

Received January 11, 2021, accepted January 15, 2021, date of publication January 18, 2021, date of current version January 27, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3052473

A Practical Estimation Method of Inland Ship Speed Under Complex and Changeable Navigation Environment

ZHI YUAN^{1,2,3}, JINGXIAN LIU^{1,2}, QIAN ZHANG³, (Member, IEEE),
YI LIU^{1,2}, (Member, IEEE), YUAN YUAN⁴, ZONGZHI LI⁵

¹Hubei Key Laboratory of Inland Shipping Technology, School of Navigation, Wuhan University of Technology, Wuhan 430063, China

²National Engineering Research Center for Water Transport Safety, Wuhan 430063, China

³Department of Electronics and Electrical Engineering, Liverpool John Moores University, Liverpool L3 3AF, U.K.

⁴ChangJiang Shipping Science Research Institute Company Ltd., Wuhan 430060, China

⁵Department of Civil, Architectural and Environmental Engineering, Illinois Institute of Technology, Chicago, IL 60616, USA

Corresponding authors: Qian Zhang (q.zhang@ljmu.ac.uk) and Yi Liu (liuyi_hy@whut.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 51709219, in part by the National Key Research and Development Program of China under Grant 2018YFC1407400, in part by the Science and Technology Department of Hubei Province under Grant 2019AHB053, in part by the China Scholarship Council under Grant 201906950086, and in part by the Qingdao Research Institute of Wuhan University of Technology under Grant 2019A02.

ABSTRACT The complexity and changefulness of inland navigation environment in space and time makes it hard to guarantee the applicability and accuracy of existing ship speed models. In this paper, a novel method for inland ship speed modelling under complex and changeable navigation environment is proposed. Firstly, an unsupervised machine learning algorithm, Density-Based Spatial Clustering of Application with Noise (DBSCAN), is utilized to cluster the environmental data including water level, water speed, wind speed and wind direction, to get the segment division information, which greatly helps reduce the influence of other uncertain environmental factors on the speed model. Then, Generalized Regression Neural Network (GRNN) is tailored and employed to build the ship speed estimation model with multiple input variables. Finally, a detailed case study of a ship sailing in the Yangtze River trunk line is conducted to validate the proposed methods. The results show that the ship speed model established based on machine learning methods works effectively in speed estimation and analysis. Moreover, compared with other regression methods and neural networks, the proposed GRNN model has the best performance in ship speed modelling.

INDEX TERMS Complex navigation environment, inland ship, speed modeling, DBSCAN, GRNN.

I. INTRODUCTION

Waterborne transportation, as a green and economical transportation mode, plays an essential role in worldwide trade. The waterway transportation system considering safe navigation, efficient transportation and energy saving puts forward higher requirements for ship speed modelling, control and optimization. Therefore, considering the influence of multiple factors to construct accurate and practical ship speed models has become a key research issue. The Yangtze River is the longest inland river in China, and also has the largest hydropower plants of the world. The freight volume of this river had achieved 1.92 billion tons in 2013, ranking first in

the world for nine consecutive years [1]. The Yangtze River trunk line has become the key water area of shipping research and is also the target water area of this paper.

The movement of a ship is caused by sufficient power output from the engine and is affected by resistance, including hydrostatic resistance, wind resistance, wave-induced resistance and shallow water resistance. In order to calculate the resistance and ship speed, some researchers used statistical analysis methods to summarize some mathematical formulas [2]–[7]. These formulas have been used in some later studies. Fang and Lin [8] used ship hydrodynamics formulas to calculate wave-induced resistance and wind loads in a ship weather-routing optimization algorithm. However, they neglected the heading errors caused by lateral forces or yawing moments due to winds and currents. Meng

The associate editor coordinating the review of this manuscript and approving it for publication was Giambattista Grusso.

et al. [9] applied the fundamentals to develop two regression models for container ship fuel efficiency. It was conducted based on the limited information conveyed by shipping logs. Yan *et al.* [10] used the methods to calculate hydrodynamic resistance forces, and established an optimization paradigm for ship energy efficiency. However, they only studied part of the voyage, and the data came from test ships. Li *et al.* [11] used Kwon's method to estimate involuntary speed loss, and established ship speed optimizations with and without voluntary speed loss for a single voyage. However, the speed studied in [11] refers to the ship speed in still water. In summary, the published formulas and methods have helped the researchers achieve certain research results. However, there are still some problems that cannot be ignored in the actual implementation of these empirical formulas: (1) For many parameters, it is difficult to select their values, and different values have a great impact on the calculation accuracy; (2) The data and information used in model elicitation are limited; (3) The influence of changes in environmental factors on the speed of the sailing ship is not considered.

In addition, it should be noted that the navigational environment of the Yangtze River is very complicated, which is mainly reflected in the following aspects:

(1) The width of the waterway is narrow and unevenly distributed. The narrowest area of the channel is only 50 meters, while the widest part reaches 500 meters.

(2) The route is often curved, as shown in Fig. 1. The bend angle of the local waterway is close to 90° , such as Yin Gongzhou Waterway in Zhenjiang City, Jiangsu Province.

(3) Across multiple elevations, the water level and water speed of different waterways are quite different.

(4) It is difficult to conduct monitoring and data collection, most of which come from hydrological stations and weather stations.

(5) Regional differences and seasonal changes are obvious.

There is no doubt that complex and changeable environmental factors and uncertain parameters in empirical formulas bring great difficulties to ship speed modelling. Unsupervised learning algorithms and Artificial Neural Networks (ANNs) have powerful learning ability and have been effectively applied in numerous fields for knowledge discovery, data classification and time series analysis [12]–[17]. Moreover, ANNs also show great advantages in data-driven modelling. Du *et al.* [18] built the feedforward ANN model for fuel efficiency of vessels based on voyage report records, and realized vessel speed optimization. Kim *et al.* [19] developed an Artificial Neural Network-based storm Surge Forecast Model (ANN-SFM) with the 5, 12 and 24 h-lead times, and applied it to the Sakai Minato region of Japan. Vieira *et al.* [20] presented an alternative method for filling missing data based on publicly available wind and wave information using ANNs. To decrease the computational complexity and increase the precision in forecasting failure envelopes of caisson foundations, Zhang *et al.* [21] proposed a method of Random Forest (RF) to study the data extracted from calibrated experiments and simulations. Uyanik *et al.* [22] applied vari-

ous machine learning methods to establish prediction models for a container ship, and realized fuel consumption estimation. Yuan *et al.* [23] analyzed three important tasks using the Long Short-Term Memory (LSTM) neural network, including engine speed and fuel consumption modelling, and vessel trajectory repair. Hence, ANNs have been well employed in the large shipping field. However, few studies have focused on implementing machine learning methods to the specific problem of Inland River ship speed modelling.

Moreover, the ANNs do not need to select subject-related parameters in advance, it can be used to solve the problems in traditional ship speed modelling, such as the difficulties in the selection of parameters and the analysis of environmental impacts. Therefore, this paper proposes a novel method for modelling the speed of inland ships. First, a salient unsupervised machine learning algorithm, Density-Based Spatial Clustering of Application with Noise (DBSCAN), is used to perform cluster analysis on the closely related environmental data, and the results are quantified into specific voyage segment information. Subsequently, an accurate and robust ANN, Generalized Regression Neural Network (GRNN), is tailored to build high-precision ship speed model for inland rivers. Then, the measured data including navigation status data, environmental data and segment information are learned in models' training to find out the optimal model parameters.

In recent years, unsupervised learning algorithms have been used for cluster analysis of multi-source data [24]. Similarly, we can use DBSCAN to analyze the environmental data of Inland River to reveal their internal connections. DBSCAN is a density-based clustering algorithm, which has only two parameters and runs very fast, because it merely need a linear number of range queries in data processing [25]. The DBSCAN algorithm has found a range of applications, as it is able to obtain clusters with arbitrary shapes and it does not require to predefine the number of clusters in advance [26]–[29]. Luchi *et al.* [30] presented two methods to generate good samples for the DBSCAN algorithm, so that it can be better applied to large data set sampling. Liu *et al.* [31] used DBSCAN to simplify the computation of a framework of regional collision risk identification. Sheridan *et al.* [32] applied the DBSCAN clustering to flight trajectory analysis, and realized flight anomaly detection during the approach phase. Wen *et al.* [33] integrated DBSCAN and ANN capable of automatic ship route design based on massive AIS data between certain ports. Liu *et al.* [34] selected DBSCAN to distinguish the normal points and unwanted outliers in data processing, and realized the accurate detection of the timestamped points degraded with random outliers in vessel trajectories.

On the other hand, GRNN has shown good performance in data-driven modelling [35]. It is a one-pass learning algorithm with a highly parallel structure [36]. Valčić and Prpić-Oršić [37] proposed a hybrid method for estimating wind loads based on elliptic Fourier descriptors (EFD) and GRNN, and obtained promising results. Borkowski [38] presented an algorithm based on GRNN, and realized the

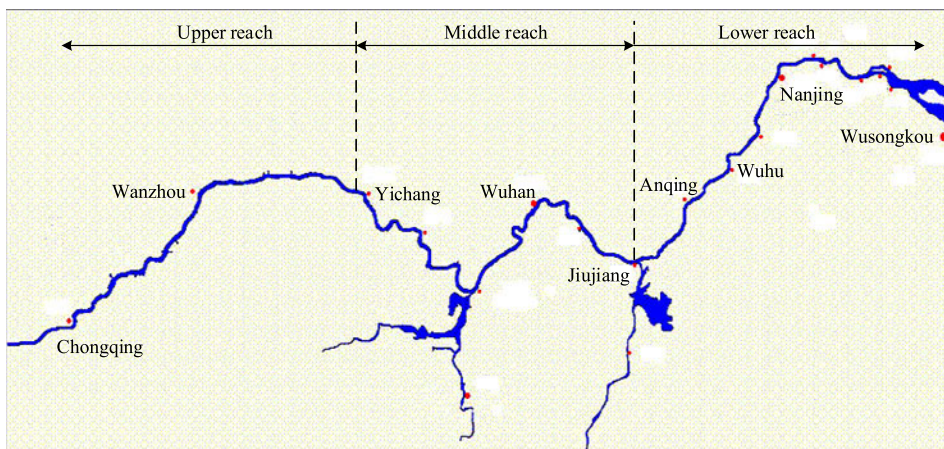


FIGURE 1. The Yangtze River trunk line: the key water area for shipping research.

prediction of ship movement trajectory. Liang *et al.* [39] used GRNN to build a hybrid model for short-term load forecasting, which enhanced prediction accuracy. Parveen *et al.* [40] developed the data-driven models to anticipate the bed depth profile of solids flowing in a rotary kiln using GRNN and other networks, and accurately predicted the parametric effects on bed depth profile. Cepowski [41] developed an ANN model to predict ship added resistance using GRNN, and obtained the wave resistance coefficient.

To sum up, the machine learning algorithms DBSCAN and GRNN have been widely used in cluster analysis and data-driven modelling, which have achieved good results. Therefore, this paper takes the ship sailing on the Yangtze River as the research object, and collects real-time status data and related environmental data. First, the DBSCAN algorithm is used to conduct cluster analysis on environmental data, which can get the information of voyage segment division. Then, GRNN is tailored and employed to build the ship speed model under the complex environment. Finally, the constructed models are verified and analyzed by the measured data set including navigation status data, environmental data and segments information. The research framework is shown in Fig. 2.

The main contributions of this study are as follows. (1) The machine learning methods are applied to build the speed model of the inland ship, which not only improves the performance of the speed model, but also increases the utilization efficiency of inland river transportation data. (2) The proposed method avoids the trouble of parameters selection in the traditional formulas for calculating ship speed, and reduces the influence of uncertain environmental factors on the ship speed model. (3) The efficient ship speed model under complex environment we constructed provides support for route planning, collision avoidance, fuel consumption optimization and operational benefit analysis. It is also helpful to promote the high-quality development of inland shipping.

The remaining of the paper is organized as follows: Section II collects the real-time status data and environmental data of the ship sailing on the Yangtze River. Section III designs a new voyage division algorithm using DBSCAN clustering, and introduces the ship speed modelling method in detail. Section IV provides the case study to verify the proposed method and compare it with other modelling approaches. Section V summarizes the study and suggests some future research directions.

II. DATA COLLECTION AND PRE-PROCESSING

In this work, the research data were collected from a bulk ship sailing on the Yangtze River trunk line, and the basic parameters of the ship are as shown in Table 1. These data include real-time status data and environmental data, which were collected from the multi-source sensors installed on the ship and hydrometeorological stations. The raw data set includes 32,143 records, which come from a complete voyage from September 11, 2019 to October 7, 2019. The departure port of the voyage is Wusongkou of Shanghai, and the destination port is Chongqing, which runs across upper reach, middle reach and lower reach areas of the Yangtze River trunk line, as shown in the Fig. 1. It is worth noting that there are many abnormal, errors and noises in the navigation status data, which were collected by multiple sensors in real-time. The environmental data were collected from hydrological and weather information released daily by hydrometeorological stations, including water level, water flow, wind speed and wind direction, which need to be further quantified. Therefore, the process of data collection and pre-processing is as shown in Fig. 3.

A. NAVIGATION STATUS DATA

In addition to static information, the navigation status data (ship dynamic data) are more important for speed modelling, which reflects the real-time status of the ship. Ship navigation status data mainly include date, time, latitude, longitude, Course Over Ground (COG), Speed Over Ground

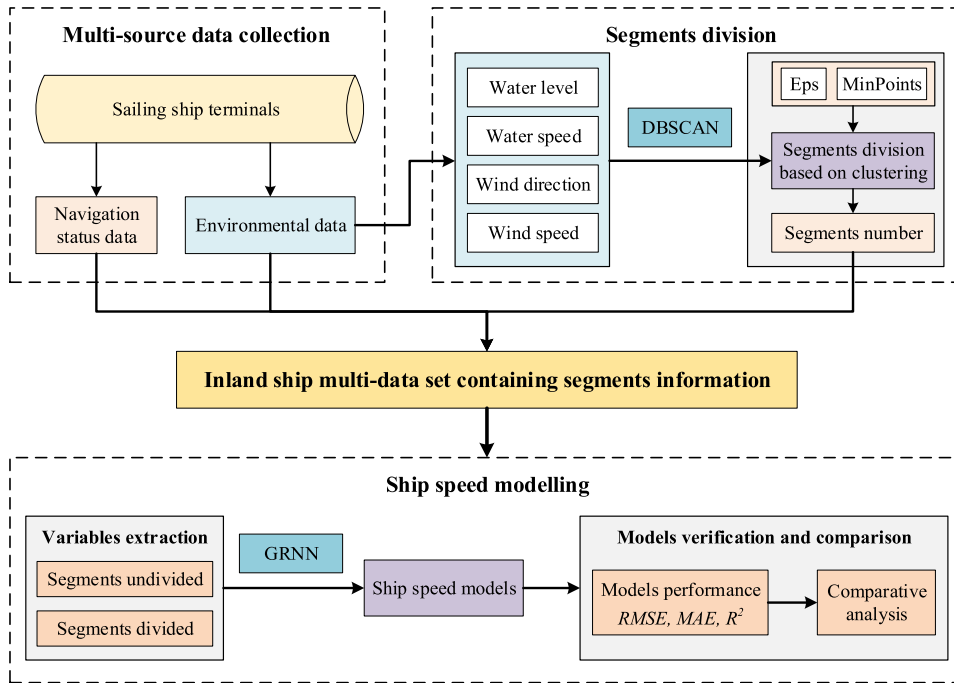



FIGURE 2. The framework diagram of speed modelling for inland ships.

TABLE 1. Basic parameters of the target ship.

Ship	Parameters	Values
	Designed length	110 m
	Moulded breadth	19.20 m
	Moulded depth	5.60 m
	Designed draught (full load)	4.65 m
	Designed hydrostatic speed (full load)	18 km/h
	Maximum engine power	735 kW × 2
	Maximum engine speed	830 rpm

(SOG), engine speed, engine temperature, mileage, main ports, etc. These data come from an inland ship monitoring system which includes shipboard multi-source sensors and the Global Positioning System (GPS) device. For the convenience of explanation, we mark two engines fitted to the target ship as left engine and right engine. It is worth noting that the navigation status data here include the temperature of two engines. Because we believe that they also truly reflect the running status of the engines and can be related to the ship speed. Like other measured data, the status monitoring data have some problems, such as data redundancy, noise interference and data missing, as shown in Fig. 4. One of the reasons lies in that the data come from different sensors and devices and this often makes the data collection asynchronous.

Fig. 4 shows the longitude, latitude, SOG, left engine speed and right engine speed of the original sampled data. The abscissa represents the sampling time with an interval of one minute, from which we can see that the original sampled

data contain a lot of noise and outliers, as marked by the red rectangles and ellipses. In Fig. 4(a), the regular range for longitude is from 105 to 125 ° E, where zeros are clearly abnormal values, while the latitude values are all normal. In Fig. 4(b), there are many zero values for SOG, which are noise. In Fig. 4(c) and Fig. 4(d), in addition to a lot of noise data with zero values, there are also some erroneous data below the normal range, which are shown in the red ellipses. Hence pre-processing the raw data becomes necessary. It must be aware the data collected by different sensors have various attribute characteristics and value ranges, and cannot be processed in the same way. For example, in a data record with abnormal longitude, the SOG is normal. In the same data record with a normal SOG, the engine speed may be abnormal. If these data records with locally abnormal are directly deleted, it will make the secondary loss of valuable information. To obtain clean data, meanwhile, keep valuable information as much as one can, the data pre-processing approach is developed as Fig. 5.

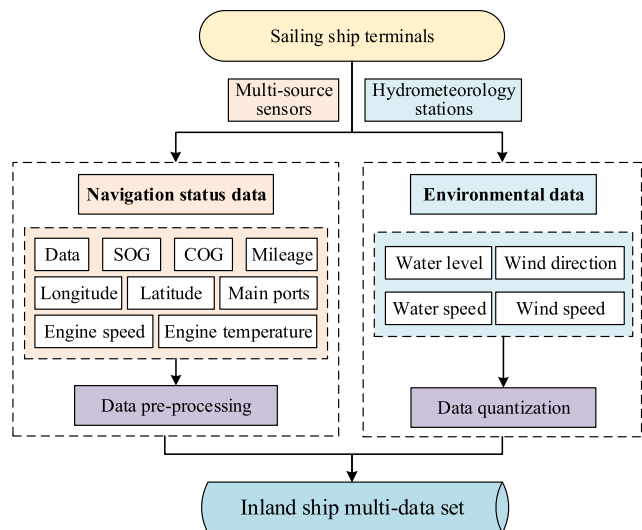


FIGURE 3. The process of inland ship multi-data collection and pre-processing.

As demonstrated in Fig. 5, the proposed data pre-processing approach consists of 5 steps, as follows:

Step 1: Sorting data according to sampling time. Firstly, all data are sorted to ensure the chronological order of the sampling data.

Step 2: Deleting duplicated data. Duplicated data should be deleted to improve data utilization.

Step 3: Extracting the characteristic data. The characteristic data include longitude, latitude, SOG, etc., which have different value ranges, need to be dealt with separately.

Step 4: Identifying and cleaning problematic data. The problematic data include abnormal data such as longitude with zero values in Fig. 4(a), error data such as engine speed in Fig. 4(c) and Fig. 4(d), and noise data such as zero values for SOG and engine speed. It must be noted that the data with different problems should be processed in different ways, such as abnormal data repair, error data removal and noise data filtering.

Step 5: Integrating multiple data. Finally, the cleaned multiple characteristic data should be integrated into one data set to prepare for ship speed modelling.

B. NAVIGATION ENVIRONMENT DATA

The navigation environment has a great influence on the status data of the sailing ship. In particular, the navigation environment of inland rivers is complicated, and the environmental factors that affect the speed of sailing ships mainly include wind and current. However, the equipment for measuring navigational environment data is expensive and cannot be reused, it is normally not installed in ships except for the test ships. However, the captain can receive the hydrometeorological data of the waterway where the ship is sailing, including real-time water level, water speed, wind speed, and wind direction. In this paper, the environmental

TABLE 2. Wind angle corresponding to the direction.

Direction	Angle	Direction	Angle
North	0°/360°	South	180°
Northeast	45°	Southwest	225°
East	90°	West	270°
Southeast	135°	Northwest	315°

data of many key waterway nodes in the actual route of the ship are collected. Among them, the water level data are based on the Yellow Sea Datum, and the unit is meter (m); the unit of water speed is meter per second (m/s). The wind direction includes north, northeast, east, southeast, south, southwest, west and northwest, and the unit of wind speed is Beaufort scale (BS). In order to be effectively used in later modelling, the wind directions are quantified into specific angles as shown in Table 2. The environmental data collected in this paper are as shown in Fig. 6.

III. METHODS

In this paper, DBSCAN and GRNN are employed to construct the accurate ship speed model under the complicated environment, as shown in Fig. 7. Firstly, DBSCAN is used for environmental data clustering to get segment numbers. Then, GRNN is tailored and implemented to build the SOG model. To analyze the influence of multiple sources variables on the SOG modelling, some groups of feature variables are extracted as inputs of the model. Initially, the variables without segment division are divided into three groups and presented to the model. Then, the information of different segments division is added to the input variables, and the corresponding results are obtained and analyzed. Finally, the performance of the constructed SOG models are verified by the measured data and compared with other methods. In order to verify the applicability of the proposed method in different scenarios, some cases with different numbers of training and testing data are studied and analyzed.

A. SEGMENTS DIVISION BASED ON DBSCAN

In Section II, we successfully obtained the ship’s navigation status data and the environmental information of the main waterways in the voyage, including water level, water speed, wind speed and wind angle. However, the Yangtze River has more complicated navigation environment. It also contains other factors that have some influence and are not easy to measure, such as waves. Some areas are also affected by tides. In order to reduce the impact of uncertain environmental factors on the ship speed model, the DBSCAN algorithm is adopted to divide the trajectory into some segments through the acquired environmental variables.

DBSCAN is an unsupervised machine learning algorithm that can find all the dense areas of the input sample points and treat these dense areas as clusters one by one [42]. The DBSCAN algorithm has three advantages: (1) It is not necessary to know the number of clusters beforehand; (2)

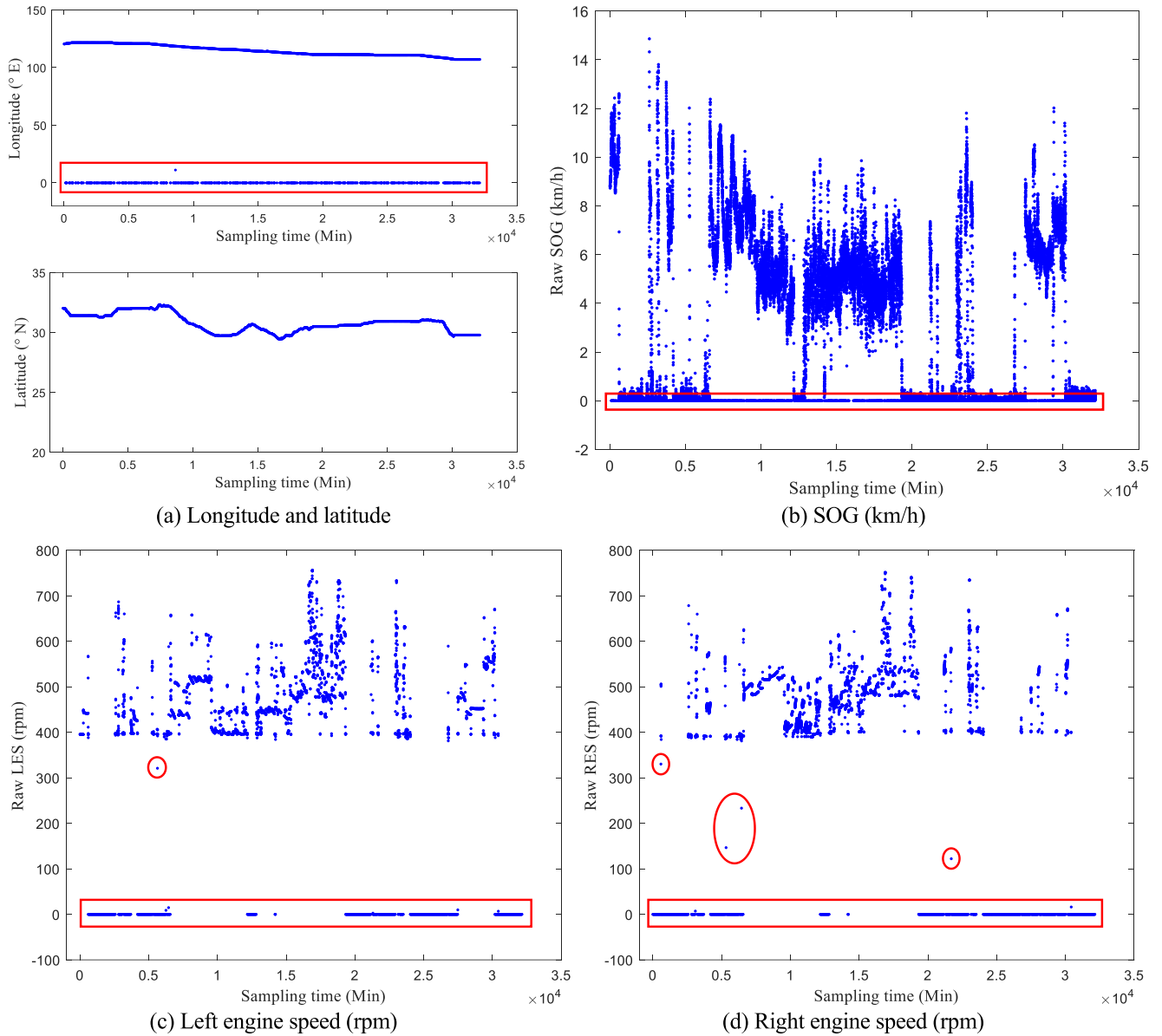


FIGURE 4. Raw sampled data: (a) longitude and latitude; (b) SOG; (c) left engine speed; (d) right engine speed.

It works on the basis of density information, which is not sensitive to abnormal points in the data set; (3) It can obtain clusters with an arbitrary shape. The content of DBSCAN can be summarized as “one, two, three and four”, as follows.

One core idea: DBSCAN clustering is a density-based spatial clustering, which divides the areas with sufficient density into different clusters. It can obtain clusters with an arbitrary shape from the data with noise.

Two algorithm parameters: Eps, neighbourhood radius, specifies the value of neighbourhood radius of each object. MinPoints, the threshold of neighbourhood density, is the minimum number of points in each cluster.

Three classes points: There are core point, border point and noise point. The core point is where the number of sample points is no less than MinPoints within Eps. The border point

is not a core point but is in the Eps of another core point. The noise point is neither the core point nor the border point. The three classes’ points of DBSCAN are as shown in Fig. 8.

Four point relationships: These are directly-density-reachable, density-reachable, density-connected and non-density-connection relationships. For the sample points p and q , if p is a core point and q is within Eps of p , which is $NEps(p) \geq MinPoints, q \in N(p)$, then the relationship between p and q is directly density-reachable. It should be noted that the relationship between any core point and itself is directly density-reachable, and the relationship of directly density-reachable is not symmetrical. For the points p_1 and q , if there are core points p_2, p_3, \dots, p_n , and the relationships of p_1 to p_2, p_2 to p_3, p_{n-1} to p_n , and p_n to q are directly density-reachable, then the relationship between

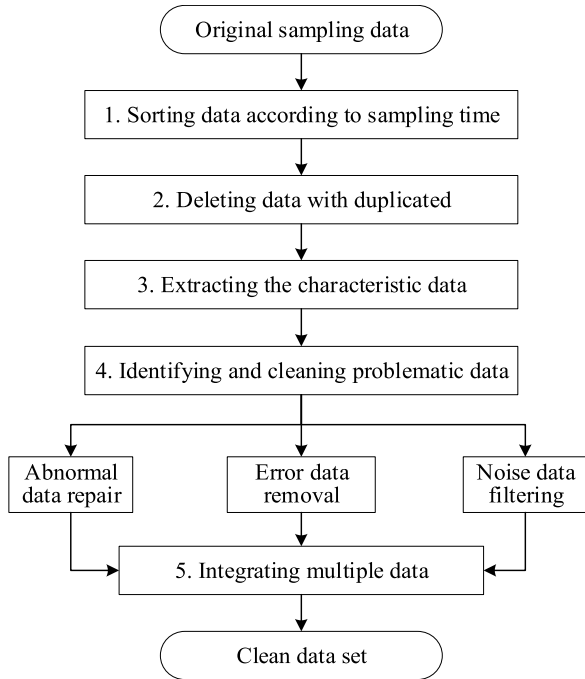


FIGURE 5. The mechanism of pre-processing inland ship status data.

p_1 and q is density-reachable. The relationship of density-reachable is also not symmetrical. If there is a core point s makes s to p and s to q density-reachable, then the relationship between p and q is density-connected. Different from directly density-reachable and density-reachable, the density-connected relationship is symmetrical, and two points that are density-connected belong to the same cluster. If two points do not belong to the relationship of density-connected, then the relationship is non-density connection. It should be noted that the two points with non-density connection belong to different clusters.

On the basis of the above definition, we regard a group of environmental variables as objects, and design the segments division algorithm based on DBSCAN as described in Algorithm 1.

B. SHIP SPEED MODELLING USING GRNN

GRNN is a radial-basis-function neural network with forward propagation. The GRNN has three advantages: (1) based on the radial basis function, it has good nonlinear approximation performance; (2) it does not need to back propagate to find model parameters, and its convergence speed is fast; (3) it has good mapping ability for few samples and unstable data. The network structure of GRNN has four layers: an input layer, a pattern layer, a summation layer and an output layer. For a data set with m samples, the feature set is $\{X_1, X_2, \dots, X_m\}$, $X_i = [x_i^1, x_i^2, \dots, x_i^n]$, $i = 1, 2, \dots, m$, the label set is $\{Y_1, Y_2, \dots, Y_m\}$, $Y_i = [y_i^1, y_i^2, \dots, y_i^k]$, $i = 1, 2, \dots, m$. The structure of GRNN can be described as Fig. 9, where n denotes the dimension of each feature sample, k denotes the dimension of each label sample.

Algorithm 1 Segment Division Algorithm Based on DBSCAN

Inputs: $EnvirDataSet$, Eps , $MinPoints$

Output: $\{S_1, S_2, \dots, S_k\}$, segments data set, k represents the clusters

- [1] Normalize the $EnvirDataSet$ to $NorEnvirDataSet$
- [2] Make all objects in $NorEnvirDataSet$ as unvisited
- [3] **for** (each object p in $NorEnvirDataSet$) **do**
- [4] **if** (p has been classified into a cluster or marked as a noise point) **then**
- [5] **continue**
- [6] **else**
- // check the $NEps(p)$ (neighbourhood of p)
- [7] **if** ($NEps(p) < MinPoints$) **then**
- [8] Make p as a border point or a noise point
- [9] **else**
- [10] Make p as a core point
- [11] Create a new segment cluster S
- [12] Add all objects in $NEps(p)$ into S
- [13] **for** (each unvisited object q in $NEps(p)$) **do**
- // check the $NEps(q)$ (neighbourhood of q)
- [14] **if** ($NEps(q) \geq MinPoints$) **then**
- [15] Add each unclassified object in $NEps(q)$ into S
- [16] **end if**
- [17] **end for**
- [18] **end if**
- [19] **end if**
- [20] **end for**

The working principle of each network layer is as follows.

Input layer: It is used to input the sample data, and the number of nodes is equal to n , the dimension of a feature sample.

Pattern layer: It is used to calculate the value of the Gaussian function of each sample in the training samples and the testing samples. The Gaussian value is the output value of the nodes in this layer. The number of nodes is m , the number of training data. The Gaussian function value of the i^{th} testing sample TeX_i and the j^{th} training sample TrX_j is calculated following Equation (1), where δ is hyperparameter of the GRNN network, which needs to be set in advance or can be obtained through an optimization process.

$$Gauss (TeX_i - TrX_j) = e^{-\frac{\|TeX_i - TrX_j\|^2}{2\delta^2}}. \quad (1)$$

Summation layer: The number of nodes is $k + 1$. The output of the summation layer includes two parts: the output of the first node is the arithmetic sum of the output of the pattern layer, and the output of the remaining k nodes is the weighted sum of the pattern layer's output of the pattern layer. Assuming that for the testing sample TeX_i , the output of the pattern layer is $\{g_1, g_2, \dots, g_m\}$, then the output of the first node and the remaining k nodes are calculated following Equation (2) and (3). Where y_{ij} is the weighting coefficient

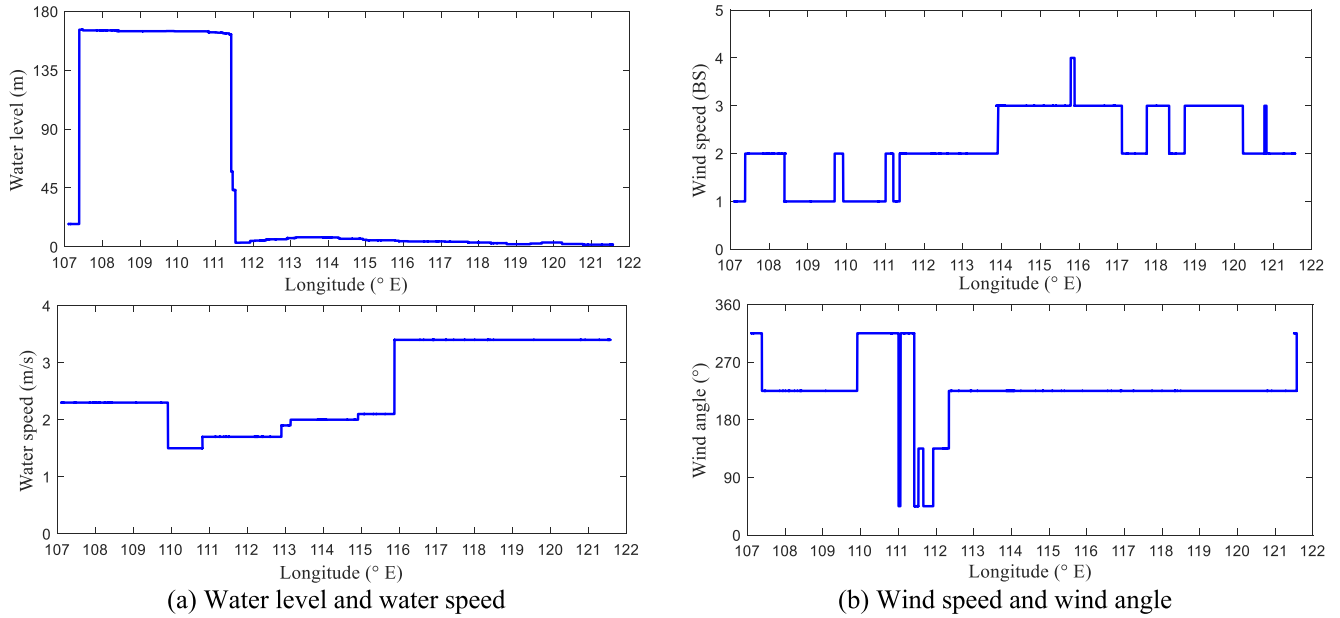


FIGURE 6. Environmental data for the entire voyage: (a) Water level and water speed, (b) Wind speed and wind angle.

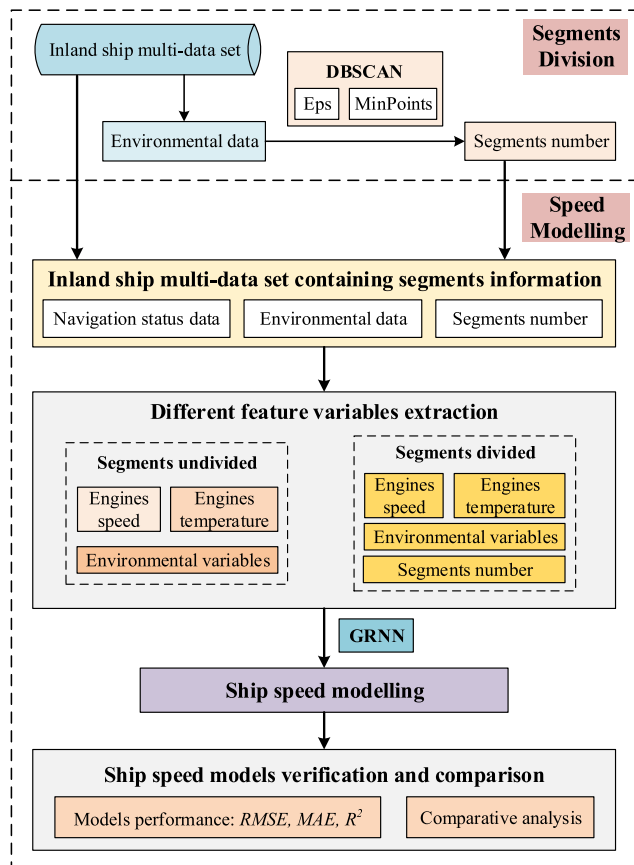


FIGURE 7. Modelling method of speed for inland ships.

of the i^{th} node of the pattern layer corresponding to the j^{th} element of the label in the training sample.

$$S_D = \sum_{i=1}^m g_i, \quad (2)$$

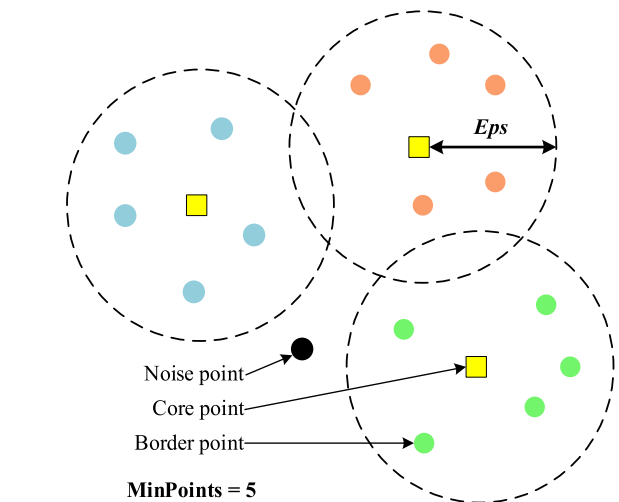


FIGURE 8. Three classes points of the DBSCAN clustering.

$$S_{Nj} = \sum_{i=1}^m y_{ij}g_i, \quad j = 1, 2, \dots, k. \quad (3)$$

Output layer: The number of nodes is k , which is the same with the dimension of the label vector. The output of each node is equal to the output of the corresponding summation layer which is divided by the output of the first node in the summation layer, as follows:

$$Y_j = \frac{S_{Nj}}{S_D}, \quad j = 1, 2, \dots, k. \quad (4)$$

The modelling process using GRNN is described as the following steps:

Step 1: Setting the feature (input) variables and the label (output) variables.

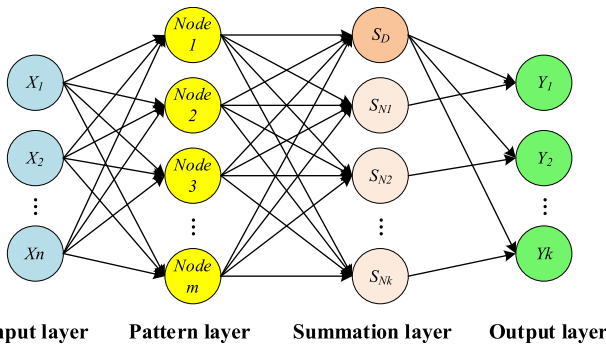


FIGURE 9. The structure of the GRNN.

Step 2: Normalizing the feature and label data to [0, 1], which can reduce the error caused by different dimensions of multi-source variables.

Step 3: Dividing the samples into a training data set and a testing data set randomly.

Step 4: Training the GRNN. Taking the cross-validation method to train the network, and finding the best “spread” through the loop training, which is the distance between the input values.

Step 5: Verifying the network with the separate testing data set, and denormalizing the results.

Step 6: Evaluating the developed model in its performance using some measures, including Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and coefficient of determination (R²).

In the above steps, the normalization, MAE, RMSE and R² can be calculated by the following.

$$x(t)^* = (x(t) - \text{Min}(x(t))) / (\text{Max}(x(t)) - \text{Min}(x(t))). \tag{5}$$

$$RMSE = \left(\frac{1}{T} \sum_{t=1}^T (y_t - \hat{y}_t)^2 \right)^{1/2}, \tag{6}$$

$$MAE = \frac{1}{T} \sum_{t=1}^T |y_t - \hat{y}_t|, \tag{7}$$

$$R^2 = 1 - \frac{\sum_{t=1}^T (y_t - \hat{y}_t)^2}{\sum_{t=1}^T (y_t - \bar{y}_t)^2}. \tag{8}$$

where, $X(t)^*$ and $X(t)$ denote the normalized data and the initial sample data, respectively. t denotes the index of a datum and T denotes the number of output data; Y_t and \hat{Y}_t represent the measured value and the forecasted value of the t^{th} datum, respectively; \bar{Y}_t denotes the average value of Y_t .

IV. CASE STUDY

The implementation platform in this work was a desktop computer, with the CPU being Inter (R) Core (TM) i5-8500, the main memory being 16.0GB RAM and the operating system being Windows 10 64-bit. Python 3.7 was the programming language, and the open-source libraries of neupy and sklearn were employed.

The original sampling data contains 32,143 records. After the proposed data pre-processing, one got a clean data set

TABLE 3. The results of environmental data clustering.

<i>Eps</i>	2.0	1.5	1.0	0.1	0.01
Clusters	2	6	10	14	16

with 15,521 records, including real-time status monitoring data and basic environmental information of the ship from the beginning to the end of the voyage. Next, we will use these measured data to conduct a detailed case analysis using the proposed approach.

To obtain the specific segments of the entire route of the sailing ship, we extracted the navigation environment data, and used Algorithm 1 to carry out cluster analysis. The input variables of Algorithm 1 are {*WaL*, *WaS*, *WiS*, *WiA*}, where *WaL* denotes water level, *WaS* denotes water speed, *WiS* denotes wind speed, and *WiA* denotes wind angle. To ensure that the segments obtained by clustering have practical significance, we believe that the data records in each segment should be more than 200. Thus the parameter *MinPoints* is set to 200. That is to say, each divided segment contains at least 200 ship trajectory points. The clustering results of different neighborhood radius *Eps* are shown in Table 3. Since the input data are normalized in Algorithm 1, the *Eps* here are relatively small.

It can be seen from Table 3 that the smaller the *Eps* is, the more clusters there are. When it is 2.0, the environment data is divided into 6 clusters. When it is 0.01, 16 clusters is obtained. Then, we correspond each cluster to a segment in the actual sailing trajectory, and the results of trajectory division of 6 segments and 16 segments are obtained, as shown in Fig. 10 and Fig. 11, respectively. In the figures, we marked the starting point of each segment with a number. From them we can conclude that: (1) The general outlines of 6 segments and 16 segments are consistent. For example, the starting point of the 2nd segment in the 6-segment case corresponds to the starting point of the 6th segment in the 16-segment case, the 3rd of the 6-segment case corresponds to the 11th of the 16-segment case, the 4th of the 6-segment case corresponds to the 12th of the 16-segment case, the 5th of the 6-segment case corresponds to the 13th of the 16-segment case, and the 6th of the 6-segment case corresponds to the 15th of the 16-segment case. It can be seen that the 16 segments division is a finer division of some segments on the basis of 6 segments. (2) The segments divided are not completely connected in sequence, such as 1-2-3-2-3-4, 11-12-13-12-13-14, and 14-15-16-15 in Fig. 11. (3) In the middle reach of the Yangtze River (as shown in Fig. 11), more segments are divided. In fact, the waterway in the middle reach is more curved, and its topography is more complicated than the lower reach and upper reach. All these proved the rationality and applicability of segments division.

In the following research, the GRNN network is tailored to build the ship speed model according to the process designed in Section 3.2, and a detailed estimation and

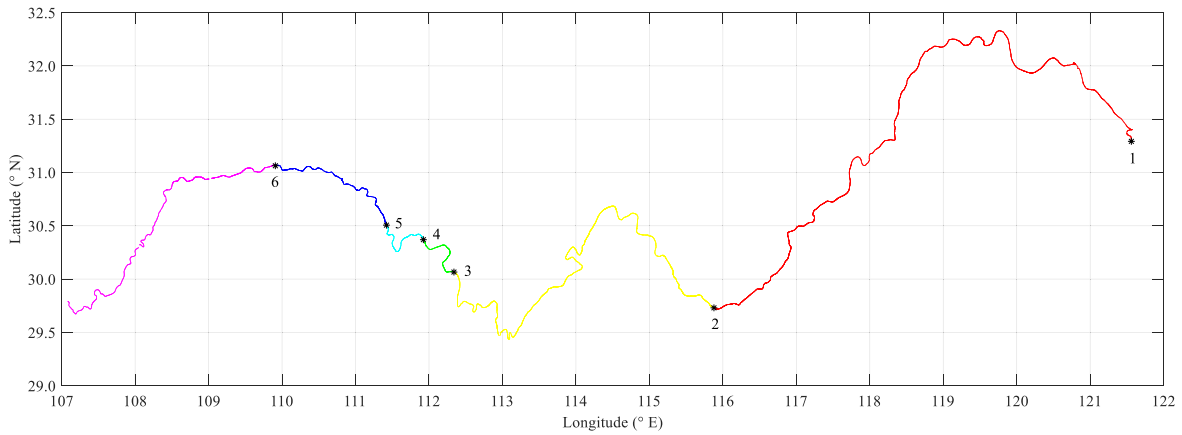


FIGURE 10. Trajectory division results of 6 segments ($Eps = 1.5$).

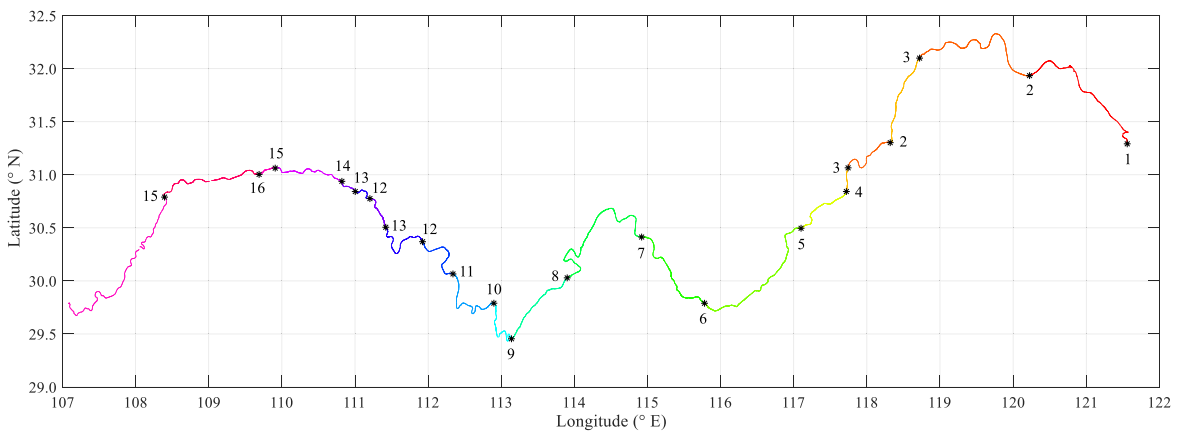


FIGURE 11. Trajectory division results of 16 segments ($Eps = 0.01$).

analysis of the ship speed are performed based on the measured data and divided segments. First of all, the SOG models of inland ship are constructed and analyzed based on the measured data without segment division. To analyze the influence of different variables on the SOG model, three groups of different feature variables were extracted as inputs, and the estimation results of the training and the testing as shown in Table 4. Among them, the output of all groups is *SOG*, the input feature variables of the first group are $\{LES, RES\}$, those of the second group are $\{LES, RES, LET, RET\}$, and those of the third group are $LES, RES, LET, RET, WaL, WaS, WiS, WiA$. *LES* denotes left engine speed, *RES* denotes right engine speed, *LET* denotes left engine temperature, *RET* denotes right engine temperature, *WaL* represents water level, *WaS* represents water speed, *WiS* represents wind speed, and *WiA* represents wind angle. 80 % (12,416 data records) of the available data were randomly selected to be the training data set and the remaining 20 % (3,105 data records) were the testing data set. To find the optimal value of the parameter “spread”, we performed a loop test with a step size of 0.001 between 0.001 and

1. After cyclic testing and cross-validation, the “spread” is determined to be 0.002. The training and testing results of third group are as shown in Fig. 12.

Secondly, we added the segment information which obtained by environmental data clustering to the data set, and carry out modelling and analysis for the SOG in the same steps as above. The SOG modelling results of different divided segments are shown in Table 5. The input feature variables of the model are *LES, RES, LET, RET, WaL, WaS, WiS, WiA, SID*, and *SID* represents the segment number. Among them, the SOG estimation results of the 6-segment division and the 16-segment division are shown in Fig. 13 and Fig. 14.

From Table 4 we can see that adding the real-time engine temperature to the input variables greatly reduces the SOG estimation error. After further increasing the navigation environmental data, the training *RMSE* and *MAE* dropped to 0.5192 and 0.3312, respectively, and the testing *RMSE* and *MAE* also dropped to 0.6844 and 0.4447, respectively. Compared with Table 5, it is found that after adding the segment information, the performance of the SOG model has been

TABLE 4. Comparison of different feature variables in SOG modelling (segment undivided).

Input variables group	Training			Testing		
	<i>RMSE</i>	<i>MAE</i>	R^2	<i>RMSE</i>	<i>MAE</i>	R^2
1	1.4088	0.9784	0.4086	1.4022	0.9835	0.3770
2	0.6993	0.4506	0.8543	0.9681	0.6064	0.7031
3	0.5192	0.3312	0.9197	0.6844	0.4447	0.8516

The input feature variables of each group are as follows:

First group: {*LES, RES*}

Second group: {*LES, RES, LET, RET*}

Third group: {*LES, RES, LET, RET, WaL, WaS, WiS, WiA*}

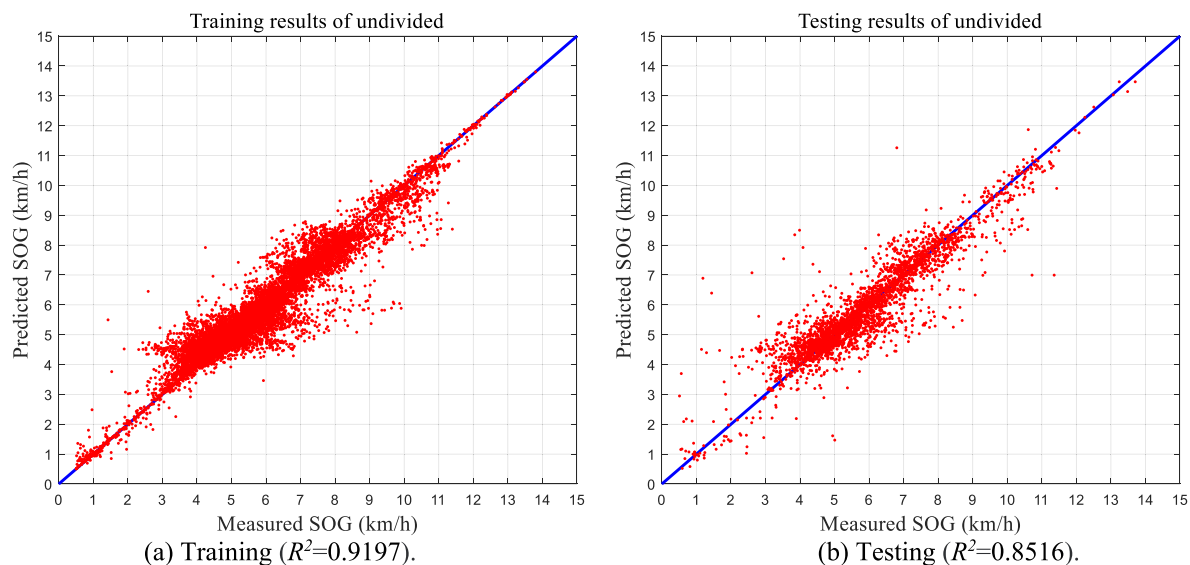


FIGURE 12. The performance of the developed SOG model (segment undivided, and the inputs feature variables being left engine speed, right engine speed, left engine temperature, right engine temperature, water level, water speed, wind speed and wind angle.): (a) measured SOG vs. predicted SOG in training and (b) measured SOG vs. predicted SOG in testing.

significantly improved, and the finer the segment division, the better the performance of the model. Especially, when the trajectory was divided into 16 segments, the *RMSE* and *MAE* of training decreased to 0.2197 and 0.1040 respectively, the *RMSE* and *MAE* of testing also decreased to 0.5264 and 0.3459 respectively, and the R^2 increased to 0.9856 and 0.9111 respectively. Compared Fig. 14 with Fig. 13 and Fig. 12, the SOG model with added segments information has better performance in both training and testing, and 16 segments are better than 6 segments.

In addition, we compared the GRNN approach with other well-known regression methods and neural networks, such as Linear Regression (LR), Interaction Linear Regression (ILR) [43], Stepwise Regression (SR) [44], Pure Quadratic Regression (PQR) [45], Coarse Tree (CT), Fine Tree (FC) [46], Elman neural network (ENN), Back Propagation neural network (BPNN), Radial Basis Function network (RBFN) [47], Recurrent Neural Network (RNN), Long-Short Term Memory (LSTM) [48] and Gate Recurrent Unit (GRU) [17]. In all comparison experiments, the inputs of the models contain 9 feature variables,

which are *LES, RES, LET, RET, WaL, WaS, WiS, WiA, SID*, and the number of training data is 12,416 (80% of the whole data) and the number of testing data is 3,105 (20% of the whole data) with 16 segments being divided. To verify the performance of the proposed method statistically, the modelling experiment was repeated 10 times. In each run, we randomly divided the data into a training data set and a testing data set and the same sets were used by all the modelling methods. In the regression methods, the time step in each input is set to 1. The neurons number of ENN and BPNN is set to 150. In RBFN, the “spread” parameter is set as 4 (the best value from many experiments). The parameter settings of RNN, GRU and LSTM are as follows: the number of neurons is 150, the time step is 1, and the batch size is 100. The epochs of all neural networks are set to 3000. The performance of different methods is shown in Table 6.

From Table 6, we can find that, all regression models and neural networks have relatively high errors in training and testing, compared with the proposed GRNN. Among all other methods, the best performance of training comes from FT, where the *RMSE* and *MAE* are 0.3936 and 0.2300

TABLE 5. The results of different segments division applied to SOG modelling.

Segments	Training			Testing		
	RMSE	MAE	R^2	RMSE	MAE	R^2
2	0.4752	0.2955	0.9327	0.6657	0.4256	0.8596
6	0.3869	0.2254	0.9554	0.6426	0.3989	0.8692
10	0.3025	0.1648	0.9662	0.6065	0.3732	0.8812
14	0.2429	0.1186	0.9773	0.5375	0.3563	0.8964
16	0.2197	0.1040	0.9856	0.5264	0.3459	0.9111

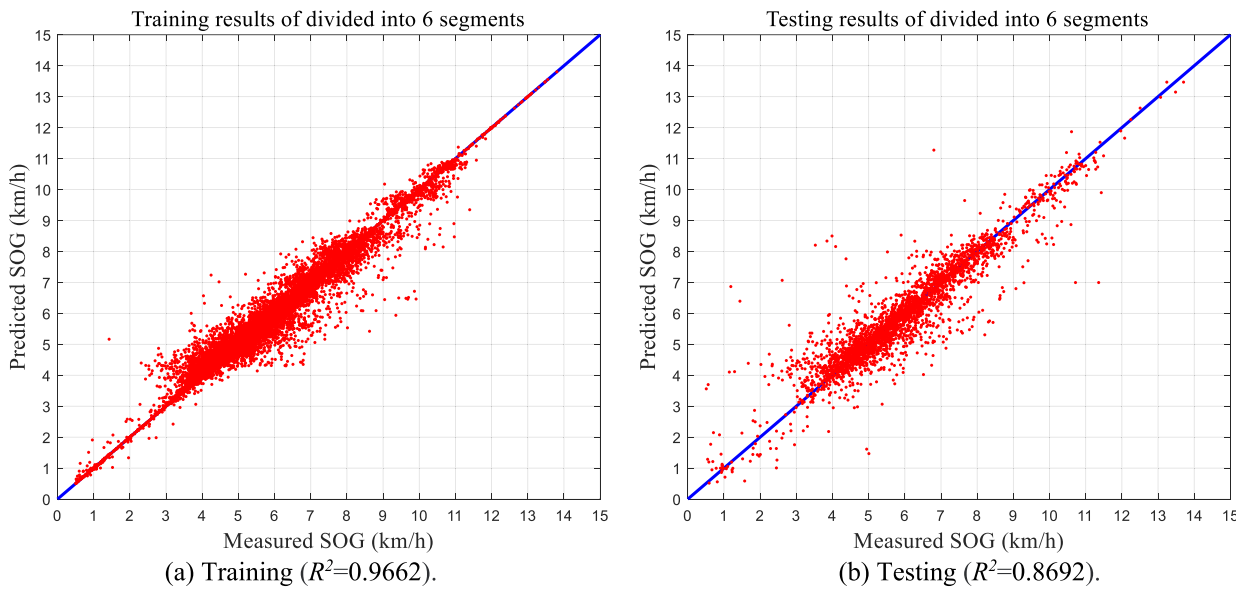


FIGURE 13. The results of SOG modelling with 6 voyage segments.

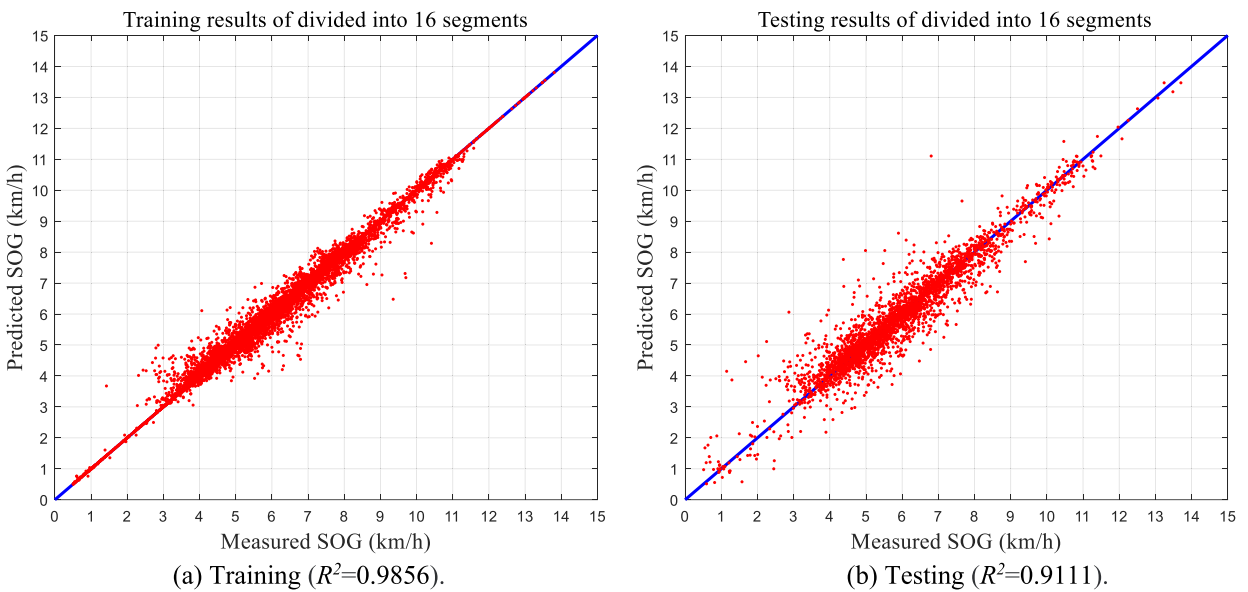


FIGURE 14. The results of SOG modelling with 16 voyage segments.

respectively, the R^2 is as high as 0.9505, but the R^2 of testing is only 0.8584. The best performance of testing comes from BPNN, where the $RMSE$ is only 0.5675 and the R^2 is 0.8981.

The training $RMSE$ of other models are between 0.7680 and 1.5136, the training R^2 are 0.3202 and 0.9055; the testing $RMSE$ of other models are between 0.6691 and 1.4583, the R^2

TABLE 6. Comparison among different methods in SOG modelling.

Method	Training			Testing		
	RMSE	MAE	R ²	RMSE	MAE	R ²
LR	1.4670±0.0061	1.0551±0.0043	0.3582±0.0024	1.4307±0.0035	1.0352±0.0101	0.3512±0.0115
ILR	1.2130±0.0034	0.8701±0.0023	0.5651±0.0016	1.2102±0.0031	0.8813±0.0124	0.5363±0.0116
RLR	1.5136±0.0052	1.0403±0.0011	0.3202±0.0012	1.4583±0.0026	1.0094±0.0131	0.3251±0.0201
SR	1.2151±0.0024	0.8704±0.0019	0.5651±0.0010	1.2103±0.0021	0.8811±0.0097	0.5368±0.0111
PQR	1.3372±0.0023	0.9620±0.0016	0.4732±0.0014	1.3144±0.0025	0.9570±0.0121	0.4528±0.0098
CT	0.7800±0.0028	0.5040±0.0021	0.8218±0.0011	0.8567±0.0025	0.5515±0.0112	0.7676±0.0126
FT	0.3936±0.0021	0.2300±0.0014	0.9505±0.0009	0.6691±0.0042	0.4445±0.0120	0.8584±0.0120
ENN	1.2541±0.0031	0.8972±0.0031	0.5307±0.0016	1.2378±0.0034	0.8998±0.0113	0.5141±0.0105
BPNN	0.5640±0.0024	0.4027±0.0018	0.9055±0.0014	0.5675±0.0030	0.4044±0.0091	0.8981±0.0092
RBFN	0.7990±0.0033	0.5862±0.0023	0.8100±0.0005	0.9530±0.0024	0.6445±0.0086	0.7120±0.0096
RNN	0.8021±0.0032	0.5975±0.0022	0.8005±0.0007	0.9580±0.0029	0.6640±0.1001	0.7050±0.1231
LSTM	0.7755±0.0026	0.5030±0.0020	0.8277±0.0012	0.8500±0.0027	0.5400±0.1011	0.7735±0.0122
GRU	0.7680±0.0025	0.4981±0.0016	0.8344±0.0006	0.8475±0.0024	0.5235±0.0096	0.7865±0.1010
Proposed GRNN	0.2196±0.0020	0.1036±0.0010	0.9857±0.0007	0.5259±0.0023	0.3459±0.0090	0.9113±0.0121

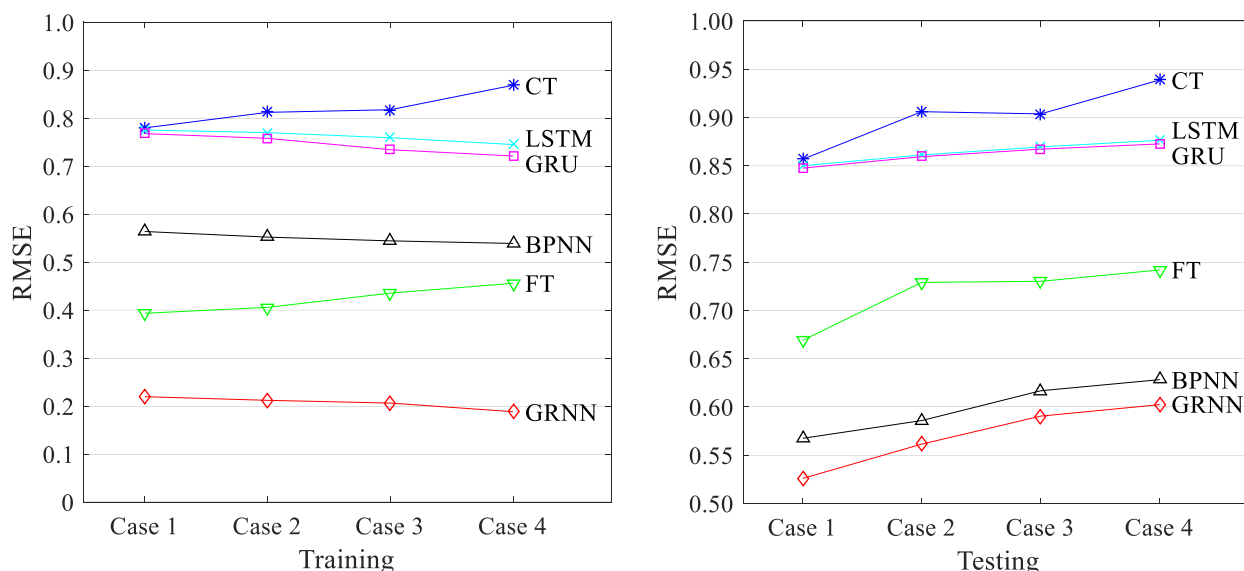


FIGURE 15. RMSE of SOG prediction for different cases (Case 1: 80% training and 20% testing; Case 2: 70% training and 30% testing; Case 3: 60% training and 40% testing; Case 4: 50% training and 50% testing).

are between 0.3251 and 0.8584. In addition, the RNN, LSTM and GRU, which are good at dealing with time series, have not exerted their advantages here, and their results are even inferior to BPNN. These results show that the proposed GRNN approach outperforms others in the ship speed modelling.

In order to further verify the applicability of the proposed method in different problem settings, we tested the modelling strategy with different proportions of training data to testing data. Five methods that perform relatively well in Table 6 were selected and used in verification and comparison. The results of four case studies are shown in Fig. 15 and Fig. 16, where the specific data division is as follows: Case 1: random 80% for training and the remaining 20% for testing; Case 2: random 70% (10,865 data records) for training and

the remaining 30% (4,656 data records) for testing; Case 3: random 60% (9,313 data records) for training and the remaining 40% (6,208 data records) for testing; Case 4: random 50% (7,761 data records) for training and the remaining 50% (7,760 data records) for testing. From Fig. 15 and Fig. 16, we can find that, with the reduction of training data, the training RMSE gradually decreases, and the training R² gradually increases. However, the testing RMSE of the corresponding case gradually increased, and the testing R² gradually decreased. Importantly, the model presented in this paper is superior to other methods in both the training data and the testing data.

In addition to the above, we used the constructed GRNN model to analyze the data of each segment of the 16 segments

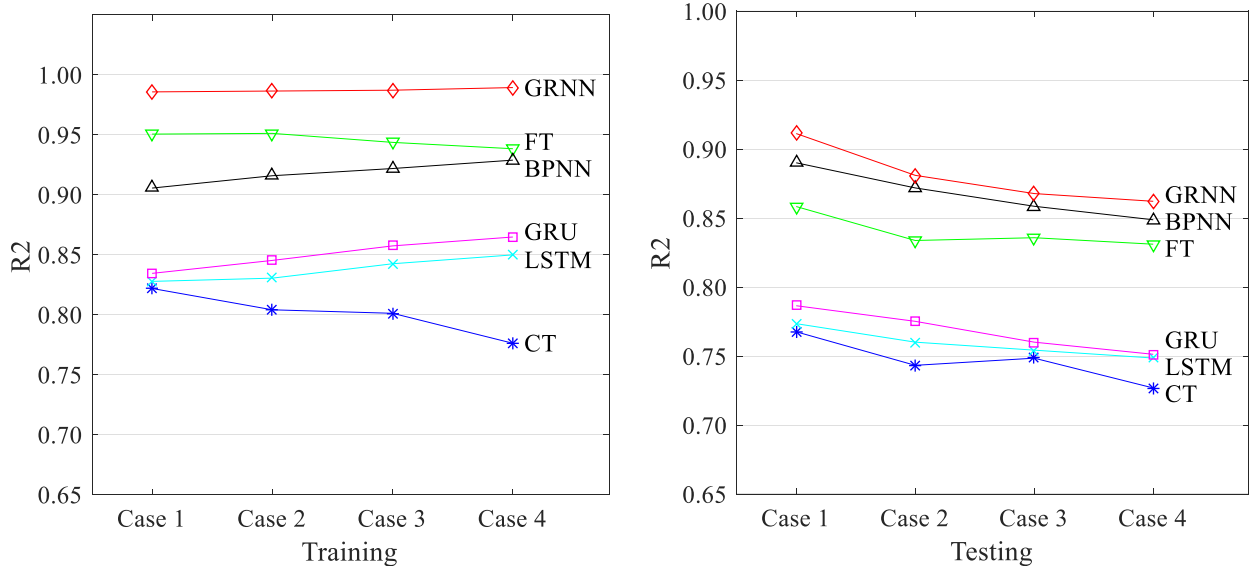


FIGURE 16. R² of SOG prediction for different cases (Case 1: 80% training and 20% testing; Case 2: 70% training and 30% testing; Case 3: 60% training and 40% testing; Case 4: 50% training and 50% testing).

TABLE 7. Ship speed modelling results of each segment (divided into 16 segments).

SID	Training		Testing		SID	Training		Testing	
	RMSE	MAE	RMSE	MAE		RMSE	MAE	RMSE	MAE
1	0.2966±	0.1887±	0.4496±	0.2665±	9	0.2081±	0.1001±	0.6220±	0.5131±
	0.0012	0.0006	0.0012	0.0010		0.0021	0.0006	0.0026	0.0013
2	0.3515±	0.2334±	0.4622±	0.3130±	10	0.2557±	0.1410±	0.6021±	0.4133±
	0.0015	0.0010	0.0016	0.0014		0.0023	0.0014	0.0022	0.0011
3	0.1810±	0.1135±	0.3701±	0.2627±	11	0.1740±	0.0951±	0.3937±	0.2714±
	0.0010	0.0008	0.0013	0.0008		0.0018	0.0011	0.0016	0.0009
4	0.1210±	0.0572±	0.4054±	0.2825±	12	0.0801±	0.0332±	0.5001±	0.3652±
	0.0007	0.0010	0.0011	0.0010		0.0012	0.0005	0.0011	0.0015
5	0.2681±	0.1661±	0.4333±	0.2554±	13	0.1927±	0.1011±	0.6108±	0.5088±
	0.0011	0.0009	0.0014	0.0011		0.0010	0.0009	0.0021	0.0012
6	0.3871±	0.2151±	0.5156±	0.3477±	14	0.1551±	0.1051±	0.2664±	0.1961±
	0.0018	0.0012	0.0021	0.0009		0.0013	0.0012	0.0016	0.0010
7	0.2889±	0.1622±	0.6021±	0.3424±	15	0.1924±	0.1132±	0.3741±	0.2582±
	0.0010	0.0009	0.0023	0.0013		0.0014	0.0009	0.0022	0.0014
8	0.2135±	0.1174±	0.6431±	0.4506±	16	0.1602±	0.1078±	0.2644±	0.1890±
	0.0009	0.0008	0.0026	0.0011		0.0017	0.0013	0.0014	0.0010

with the same steps, and all the results (mean and standard deviation of 10 experiments) are good, as shown in Table 7. 80% of the data were randomly selected and set as training data and the remaining 20% were set as testing data in each segment. These results fully demonstrate that: (1) In addition to engine speed, the engine temperature and navigation environment data also have a significant impact on ship speed model of inland ships. (2) The segment information obtained by environmental data clustering can be well applied to ship speed modelling.

V. CONCLUSION AND FUTURE WORK

In this study, an attempt has been made to build the ship speed model under the complex navigation environment using

machine learning algorithm and neural networks in the field of inland waterway transportation. First, the unsupervised learning algorithm DBSCAN was used for clustering analysis of navigation environment data to obtain the segment division information of the entire voyage. Then, GRNN was tailored and employed in ship speed modelling. Finally, these works were successfully validated using measured data. The case study results show that the accuracy of the SOG model was greatly improved by adding environment data and segment information to input feature variables. When the route was divided into 16 segments, the testing RMSE and MAE dropped by 23.07 % and 22.21 %, respectively. In addition, compared with other regression methods and neural networks, the model we built has better performance in ship speed

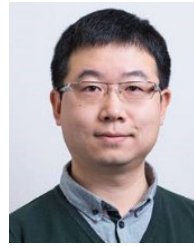
modelling. Moreover, the contribution of engine temperature to ship speed modelling has also been verified. To the best of our knowledge, for the first time, engine temperature was used for ship speed modelling.

The modelling methods proposed in this paper can also be applied to other inland rivers, as long as the relevant measured data are available. The obtained ship speed models can be exploited in navigation planning and optimization of fuel consumption, where ship speed is not only an important control variable for transportation time, but also an important decision variable for the fuel consumption optimization problem. Applying the method and results of segment division to constructing the fuel consumption model for inland ships will be the next step of this study. In addition, exploring more machine learning methods for data modelling in inland waterway transport is also important content of future work.

REFERENCES

- [1] G. Tang. (2014). *Volume of the Yangtze River's Cargo Transportation is Around 1.92 Billion Ton in 2013*. Chang Jiang River Administration of Navigation Affairs. MOT, China. Accessed: Jan. 2014. [Online]. Available: http://www.hubei.gov.cn/zwgk/rdgz/rdgzqb/201401/t20140114_486170.shtml
- [2] J. Ström-Tejsten, "Added resistance in waves," in *Naval Ship Research and Development Center, Bethesda, Maryland, USA, Research and Development Report, Ship Performance Department, Paper 3 of The Annual Meeting New York of The Society of Naval Architects and Marine Engineers, SNAME Transactions*. Delft, The Netherlands: Technische Universiteit Delft, 1973. [Online]. Available: <https://repository.tudelft.nl/islandora/object/uuid:3281ac5c-d684-46f5-bed4-7cc0449c532d>
- [3] R. L. Townsin, B. Moss, J. B. Wynne, and I. M. Whyte, "Monitoring the speed performance of ships," *North East Coast Inst Eng. Shipbuilders Trans.*, vol. 91, no. 5, pp. 159–175, 1975.
- [4] J. Holtrop and G. G. J. Mennen, "An approximate power prediction method," *Int. Shipbuilding Prog.*, vol. 29, no. 335, pp. 166–170, Jul. 1982.
- [5] X. Hu, "The influence of shallow water channel and narrow channel on ship resistance," *Water Transp. Eng.*, vol. 6, pp. 27–29, 1986.
- [6] F. Pérez Arribas, "Some methods to obtain the added resistance of a ship advancing in waves," *Ocean Eng.*, vol. 34, no. 7, pp. 946–955, May 2007.
- [7] Y. I. Kwon, "Speed loss due to added resistance in wind and waves," *Nav. Architect*, no. 3, pp. 14–16, 2008.
- [8] M.-C. Fang and Y.-H. Lin, "The optimization of ship weather-routing algorithm based on the composite influence of multi-dynamic elements (II): Optimized routings," *Appl. Ocean Res.*, vol. 50, pp. 130–140, Mar. 2015.
- [9] Q. Meng, Y. Du, and Y. Wang, "Shipping log data based container ship fuel efficiency modeling," *Transp. Res. B, Methodol.*, vol. 83, pp. 207–229, Jan. 2016.
- [10] X. Yan, K. Wang, Y. Yuan, X. Jiang, and R. R. Negenborn, "Energy-efficient shipping: An application of big data analysis for optimizing engine speed of inland ships considering multiple environmental factors," *Ocean Eng.*, vol. 169, pp. 457–468, Dec. 2018.
- [11] X. Li, B. Sun, C. Guo, W. Du, and Y. Li, "Speed optimization of a container ship on a given route considering voluntary speed loss and emissions," *Appl. Ocean Res.*, vol. 94, Jan. 2020, Art. no. 101995.
- [12] A. Filippo, A. R. Torres, B. Kjerfve, and A. Monat, "Application of artificial neural network (ANN) to improve forecasting of sea level," *Ocean Coastal Manage.*, vol. 55, pp. 101–110, Jan. 2012.
- [13] A. Montanari, A. Londei, and B. Staniscia, "Can we interpret the evolution of coastal land use conflicts? Using artificial neural networks to model the effects of alternative development policies," *Ocean Coastal Manage.*, vol. 101, pp. 114–122, Nov. 2014.
- [14] C. Stokes, G. Masselink, M. Revie, T. Scott, D. Purves, and T. Walters, "Application of multiple linear regression and Bayesian belief network approaches to model life risk to beach users in the UK," *Ocean Coastal Manage.*, vol. 139, pp. 12–23, Apr. 2017.
- [15] H. U. Dike, Y. Zhou, K. K. Deveerasetty, and Q. Wu, "Unsupervised learning based on artificial neural network: A review," in *Proc. IEEE Int. Conf. Cyborg Bionic Syst. (CBS)*, Oct. 2018, pp. 322–327.
- [16] X. Chen, D. Huang, X. Chen, W. Lian, L. Gu, Y. Zheng, and L. Xu, "Risk assessment for dangerous sections of the levees: A case study in guangdong province, China," *Ocean Coastal Manage.*, vol. 185, Mar. 2020, Art. no. 105061.
- [17] Y. Huang, Y. Li, Z. Zhang, and R. W. Liu, "GPU-accelerated compression and visualization of large-scale vessel trajectories in maritime IoT industries," *IEEE Internet Things J.*, vol. 7, no. 11, pp. 10794–10812, Nov. 2020.
- [18] Y. Du, Q. Meng, S. Wang, and H. Kuang, "Two-phase optimal solutions for ship speed and trim optimization over a voyage using voyage report data," *Transp. Res. B, Methodol.*, vol. 122, pp. 88–114, Apr. 2019.
- [19] S. Kim, S. Pan, and H. Mase, "Artificial neural network-based storm surge forecast model: Practical application to sakai minato, japan," *Appl. Ocean Res.*, vol. 91, Oct. 2019, Art. no. 101871.
- [20] F. Vieira, G. Cavalcante, E. Campos, and F. Taveira-Pinto, "A methodology for data gap filling in wave records using artificial neural networks," *Appl. Ocean Res.*, vol. 98, May 2020, Art. no. 102109.
- [21] P. Zhang, Y.-F. Jin, Z.-Y. Yin, and Y. Yang, "Random forest based artificial intelligent model for predicting failure envelopes of caisson foundations in sand," *Appl. Ocean Res.*, vol. 101, Aug. 2020, Art. no. 102223.
- [22] T. Uyanık, Ç. Karatug, and Y. Arslanoğlu, "Machine learning approach to ship fuel consumption: A case of container vessel," *Transp. Res. D, Transp. Environ.*, vol. 84, Jul. 2020, Art. no. 102389.
- [23] Z. Yuan, J. Liu, Y. Liu, Q. Zhang, and R. W. Liu, "A multi-task analysis and modelling paradigm using LSTM for multi-source monitoring data of inland vessels," *Ocean Eng.*, vol. 213, Oct. 2020, Art. no. 107604.
- [24] K. P. Sinaga and M.-S. Yang, "Unsupervised K-means clustering algorithm," *IEEE Access*, vol. 8, pp. 80716–80727, 2020.
- [25] A. Gunawan, "A faster algorithm for DBSCAN," M.S. thesis, Dept. Math. Comput. Sci., Eindhoven Univ. Technol., Eindhoven, The Netherlands, 2013. [Online]. Available: <http://alexandria.tue.nl/extra1/afstversl/wsk-i/gunawan2013.pdf>
- [26] J. Shen, X. Hao, Z. Liang, Y. Liu, W. Wang, and L. Shao, "Real-time superpixel segmentation by DBSCAN clustering algorithm," *IEEE Trans. Image Process.*, vol. 25, no. 12, pp. 5933–5942, Dec. 2016.
- [27] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu, "DBSCAN revisited: Why and how you should (Still) use DBSCAN," *ACM Trans. Database Syst.*, vol. 42, no. 3, pp. 1–21, Aug. 2017.
- [28] Y. Chen, L. Zhou, S. Pei, Z. Yu, Y. Chen, X. Liu, J. Du, and N. Xiong, "KNN-BLOCK DBSCAN: Fast clustering for large-scale data," *IEEE Trans. Syst., Man, Cybern. Syst.*, early access, Dec. 18, 2020, doi: 10.1109/TSMC.2019.2956527.
- [29] Y. Wang, Y. Gu, and J. Shun, "Theoretically-efficient and practical parallel DBSCAN," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2020, pp. 2555–2571.
- [30] D. Luchi, A. L. Rodrigues, and F. M. Varejão, "Sampling approaches for applying DBSCAN to large datasets," *Pattern Recognit. Lett.*, vol. 117, pp. 90–96, Jan. 2019.
- [31] Z. Liu, Z. Wu, and Z. Zheng, "A novel framework for regional collision risk identification based on AIS data," *Appl. Ocean Res.*, vol. 89, pp. 261–272, Aug. 2019.
- [32] K. Sheridan, T. G. Puranik, E. Mangorrey, O. J. Pinon-Fischer, M. Kirby, and D. N. Mavris, "An application of DBSCAN clustering for flight anomaly detection during the approach phase," in *Proc. AIAA Scitech Forum*, Jan. 2020, p. 1851.
- [33] Y. Wen, Z. Sui, C. Zhou, C. Xiao, Q. Chen, D. Han, and Y. Zhang, "Automatic ship route design between two ports: A data-driven method," *Appl. Ocean Res.*, vol. 96, Mar. 2020, Art. no. 102049.
- [34] R. W. Liu, J. Nie, S. Garg, Z. Xiong, Y. Zhang, and M. S. Hossain, "Data-driven trajectory quality improvement for promoting intelligent vessel traffic services in 6G-enabled maritime IoT systems," *IEEE Internet Things J.*, early access, Oct. 5, 2020, doi: 10.1109/JIOT.2020.3028743.
- [35] A. Li, H. Jiang, Z. Li, J. Zhou, and X. Zhou, "Human-like trajectory planning on curved road: Learning from human drivers," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 8, pp. 3388–3397, Aug. 2020.
- [36] D. F. Specht, "A general regression neural network," *IEEE Trans. Neural Netw.*, vol. 2, no. 6, pp. 568–576, Nov. 1991.
- [37] M. Valčić and J. Prpić-Oršić, "Hybrid method for estimating wind loads on ships based on elliptic Fourier analysis and radial basis neural networks," *Ocean Eng.*, vol. 122, pp. 227–240, Aug. 2016.
- [38] P. Borkowski, "The ship movement trajectory prediction algorithm using navigational data fusion," *Sensors*, vol. 17, no. 6, p. 1432, Jun. 2017.

- [39] Y. Liang, D. Niu, and W.-C. Hong, "Short term load forecasting based on feature extraction and improved general regression neural network model," *Energy*, vol. 166, pp. 653–663, Jan. 2019.
- [40] N. Parveen, S. Zaidi, and M. Danish, "Development and analyses of data-driven models for predicting the bed depth profile of solids flowing in a rotary kiln," *Adv. Powder Technol.*, vol. 31, no. 2, pp. 678–694, Feb. 2020.
- [41] T. Cepowski, "The prediction of ship added resistance at the preliminary design stage by the use of an artificial neural network," *Ocean Eng.*, vol. 195, Jan. 2020, Art. no. 106657.
- [42] S.-S. Li, "An improved DBSCAN algorithm based on the neighbor similarity and fast nearest neighbor query," *IEEE Access*, vol. 8, pp. 47468–47476, 2020.
- [43] A. F. Schmidt and C. Finan, "Linear regression and the normality assumption," *J. Clin. Epidemiology*, vol. 98, pp. 146–151, Jun. 2018.
- [44] S. Abraham, M. Raisee, G. Ghorbaniasl, F. Contino, and C. Lacor, "A robust and efficient stepwise regression method for building sparse polynomial chaos expansions," *J. Comput. Phys.*, vol. 332, pp. 461–474, Mar. 2017.
- [45] H. Jiang and Y. Dong, "Global horizontal radiation forecast using forward regression on a quadratic kernel support vector machine: Case study of the tibet autonomous region in China," *Energy*, vol. 133, pp. 270–283, Aug. 2017.
- [46] F. Duan, Y. Wan, and L. Deng, "A novel approach for coarse-to-fine windthrown tree extraction based on unmanned aerial vehicle images," *Remote Sens.*, vol. 9, no. 4, p. 306, Mar. 2017.
- [47] A. Asgharnia, A. Jamali, R. Shahnazi, and A. Maheri, "Load mitigation of a class of 5-MW wind turbine with RBF neural network based fractional-order PID controller," *ISA Trans.*, vol. 96, pp. 272–286, Jan. 2020.
- [48] Z. Yuan, J. Liu, Y. Liu, Y. Yuan, Q. Zhang, and Z. Li, "Fitting analysis of inland ship fuel consumption considering navigation status and environmental factors," *IEEE Access*, vol. 8, pp. 187441–187454, 2020.



QIAN ZHANG (Member, IEEE) received the B.Eng. degree in automatic control from Zhejiang University, Zhejiang, China, in 2003, and the Ph.D. degree from the University of Sheffield, U.K., in 2008. He is currently a Senior Lecturer of electronics and electrical engineering with Liverpool John Moores University (LJMU). His research interests include data-driven modeling, multi-objective optimal design, robotics and control, and their applications in engineering and transport systems.



YI LIU (Member, IEEE) received the Ph.D. degree in transportation engineering from the Department of Civil, Architectural and Environmental Engineering, Illinois Institute of Technology, in 2015. He is currently an Associate Professor with the School of Navigation, Wuhan University of Technology (WUT). His current research interests include vessel traffic flow theory, area wide traffic dynamics, transportation system optimization, and intelligent traffic organization.



ing, maritime transport, computational transportation science, and artificial intelligence.

ZHI YUAN received the M.Sc. degree from the Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, China, in 2016, and the Joint Ph.D. degree from the Department of Electronics and Electrical Engineering, Liverpool John Moores University, U.K., in 2019. He is currently pursuing the Ph.D. degree with the School of Navigation, Wuhan University of Technology (WUT). His research interests include data-driven modeling,



YUAN YUAN received the B.Eng. degree in energy, power system and automation and the M.Sc. degree in marine engineering from the Wuhan University of Technology (WUT), in 2013 and 2017, respectively. She is currently an Engineer with ChangJiang Shipping Science Research Institute Company Ltd. Her research interest is marine technology.



JINGXIAN LIU received the M.Sc. degree in traffic information engineering and control from the Wuhan University of Technology (WUT), Wuhan, China, in 2004, and the Ph.D. degree from the School of Energy and Power Engineering, WUT, Wuhan, China, in 2009. He is currently a Full Professor with the School of Navigation (WUT). His current research interests include intelligent transportation systems, vessel traffic flow, vessel navigation safety, risk assessment, and intelligent traffic organization.



ZONGZHI LI received the B.Eng. degree from Chang'an University, in 1992, and the Ph.D. degree in transportation and infrastructure systems engineering from Purdue University, in 2003. He is currently a Professor with the Department of Civil, Architectural and Environmental Engineering, Illinois Institute of Technology. His research interests include multimodal transportation system performance modeling, performance-based evaluation and investment decision-making, and transportation network economics.

...