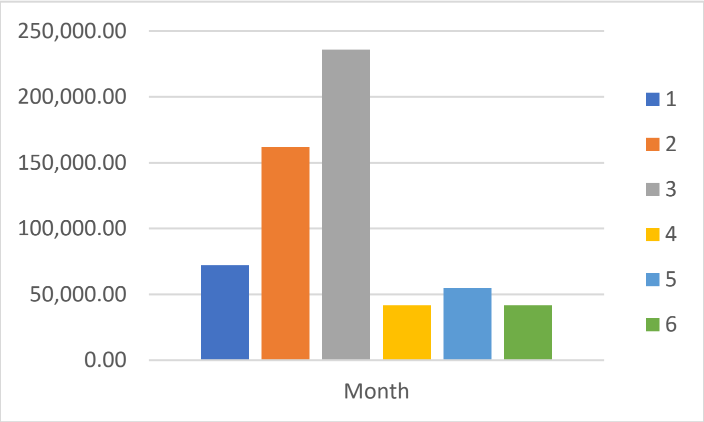


Figure 3.16 Monthly distribution of greenhouse gas emissions in the Straits of Malacca

The most GHG emissions are in March, and the traffic flow in March is about 7 ships/hour, which is also the largest ship traffic flow of all months. The overall emissions from January to March are greater than from April to June, and the emissions from March to April have dropped sharply. Traffic flow in April is also one of the smallest months.

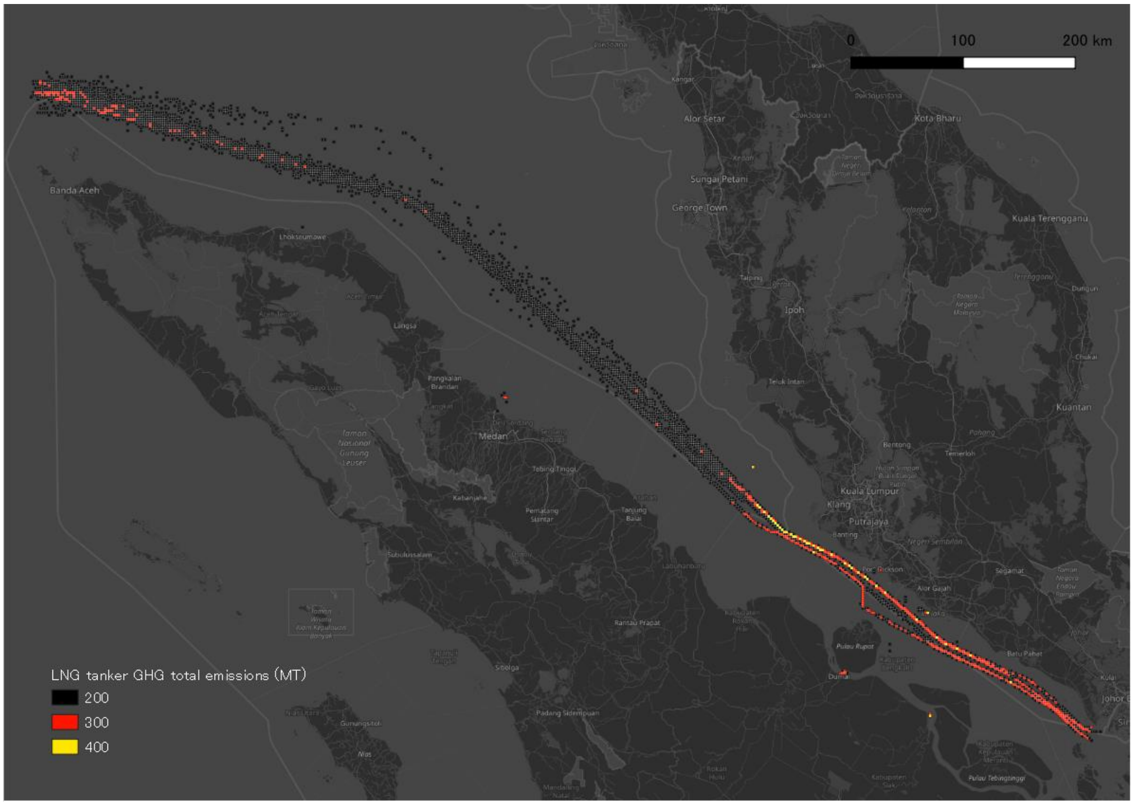


Figure 3.17 Geographical distribution of greenhouse gas emissions

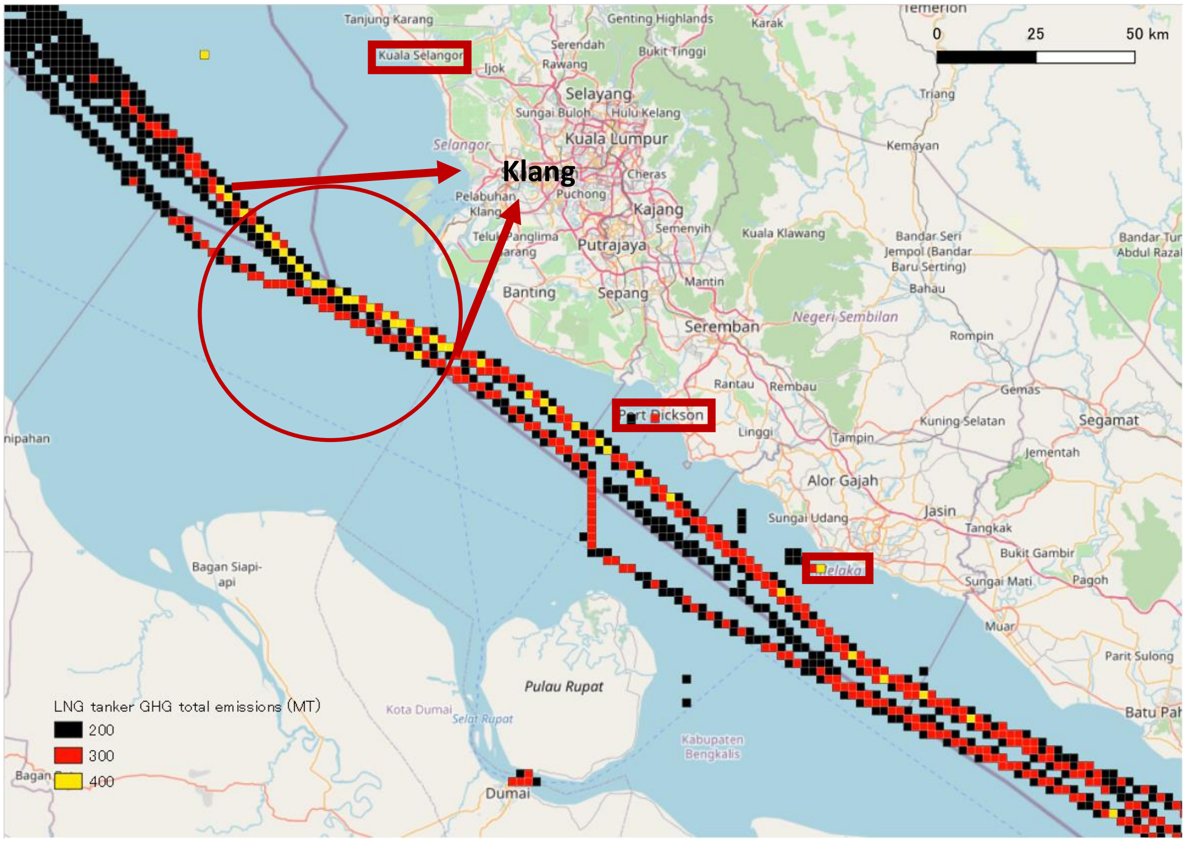


Figure 3.18 Geographical distribution of greenhouse gas emissions

We use a 5km*5km grid to cover the Strait of Malacca and sum up the greenhouse gas emissions of LNG ships in each grid to obtain the GHG spatial distribution map generated by the LNG ships in the Malacca Strait, as shown in Figure 3.17. The yellow area represents a large amount of GHG emissions in this sea area. We zoom in on this sea area. As shown in Figure 3.18, we find that this sea area has the largest port in Malaysia-Port Klang, so there are many ships in this sea area. In this area, the channel is narrow, and the concentration of ships is high. The yellow area belongs to the northwest channel of the Strait of Malacca. The speed of the ship is faster. Therefore, the GHG emissions in this area are higher.

In order to verify the validity of the estimated value, we use the environmental indicators of the Malaysian Environmental Department for judgment. First, we selected three coastal cities that extended outward from the Port of Klang as the center: Bandararay Melaka, Port Dickson, Kuala Selangor. Then, we queried the environmental indicators of these four cities from January to June 2016, as shown in the figure 3.20-3.23 shown.

Legend of environmental index is shown in Figure 3.19. The larger the index, the worse the environment.

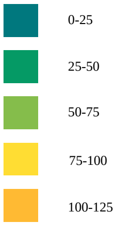
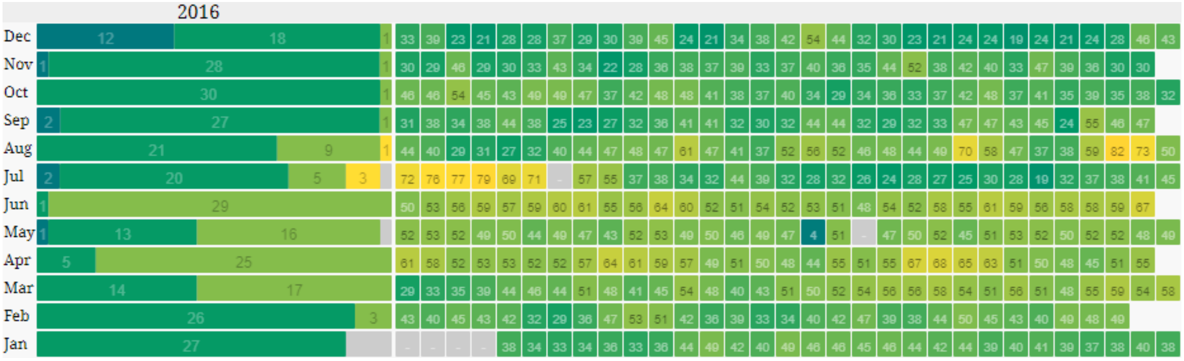
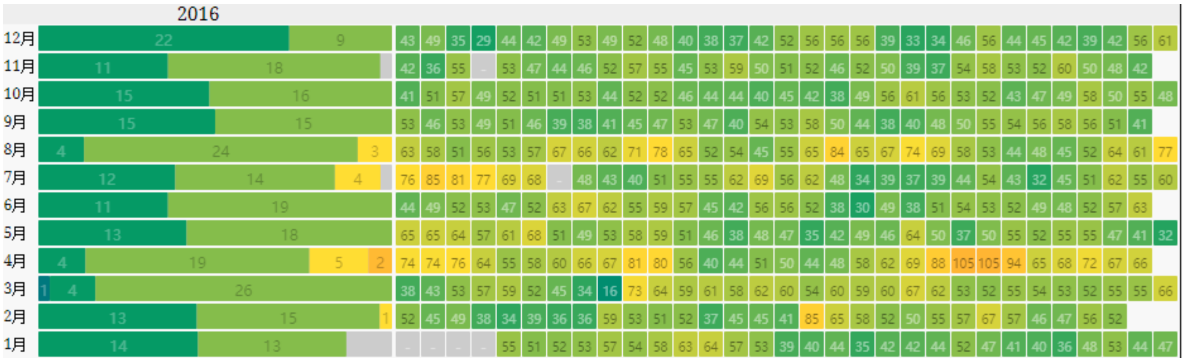


Figure 3.19 Legend of environmental index



	2016																															
Dec	12 18 1	33	39	23	21	28	28	37	29	30	39	45	24	21	34	38	42	54	44	32	30	23	21	24	24	19	24	21	24	28	46	43
Nov	1 28 1	30	29	46	29	30	33	43	34	22	28	36	38	37	39	33	37	40	36	35	44	52	38	42	40	33	47	39	36	30	30	
Oct	30 1	46	46	54	45	43	49	49	47	37	42	48	48	41	38	37	40	34	29	34	36	33	37	42	48	37	41	35	39	35	38	32
Sep	2 27 1	31	38	34	38	44	38	25	23	27	32	36	41	41	32	30	32	44	44	32	29	32	33	47	47	43	45	24	55	46	47	
Aug	21 9 1	44	40	29	31	27	32	40	44	47	48	47	61	47	41	37	52	56	52	46	48	44	49	70	58	47	37	38	59	82	73	50
Jul	2 20 5 3	72	76	77	79	69	71	-	57	55	37	38	34	32	44	39	32	28	32	26	24	28	27	25	30	28	19	32	37	38	41	45
Jun	1 29	50	53	56	59	57	59	60	61	55	56	64	60	52	51	54	52	53	51	48	54	52	58	55	61	59	56	58	58	59	67	
May	1 13 16	52	53	52	49	50	44	49	47	43	52	53	49	50	46	49	47	4	51	-	47	50	52	45	51	53	52	50	52	52	48	49
Apr	5 25	61	58	52	53	53	52	52	57	64	61	59	57	49	51	50	48	44	55	51	55	67	68	65	63	51	50	48	45	51	55	
Mar	14 17	29	33	35	39	44	46	44	51	48	41	45	54	48	40	43	51	50	52	54	56	56	58	54	51	56	51	48	55	59	54	58
Feb	26 3	43	40	45	43	42	32	29	36	47	53	51	42	36	39	33	34	40	42	47	39	38	44	50	45	43	40	49	48	49		
Jan	27	-	-	-	-	38	34	33	34	36	33	36	44	49	42	40	49	46	46	45	46	44	42	44	39	40	41	39	37	38	40	38

Figure 3.20 Bandaraya Melaka of Environmental Index



	2016																															
12月	22 9	43	49	35	29	44	42	49	53	49	52	48	40	38	37	42	52	56	56	56	39	33	34	46	56	44	45	42	39	42	56	61
11月	11 18	42	36	55	-	53	47	44	46	52	57	55	45	53	59	50	51	52	46	52	50	39	37	54	58	53	52	60	50	48	42	
10月	15 16	41	51	57	49	52	51	51	53	44	52	52	46	44	44	40	45	42	38	49	56	61	56	53	52	43	47	49	58	50	55	48
9月	15 15	53	46	53	49	51	46	39	38	41	45	47	53	47	40	54	53	58	50	44	38	40	48	50	55	54	56	58	56	51	41	
8月	4 24 3	63	58	51	56	53	57	67	66	62	71	78	65	52	54	45	55	65	84	65	67	74	69	58	53	44	48	45	52	64	61	77
7月	12 14 4	76	85	81	77	69	68	-	48	43	40	51	55	55	62	69	56	62	48	34	39	37	39	44	54	43	32	45	51	62	55	60
6月	11 19	44	49	52	53	47	52	63	67	62	55	59	57	45	42	56	56	52	38	30	49	38	51	54	53	52	49	48	52	57	63	
5月	13 18	65	65	64	57	61	68	51	49	53	58	59	51	46	38	48	47	35	42	49	46	64	50	37	50	55	52	55	55	47	41	32
4月	4 19 5 2	74	74	76	64	55	58	60	66	67	81	80	56	40	44	51	50	44	48	58	62	69	88	105	105	94	65	68	72	67	66	
3月	1 4 26	38	43	53	57	59	52	45	34	16	73	64	59	61	58	62	60	54	60	59	60	67	62	53	52	55	54	53	52	55	55	66
2月	13 15 1	52	45	49	38	34	39	36	36	59	53	51	52	37	45	45	41	85	65	58	52	50	55	57	67	57	46	47	56	52		
1月	14 13	-	-	-	-	55	51	52	53	57	54	58	63	64	57	53	39	40	44	35	42	42	44	52	47	41	40	36	48	53	44	47

Figure 3.21 Klang of Environmental Index

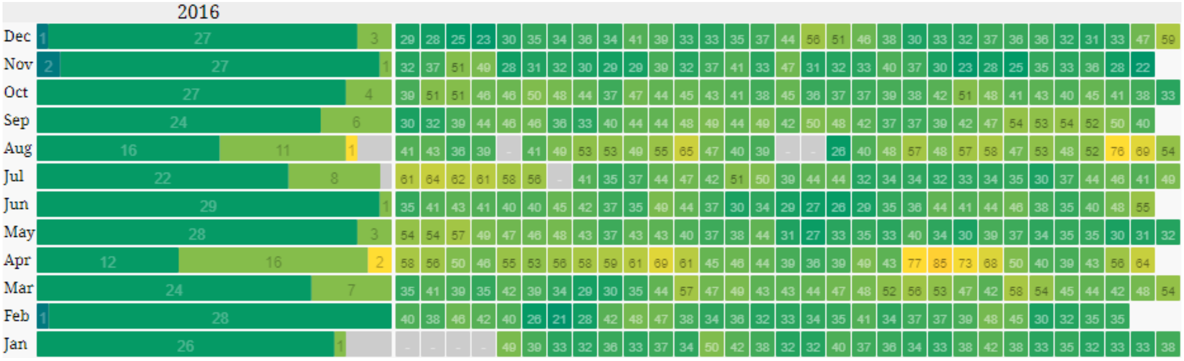


Figure 3.22 Port Dickson of environmental index

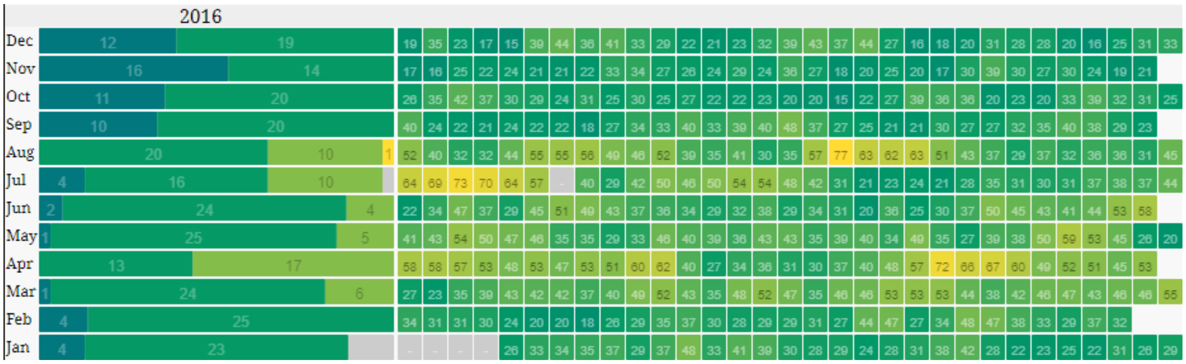


Figure 3.23 Kuala Selangor of Environmental Index

Refer to Figure 3.19, we can observe Figure 3.20-3.23, Klang has an environmental index of 50-75 for 110 days, and 75-100 for 8 days, and 0-25 for only 1 day. Bandaraya Melaka has an environmental index of 50-75 for 90 days, Port Dickson has a 28-day environmental index of 50-75, and Kuala Selangor has a 32-day environmental index of 50-75, so Klang is the worst coastal city in the environment, which is related to the GHG spatial distribution estimated in this study.

ITTC can be used not only to estimate the greenhouse emissions in a certain area, but also to track the GHG footprint of the ship, which allows us to clearly see the ship's greenhouse gas emissions at each location, which is Great help, especially in restricted emission areas. We tried to select 4 LNG ships of different sizes in the Strait of Malacca to track their greenhouse gas emissions as they cross the Strait. The size data is shown in table 3.8. The tracking effect is shown in Figure 3.24-3.27.

Table 3.8 LNG Tanker Dimensions

IMO	LWL/M	Beam/M	Gross weight/MT	DWT/MT	Year	Draught Summer/M	GHG/MT	Time/H	Rate/MT/H	COG
9064073	212	32	46555	35760	1996	10	297.1	36.45	8.2	ES
9311567	275	44	98798	83961	2007	12	372.5	38.45	9.7	WN
9307176	296	44	95824	78594	2005	12	675.0	32.7	20.6	WN
9418365	331	54	163922	130102	2009	12	1060	34.4	31.0	ES

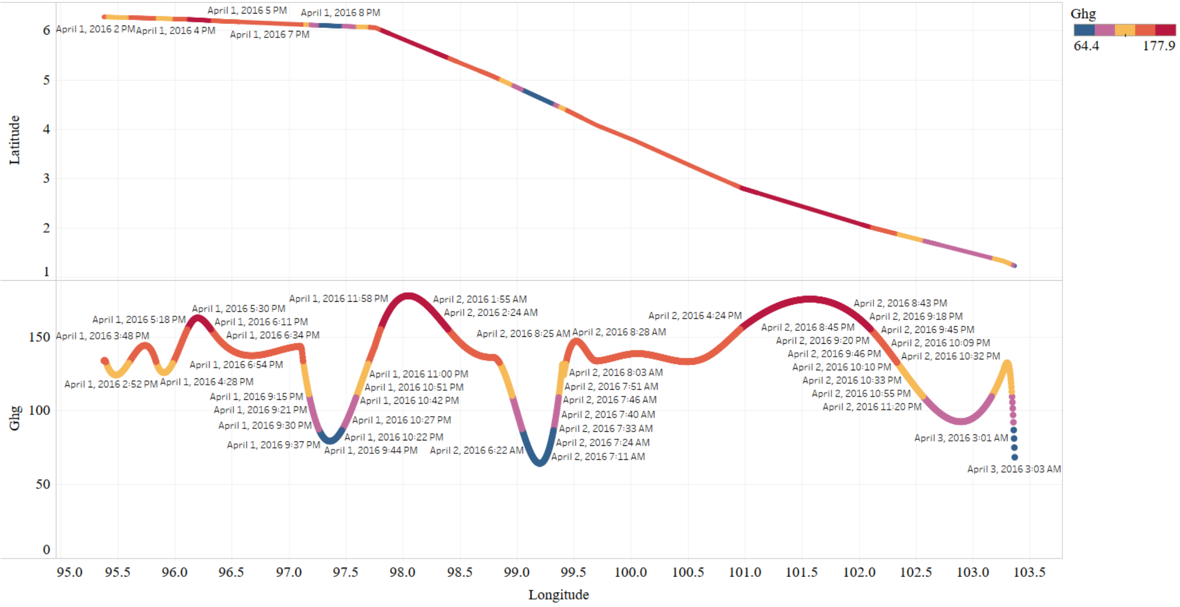


Figure 3.24 9064073-LNG Ship GHG Footprint

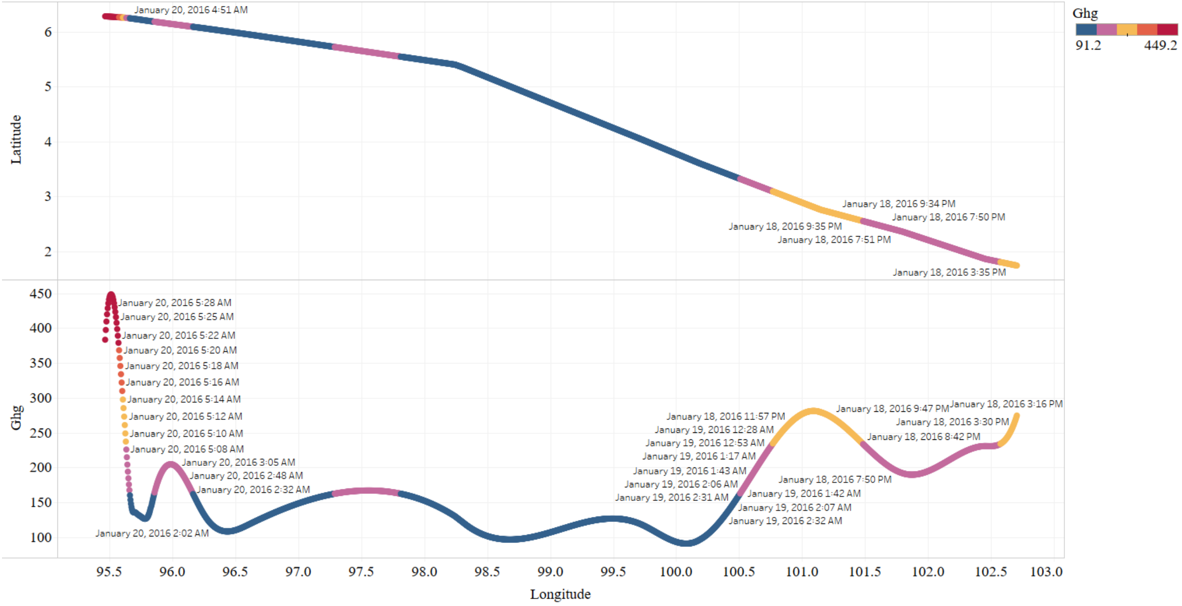


Figure 3.25 9311567-LNG Ship GHG Footprint

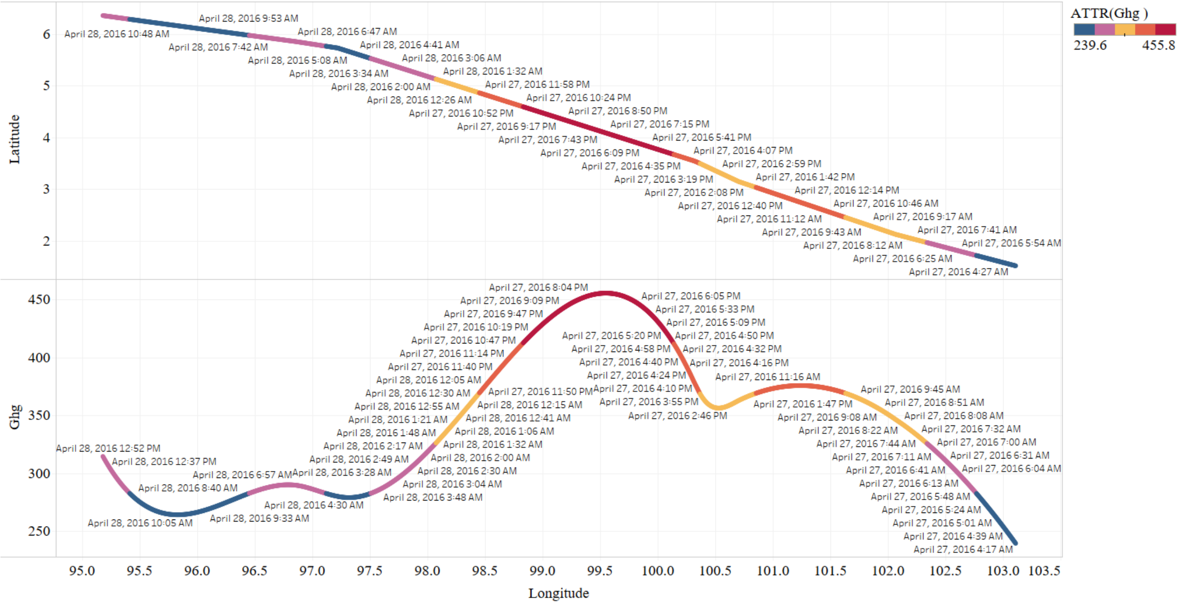


Figure 3.26 9307176-LNG Ship GHG Footprint

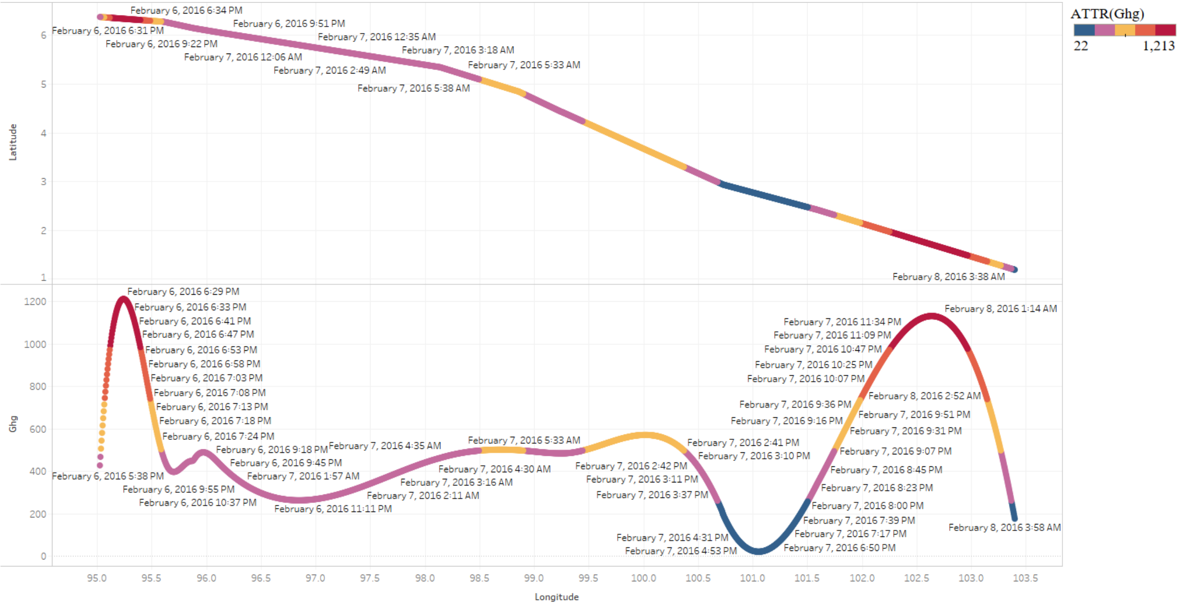


Figure 3.27 9418365-LNG Ship GHG Footprint

We found that the size of the ship (LWL and Beam) has a positive correlation with the GHG emission rate, which means that the larger the size of the ship, the greater the GHG emissions. It should be noted that the speed of the energy ship in the Strait of Malacca is characterized by the overall stability of the speed, and the difference is not large, so this theory needs to add the precondition of speed stability.

4. Ship Status Prediction

After database analysis (Chapter 3, 3.3), taking the data integrity and continuity as the principle, we extracted the LNG ship trajectory data of MMSI 310028000 from the trajectory database as a sample to illustrate the feasibility of the prediction model. The vessel trajectory over 6 months was shown in Figure 4.1. The vessel travels mainly between Japan and Australia, IMO number is "8913174", built in 1992. To simplify calculations, carbon dioxide is used instead of GHG in this chapter (CO2 accounts for 90% of GHG).

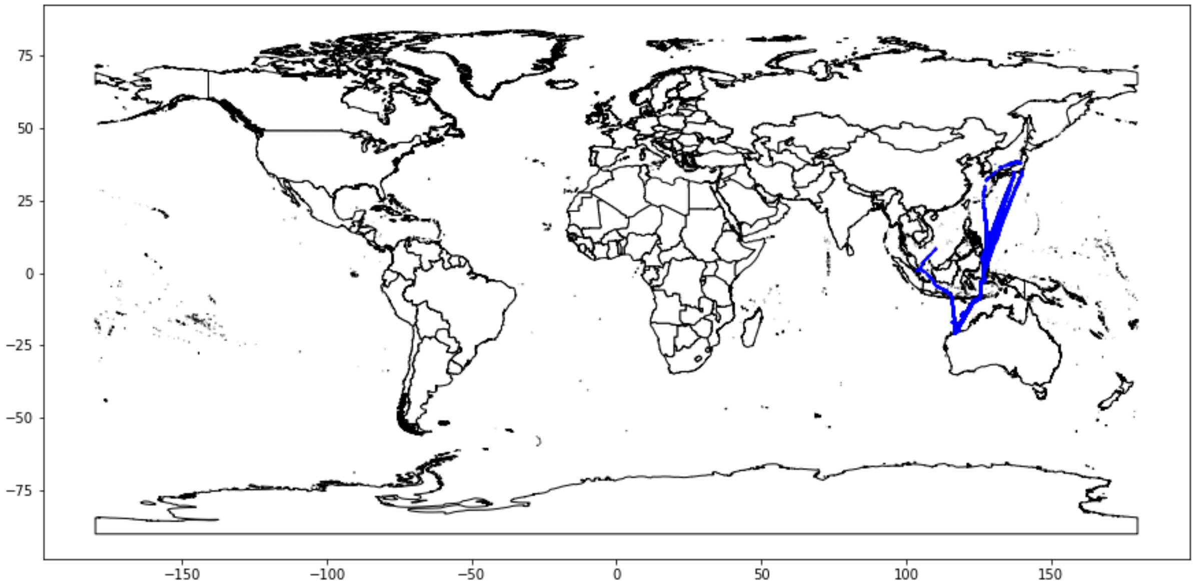


Figure 4.1 Distribution of Sample ship trajectory

4.1 Interpolation Calculation

The focus of this research was to grasp the carbon dioxide emissions of vessels in real-time. There should not be too much historical data. That's why we chose Deep learning. Deep learning had the function of learning features from a small amount of data, which was suitable for real-time analysis of data. Therefore, we chose to use the vessel data of January 5, 2016 at 00: 35: 10-00: 59: 40 for analysis. The total time interval was 26 minutes, but there are only 23 trajectory data, as shown in table 4.1.We could find that the data intervals are different, which were 1 second, 7 seconds, 11 seconds, 12 seconds, 7 seconds. Therefore, this study proposes to use the cubic spline interpolation method to resample the data at 1 second intervals.

Table 4.1 Partial Sample Data

MMSI	TIME	LONGITUDE	LATITUDE	SOG	COG
310028000	00:35:10	118.6957	-17.0403	15.9	24.7
310028000	00:35:11	118.6961	-17.0395	15.8	25.1
310028000	00:35:18	118.6963	-17.0391	15.9	25.1
310028000	00:35:29	118.6966	-17.0383	15.8	25
310028000	00:35:41	118.697	-17.0375	15.8	24.8
310028000	00:35:48	118.6972	-17.0371	15.8	25.2
……	……	……	……	……	……

We used equation (1) to (14) to calculate, took longitude as an example, $x_i = longitude_i$, The partial calculation results were shown in Figure 4.2, We could find that the line connecting the original data points was no longer a simple straight orange line, but had become a blue curve, which was more in line with the actual situation. In the actual geographic location, we could see the restoration trajectory after interpolation calculation from Figure 4.3, where the yellow point was the trajectory point of the vessel at 00:59:40, the red point was the trajectory point on January 5, orange point was point on other days in January, green point was interpolation points.

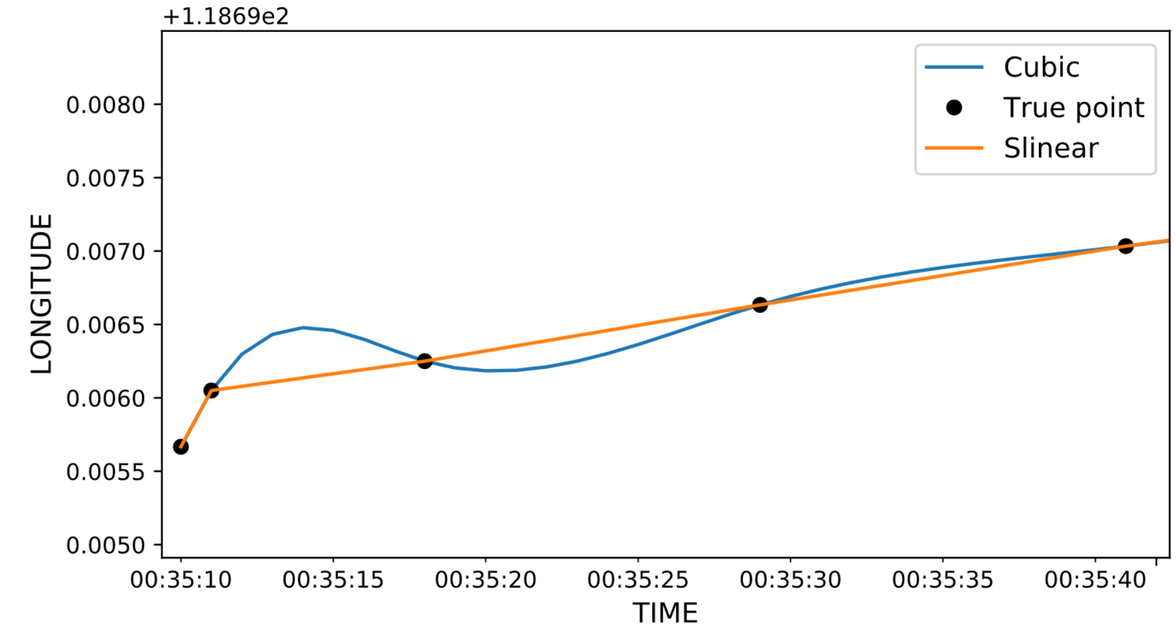


Figure 4.2 Distribution of Longitude Interpolation

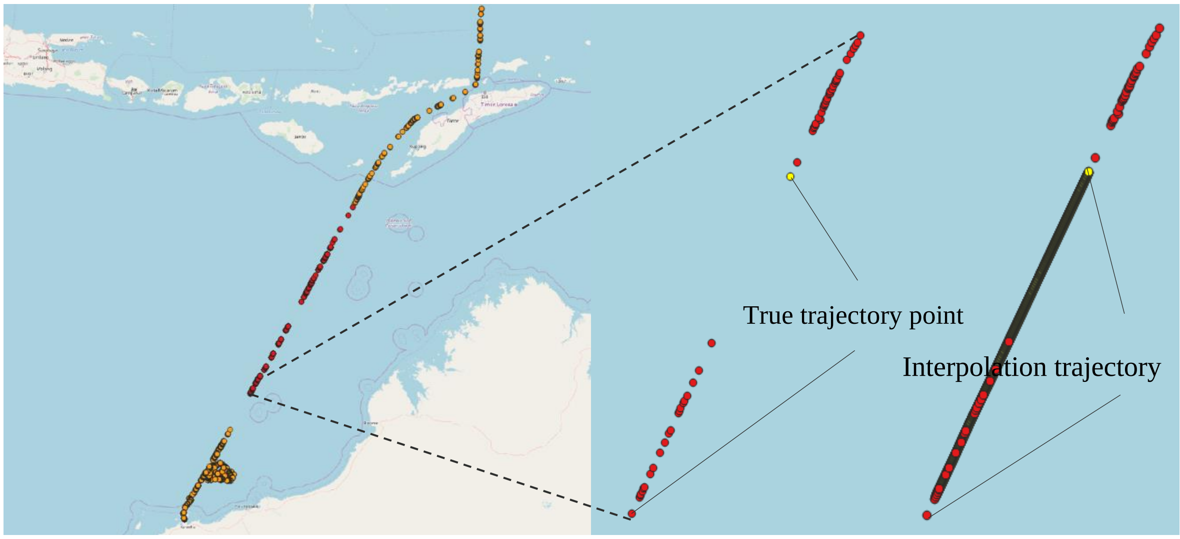


Figure 4.3 Geographical Distribution of Trajectory Interpolation

4.2 Vessel trajectory prediction

The recurrent neural network is a typical framework for deep learning, it can be used to deal with Spatial-temporal sequences problems. The characteristic is that the output of the current moment depends on the calculation result of the previous moment and the timing is strong. The LSTM model is an improvement of the recurrent neural network, also analyzes historical calculation results that are much older, automatically loses invalid historical calculation results, and remembers useful historical calculation results. The timeliness is stronger than the recurrent neural network.

As described in Chapter 1 and 2, for a vessel, its trajectory characteristics at time *t* could be expressed as $Y_{(t)}$ ={(p_1, a_1, t_1),(p_2, a_2, t_2),(p_3, a_3, t_3),…,(p_n, a_n, t_n)}, we could use it as input value of LSTM model. The amount of data after interpolation calculation changed from 23 to 1,472. The experimental environment for this study was the DELL OptiPlex 7050 desktop computer. CPU: Intel(R) Core (TM) i7-7700 CPU @3.60GHz, memory was 16.0GB, operating system was Windows10 Pro, program development environment was Pycharm (Python 3.7), using LSTM model provided by Keras. After experiments and manual adjustment of parameters, the optimal parameters of the LSTM model in this study were finally determined. As shown in table 4.2, the time required to run the model once was 195 seconds. We think that it was better to use the image to display the model, the loss function was shown in Figure 4.4, the prediction results were shown in Figures 4.5 and 4.6.

Table 4.2 LSTM Model Parameters

Base learning rate	0.001	LSTM_layer_1	256
Optimizer	Adaptive Moment Estimation	LSTM_layer_2	128
Epoch	125	Dropout_layer _1	128
Batch size	138	Dense_layer _1	128
Loss function	Mean Square Error	Dropout_layer _2	128
Activation_1	Tanh	Dense_layer _2	4
Activation_2	Linear	Kernel_initializer	Orthogonal
Train set	1,173	Validation set	235
Test set	293		

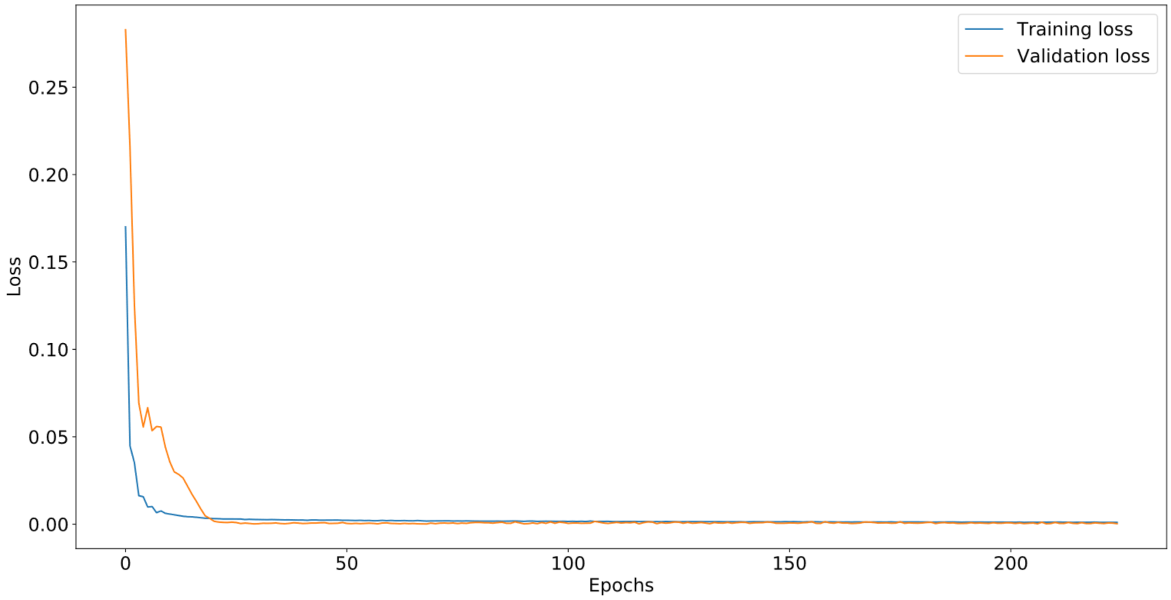


Figure 4.4 Training and Validation Loss of LSTM Model

From Figure 4.5, the green, red, and blue points indicated by the arrows were the trajectory points of the vessel at 00:59:40, where the green point was the true trajectory point, the red point was the trajectory point predicted by the LSTM model, and the blue point was the trajectory points predicted by the RNN model. The trajectory point predicted by the LSTM

(118.7394, -16.9514) was closer to the true value (118.7439,-16.9425) than the trajectory points predicted by the RNN model (118.7369,-16.9564).

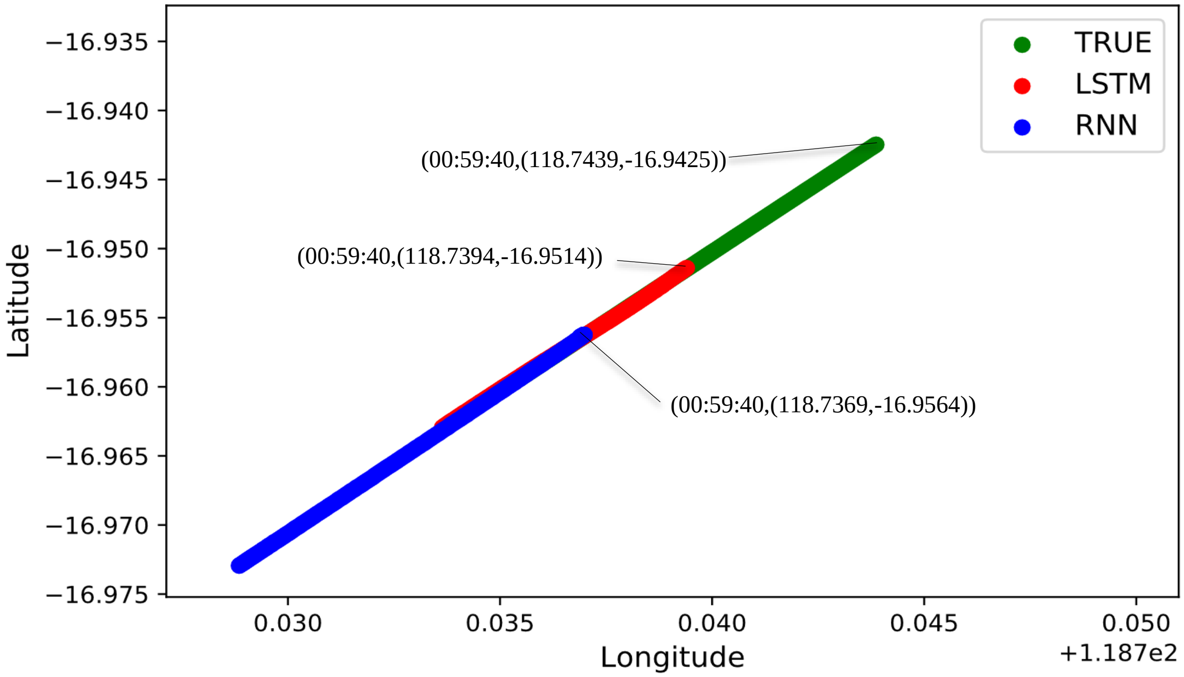


Figure 4.5 Coordinate Distribution of True and Prediction Trajectory Results

The LSTM model error was 0.593 nautical miles (1.097 *km*), the RNN model error was 0.928 nautical miles (1.718 *km*). From Figure 4.5, it can be found that the three trajectory trends were generally consistent, but the predicted endpoints were different. This perspective was like the actual geographic perspective. From Figure 4.6, it can be found that the trajectory points predicted by RNN were more concentrated at the end point, causing the trajectories to overlap, but the LSTM model did not have this defect. The time step was that time difference from 00:54:49 (in seconds). From the perspective of actual geography, Figure 4.7, the trajectories predicted by the two models roughly coincided with the actual trajectories.

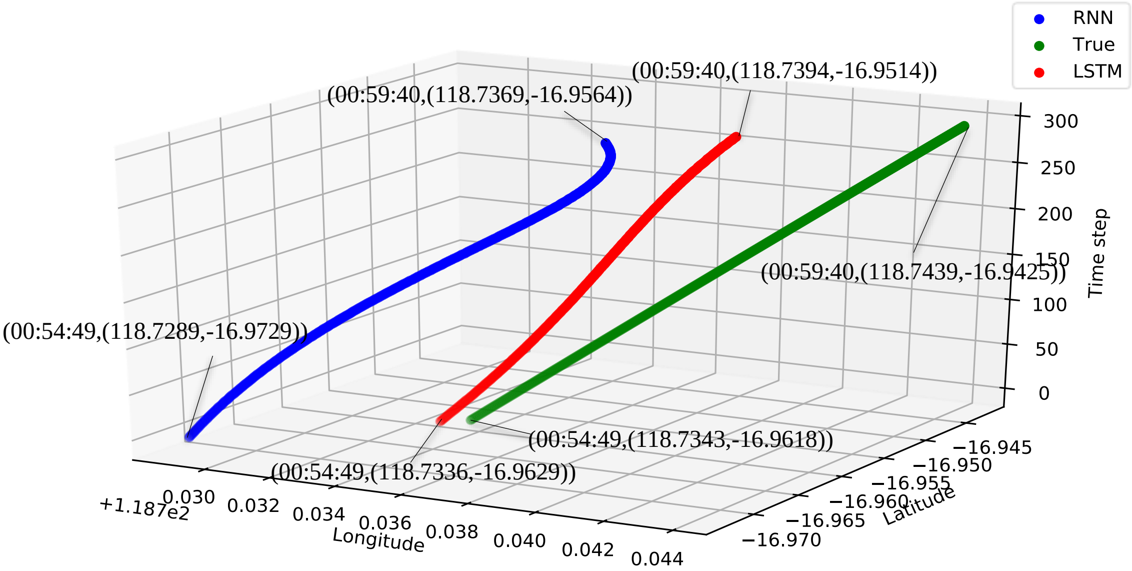


Figure 4.6 Spatial-temporal Distribution of True and Prediction Trajectory Results

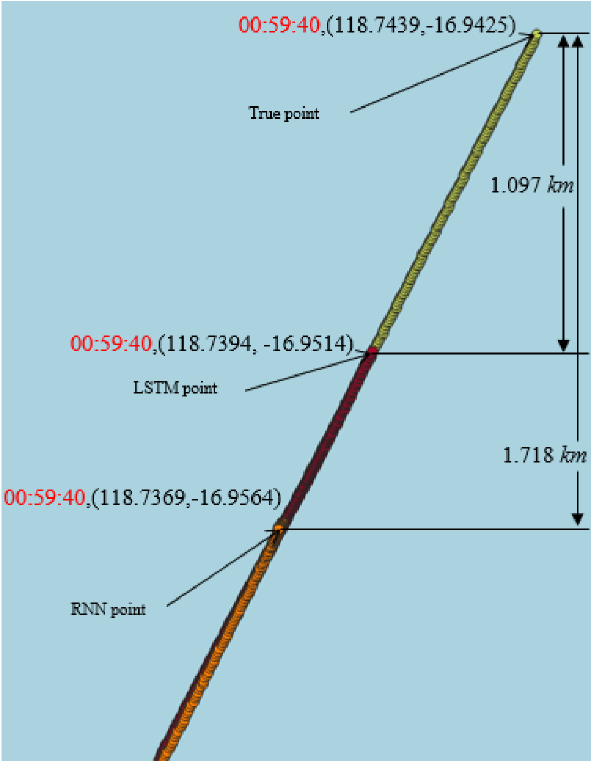


Figure 4.7 Geographical Distribution of Prediction Results Comparison

4.3 CO2 Emission Prediction

We calculated the carbon dioxide emissions of the vessel without interpolation by ITTC and used formula (28), we could get the emissions of the vessel at 00: 35: 10-00: 59: 40 was 47,169 kg, but as shown in Figure 4.8 and Figure 4.9, we calculated the vessel carbon dioxide emissions after interpolation was 74,926 kg. This was 27,757 kg more carbon dioxide emissions than traditional estimation methods. Therefore, our proposed model improves the accuracy of the vessel carbon dioxide emissions. Observe the vessel carbon dioxide emissions from spatial-temporal perspective, it goes down first and then goes up.

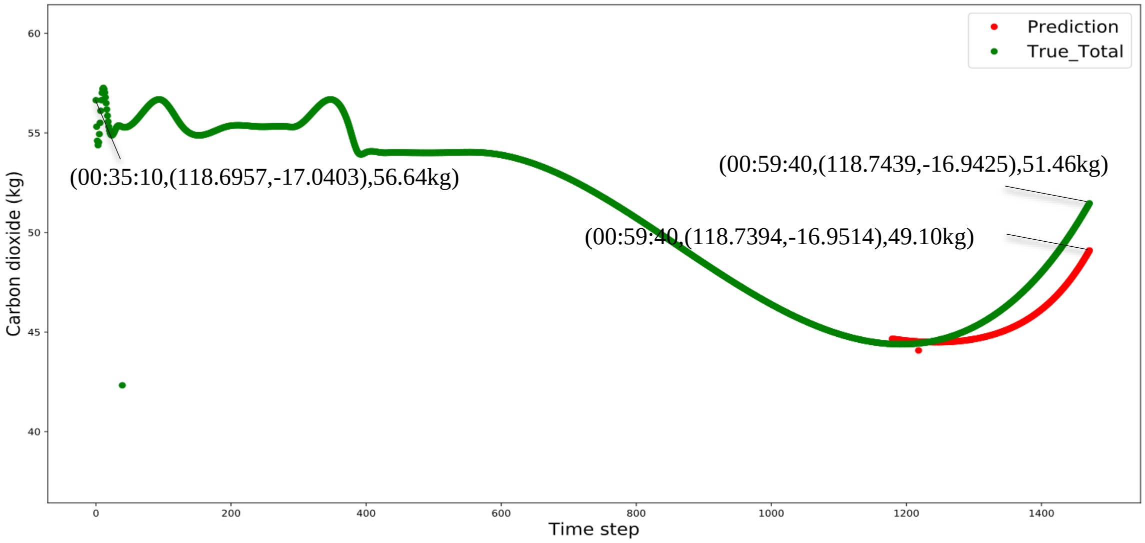
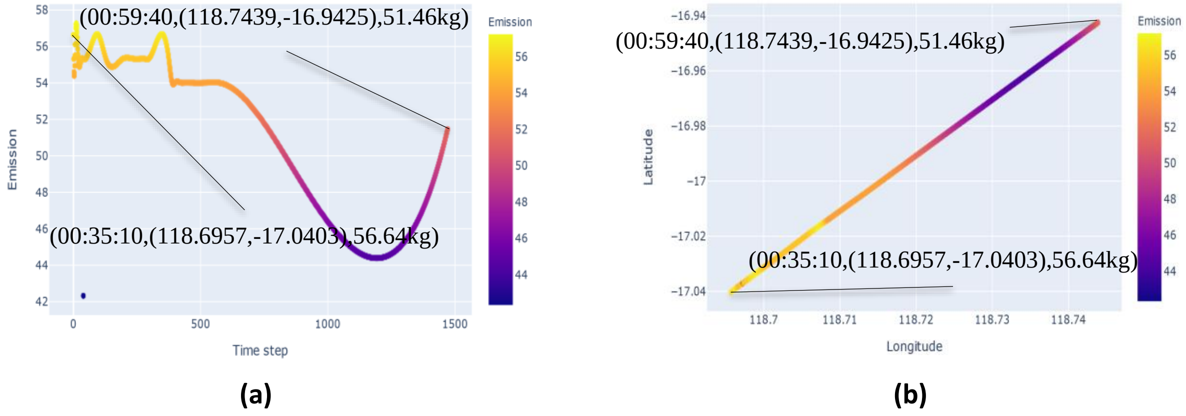


Figure 4.8 Temporal Distribution of Total True and Prediction Results Comparison

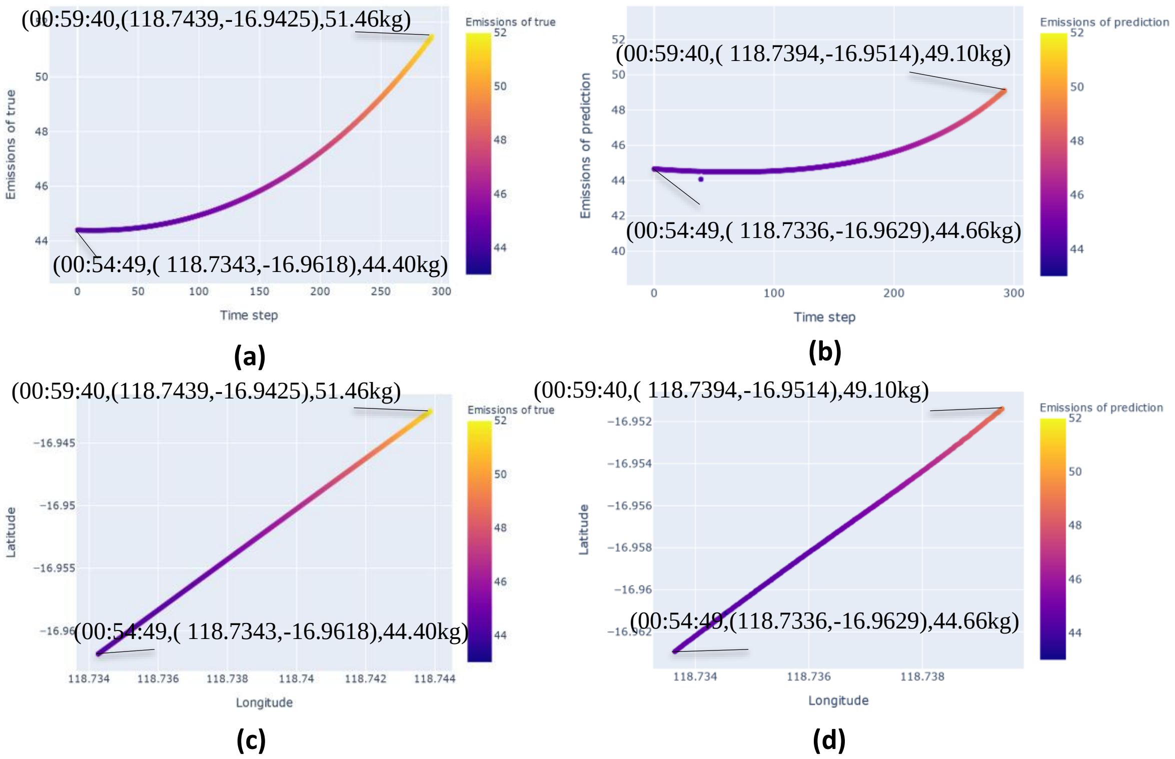


(a)

(b)

Figure 4.9 Spatial-temporal Distribution of Total True Results

Next, we used the trajectories predicted by the LSTM model to estimate carbon dioxide emissions to obtain the spatial-temporal distribution of future carbon dioxide emissions from vessels. As shown in Figure 4.8 and Figure 4.10,the predicted carbon dioxide emissions were 13,315 kg and the true was 13,616 kg, error was 301 kg. We could see heat map of spatial-temporal distribution of carbon dioxide emissions from vessels from 00:54:49 to 00:59:40 every second through Figure 4.10. Observing the vessel carbon dioxide emissions from a time perspective, the actual and predicted values had been increasing during this period. We can find that the growth rate of the predicted value was slower than the growth rate of the real value. At 00:59:41, the error was 2.36 kg.



(a) (b)

(c) (d)

Figure 4.10 Spatial-temporal Distribution of True and Prediction Results

Because we knew the vessel emissions per second, we calculated what the cumulative carbon dioxide emissions from vessels were every 1 second. The vessel cumulative carbon dioxide emissions from 00: 54: 49-00: 59: 41 were 13315 kg. It was predicted that the highest value will be reached in this area (118.7439, -16.9425).

Besides, it was found that the vessel emissions in the future period would continue to increase. The deceleration trend indicates that the vessel will accelerate to the northeast in the future. On the other hand, because we had the data of the carbon dioxide emissions of the vessel every second, we can re-establish any time interval to reduce the number of vessel trajectory points, the larger the trajectory point interval, the more conducive to trajectory aggregation. This was useful for calculating a larger number of vessel carbon emission trajectories for example, to calculate the carbon dioxide emissions of the vessel every 3 seconds, 6 seconds or 10 seconds. We can observe the temporal distribution of the carbon dioxide emissions of the vessel every 3 seconds, every 6 seconds, and every 10 seconds from Figure 4.11.We can see from Figure 4.11 that the larger the number of time steps, the faster and faster the rate of increase of carbon dioxide emissions, which also showed that the SOG of vessel was getting faster and faster. At the same time, the number of trajectory points decreases as the time interval decreases, the vessel carbon dioxide emissions represented by each track point were gradually increasing, but the overall trajectory trend remained the same.

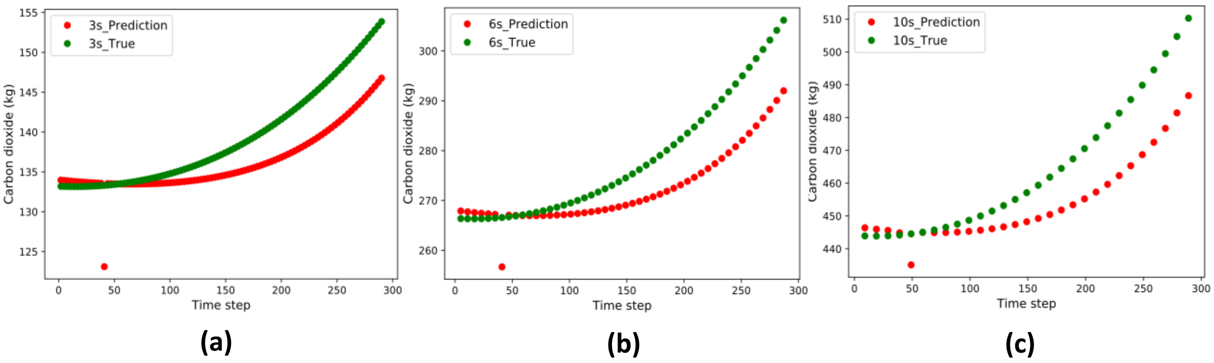


Figure 4.11 Spatial Distribution of True and Predicted Carbon Dioxide Emissions

AIS had too long time interval for vessel data collection due to equipment and man-made reasons, it made the vessels carbon dioxide emissions estimated based on AIS data had large error, therefore, this study proposes using cubic spline interpolation to resample missing AIS data with smooth curves, compared with the simple linear (slinear) interpolation method, it was more in line with the actual situation. According to Figure 8, the trajectory of the repaired vessel was basically the same as the original trajectory, so we can successfully obtain AIS data with a time interval of 1 second. Based on the trajectory data of the repaired vessel (1,472 pieces of data), we estimated the carbon dioxide emissions of a vessel. The value was 74,926 kg. Compared with the unrepaired trajectory data (23 pieces of data), 27,757 kg of carbon dioxide were added. This showed that with the improvement of the accuracy of carbon dioxide estimation, the actual carbon dioxide emissions will become more, which was conducive to the management of vessel carbon emissions by maritime management agencies. In addition, we also used the spatial-temporal data of vessels that have been constructed at regular intervals, combined with deep learning for trajectory prediction, and compared the LSTM model with the RNN model, confirming that the performance of the LSTM model was better than the RNN model. Therefore, we used 938 pieces of data to train the LSTM model, and finally, 235 pieces of data were used to verify the prediction effect. At an equal time (00:59:40), the predicted trajectory point of the LSTM model differs from the actual trajectory point by 0.593 nm, which was in line with expectations. Besides, we used the predicted SOG to predict the future carbon dioxide emissions of the vessel. The error was 301 kg, which was in line with expectations. This showed that the LSTM model was suitable for spatial-temporal data prediction and had excellent performance

Therefore, the method proposed in this study can further increase the estimated amount of vessel carbon dioxide emissions and can monitor the vessel carbon dioxide emissions and vessel trajectory in real-time and can provide vessels with early warning services of carbon dioxide emissions. In addition, the predicted value can also be used as an alternative. If the vessel to shut down AIS for artificial reasons (including active or passive) or that the AIS collection data fails at 00:54:49 to 00:59:40, managers can use this method to re-import the carbon dioxide emissions trajectory.

5. Conclusion

The purpose of this study is to assess and monitoring the maritime traffic conditions, using AIS data mining technology to investigate the ship traffic characteristics within the area and between single ships and summarize the sailing characteristics. In addition, on the basis of the survey on the basic characteristics of marine traffic, it proposes the use of ITTC to estimate the ship's greenhouse gas emissions, to monitor and track the ship's GHG emissions behavior, at the same time, it also proposes to use the LSTM method to predict the ship's future location and carbon dioxide emissions, to achieve the performance of enhancing the safe navigation of ships in narrow seas, and to enhance the supervision of ships' greenhouse gas emissions.

Mainly do the following theoretical research:

1) Point out the quality of AIS data and summarize the cleaning method of AIS data;
2) Summarize the repair method of AIS data-interpolation method;
3) Summarize the role of deep learning methods in AIS data prediction;
4) Summarize ITTC's algorithm characteristics and calculation process;
5) The literature is combed from three aspects: ship emission inventory, in-depth learning, and marine traffic flow.

Mainly made the following case studies:

1) Use K-mean to study the relationship between the Malacca Strait ship course-speed-draught-heat and other traffic characteristics;
2) Using the method of maritime traffic flow to study the traffic flow of the Strait of Malacca;
3) Distribution of greenhouse gas emissions from LNG vessels in the Straits of Malacca from January to June 2016;
4) Track the greenhouse gas emission footprint of four ships in the Strait of Malacca;
5) Study the prediction of single ship trajectory and navigation status (position and CO2 emissions) on the Japan-Australia route.

Finally, with the economic development and global warming, we believe that the prediction of the ship's exhaust emission status and ship's navigation status is becoming more and more important and useful in the assessment or investigation of marine traffic conditions.

1) Through the AIS data, the characteristics of ship traffic within a certain area and between single ships can be truly investigated.
2) In addition, based on the investigation of the basic characteristics of maritime traffic, ITTC is used to estimate the ship’s greenhouse gas emissions to monitor and track the ship’s GHG emissions behavior.
3) It is also recommended to use the LSTM model to predict the future location and carbon dioxide emissions of the ship. In order to enhance the safety of ships in narrow seas and strengthen the supervision of ships greenhouse gas emissions.

Future Study:

1) We can use AIS data combined with deep learning to predict the arrival time of ships, and predict the harmful gas emissions of ships within the port area in advance, decide whether to let the ship enter the port area, and notify the ship in advance of the tax on harmful gas that needs to be paid;
2) We can also use cubic spline interpolation to repair a wider range of ship routes, combined with ITTC to estimate the harmful gas emissions of the ship throughout the voyage cycle, which can determine which country the ship emits the most in the sea area, and thus judge the environmental protection of each country Management level at work.
3) We can also estimate the payload of ships, especially energy ships, from the draft data in the AIS data. Combined with the destination of the ship, this can estimate the energy imports of each country.

Acknowledgement

I would like to express my sincere gratitude to my advisor. Assoc. Prof. Daisuke Watanabe has given me great support, helping me to participate in academic conferences in research, providing various research books, and guiding my thesis writing. In life, he helped me to actively integrate into the life of the laboratory, and led me to experience the Japanese culture, so, with his help, I grew up quickly, the only thing is that I did not learn Japanese etiquette. There will be some mistakes in life, but Professor Daisuke Watanabe taught me patiently, thank you very much! Next I want to thank all the people who supported me to study in Japan. With your encouragement, I can persevere. At last I would like to thank the Japanese classmates in the laboratory. I am very happy to study with them.

References

Altan, Y., Otay, E., 2017, Maritime Traffic Analysis of the Strait of Istanbul based on AIS data, The Journal of Japan Institute of Navigation (2017), 70, 1367–1382.

Dertat. A., 2017, Applied Deep Learning-Part 4: Convolutional Neural Networks,

https://towardsdatascience.com/@ardendertat.

Dertat. A., 2017, Applied Deep Learning-Part 3: Autoencoders,

https://towardsdatascience.com/applied-deep-learning-part-3-autoencoders-1c083af4d798.

Borkowski, Piotr, 2017, The Ship Movement Trajectory Prediction Algorithm Using Navigational Data Fusion. Sensors, 10.3390/s17061432, 1432.

Bartels, R. H, Beatty, J. C, and Barsky, B. A, Hermite and Cubic Spline Interpolation: Ch. 3 in An Introduction to Splines for Use in Computer Graphics and Geometric Modelling, San Francisco, Morgan Kaufmann, pp. 9-17, 1998.

Coello, J., Williams, I., Hudson, D. A., Kemp, S., 2015, An AIS-based approach to calculate atmospheric emissions from the UK fishing fleet, Atmospheric Environment, 114, 1-7. doi:10.1016/j. atmosenv.

Fujii, Y., 1970, Data processing of marine traffic survey, The Journal of Japan Institute of Navigation (1970), 32,67-68.

Hubel D H, Wiesel T N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. [J]. Journal of Physiology, 1962, 160(1):106.

Hopfield, John. (1982). Neural Networks and Physical Systems with Emergent Collective Computational Abilities. Proceedings of the National Academy of Sciences of the United States of America. 79. 2554-8. 10.1073/pnas.79.8.2554.

Hochreiter, S., Schmidhuber, J., 1997, Long short-term memory, Neural Comutation, 9(8): 1735-1780.

Hinton G E, Osindero S, Teh Y W. A Fast Learning Algorithm for Deep Belief Nets[J]. Neural Computation, 2014, 18(7):1527-1554.

Huang, Y., Yip, T., al et., 2019, Comparative analysis of marine traffic flow in classical models, Ocean Engineering, 187(2019): 106-123.

Kim, H., Watanabe, D and Toriumi, S., Spatial analysis of AIS-based LNG fleet emission inventory, Proceedings of International Forum on Shipping, Ports and Airports (IFSPA2019), Paper ID M53, 1-13, 2019.

Laxhammar, R., 2009, Anomaly detection in sea traffic-a comparison of the gaussian mixture model and the kernel density estimator. IEEE, 756-763.

Ljunggren, Henrik, 2018, Using Deep Learning for Classifying Ship Trajectories.

Li Y.P., 2019, Research on Major Characteristics of Marine Traffic Based on AIS Data[D], Dalian Maritime University.

Minami, M., Kikuchi T., Itoh H., 2014, The Function of the Hazard map to reduce Marine Accidents, The Journal of Japan Institute of Navigation, 131(129): 100-105.

Dewan. M. H., 2014, Ship Construction- Ship Dimensions, 6, https://www.slideshare.net/MohammudHanifDewan/ship-construction-ship-dimensions

MAN Diesel & Turbo (2011), Basic Principles of Ship Propulsion, https://marine.mandieselturbo.com/docs/librariesprovider6/propeller-aftship/basic-principles-of-propulsion.pdf?sfvrsn=0

Molland, A. F., Turnock, S. R., Hudson, D. A., 2016, Ship resistance and propulsion: Practical estimation of ship propulsive power, Cambridge: Cambridge University Press.

Meijer, R., 2017, Predicting the ETA of a container vessel based on route identification using AIS data, MOT TU Delft, Master Thesis.

Nguyen, D. D., Van, C. L., Ali, M. I., 2018, Vessel Trajectory Prediction using Sequence-to-Sequence Models over Spatial Grid, The 12th ACM International Conference on Distributed and Event-based Systems, Article 4, 4 pages.

Olah. C., 2015, Understanding LSTM Networks, https://colah.github.io/posts/2015-08-Understanding-LSTMs/.

Quan B., Yang B. C., Hu K. Q., Guo C. X., Li Q. Q., 2018, Prediction model of ship trajectory based on lstm, Computer Science, Vol.45 No.11A.

Rumelhart, D., Hinton, G. & Williams, R. Learning representations by back-propagating errors. Nature 323, 533–536 (1986).

Sathasivam, Saratha & Wan Abdullah, W.A.T. (2009). Logic Learning in Hopfield Networks. Modern Applied Science. 2. 10.5539/mas. v2n3p57.

Sérgiomabunda, A., Astito, A., Hamdoune, S., 2014, Estimating Carbon Dioxide and Particulate Matter Emissions from Ships using Automatic Identification System Data, International Journal of Computer Applications, 88(6), 27-31. doi:10.5120/15358-3823.

Smith, T. W. P., Jalkanen, J. P., Anderson, B. A., Corbett, J. J., Faber, J., Hanayama, S., Pandey, A., 2015, Third IMO Greenhouse Gas Study 2014.

Sang, L. Z, Yan, X. P., Wall, A., Wang, J. Mao, Z., 2016, CPA Calculation Method Based on AIS Position Prediction, LJMU Research.

Shirai, T., Kubo, N., et al., 2016, A Basic Study on Prediction of Ship Status at Tokyo Bay, The Journal of Japan Institute of Navigation, 135(133): 108-114.

Tasseda, E., Shoji, R., 2014, Trip distribution modeling in tokyo bay based on ais data, The Journal of Japan Institute of Navigation, 131(128): 1-8.

Wang Z. H., Wang W. X., SHI X., 2019b, AIS data-based ship emission estimation model and real ship verification, Journal of Shanghai Maritime University, DOI:10．13340/j．jsmu．2019．04．003.

Wang L. L., Liu, J., 2019b, Ship behavior recognition method based on multi-scale convolution. Journal of Computer Applications, 39(12): 3691-3696.

Winther, M., Christensen, J. H., Plejdrup, M. S., Ravn, E. S., Eriksson, Ó. F., and Kristensen, H. O, 2014, Emission inventories for ships in the Arctic based on AIS data, Atmos. Environ., 91, 1–14.

Yamin, H., Linying, C., Chena, P. F., 2019, Ship collision avoidance methods: State-of-the-art, Safety Science, 121(2020):451-473.

Yao, X., Mou, J., Chen, P., Zhang, X, 2016, Ship Emission Inventories in Estuary of the Yangtze River Using Terrestrial AIS Data, TransNav, the International Journal on Marine Navigation and Safety of Sea Transportation, 10(4), 633-640. doi:10.12716/1001.10.04.13.

Zhao S. B., Tang C., Liang S., Wang D. J., 2012, Track prediction of vessel in controlled waterway based on improved Kalman filter. Journal of Computer Applications, 32(11): 3247-3250.

Zhang, C., 2019, AIS data -driven general vessel destination prediction: a trajectory similarity-based approach, THE UNIVERSITY OF BRITISH COLUMBIA (Okanagan), Master Thesis.