Ship classification based on ship behavior clustering from AIS data

Zhou, Yang; Daamen, Winnie; Vellinga, Tiedo; Hoogendoorn, Serge P.

DOI
10.1016/j.oceaneng.2019.02.005

Publication date
2019

Document Version
Accepted author manuscript

Published in
Ocean Engineering

Ship Classification based on Ship Behavior Clustering from AIS Data

Yang Zhou[1*], Winnie Daamen[2], Tiedo Vellinga[1], Serge P. Hoogendoorn[2]

[1] (*Department of Hydraulic Engineering, Faculty of Civil Engineering and Geosciences, Delft University of Technology, Netherlands*)

[2] (*Department of Transport and Planning, Faculty of Civil Engineering and Geosciences, Delft University of Technology, Netherlands*)

Abstract

Since the introduction of the Automatic Identification System (AIS), AIS data has proven to be a valuable source of ship behavior analysis using data mining. It records ship position, speed and other behavior attributes at specific time intervals in all voyages at sea and in ports. However, the current studies in ship behavior analyze the behavior patterns either with a subjective choice of classification for behavior differences among the groups of ships or without any classification at all. In order to fill this gap, a new methodology for ship classification in ports based on behavior clustering is developed by analyzing AIS data from the port of Rotterdam. Besides a proper data preparation, the proposed methodology consists of two steps: step I, clustering ship behavior in a port area and identifying the characteristics of the clusters; step II, classifying ships to such behavior clusters based on the ship characteristics. The clustering results present both the behavior patterns and the behavior change patterns for ship path and speed over ground, which are the dominant behavior attributes for ships in ports. Some patterns of integral ship behavior can also be revealed by investigating the correlation between the two behavior attributes. Our research has shown that length and beam can be adopted as explanatory variable to classify ships to the corresponding behavior clusters. The classifiers are developed based on both unsupervised discretization (equal width binning) and supervised discretization (Chi2). The performances of classifiers are compared by three evaluation metrics, including Average Accuracy, F_1score , and AUC. We found that the classification based on multi-criteria is more accurate than using a single criterion. The classifications based on Chi2 discretization

* Corresponding author:
E-mail address: Y.Zhou-5@tudelft.nl; y.zhou_navi@outlook.com
Tel.: +31 (0) 15 27 82520

outperform the ones with equal width discretization. The outcome leads to a systematic understanding of ship behavior in a port area and can be used to predict the ship behavior pattern based on their characteristics and simulate the ship behavior.

Keywords: Data mining; AIS data; Behavior clustering; Ship classification; Ports and waterways

1. Introduction

Waterborne transport has become an increasingly important means of international freight transport. Due to a large amount of cargo carried by individual ships and the high frequency of ships visiting the hub ports, the safety of ships and the capacity of ports have been global challenges with high priority for nautical traffic management and port authorities. Both efficient traffic management and predictive port design require a systematic and thorough understanding of ship behavior in port and inland waterways. For individual ships, the behavior is always different due to the officers on board and the different sailing situations. However, it is assumed to exist some behavior patterns for the total ship traffic in an area. The ship behavior patterns in an area are revealed by clusters of similar ship behavior, while the clusters are distinctive with each other over the area. For the macroscopic ship traffic flow, the behavior patterns show the characteristics of the ship behavior in the area. For the individual ships, the behavior patterns indicate the range that the ships will behave. accurate behavior pattern prediction will support the port operation and the maritime surveillance. As only static ships characteristics are known before the ships enter a port area, a relation between these ship characteristics and the behavior patterns needs to be found. Therefore, ships are classified according to such patterns based on the ship characteristics.

Currently, in the field of ship behavior analysis, Automatic Identification System (AIS) data has proven to be a valuable source of big data. Many researchers have used AIS data for general behavior pattern recognition and anomaly detection, without a clear classification of the involved ships (Gunnar Aarsæther and Moan, 2009; Pallotta et al., 2013; Ristic et al., 2008). In these studies, the behavior differences of individual ships are ignored. However, such individual differences have been revealed in other studies, in which a ship classification is pre-defined before analyzing the ship behavior using

AIS data. Silveira et al. (2013), Goerlandt and Kujala (2011), and Mascaro et al. (2010) classify ships based on their type, while not considering the size of ships. For the purpose of ship behavior investigation, this classification method lacks a detailed description of the behavior differences within the same ship type. De Boer (2010) proves behavior differences of container ships with different deadweight tonnages (DWT). However, the classification thresholds are determined such that approximately the same amount of available trajectories is included in every data set. This classification method only explains the behavior differences when comparing groups of ships, without a clear recognition of the actual behavior patterns in the area. The results of behavior comparison depend on the choice of classification criterion. The same drawback holds for the studies of Shu et al. (2013) and Xiao et al. (2015), which classify ships according to gross tonnage (GT). Moreover, the GT is a measure of ship's overall internal volume without information on ship's size which seems to be relevant to the behavior. Both classifications are based on the presence frequencies of ships with different GT in the data set. In the study for ship behavior during collision avoidance based on AIS data by Mou et al. (2010), the overall length is selected as the criterion to distinguish ship size and investigate the correlation between ship length and the closest point of approach (CPA). However, the reason for choosing such a criterion is not explicitly explained. Thus, in the studies of predefined classification for ship behavior, the choice of the classification criterion would result in a subjective explanation of the ship behavior patterns. It would be better to recognize the patterns of ship behavior directly from the behavior clusters, which then form the basis of ship classification.

To the best of the authors' knowledge, there is no dedicated research on the methodology of ship classification based on their behavior in a port or waterway. In our preliminary research on ship classification methodology (Zhou et al., 2015), ships are classified based on the relation between ship characteristics and behavior when passing a specific cross-section. However, the proposed method cannot be applied in a waterway or port area since it cannot handle the behavior in time series. The behavior patterns are not recognized, either. In a recent study, ship trajectories in a coastal area have been classified to predict ship type (Sheng et al., 2018). However, the classification only considers two types of ships (cargo ships and fishing ships) without other ship characteristics, and the method cannot be applied in a port area since detailed paths and speeds are not analyzed. In real-life port operations or

maritime traffic management, the general ship classification in a port, a water area, or a country is subjectively determined by the local port authority or other equivalent parties for the purpose of port dues calculation or ship registration. Ship classifications based on guidelines or rules will not support the accurate prediction of port operations with respect to maritime traffic.

The objective of this paper is to develop a ship classification method based on ship behaviors revealed by AIS data, using static ship characteristics as explanatory variables. The first contribution is to develop a methodology to cluster ship behavior using AIS data. The resulting ship behavior patterns form the classification basis. The second contribution is the development of a ship classifier able to predict the ship behavior based on the static ship characteristics. To apply clustering and classification techniques in the domain of maritime traffic, existent methods in data mining will be modified and improved according to the characteristics of the data. The research result will support the behavior pattern recognition from AIS data and simulate the behavior in a systematic way considering the differences of ship characteristics. From the practical application perspective, the port authority would be able to predict the behavior pattern based on the available ship characteristics from AIS data or Vessel Traffic Services (VTS) report when a ship is about to approach. In case of special circumstances, the prediction can provide theoretical references for traffic control measures regarding ship behavior suggestions (e.g. position control or speed limit). In this paper, the data sets are introduced in Section 2 to give an overview of the characteristics of maritime traffic in ports. Section 3 explains the proposed methodology for behavior clustering and ship classification. The clustering results and the classifier performance are presented and discussed in Section 4. Finally, Section 5 concludes the paper with recommendations for further research.

2. Data description

The study area is a nearly straight waterway, Nieuwe Waterweg, located at the entrance of the port of Rotterdam, the Netherlands, as shown in Figure 1. The reason for choosing a straight waterway for behavior clustering is to eliminate the impact of a specific waterway layout on ship behavior. Thus, the proposed methodology is expected to be generically applicable in other (straight) waterways. This study focuses on ship behavior in unhindered situations, as it is assumed that ships with the same

behavior pattern in an unhindered situation would behave similarly when under the impacts of external factors, such as encounters, wind, and current. Based on such behavior clustering and ship classification, the behavior in other waterway layout or hindered situation can be systematically analyzed. The length of the study area is 2300 m, and its width is about 650 m. For the inbound traffic (sailing from the sea towards the port of Rotterdam), the Maasgeul channel (see Figure 1) splits into Nieuwe Waterweg and Calandkanaal, which are physically separated by a slightly bent mole, named the Splitsingsdam.

The X-Y coordinate system in Figure 1 corresponds to the Dutch geographical coordinate system, the Rijksdriehoeksmeting (RD system), with the origin located in the southwest of the study area. In order to facilitate the analysis of ship behavior, the coordinate system is transposed to a local reference system. The transposed origin lies at the bottom left corner of the study area, which corresponds to the west end of the Splitsingsdam.

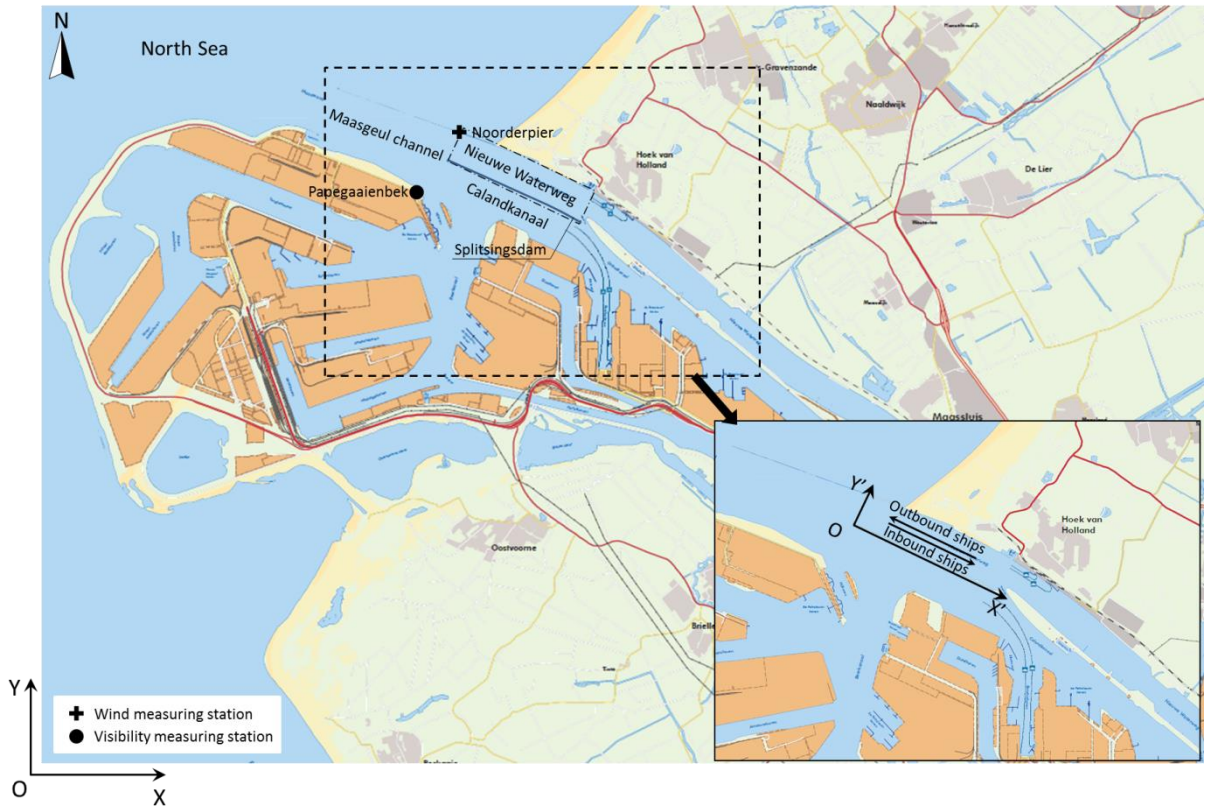

Figure 1. Location of the study area and the wind and visibility measuring stations in the port of Rotterdam. The X-Y coordinate system is RD system. In the cutout area, the transposed system is indicated, so the inbound ships sail in the X'-direction, while the lateral deviations from the straight path are visible in Y'-direction.

The data for the whole year of 2014 have been collected from the port authority of Rotterdam, including the ship behavior records and the data of external environmental factors. The raw AIS data

reveal the ship behavior in the study area, while the meteorological and hydrological data present the external conditions, including visibility, wind, and current. In this section, the collected data are introduced.

2.1. AIS data

The AIS system is an automated tracking system onboard ships, which is designed to automatically provide information about the ship to other ships and to coastal authorities. In 2000, International Maritime Organization (IMO) issued an amendment adopting a new requirement regarding the introduction of AIS system in the International Convention for the Safety of Life at Sea (SOLAS) (IMO, 1974). The AIS system is mandatory by the end of 2004 for all ships of 300 Gross Tonnage (GT) and more engaged on international voyages, cargo ships of 500 GT and more not engaged in international voyages and all passenger ships irrespective of size. In the study area, every seagoing ship, even below the GT limit of IMO regulation, has installed AIS equipment and used it in all voyages. For the inland ships, both commercial and recreational ships, and sailing vessels longer than 20 meters are mandatory to use AIS since December 1st, 2014 according to the resolution of the Central Commission for the Navigation of the Rhine. The regulation applies to most of the inland vessels in the Netherlands. Since the year of 2014 is still a transition period, the majority of the collected AIS data of 2014 are seagoing ships.

According to the guidelines by IMO (2003), the AIS data contain three types of information: (1) static information (Maritime Mobile Service Identity (MMSI) number, IMO vessel number, ship name, radio call sign, ship type, overall length, beam, etc.); (2) dynamic information (UTC time, ship position, speed over ground (SOG), course over ground (COG), heading, navigational status, etc.); (3) voyage-related information (draught, destination, etc.). The static information is entered into the AIS system by the equipment provider when the equipment is initially installed or after a major change of the ship structure. The dynamic information is updated automatically based on the sensor data. The update time interval of dynamic information depends on the speed of the ship according to the regulation by the International Telecommunication Union (ITU, 2014). The time interval is short when the SOG is high, and vice versa. The voyage-related information should be manually updated to the

real-time situation by the officers onboard. The actual draught may indicate the loading condition of the ship, which affects the ship's maneuverability. However, in the collected data set, errors are found in the voyage-related information. For most of the ships in the data set, the draught is not updated in each voyage. For some ships, the value of draught equals the molded draught in the registration. Other ships are recorded with a draught of 0 meter in the data set. This implies that the data of ship draught are not reliable, thus these are not analyzed.

Since the port authority of Rotterdam only stores the mandatory fields of static information in AIS data, the ships characterization is limited to type, length, and beam. In the collected AIS data set (2,299,842 messages), numerous ship types occur, including cargo ships, tankers, passenger ships, pilot ships, tugs, and dredgers. For the specific purpose ships (pilot ships, tugs, and dredgers), their working status is not indicated in the AIS messages. However, their behaviors in working and non-working states are different, and as such, these ships are excluded from this research. Thus, only three main types of ships are analyzed in this research, being cargo ships (993,566 messages), tankers (522,614 messages) and passenger ships (77,724 messages). Since the cargo ships are not further categorized (e.g., into container, general cargo ship or bulker), the ship type is not included as a characteristic to classify the ships, even though it is assumed that the behavior of these types of cargo ships vary considerably. Therefore, the static ship characteristics to classify ships are length and beam in this research. The ratio between length and beam is also considered as a characteristic, as it is one of the ship dimension ratios indicating ship maneuverability. The ratio is calculated with an accuracy of 0.1 considering the rules of significant figures in division calculation (Harris and Stöcker, 1998), since the length and beam are registered with an accuracy of 0.1 meter in AIS data.

In the AIS data, the dynamic ship behavior is recorded through four behavior attributes (see Figure 2): position, heading, COG, and SOG. Since the study area is a straight waterway, the COGs of the ships mostly follow the direction of the waterway. The headings of the ships change seldom, either. Thus, in this research, the ship position and SOG are the main behavior attributes to be analyzed and clustered. Based on the transposed coordinate system (see Figure 1), the ship position is represented by the distance to the Y'-axis (the northwest boundary of the study area) and the lateral distance to the X'-axis (the southwest boundary). Since the Splitsingsdam is a slightly bent mole, the lateral distance

to the X'-axis does not exactly correspond to the distance to the dam.

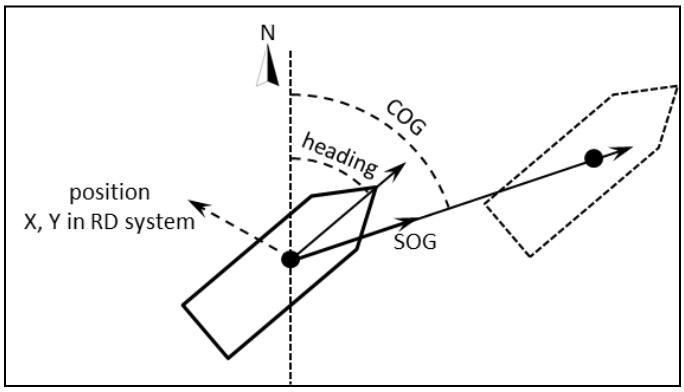

Figure 2. Illustration of behavior attributes in AIS data (RD system is the geographical coordinate system in the Netherlands).The ship with a solid outline is the current position, while the ship with a dotted outline is the position at next time step. The position of the ship is represented as the center in the figure, but the exact position for each ship depends on the location of transmitter installed on board.

2.2. Meteorological and hydrological data

Since the ship behavior is highly influenced by the external environmental conditions (Shu et al., 2017), meteorological and hydrological data are collected to describe the external conditions of ship behavior.

The meteorological condition refers to wind and visibility. Both are measured during the same period as for which the AIS data have been collected (2014). The location of the measuring stations are presented in Figure 1. The measured wind velocity data are stored at an interval of 5 minutes, while the visibility is presented every minute. In non-extreme weather conditions, there is no sudden change of wind within 5 minutes. Thus, the data are reliable and sufficiently accurate in presenting the external conditions. Since there are no obstructions in the study area, the wind and visibility are deemed to be the same for the whole area.

The hydrological condition refers to the current, in particular, the current velocity. Unlike wind and visibility, the measured current velocity at a specific measuring station is not representative for the whole area, due to the propagation of flow and the velocity variation over the water-depth. Thus, the data of current velocity are calculated by the port authority using the SIMONA model (Vollebregt et al., 2003) using the measured water level from eight stations around the port as input. The modeled velocity has been validated by comparing to the measured velocity at one station in the area. The collected data describe the current velocity in 41×7 orthogonal curvilinear grids with a resolution of about 85 meters. The current velocity in each grid cell is presented by 10 layers with the same depth

averaged by the water depth of the grid at an interval of 15 minutes. For most of the ships, the length is larger than 85 meters, so the grid resolution is sufficiently accurate. The data represent the current situation along the voyage of ships.

3. Research methodology

The goal of this research consists of two parts, being distinguishing behavior clusters and classifying ships. Since the data mining in this research involves multi data sources, data preparation is necessary. The flow diagram in Figure 3 illustrates the three steps of the research methodology, which are further explained in this section.

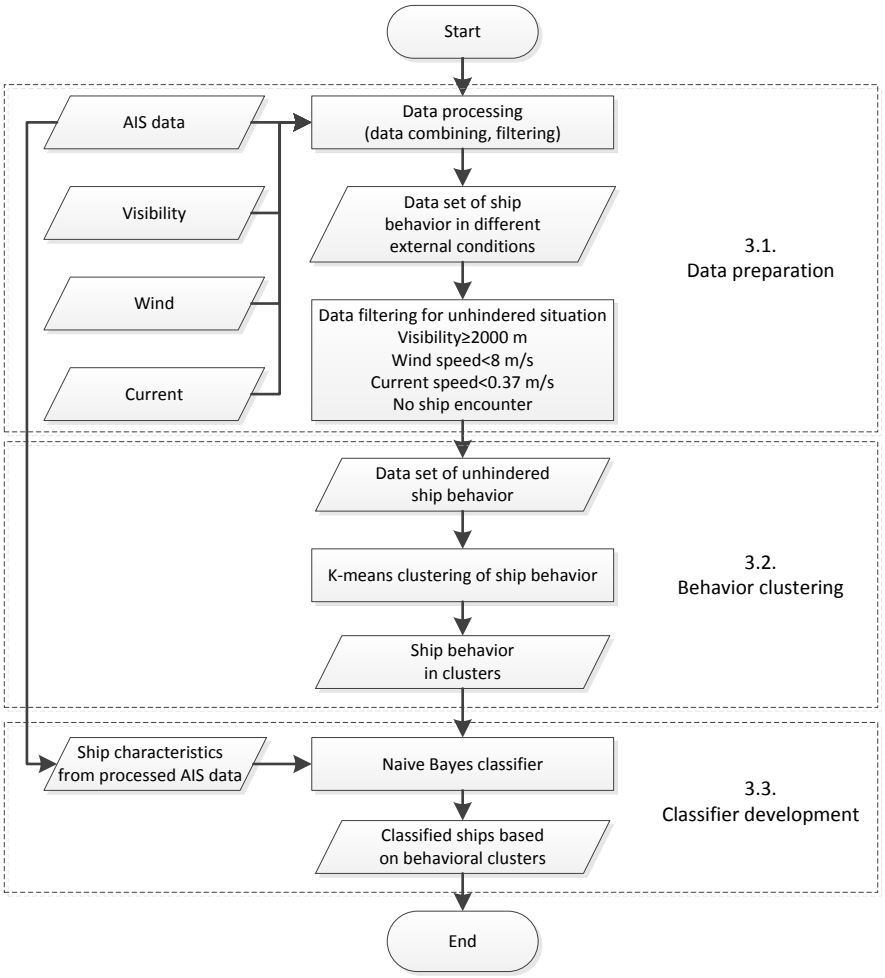

Figure 3. Flow diagram of the ship classification based on ship behavior clustering.

3.1. Data preparation

Since ship behavior is influenced by the external conditions (Shu et al., 2017), the clustering of ship behavior is assumed to be based on the unhindered situation. This way, the impacts of external

factors are eliminated, which could reveal the behavior patterns mainly due to the ship characteristics. In the unhindered situation, the bridge team is assumed to behave with good seamanship, which is mostly based on the ship characteristics and maneuverability. The clustered ships are also expected to behave similarly when they are in some specific hindered situation. Therefore, multiple data sources are integrated.

In the data processing of this research, the raw AIS data are integrated with the data of visibility, wind, and current based on the time and ship position in each AIS message. To eliminate the impacts of environmental factors, thresholds are used to filter out the unhindered situation. The thresholds have been previously analyzed in the impact analysis of external factors using the same data set (Zhou et al., 2017). An unhindered situation is characterized by visibility $\geq$ 2000 m, wind speed $<$ 8 m/s, and current speed $<$ 0.37 m/s. Besides, in order to exclude the impact of ship encounters, the processed data set has filtered the trajectories of ships with any encounter with another ship in the study area. The three main types of ship encounter identified in the International Regulations for Preventing Collisions at Sea (COLREGs) are taken into account: head-on, overtaking, and crossing. The encounters of overtaking and crossing can be easily distinguished by the speed and position changes. The head-on situation at sea is defined when one ship is coming towards the other one roughly within 6 degrees on either side of the heading. For the bi-directional waterway in port area, any two ships sailing in opposite directions will be deemed as head-on situation, since the inland waterway is narrow. When the two ships sail close to each other, they may change their course to avoid collision risk. Even without the collision risk, both ships will sail as near to the outer limit of the waterway on her starboard side as is safe and practicable, according to the rule of narrow channels in COLREGs. Thus, when two ships pass by each other, such encounter will influence the behavior of both ships. The head-on situation is identified and filtered, when one ship encountering the other from the opposite direction in the study area. So far, the data set of ship behavior in the unhindered situation is generated.

The AIS messages are transmitted at different time intervals due to the different speeds of ships. In the collected AIS data, the duration of the time interval is between 6 seconds and 15 seconds. A set of cross-sections has been developed parallel to the Y'-axis in order to analyze the behavior pattern of all ships when passing the same cross-section. The distance between cross-sections is equal, and is

determined by calculating the proceeded distance of ships between two adjacent AIS messages. The distance should guarantee that there is at least one AIS message in between two adjacent cross-sections for most of the ships in order to reduce the inaccuracy introduced by interpolation. This results in a distance between cross-sections of 65 meters, with 35 cross-sections in total. The data of ship behavior attributes are linearly interpolated by the last message before and the first message after the cross-section. Considering the ships cannot make sudden changes in behavior in port areas due to the maneuverability and large inertia, linear interpolation of ship behavior within short distance will not decrease the data accuracy or influence the results.

3.2. Behavior clustering

The AIS data in unhindered situations are used to form classes for each behavior attribute using clustering techniques. Clustering analysis is an unsupervised technique in data mining. The data set without any pre-classification can be grouped into multiple clusters, so that the objects within a cluster have high similarity with each other, but are distinctive to the objects in other clusters (Han et al., 2011). Clustering methods can be divided into two groups: hierarchical and partitioning techniques (Saxena et al., 2017). In the hierarchical clustering methods, clusters are revealed by iteratively dividing the groups using a top-down method or forming the groups by a bottom-up approach. The result of such methods usually leads to a dendrogram among the data objects. As the behavior of individual ships is assumed to be independent, the hierarchical clustering method is not appropriate for this research. Therefore, the partitioning method is adopted, with the aim to assign the data objects into clusters without hierarchical structure by optimizing some criterion function. The criterion is usually expressed by the dissimilarity between each data object and the corresponding cluster center. In the clustering of ship behavior, the behavior attributes are all scalable and each ship can only be assigned to one cluster. Thus, the centroid-based portioning technique, *k*-means algorithm, satisfies the requirements in this research.

The general procedure of *k*-means clustering includes: step 1, choosing initial cluster centers for a given number of clusters; step 2, assigning each data object to the cluster with least dissimilarity to the cluster center; step 3, updating the cluster center after all objects being assigned and repeat step 2 until

there is no change in cluster center. The limitations of this method can be found through its procedure (Saxena et al., 2017): (i) strong reliance on the user to define the number of clusters in advance; (ii) high sensitivity to the initialization phase; (iii) high sensitivity to the outliers in the assigning process; (iv) high sensitivity to the definition of stopping criterion. In this research, the disadvantages are improved or overcame in the ship behavior clustering algorithm.

(i) decision on the number of clusters

Given the general ship behavior data set, the number of behavior clusters is unknown beforehand. Too few clusters will lead to an obscure recognition of behavior patterns, as some behavior patterns might be combined. Meanwhile, choosing too many clusters will lead to a lack of general representativeness of behavior patterns, and the clustering patterns are possibly indicated by statistical artificial differences.

To deal with the influence of the number of clusters, the *k*-means clustering method is performed using different numbers of clusters as input, starting with 2 clusters, and increasing the number of clusters until the ship behavior data in each cluster are significantly different with the data objects in other clusters over the whole area, which can be compared at all cross-sections. The statistical t-test is performed to compare the ship behavior patterns with a significance level of 95% (corresponding to a p-value of 0.05). This condition guarantees that the clustering results represent the ship behavior patterns in the whole study area. When the data in two clusters are not significantly different at some cross-sections, these two clusters cannot be deemed to have a different behavior pattern in the whole area.

(ii) the defined initialization phase

The common *k*-means clustering starts with an arbitrary choice of the centers of initial clusters. However, the random initialization would lead to different clustering results in different runs, where each run starts with different centers of initial clusters. In order to get a unique clustering result representing behavior patterns, the centers of initial clusters are defined to be distinct. As the ship behavior is assumed to be smooth without sudden changes due to its maneuverability, the initial centers can be calculated from the data objects. For any ship trajectory n, a general indicator of behavior attribute B_n is defined as the mean value of this behavior attribute on all cross-sections. Two

of the cluster centers are the minimum and maximum B_n, respectively. The other cluster centers are the trajectories with the corresponding percentile value of B_n. For instance, when the number of clusters is 4, the initial cluster centers are the minimum, 33rd percentile, 67th percentile and maximum of B_n. This way, the initial centers are unique and distinct to each other.

(iii) the measure of overall dissimilarity between clusters

In the *k*-means algorithm, every data object will be assigned to a cluster and influence the cluster center, which makes the algorithm sensitive to the outliers. Outliers in ship behavior in AIS data are mostly caused by the occasional measurement error during one message transmission. The clustering based on differences of behavior when passing a specific cross-section is sensitive to such data outliers. Besides, the cluster of ship behavior should represent the general behavior pattern over an area. Thus, to express such overall dissimilarity of behavior and overcome the sensitivity to the outliers at some point, the measure of distance between a data object to the cluster center is defined considering the ship behavior on all cross-sections. For any ship trajectory *n*, the distance to the center of cluster *i* $D_{(n,i)}$ is defined as

$$D_{(n,i)} = \sum_{j=1}^{m} \left(B_{n(j)} - B_{c(i,j)} \right)^2 \tag{1}$$

where $B_{n(j)}$ denotes the behavior attribute of trajectory *n* on cross-section *j*, $B_{c(i,j)}$ denotes the center of cluster *i* for this behavior attribute on cross-section *j*, and *m* is the total number of cross-sections.

The difference is squared for each cross-section to avoid difference compensation and to weigh the difference (larger differences have a larger effect). Based on the calculation of distance to all cluster centers, each ship trajectory will be assigned to the cluster with the minimum distance to the corresponding center.

(iv) the data-based stopping criterion

After each iteration of assigning ships to clusters, the centers are updated by calculating the mean value of the behavior attributes on each cross-section in each cluster. The clustering stops when the centers do not longer change. In the application to practical problems, a specific stopping criterion needs to be defined, which is usually indicated by the number of iteration steps. However, in the

clustering of ship behavior to recognize the behavior pattern in an area, such stopping criterion (no change of cluster centers or a definite number of iterations) lead to heavy computation load or incomplete clustering results. In this research, the stopping criteria are decided based on the significant figures of behavior data. When the maximum change of cluster centers is less than a threshold value, further clustering is no longer significant in practice. Then, the clustering repetition stops. In this research behavior, the threshold of position is set as 0.1 meter, while it is set to 0.01 knot for SOG.

When the clustering results satisfy the aforementioned stopping criteria, the formed clusters represent the general ship behavior patterns in the area. The clustering result for each ship trajectory is input as label for its behavior pattern in the process of ship classification, see the next section.

3.3. Classifier development

The clustering result reveals the behavior patterns and identifies clusters for each type of ship behavior. In this section, the proposed classifier(s) is identified to predict to which behavior cluster a ship belongs to, based on the ship characteristics. To discover the most appropriate criteria to classify ships using available data, four combinations of ship characteristics are selected to develop classifiers, and their performances need to be tested and compared. These combinations are: (1) ship length; (2) ship beam; (3) ratio between ship length and ship beam; (4) ship length and ship beam.

The data classification is a two-step process, consisting of a learning step to develop a classifier or a classification model and a classification step where the classification model is used to predict class labels for given data objects (Han et al., 2011). In the second step, the performance of the developed classifier can be tested. In this research, the holdout method with random subsampling is used to compare the performances of different classifiers. Two-third of the ship trajectories are allocated to the training set, while the remaining one-third of the trajectories are used in the test set. To avoid the impact of seasonality of port operation on maritime traffic, random subsampling is performed to the data per month over the year. This way, both training and test sets cover the whole year of collected data.

Five main categories of classification algorithms are distinguished (Kotsiantis et al., 2007): logic-based algorithms, perceptron-based techniques, statistical learning algorithms, instance-based learning,

and Support Vector Machines. In this research, instead of a direct causal relationship between ship characteristics and ship behavior, there is a possibility for any ship to belong to any behavior pattern. It means that the class of a data object is predicted by the highest possibility to belong to a single class among all classes. Besides, it is expected that more than two classes of ship behavior will be distinguished. Thus, in this research, the statistical learning algorithm underlying a probability model is adopted. Among the statistical learning algorithms, the naive Bayesian classifier has been found to be comparable in the performance with other neural network classifiers, and it has a high accuracy and short computational time when applied to large data sets (Han et al., 2011). As the effects of ship characteristics on the prediction of behavior class are assumed to be independent, it also follows the basic assumption in Naive Bayesian classification. Thus, the Naive Bayesian classifier is appropriate for this research. In the following, the two steps of classification using a Naive Bayesian classifier are further elaborated upon.

Step 1: learning step to develop classifier

Let any ship $S_n \in S$ be represented by its characteristics, $\mathbf{X} = \{x_1, x_2, ..., x_m\}$. Suppose that there are *k* behavior clusters (the results from ship behavior clustering), $C_1, C_2, ..., C_k$. The classifier will assign S_n to class *i* with the highest posterior probability, which is the maximum posteriori probability (MAP) presented as

$$C_{MAP} = \arg\max_{S_n \in S} P(C_i \,|\, \mathbf{X}) \tag{2}$$

The posterior probability is calculated according to Bayes' theorem:

$$P(C_i \,|\, \mathbf{X}) = \frac{P(\mathbf{X} | C_i) P(C_i)}{P(\mathbf{X})} \tag{3}$$

Since $P(\mathbf{X})$ is the same for all classes, only $P(\mathbf{X} | C_i) P(C_i)$ needs to be calculated and maximized. The prior probabilities of all behavior classes $P(C_i)$ are known based on the clustering results for each behavior attribute. Thus, only the prior probability of different characteristics should be estimated. For numerical characteristics, there are two alternatives to estimate the prior probability:

(i) For each ship characteristic, if a continuous distribution can be fitted based on the collected data, the prior probability can be computed using the distribution function.

(ii) If there is no fitted distribution for the ship characteristics, the data need to be discretized into bins to generate the prior probability from the training data set. The techniques of discretization can be categorized as supervised or unsupervised (Witten et al., 2016). Since neither method always yields better results than the other, both discretization methods have been tested in this research.

(a) *unsupervised discretization*

Among the unsupervised discretization methods, the equal-width binning and equal-frequency binning are the basic ones, while the discretization based on clustering is more sophisticated (Joiţa, 2010). Since the number of intervals for each ship characteristic is unknown, the discretization by clustering analysis, such as *k*-means discretization, cannot be applied. As the purpose of discretization in this paper is to calculate the prior probability, the equal-frequency interval binning is not appropriate. Thus, the equal-width binning is chosen as unsupervised discretization method.

In the equal-width interval binning, the data range $(a_1, a_2, ..., a_n)$ is divided into k intervals of an equal width determined by $(a_{\max} - a_{\min})/k$. This way, the interval is determined by the data objects and the desired number of intervals. However, as stated before, the number of intervals in this research is unknown. Thus, the interval width for each ship characteristic is given in the discretization instead of the number. Since an empty bin of ship characteristics in the training set means the prior probability equal to 0, the ships with such characteristics cannot be properly classified. To avoid empty bin and consider the values of different ship characteristics, the interval for length is determined as 10 meters, 2 meters for beam, and 0.5 for the ratio between length and beam.

(b) *supervised discretization*

Compared to the unsupervised discretization methods, the supervised methods make use of the class labels when partitioning the characteristics. Several discretization methods exist, and the supervised discretization methods in classification have been tested and compared (Lavangnananda and Chattanachot, 2017). They show that the Chi2 algorithm yielded the best performance in most datasets compared to other supervised methods. In the classifiers with Naive Bayes, Chi2 also yielded the best performance. Thus, in this research, Chi2 is adopted, which is a bottom-up discretization method using the χ^2 value to determine the merging point (Liu and Setiono, 1995).

The discretization process starts with sorting the values of a characteristic. Then, the following steps are performed: (1) each value forms one interval; (2) the χ^2 value for every pair of adjacent intervals are calculated according to equation 4; (3) the pair of adjacent intervals with the lowest χ^2 value are merged into one interval. Since the main behavior class in adjacent intervals of ship characteristics might be different, the merging only considering the χ^2 value may change the distribution of behavior classes within the interval. To avoid such unexpected variation, one more criterion for merging is required, being that the pair of adjacent intervals should be consistent in the main behavior class.

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{k}\frac{\left(A_{ij}-E_{ij}\right)^2}{E_{ij}} \tag{4}$$

where:

k is the number of classes;

A_{ij} is the number of data objects in the i^{th} interval, j^{th} class;

E_{ij} is the expected frequency of A_{ij}, $E_{ij} = R_i * C_j / N$;

R_i is the number of data objects in the i^{th} interval, $R_i = \sum_{j=1}^{k} A_{ij}$;

C_j is the number of data objects in the j^{th} class, $C_j = \sum_{i=1}^{2} A_{ij}$;

N is the total number of data objects, $N = \sum_{i=1}^{2} R_i = \sum_{j=1}^{k} C_j$.

Instead of setting a threshold of χ^2, Chi2 introduces the inconsistency rate. For the data objects with the same value of a characteristic, the inconsistency count is the total number of objects minus the largest number of objects with the same class label. The inconsistency rate is the sum of the inconsistency count divided by the total number of the objects in the interval. For example, there are n data objects with the same value of a characteristic, c_1 objects belong to class 1, c_2 objects to class 2, and c_3 objects to class 3, where $c_1 + c_2 + c_3 = n$. If c_1 is the largest among the three classes, the inconsistency count is $n - c_1$. In the original Chi2 algorithm, the merging process stops when the inconsistency rate exceeds a certain value δ. However, an appropriate value of δ can only be given

after some tests on the data set. The same value of δ for all intervals also ignores the different portion of inconsistent objects in different intervals. Thus, the value of δ is set as a dynamic criterion, which is the lowest initial inconsistency rate among the involved initial intervals. For instance, if the i^{th} interval is merged from the 4^{th}, 5^{th}, and 6^{th} initial intervals, the value of δ for the i^{th} interval is the minimum of the inconsistency rate among the three intervals at the initial step without any merging. This is still to avoid the change of the behavior class distribution after merging the intervals.

Since the supervised discretization considers the class labels, the discretization for the same ship characteristics for different ship behavior might be different. Thus, in the research, the discretization would perform 6 times, which are for 2 ship behavior attributes with 3 ship characteristics.

Step 2: classification step to measure performance

The performance of a classifier can be evaluated by comparing the predicted class labels by the classifier and the actual class labels by behavior clustering (Sokolova and Lapalme, 2009). The confusion matrix in this research is defined in a one-versus-others method based on the classical matrix for binary classification, as listed in Table 1.

Table 1. Confusion matrix for class *i* in ship classification.

Actual ship behavior class	Predicted as class *i*	Predicted as other classes
Class *i*	true positive (TP_i)	false negative (FN_i)
Other classes	false positive (FP_i)	true negative (TN_i)

Considering the characteristics of different performance measures for classification (Sokolova and Lapalme, 2009), three evaluation metrics are selected for different purposes: (1) Average Accuracy to represent the average per-class effectiveness of the classifier; (2) F_1score , which is the harmonic mean of precision (a measure of exactness) and recall (a measure of completeness) with equal weights, to represent the effectiveness of the classifier to identify positive class; (3) Area Under the Curve (AUC), which is also referred as balanced accuracy, to represent the classifier's ability to avoid false classification. Usually, the AUC of a classifier should be in the interval [0.5,1]. When the AUC is equal to 0.5, it means random guessing in binary classification. When the AUC equals to 1, the prediction of the classifier perfectly matches the actual labels.

$$\text{Average Accuracy} = \left(\sum_{i=1}^{k} \frac{\text{TP}_i + \text{TN}_i}{\text{TP}_i + \text{TN}_i + \text{FP}_i + \text{FN}_i} \right) / k \tag{5}$$

$$\text{Average Precision} = \left(\sum_{i=1}^{k} \frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i} \right) / k \tag{6}$$

$$\text{Average Recall} = \left(\sum_{i=1}^{k} \frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i} \right) / k \tag{7}$$

$$\text{F}_1\text{score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{8}$$

$$\text{AUC} = \frac{1}{2} \left(\sum_{i=1}^{k} \frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i} / k + \sum_{i=1}^{k} \frac{\text{TN}_i}{\text{TN}_i + \text{FP}_i} / k \right) \tag{9}$$

The performance test is to investigate the developed classifiers from two aspects: (1) to compare the classification based on two discretization methods; (2) to discover the most appropriate criterion(-a) to classify ships regarding the ship behavior.

In the next chapter, the methodology introduced in this chapter will be applied to the data set introduced in chapter 2.

4. Results and analyses

For the application of the proposed methodology, the behavior of both inbound and outbound ships in the study area have been analyzed. Since the sailing direction (approach to or departure from a port) might influence the ship behavior, the data of inbound and outbound ships are handled as two independent data sets in the research. In this section, the behavior clustering results are discussed (section 4.1) and the performances of classifiers based on different ship characteristics are compared (section 4.2).

4.1. Ship behavior clustering and statistical test results

The behavior clustering results of ship path and SOG are discussed in this section. The results of inbound ships are presented with detailed explanations, while the results of outbound ships are briefly provided, as they only deviate slightly from the results of the inbound ships.

According to the methodology proposed in section 3.2, the ship behavior attributes are clustered into 2 to 10 clusters. The statistical t-test is performed to the behavior data of all clusters on all cross-sections, as presented in Table 2.

Table 2. Number of cross-sections with significantly different behavior in all clusters of inbound ships.

No. of clusters	2	3	4	5	6	7	8	9	10
Path	35	35	35	35	35	33	32	29	28
SOG	35	35	35	35	34	33	32	30	29

* The total number of cross-sections is 35. The gray shading indicates the formed clusters are significantly different on some cross-sections, not over the whole area.

For the ship path, the clusters are significantly different from each other on all cross-sections, when the number of clusters is less than 7. With a further increase of the number of clusters, the clusters are not significantly different on all cross-sections anymore. Thus, the number of path clusters of inbound ships is determined as 6, for which the different behavior patterns (indicated as clusters) over the whole area can be recognized. The ship paths for each cluster are shown in Figure 4. Since the mole locating at the south of the study area is slightly bent with an irregular outline, the shape is extracted from the map and indicated as an irregular black bar at the bottom of each plot. The solid line shows the center value of each cluster, while the dashed line shows the 95% confidence interval of the ship path of each cluster on each cross-section. The lines do not refer to the behavior of a single ship, but the behaviors of all ships within that cluster. The vessel behavior of clusters varies within the range. Thus, an overlap of the range can be observed between clusters, which refers to the ships with the behaviors on the edges of the clusters. For the individual ships from two clusters may have similar behaviors. However, the defined function of distance to the cluster center (in section 3.2) can mathematically guarantee that every ship will be clustered to the group with least dissimilarity over the whole area. When comparing the cluster centers of all clusters, the ship paths are generally different over the whole area, with the ship path of cluster #1 closest to the starboard bank while the ship path of cluster #6 is farthest away from the starboard bank. The range and distribution of paths in clusters are also different, when looking at the ship path at 95% confidence interval. Along the sailing direction of inbound ships (from the left to the right of the figure), the ship paths in all clusters move farther away from the X'-axis. The reason is that the waterway becomes narrower outside the right boundary of the study area as shown in Figure 1. Thus, the ships sail farther away from the starboard

bank to enter the neighboring waterway smoothly. Similarly, the clustering results for outbound ships consist of 4 clusters of path.

The comparisons of ship path centers between clusters of inbound and outbound ships are presented in Figure 5. The ship path centers for outbound clusters (see Figure 5(b)) are obviously different with each other over the study area. However, for the inbound ships, the path centers of cluster #3 and cluster #4 cross in the middle area. The ships in cluster #3 sail closer to the starboard bank when entering the study area from the sea compared to the ships in cluster #4, while farther away from the bank when leaving the study area. When investigating the composition of ships in these two clusters, approximately half of ships in both clusters have medium length (100-170 meters) and medium beam (16-26 meters). Of the remaining ships in cluster #3, more ships have small lengths and beams than large ones. The percentage of ships with small and large length and beam in cluster #4 are around the same. As the majority of ships in these two clusters are similar in ship characteristics and the sailing situation is unhindered, the behavior difference in path is possibly due to the maneuvering preferences of officers onboard when sailing in a narrowing waterway. Especially for the ships with medium size, the space for sailing in the waterway is more flexible than smaller ships (which need to sail closer to the starboard bank due to the navigation rules in narrow channels) or larger ships (which need to sail farther to the starboard bank due to the bigger draught). Some officers prefer to change their position more to the center in advance in case of unexpected situation in the neighboring narrower waterway, which is presented as cluster #3. The other officers prefer to keep themselves as close to the starboard bank as possible according to the advice of navigation rules, shown as cluster #4. Thus, not only the different behavior patterns over the whole area can be recognized by the proposed clustering method, the behavior change patterns are also identified in different clusters.

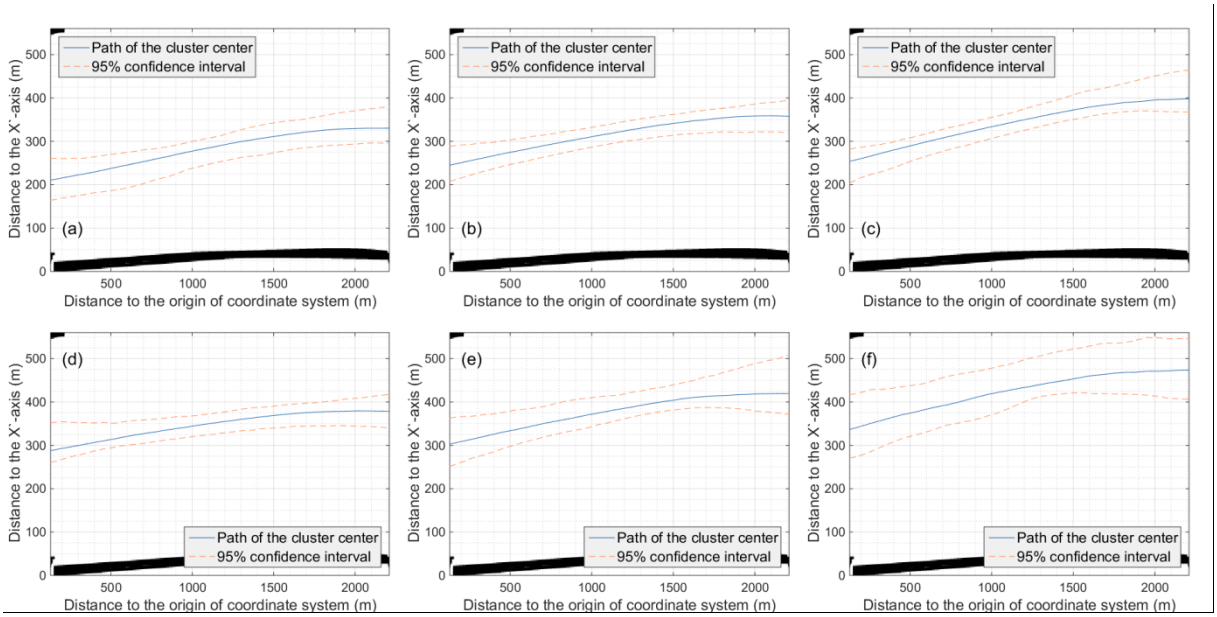

Figure 4. Clustered inbound ship paths: (a) cluster #1; (b) cluster #2; (c) cluster #3; (d) cluster #4; (e) cluster #5; (f) cluster #6 (The distance to the X'-axis as a function of the distance to the origin of the coordinate system with the sailing direction from the left to the right of the figure). The black bar at the bottom of each plot indicates the mole, Splitsingsdam, as shown in Figure 1.

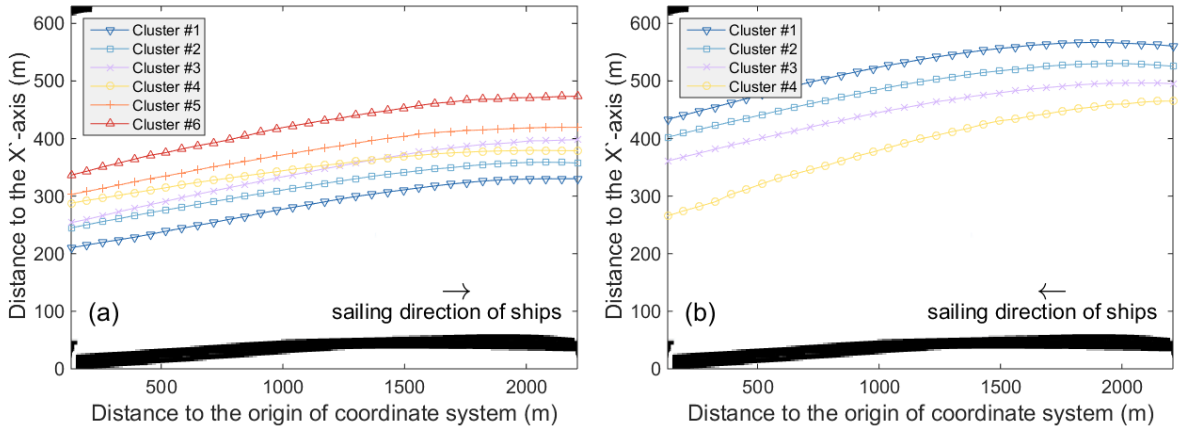

Figure 5. Comparison of ship path in clusters: (a) inbound ships; (b) outbound ships. The black bar at the bottom of each plot indicates the mole, Splitsingsdam, as shown in Figure 1.

The method to decide the number of clusters for SOG is the same as the method used for ship path, see Table 2. The five clusters of SOG for inbound ships are presented in Figure 6. The ships in SOG cluster #1 sail with the lowest speed, while the SOG for cluster #5 is highest. For the ships in cluster #3, #4, and #5, the range of the 95% confidence interval is larger at the boundaries than in the middle of the area. The ships with high speed tend to have more variation of speed when entering a waterway, while behaving similarly when there is no change in the sailing environment. The behavior patterns of SOG are significantly different in all clusters over the whole study area for both inbound and outbound ships, as shown in Figure 7. Most of the ships keep a stable speed during the whole voyage in the area (cluster #1, #2, #3 for inbound ships, cluster #1, #2, #3, #4 for outbound ships).

However, for the ships with high speed, the closer to the neighboring narrow waterway (the right of the figure), the lower speed they sail (cluster #4, #5 for inbound ships, cluster #5, #6 for outbound ships). A lower speed in the narrow waterway reduces the impact of ship wave on other ships and guarantees sufficient time for maneuvering in case of unexpected situations.

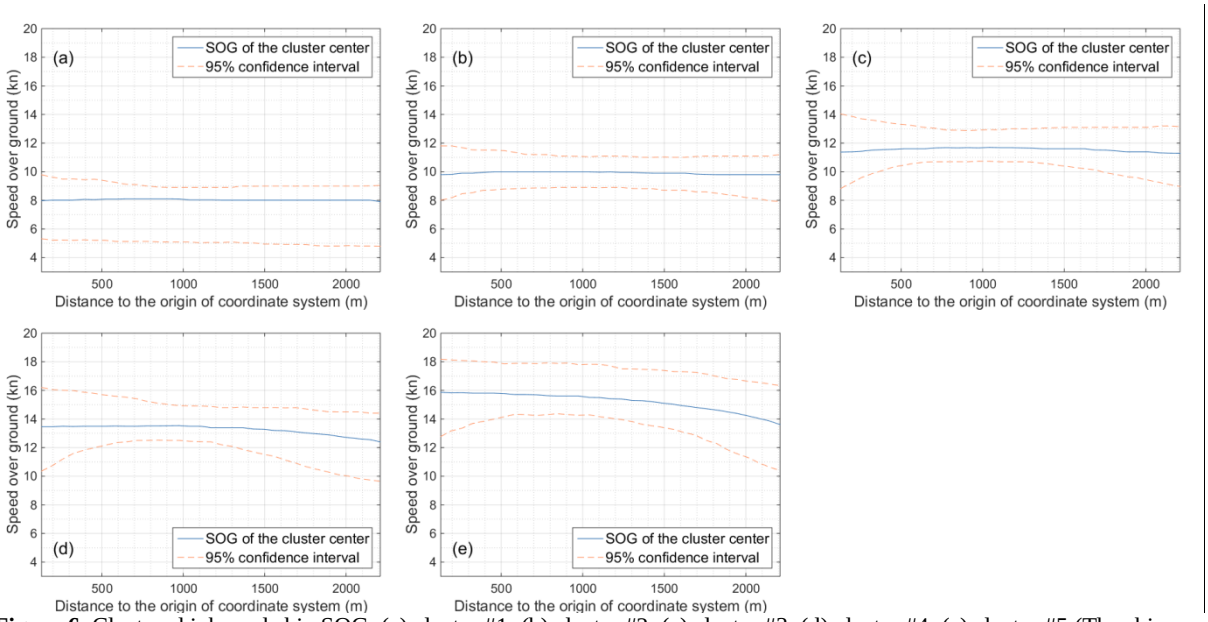

Figure 6. Clustered inbound ship SOG: (a) cluster #1; (b) cluster #2; (c) cluster #3; (d) cluster #4; (e) cluster #5 (The ship SOG as a function of the distance to the origin of the coordinate system with the sailing direction from the left to the right of the figure)

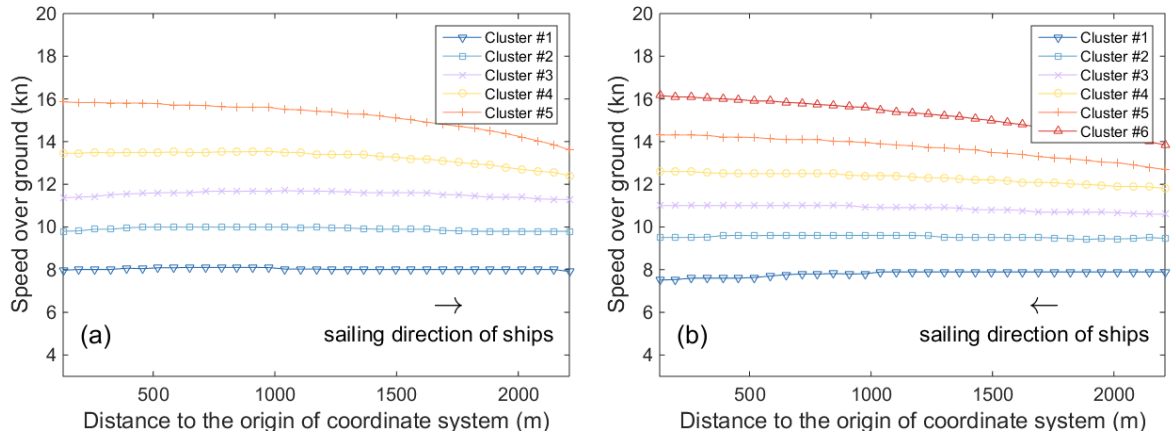

Figure 7. Comparison of ship SOG in clusters: (a) inbound ships; (b) outbound ships.

To further interpret the clustering results with respect to the correlation between ship path and SOG (the integral ship behavior), the number of ships in each path and SOG cluster for both inbound and outbound directions are shown in Figure 8. Based on the number of ships in clusters for path and SOG, some patterns of the integral behavior can also be revealed for both inbound and outbound ships. Generally, ships sailing with low speed (SOG #1) mostly sail close to the starboard bank (Path #1). With an increase of speed, ships sail farther from the starboard bank. However, the ships with the

highest SOG (cluster #5 of inbound ships and cluster #6 of outbound ship) keep a proper distance to both sides of the bank, instead of sailing farthest to the starboard bank. It is possibly in consideration for sufficient maneuvering space for ship encounter or other special circumstances.

When comparing the behavior pattern for inbound and outbound ships, the speed range for both are similar, but the distribution of speed for the ships sailing farthest away from the starboard bank is different. For the inbound ships (path cluster #6), most of them sail with a high speed as in SOG cluster #4 and #5. Meanwhile, the speed for outbound ships in path cluster #4 is evenly distributed. When investigating for both directions the composition of ships sailing farthest to starboard bank, most of them are large-size ships. In narrow waterway, large ships need a relatively high speed to maintain the rudder effectiveness, even it may consume more fuel. For inbound ships, they need to keep a relatively high speed (SOG cluster #4 and #5) for maneuvering in the neighboring narrow waterway. However, for outbound ships, they can sail at their desired speed maintaining the basic maneuverability without consuming much fuel.

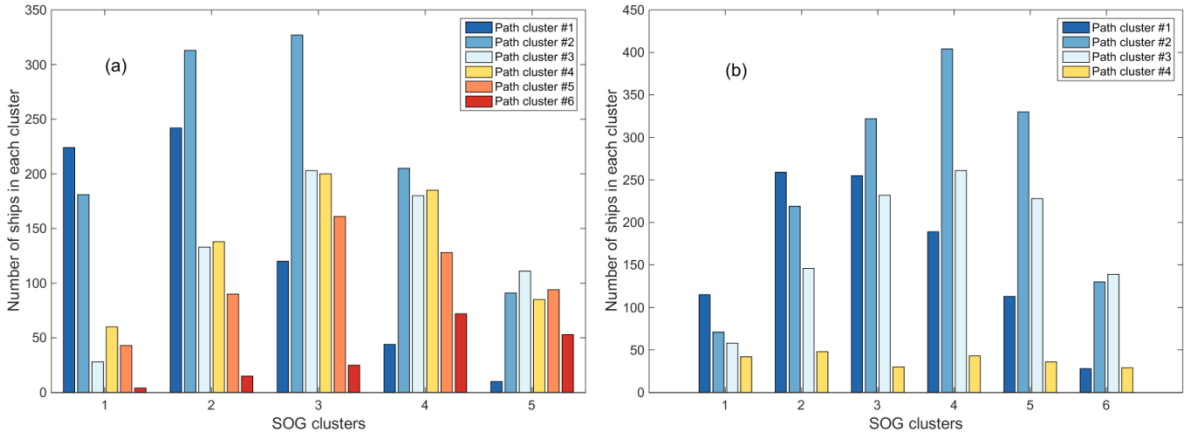

Figure 8. The number of ships in different path and SOG clusters: (a) inbound ships; (b) outbound ships.

4.2. Ship classification and performance measures results

As stated in the methodology of ship classification (section 3.3), the first step is to estimate the prior probability of different ship characteristics. Since the characteristics of length, beam and the ratio between length and beam are all ratio variables, the ship data are analyzed using Arena Input Analyzer to fit common statistical distributions, including normal, lognormal, gamma, Erlang, beta, Weibull, uniform and exponential distributions (Takus and Profozich, 1997). The results show that the lognormal distribution is the most likely distribution for the beam and the ratio between length and

beam of inbound ships and all three characteristics for outbound ships, while the gamma distribution best describes the length of inbound ships. Based on the fitting results, the Kolmogorov-Smirnov test (K-S test) is performed to the data and the most likely distribution. The null hypothesis is that the data are drawn from the corresponding distribution. However, the null hypotheses are rejected for all characteristics of inbound and outbound ships with the p-value of 0.05. Thus, the most likely distribution cannot represent the real situation. Therefore, the second solution to estimate the prior probabilities for each characteristic is adopted consisting of computing the observations in the data set using supervised and unsupervised discretization methods, respectively. The number of intervals for different ship characteristics resulting from these two discretization methods is shown in Table 3. The supervised discretization method Chi2 results in different intervals of ship characteristics when considering the classes of behavior attributes.

Table 3. Number of intervals for different ship characteristics by two discretization methods.

Ship characteristics	Unsupervised discretization (EWB)	Supervised discretization (Chi2)			
		Inbound ships		Outbound ships	
		Path	SOG	Path	SOG
Length	18	23	6	21	26
Beam	14	3	6	4	4
Length/Beam	12	6	8	17	9

To test the classifiers, 50 runs of classification are performed with the holdout method using the collected data. The values of three evaluation metrics in 50 classification runs for each classifier follow a normal distribution, which has been statistically tested. The mean value of evaluation metrics of the classifiers for inbound and outbound ships are shown in Table 4.

Table 4. Mean value of evaluation metrics for the ship classification in 50 runs.

Data set of ships	Behavior attributes	Ship characteristics	Classification based on unsupervised discretization (EWB)			Classification based on supervised discretization (Chi2)		
			Average Accuracy	F_1score	AUC	Average Accuracy	F_1score	AUC
Inbound ships	Path	Length	0.7830	0.6233	0.7077	0.7943*	0.6781*	0.7435
		Beam	0.7838*	0.6256*	0.7198	0.7805	0.6567	0.7354
		Length/Beam	0.7734	0.6188	0.6899	0.7741	0.6602	0.7069
		Length & Beam	0.7837	0.6256*	0.7265*	0.7901	0.6687	0.7487*
	SOG	Length	0.7411	0.5881	0.6829	0.7391	0.6244	0.7025

		Beam	0.7465*	0.5898	0.6941	0.7462*	0.6296*	0.7031
		Length/Beam	0.7304	0.5791	0.6658	0.7286	0.6126	0.6808
		Length & Beam	0.7450	0.5900*	0.7013*	0.7412	0.6140	0.7182*
Outbound ships	Path	Length	0.7266	0.6443	0.6801	0.7333	0.6771	0.7028
		Beam	0.7344*	0.6506*	0.6927*	0.7328	0.6774	0.7019
		Length/Beam	0.7036	0.5886	0.6373	0.7087	0.6358	0.6589
		Length & Beam	0.7295	0.6486	0.6916	0.7346*	0.6802*	0.7065*
	SOG	Length	0.7756	0.6149	0.7159	0.8001*	0.6868*	0.7602
		Beam	0.7820	0.6209	0.7230	0.7788	0.6545	0.7311
		Length/Beam	0.7635	0.5984	0.6988	0.7686	0.6418	0.7200
		Length & Beam	0.7865*	0.6277*	0.7318*	0.7988	0.6797	0.7621*

* In the classification for each behavior attribute in each data set of ships (e.g. for the path of inbound ships), the highest values of each evaluation metrics in the developed classifiers are marked with a star(*), respectively. The gray shading indicates the classifiers outperforming others in all three evaluation metrics.

For all of the developed classifiers, the AUC values are larger than 0.5, which suggests the classifiers perform better than a random class assignment. The classifiers with gray shading in Table 4 are the ones with an overall good performance in classifying ships to the corresponding behavior classes. For the classifiers marked with two stars, the differences of the remaining evaluation metrics to the highest value are all less than 0.01. The performance of such classifier is also deemed as comparable and adoptable in practice.

To compare the classification based on two discretization methods, the classifiers based on Chi2 outperform the ones with EWB. The reason is that the behavior classes of ships have been considered during the Chi2 discretization. Thus, the estimate of prior probabilities based on such discretized intervals will lead to a better performance in classification. Especially when looking at F_1score and AUC which indicate the ability to identify the correct classes, the classifiers based on Chi2 discretization are better than the other algorithm in all cases. Thus, the Chi2 algorithm is recommended to perform discretization of ship characteristics, though the complexity of such classification increases than the one with EWB.

To discover the most appropriate criterion(-a) of ships characteristics, the performances of classifiers based on different characteristics in both discretization methods are analyzed. When investigating the classification based on a single criterion, the classifiers based on beam outperform

the others, considering the evaluation metrics and the number of intervals in Chi2 discretization. In the classification based on EWB discretization, the classifier based on beam also outperforms the other classifiers with a single criterion. Comparing the performances of all classifiers, the ones based on length & beam perform well with two or three stars in three cases (path in inbound ships, SOG in inbound ships, SOG in outbound ships) in the classification with EWB discretization. In the classification with Chi2 discretization, the classifiers based on length & beam are marked with the highest AUC value in all cases. The AUC value represents the classifier's ability to avoid false classification. Considering the other two evaluation metrics, the performances of classifier based on length & beam are also comparable to the ones with highest value. It could also be expected that with more ship characteristics for classification, the performance is better, since the ship can be characterized from different aspects.

The classification for inbound and outbound ships are developed and tested independently. If to choose a classifier for both inbound and outbound ships in practice for the study area, the one with Chi2 discretization based on length & beam is recommended.

5. Conclusions

This paper presents a methodology for clustering ship behavior in an area and classifying ships into these clusters based on the static ship characteristics. The ship behavior clustering methodology is based on the *k*-means theory and modified to overcome its drawbacks in subjective decision on number of clusters and sensitivity to initialization and stopping criteria. The proposed algorithm is stable in clustering results without subjective decisions in the initialization phase. The ship classifier is developed according to the principle of Naive Bayesian classification. Instead of assuming a distribution to estimate the prior probability, two discretization methods (unsupervised Equal Width Binning and supervised Chi2) are tested to calculate the probability. The most appropriate classifier can be indicated by the evaluation metrics.

The methodology has been independently applied to two subsets: inbound ships and outbound ships in the study area in the port of Rotterdam. The clustering result can recognize both the fully different behavior patterns over the whole research area and the different behavior change patterns for

some clusters of ships. The integral ship behavior pattern can also be revealed. With the holdout method, the developed classifier (based on training data set) has been used to classify ships (in testing data set) to the corresponding behavior cluster. The evaluation results of classification show that the Chi2 algorithm tends to perform better than EWB. The classifications based on length & beam outperform the ones based on a single criterion. The results reveal the underlying relation between ship characteristics and behavior patterns.

Both the port authority and the researchers could benefit from the proposed methodology and the identified clusters. For the port authority, the ship behavior clustering results reveal the ship behavior patterns and the ship behavior change patterns in a specific area, which helps the port operation and traffic management. The developed classifiers can be used to predict the behavior patterns of ships. For nautical researchers, this paper provides an integrated method of behavior pattern recognition based on AIS data and the corresponding ship classification. The results could also help to simulate the behavior of different ships in a systematic way. For data mining researchers, the method of deciding the number of clusters in *k*-means clustering can be applied to other problems. The results of classification based on two discretization methods indicate the applicability and effectiveness in the domain of maritime traffic.

Future research can be to include more ship characteristics in the classification. With a more comprehensive data set of ship particulars (e.g., GT, DWT, actual draught), the classifier performances can be compared to choose the criterion best indicating ship behavior patterns. In a later stage, the results will be applied in traffic model development to simulate such ship behavior. Given the detailed recognition of ship behavior patterns, the simulation results will be closer to the reality.

Acknowledgement

This work is initiated by the project, Nautical traffic model based design and assessment of safe and efficient ports and waterways, under the Netherlands Organization for Scientific Research (NWO). The fellowship of Yang Zhou is supported by China Scholarship Council and Delft University of Technology. The support from SmartPort, both financially and by embedding the research in the practical context of the Port of Rotterdam, is highly appreciated. The authors would also like to thank

the department of Data Management in the port of Rotterdam during the data collection, and appreciate Frank Cremer for accessing AIS data, Cor Mooiman for providing wind and visibility data, Bob van Hell and Lamber Hulsen for simulating current data.

References

De Boer, T., 2010. An analysis of vessel behaviour based on AIS data. TU Delft, Delft University of Technology.
Goerlandt, F., Kujala, P., 2011. Traffic simulation based ship collision probability modeling. Reliability Engineering & System Safety 96 (1), 91-107.
Gunnar Aarsæther, K., Moan, T., 2009. Estimating navigation patterns from AIS. Journal of Navigation 62 (04), 587-607.
Han, J., Pei, J., Kamber, M., 2011. Data mining: concepts and techniques. Elsevier.
Harris, J.W., Stöcker, H., 1998. Handbook of mathematics and computational science. Springer Science & Business Media.
IMO, 1974. International Convention for the Safety of Life at Sea, London.
IMO, 2003. Guidelines for the Installation of a Shipborne Automatic Identification System (AIS), SN/Circ.227, London.
ITU, 2014. Technical characteristics for an automatic identification system (AIS) using time division multiple access in the VHF maritime mobile frequency band, in: ITU (Ed.), M.1371-5, Geneva.
Joiţa, D., 2010. Unsupervised static discretization methods in data mining. Titu Maiorescu University, Bucharest, Romania.
Kotsiantis, S.B., Zaharakis, I., Pintelas, P., 2007. Supervised machine learning: A review of classification techniques.
Lavangnananda, K., Chattanachot, S., 2017. Study of discretization methods in classification, Knowledge and Smart Technology (KST), 2017 9th International Conference on. IEEE, pp. 50-55.
Liu, H., Setiono, R., 1995. Chi2: Feature selection and discretization of numeric attributes, Tools with artificial intelligence, 1995. proceedings., seventh international conference on. IEEE, pp. 388-391.
Mascaro, S., Korb, K.B., Nicholson, A.E., 2010. Learning abnormal vessel behaviour from ais data with bayesian networks at two time scales. Tracks A Journal Of Artists Writings.
Mou, J.M., Tak, C.v.d., Ligteringen, H., 2010. Study on collision avoidance in busy waterways by using AIS data. Ocean Engineering 37 (5), 483-490.
Pallotta, G., Vespe, M., Bryan, K., 2013. Vessel pattern knowledge discovery from ais data: A framework for anomaly detection and route prediction. Entropy 15 (6), 2218-2245.
Ristic, B., La Scala, B., Morelande, M., Gordon, N., 2008. Statistical analysis of motion patterns in AIS data: Anomaly detection and motion prediction, Information Fusion, 2008 11th International Conference on. IEEE, pp. 1-7.
Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O.P., Tiwari, A., Er, M.J., Ding, W., Lin, C.-T., 2017. A review of clustering techniques and developments. Neurocomputing 267, 664-681.
Sheng, K., Liu, Z., Zhou, D., He, A., Feng, C., 2018. Research on Ship Classification Based on Trajectory Features. The Journal of Navigation 71 (1), 100-116.
Shu, Y., Daamen, W., Ligteringen, H., Hoogendoorn, S., 2013. Vessel Speed, Course, and Path Analysis in the Botlek Area of the Port of Rotterdam, Netherlands. Transportation Research Record: Journal of the Transportation Research Board (2330), 63-72.
Shu, Y., Daamen, W., Ligteringen, H., Hoogendoorn, S.P., 2017. Influence of external conditions and vessel encounters on vessel behavior in ports and waterways using Automatic Identification System data. Ocean Engineering 131, 1-14.
Silveira, P., Teixeira, A., Soares, C.G., 2013. Use of AIS data to characterise marine traffic patterns and ship collision risk off the coast of Portugal. Journal of Navigation 66 (06), 879-898.
Sokolova, M., Lapalme, G., 2009. A systematic analysis of performance measures for classification tasks. Information Processing & Management 45 (4), 427-437.

Takus, D.A., Profozich, D.M., 1997. ARENA software tutorial. Institute of Electrical and Electronics Engineers (IEEE).
Vollebregt, E.A., Roest, M., Lander, J., 2003. Large scale computing at Rijkswaterstaat. Parallel Computing 29 (1), 1-20.
Witten, I.H., Frank, E., Hall, M.A., Pal, C.J., 2016. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.
Xiao, F., Ligteringen, H., van Gulijk, C., Ale, B., 2015. Comparison study on AIS data of ship traffic behavior. Ocean Engineering 95, 84-93.
Zhou, Y., Daamen, W., Vellinga, T., Hoogendoorn, S., 2015. Vessel classification method based on vessel behavior in the port of Rotterdam. Scientific Journals of the Maritime University of Szczecin, 42 (114), 2015.
Zhou, Y., Daamen, W., Vellinga, T., Hoogendoorn, S.P., 2017. AIS data analysis for the impacts of wind and current on ship behavior in straight waterways, in: GUEDES SOARES, C., Teixeira, A. (Eds.), 17th International Congress of the International Maritime Association of the Mediterranean. CRC Press, pp. 265-272.