Extraction of Key Factors of Ship Following Behavior based on Rough Set Theory

Yihua Liu, Bo Tu, Xinyue Li, and Shiyu Tu

***Abstract* — In order to clarify the main influencing factors of ship following behavior, an algorithm for extracting key factors of ship following behavior based on ship AIS data in straight channel waters is proposed. After the statistical analysis of the obtained ship following trajectory data, the correlation of each influencing factor of ship following behavior and the importance of their influence on the ship following distance are analyzed by applying the rough set dependency and importance assignment methods, and the attribute simplification algorithm based on the rough set dependency and the decision rule algorithm based on the rough set attribute simplification are used to mine the strong correlation rules between each influencing factor and the following distance. Finally, set the straight section of the south channel of the Yangtze River estuary as the study water for case study. The results show that the average ship speed is 9.41 knots, the average space of ship is 2110 m, the average bow time is 450s, and the average relative speed is 0 knot; the weights of ship speed, ship length, relative speed, ship type, density and pilot class in the factors affecting ship following behavior are 0.1953, 0.0836, 0.2092, 0.2265, 0.1852 and 0.1058, respectively; the correlation between ship speed and relative speed is large, no strong correlation rule was found with others; ship following behavior should focus on ship speed and density, and the distance of [1000, 2000 m) is the key concern interval.**

***Keywords* — AIS data, association rules, attribute importance, MASS, rough set theory, ship following behavior.**

I. Introduction

At present, the high-speed development of communication, information, sensing, computer, and other high-techs have provided strong support for the development of intelligent ships.MASS (Maritime Autonomous Surface Ships) have also become the focus of scholars at home and abroad. Internationally, IMO (International Maritime Organization) put forward the concept of the MASS on water for the first time in the 99th MSC, and then the EU, Norway, France, Japan, and Russia have also carried out research projects, related support laws and regulations for autonomous navigation of ships on water. In China and in 2015, the China Classification Society (CCS) released the "Intelligent Ship Specification", which put forward specific requirements for 6 major functions of intelligent ships. 2018, the Ministry of Industry and Information Technology (MIIT) released the Intelligent Ship 1.0 special project to carry out research and development of ship intelligence technology. 2019, the Ministry of Science and Technology launched the National Key Research and Development Program Project – Ship Intelligent Navigation and Control Based on Ship-Shore Collaboration Key Technology Research.

MASS are mainly divided into intelligent navigation, shore-based remote driving control, ship formation navigation, and other intelligent modules, among which intelligent navigation is particularly important. In an intelligent navigation, the ship needs to complete the process of situational awareness, intelligent collision avoidance, and intelligent control. A ship following model is the key technology in ship autonomous navigation and the painful and difficult point that needs to be solved urgently. In establishing and optimizing the ship following model, it is necessary to accurately extract the marine ship following data, consider the important factors influencing the following behavior, and analyze the relationship between the factors influencing the ship tracking behavior. Therefore, this paper carries out the work related to the extraction of the ship following behavior.

In this paper, the data extraction algorithm for ship following is firstly designed and the data is analyzed descriptively. Then, the relationship between the influencing factors of ship following behavior was analyzed, and the importance of each influencing factor on the ship following spacing was determined. Finally, a decision rule algorithm based on rough set theory is applied to mine the association rules in ship following behavior.

II. Related Work

The following models have a long history of development in road traffic engineering. The concept of car-following was introduced by Reuschel [1] and Pippes [2] in the 1950s. After nearly 70 years of development, car-following models are broadly classified into two categories, traffic engineering and statistical physics, depending on the research focus. The main heeling models are the safety distance model [3], stimulus-response model [4], psycho-physiological model [5], optimal speed model [6], and intelligent driving model [7]. However, in marine traffic engineering, the ship following theory is still in its infancy. Zhu Jun [8] used GM model to establish the ship following model and the relationship between bow spacing and ship speed. He Liangde [9] determined the safe spacing of inland river vessels, substituted the traffic flow-

Submitted on April 09, 2022.
Published on May 21, 2022.
Y. Liu, College of Merchant Marine, Shanghai Maritime University, China.
(e-mail: liuyh@shmtu.edu.cn)
B. Tu, College of Merchant Marine, Shanghai Maritime University, China.
(e-mail: tubo2580@163.com)
X. Li, College of Merchant Marine, Shanghai Maritime University, China.
(e-mail: 743792185@qq.com)
Shiyu Tu, College of Merchant Marine, Shanghai Maritime University, China.
(e-mail: 532238819@qq.com)

density-velocity constant equation, obtained the ship following spacing model, and used it to introduce the ship flow limit. Mingli [10] established a longitudinal spacing calculation model for mega-ships according to the requirements of ship safety spacing. Li Zhenfu [11] established a model for the Arctic route based on Gippes' theory of healing based on the safety distance and the ship spacing. Most of the above researchers applied the road-following model directly to the marine traffic research, which did not sufficiently extract the ship-following data and consider the influencing factors of ship-following.

Rough set theory is a kind of incomplete data without any prior knowledge, and it can find the importance of each attribute and perform attribute simplification without any a priori knowledge other than the original data. And it can also find out the dependency between data, reveal the pattern between data, and mine the association rules between data. Due to these characteristics, rough set theory is widely used in weight analysis and data association rule mining in various fields. And it can also be well applied in weight assignment and association rule mining of factors influencing ship following behavior in this paper. Su YH [12] introduced the rough set center of gravity theory to calculate the objective weight set, applied the decision table method in the rough set theory to calculate the objective weights of each factor. Zhang [13] combined rough set theory and AHM, determined the evaluation index weights of innovation capability of equipment manufacturing enterprises and proved the feasibility of the method by example. Based on rough set theory and particle calculation, Zhou Danchen [14] proposed an objective assignment method integrating fuzzy quotient space theory and rough set theory, and the example was applied to job performance evaluation.Dan Zhou [15] applied the algorithm related to rough set theory to analyze the influencing factors in the field of ships and instantiated 2017 AIS data to approximate and determine the respective weights of the influencing factors. Zhen Shi [16] combined rough set theory and knowledge granularity to accurately calculate the weights of knowledge features, which provided a feasible method for accurate knowledge cognition and effective knowledge management. Wu Sen [17] applied the rough set theory of weight determination to the analysis of pressure in the mining site, and accurately and scientifically analyzed the importance of ground pressure revealing factors. Yanping Zhang [18] proposed a method to evaluate the importance of rail transit nodes based on rough set theory and proved by example that the importance determined by this method is more accurate. Shi Fuqian [19] proposed an association rule mining method for perceptual knowledge based on rough set and obtained a strong association rule set of key features to perceptual descriptions. Y. Zhang [20] proposed a method of applying rough set theory to obtain rules by selecting important features as conditional attributes and implementing the algorithm to obtain final rules with good results in example analysis. Zheng [21] proposed a data mining method combining rough set analysis and classical association rules, and the scientificity and effectiveness of the method were verified by practical cases of road transportation management. Wang Ning [22] constructed a method for mining probabilistic rules in emergency cases based on rough sets, applied a genetic algorithm for attribute approximation of emergency case decision table and then obtained probabilistic rules.

III. Datasets

A. Data Sources

The waters of the south channel of the Yangtze River estuary in Shanghai (121.821127E-121.869621E), (31.164485N-31.20507N) were selected as the test waters for this experiment. This water is free for ships to navigate and there is no traffic control, so the ship's following behavior can be well studied.

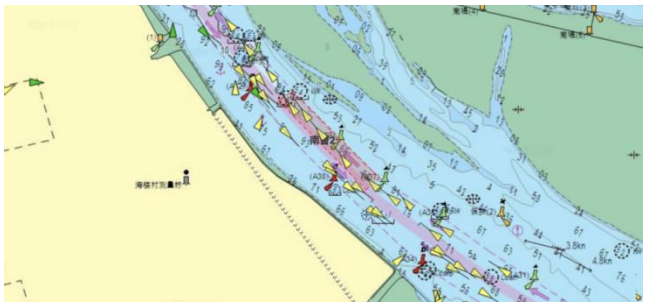

Fig. 1. Test waters.

The data in this paper are derived from the AIS (Automatic Identification System) data in the East China Sea of Shanghai from January to October 2018, from which the above-mentioned data of the South Channel test waters were screened. According to the requirements of the China Maritime Administration, all international sailing ships of 300 GT and above, non-international sailing coastal ships of 500 GT and above, all passenger ships are rigidly equipped with A-class AIS; coastal sailing ships of 200 GT to 500 GT, all harbor craft tugboats and self-propelled ships involved in underwater construction operations, 100 GT and above sailing in the Yangtze River Main Line, Pearl River Main Line, Beijing-Hangzhou Canal and Huangpu River of the inland river. The AIS data contains both static and dynamic information of the ship, the static information includes ship MMSI code, ship name, call sign, ship type, ship length, ship width, antenna distance from bow and stern, antenna distance from port and starboard, etc. The dynamic information includes time, ship-to-ground speed, ship-to-ground course, latitude, longitude, etc. The popularization of AIS equipment, the perfect dynamic and static messages in the AIS data provide strong data guarantee for the research of ship following behavior. The data format of the AIS part is shown in Table I, where *a* represents the antenna distance from the bow and *c* represents the antenna distance from the port side of the ship.

B. Data Preparation

Data Cleaning. In the process of processing AIS data, some abnormal data were found, such as negative ship speed, excessive ship speed, and duplicate data. For the duplicated data, the direct deletion method is used; for the abnormal speed data, the speed in the AIS data of the previous moment and the next moment of that ship is used for replacement.

Data interpolation. AIS device messages are sent at non-equal intervals, resulting in receiving AIS data at non-equal

TABLE I: AIS PARTIAL DATA FORMAT TABLE

MMSI	Time	Longitude	Latitude	speed	type	length	width	a	c
413443120	2018-01-12 12:21:32	121.85442	31.17772	8.8	Petroleum chemical ships	97	15	80	10
413443120	2018-01-12 12:25:49	121.84611	31.18535	8.6	Petroleum chemical ships	97	15	80	10

TABLE II: EFFECTIVE SHIP FOLLOWING DATA TABLE

Speed	Length	Type	R_speed	K	Grade	T_gap	Space
12.19	139	Container	-0.84	1	2	527.88	3310.39
5.91	132	Dry Cargo ship	-0.12	1	1	529.4	1609.58
9.44	148	Container	1.47	2.33	1	359.2	1744.42
10.7	115	Bulk	4.65	1	2	937.09	5158.25
6.58	88	Dry Cargo ship	2.05	1	1	598.04	2024.38
5.5	112	chemical ship	2.81	1	2	1077.02	3047.37

intervals as well. The ship following study is about the behavior of two ships at the same time and space, so the interpolation method is needed to process the AIS data with equal intervals. In this experiment, the speed and position information in the AIS data is interpolated at equal time intervals by using the cubic spline interpolation method. Finally, the processed data are stored in CSV format, and then the required valid ship following data are extracted.

Extraction of the valid ship following data. The pre-processed ship AIS data of the selected area is applied to the ship following behavior extraction algorithm, and finally 3426 ship following trajectories are obtained. The following distance and following time distance between ships are obtained according to equations (3) and (4), and the data are unified and summarized in a CSV file, part of which is shown in Table II.

IV. METHODOLOGY

A. Rough Set Theory

1) Upper approximation, lower approximation

Given the decision table $S = (U, C \cup D, V, f), X \subseteq U$和$R \subseteq C$. The upper approximation set:$R^*(X) = \cup \{Y | Y \in U/R$且$Y \cap X \neq \Phi\}$, the upper approximation denotes the set of all objects that may belong to *X*. The lower approximation set ； $R_*(X) = \cup \{Y | Y \in U/R$且$Y \subseteq X\}$ ， the lower approximation denotes the set of all objects that must belong to *X*. The positive region of the set *S* over *R* is denoted as $POS_R(S)$.

2) Attribute Dependency

Given a knowledge system, $S = (U, A, V, f)$ $\forall P, Q \in IND(U)$, define:

$$\gamma_p(Q) = k = \frac{|pos_p(Q)|}{|U|} \tag{1}$$

Define (1) is the degree to which knowledge *Q* depends on knowledge *P*.

3) Relative Attribute Importance

Let a knowledge system $S = (U, A, V, f)$, $A = C \cup D$, $C \cap D = \Phi$, $\forall B \subseteq C$, $\forall \beta \in C$, $\forall \alpha \in C - B$, when $D \neq \Phi$,

$$sig(\beta, C; D) = \gamma_C(D) - \gamma_{C-\{\beta\}}(D) \tag{2}$$

That is called the importance of the conditional attributes to the full set of conditional attributes *C* with respect to the decision attribute *D*.

4) Relative property approximation and relative kernels

P and *Q* are two sets of equivalence relations on *U*, *P*∈*Q*, *P* is a *B*-approximation of *Q*, *P* is a B-independent subset of *Q*, and $POS_P(B) = POS_Q(B)$, then P is said to be a relative approximation of *Q*, or relative approximation for short. The set consisting of all properties in *M* that form an equivalence relation with *B* is called the *B*-core of *M*, denoted as $CORE_B(Q)$.

5) Rule Definition

In the knowledge system, let $U / IND(O.C) = \{X_1, X_1, \cdots X_P\}$, $U / IND(O.D) = \{Y_1, Y_2, \cdots, Y_q\}$, $Des_{O.C}(X_i)$ and $Des_{O.C}(Y_i)$ denote the description of the equivalence class $X_i (i = 1, 2, \cdots, p)$ and $Y_j (j = 1, 2, \cdots, q)$ respectively. $X_i \bigcap Y_j = \phi$ and the rule is defined as: $r_{ij} : Des_{O.C}(X_i) \to Des_{O.D}(Y_i)$. Rules can be evaluated by the following three metrics: (1) support $sup(r_{ij}) = |X_i \bigcap Y_j| / |U|$, which defines the strength of a decision rule; (2) confidence $cer(r_{ij}) = |X_i \bigcap Y_j| / |X_i|$, when $cer(r_{ij}) = 1$, the rule is deterministic and the conditional attributes uniquely describe the decision attributes, and when $0 < cer(r_{ij}) < 1$, the conditional attributes probabilistically describe the decision attributes; (3) coverage $cov(r_{ij}) = |X_i \bigcap Y_j| / |Y_i|$, it is the proportion of data objects that satisfy both the antecedent and the consequent parts of the rule among the data objects that satisfy its consequent parts.

B. Mutual Information Theory

Information entropy can measure the uncertainty of a variable, let *X* be a discrete random variable with a value domain of *V* and its probability distribution function is $p(x) = P(X = x), x \in V_x$. The information entropy is defined as:

$$H(X) = -\sum_{x \in V_x} p(x) \log_2 p(x) \tag{3}$$

Given a variable X, the uncertainty of variable Y can be measured by the following conditional entropy measure.

$$H(Y \mid X) = -\sum_{x \in V_x} p(x) H(Y \mid X = x) = -\sum_{x \in V_x} \sum_{x \in V_x} p(x,y) \log_2 p(y \mid x) \quad (4)$$

The degree of correlation between two variables X and Y can be measured by the mutual information, which is defined as follows:

$$I(X \mid Y) = -\sum_{x \in V_x} \sum_{x \in V_x} p(x,y) \log_2 \frac{p(x \mid y)}{p(x)p(y)} \quad (5)$$

The mutual information values of X and Y can reflect the correlation between them. If the mutual information values of X and Y are large (small), then the correlation between X and Y is large (small). If the mutual information of X and Y is 0, then X and Y are not correlated at all.

C. Algorithm for Extracting Correlation Factors of the Ship Following Behavior

1) Definition of the ship following behavior

When two ships are at the same time and space, the rear ship follows the front ship and changes according to its change (acceleration and deceleration), and this process is called "the ship's following behavior". The ship following distance is the actual distance between the stern of the front ship and the bow of the ship at the same time. Firstly, the ship's latitude (φ) and longitude (λ) are converted into coordinates under the X-Y coordinate system, and the positive half-axis of the X-axis represents the longitudinal distance of the ship moving from the origin, and the positive half-axis of the Y-axis represents the lateral distance of the ship moving from the origin. 5 times of the ship's width is selected as a splitting course on Y-axis, and the ship's space-time trajectory in the X-axis direction is drawn according to the change of time, and then 20 s time is used as the interval to make truncation on the space-time trajectory. The distance between two ships in the X-axis direction before and after the same moment is the gap.

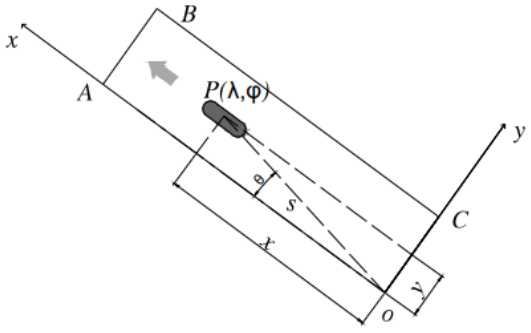

Fig. 2. X-Y coordinate system.

The ship's following distance and ship time interval satisfy the following relationship equations.

$$Gap = X_1 - X_2 - P_1 - P_2 \quad (6)$$

$$T\_gap = Gap / v \quad (7)$$

where Gap indicates the following distance between the front and rear ships, X_1 indicates the distance of the front ship in the longitudinal direction, X_2 indicates the distance of the rear ship in the longitudinal direction, P_1 indicates the distance between the AIS antenna and the transom of the front ship, P_2 indicates the distance between the AIS antenna and the bow of the rear ship, v indicates the sailing speed of the rear ship, and T_gap indicates the following time interval of the ship.

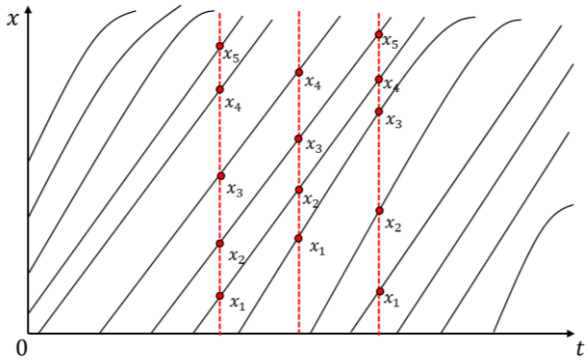

Fig. 3. Schematic diagram of ship following behavior.

2) Ship following behavior extraction algorithm design

First, filter the ship following data according to the definition of ship following behavior. Then judge whether the two ships before and after are on the same channel, and judge whether both ships pass the beginning and the end of that channel. Next, judge whether their course change angle is less than 5 degrees (the ship straight deviation angle is generally not more than 5 degrees), and finally judge whether the ship following time is more than 2 minutes (the ship pilot decision time is generally 2 minutes). The specific algorithm is shown as follows.

STEP 1: Input ship AIS data.

STEP 2: Clean the abnormal AIS data.

STEP 3: Apply the third spline interpolation method to interpolate the AIS data with a 20 s time interval.

STEP 4: Sort the AIS data in chronological order and selecting the ships with time intersection.

STEP 5: Convert latitude and longitude in AIS data into X and Y values in the X-Y coordinate system.

STEP 6: Judge whether the ships are in the same segmented channel according to the Y value, if yes, keep them, otherwise delete them.

STE P7: Save the data of the two ships before and after into a group according to the size of X value sorting.

STEP 8: Judge whether the deviations of the two ships' heading angle when entering and leaving the water in the group data are both less than 5°, if yes, keep them, otherwise delete them.

STEP 9: Judge whether the sailing time of two ships in the group data is more than 2 min, if yes, keep, otherwise delete.

STEP 10: Output the AIS data of the ship following track.

Attribute dependency-based rough set approximation and decision rules

Input: Decision table $T = (U, C, d, V, f)$ support,

confidence, coverage functions $(\min\_\sup, \min\_cer, \min\_\text{cov})$.

① Calculate the dependence $\gamma(D)$ of the decision attribute *D* on the conditional attribute *C* according to (1).

② Calculate the importance of each attribute $c \in C$ with respect to *D* in *C* $sig^{D}_{C-\{c\}}(c)$ according to (2) and let $core_D(C) = \Phi$. If $sig^{D}_{C-\{c\}}(c)$ is not 0, then $core_D(C) = core_D(C) \bigcup \{c\}$, get $core_D(C)$ as the relative kernel of *C* with respect to *D*. If $sig^{D}_{C-\{c\}}(c)$ is 0, then terminate the calculation, otherwise execute ③.

③Make $E = core_D(C)$, and repeat for the set of attributes *C-E*.

1) for each attribute $c \in C - E$, calculate the importance $sig^{E}_{D}(c)$ of attribute *c* with respect to *E* with respect to *D* by (2).

2) select the attribute *c* such that it satisfies $sig^{E}_{D}(c) = \max_{c' \in C-E} sig^{E}_{D}(c'), \mathrm{E} = \mathrm{E} \bigcup \{\mathrm{c}\}$.

3) If $\gamma_E(D) = \gamma_C(D)$, then terminate (at this point *E* is a simplification of *C*), otherwise turn ①.

④ According to the result of the minimum relative simplification, the association rule that satisfies $(\min\_\sup, \min\_cer, \min\_\text{cov})$ is obtained.

Output: Minimum relative attribute approximation and association rules satisfying confidence, support, and coverage.

V. Experiment

A. A. Correlation Factor Analysis

1) Data analysis

In this paper, the width of the selected waters is 800 m, the average length of the ships passing through the waters in 2018 is 100 m and the average ship width is 20 m. The whole waters are divided into eight segmented channels according to five times the ship width, and the data in the segmented channels with *Y*-values in the 400-500 m range are selected for subsequent analysis. After applying the ship following behavior extraction algorithm to extract the effective following data of the south channel of the Yangtze River Estuary, a descriptive analysis of these data was conducted. The average ship following speed is 9.41 knots, the average following distance is 2110 m, the average following time distance is 450 s, and the average relative speed is 0 knots. The peak of the ship the following speed appears at 10 knots, the data are mainly concentrated in 6-12 knots, and the data distribution is normal. The peak of the ship following spacing appears at 1200 m, the data are mainly concentrated in 500-3000 m, and the peak of the ship following time distance appears at 250 s. The data are mainly concentrated in 100-700 s, and both distribution characteristics show the same trend and left skew.

TABLE III: Descriptive Statistics of Effective Follow-up Data

	Speed (Kn/h)	Gap (m)	Time-Gap (s)	R-speed (kn/h）
Average value	9.41	2110	450	0.00
Variance	1.97	1886	279	1.75

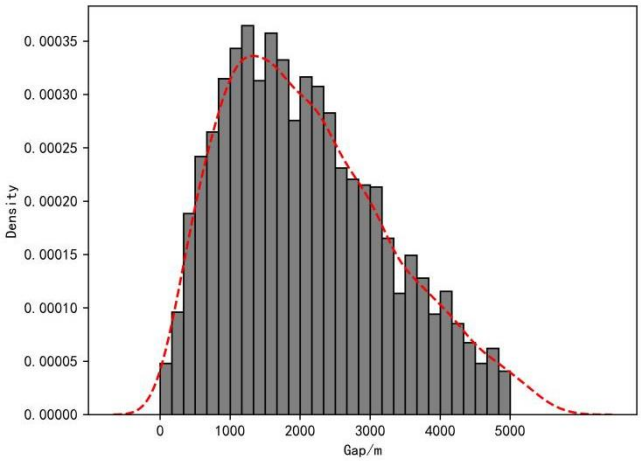

Fig. 4. Ship gap distribution map.

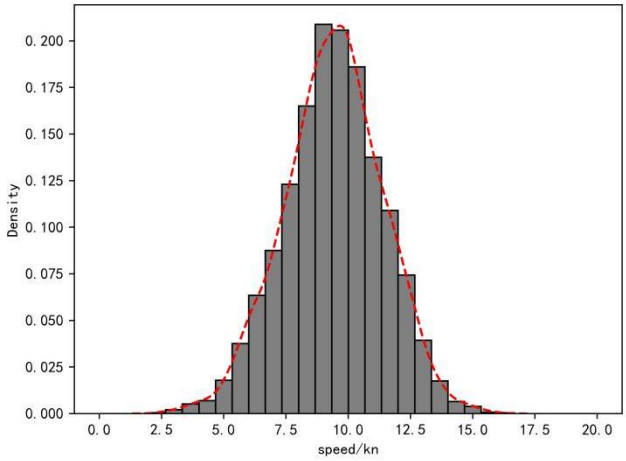

Fig. 5. Ship speed distribution chart.

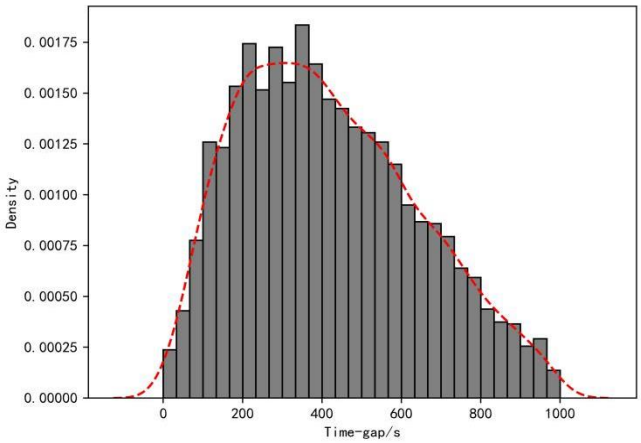

Fig. 6. Ship time distance distribution.

2) *The relationship between ship following speed and bow* spacing

The ship's speed is relatively stable with a small change in the following behavior, but it will also change with the change of bow spacing. As shown in the figure, when the following distance is 0-4000 m, the ship's following speed increases as the ship's following distance increases. When the following distance is larger than 4000 m, the average ship's following speed is rising, but the peak value is falling. This shows that when the following distance is greater than 4000 m, the ship pilot will drive according to the established speed of the ship and basically will not be affected by the previous ship.

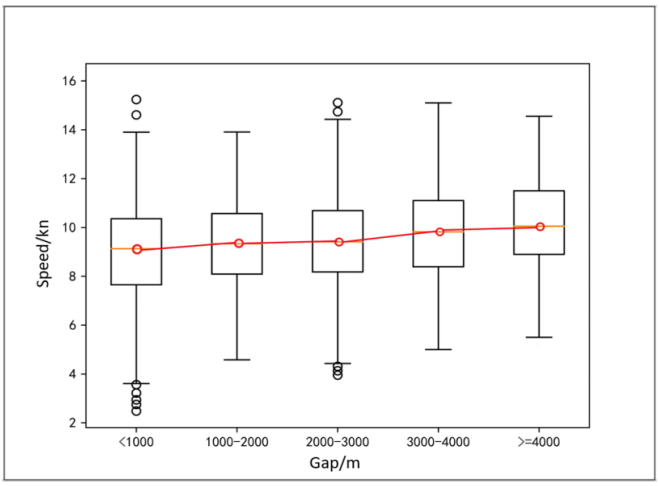

Fig. 7. Relationship between speed and gap.

3) Relationship between bow spacing and ship length

Therefore, the ship length is divided into four intervals: 0-50m, 50-100 m, 100-150 m, and 150-200 m, and the relationship between bow spacing and ship length can be further analyzed. 50 m, 50-100 m, 100-150 m, 150-20 m ship length average following bow spacing is 2062, 2073, 2132, 2296 m, as shown in the figure, the ship spacing becomes larger as the ship length becomes longer, but its change trend is weak.

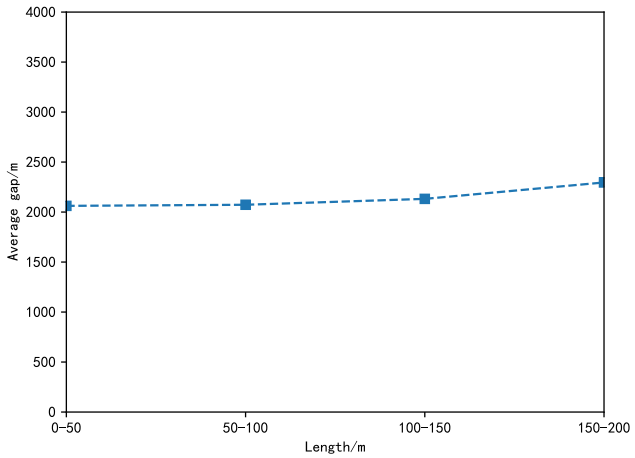

Fig. 8. Relationship between gap and length.

B. Importance of Factors Influencing Ship Following Behavior and Association Rule Mining

In the previous section, the relationship between bow spacing, ship speed, and ship length was analyzed, but this relationship was not obvious. Therefore, ship type, relative ship speed, ship density, and pilot class were introduced to analyze the factors affecting ship following behavior. And to determine the relationship between each factor, clarify the importance of each factor on the following behavior, and dig out the strong correlation rule between each factor and ship following spacing.

The main influencing factors of the ship's following behavior are ship to ground speed, the relative speed between the ship and the previous ship, ship type, ship length, ship density, and driver class, as shown in Table IV. Among them, ship type and ship length represent ship maneuvering performance (acceleration and deceleration performance) to a certain extent. Ship speed and relative speed can reflect the emergency braking distance of the ship. Ship density can reflect the busy degree of the ship in the channel, and the driver level can roughly reflect the driver's reaction time.

TABLE IV: THE MAIN FACTORS AFFECTING THE SHIP'S FOLLOWING DISTANCE

Factors	Symbols	Unit	Symbols
Speed	*V*	kn/h	C1
Length	*L*	m	C2
Relative speed	*RV*	Kn/h	C3
Type	*T*	dimensionless	C4
Density	K	Vessel/kn	C5
Pilot class	G	dimensionless	C6

1) Discretization of influencing factors of ship following behavior

Ship speed: The minimum speed is 2 knots and the maximum speed is 16 knots in the effective tracking data, divided according to the approximate frequency of the speed. 1=[2,6], 2=[6,8), 3=[8,10), 4=[10,12), 5=[12,16).

Ship length: The shortest 29m and the longest 200 m of the effective following data, divided by equal frequency interval. 1=[0,80), 2=[80,120), 3=[120,200].

Relative speed: 98% of the relative speeds of ships and previous ships are concentrated between [-4,4], so the relative speeds are divided into equal intervals of 2 knots. 1=[-4,-2), 2=[-2,0), 3=[0,2), 4=[2,4).

Ship type: There are 15 ship types in total in the statistical area, which can be divided into six sub-categories according to ship functions. 1=[cargo ship], 2=[container ship], 3=[petrochemical ship], 4=[engineering service ship], 5=[passenger ship], 6=[fishing ship].

Ship density: According to the size of ship density, the minimum density of statistical area is 0.33 ships/nautical mile, and the maximum density is 3.67/nautical mile, so the ship density is divided into equal intervals with 1 ship/nautical mile as the unit. 1=[0,1), 2=[1,2), 3=[2,3), 4=[3,4).

Driver class: 1=[third mate], 2=[second mate], 3=[first mate].

Vessel following spacing: Follow spacing in equal intervals of 1000m. 1=[0,1000), 2=[1000,2000), 3=[2000,3000), 4=[3000,4000), 5=[>=4000].

2) Correlation analysis of factors influencing ship following behavior

The correlation between the ship speed, ship length, ship type, relative speed, ship density, and a pilot class of the discrete effective following data is applied to the formula between the two, and their mutual information values are obtained to reflect the correlation between them, and the results are shown in Table V.

TABLE V: CORRELATION OF FACTORS INFLUENCING SHIP FOLLOWING

Factor	C1	C2	C3	C4	C5	C6
C1	1	0.0007	0.0224	0.0718	0.0049	0.0018
C2	0.0007	1	0.0013	0.0003	0.0058	0.0123
C3	0.0224	0.0013	1	0.0176	0.0028	0.0021
C4	0.0718	0.0003	0.0176	1	0.0019	0.0005
C5	0.0049	0.0058	0.0028	0.0019	1	0.0001
C6	0.0018	0.0123	0.0021	0.0005	0.0001	1

As it can be seen in Table V, the factors have almost no influence on each other, except for the correlation between ship speed and ship type.

3) Importance of factors influencing ship following behavior

Firstly, construct the information system $S=(U,A,V,f)$, where $A=C\bigcup D,\ C\bigcap D=\Phi$. The attributes in C are called conditional attributes and those in D are called decision attributes. The factors affecting the ship following behavior correspond to the conditional attributes, and the ship following distance corresponds to the decision attributes. Where C1 represents ship speed, C2 represents ship length, C3 represents ship type, C4 represents ship relative speed, C5 represents ship density, C6 represents pilot class, and D represents ship following distance.

Classifying the thesis domain according to the conditional and decision attributes separately yields.

$$U/IND(C)=\left\{u\middle|x\in U,x\in u_{c_i},u_{c_i}\in U/c_i,u=u_{c_1}\bigcap u_{c_2}\bigcap\cdots\bigcap u_{c_7}\right\}$$
$$c_i\in C,i=1,2,\cdots,7 \tag{8}$$

$$POS_C(D)=\{u_c\mid u_c\in U/C,u_d\in U/D,u_c\in u_{\mathrm{d}}\} \tag{9}$$

$POS_{C\text{-}\{c\}}(D)$ calculated in the same way as $POS_C(D)$.

Calculate the importance of each conditional attribute with respect to the decision attribute.

$$\gamma_C(D)=\frac{\left|POS_C(D)\right|}{U} \tag{10}$$

$$\sigma_D(C_i)=\gamma_C(D)-\gamma_{C-\{C_i\}}(D),i=1,2,3,4,5,6 \tag{11}$$

The importance of each condition attribute is shown in the following table.

TABLE VI: IMPORTANCE OF EACH INFLUENCING FACTOR OF SHIP FOLLOWING BEHAVIOR

	C1	C2	C3	C4	C5	C6
Importance	0.1321	0.0566	0.1415	0.1357	0.1253	0.0851

After normalization, the weights of each influencing factor of ship following behavior are 0.1953, 0.0836, 0.2092, 0.2265, 0.1852 and 0.1058 respectively. It can be seen that, among the influencing factors of the ship following behavior, the relative speed of two ships, ship type, speed, and ship density are more important, and the ship length and pilot class have less influence on the maintained following distance. The influence of vessel length and pilot class on the distance of following is relatively small. When building the ship intelligent following model, the relative speed and speed should be considered, and the ship type and ship density should be coefficient to ensure a more accurate model.

Then the attribute simplification is performed by applying the rough set attribute simplification algorithm to the above decision table. And the simplified attribute results are used in the subsequent association rule mining to improve the computational efficiency and rule accuracy. The results of the simplified attributes are shown in Table VI.

TABLE VII: RESULTS OF SIMPLIFIED ATTRIBUTES

Minimalist attributes	Support	Length
{speed, length, type, r_speed, k, grade}	100	6
{length, type, r_speed, k, grade}	100	5
{speed, length, r_speed, k, grade}	100	5
{speed, type, r_speed, k, grade}	100	5
{speed, length, type, k, grade}	100	5
{speed, length, type, r_speed, grade}	100	5
{speed, length, type, r_speed, k}	100	5
{type, r_speed, k, grade}	100	4
{speed, length, r_speed, k}	100	4
{length, r_speed, k, grade}	100	4
{length, type, r_speed, grade}	100	4
{speed, length, type, k}	100	4
{speed, length, r_speed, grade}	100	4
{speed, length, type, r_speed}	100	4
{speed, length, type, grade}	100	4
{length, type, k, grade}	100	4
{length, type, r_speed, k}	100	4
{speed, type, k, grade}	100	4
{speed, type, r_speed, grade}	100	4
{speed, r_speed, k, grade}	100	4
{speed, type, r_speed, k}	100	4
{speed, length, k, grade}	100	4
{speed, type, grade}	100	3
{speed, length, type}	100	3
{length, type, grade}	100	3
{speed, type, r_speed}	100	3
{speed, r_speed, k}	100	3
{speed, type, k}	100	3
{speed, length, r_speed}	100	3
{type, k, grade}	100	3
{type, r_speed, grade}	100	3
{length, type, r_speed}	100	3
{type, r_speed, k}	100	3
{length, type, k}	100	3
{speed, length, k}	100	3
{length, r_speed, grade}	100	3
{speed, r_speed, grade}	100	3
{speed, k, grade}	100	3
{length, type}	100	2
{type, r_speed}	100	2
{speed, r_speed}	100	2
{speed, type}	100	2

The minimum support, confidence and coverage of the association rules are set to 8%, 50%, and 10%, respectively. The association rules between the ship following spacing and each influencing factor are shown in Table VII.

The ship length does not appear in the probability rule set, but the density appears in the rule. Therefore, the density should be considered in the subsequent analysis of the heeling spacing.

TABLE VIII: PROBABILISTIC RULE SET

Rule	Probabilistic rule sets	Support	Confidence	Coverage
1	k=3→space=2	10%	62%	10%
2	speed=2∧grade=2∧type=1→space=2	8%	60%	10%
3	k=3∧type=1→space=2	9%	58%	10%
4	Grade=1∧r_speed=3∧k=2∧type=1→space=2	8%	57%	10%

VI. Conclusion

In this paper, an effective ship following behavior data extraction algorithm is designed and 3426 valid ship following behavior data are extracted from a large amount of AIS data. It offers theoretical support for the study of tracking behavior in marine traffic.

Then, the relationship between various influencing factors and the following distance was analyzed by the extracted effective following behavior data, and it was found that: when the following distance was 0-4000 m, the following speed increased with the following distance, and when the following distance exceeded 4000 m, the average following speed was increasing, but the peak speed was decreasing. This relationship between speed and space of vessels may help the department of marine administrator to control a reasonable speed threshold to make the safety and efficiency for the waterway transportation on the channel.

Finally, the weights of each influencing factor of ship speed, length, relative speed, type, density, and pilot class are determined as 0.1953, 0.0836, 0.2092, 0.2265, 0.1852 and 0.1058 respectively, and the strong correlation law between each influencing factor of ship following behavior and space were extracted to clarify the key factors of ship following behavior. It helps to establish and improve the data-based maritime ship following model, solve some problem in the autonomous navigation of ships, and promote the development of MASS.

After extracting the important factors and strong correlation rules of the ship following behavior, which focuses on these factors closely and take account into the differences between waterways and lanes, ships and vehicles, the mode of decision making of ships and vehicles, which will make the ships following model more suitable in the water traffic environment and promote the development of MASS.

Acknowledgment

The authors are grateful to the China Maritime Bureau for providing the AIS data support and to the Shanghai Commission of Science and Technology Project (Grant numbers 21DZ1201004) for financial support.

Funding

This work was partially supported by grants from National Natural Science Foundation of China (Grant numbers 51509151), Shandong Province Key Research and Development Project (SPKR \& DP) (Grant numbers 2019JZZY020713), Shanghai Commission of Science and Technology Project (Grant numbers 21DZ1201004), Anhui Provincial Department of Transportation Project (Grant numbers 2021-KJQD-011).

Conflict of Interest

Author Yihua Liu receives a salary from Shanghai Maritime University. Shanghai Maritime University where he is the Associate Professor. Author Bo Tu, Shiyu Tu, Xinyue Li, are master of Shanghai Maritime University, who are students mentored and sponsored by Yihua Liu.

References

[1] Reuschel A. Vehicle Movements in a platoon with uniform acceleration or deceleration of the Lead vehicle. *Ocsterrich Ingr Arch*, 1950;4:193-215.
[2] Pipes L A. An Operational Analysis of traffic dynamics. *Journal of Applied Physics*, 1953;24(3):274-281.
[3] Kometani E, Sasaki T. Dynamic behaviour of traffic with a nonlinear spacing-speed relationship. Elsevier. *Proceedings of the Symposium on Theory of Traffic Flow*. New York: Elsevier, 1959: 105-119.
[4] Michaels RM. Perceptual factors in car following. JOYCE A. *Proceedings of the Second International Symposium on the Theory of Traffic Flow*. Paris: OECD, 1963: 44-59.
[5] Chandler RE, Montroll H. Traffic Dynamics: Studies in Car Following. *Operations Research*, 1958;6(2):165-184.
[6] Bando, Hasebe, Nakayama, *et al.* Dynamical model of traffic congestion and numerical simulation. *Physical review. E, Statistical physics, plasmas, fluids, and related interdisciplinary topics*, 1995: 1035-1042.
[7] Treiber M, Hennecke A, Helbing D. Congested traffic states in empirical observations and microscopic simulations. *Physical Review E*, 2000;62:1805-1824.
[8] Zhu Jun, Zhang Wei. Calculation model of inland waterway transit capacity based on ship-following theory. *Journal of Traffic and T ransportation Engineering*, 2009,9(05):83-87.
[9] He Liangde, Jinag Ye, Yin Zhaojin, Zhou Bo, Tang Hui. Following distance model of inland ship. *Journal of Traffic and Transportation Engineering*, 2012;12(01):55-62.
[10] Ming li, Liu Jingxian, Wang Xianfeng. Calculation model of safe longitudinal distance for very large vessels. *Navigation of China*, 2014;37(04):40-43.
[11] Li Zhen-fu, Sun Yue, Wei Bo-wen. Car following model for navigation safety on "Polar Silk Road". *Journal of Dalian Maritime University*, 2018;44(03):22-27.
[12] Su Yonghua, Liu Kewei, Zhang Jinhua. Fuzzy evaluation of collapse incidents in highway tunnel construction based on rough set and Barycenter theory. *Journal of Hunan University (Natural Sciences)*, 2013;40(01):21-26.
[13] Zhang Xiaoming. Study on evaluation index weight of equipment manufacturing enterprises innovation capability based on rough set and AHM. *China Soft Science Magazine*, 2014;06:151-158.
[14] Zhou Danchen. A method for ascertaining the weight of attributes based on granular computing. *CAAI Transactions on Intelligent Systems*, 2015;10(02):273-280.
[15] Zhou Dan, Zheng Zhongyi. Relationship between the ship domain and its influence factors. *Journal of Dalian Maritime University*, 2016;42(03):25-30.
[16] Shi Zhenquan, Chen Shiping. A new method of knowledge characteristics weighting based on rough set and knowledge granulation. *Science and Technology Management Research*, 2018;38(12):248-253.
[17] Wu Sen, Wang Ping, Xu Mengguo, Cheng Aiping, Li Xingdong. Weight analysis of ground pressure developing factors in stope based on rough set theory. *Industrial Minerals & Processing*, 2019;48(02):1-4.
[18] Zhang Tingping, Wan Di. Evaluation Algorithm of Rail Transit Node Importance Based on Rough Set Theory. *Journal of Jilin University* (Engineering and Technology Edition), 2022-04-04:1-8.
[19] Shi Fuqian, Sun Shouqian, Xu Jiang. Association Rule Mining of Kansei Knowledge Based on Rough Set. *Computer Integrated Manufacturing Systems*, 2008;02:407-411+416.
[20] Zhang Yong, Yang Zhiyong. Rule acquisition for web logs based on rough set. *Computer Engineering*, 2006;20:84-85+146.
[21] Zheng Xiaofeng, Wang Shu. Data mining method of road transportation management information based on rough set and association rule. *Journal of South China University of Technology* (Natural Science Edition), 2014;42(02):132-138.
[22] Wang Ning, Liu Haiyuan, Zhou Xueke. Probability Rule Mining in Emergency Cases Based on Rough Set. *Operations Resea Rchand Management Science*, 2018;27(12):84-94.