Journal Pre-proof

Ship behavior prediction via trajectory extraction-based clustering for maritime situation awareness

Brian Murray, Lokukaluge Prasad Perera

PII: S2468-0133(21)00022-X
DOI: https://doi.org/10.1016/j.joes.2021.03.001
Reference: JOES 211

To appear in: *Journal of Ocean Engineering and Science*

Received date: 15 October 2020
Revised date: 14 February 2021
Accepted date: 18 March 2021

Please cite this article as: Brian Murray, Lokukaluge Prasad Perera, Ship behavior prediction via trajectory extraction-based clustering for maritime situation awareness, *Journal of Ocean Engineering and Science* (2021), doi: https://doi.org/10.1016/j.joes.2021.03.001

Highlights

- This study presents an approach to predict the future behavior of a vessel using historical AIS data. Such predictions can subsequently be utilized to evaluate the potential future collision risk, supporting proactive collision avoidance. In this manner, enhanced situation awareness can be facilitated.
- By utilizing machine learning techniques, historical local behavior clusters can be extracted from historical AIS data to describe the possible future 30 minute behavior of a vessel. These clusters are discovered using the Karhunen-Loeve transform and Gaussian Mixture Models.
- The developed technique allows the vessel to be classified to a cluster of behavior, and conducts a trajectory prediction with respect to the behavior in this cluster.
- The method is evaluated using test cases from the region surrounding Tromsø, Norway. The results indicate that the technique has good performance in predicting future ship behavior.

Ship behavior prediction via trajectory extraction-based clustering for maritime situation awareness

Brian Murray*, Lokukaluge Prasad Perera

UiT The Arctic University of Norway, Tromsø, Norway

ARTICLE INFO

Keywords:
Maritime Situation Awareness
Ship Navigation
Trajectory Prediction
Collision Avoidance
Machine Learning
Unsupervised Learning
AIS

ABSTRACT

This study presents a method in which historical AIS data are used to predict the future trajectory of a selected vessel. This is facilitated via a system intelligence-based approach that can be subsequently utilized to provide enhanced situation awareness to navigators and future autonomous ships, aiding proactive collision avoidance. By evaluating the historical ship behavior in a given geographical region, the method applies machine learning techniques to extrapolate commonalities in relevant trajectory segments. These commonalities represent historical behavior modes that correspond to the possible future behavior of the selected vessel. Subsequently, the selected vessel is classified to a behavior mode, and a trajectory with respect to this mode is predicted. This is achieved via an initial clustering technique and subsequent trajectory extraction. The extracted trajectories are then compressed using the Karhunen-Loéve transform, and clustered using a Gaussian Mixture Model. The approach in this study differs from others in that trajectories are not clustered for an entire region, but rather for relevant trajectory segments. As such, the extracted trajectories provide a much better basis for clustering relevant historical ship behavior modes. A selected vessel is then classified to one of these modes using its observed behavior. Trajectory predictions are facilitated using an enhanced subset of data that likely correspond to the future behavior of the selected vessel. The method yields promising results, with high classification accuracy and low prediction error. However, vessels with abnormal behavior degrade the results in some situations, and have also been discussed in this study.

thumbnails/cas-email.jpeg brian.murray@uit.no (B. Murray)
ORCID(s):

Nomenclature

$\mathbf{a}$	Arbitrary AIS Parameter Vector
$\mathbf{A}$	Set of AIS Data
c	Trajectory Class
$\mathbf{C}$	Data Cluster
d	Euclidean Distance
e	Eigenvector
$\mathbf{E}$	Eigenvector Matrix
$\mathbf{f}$	Trajectory Feature Vector
$\mathbf{I}$	Identity Matrix
J_3	Class Separability Criterion
k	Hyper-parameter for kNN classifier
K_M	Number of Free Parameters in Mixture Model
L	Number of Data Points in Selected Trajectory
$LL(\cdot)$	Log-likelihood Function
M	Number of Models in Mixture Model
N	Number of Trajectories
$\mathbf{p}$	Arbitrary Vessel Position
$\mathbf{q}$	Selected Vessel Position
r	Search Radius $[m]$
$\mathbf{R}$	Rotation Matrix
$\mathbf{s}$	Vessel State
$\mathbf{S}_b$	Between-class Scatter Matrix
$\mathbf{S}_w$	Within-class Scatter Matrix
t	Timestamp
T	Elapsed Time $[s]$
T_δ	Additional Time Period $[s]$
T_p	Desired Prediction Time Horizon $[s]$
v	Speed over Ground $[m/s]$

x	UTM x-coordinate $[m]$
$\mathbf{x}$	Reduced Feature Vector
$\mathbf{X}$	Set of Reduced Feature Vectors
y	UTM y-coordinate $[m]$
$\mathbf{z}$	Class Membership
$\mathbf{Z}$	Spatial Data Matrix
ΔL	Step Size $[m]$
$\mathbf{\Lambda}$	Eigenvalue Matrix
$\boldsymbol{\mu}$	Mean Vector
π	Prior Distribution
$\mathbf{\Sigma}$	Covariance Matrix
θ	Rotation Angle [°]
$\mathbf{\Theta}$	Model Parameters
χ	Course over Ground [°]
Subscripts	
0	Initial State
i	Sample Number
g	Global
j	Class Number
k	k^{th} State
l	Number of Eigenvectors
m	Model Number in Gaussian Mixture
δ	Maximum Offset
Superscripts	
$\hat{}$	Estimated Parameter / State
Acronyms	
AIS	Automatic Identification System
BIC	Bayesian Information Criterion
EM	Expectation Maximization
GMM	Gaussian Mixture Model
KL	Karhunen-Loéve
LDA	Linear Discriminant Analysis

1. Introduction

Technological advances are permeating almost every industry. Artificial intelligence, increased computational power and wireless communication capabilities have the potential to allow for disruptive innovations that can change business models drastically. Many argue that there is a digital revolution underway and are calling it Industry 4.0 (Hermann et al., 2016). If one looks to the automotive industry for instance, significant innovations related to autonomous cars are being developed at an exponential rate. Autonomous cars are already being tested in general traffic areas and there are claims that mass production could be possible by 2021 (Chan, 2017).

Similarly, it can be argued that shipping is currently on its way into a fourth technical revolution, Shipping 4.0 (Rødseth et al., 2015). The first revolution in shipping can be argued to be the transition from sail to steam in at the turn of the 19^{th} century, the second from steam to diesel around 1910, and the third came with the introduction of automated systems, made possible through the advent of computers around 1970. Like the car industry, the shipping industry is looking to autonomy as a possible disruptive element. The shipping industry has, however, historically been considered conservative, with innovations being implemented at a slower rate than in similar industries. As such, technologies associated with autonomous ships are not as developed as those for autonomous cars. Nonetheless, many companies are working on the development of autonomous ships. The first autonomous ships, e.g. Yara Birkeland, are planned to be launched in 2020 and fully autonomous by 2022 (Yara, 2019). It can be argued that if the required technologies are available, autonomous ships will be safer and more efficient than conventional vessels, and that because of this fact they should be adopted by the industry (Levander, 2017). For this to occur, however, autonomous ships must be proven to operate at a level of safety comparable to, or better than, conventional manned vessels.

1.1. Maritime Situation Awareness

For autonomous ships to be introduced into commercial shipping lanes, effective collision avoidance systems (Perera et al., 2015) must be in place to ensure that the autonomous operations have the required level of safety. Given that the vessels are unmanned, an autonomous ship must be able to make decisions based on its understanding of its surroundings, i.e. its own situation awareness. Situation awareness is defined as *"Being aware of what is happening around you and understanding what that information means to you now and in the future"* (Endsley et al., 2003), and is separated into three levels (Endsley, 1995):

1. Perception of the elements in the current situation
2. Comprehension of the current situation
3. Projection of the future status

For an autonomous vessel, situation awareness will primarily entail obstacle detection and prediction of close-range encounter situations. Other vessels are the most common obstacle an autonomous ship will encounter, and are referred to as target vessels in an encounter situation. The autonomous vessel in this case is referred to as the own ship. Such situations will require collision avoidance maneuvers.

1.1.1. Perception of Elements in The Current Situation

To effectively conduct collision avoidance maneuvers with respect to target vessels, an own ship will need to be able to first detect the target vessel, and evaluate relevant parameters such as its position, course over ground and speed over ground. This can be considered as the first level of situation awareness. An autonomous ship must, therefore, first define its current state, where all obstacles and their current states are known. In order to perceive the relevant obstacles, an autonomous ship must be able to observe them. Since there is no navigator on-board, collision avoidance technologies will rely heavily on the sensor suite available on-board the vessel, as they must in essence replace the eyes of the navigator. An advanced obstacle detection and tracking system, which utilizes sensor fusion to enhance detection capabilities, should be utilized. Relevant sensors will likely include RADAR (Radio Detection and Ranging) and electro-optical sensors (Prasad et al., 2017). Some examples include, LIDAR (Light Detection and Ranging), stereo cameras and infra-red cameras.

1.1.2. Comprehension of The Current Situation

Based on its current state, the own ship must be capable of evaluating the risk of collision. If there is a risk of collision, the own ship must conduct a collision avoidance maneuvers that adhere to the COLREGS as outlined in Perera et al. (2010). This corresponds to level two of Endsley's situation awareness, where the ship must now make sense of its current state, and the immediate implications it has for the safety of the operation. Fujii and Tanaka (1971) and Goodwin (1975) introduced the concept of the ship domain, where a safety region around a relevant vessel is introduced to indicate the collision risk. A thorough review of collision avoidance methods can be found in Tam et al. (2009). These methods are designed with respect to ships in close-range encounters, where the collision risk is high enough to require collision avoidance maneuvers.

1.1.3. Projection of The Future Status

Level three situation awareness addresses the projection of the future state of the vessel. In a collision avoidance setting, this entails predicting both the future states of the own ship, as well as the future states of target vessels. Previous studies relating to collision avoidance techniques entail predicting the future state of a target vessel via calculations using constant course and speed values. Based on this, collision risk parameters relating to the closest point of approach (CPA) such as the distance (DCPA) and time (TCPA) can be determined, and necessary collision avoidance maneuvers conducted on this basis.

Ships have a slow response time when control actions are sent to change the speed or course over ground. Cars for instance can make changes almost instantaneously, depending on their speed. The inertia forces of a ship are, however, much higher, and resultant collision avoidance maneuvers will take much longer to conduct. Therefore, it is desirable to predict the risk of collision as far as possible in advance. This entails predicting the future trajectories of both the own ship and target vessels accurately. Methods such as Perera et al. (2011), where a fuzzy logic based decision making system for collision avoidance was introduced, and Yang et al. (2019), where parallel trajectory planning was proposed for autonomous collision avoidance, can improve the ability of an autonomous vessel to make decisions. Additionally, work on more advanced prediction algorithms, e.g. Perera et al. (2012), where extended Kalman filters were utilized to estimate ship trajectories, can enhance the situation awareness of autonomous vessels to aid in effective collision avoidance. However, predictions under such methods are only useful up to rather short prediction horizons (order of seconds to minutes). These methods are, therefore, useful in the case of a close-range encounter situation. In such situations, the own ship must make decisions based on input from the sensor system, and plan effective collision avoidance maneuvers. This, however, entails that there is a risk of collision.

This study suggests an approach in which the trajectory of a target vessel is predicted far in advance, such that potential close-range encounter situations are prevented from occurring. With an enhanced level of situation awareness, an autonomous vessel can predict its own future states, as well as those for relevant target vessels, for a period up to 30 minutes into the future. Based on this level of situation awareness, intelligent decisions can be made to identify possible future close-range encounter situations, and optimally implement simple proactive collision avoidance strategies. Examples of such strategies may include minor speed or course alterations, such that the future trajectory of the own ship is altered. This is unfortunately not straight forward. It can be assumed that the majority of vessels will be manned in the foreseeable future. As such, the behavior of potential target vessels is highly unpredictable for an autonomous agent. Such a strategy, therefore, requires a system intelligence based approach to maritime situation awareness.

1.2. System Intelligence Based Ship Trajectory Prediction

Data from the Automatic Identification System (AIS) provide a powerful data set upon which analytics can be conducted. Historical AIS data provide insight into historical ship behavior that can be used to gain insight into patterns in maritime traffic. A myriad of ship parameters are recorded in the stored ship trajectories, including positional data, speed over ground values, and course over ground values for various time instances. AIS data provide an ideal data set upon which machine learning techniques can be applied to yield insight into patterns for subsequent use in maritime traffic analysis. Machine learning is a powerful tool, where insight can be extracted from data for a variety of purposes. Examples in the maritime field include Xu et al. (2020), where an optimal truncated least square support vector was utilized to estimate parameters for nonlinear maneuvering models, and Shen et al. (2019) where deep reinforcement learning was used to facilitate automatic collision avoidance.

This study suggests to provide future vessels with a degree of system intelligence, facilitated by historical knowledge that is extrapolated via machine learning techniques from AIS data. Using the historical knowledge available, such system intelligence will provide predictions of vessel trajectories, allowing for subsequent collision risk assessment. The purpose is to enhance the safety of both future autonomous ship operations, as well as provide decision support to conventional vessels. This section presents relevant related work and the contributions of this study.

1.2.1. Related Work

An increasing amount of research is being conducted on methods to utilize AIS data. Zhang et al. (2017) analyzed AIS data to gain insight into the spatial-temporal dynamics of ship traffic around ports. Additionally, Liu et al. (2019) used AIS data to evaluate regional collision risk, and Wen et al. (2020) utilized AIS data to automatically generate ship routes. Tu et al. (2017) provided a comprehensive review of methods to exploit AIS data for maritime navigation. Most work in the field has previously focused on predicting vessel trajectory patterns and general traffic behavior e.g. Aarsæther and Moan (2009). Identifying anomalous behavior based on general vessel patterns, e.g. Laxhammar et al. (2009), has also been of focus. These methods are useful for general behavior analysis, but are of limited use with respect to aiding in collision avoidance.

Of most interest in a collision avoidance setting is the work done on utilizing AIS data to predict the future trajectory of a vessel. The idea is to infer the future trajectory of a vessel based on the historical behavior of vessels in the same region. This information is stored in historical AIS data. Ristic et al. (2008) presented a method to predict the future motion of a vessel utilizing a particle filter approach, but the accuracy is limited for use in collision avoidance. Pallotta et al. (2013) presented the TREAD (Traffic Route Extraction and Anomaly Detection) methodology to cluster all trajectories in a defined region in an unsupervised manner, and subsequently classify a selected vessel to one of the clusters, each representing a traffic route for the purpose of anomaly detection. Pallotta et al. (2014) subsequently utilized the TREAD methodology to identify traffic routes, classify a vessel to a route, and predict the vessel position along this route using the Ornstein-Uhlenbeck stochastic process. The TREAD technique, however, clusters way-points, entry points, and stationary points such that the data for the entire region is utilized to differentiate between the vessels. As such, there can be significant discrepancies between sub-paths for trajectories belonging to the same class. This is of limited importance for long-term predictions (order of hours), and the method using the Ornstein-Uhlenbeck stochastic process is effective in such cases. The method's mean and variance functions do not change over time, however, which can be considered a strict assumption for real applications. Short-term predictions (order 5-30 minutes) of high accuracy and resolution, however, are arguably of more interest for collision avoidance purposes. For such predictions, the method will not be as effective. Mazzarella et al. (2015) also presented a prediction method using a Bayesian network-based algorithm with a particle filter for prediction horizons in the order of hours. However, this method also has limited efficacy in short-term trajectory predictions relevant for collision avoidance purposes.

Hexeberg et al. (2017) presented an AIS-based approach to predict short-term vessel trajectories. The method utilizes a single point neighbor search method to predict a vessel trajectory based on the underlying AIS data. The method, however, is unable to handle branching, and Dalsnes et al. (2018) expanded on this work to provide multiple predictions via a prediction tree, where samples are drawn from close neighbors in the underlying data. In this manner, a probability estimate can be evaluated for the future position at a given point in time, facilitated via Gaussian mixture models. As opposed to previous methods, these methods do not utilize clustering to identify traffic routes. All predictions are based on the AIS data in the neighborhoods of predicted states. As such, these methods do not take into consideration the relationship between data points. Future states are predicted iteratively from an initial state based on the AIS data in the neighborhood of a predicted position. These data, however, may include data points that have no relationship to the initial, or previous, predicted states, and as such may degrade the accuracy. Rong et al. (2019) also presented an approach using a Gaussian process model, where a probabilistic trajectory prediction method is outlined which, in addition to predicting the future positions of a vessel, also describes the uncertainty of the predicted position. The method, however, is only evaluated with using regular ship routes and offers no method to identify multiple possible future routes the vessel may follow, and classify it to one.

1.2.2. Contribution

In this study, a method to provide system intelligence to future autonomous ships is suggested for the purpose of enhanced situation awareness. The method is facilitated by leveraging historical AIS data via machine learning techniques to predict the future trajectory of a vessel based on its initial state. The method provides short-term trajectory predictions (order 5-30 minutes) that can provide a basis for collision risk assessments. In this manner, possible close-range encounter situations can be avoided, and the overall safety associated with autonomous operations can be increased.

The method presented in this study is based on a similar structure to that of previous techniques, in that trajectories are first clustered, a selected vessel is classified to a given cluster of trajectories, and a subsequent trajectory prediction is determined. However, this method is designed to aid in short-term trajectory predictions. As such, an alternative approach is suggested, where an initial clustering technique is utilized to extract a subset of data from a historical AIS data set, centered about the initial vessel state. This cluster contains AIS data that has a high degree of similarity to the initial state of the selected vessel. Using this initial cluster, all unique future, i.e. forward, trajectories are extracted from the cluster. The length of these is defined by the desired prediction time horizon. These trajectories represent all future paths of ships that had similar states to the initial state of the selected vessel. This data set will, therefore, only contain data that are related to the initial vessel state, as well as retain the relationship between data points.

The extracted forward trajectories represent the possible future behavior of the selected vessel for a given prediction horizon. In this study, it is of interest to identify all possible trajectory modes of the historical ship behavior, such that a high fidelity trajectory prediction can be conducted to support collision avoidance. Identifying such modes can be facilitated by clustering the forward trajectories. It is only of interest to differentiate between different possible modes for the duration of the desired prediction horizon. As such, clustering the extracted forward trajectories will provide a better basis for relevant route identification compared to other methods where entire trajectories for regions are considered.

A clustering technique is suggested based on all relevant data in each unique extracted forward trajectory. Dimensionality reduction via the Karhunen-Loéve transform is first conducted in order to compress the trajectories, whilst retaining the most important information relevant for differentiating the ship behavior. Such dimensionality reduction should support the clustering performance. Clustering is then facilitated via unsupervised Gaussian Mixture Modeling. A selected vessel is then classified to a cluster based on its past behavior. This is achieved backward trajectory extraction, and optimally generating features for class separation using Linear Discriminant Analysis. Finally, a trajectory prediction is conducted with respect to the trajectory data in the cluster of historical ship behavior.

The method has enhanced performance as it can discover the cluster of most similar ship behavior. This allows for predictions with a higher degree of fidelity than other methods with respect to collision avoidance. This is effective for challenging regions with more complex traffic, i.e. multiple possible routes with various speeds. The trajectory prediction also provides an increased level of accuracy given the relationship between data points in the underlying data. Similar methods use techniques that introduce time invariance, such as dynamic time warping. These result in effective clustering of trajectories of similar shapes for a given region, but can not capture the relationship between when various maneuvers takes place. Furthermore, clustering ship trajectories for entire regions will yield different results than clustering the extracted trajectories as suggested in this study. Therefore, the technique in this study provides a better basis for differentiating relevant historical ship behavior, supporting ship trajectory prediction. The method can also be applied in any geographical region, as the algorithm only requires access to raw AIS data of sufficient density for the region of interest.

An initial version of this work was presented in Murray and Perera (2019). Furthermore, in Murray and Perera (2020), a Dual Linear Autoencoder approach was introduced to facilitate trajectory prediction, that utilized similar clustering and classification regimes to those in this study. The clustering and classification techniques utilized in Murray and Perera (2020) were, however, not the focus of the study, and, therefore, not addressed in detail. This study can, therefore, be considered a parallel study, where the methods introduced in Murray and Perera (2019) are expanded upon, and addressed in detail.

2. Methodology

This section outlines the methodology utilized to facilitate trajectory predictions via the proposed system intelligence approach. First, the general approach to facilitate a trajectory prediction is presented. Second, the trajectory clustering module is outlined. Next, the trajectory classification module is discussed. Finally, the methodology involved in the trajectory prediction module is presented.

2.1. General Prediction Approach

The objective of the method presented in this study is to facilitate a prediction of the future trajectory of a target vessel, hereafter referred to as a selected vessel. In this study, a prediction horizon of 30 minutes is investigated. This, however, can be varied based on the desires of the user. It is further assumed that the past 10 minute behavior of the selected vessel is available. The architecture of the method can be split into three modules; the trajectory clustering module, the trajectory classification module and the trajectory prediction module. This is illustrated in Fig. 1.

The trajectory clustering module first employs an initial clustering technique. Based on the current state of the selected vessel, the technique identifies historical AIS messages in a defined region surrounding the current position of the selected vessel. Furthermore, data are filtered such that they have a similar speed and course over ground values to that of the selected vessel. In this manner, ships with similar behavior in the past are identified. The forward, i.e. future, trajectories are then extracted 30 minutes into the future from their initial data points. These represent the distribution of possible 30 minute behavior for the selected vessel. Next, the forward trajectories are clustered. In this manner, each cluster represents a mode of ship behavior, where each is comprised of similar trajectories. It is, therefore, of interest to identify the most likely mode of future ship behavior the selected vessel may belong to, such that a prediction of enhanced fidelity can be facilitated.

In the classification module, the trajectories identified in the clustering module are extended 10 min-

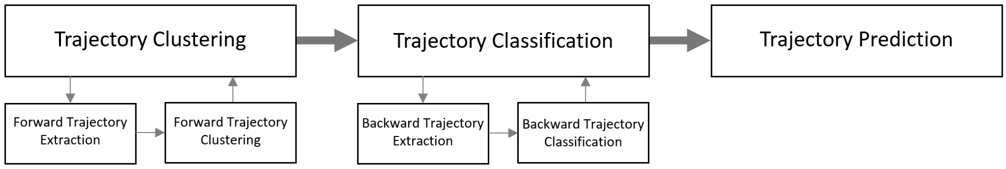

Figure 1: Method architecture.

utes into the past. This is referred to as a backward trajectory extraction. In this manner, each backward trajectory is an extension of one of the forward trajectories from the clustering module, and have corresponding class labels. By comparing the past 10 minute behavior of the selected vessel to the backward trajectories, the behavior can be compared, and used to classify the selected vessel to one of the clusters of future behavior.

In the prediction module, the subset of forward trajectories belonging to the cluster identified in the classification model are utilized to conduct a prediction. In this manner, only the specific behavior in the cluster is used to conduct the prediction. This should enhance the accuracy of the prediction compared to cases in which the trajectories diverge.

2.2. Trajectory Clustering Module

Machine learning can be split into two groups, supervised and unsupervised learning. Supervised learning deals with techniques where class labels are available, and one wishing to train an algorithm to correctly classify an unseen data point to a given class. Unsupervised learning, however, deals with data where the class labels are unavailable. In such a case, it is desirable to discover underlying groupings, or clusters, in the data. Clustering is, therefore, a form of unsupervised learning. In this study, the class labels for the extracted trajectories are unavailable, requiring the use of unsupervised learning. As such, clustering is investigated to discover groupings, or clusters, of historical ship trajectories that represent represent behavior modes that a selected vessel may belong to. This section covers the methodology utilized to cluster the historical trajectories.

2.2.1. Initial Clustering

The input to the algorithm is the initial state of a selected vessel, and is defined in (1). This state can be thought of as the current state of a target vessel, whose future trajectory is of interest to predict. Such parameters can be acquired from on board sensors e.g. radar, or from external sources e.g. AIS.

$$\mathbf{s}_0 \rightarrow [x_0, y_0, \chi_0, v_0, T_0] \tag{1}$$

It is of interest to identify similar vessels in the historical AIS database, i.e. data points with a high degree of similarity to $\mathbf{s}_0$. It can be argued that AIS data points similar to $\mathbf{s}_0$ will have a higher probability of having similar trajectories than dissimilar data points. It is, therefore, assumed that ships that were in a similar geographical location, with a similar course and speed over ground, will likely have behaved in a similar manner. As such, the trajectories of these vessels can be thought of as representing the distribution of the possible future behavior of the selected vessel. It is, therefore, assumed that these trajectories can be used to estimate the future behavior of the selected vessel. The discovery of such similar vessels is achieved via the initial clustering technique described in this section.

A matrix $\mathbf{Z}$ can be defined as the subset of spatial data in the AIS data set. The spatial data is converted from longitude and latitude values to UTM coordinates (x, y) prior to clustering. A rotational affine transformation can be defined to rotate $\mathbf{Z} = [x_z, y_z]$ by $\theta = \chi_0$ to $\mathbf{Z}' = [x_{z'}, y_{z'}]$. This transformation is defined in (2).

$$\mathbf{Z}' = \mathbf{R}\,\mathbf{Z}^T \tag{2}$$

Where $x_z \in \mathbb{R}$, $y_z \in \mathbb{R}$, $x_{z'} \in \mathbb{R}$, $y_{z'} \in \mathbb{R}$ and $\mathbf{R}$ is the rotation matrix defined as:

$$\mathbf{R} = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \tag{3}$$

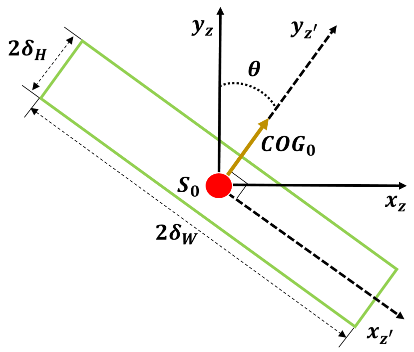

Figure 2: Initial cluster $\mathbf{C}_0$.

The new matrix, $\mathbf{Z}'$, will have a basis comprised of a vector in the direction of χ_0, and one orthogonal to χ_0. An initial cluster $\mathbf{C}_0$ is then created using data in the space spanned by these basis vectors in (4). This clustering operation results in a rectangular cluster $\mathbf{C}_0$ with a height of $2\delta_H$ and width $2\delta_W$ centered about $\mathbf{s}_0$ as illustrated in Fig. 2, which is adapted from that presented in Murray and Perera (2019). The cluster also only contains data points with similar χ and v values that were at a similar position to the selected vessel at some previous time point. The rectangular shape of the cluster orthogonal to χ_0 should capture most vessels that have similar trajectories to that of $\mathbf{s}_0$.

$$\mathbf{C}_0 = \{\mathbf{a}_i \in \mathbf{A} : (|x_{z'i} - x_{z'0}| \leq \delta_W \wedge |y_{z'i} - y_{z'0}| \leq \delta_H) \wedge (|\chi_i - \chi_0| \leq \chi_\delta \wedge |v_i - v_0| \leq v_\delta)\} \quad (4)$$

2.2.2. *Forward Trajectory Extraction*

Based on the initial cluster $\mathbf{C}_0$, unique instances of vessel trajectories are identified, given that multiple data points in $\mathbf{C}_0$ may belong to the same trajectory. Once unique trajectory instances have been identified, the nearest point of each trajectory to $\mathbf{s}_0$ in geographical space is defined as its initial point. The forward trajectories of all instances are then extracted from this point and a period of time into the future corresponding to the desired prediction horizon T_p. An additional time period, T_δ, is extracted to ensure sufficient data density for the trajectory prediction module at the culmination of the prediction. The trajectories belonging to $\mathbf{C}_0$ represent the possible behavior of the selected vessel, as their initial points have a high degree of similarity to $\mathbf{s}_0$. In other words, it is likely that the future trajectory of the selected vessel will be similar to one of the trajectories in $\mathbf{C}_0$.

2.2.3. *Trajectory Feature Generation*

Assuming that the trajectories in $\mathbf{C}_0$ represent the distribution of the possible future behavior of the selected vessel, it is desirable to discriminate between the various possibilities, i.e. discover groupings of behavior. In this sense, one wishes to cluster the trajectories into classes of behavior. To achieve this, each unique trajectory must be described by a set of features. The term feature in this case refers to an individual measurable parameter that describes the trajectory. Each trajectory is to be clustered in an unsupervised manner based on these features. As such, a trajectory feature vector is constructed comprising relevant parameters.

The first step in the generation of the feature vectors is to linearly interpolate each trajectory at 30 second intervals. This is done to generate higher density data, as well as provide a common time index with which the trajectories can be compared. The initial point of each trajectory is defined as T_0. Subsequent data points are, therefore, 30 seconds apart, starting at this point. In this manner, the trajectories can be directly compared at the same time instance relative to T_0. Using the interpolated data, each trajectory feature vector is constructed by flattening the matrix containing the positional and speed data (x, y, v) of the trajectory. If each trajectory is of length L, the resultant trajectory feature vector is defined as $\mathbf{f} \in \mathbb{R}^{3L\times 1}$. Utilizing the positional data, $\mathbf{f}$ will incorporate the shape of the trajectory and the inherent course alterations between data points. The speed of the vessel along the trajectory will also be inherent in the positional data. Nonetheless, the speed over ground values at each time instance were deemed relevant to include to enhance the information stored in each vector.

As mentioned, the objective of the module is to cluster the trajectories, and as such, the respective feature vectors should provide a basis for discriminating between the classes of behavior. In general, including as much information as possible, i.e. increasing the dimensionality of the feature vector, should enhance the discriminatory properties of the data set. This is true, but in a clustering setting may encounter issues relating to the curse of dimensionality (Bellman, 1961).

Clustering is based on grouping data points via some distance measure. Points that are closer together are more likely to be considered part of the same cluster. The curse of dimensionality in relation to clustering was discussed in Steinbach et al. (2004), where it was pointed out that a fixed number of data points will become increasingly sparse as the dimensionality increases. Data points can in a sense be lost in space as the dimensionality increases, as the distance between points with respect to a given dimension can be large. As a result, clustering data using standard techniques in a high dimensional space will degrade the results, as the algorithms are unable to find groupings in the data. One method to ameliorate this effect is to reduce the dimensionality of the data.

A common method for dimensionality reduction is the Karhunen-Loéve (KL) transform (Karhunen, 1946). The purpose of the transform is to attain uncorrelated features and is shown in (5). First, the set of all feature vectors is centered such that all features have mean zero within the set. Subsequently, the covariance matrix $\mathbf{\Sigma}$ of the set of all feature vectors is calculated. Matrix $\mathbf{E}$ consists of the eigenvectors of $\mathbf{\Sigma}$, and $\mathbf{\Lambda}$ is the eigenvalue matrix, where the relationship is shown in (6). (5) projects the feature vector $\mathbf{f}$ onto the space spanned by the eigenvectors of the covariance matrix. The covariance of the data inherently describes the correlation among the respective parameters. As such, the eigenvectors of the covariance matrix will describe the directions in which the data has the highest degree of variation orthogonal to each other.

$$\mathbf{x} = \mathbf{E}^T \mathbf{f} \tag{5}$$

Where $\mathbf{x} \in \mathbb{R}^{3L \times 1}$, $\mathbf{f} \in \mathbb{R}^{3L \times 1}$ and $\mathbf{E} \in \mathbb{R}^{3L \times 3L}$

$$\mathbf{\Sigma} = \mathbf{E} \Lambda \mathbf{E}^T \tag{6}$$

Where $\mathbf{\Sigma} \in \mathbb{R}^{3L \times 3L}$ and $\mathbf{\Lambda} \in \mathbb{R}^{3L \times 3L}$

In a high dimensional space, however, many of the eigenvectors will describe very little variation in the data. The KL-transform, therefore, projects $\mathbf{f}$ onto the subspace spanned by the l eigenvectors with the l largest eigenvalues in (7).

$$\mathbf{x} = \mathbf{E}_l^T \mathbf{f} \tag{7}$$

Where $\mathbf{x} \in \mathbb{R}^{l \times 1}$ and $\mathbf{E}_l \in \mathbb{R}^{3L \times l}$

This will inherently preserve the most important covariance information in the data whilst reducing the dimensionality to l. This may be abstract for the case of the trajectory feature vector, $\mathbf{f}$, as each dimension represents a position or speed value at a given time instance. Take for instance the case of a 30 minute prediction with five minutes added to allow for sufficient data density. The dimensionality of $\mathbf{f}$ will then be 210. The eigenvectors of $\mathbf{\Sigma}$ will point in the directions within this 210-dimensional space where there is a high degree of variation between the trajectories. As such, it is difficult to gain a direct physical interpretation of the eigenvectors, as the projection onto them represents a combination of multiple parameters. By choosing the l largest eigenvalues, one chooses the l directions where the variation in the data is greatest. When projecting the feature vectors onto the subspace spanned by the eigenvectors corresponding to the largest eigenvalues, one is in fact generating new features with a high degree of variation that can be used for further analysis.

In this study, the projection of $\mathbf{f}$ onto the eigenvectors corresponding to the three largest eigenvalues was chosen as a representation for each trajectory. Generally, the projection should retain at least 95 % of the variance in the data. This is evaluated by investigating the sum of the chosen eigenvalues over the sum of all eigenvalues (Hyvärinen, 2009). It was found that using the eigenvectors corresponding to the three largest eigenvalues fulfilled this requirement when evaluating the results. Additionally, a three-dimensional vector can easily be visualized when evaluating the performance of the clustering algorithm.

2.2.4. *Unsupervised Gaussian Mixture Model Clustering*

Using the reduced trajectory feature vectors generated via the KL-transform, the trajectories can be clustered. Depending on $\mathbf{s}_0$, the number of true clusters, i.e. classes, will vary. As such, a flexible clustering algorithm is required that can adapt to the data in each prediction. Unsupervised Gaussian Mixture Model Clustering was chosen for use in this study. A Gaussian Mixture Model (GMM) (Reynolds et al., 2000) is a flexible model that adapts to the underlying data. GMMs assume that a data set $\mathbf{X}$ consists of a mixture of M different Gaussian distributions. Each distribution has its own mean vector $\boldsymbol{\mu}_m$, covariance matrix $\boldsymbol{\Sigma}_m$ and prior distribution π_m. As such, each distribution will describe that particular class or cluster, i.e. class m. The class membership parameter, $\mathbf{z}_i$, is introduced for each data point $\mathbf{x}_i$ where:

$$\mathbf{z}_{ik} = \begin{cases} 1 & \text{if } k = m \\ 0 & \text{otherwise} \end{cases}$$

Where $\mathbf{z}_i \in \mathbb{R}^{M\times 1}$

The class conditional probability is shown in (8). The most likely model is estimated by maximizing the log-likelihood with respect to the various model parameters.

$$p(\mathbf{x}_i|\mathbf{z}_{im} = 1) \sim N(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) \tag{8}$$

The class membership of the trajectories is, however, unknown. As such, the Expectation Maximization (EM) algorithm is utilized to conduct the unsupervised GMM clustering. The GMM requires that a specified number of underlying models, M, is input. Based on this, the EM algorithm initializes all model parameters. A common method is to initialize all $\boldsymbol{\mu}_m$ as randomly chosen data points, the priors as $\pi_m = \frac{1}{M}$ and $\boldsymbol{\Sigma}_m = \mathbf{I}$. This initialization is unlikely to model the underlying data correctly. As such, the algorithm conducts what is known as the expectation step. In this step, the expected class membership $\langle \mathbf{z}_{im} \rangle$ is evaluated in (9), based on the current model parameters, $\boldsymbol{\Theta}$. All data points will, therefore, have updated class memberships based on the current model parameters..

$$\langle \mathbf{z}_{im} \rangle = \frac{p(\mathbf{x}_i|\mathbf{z}_{im} = 1; \boldsymbol{\Theta})\pi_m}{\sum_{k=1}^{M} p(\mathbf{x}_i|\mathbf{z}_{ik} = 1; \boldsymbol{\Theta})\pi_k} \tag{9}$$

The next step in the EM algorithm is known as the maximization step. In this step, the model parameters are updated based on the new distribution resulting from the expectation step. This is done by maximizing the log-likelihood with respect to $\boldsymbol{\Theta}$. The estimated parameters in the maximization step are calculated in (10), (11) and (12).

$$\hat{\boldsymbol{\mu}}_m = \frac{\sum_{i=1}^{N} \langle \mathbf{z}_{im} \rangle \mathbf{x}_i}{\sum_{i=1}^{N} \langle \mathbf{z}_{im} \rangle} \tag{10}$$

$$\hat{\boldsymbol{\Sigma}}_m = \frac{\sum_{i=1}^{N} \langle \mathbf{z}_{im} \rangle (\mathbf{x}_i - \boldsymbol{\mu}_m)(\mathbf{x}_i - \boldsymbol{\mu}_m)^T}{\sum_{i=1}^{N} \langle \mathbf{z}_{im} \rangle} \tag{11}$$

$$\hat{\pi}_m = \frac{\sum_{i=1}^{N} \langle \mathbf{z}_{im} \rangle}{N} \tag{12}$$

The EM algorithm now repeats, where the expected class memberships are updated, as well as the model parameters. The algorithm is in a sense adapting to the data, where the most likely distribution of the data is discovered. This iterative process continues in a loop until a stopping criteria is met. One common stopping criteria is the convergence of the total log-likelihood. Alternatively, one can terminate the algorithm if there is little to no change in the model parameters, i.e. the parameters themselves converge. The parameter convergence criteria was utilized in this study. Often times, the EM algorithm can haves issues with convergence, due to poor initialization. To avoid divergence issues, a technique is employed where a number of random initializations are run for a number of iterations. The best run, i.e. the run with the greatest log-likelihood score, is then chosen and run for further iterations. The mixture model will, upon convergence, consist of M distinct Gaussian distributions which describe the class conditional probabilities, $p(\mathbf{x}|c_m)$, of the data, along with an associated prior distribution, π_m. The posterior probability $p(c_m|\mathbf{x})$ can be found via Bayes Rule in (13) using the resultant conditional probabilities and priors from the algorithm.

$$p(c_m|\mathbf{x}) = \frac{p(\mathbf{x}|c_m)\pi_m}{p(\mathbf{x})} \tag{13}$$

$$p(c_m|\mathbf{x}) > p(c_j|\mathbf{x}) \ \forall\ j \neq m, j = 1...M \tag{14}$$

Clustering of the dataset is then conducted via Bayesian classification, where each feature , $\mathbf{x}_i$, is classified to class m according to (14). However, the number of underlying classes, M, is as previously mentioned unknown. In order to determine the most likely number of clusters, the Bayesian Information Criterion (BIC) (Schwarz et al., 1978) defined in (15), is utilized.

$$BIC = -2LL(\mathbf{\Theta}_M) + K_M ln(N) \tag{15}$$

For a GMM with M underlying distributions, $LL(\mathbf{\Theta}_M)$ is the total log-likelihood function computed at the optimum, K_M the number of free parameters in the mixture model, and N the number of data points. The EM algorithm can be run for various GMMs by altering M. By calculating the BIC for each resultant GMM, the most likely GMM is that with the lowest BIC. This is due to it having the highest likelihood and least complexity. In this study, it was assumed that there will be no more than 20 unique clusters in the trajectory data, and the BIC was, therefore, evaluated for values of M up to 20.

This process discovers the best GMM to fit the data and provides the number of possible routes, or trajectory behavior modes, a selected vessel may belong to. By classifying all the extracted forward trajectories, class labels can assigned. These labels are used for further analysis in the subsequent modules.

2.3. Trajectory Classification Module

The trajectory clustering module has now clustered all trajectories present in $\mathbf{C}_0$ to M classes. Each class represents a group of trajectories that have a high degree of similarity. As such, each class represents a possible future route, or behavior mode, the selected vessel may belong to. It is now of interest to classify the selected vessel to the most likely class of the M possibilities. In this sense, an estimate of the distribution of the possible future behavior of the selected vessel can be made. Using the data in the class of trajectory behavior, a trajectory prediction can be made. This section presents the method utilized to achieve such a classification.

2.3.1. Backward Trajectory Extraction

One possible method to conduct the aforementioned classification is to utilize the current vessel state, $\mathbf{s}_0$, and compare it to the data points in $\mathbf{C}_0$. This, however, will have limited predictive power, as the classification will be based solely on one time instance of the selected vessel. An alternative approach is, therefore, suggested, where the previous 10 minutes of the selected vessel's trajectory are be compared to the previous 10 minutes of data for all trajectories in $\mathbf{C}_0$. This in a sense is the inverse of the forward trajectory extraction process described in Sec. 2.2.3. Instead of extracting the trajectories from T_0 and for instance 30 minutes into the future, the past trajectories are extracted from the same initial point, i.e. from T_0, and 10 minutes into the past from that time instance. It is assumed that at least 10 minutes of behavior for the selected vessel should be available via the on board sensors of the own ship, or via external sources e.g. AIS. The method is otherwise identical to that described in Sec. 2.2.3. All the backward trajectories extracted from $\mathbf{C}_0$ will have the same labels as those determined by the clustering technique in Sec. 2.2.4. As such, a labeled data set is available that can be used to classify the observed trajectory of selected vessel.

2.3.2. Optimal Feature Generation

Each backward trajectory feature vector is represented by flattening the matrix containing all position and speed over ground data, in the same manner as for the forward trajectories in Sec. 2.2.3. This will result in a vector $\mathbf{f} \in \mathbb{R}^{3L\times 1}$. In the case of a 10 minute trajectory this will be a 60-dimensional space within which the classification must take place. This can be a challenging task, as it is likely that the features are quite similar, given that the vessels in $\mathbf{C}_0$ generally will have similar trajectories for the past 10 minutes.

To improve the classification accuracy, Linear Discriminant Analysis (LDA) (Fischer, 1936) is utilized. LDA provides a method to generate features with optimal separation between classes in a supervised manner. Using the class separability measure J_3 in (16), one can optimize a transformation such that features are generated to optimize class separability.

$$J_3 = trace\{\mathbf{S}_w^{-1}\mathbf{S}_m\} \tag{16}$$

$\mathbf{S}_m$ is the mixture scatter matrix defined as $\mathbf{S}_m =$

$\mathbf{S}_w + S_b$, where $\mathbf{S}_w$ is the within-class scatter matrix and $\mathbf{S}_b$ the between-class scatter matrix. $\mathbf{S}_w$ and $\mathbf{S}_b$ are defined in (17) and (19) respectively. $\mathbf{S}_w$ describes how compact the data within each class is, whilst $\mathbf{S}_b$ describes how spread out each class is with respect to the global mean, $\boldsymbol{\mu}_g$. In a classification setting, one wishes to minimize the trace of $\mathbf{S}_w$, i.e. data are more compact within each class, and maximize the trace of $\mathbf{S}_b$, i.e. the classes are more spread out. This corresponds to maximizing the class separation criterion J_3.

$$\mathbf{S}_w = \sum_{m=1}^{M} \pi_m \boldsymbol{\Sigma}_m \tag{17}$$

$$\boldsymbol{\mu}_g = \sum_{m=1}^{M} \pi_m \boldsymbol{\mu}_m \tag{18}$$

$$\mathbf{S}_b = \sum_{m=1}^{M} \pi_m (\boldsymbol{\mu}_m - \boldsymbol{\mu}_0)(\boldsymbol{\mu}_m - \boldsymbol{\mu}_0)^T \tag{19}$$

It is desirable to find a transformation $\mathbf{x} = \mathbf{A}^T \mathbf{f}$ such that J_3 is maximized in the transformed space. The optimal transformation with respect to class separability is found to be $\mathbf{A} = \mathbf{E}$, where $\mathbf{E}$ is the matrix of eigenvectors of $\mathbf{S}_w^{-1}\mathbf{S}_b$ in the original vector space. This relationship is shown in (21), where $\boldsymbol{\Lambda}$ is the corresponding diagonal eigenvalue matrix. The transformation is shown in (20). However, $\mathbf{S}_b$ is of rank $M-1$, and correspondingly $\mathbf{S}_w^{-1}\mathbf{S}_b$ is also of rank $M-1$. As such, there will be $M-1$ nonzero eigenvalues. (20) will, therefore, project $\mathbf{f}$ onto the subspace spanned by the l largest eigenvectors in a similar manner to the KL-transform. If $l = M-1$, optimality with respect to J_3 will be preserved. Further dimensionality reduction can still be conducted by choosing a value $l < M-1$. This will, however, be a sub-optimal solution. Further details on LDA can be found in Theodoridis and Koutroumbas (2009).

$$\mathbf{x} = \mathbf{E}^T \mathbf{f} \tag{20}$$

Where $\mathbf{x} \in \mathbb{R}^{3L \times 1}$, $\mathbf{f} \in \mathbb{R}^{3L \times 1}$ and $E \in \mathbb{R}^{3L \times l}$

$$\mathbf{S}_w^{-1}\mathbf{S}_b = \mathbf{E}\boldsymbol{\Lambda}\mathbf{E}^T \tag{21}$$

Where $\mathbf{S}_w^{-1}\mathbf{S}_b \in \mathbb{R}^{3L \times 3L}$ and $\boldsymbol{\Lambda} \in \mathbb{R}^{l \times l}$

2.3.3. Classification

Despite utilizing the optimal features described in Sec. 2.3.2, the classification task is highly nonlinear, and likely with significant overlap between classes in most cases. This is due to the high degree of similarity between the past trajectories. As a result, the k-Nearest Neighbor (kNN) classifier (Dasarathy, 1991) is utilized due to its nonlinear predictive power.

Given a data point $\mathbf{x_0}$, the kNN classifier will measure the distance to all other data points, $\mathbf{x}_i$, in the dataset $\mathbf{X}$ using the Euclidean distance as shown in (22).

$$d_i = ||\mathbf{x}_i - \mathbf{x}_0||_2 \tag{22}$$

The kNN classifier will then identify the k nearest data points using distance measures from (22). Based on this subset of data, the algorithm then identifies the class with the most data points in the subset, and classifies $\mathbf{x_0}$ to the majority class.

In this study, $\mathbf{x}_0$ is the projection of the backward trajectory feature vector $\mathbf{f}_0$ of the selected vessel onto the LDA subspace according to (20). The kNN classification is then conducted in the LDA subspace, where the k nearest trajectories, i.e. most similar, are found, and the majority class is defined as the class, j, of the selected vessel. The corresponding subset for class j, $\mathbf{A}_j \subset \mathbf{A}$, represents the set of AIS data points that belong to the trajectories in this class.

2.4. Trajectory Prediction Module

Once the past trajectory of the selected vessel has been classified to one of the classes, a trajectory prediction can be conducted with respect to that class. The input data to this module are all trajectory data belonging to the class determined in Sec. 2.3.3, i.e. $\mathbf{A}_j$. This approach assumes that the classification is accurate, and as such only predicts one unique trajectory.

The algorithm utilized for the trajectory prediction is adapted from that presented in Murray and Perera (2018). The algorithm is based on the method outlined in Hexeberg et al. (2017), where a Single Point Neighbor Search Method was presented to predict vessel trajectories based on historical AIS data.

Given the initial state of the selected vessel, $\mathbf{s}_0$, the prediction algorithm estimates the future states of the selected vessel. This is an iterative process where the state in the k^{th} iteration is defined in (23).

$$\hat{\mathbf{s}}_k \rightarrow [\hat{x}_k, \hat{y}_k, \hat{\chi}_k, \hat{v}_k, \hat{T}_k] \tag{23}$$

The estimated future position in state k, i.e. $[\hat{x}_k, \hat{y}_k]$, is estimated given the parameters in state $\hat{\mathbf{s}}_{k-1}$ as a distance Δ_L from $[\hat{x}_{k-1}, \hat{y}_{k-1}]$, in the direction of $\hat{\chi}_{k-1}$. The time parameter $\hat{T}_k$ is then updated according to (24).

$$\hat{T}_k = \hat{T}_{k-1} + \frac{\Delta L}{\hat{v}_{k-1}} \tag{24}$$

Once the position parameters $[\hat{x}_k, \hat{y}_k]$ are updated, $\hat{\chi}_k$ and $\hat{v}_k$ are updated using a circular distance based clustering technique. A cluster $\mathbf{C}_k$ can be defined according to (25) where $\mathbf{p}_i$ is an arbitrary vessel position, and $\mathbf{q}_k$ is the selected vessel position at $\hat{\mathbf{s}}_k$. The clustering is conducted on the set of data points in the subset of AIS data that corresponds to the classified class, i.e. $\mathbf{A}_j$. $\mathbf{C}_k$ will, therefore, comprise the data points within a radius r from the predicted position.

$$\mathbf{C}_k = \{\mathbf{a}_i \in \mathbf{A}_j \, : \, ||\mathbf{p}_i - \mathbf{q}_k|| \leq r\} \tag{25}$$

$\hat{\chi}_k$ and $\hat{v}_k$ are estimated as the median values of the data points in cluster $\mathbf{C}_k$ according to (26) and (27). The median values were chosen as opposed to the mean as that they are less sensitive to outliers.

$$\hat{\chi}_k = median(\chi_i \in \mathbf{C}_k) \tag{26}$$

$$\hat{v}_k = median(v_i \in \mathbf{C}_k) \tag{27}$$

This iterative process continues until the desired prediction horizon is reached, i.e. $\hat{T}_k \geq T_p$. The set of all estimated states will constitute the predicted trajectory of the selected vessel. The predicted trajectory is subsequently linearly interpolated at 30 second intervals for comparative analysis. The method is illustrated in Fig. 3, adapted from that presented in Murray and Perera (2019).

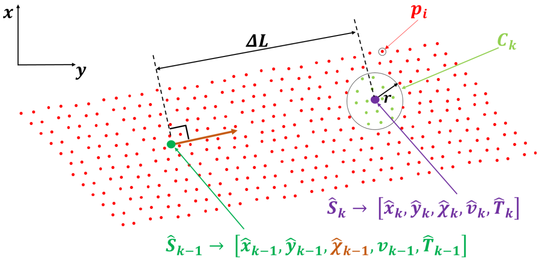

Figure 3: Illustration of trajectory prediction technique.

3. Results and Discussion

In this section, the results of a case study to predict the future trajectory of a selected vessel are presented and discussed using the method in this study. First, the historical AIS data set utilized in the study is presented. Next, the results of the trajectory clustering module are outlined. Subsequently, the results from the trajectory classification module are presented. Then, the results of the trajectory prediction module are discussed. Finally, the prediction accuracy of the approach is presented, where the classification and position accuracy are discussed.

3.1. Historical AIS Data Set

In this study, a data set corresponding to one year of historical AIS data from January 1^{st} 2017 to January 1^{st} 2018 for the region around the city of Tromsø, Norway was investigated. This data set corresponded to approximately 15 million AIS messages, made available by the Norwegian Coastal Administration. The ship behavior in the region relates to that of inland waterways and around ports. As such, the region can be considered to represent more complex ship traffic that that of the open ocean, where more linear behavior is likely observed. As a result, the region is considered to be relevant to test the methodology outlined in this study.

AIS messages contain a variety of information including static, dynamic and voyage related information. The included information is summarized in Tab. 1. However, not all parameters are utilized in the study. The only parameters used were the position converted to UTM coordinates, (x, y) speed over ground, v, course over ground, χ, timestamp, t and MMSI number.

To evaluate the performance of the method, 100

Static	Dynamic	Voyage Related
MMSI	Navigational status	Draught
Call-sign	Latitude position	Hazardous cargo
Name	Longitude position	Destination
IMO number	Timestamp	Estimated time of arrival
Length	Course over ground	
Beam	Speed over ground	
Ship type	Heading	
Location of antenna	Rate of turn	

Table 1
AIS data.

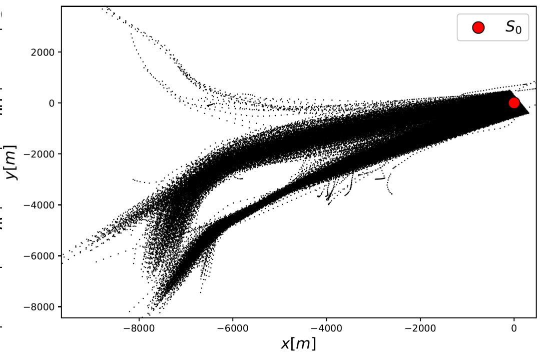

Figure 4: Illustration of extracted forward trajectories.

different trajectories were used as a test set. To facilitate this, 100 data points were randomly selected to represent the initial state of 100 selected vessels. The case study predicted the future 30 minute trajectory for each selected vessel, where the predictions were validated using the true trajectories stored in the historical AIS data. The performance of each module was investigated, as well as the overall performance of the method in predicting future ship trajectories.

3.1.1. Reproducibility

In order to reproduce the results of this study, one must have access to a historical AIS database that contains the data for the region surrounding Tromsø, Norway for the year of 2017. Once this is attained, test trajectories can be extracted, and the method applied to predict these trajectories. As such, the method can be validated by any interested party with access to the historical AIS data for the region.

3.2. Trajectory Clustering Module

In this subsection, the results of the trajectory clustering module are presented. The results presented are for a specific test case to illustrate the performance of the method.

3.2.1. Extracted Trajectories

As outlined in Sec. 2.2.2, all trajectories present in the initial cluster $\mathbf{C}_0$ centered about the initial vessel state, $\mathbf{s}_0$ were extracted. An example of the interpolated extracted trajectories is visualized in Fig. 4. The illustrated position data are defined with respect to $\mathbf{s}_0$ (i.e. $[x_0, y_0] = [0, 0]$) to more easily visualize the distances involved. To the human eye, it is evident that there are two main routes the vessel may follow, with a few outliers. This information may also be what a navigator on the bridge might be aware of, and base his future decisions upon.

3.2.2. Clustering Results

The first phase of the clustering technique is to reduce the dimensionality according to (7). Subsequently, a GMM is fit to the projection of the trajectory data in the subspace spanned by the three eigenvectors with the largest eigenvalues, as outlined in Sec. 2.2.4. This technique was found to be quite effective in generating new features with a high degree of variation between data points. The result of the clustering of the extracted forward trajectories in Fig. 4 is visualized in Fig. 5. This figure illustrates the clusters in the reduced subspace. Using the discovered classes, the labeled trajectories are visualized in Fig. 6. The method in this example has discovered eight unique clusters. This implies that the vessel may have one of eight behavior modes. It is evident that the algorithm has primarily focused on differences in the spatial aspects of the trajectories, i.e. the upper and lower routes. However, the results indicate that the algorithm also discovers sub-routes within the main routes. These indicate vessels traveling along the prevailing route at various speeds. As such, the algorithm is in fact discovering behavior modes within the data. The method, therefore,

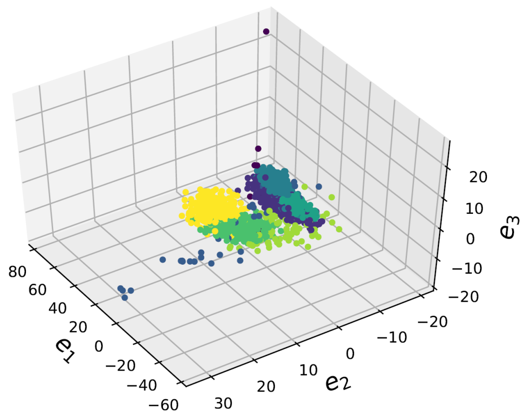

Figure 5: Clustering results in KL-subspace.

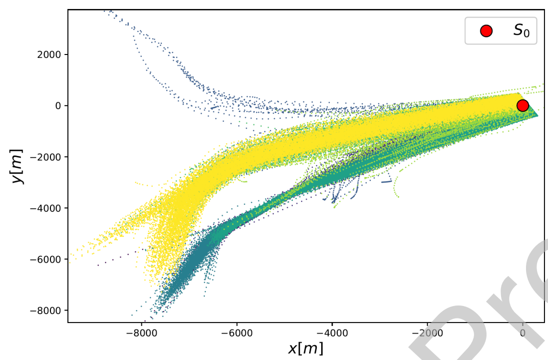

Figure 6: Labeled forward trajectories.

is effective in regions with more complex traffic. In regions where vessels have a high degree of regularity, such advanced clustering will likely not be as necessary.

The results indicate that KL-transform effectively stores the most important information from the trajectories by projecting a 210-dimensional vector to a 3-dimensional vector. This data compression subsequently allows for effective clustering in the lower-dimensional subspace, where multiple trajectory groupings can be discovered, providing a more accurate dataset upon which a trajectory prediction can be conducted. Given that the true future trajectory of the selected vessel is also available in the historical data, it can be classified to one of the clusters via the GMM. This provides the true class of the selected vessel for subsequent accuracy analysis.

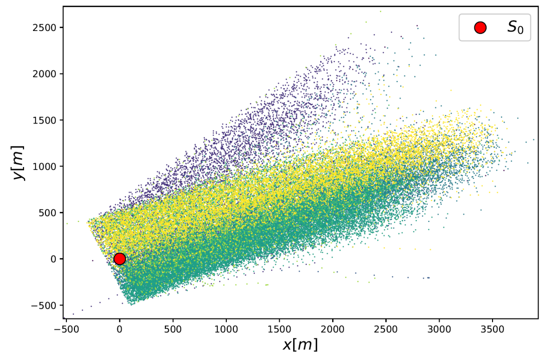

Figure 7: Backward trajectories with labels from the corresponding forward trajectories.

3.3. Trajectory Classification Module

In this subsection, the results of the trajectory classification module are presented. The results presented are for a specific test case to illustrate the performance of the method.

3.3.1. *Optimal Feature Representation*

In this phase of the method, the backward trajectories of all vessels present in the initial cluster, $\mathbf{C}_0$, are extracted. These trajectories are visualized for the example in Fig. 7 with labels from the corresponding forward trajectories. The motivation now is to classify the past trajectory of the selected vessel to one of the classes. Using (20), the trajectory features are projected onto the LDA-subspace. In this subspace, the trajectories are optimally separated, making it easier for classification. The projection onto the three largest components for the previous example is visualized in Fig. 8.

3.3.2. *Classification*

The clusters in this study have a significant overlap in many cases, as seen in the case presented in Fig. 8. The data points in this figure are representations of the trajectories in Fig. 7, generated via LDA. As evident from Figure 7, many of the backward trajectories are quite similar across classes. Discriminating between classes is, therefore, challenging, even when applying LDA to generate more optimal representations.

As a result, it was found during the study that other classifiers e.g. Support Vector Machines had degraded performance. For the case of kNN classifiers, however, a local estimate in the region of

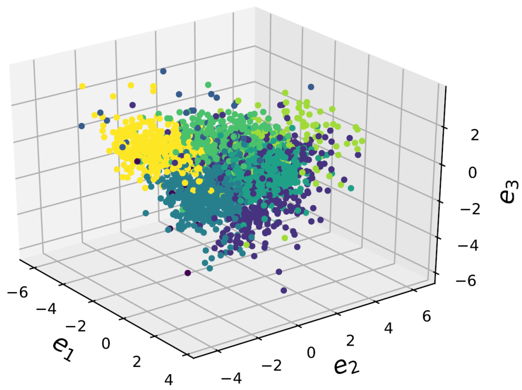

Figure 8: LDA projection of backward trajectory data.

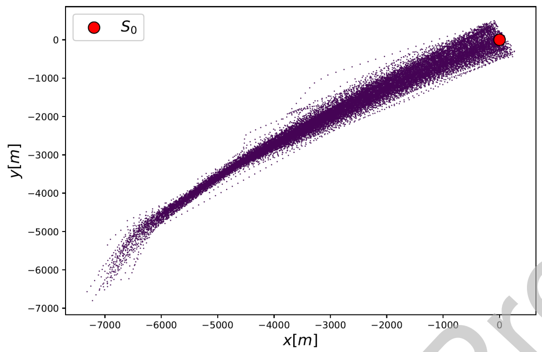

Figure 9: Forward trajectories in the cluster corresponding to the class prediction.

interest is utilized for the classification. Decision boundaries are, therefore, not utilized. Such decision boundaries will result in many misclassifications due to the degree of overlap between classes. By applying a kNN classifier, however, the most likely class can be found based on the neighborhood surrounding the data point, even in cases with high overlap.

Using a sensitivity analysis with respect to the k-parameter, it was found that the best results were attained for a value of k=7 in this study. As a result, a kNN classifier with $k = 7$, was used to classify the projection of the backward trajectory of the selected vessel to one of the clusters. This resulted in the selected vessel being classified to the purple class previously illustrated. The extracted forward trajectory data corresponding to this class is illustrated in Fig. 9.

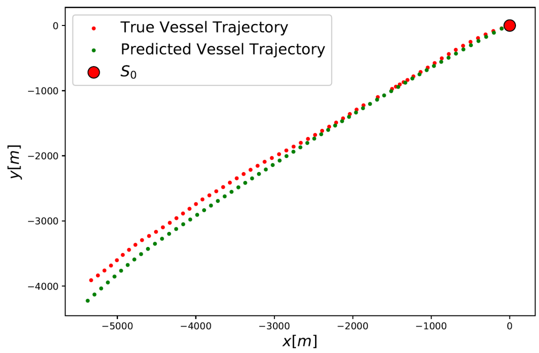

Figure 10: Trajectory prediction.

3.4. Trajectory prediction module

Using the data visualized in Fig. 9, a prediction can be conducted utilizing the methodology outlined in Sec. 2.4. Fig. 10 illustrates the resultant prediction for the previous example. For this case, the prediction appears to closely correspond to the true vessel trajectory.

3.5. Prediction Accuracy

In order to evaluate the overall performance of the method, 100 random data points were chosen from the AIS data set. Each data point was defined as the initial state of a selected vessel, i.e. $\mathbf{s}_0$. The method outlined in Sec. 2 was then run on each selected vessel to predict its future trajectory. In this section, the accuracy of the classification module and the position error of the resultant trajectory predictions are evaluated.

3.5.1. *Classification Accuracy*

The true class label of the selected vessel, i.e. the ground truth, was evaluated using the fitted GMM for each tested vessel. The predicted classes for all vessels were then compared with the ground truth, and an overall classification accuracy calculated. It was found that for the 100 cases tested in this study, the classification accuracy was of 70 %. This indicates that the features generated via LDA from the backward trajectories provided a basis to correctly classify 70 of the 100 tested vessels.

3.5.2. *Position Accuracy*

The position accuracy of the trajectory predictions was also investigated. The accuracy was evaluated as a function of time, where the distance between the true and predicted position of the selected vessel define the error. The position error was calculated for three cases; the overall error for all vessels, the error for incorrectly classified vessels, and the error for correctly classified vessels. Given that the true trajectories of the selected vessels are of various lengths, the position error is evaluated as a percentage of the true distance traveled for each time instance. The distance traveled for each selected vessel was estimated as the sum of trajectory segments extracted from the true trajectory.

The median position errors for all cases are illustrated in Fig. 11. The median error was chosen for presentation as opposed to the root mean squared error due to the sensitivity of the root mean squared error to outliers. It is clear that the error is significantly higher for the incorrectly classified vessels. However, for those vessels which are classified to the correct class, the median error is quite reasonable with a value of approximately 4 % of the true distance traveled for a 30 minute prediction. The error appears to increase rather linearly. This is to be expected, as errors will accumulate as a function of time.

The position error of the incorrectly classified vessels is also investigated, as the method incorrectly classifies 30 % of the vessels, and as such, will have a corresponding performance in these cases. Fig. 12 illustrates an example of a trajectory prediction when the selected vessel was incorrectly classified. It is evident that the subsequent error can grow to be quite high for the case of a 30 minute prediction. For the case of the incorrectly classified vessels in Fig. 11, a sudden dip is observed around a prediction horizon of 15 minutes. This effect is not observed for the correctly classified vessels. It is likely due to the nonlinear nature of many of the ship trajectories. At this point, certain ships that are predicted to travel along an incorrect route, turn and approach the true route, causing the position error to decrease for a short period of time, before once again linearly increasing.

In order to reduce the error associated with the incorrect classifications, one needs to improve the classification module to provide either a better representation to conduct classification on, or utilize another classifier. Additionally, anomalies can be filtered out as they will have a degrading effect on the classification. The error with respect to the correctly classified predictions is, therefore, of greater interest for further investigation. Fig. 13 illustrates the box plots for the position error at five minute intervals. The horizontal green line illustrates the median error. It appears that the lower 50 % of the predictions are rather tightly bounded, whilst the upper 50 % have a higher variance. The variance of the error also increases significantly as a function of time. This is to be expected, as the predictions are dependent on both the speed estimates, as well as the degree of variation within the cluster.

Nonetheless, the degree of variance observed for the 30 minute predictions is quite high for the upper quartiles. The correctly classified vessels with poor predictions were, therefore, investigated. Fig. 14 visualizes one such case. It appears that the vessel has been classified correctly, as the predicted and true trajectories are similar at first. However, approximately half way along the predicted trajec-

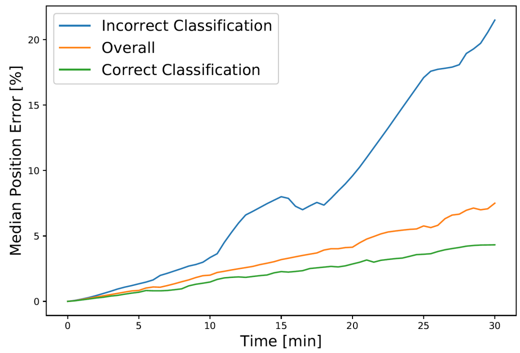

Figure 11: Median position error of trajectory predictions evaluated as a percent of the distance traveled.

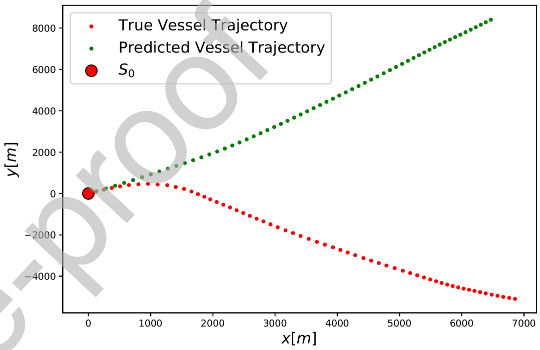

Figure 12: Incorrectly classified trajectory prediction.

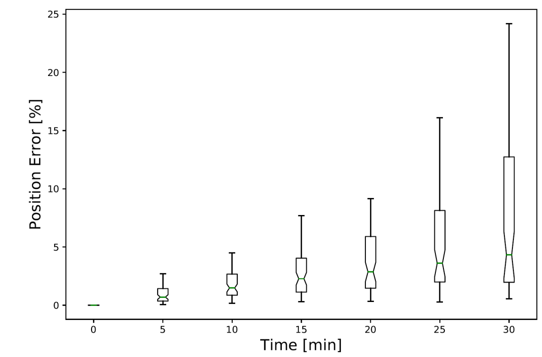

Figure 13: Prediction error for correctly classified vessels evaluated as a percent of the distance traveled.

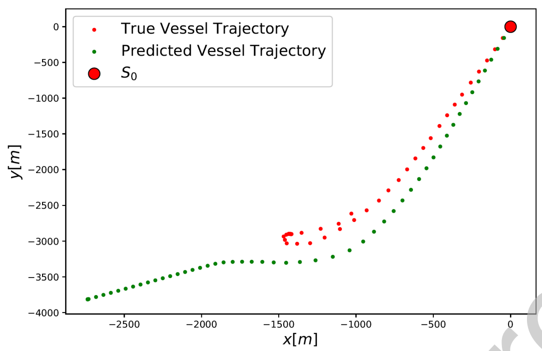

Figure 14: Correctly classified trajectory with high position error.

tory, the true trajectory of the vessel stops and turns around. Such irregular ship trajectories are difficult to predict, and the algorithm is unable to identify and recreate such patterns. These irregularities are the source of much of the high positional error illustrated in Fig. 13. The algorithm is, nonetheless, effective in predicting what can be considered regular ship trajectories, represented by the lower quartiles in Fig. 13. In general, it appears that the approach yields successful results with respect to predicting the future trajectory of a selected vessel.

4. Conclusion and Further Work

This study presents a method to provide system intelligence to future autonomous ships such that they can attain a high level of maritime situation awareness. This is facilitated through the use of historical AIS data and machine learning. Relevant trajectories are extracted from historical AIS data, and commonalities in the data are discovered via Gaussian Mixture Model clustering. These clusters represent modes of historical ship behavior. When predicting the future trajectory of a target vessel, it is likely that its future behavior will belong to one of these modes. Therefore, the observed behavior of a selected vessel is classified to one the modes to improve the fidelity of a trajectory prediction. Such high fidelity predictions can then be used to aid in collision avoidance.

Assuming that the previous behavior of the selected vessel is known, the method has a high classification accuracy. The results indicate that the use of Linear Discriminant Analysis provides a more optimal basis for classification. However, if the previous behavior is unknown, the classification accuracy will likely be degraded.

The results for the trajectory predictions indicate that the method was able to successfully predict the future trajectory of a selected vessel, with low error. For incorrectly vessels, however, the performance was degraded. This is to be expected, given that predictions are conducted with respect to incorrect behavior modes in these cases.

For the cases investigated in this study, correctly classified vessels had low prediction error for time horizons up to 30 minutes. The median error value was approximately 4 % of the true distance traveled after 30 minutes. This is likely aided by the direct relationship between data points in the clusters utilized in the predictions, as well as the ability to discover ship behavior modes that match the selected vessel. Certain vessels, however, had anomalous behavior, which the method was unable to accurately predict.

The method presented in this study is generic, and can be applied to any geographical region, given sufficient density of the historical AIS data. As a result, the algorithm can be implemented in a generic form on any vessel, and will run on the raw AIS data for that region. Seeing as the method is data-driven, the amount of data available will enhance the results. The accuracy of the predictions will also be location specific, as the number of possible behavior modes that exist will vary. The method will likely have better performance in open waterways with fewer possible routes, and a generally high degree of regularity in ship behavior compared to more complex waterways such as coastal regions and ports. The sensitivity of location has, however, not been investigated. Nonetheless, the ability of the method to discover behavior modes in the historical data will improve the performance in complex waterways compared to other methods. The predictions are also conducted without considering the prevailing weather conditions. These will likely have a significant effect of the behavior of a vessel, and should be included in the prediction method.

Further work will include enhancing the classification accuracy of the method, as well as including weather parameters into the data set. The classification accuracy can likely be further increased by using more advanced architectures e.g. kernel support vector machines (Murty and Raghava, 2016). Alternative trajectory prediction methodologies applying additional machine learning techniques, e.g. deep leaning, will also be investigated to further enhance the predictions. It is also vital to connect the trajectory predictions to existing collision avoidance frameworks and regulations. This will be addressed in the future, where such predictions will be applied in a collision avoidance setting.

Acknowledgments

An initial version of this paper was presented at the 38th International Conference on Ocean, Offshore and Arctic Engineering (OMAE 2019), June, 2019, Glasgow, Scotland. This work was supported by the Norwegian Ministry of Education and Research and the MARKOM2020 project, a development project for maritime competence established by the Norwegian Ministry of Education and Research in cooperation with the Norwegian Ministry of Trade, Industry and Fisheries. The authors would also like to express gratitude to the Norwegian Coastal Administration for providing access to their AIS database.

Declaration of Competing Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Aarsæther, K.G., Moan, T., 2009. Estimating Navigation Patterns from AIS. Journal of Navigation 62, 587–607. doi:10.1017/S0373463309990129.

Bellman, R.E., 1961. Adaptive control processes: a guided tour. volume 2045. Princeton university press.

Chan, C.Y., 2017. Advancements, prospects, and impacts of automated driving systems. International Journal of Transportation Science and Technology 6, 208–216. doi:10.1016/J.IJTST.2017.07.008.

Dalsnes, B.R., Hexeberg, S., Flåten, A.L., Eriksen, B.O.H., Brekke, E.F., 2018. The neighbor course distribution method with gaussian mixture models for ais-based vessel trajectory prediction, in: 2018 21st International Conference on Information Fusion (FUSION), IEEE. pp. 580–587.

Dasarathy, B.V., 1991. Nearest neighbor (NN) norms : NN pattern classification techniques. IEEE Computer Society Press, Los Alamitos, California.

Endsley, M.R., 1995. Toward a Theory of Situation Awareness in Dynamic Systems. Human Factors: The Journal of the Human Factors and Ergonomics Society 37, 32–64. doi:10.1518/001872095779049543.

Endsley, M.R., Bolte, B., Jones, D.G., 2003. Designing for situation awareness : an approach to user-centered design. Taylor & Francis.

Fischer, R.A., 1936. The use of multiple measurements in taxonomic problems. Annals of Eugenics 7, 179–188. doi:10.1111/j.1469-1809.1936.tb02137.x.

Fujii, Y., Tanaka, K., 1971. Traffic Capacity. Journal of Navigation 24, 543–552. doi:10.1017/S0373463300022384.

Goodwin, E.M., 1975. A Statistical Study of Ship Domains. Journal of Navigation 28, 328–344. doi:10.1017/S0373463300041230.

Hermann, M., Pentek, T., Otto, B., 2016. Design Principles for Industrie 4.0 Scenarios, in: 2016 49th Hawaii International Conference on System Sciences (HICSS), IEEE. pp. 3928–3937. doi:10.1109/HICSS.2016.488.

Hexeberg, S., Flaten, A.L., Eriksen, B.O.H., Brekke, E.F., 2017. AIS-based vessel trajectory prediction, in: 2017 20th International Conference on Information Fusion (FUSION), IEEE. doi:10.23919/ICIF.2017.8009762.

Hyvärinen, A., 2009. Principal component analysis. URL: http://www.mv.helsinki.fi/home/amoaning/movies/uml/pca{_}handout.pdf.

Karhunen, K., 1946. Zur spektraltheorie stochastischer prozesse. Annales Academiae Scientiarum Fennicae 37.

Laxhammar, R., Falkman, G., Sviestins, E., 2009. Anomaly detection in sea traffic-a comparison of the gaussian mixture model and the kernel density estimator, in: Information Fusion, 2009. FUSION'09. 12th International Conference on, IEEE. pp. 756–763.

Levander, O., 2017. Autonomous ships on the high seas. IEEE Spectrum 54, 26–31.

Liu, Z., Wu, Z., Zheng, Z., 2019. A novel framework for regional collision risk identification based on AIS data. Applied Ocean Research 89, 261–272. doi:10.1016/j.apor.2019.05.020.

Mazzarella, F., Arguedas, V.F., Vespe, M., 2015. Knowledge-based vessel position prediction using historical AIS data, in: 2015 Sensor Data Fusion: Trends, Solutions, Applications (SDF), IEEE. pp. 1–6. doi:10.1109/SDF.2015.7347707.

Murray, B., Perera, L.P., 2018. A Data-Driven Approach to Vessel Trajectory Prediction for Safe Autonomous Ship Operations, in: Proceedings of The Thirteenth International Conference on Digital Information Management (ICDIM), Berlin,. pp. 240–247. doi:10.1109/ICDIM.2018.8847003.

Murray, B., Perera, L.P., 2019. An AIS-Based Multiple Trajectory Prediction Approach for Collision Avoidance in Future Vessels, in: 38th International Conference on Ocean, Offshore & Arctic Engineering (OMAE 2019), Glasgow, Scotland. doi:10.1115/OMAE2019-95963.

Murray, B., Perera, L.P., 2020. A dual linear autoencoder approach for vessel trajectory prediction using historical AIS data. Ocean Engineering 209, 107478. doi:10.1016/j.oceaneng.2020.107478.

Murty, M.N., Raghava, R., 2016. Kernel-Based SVM. Springer International Publishing, Cham. pp. 57–67. doi:10.1007/978-3-319-41063-0_5.

Pallotta, G., Horn, S., Braca, P., Bryan, K., 2014. Context-Enhanced Vessel Prediction Based On Ornstein-Uhlenbeck Processes Using Historical AIS Traffic Patterns : Real-World Experimental Results. Information Fusion (FUSION), 2014 17th International Conference on , 1–7.

Pallotta, G., Vespe, M., Bryan, K., 2013. Vessel Pattern Knowledge Discovery from AIS Data: A Framework for Anomaly Detection and Route Prediction. Entropy 15, 2218–2245. doi:10.3390/e15062218.

Perera, L.P., Carvalho, J.P., Guedes Soares, C., 2010. Autonomous guidance and navigation based on the COLREGs rules and regulations of collision avoidance. Advanced Ship Design for Pollution Prevention , 205–216doi:10.1201/b10565-26.

Perera, L.P., Carvalho, J.P., Guedes Soares, C., 2011. Fuzzy logic based decision making system for collision avoidance of ocean navigation under critical collision conditions. Journal of Marine Science and Technology 16, 84–99. doi:10.1007/s00773-010-0106-x.

Perera, L.P., Ferrari, V., Santos, F.P., Hinostroza, M.A., Guedes Soares, C., 2015. Experimental Evaluations on Ship Autonomous Navigation and Collision Avoidance by Intelligent Guidance. IEEE Journal of Oceanic Engineering 40, 374–387. doi:10.1109/JOE.2014.2304793.

Perera, L.P., Oliveira, P., Guedes Soares, C., 2012. Maritime Traffic Monitoring Based on Vessel Detection, Tracking, State Estimation, and Trajectory Prediction. IEEE Transactions on Intelligent Transportation Systems 13, 1188–1200. doi:10.1109/TITS.2012.2187282.

Prasad, D.K., Rajan, D., Rachmawati, L., Rajabally, E., Quek, C., 2017. Video Processing From Electro-Optical Sensors for Object Detection and Tracking in a Maritime Environment: A Survey. IEEE Transactions on Intelligent Transportation Systems 18, 1993–2016. doi:10.1109/TITS.2016.2634580.

Reynolds, D.A., Quatieri, T.F., Dunn, R.B., 2000. Speaker verification using adapted gaussian mixture models. Digital signal processing 10, 19–41.

Ristic, B., Scala, B.L., Morelande, M., Gordon, N., 2008. Statistical analysis of motion patterns in AIS Data: Anomaly detection and motion prediction. 2008 11th International Conference on Information Fusion , 40–46doi:10.1109/ICIF.2008.4632190.

Rødseth, Ø.J., Perera, L.P., Mo, B., 2015. Big Data in Shipping - Challenges and Opportunities, in: 15th International Conference on Computer and IT Applications in the Maritime Industries - COMPIT '16, pp. 361–373.

Rong, H., Teixeira, A., Guedes Soares, C., 2019. Ship trajectory uncertainty prediction based on a Gaussian Process model. Ocean Engineering 182, 499–511. doi:10.1016/J.OCEANENG.2019.04.024.

Schwarz, G., et al., 1978. Estimating the dimension of a model. The annals of statistics 6, 461–464.

Shen, H., Hashimoto, H., Matsuda, A., Taniguchi, Y., Terada, D., Guo, C., 2019. Automatic collision avoidance of multiple ships based on deep Q-learning. Applied Ocean Research 86, 268–288. doi:10.1016/j.apor.2019.02.020.

Steinbach, M., Ertöz, L., Kumar, V., 2004. The Challenges of Clustering High Dimensional Data, in: New Directions in Statistical Physics. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 273–309.

Tam, C., Bucknall, R., Greig, A., 2009. Review of collision avoidance and path planning methods for ships in close range encounters. Journal of Navigation 62, 455–476. doi:10.1017/S0373463308005134.

Theodoridis, S., Koutroumbas, K., 2009. Pattern recognition. 4th ed. ed., Elsevier, Amsterdam.

Tu, E., Zhang, G., Rachmawati, L., Rajabally, E., Huang, G.B., 2017. Exploiting AIS Data for Intelligent Maritime Navigation: A Comprehensive Survey From Data to Methodology. IEEE Transactions on Intelligent Transportation Systems , 1–24doi:10.1109/TITS.2017.2724551.

Wen, Y., Sui, Z., Zhou, C., Xiao, C., Chen, Q., Han, D., Zhang, Y., 2020. Automatic ship route design between two ports: A data-driven method. Applied Ocean Research 96, 102049. doi:10.1016/j.apor.2019.102049.

Xu, H., Hassani, V., Guedes Soares, C., 2020. Comparing generic and vectorial nonlinear manoeuvring models and parameter estimation using optimal truncated least square support vector machine. Applied Ocean Research 97, 102061. doi:10.1016/j.apor.2020.102061.

Yang, R., Xu, J., Wang, X., Zhou, Q., 2019. Parallel trajectory planning for shipborne Autonomous collision avoidance system. Applied Ocean Research 91, 101875. doi:10.1016/j.apor.2019.101875.

Yara, 2019. Yara Birkeland press kit. URL: https://www.yara.com/news-and-media/press-kits/yara-birkeland-press-kit/.

Zhang, L., Meng, Q., Fang Fwa, T., 2017. Big AIS data based spatial-temporal analyses of ship traffic in Singapore port waters. Transportation Research Part E: Logistics and Transportation Review doi:10.1016/J.TRE.2017.07.011.