https://dx.doi.org/10.1016/j.oceaneng.2024.117087

Multi-state ship traffic flow analysis using data-driven method and visibility graph

Zhongyi Sui[a], Shuaian Wang[a], Yuanqiao Wen[b, c], Xiaodong Cheng[b, d,*]
Gerasimos Theotokatos[d]

[a] *Department of Logistics and Maritime Studies, The Hong Kong Polytechnic University, Hong Kong, China*
[b] *State Key Laboratory of Maritime Technology and Safety, Wuhan, China*
[c] *Sanya Science and Education Innovation Park, Wuhan University of Technology, Sanya, China*
[d] *Department of Naval Architecture, Ocean & Marine Engineering, University of Strathclyde, Glasgow, United Kingdom*

Abstract: Ship traffic flow characteristics play a crucial role in enhancing the effectiveness and efficiency of intelligent maritime traffic management systems. The primary objective of this study is to establish a comprehensive framework for analyzing multi-state traffic flow based on the automatic identification system (AIS). The collected AIS data undergoes preprocessing to calculate traffic flow density, velocity, and intensity. Subsequently, clustering techniques, specifically the *K*-medoids algorithm and silhouette coefficient analysis, are applied to classify traffic states ranging from least congested to highly congested. The datasets corresponding to each cluster are then utilized to construct visibility graphs, which enable a graphical representation of the traffic flow dynamics. Statistical analysis is conducted to examine the topological characteristics of the network. To illustrate the applicability of the proposed framework, a case study of the Meishan island water areas is conducted, allowing for an in-depth analysis of ship traffic flow characteristics and the identification of distinct traffic flow states. The findings of this study demonstrate the effectiveness of the visibility graph method in analyzing multi-state ship traffic flow. Additionally, the statistical characteristics derived from the developed complex

networks adeptly capture the inherent maritime traffic flow characteristics. The insights gained from this study contribute to the advancement of maritime traffic management by providing a deeper understanding of complex traffic flow patterns and delineation.

Keywords: Maritime traffic management, AIS data, Ship traffic flow, Visibility graph, Complex network, Clustering

1. Introduction

The traffic state is influenced by multiple factors, including the presence of ships, environmental conditions, navigation rules, and other variables. The mandatory installation of Automatic Identification System (AIS) equipment has facilitated the collection of a vast volume of near real-time traffic data, enabling a comprehensive depiction of the dynamic behavior of ships.

While numerous traffic flow models have been developed and utilized in the maritime domain, limited attention has been given to the evolutionary features of traffic flow (Wen et al., 2015; Huang et al., 2019; Wang et al., 2019; Zhang et al., 2022b). Modeling and predicting the real-world characteristics of maritime traffic flow remain challenging. As waterway traffic continues to grow, a substantial amount of traffic data is generated. The analysis of ship traffic flow data is deemed crucial for waterway traffic analysis and decision-making. Traditional traffic management systems rely on rudimentary information release and basic traffic flow prediction. However, with the advent of maritime traffic management, situational awareness has gained increasing importance. Vessel traffic services operators must swiftly identify traffic states and analyze traffic flow characteristics based on traffic flow data to provide intelligent management support. Unfortunately, the current AIS data quality fail to meet the requirements of modern intelligent maritime traffic management systems (Zhu et al., 2018). The analysis capabilities of ship traffic flow data directly influence the service and quality of intelligent transportation management systems.

Stochastic process theory, statistical analysis, and machine learning techniques have been extensively employed to analyze traffic flow datasets. These methods, in conjunction with traffic theory, enable the derivation of dynamic characteristics of traffic flow (Meena et al., 2020; Li et al., 2021). However, these approaches primarily focus on the mathematical analysis of time series data and often overlook the

consideration of real-world traffic situations. In a pioneering study, Zhang et al. (2022b) proposed a microscopic perspective for predictive analytics to assess the complexity of maritime traffic flow. This method also evaluates the impact of traffic flow complexity on the likelihood of accidents, providing valuable insights into the relationship between the two. Furthermore, Zhang et al. (2021 and 2022a) introduced big data analytics techniques to evaluate traffic risks influenced by hydrometeorological conditions. These techniques have proven instrumental in identifying potential undesired scenarios across different voyages and establishing universally accepted risk criteria. Consequently, these criteria can trigger alerts for managing maritime traffic flow effectively. Collectively, these approaches offer innovative insights into the analysis of maritime traffic under real operational conditions. More recently, the application of complex network theory has gained significant attention for analyzing dynamic characteristics of datasets and has been successfully implemented in various fields (Sui et al., 2020 and 2022). Tang et al. (2014), Yan et al. (2017), and Sui et al. (2023) have reported the application of complex networks in studying the characteristics of traffic flow datasets, highlighting the advantages of this method over alternative approaches.

Despite the existence of various traffic flow models used in maritime research, only a limited number of them have focused on the fundamental characteristics of traffic flow states or the dynamic nature of ship traffic flow time series. To address this research gap, the objective of this study is to employ the visibility graph (VG) technique to gain insights into the dynamic properties of ship traffic flow time series. The visibility graph approach, as defined by Lascasa et al. (2008), represents the data points within a time series as nodes in a complex network. The statistical features of this complex network derived from the time series effectively capture the characteristics of the underlying data (Barabási, 2009).

The subsequent sections of this study are organized as follows: Section 2 provides a comprehensive review of relevant research on ship traffic flow. Section 3 outlines the methodology employed in this study, including the adopted approaches and techniques. Section 4 presents a detailed case study, along with the results obtained and their subsequent discussion. Finally, Section 5 summarizes the main findings and conclusions drawn from this research.

2 Literature review

Traffic flow models play a crucial role in comprehensively analyzing the characteristics of traffic flow and accurately describing its dynamic evolution. These models often assume that traffic flow exhibits similarities with the flow of liquids or gas dynamics (Jia et al., 2014). One of the earliest studies on traffic flow models was conducted by Greenshields, who analyzed the relationship between speed and density (Greenshields, 1934). In general, traffic flow simulation models can be categorized into two main types: microscopic and macroscopic models. These models establish relationships between fundamental traffic flow characteristics, such as speed, density, and flow, thereby representing the prevailing traffic conditions.

Microscopic models are used to establish the connections between aggregated data, such as average density and velocity, in traffic flow analysis. These models primarily focus on analyzing the irregularity and unpredictability of traffic time series. Various methods are employed for this purpose, including the approximate entropy method (Menelaou et al., 2017), the power spectrum method (Frazier et al., 2004), the phase space reconstruction method (Zhang et al., 2009), and the Lyapunov exponent method (Kwasniok et al., 2019). Yip (2013) proposed a macroscopic model based on classical traffic flow theory, which provides insights into traffic characteristics. By analyzing the average vessel speed and flow distribution within a designated area, it becomes possible to predict the probability of collision accidents (Pedersen, 2010). However, macroscopic models do not consider individual ship behavior, which is a limitation of these models.

Traditional ship traffic evaluations have often relied on visual and radar observations, which are considered inefficient (Yamaguchi et al., 1971). In recent times, the utilization of the Vessel Traffic Services System (VTS) and AIS has become prevalent, enabling the acquisition of extensive ship traffic data. Consequently, a greater volume of datasets is now available for analyzing ship traffic flow. Ship traffic analysis using VTS and AIS data has demonstrated its potential in accurately capturing ship navigation activities and revealing detailed ship traffic conditions (Tsou, 2010; Shelmerdine, 2015). Silveira et al. (2013) utilized AIS data to create a visual representation of ship traffic patterns off the coast of Portugal. Other researchers have employed observation lines to collect information on the number of ships crossing different locations, capturing ship positions and speeds (Xiao et al., 2015; Hou et al.,

2014; Suo et al., 2022). Additionally, AIS data has been directly utilized by researchers to analyze ship trajectories in specific ocean regions and identify trajectory trends, analogous to the observation of automobile traffic on road networks (Zhou et al., 2019b; Rong et al., 2022). Furthermore, to enhance ship navigation safety and enable real-time traffic management, Zhang et al. (2019) conducted a comprehensive examination of ship traffic demand and the spatial-temporal dynamics of ship traffic in the waters of Singapore port.

Traffic flow analysis methodologies can generally be classified into three main types: data-driven, statistical, and traffic model-based approaches (Fangce et al., 2013; Meng et al., 2015; Zhao et al., 2022a). When it comes to analyzing ship traffic flow characteristics, traffic model-based approaches (Chen et al., 2020; Zhao et al., 2022b) are commonly employed, utilizing complex mathematical descriptions of the fundamental rules governing traffic phenomena. For instance, Kang et al. (2018) utilized traditional fundamental traffic flow diagrams and least-squares methods to estimate the speed-density relationship in the Singapore Strait, ultimately providing an estimate of the canal system's capacity. However, accurately representing real-world traffic situations in a model poses challenges due to the complexity and uncertainty inherent in ship traffic flow (Yu et al., 2016). The dynamic nature of ship traffic and the multitude of factors influencing it make it difficult to develop a model that precisely captures the intricacies of real-world scenarios.

As a result, statistical methods have gained more prominence in traffic flow analysis (Sun et al., 2003; Williams and Hoel, 2003; Guo et al., 2014; Vaishnav et al., 2018). These techniques are particularly suitable for analyzing stationary or linear time series data (Meng et al., 2015). However, to capture the non-linear characteristics of traffic flow data, data-driven methods are widely employed. Examples of such methods include Support Vector Regression (SVR) models (Asif et al., 2014), K-Nearest Neighbors models (Zhou et al., 2019a), and neural network-based models (Li, 2016). Several studies have compared data-driven methods with Kalman filtering and Autoregressive Integrated Moving Average model (ARIMA), consistently demonstrating their superiority in terms of prediction accuracy and overall model performance improvement (Lam et al., 2010; Yoon and Chang, 2014).

Based on the literature review conducted earlier, previous studies on ship traffic flow have primarily concentrated on traffic flow models, traffic flow characterization, and traffic flow prediction. In the traditional traffic flow theory, the interrelationships

among various traffic flow parameters are analyzed through the fundamental diagram, enabling the determination of the traffic state. In this study, the visibility graph method is employed to analyze the actual traffic flow data. A detailed comparison between the traffic flow methods and the visibility graph method is presented in Table 1.

Table 1. Comparison between traffic flow methods

Method		Research topic				
		Statistical characteristics	Spatial characteristics	Temporal characteristics	Dynamic characteristics	Prediction
Traffic flow simulation model	Microscopic models		X	X		
	Macroscopic models		X	X		
Traffic flow analysis method	Traffic model-based methods		X	X		X
	Statistical methods	X				X
	Data-driven methods	X	X	X		X
Visibility Graph		X	X	X	X	

The existing literature primarily focuses on traffic flow models, traffic flow characterization, and traffic flow prediction in the context of ship traffic. However, several research gaps have been identified based on the literature review:

(i) Both microscopic and macroscopic models are capable of representing key features of real traffic, such as the relationship between density, velocity, and flow rate. However, these models fail to adequately capture the current state of traffic flow.

(ii) Statistical analysis methods provide insights into traffic flow characteristics, including ship velocity, traffic density, and traffic flow distribution. While these methods offer valuable statistics based on historical data analysis, they are unable to predict the temporal dependence between traffic flow characteristics or identify the relationship between traffic flow state and influencing factors.

(iii) Data-driven methods can predict traffic flow volume, but their output parameters heavily rely on available historical data. These models, however, do not provide a comprehensive understanding of traffic flow dynamics.

To address these gaps, this paper aims to investigate the dynamic characteristics of ship traffic flow time series using complex network theory and VG algorithms. The

visibility graph method directly represents time series data as nodes in a complex network, with connections between nodes determined by the linear visual relationships between the data (Lacasa et al., 2008). Consequently, the visibility graph method can transform any time series into a network, and the statistical properties of the resulting complex network can effectively reflect the characteristics of the time series. For instance, a periodic time series can be transformed into a regular graph, a random time series into a random graph, and a fractal time series into a scale-free network (Barabási, 2009).

3. Methodology

3.1 Ship traffic flow parameters and data standardization

The AIS data, which is received in the form of a sequence of messages with irregular time intervals and a nonstandard format, requires preprocessing due to its inherent noise. To address this, a trajectory restoration method (Sang et al., 2015) is employed to handle abnormalities in ship location, speed, and course. Subsequently, traffic flow parameters can be derived from the AIS data using a time-space diagram (Huang et al., 2019). The key parameters used to characterize ship traffic flow are density, velocity, and intensity. Density represents the average number of vessels per unit length of the channel, velocity denotes the average vessel speed per unit length of the channel, and intensity refers to the average number of vessels per unit time.

Before data analysis, it is necessary to collect a large number of indicators. The characteristics of each indicator, such as its nature, dimension, and order of magnitude, may vary, making it impossible to directly analyze the characteristics and laws of the research object. In this research, to homogenize the available datasets, data standardization should be performed before clustering, and the absolute values of variables should be turned into relative values. The object of data standardization are three parameters of ship traffic flow, density (k), velocity (u) and intensity (q). The max-min standardization method is adopted herein, which employees the following equation:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{1}$$

where x' is the normalized value, $x \in \{k, u, q\}$, $\max(x)$ is the maximum value of

the parameter x, and $\min(x)$ is the minimum value of the parameter x.

3.2 Clustering of multi-state ship traffic flow

To classify the traffic congestion level, the *K*-medoids clustering algorithm has been used to cluster the multi-state traffic flow (Park et al., 2009). The traffic flow parameters are employed as input, $\left(k_t, u_t, q_t\right), t = 1, 2, \cdots r$, r represents the amount of sample data. The dataset is split into K classes, and the centres of the K clusters with the least square difference are chosen from it. The *K*-medoids method's primary steps for classifying traffic flow states are as follows:

(1) Choose *K* samples at random from the dataset to serve as the starting centre point, $o_i (i = 1, 2, \cdots, K)$

(2) Based on Manhattan distance, allocate the remaining samples to the cluster symbolised by the closest centre point.

$$d_{R,o_i} = \left|k_R - k_{o_i}\right| + \left|u_R - u_{o_i}\right| + \left|q_R - q_{o_i}\right| \tag{2}$$

where R is one of the remaining samples.

(3) Determine the square difference function w:

$$w = \sum_{i=1}^{K} \left(\left|k_p - k_{o_i}\right| + \left|u_p - u_{o_i}\right| + \left|q_p - q_{o_i}\right| \right)^2 \tag{3}$$

where p represents the sample in the cluster L_i and 0_i represents the center point of the cluster L_i.

(4) Update each cluster's centre based on the direction in which the square difference function decreases, then repeat steps (1) through (3) until the cluster stops changing.

K-medoids algorithm cannot identify the number of clusters. To determine the number of clusters, the silhouette coefficient (Rousseeuw, 1987) is utilised. The silhouette coefficient produces a succinct graphical representation for each item grouping and is a technique for analyzing and verifying consistency among data clusters (Rousseeuw, 1987). For the *i*th data point in a cluster, the silhouette coefficient is calculated according to the following:

$$S(i) = \frac{b(i) - a(i)}{\max\left\{a(i), b(i)\right\}} \tag{4}$$

where $a(i)$ represents the average distance between i and all other data points in the same cluster, $b(i)$ is the shortest mean distance between i and any other cluster of

which i is not a part. From the equation (4), $S(i) \in [-1,1]$. Silhouette coefficient of clustering results can be calculated by following equation:

$$s = \frac{1}{r}\sum_{i=1}^{r} s(i) \tag{5}$$

where r represents the amount of sample data.

The amount of clusters corresponding to maximum silhouette coefficient is optimal.

3.3 Visibility graph of ship traffic flow

This section describes the development of the VG for the ship traffic flow. If any other data $(t_i, y(t_i))$ is positioned between two arbitrary data $(t_a, y(t_a))$ and $(t_b, y(t_b))$, they will become visible and, as a result, two linked nodes of the related graph if and only if they satisfy the following conditions:

$$y(t_i) < y(t_a) + \frac{t_i - t_a}{t_b - t_a}[y(t_b) - y(t_a)] \tag{6}$$

where t represents the time, y represents the value of traffic parameter at time t.

Several parameters are required to characterise the ship traffic flow state. As VG can only deal with single variable time series, they cannot represent the overall features of ship traffic flow. To address this challenge, the network matrix of the ship traffic flow multi-parameter time series is developed through matrix superposition. The three adjacency matrices corresponding to the three traffic parameters (density, velocity, intensity) time series for each cluster are superimposed, so that the result matrix can reflect the ship traffic flow characteristics to the greatest extent. Equation (6) is employed to obtain the network's adjacency matrix for each parameter. Since each matrix is a Boolean matrix, the adjacency matrix of the corresponding network for each type of traffic state is $M = M_k \wedge M_u \wedge M_q$.

The process of developing the VG can be seen in Fig. 1. The multi-state ship traffic flow cluster is performed and the time series of the three parameters of interest for each cluster are identified. Subsequently, the adjacency matrices are calculated and used to determine the ship traffic flow adjacency matrix. The corresponding time series are transformed into a VG by employing the Gephi software.

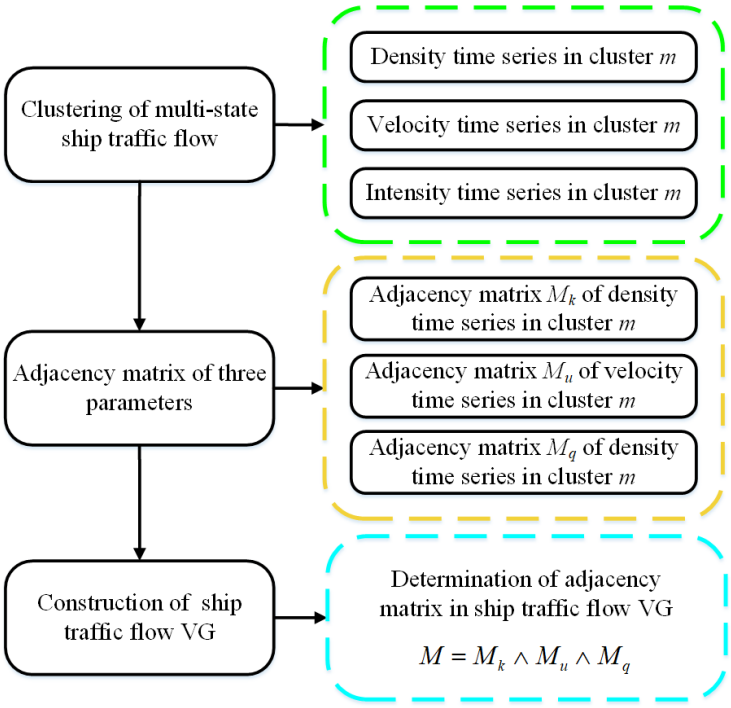

Fig. 1. Process to develop the ship traffic flow Visibility Graph

3.4 Ship traffic flow network characteristics

The degree and degree distribution are the key indicators of the network. A high node degree reflects a sudden increase in ship traffic volume. The average degree is defined according to the following equation:

$$\bar{k} = \frac{1}{N}\sum_{i=1}^{N} k_i \tag{7}$$

where k_i represents the node i's degree, $\bar{k}$ represents the network's average degree, and N represents the amount of nodes in a complex network.

Nodes in real-world networks typically form close-knit communities with a comparatively high tie density. This metric comes in two flavours: global and local. While the local reveals the embeddedness of individual nodes, the global clustering coefficient is intended to provide an overall indicator of the clustering in the network (Watts and Strogatz, 1998). The definitions of clustering coefficients are provided by the following equations:

$$C_i = \frac{E_i}{C_{k_i}^2} \tag{8}$$

$$\bar{C} = \frac{1}{N}\sum_{i=1}^{N} C_i \tag{9}$$

where C_i represents the local clustering coefficient of node i, E_i represents the number of connection among node *i*'s neighbours which are realised, $C_{k_i}^2$ represents the number of all possible connections among node *i*'s neighbours, and $\overline{C}$ represents the global clustering coefficient.

The longest of all the shortest paths in a network is known as the network diameter, and it is another way to measure network graphs (Pandurangan et al., 2003). A network with a larger clustering coefficient and smaller network diameter will be called a small-world network. The network diameter is calculated by:

$$D = \max_{1 \le i < j \le N} d_{ij} \tag{10}$$

where d_{ij} is the shortest path between node i and node j.

3.5 Evaluation criterion

The accuracy (ACC), mean absolute error (MAE), maximum absolute error (MAXAE), and root mean squared error (RMSE) were chosen as the evaluation indices to assess the effectiveness of the method proposed in this study.

The accuracy of the ship traffic flow state identification is calculated by:

$$\mathrm{ACC} = \frac{n_a}{r} \times 100\% \tag{11}$$

where r represents the total number of samples, n_a represents the number of samples that are accurately recognised.

The MAE, MAXAE and RMSE of ship traffic flow state identification are calculated according to the following equations, respectively:

$$\mathrm{MAE} = \frac{1}{r}\sum_{t=1}^{r}\left|p_t - p_t\right| \tag{12}$$

$$\mathrm{MAXAE} = \max_{t \in \{1,2,\cdots,r\}}\left\{\left|p_t - p_t\right|\right\} \tag{13}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{r}\sum_{t=1}^{r}\left(p_t - p_t\right)^2} \tag{14}$$

where p_t is the actual ship traffic flow state, p_t is the predicted ship traffic flow state.

4. Case study

4.1 Case environment

The case study focuses on the ship traffic flow characteristics in the Ningbo Meishan port area, which is renowned as one of the busiest ports globally. The observation areas within the port are depicted in Figure 2. AIS data for this specific area were collected over a ten-day period in April 2021. The observed channel has a length of 2 nautical miles, and each observation interval spans 5 minutes. Utilizing the AIS data, calculations were performed to determine the density, velocity, and intensity of ship traffic flow within the research area. A partial representation of the preprocessed AIS data is presented in Table 2.

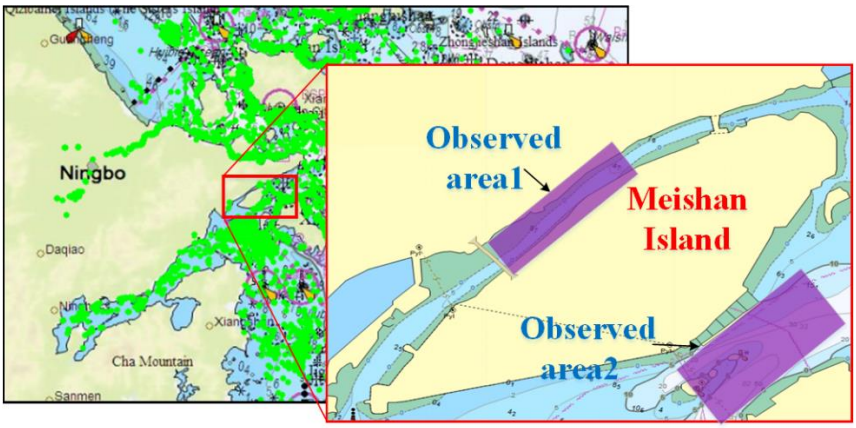

Fig. 2. Research area in Meishan island water area

Table 2. Portion of preprocessed AIS data

MMSI	Time	Latitude	Longitude	Speed	Course
41337****	2021-04-21 00:00:00	29°48.325N	121°58.368E	10.55kn	219.0°
41343****	2021-04-21 00:05:00	29°48.127N	121°58.116E	10.02kn	218.6°
63601****	2021-04-21 00:10:00	29°47.79N	121°57.812E	9.65kn	218.6°
41344****	2021-04-21 00:15:00	29°47.57N	121°57.529E	9.74kn	207.8°
…	…	…	…	…	……
41276****	2021-04-21 23:55:00	29°47.804N	121°57.828E	10.26kn	262.4°
41327****	2021-04-21 24:00:00	29°47.651N	121°57.637E	11.2kn	254.5°

4.2 Multi-state traffic flow time series analysis based on VG

Figure 3 illustrates the silhouette coefficient corresponding to the number of traffic flow states in the two observation areas. Based on the highest value of the silhouette coefficient, it is determined that the traffic flow states can be effectively separated into three distinct clusters.

The clustering results and the relationships between the traffic variables in the two observation areas are presented in Figure 4 and Figure 5. Figure 4a and Figure 5a display the clustering results of the traffic flow datasets, where the position of each sample point is determined by the density, velocity, and intensity of the traffic flow. The levels 1 to 3 indicate the classification of traffic flow states, with level 1 representing the least congested ship traffic flow and level 3 representing the most congested traffic flow. The relationships between any two traffic flow variables are depicted in Figure 4b, Figure 4c, Figure 4d, Figure 5b, Figure 5c, and Figure 5d, respectively.

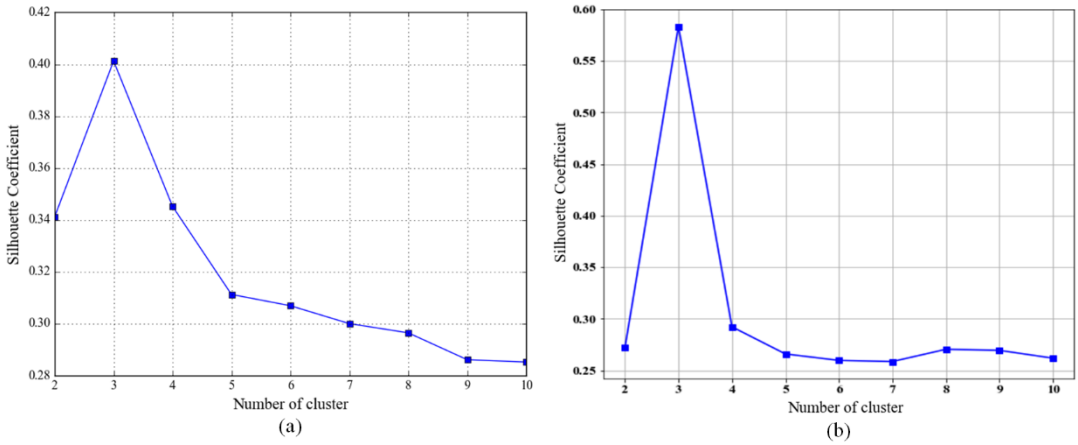

Fig. 3. Silhouette coefficient corresponding to the number of clusters

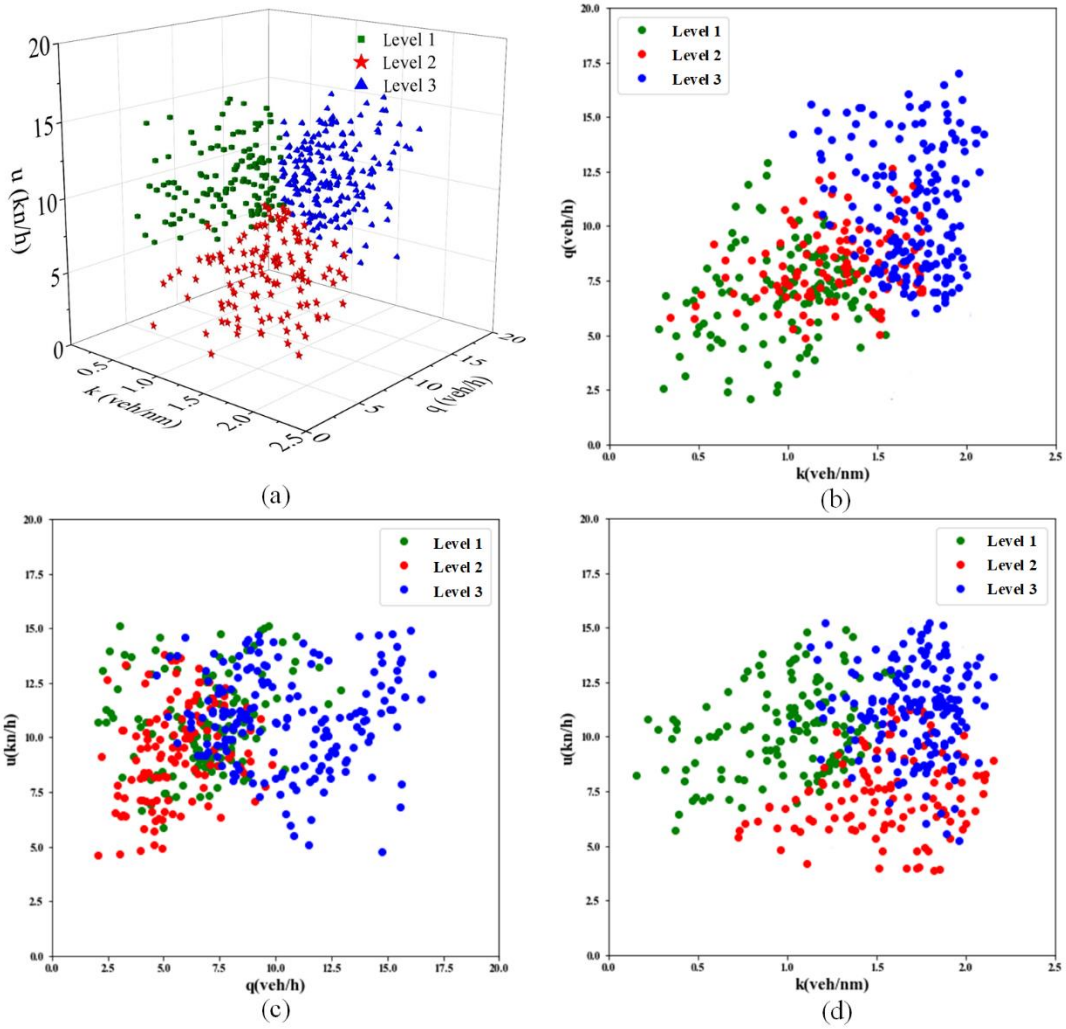

Fig. 4. Clustering results and relationships between the traffic variables in area 1 (a) Clustering results (b) Relationship between intensity and density (c) Relationship between velocity and intensity (d) Relationship between velocity and density

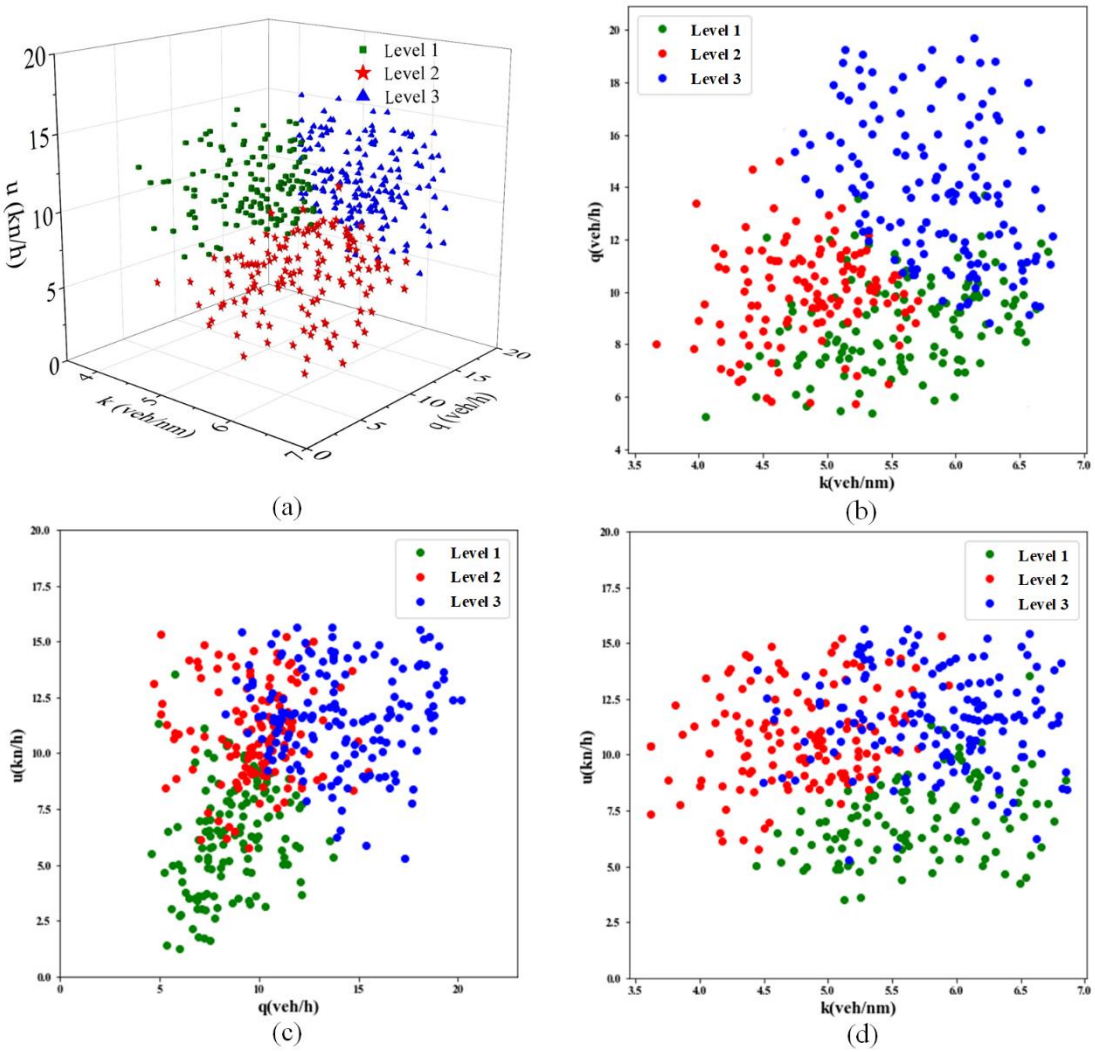

Fig. 5. Clustering results and relationships between the traffic variables in area2 (a) Clustering results (b) Relationship between intensity and density (c) Relationship between velocity and intensity (d) Relationship between velocity and density

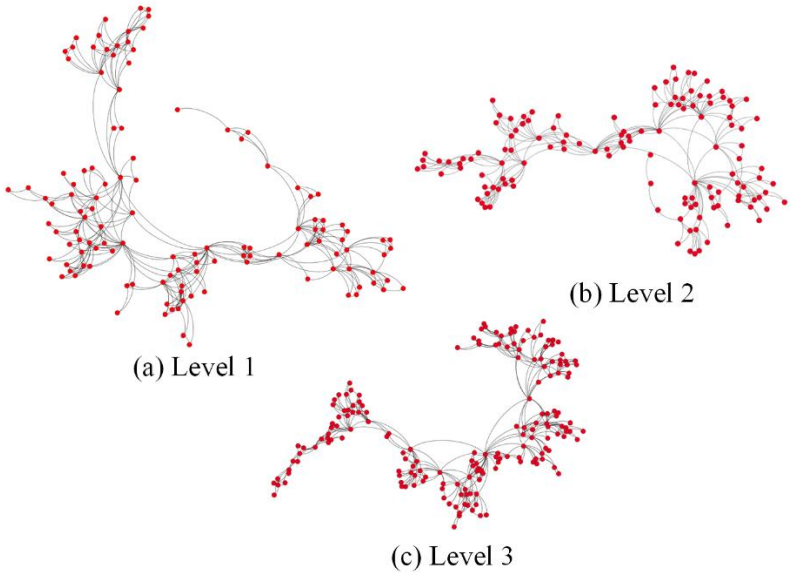

Fig. 6. Complex network diagram of traffic time series based on different states in area 1

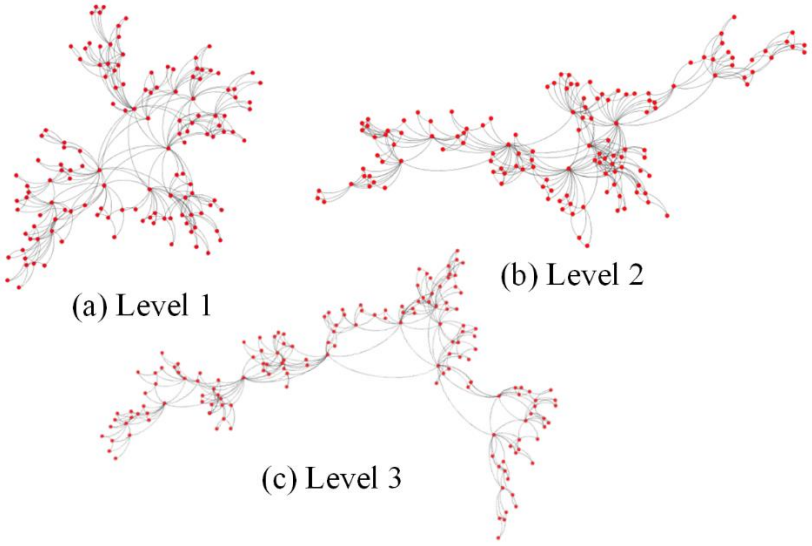

Fig. 7. Complex network diagram of traffic time series based on different states in area 2

Figures 6a and 7a visually demonstrate that the edges and nodes of the constructed VG are evenly distributed, with only local aggregation effects observed. This phenomenon suggests that, under conditions of smooth traffic (i.e., least dense traffic flow), the traffic flow parameters exhibit slow variations. In contrast, Figures 6b and 7b reveal that as the traffic state becomes moderately dense, the distribution of edges and nodes in the VG starts exhibiting aggregation effects. This indicates a higher degree of correlation among the data points in the datasets. Furthermore, Figures 6c and 7c

illustrate a gradual increase in the number of agglomerative communities within the VG, with more pronounced characteristics of these agglomerative communities. These findings suggest a strong association between the traffic flow states and the network characteristics derived from the VG analysis. Based on the foregoing discussion, it can be deduced that the traffic flow states are closely related to the observed network characteristics.

The dynamic characteristics of traffic flow are further explored by analyzing parameters after developing the VG. The degree distribution of the traffic flow time series network in the two areas is presented in Fig. 8. Fig. 8a shows the results for Area 1, while Fig. 8b displays the results for Area 2. In these figures, the blue markers represent the degree distribution of the Level 3 cluster, and the blue line represents the fitting of the Level 3 degree distribution. Similarly, the red markers and line represent the Level 2 cluster, and the green markers and line represent the Level 1 cluster. The degree distribution exhibits variations corresponding to changes in traffic flow states, ranging from less dense to congested, as depicted in Fig. 8. This phenomenon indicates that the degree can serve as a key indicator for assessing the traffic flow state. However, due to the influence of various factors, the degree distribution of traffic flow time series differs across different regions. For example, in the two selected areas in this research, there is a significant difference in degree distribution, as shown in Fig. 8. Moreover, the degree distribution varies for different traffic flow states within the same region. For instance, in Fig. 8b, the three curves demonstrate different patterns, making it challenging to use a unified degree distribution to measure the traffic flow states in different regions. Therefore, this study employs the average value of traffic flow time series complex network topological indicators as an indicator for evaluating the traffic flow state level, such as average degree, clustering coefficient and network diameter. These indicators can be obtained from historical AIS data within different research areas.

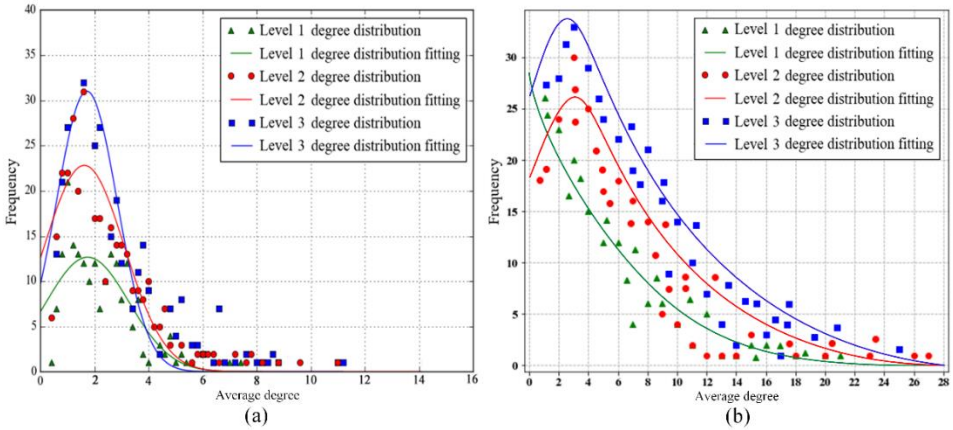

Fig. 8. Degree distribution of the developed VG for the three identified clusters

Figure 9 presents the statistical characteristics of the developed VG for the three identified clusters. In Figure 9a, 9b, and 9c, the results for Area 1 are displayed, while Figure 9d, 9e, and 9f show the results for Area 2. As the congestion level increases, the statistical characteristics of the VG exhibit variations. Therefore, it can be inferred that ship traffic states can be identified by visualizing the corresponding datasets and examining the VG's characteristics. During traffic congestion, the VG exhibit high connectivity. At this state, the respective datasets can be considered as a distinct cluster, as evidenced by an increase in the clustering coefficient and a decrease in the network diameter.

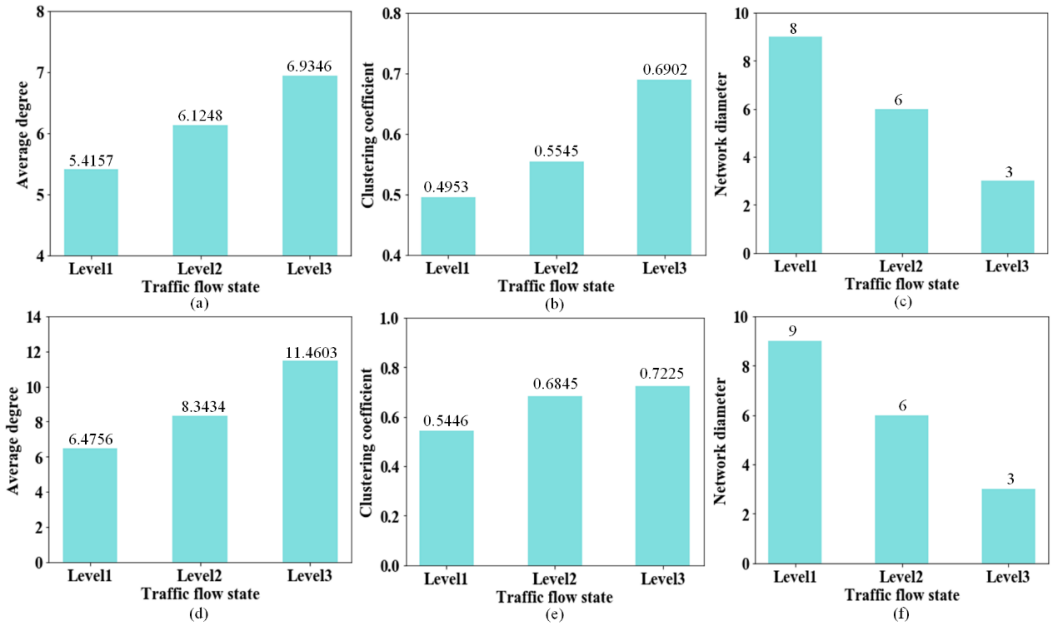

Fig. 9. Statistical characteristics of the developed VG for the three identifies clusters

4.3 Traffic flow state identification

This section focuses on the identification of traffic flow states. After undergoing data cleaning and filtering procedures, the density, velocity, and intensity parameters are calculated. Subsequently, a clustering algorithm is employed to estimate the traffic flow state labels. Table 3 displays a portion of the traffic flow data.

Table 3. Portion of preprocessed traffic flow data

Time	q	u	k	Label
2021-04-21 00:05:00	5.221285	14.69481	1.787577	Level 1
2021-04-21 00:10:00	9.079542	12.94359	1.715332	Level 1
2021-04-21 00:15:00	8.751987	15.50148	1.781311	Level 1
2021-04-21 00:20:00	8.069001	14.17205	1.813096	Level 2
…	…	…	…	…
2021-04-21 23:55:00	6.489041	10.16565	1.876292	Level 1
2021-04-21 24:00:00	5.176388	13.20179	1.892241	Level 1

Next, the sliding time window method was employed to partition the traffic flow data into fragments. The length of each fragment, denoted as ω, represents the number of data points within the fragment. Since the traffic flow state is influenced by previous time slices, each fragment incorporates the data from the previous one hour, comprising 12 time slices, as well as the current time slice data. In order to accommodate the sliding window format, the datasets are divided into multiple segments, with each segment having a length of 12 time slices and a step size of 1. Figure 10 displays the sliding windows of the analyzed datasets.

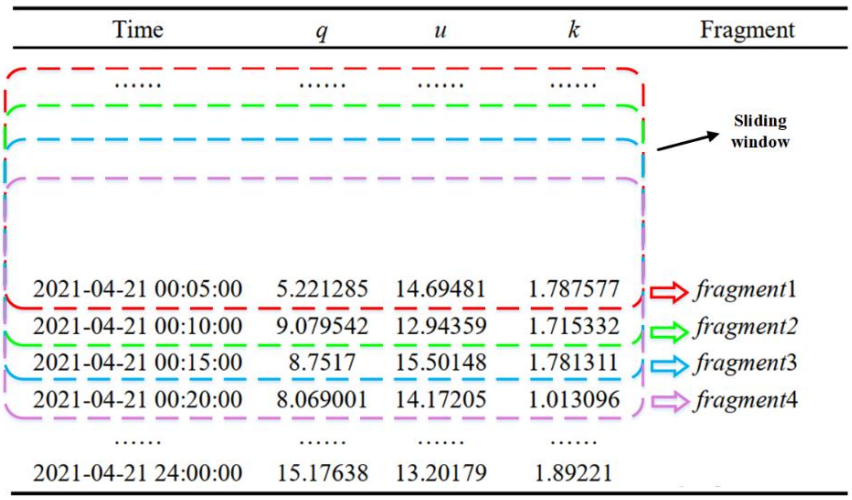

Fig. 10. Sliding window to divide the traffic flow data into fragments.

Based on the analysis, a total of 288 traffic flow states were identified, and corresponding traffic flow VG were constructed. The average degree, clustering

coefficient, and network diameter of each VG were subsequently computed. These three indicators were then compared against predefined thresholds (as shown in Figure 9). For a specific time period, the ship traffic flow state is identified if all three indicators simultaneously satisfy the real-time preset thresholds. The variations of these three indicators as a function of the time slice number in Area 1 are presented in Figure 11, while the variations in Area 2 are shown in Figure 12. Additionally, the average values of these indicators corresponding to the three clusters (Levels 1 to 3) are provided in these figures. When the traffic congestion status changes within the research area, the traffic flow state level vary in accordance with the changes in the complex network parameters proposed in this paper. As shown in Fig. 11a and 12a, the degree is directly proportional to the traffic flow state level, and when the degree exceeds the set threshold, it indicates that the traffic flow has reached a specific level. Similarly, the network clustering coefficient characterizes the aggregation characteristics of the traffic flow time series network by measuring its clustering properties. As depicted in Fig. 11b and 12b, the clustering coefficient also changes with the variation in traffic flow state levels, showing a direct positive correlation. On the contrary, the network diameter exhibits a negative correlation with the traffic flow state level. When the traffic flow is smooth, the time series network demonstrates an increased diameter. Conversely, as the traffic flow becomes congested, the diameter of the time series network decreases. Therefore, this characteristic can be utilized to assess the traffic flow status within the research area, as illustrated in Fig. 11c and 12c.

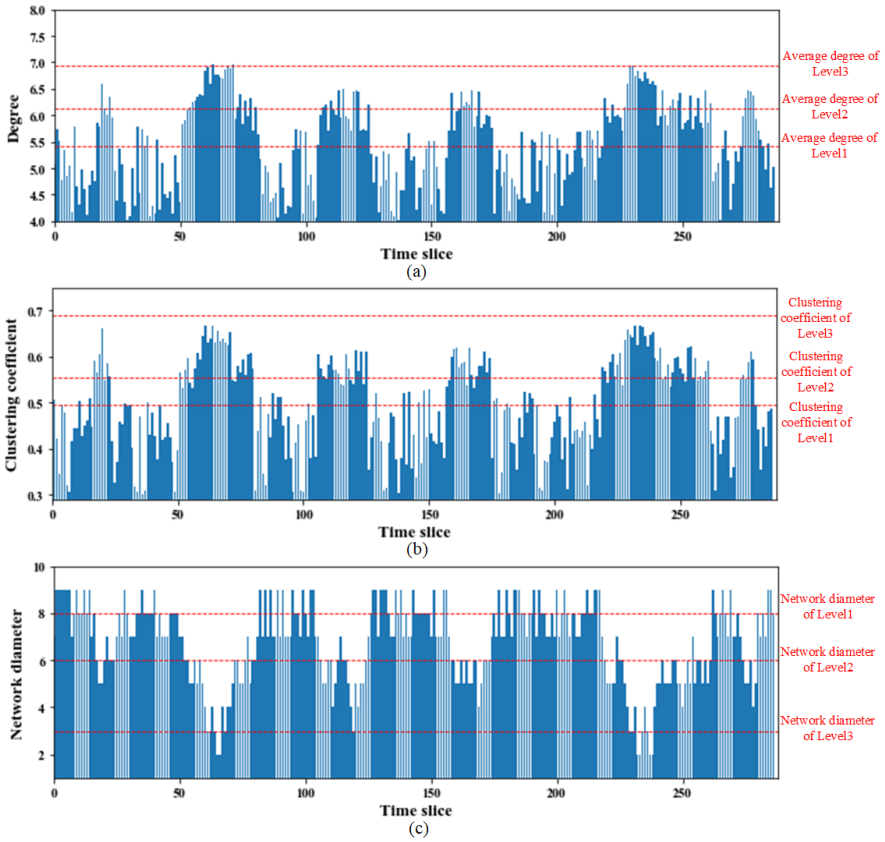

Fig. 11. Indicators of ship traffic flow complex network in area1

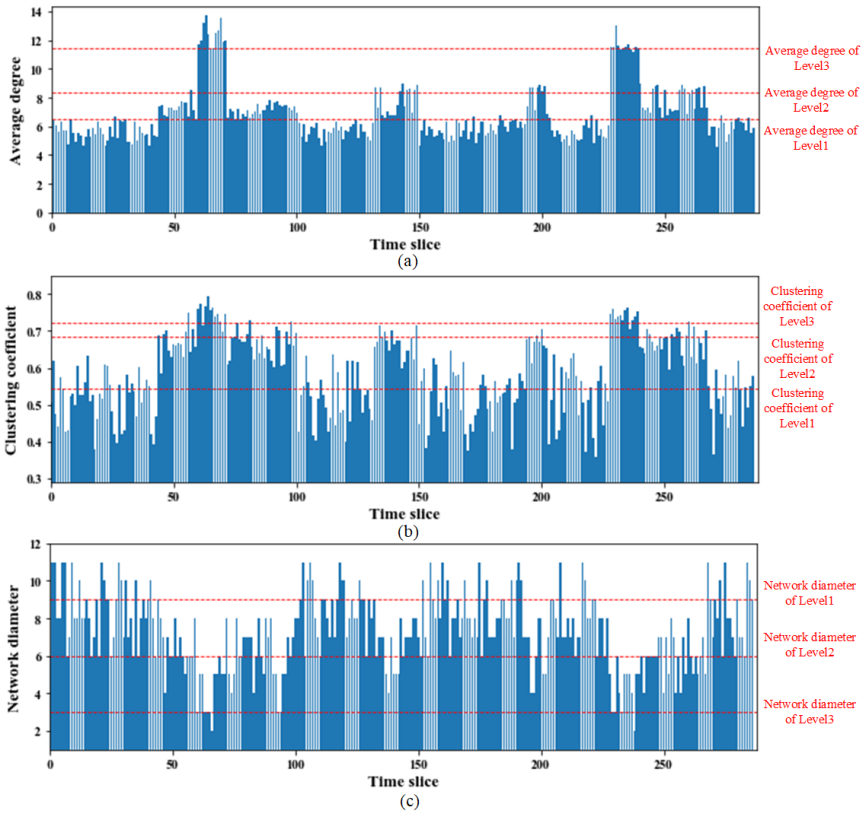

Fig. 12. Indicators of ship traffic flow complex network in area 2

4.4 Verification

The ship traffic flow state identification results based on VG in two areas are presented in Figure 13a and Figure 14a. In Area 1, the identification process achieved an ACC of 86.8%, a MAE of 0.1324, a MAXAE of 1, and a RMSE of 0.3639. In Area 2, the identification process achieved an ACC of 85.76%, an MAE of 0.1465, a MAXAE of 1, and an RMSE of 0.3716. The identification performance of the proposed method is summarized in Table 4. The results demonstrate that the proposed method possesses the capability to effectively identify the ship traffic flow state.

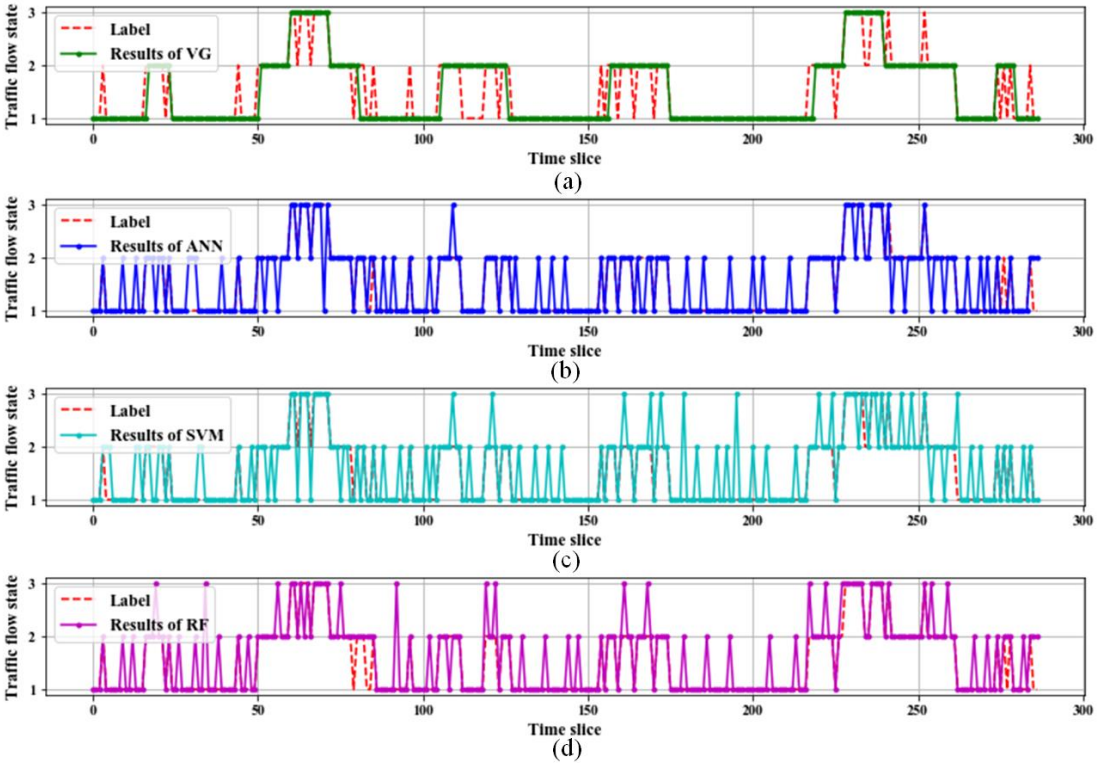

Fig. 13. Identification of the ship traffic flow states in area1

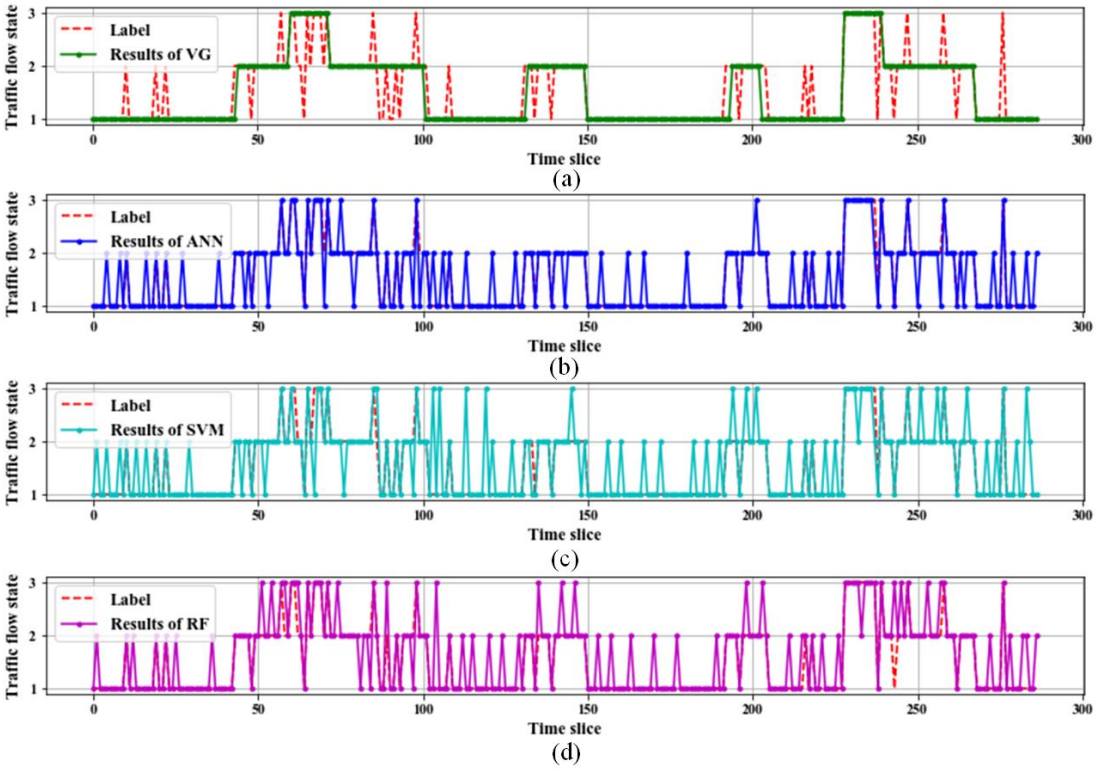

Fig. 14. Identification of the ship traffic flow states in area2

Table 4. Performance of the proposed model in two research areas

Performance index	Area 1	Area 2
ACC	86.8%	85.76%
MAC	0.1324	0.1465
MAXAE	1	1
RMSE	0.3639	0.3716

In this study, comparison tests were conducted to validate the effectiveness of the proposed method. Ship traffic flow state identification was performed using Artificial Neural Networks (ANN), Support Vector Machines (SVM), and Random Forest (RF). After conducting multiple trials, the hyperparameters for each model utilized in the experiments are presented in Table 5. The results of these comparative experiments are illustrated in Figure 13b, Fig. 13c, Fig. 13d, and Fig. 14b, Fig. 14c, Fig. 14d. The results demonstrate that the VG approach outperforms other conventional techniques, such as ANN, SVM, and RF, in terms of traffic flow prediction performance.

Table 5. Hyperparameter setting of models

Model	Hyperparameter	
ANN	Neuron	12
	Time step	5
	Number of iterations	300
SVM	Kernel function	Radial basis function
	Penalty factor	0.8
	Number of iterations	500
RF	*n*_estimators	100
	bootstrap	True
	oob_score	True

Table 6. Performance comparison of different models

Research area	Models	Evaluation indexes			
		ACC	MAC	MAXAE	RMSE
Area1	VG	86.8%	0.1324	1	0.3639
	ANN	72.34%	0.1763	2	0.5134
	SVM	70.45%	0.2374	2	0.6235
	RF	77.93%	0.1639	2	0.4923
Area2	VG	85.76%	0.1465	1	0.3716
	ANN	75.45%	0.1697	2	0.5234
	SVM	69.45%	0.2483	2	0.6723
	RF	78.65%	0.1592	2	0.5982

To further validate the effectiveness of the proposed model, a panel of 10 experts was invited to evaluate the predicted traffic flow states corresponding to the 288 time slices. The characteristics of these experts are listed in Table 7. The evaluation results obtained from the experts were compared to the results obtained from the VG, as illustrated in Figure 15. The comparison was performed separately for Area 1 and Area 2. In Figure 15, the horizontal axis represents the VG results, while the vertical axis represents the experts' evaluation results. The numbers shown in the figure indicate the time slices where consistent results were observed between the proposed method and the experts' evaluation. The size of the red circles in the figure is proportional to the corresponding numbers. The consistency rate between the VG results and the experts' evaluation in Area 1 was found to be 77.08%. Similarly, the consistency rate between the VG results and the experts' evaluation in Area 2 was found to be 79.22%. The results obtained from the VG analysis and the experts' evaluation mutually verify each other.

Therefore, it can be inferred that the experts' evaluation results confirm the sufficient performance of the proposed method in identifying the traffic flow state.

Table 7. Experts' characteristics

No.	Expertise area	Academic degree	Years of experience
S1	Navigation Technology	Master	25
S2	Maritime management	Master	20
S3	Transportation engineering	Master	23
S4	Navigation Technology	Bachelor	25
S5	Navigation Technology	Master	17
S6	Navigation Technology	Bachelor	19
S7	Maritime management	Master	18
S8	Maritime management	Master	15
S9	Maritime management	Bachelor	13
S10	Maritime management	Bachelor	11

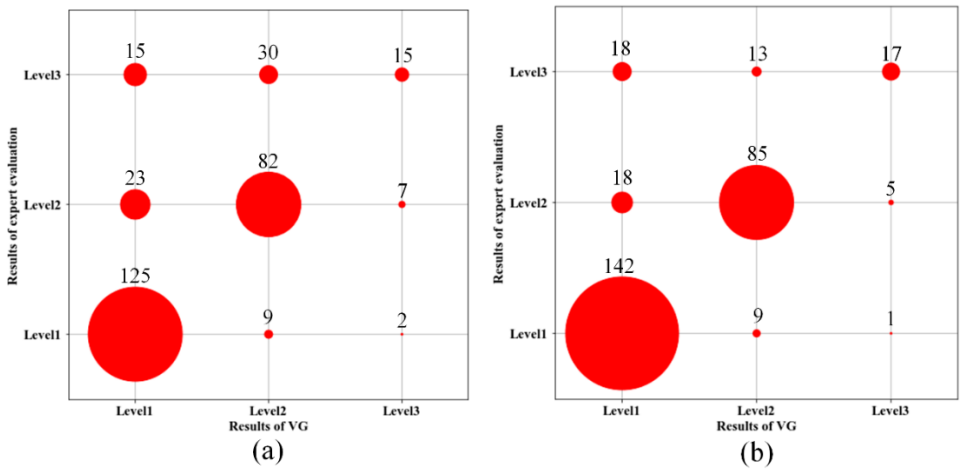

Fig. 15. (a) Experts' evaluation and VG analysis results in area 1. (b) Experts' evaluation and VG analysis results in area 2.

5. Conclusions

This research introduces a novel framework for the analysis of multi-state traffic flow using the visibility graph technique in conjunction with AIS data, with the aim of supporting the development of intelligent maritime traffic management systems. In this study, we processed actual AIS data and classified ship traffic states in the case study area using the *K*-medoids method. This allowed us to obtain datasets that accurately

characterize the multistate ship traffic flow. We then developed a complex network of ship traffic flow time series and combined it with visibility graphs to analyze the topological characteristics of ship traffic flow in different states. The key contributions of this study can be summarized as follows:

(i) Development of a framework that establishes a connection between time series data and complex networks through the visibility graph approach. By employing this method, certain characteristics of the original data are preserved when mapping the time series to networks.

(ii) Unraveling the characteristics of traffic flow time series complex networks using the visibility graph. This analysis provides valuable insights into the complex nature of traffic flow dynamics.

(iii) A novel evaluation method is proposed in this study, which utilizes data-driven method and visibility graph to evaluate the ship traffic flow states in order to support maritime transportation safety management.

Unlike traditional research methods, our study employs complex network analysis as a tool to comprehensively represent the attributes of ship traffic flow. While the visibility graph method is widely recognized for analyzing ship traffic flow states, it is important to acknowledge the limitations of the data collection method. Additionally, incorporating data from other sensors or sources such as radar and GPS, and considering factors like environmental conditions and ship types in the observed area, presents opportunities for future research and exploration.

In the context of our study, the conversion of time series data into a network structure using the VG approach requires a substantial amount of historical data. These longer data periods are essential for training the model and facilitating the analysis of various traffic flow states. Furthermore, it is crucial to note that the selection of training data at different scales can significantly influence the outcomes obtained from the VG analysis. Therefore, careful consideration should be given to the choice of training data to ensure reliable and accurate results.

To construct a training dataset encompassing multi-state traffic flows, we employed the *K*-medoids clustering algorithm. However, it is important to recognize that the boundaries between different traffic flow states can sometimes be indistinct, posing a challenge for clustering. Determining the appropriate number of clusters is crucial, as the types of traffic flow states can vary across different research areas. In this study, we utilized the contour coefficient method to estimate the number of clusters.

Nevertheless, it is important to acknowledge that some areas may lack clear-cut characteristics for analysis, making the application of this method somewhat impractical. For instance, the silhouette coefficient, commonly used to identify the optimal number of clusters, may yield a value lower than 0.5, indicating an unfavorable clustering outcome. Thus, it is essential to combine the silhouette coefficient method with expert opinions or domain knowledge in future research to determine the number of traffic flow states in a specific area, thereby enhancing the reliability and accuracy of the analysis.

Acknowledgments

The authors would like to thank the anonymous reviewers and editors for their constructive comments, which is very helpful to improve the paper. This work was supported by the National Natural Science Foundation of China (NSFC) through Grant No.52072287 and the Research Program of Sanya City through Grant No. 2022KJCX36.

Appendix 1

Table A1. List of Abbreviations

AIS	Automatic Identification System
VG	Visibility Graph
ARIMA	Auto Regressive Integrated Moving Average
SARIMA	Seasonal Auto Regressive Integrated Moving Average
ANN	Artificial Neural Networks
SVM	Support Vector Machines
RF	Random Forest
ACC	Accuracy
MAC	Mean Absolute Error
MAXAE	Maximum Absolute Error
RMSE	Root Mean Square Error

546 Table A2. List of notations

k	Density	N	The number of nodes in a complex network
q	Intensity	$\bar{k}$	The average degree of a complex network
u	Velocity	C_i	The local clustering coefficient of node i,
x_{ij}'	The normalized value	$C_{k_i}^2$	The number of all possible connections among node i's neighbours.
K	The number of cluster	$\overline{C}$	The global clustering coefficient
w	The square difference function	d_{ij}	The shortest path length between node i and node j
$S(i)$	The silhouette coefficient	D	Network diameter
k_i	The node i's degree		

547 # Appendix 2

548 Related data and model: https://github.com/zysui/visibility-graph-ship-traffic-
549 flow-time-series-analysis

550 # References

551 Asif, M.T., Dauwels, J., Chong, Y.G., Oran, A., Fathi, E., Xu, M., 2014. Spatiotemporal
552 patterns in large-scale traffic speed prediction. IEEE Trans. Intell. Transp. Syst. 15
553 (2), 794–804.
554 Barabási, A., 2009. Scale-free networks: a decade and beyond. Science 325 (5939),
555 412-413.
556 Chen, X., Wu, S., Shi, C., Huang, Y., Yang., Y., Ke, R., Zhao, J., 2020. Sensing data
557 supported traffic flow prediction via denoising schemes and ANN: A comparison.
558 IEEE Sens. J. 20 (23), 14317-14328.
559 Fangce, G., Rajesh, K., John, P., 2013. A computationally efficient two-stage method
560 for short-term traffic prediction on urban roads. Transport. Plan. Technol. 36 (1),
561 62–75.
562 Frazier, C., Kockelman, K.M., 2004. Chaos theory and transportation systems:
563 instructive example. Transp. Res. Rec. 1897 (1), 9–17.

564 Greenshields, B.D., 1934. The photographic method of studying traffic behaviour. In:
565 Proceedings of the 13th Annual Meeting of the Highway Research Board, pp. 382–
566 399.
567 Guo, J., Huang, W., Williams, B.M., 2014. Adaptive Kalman filter approach for
568 stochastic short-term traffic flow rate prediction and uncertainty quantification.
569 Transport. Res. Part C: Emerg. Technol. 43, 50–64.
570 Hou, H.Q., Li, Y.C., He, W., Liu, X.L., 2014. Vessel traffic flow distribution model of
571 bridge area waterway in the middle stream of Yangtze River. Appl. Mech. Mater.
572 551, 127–133.
573 Huang, Y., Yip, T. L., Wen, Y., 2019. Comparative analysis of marine traffic flow in
574 classical models. Ocean Eng. 187, 106195.
575 Jia, L., Teng, W., Pan, W., Xue, Y., He, H., 2014. Correlation analysis of
576 synchronization flow at a traffic bottleneck. Nonlinear Dyn. 78, 1801–1809.
577 Kang, L., Meng, Q., Liu, Q., 2018. Fundamental diagram of ship traffic in the Singapore
578 Strait. Ocean Eng. 147, 340–354.
579 Kwasniok, F., 2019. Fluctuations of finite-time Lyapunov exponents in an intermediate
580 complexity atmospheric model: a multivariate and large-deviation perspective.
581 Nonlinear Proc. Geoph. 26 (3), 195–209.
582 Lacasa, L., Luque, B., Ballesteros, F., Luque, J., Nuno, J., 2008. From time series to
583 complex networks: The visibility graph. P. Natl. A. Sci. 105 (13), 4972-4975.
584 Lam, W.H.K., Tang, Y.F., Tam, M.L., 2010. Comparison of two non-parametric models
585 for daily traffic forecasting in Hong Kong. J. Forecasting 25 (3), 173–192.
586 Li, H., 2016. Research on prediction of traffic flow based on dynamic fuzzy neural
587 networks. Neural Comput. Appl. 27 (7), 1969–1980.
588 Li, J., Boonaert, J., Doniec, A., Lozenguez, G., 2021. Multi-models machine learning
589 methods for traffic flow estimation from Floating Car Data. Transport. Res. Part
590 C: Emerg. Technol. 132, 103389.
591 Meena, G., Sharma, D., Mahrishi, M., 2020. Traffic prediction for intelligent
592 transportation system using machine learning. In 2020 3rd International
593 Conference on Emerging Technologies in Computer Engineering: Machine
594 Learning and Internet of Things (ICETCE) (pp. 145-148). IEEE.
595 Menelaou, C., Kolios, P., Timotheou, S., Panayiotou, C.G., 2017. Controlling road
596 congestion via a low-complexity route reservation approach. Transport. Res. Part
597 C: Emerg. Technol. 81, 118-136.

598 Meng, M., Shao, C.F., Wong, Y.D., Wang, B.B., Li, H.X., 2015. A two-stage short-term
599 traffic flow prediction method based on AVL and AKNN techniques. J. Cent.
600 South Univ. 22 (2), 779–786.
601 Pandurangan, G., Raghavan, P., Upfal, E., 2003. Building low-diameter peer-to-peer
602 networks. IEEE J. Sel. Area. Comm. 21 (6), 995-1002.
603 Park, H., Jun, C., 2009. A simple and fast algorithm for *K*-medoids clustering. Expert
604 Syst. Appl. 36 (2), 3336-3341.
605 Pedersen, P. T., 2010. Review and application of ship collision and grounding analysis
606 procedures. Mar. Struct. 23 (3), 241-262.
607 Rong, H., Teixeira, A. P., Soares, G., 2022. Maritime traffic probabilistic prediction
608 based on ship motion pattern extraction. Reliab. Eng. Syst. Safe. 217, 108061.
609 Rousseeuw, P., 1987. Silhouettes: A Graphical Aid to the Interpretation and Validation
610 of Cluster Analysis. J. Comput. Appl. Math. 20, 53-65.
611 Sang, L. Z., Wall, A., Mao, Z., Yan, X. P., Wang, J., 2015. A novel method for restoring
612 the trajectory of the inland waterway ship by using AIS data. Ocean Eng. 110, 183-
613 194
614 Sun, H., Liu, H., Xiao, H., He, R., Ran, B., 2003. Use of local linear regression model
615 for short-term traffic forecasting. Transport. Res. Rec. J. Transport. Res. Board
616 1836 (1), 143–150.
617 Shelmerdine, R.L., 2015. Teasing out the detail: how our understanding of marine AIS
618 data can better inform industries, developments, and planning. Mar. Policy 54, 17–
619 25.
620 Silveira, P.A.M., Teixeira, A.P., Soares, C.G., 2013. Use of AIS data to characterise
621 marine traffic patterns and ship collision risk off the coast of Portugal. J. Navig.
622 66, 879–898.
623 Sui, Z., Wen, Y., Huang, Y., Song, R., Piera, M. A., 2023. Maritime accidents in the
624 Yangtze River: A time series analysis for 2011–2020. Accident Anal. Prev. 180,
625 106901.
626 Sui, Z., Wen, Y., Huang, Y., Zhou, C., Du, L., Piera, M. A., 2022. Node importance
627 evaluation in marine traffic situation complex network for intelligent maritime
628 supervision. Ocean Eng. 247, 110742.
629 Sui, Z., Wen, Y., Huang, Y., Zhou, C., Xiao, C., Chen, H., 2020. Empirical analysis of
630 complex network for marine traffic situation. Ocean Eng. 214, 107848.
631 Suo, Y., Ji, Y., Zhang Z., Chen, J., Claramunt, C., 2022. A formal and visual data-mining

model for complex ship behaviors and patterns. Sensors 22.14, 5281.
Tang, J., Wang, Y., Wang, H., Zhang, S., Liu, F., 2014. Dynamic analysis of traffic time series at different temporal scales: A complex networks approach. Physica A. 405, 303-315.
Tsou, M., 2010. Discovering knowledge from AIS database for application in VTS. J. Navig. 63, 449-469.
Vaishnav, V., Vajpai, J., 2018. Seasonal time series forecasting by group method of data handling. In 2018 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS) (pp. 1-5). IEEE.
Wang, X., Li, J., Zhang, T., 2019. A machine-learning model for zonal ship flow prediction using AIS data: A case study in the south atlantic states region. J. Mar. Sci. Eng. 7 (12), 463.
Watts, D., Strogatz, S., 1998. Collective dynamics of 'small-world' networks. Nature 393 (6684), 440-442.
Wen, Y., Huang, Y., Zhou, C., Yang, J., Xiao, C., Wu, X., 2015. Modelling of marine traffic flow complexity. Ocean Eng. 104, 500-510.
Williams, B.M., Hoel, L.A., 2003. Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: theoretical basis and empirical results. J. Transp. Eng. 129 (6), 664-672.
Xiao, F., Ligteringen, H., van Gulijk, C., Ale, B., 2015. Comparison study on AIS data of ship traffic behavior. Ocean Eng. 95, 84-93.
Yamaguchi, A., Sakaki, S., 1971. Traffic surveys in Japan. J. Navig. 24, 521–534.
Yan, Y., Zhang, S., Tang, J., Wang, X., 2017. Understanding characteristics in multivariate traffic flow time series from complex network structure. Physica A. 477, 149-160.
Yip, T.L., 2013. A marine traffic flow model. TransNav: Int. J. Mar. Navig. Saf. Sea Transp. 7 (1), 109-113.
Yoon, B., Chang, H., 2014. Potentialities of data-driven nonparametric regression in urban signalized traffic flow forecasting. J. Transp. Eng. 140 (7), 04014027.
Yu, B., Song, X., Guan, F., Yang, Z., Yao, B., 2016. K-nearest neighbor model for multiple-time-step prediction of short-term traffic condition. J. Transp. Eng. 142 (6), 04016018.
Zhang, L., Meng, Q., Fwa, T. F., 2019. Big AIS data based spatial-temporal analyzes of ship traffic in Singapore port waters. Transport. Res E: Logist. Transport. Rev. 129,

287-304.

Zhang, M., Kujala, P., Hirdaris, S., 2022. A machine learning method for the evaluation of ship grounding risk in real operational conditions. Reliab. Eng. Syst. Safe. 226, 108697.

Zhang, M., Montewka, J., Manderbacka, T., Kujala, P., Hirdaris, S., 2021. A big data analytics method for the evaluation of ship-ship collision risk reflecting hydrometeorological conditions. Reliab. Eng. Syst. Safe. 213, 107674.

Zhang, M., Zhang, D., Fu, S., Kujala, P., Hirdaris, S., 2022. A predictive analytics method for maritime traffic flow complexity estimation in inland waterways. Reliab. Eng. Syst. Safe. 220, 108317.

Zhang, Y., Guan, W., 2009. Complexity measure of traffic flow based on union entropy and Co-complexity. Comput. Eng. Appl. 46(15): 22-4.

Zhao, C., Li, X., Zuo, M., Mo, L., Yang, C., 2022a. Spatiotemporal dynamic network for regional maritime vessel flow prediction amid COVID-19. Transport Policy 129: 78-89.

Zhao, J., Chen, Y., Zhou, Z., Zhao, J., Wang, S., Chen, X., 2022b. Extracting vessel Speed based on machine learning and drone images during ship Traffic Flow Prediction. J. Adv. Transpor. 2022, 3048611.

Zhou, T., Jiang, D., Lin, Z., Han, G., Xu, X., Qin, J., 2019a. Hybrid dual Kalman filtering model for short-term traffic flow forecasting, IET Intell. Transp. Syst. 13 (6), 1023–1032.

Zhou, Y., Daamen, W., Vellinga, T., Hoogendoorn, S.P., 2019b. Ship classification based on ship behavior clustering from AIS data. Ocean Eng. 175, 176-187.

Zhu, L., Yu, F. R., Wang, Y., Ning, B., Tang, T., 2018. Big data analytics in intelligent transportation systems: A survey, IET Intell. Transp. Syst. 20(1), 383-398.