IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2023.1120000

Spatial Clustering Approach for Vessel Path Identification

MOHAMED ABUELLA[1], M. AMINE ATOUI[1], SŁAWOMIR NOWACZYK[1], SIMON JOHANSSON[2], ETHAN FAGHANI[2]

[1]Center for Applied Intelligent Systems Research (CAISR), Halmstad University, Halmstad, 30118, Sweden (e-mail: mohamed.abuella, amine.atoui, slawomir.nowaczyk}@hh.se)
[2]CetaSol AB, Gothenburg, 41251, Sweden (e-mail: simon.johansson and ethan.faghani}@cetasol.com)

Corresponding author: Mohamed Abuella (e-mail: mohamed.abuella@hh.se).

This research project is funded by Sweden's innovation agency (Vinnova)

ABSTRACT This paper addresses the challenge of identifying the paths for vessels with operating routes of repetitive paths, partially repetitive paths, and new paths. We propose a spatial clustering approach for labeling the vessel paths by using only position information. We develop a path clustering framework employing two methods: a distance-based path modeling and a likelihood estimation method. The former enhances the accuracy of path clustering through the integration of unsupervised machine learning techniques, while the latter focuses on likelihood-based path modeling and introduces segmentation for a more detailed analysis. The result findings highlight the superior performance and efficiency of the developed approach, as both methods for clustering vessel paths into five clusters achieve a perfect F1-score. The approach aims to offer valuable insights for route planning, ultimately contributing to improving safety and efficiency in maritime transportation.

INDEX TERMS Spatial clustering, vessel path identification, maritime transportation, average nearest neighbor distance, hierarchical clustering, likelihood estimation.

I. INTRODUCTION

A. BACKGROUND AND MOTIVATION

Maritime transportation is crucial for global trade, generating extensive vessel trajectory data that reveals intricate spatial and temporal navigation patterns. Understanding these patterns is vital for effective maritime traffic surveillance and management [1].

It is crucial to distinguish between trajectory and path when studying movement data.
The term *path* refers to the specific course taken by the object. While, a *trajectory* refers to a sequence of consecutive geographical points, each representing a specific location at a given timestamp [2]. Thus, the trajectory typically denotes the movement of an object over time. In other words, a trajectory is a path with schedule and speed information. For instance, if a vessel travels from origin port to destination port, its trajectory is the sequence of geographical points it passes through, while its path is the specific course taken, such as a fairway in a small river or through a specific canal. While the vessel trajectory is its navigation including time and space information, such as time schedule, speed,

and location [3], [4].
The primary distinction between a path and a trajectory lies in the absence of temporal data within the path, which solely encompasses spatial and sequential information.

Trajectories, observed in various scenarios such as pedestrian movements, vehicular routes, and natural events like wildlife migrations or hurricanes, involve time-evolving position data. Trajectory mining aims to uncover significant patterns within datasets, enabling tasks like path classification, anomaly detection for accidents or traffic congestions, surveillance for suspicious activities, and prediction of vessel trajectories in different landscapes [5].

Additionally, the term *route* it is the path or trajectory (according to their respective data) that shares the same origin and destination. Conversely, any difference in either the origin or destination represents another route. For example, when a vessel travels from an origin port to three destination ports, we say this vessel has three distinct routes to reach its three destinations.

The visual example for the aforementioned definitions are illustrated by Figure 1. The waypoints in the graphs represent the type of information that are encompassed with the path are positional data (p_i), while in the trajectory are spatial and temporal data (f_i).

Moreover, a *Voyage* is a contextual term, generally referring to the period between a departure from a port to the departure from the next port. Voyage is commonly used in reference to sea travel, much like the term 'trip' is used in the context of air travel.

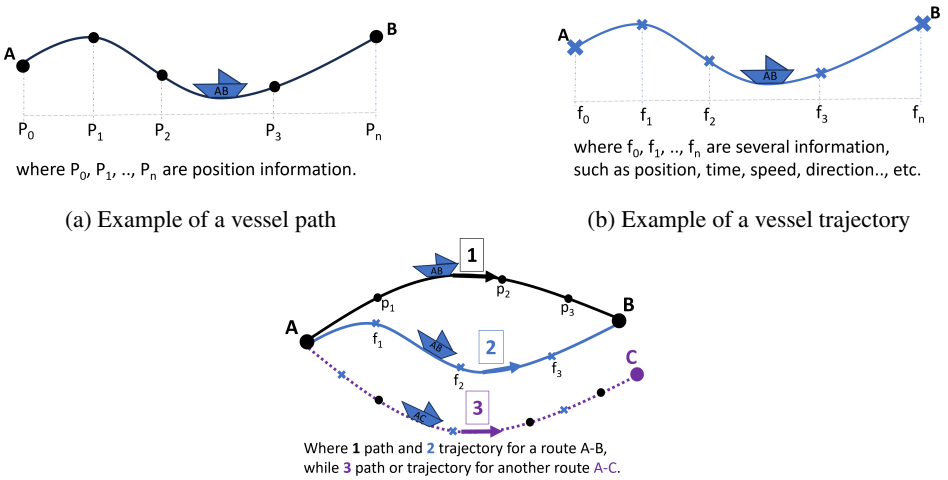
Path clustering - same as trajectory clustering, but we do not consider the time information. Our focus is primarily on spatial information of the moving object. Path clustering, a versatile technique, involves grouping paths into clusters based on their similarity, demonstrating its effectiveness in a myriad of practical applications. In the realm of navigation, path identification empowers systems to generate clear and detailed instructions for users seeking their way. Traffic analysis benefits from path clustering as it facilitates the identification of diverse traffic patterns, such as the smooth flow of traffic on highways and the congestion often encountered on city streets. Path identification proves equally valuable in route planning, enabling the optimization of routes for transportation systems, including public and maritime transportation services [6].

In the scope of the maritime industry, path identification from Automatic Identification System (AIS) data is a challenging task due to the high spatial freedom and, especially in coastal areas, the high frequency of ship's navigation maneuvers. Thus, it is imperative to develop a path identification tool that integrates with route planning systems for improving maritime safety and optimizing vessel routing. As data-driven approaches from AIS data continue to grow and evolve, path clustering will undoubtedly play an increasingly important role in understanding vessel behavior and supporting decision-making in maritime transportation [7].

B. AIMS AND CONTRIBUTIONS

This paper aims to address the challenge of identifying vessel paths in scenarios characterized by repetitive, semi-repetitive, and novel operations. In general, the aims and contributions of the proposed approach in this paper can be outlined as follows:

- The proposed clustering approach of vessel paths requires only position information, specifically longitude and latitude.
- The clustering approach has a proven added value for clustering challenging unseen or unknown paths.
- The approach is robust and interpretable by applying a similarity measure that reduces the influence of noise or outliers and offers a clear interpretation of path clustering.
- The approach has a customizable parameter to determine the number of path clusters, thereby enhancing the flexibility and adaptability of the framework and allowing users to tailor it to their specific needs.
- The approach also includes a method to



(a) Example of a vessel path

(b) Example of a vessel trajectory

(c) Example of vessel routes

FIGURE 1: Illustration for the definitions of (a) path, (b) trajectory, and (c) route of a vessel.

study and analyze the patterns within specific segments of a path.

- It is a data-driven solution that can be used as a valuable asset for informed decision-making in route planning and optimization, traffic management, and resource allocation.

The rest of this paper is organized as follows. Section II reviews the related work on vessel path identification. Section III describes our proposed spatial clustering approach in detail. The real-world vessel data of our case study is described in Section IV. Section V presents the results of our experimental evaluation. Section VI concludes the paper and discusses future work.

II. RELATED WORK

Path clustering can be done using a variety of different methods [8]. We will explore the related works of these various methods.

Clustering is gaining popularity for route extraction. Machine learning (ML) has recently been applied extensively for vessel path identification by learning patterns from historical data. Lee et al., in [5] introduced TRACLUS, a trajectory clustering algorithm employing a partition-and-group framework to discover common sub-trajectories. Demonstrating efficacy through formal trajectory partitioning and density-based clustering and efficiently identifies shared patterns in real trajectory data. The study presented in [9] introduced a framework, Traffic Route Extraction and Anomaly Detection (TREAD), which utilizes unsupervised learning for maritime route extraction. The primary emphasis is on anomaly detection and route prediction, highlighting the crucial role of AIS data in enhancing maritime situational awareness. The work specifically addresses challenges related to intermittency and persistence in AIS data. Another method of route extraction was proposed in [10], transforming ship trajectories into ship trip semantic objects (STSO) and utilizing graph theory for route extraction. The method proves robustness in extracting traffic routes for merchant ships but may have limitations for vessels with frequent navigation behavior changes, such as fishing vessels. The approach in [11], on the other hand, adopts a dynamic time warping (DTW) distance as a similarity measure and

considers vessel course changes to analyze its trajectories. Experiments demonstrated its high accuracy in distinguishing and detecting similar vessel trajectories, outperforming existing methods in accuracy and cluster degree evaluation. Moreover, [12] presents a machine learning framework for maritime vessel trajectory analysis, incorporating clustering, classification, and outlier detection. It employs piecewise linear segmentation for compression and alignment kernels to integrate geographical domain knowledge, enhancing task performance. Results show reduced computation time without compromising accuracy.

Capobianco et al. [13] proposed a deep learning approach using recurrent neural networks, employing a Bidirectional Long Short-Term Memory (BiLSTM) layer as an encoder and a Unidirectional Long Short-Term Memory (LSTM) layer as a decoder, for vessel trajectory prediction. Their model outperforms baseline approaches, showcasing the effectiveness of sequence-to-sequence neural networks. In their study, Li et al. [14] present an AIS data-based machine learning method for feature extraction and unsupervised route planning for Maritime Autonomous Surface Ships (MASS). The method uses Automatic and Adaptive Dynamic Time Warping (AADTW), Spectral Clustering with Affinity Feature (SCAF), and a route optimization algorithm based on dynamic programming to extract features, obtain movement patterns, and plan routes. The proposed method outperforms existing methods by considering the impact of hidden factors and providing different routes for different types of MASS. The work in [15] systematically analyzes the performance of twelve ship trajectory prediction methods, including classical machine learning and emerging deep learning techniques. It compares twelve methods across three AIS datasets, representing different maritime traffic scenarios, and evaluates their effectiveness based on six indexes. The study concludes that traditional machine learning-based trajectory prediction methods struggle to meet the rising demands for accuracy and real-time performance, leading to increased interest in and promising results from deep learning-based approaches. EnvClus*, introduced in [16], it is an innovative unsupervised data-driven framework for vessel trajectory forecasting, achieving a 33% improvement over state-of-the-art methods. EnvClus* excels in accurately predicting vessel routes, particularly in long trips, showcasing its effectiveness in mobility analytics and trajectory prediction.

A maritime traffic route extraction approach based on multi-dimensional density-based spatial clustering of applications with noise (MD-DBSCAN) was developed in [17]. The approach incorporates trajectory compression, similarity measures, and extraction of ship trajectory clusters. The approach demonstrates effectiveness in noise reduction and route extraction. The authors in [18] proposed a trajectory clustering method based on Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) and Hausdorff distance to generate a similarity matrix. The method adapts to shape characteristics and exhibits good clustering scalability and improved clustering results compared to DBSCAN, k-means, and spectral clustering algorithms.

Eljabu et al. proposed spatial clustering methods (SPTCLUST and SPTCLUST-II) in [19] and [20] respectively, for maritime traffic routes extraction from AIS data. The approach consists of data preprocessing, pathfinding, and route extraction without using traditional clustering algorithms. It achieved high F1-scores, 97% and 99%, for tankers and cargo maritime traffic routes.

The study in [21] enhanced the DBSCAN method by integrating the Mahalanobis distance metric for vessel behavior modeling. The proposed methodology includes clustering historical AIS data and detecting anomalies. The study showcases applicability to diverse water regions, contributing to situational awareness, collision prevention, and route planning.

Farahnakian et al. [22] conducted a comprehensive examination of clustering-based techniques, including k-means, DBSCAN, Affinity Propagation (AP), and the Gaussian Mixtures Model (GMM), for detecting abnormal vessel behaviors from AIS data. Results indicate that k-means is particularly effective in detecting dark ships and spiral vessel movements, which is crucial for enhancing maritime safety. Furthermore, the study [23] proposed two methods for trajectory outlier detection, with the first utilizing DBSCAN clustering and Hausdorff distance, and the second employing Support Vector Machine (SVM) classifier and the Generalized Sequence Pattern algorithm. Both models outperform the baselines, with the SVM approach demonstrating superior performance in the identification of traffic patterns and outliers. The study [3] developed MUITAS, a novel trajectory similarity measure addressing limitations in existing approaches for multiple-aspect trajectories enriched with heterogeneous semantic dimensions. Through evaluation on real datasets, MUITAS demonstrates robustness and outperforms current methods in precision at recall and clustering techniques for diverse mobility data. Moreover, the authors in [24] presented a methodology for extracting navigation network information from vessel trajectories, utilizing AIS data. The proposed model identifies key areas, speed, and course patterns, forming a network abstraction for optimizing ship routing and scheduling in the maritime industry, demonstrated through analysis in the eastern Mediterranean sea. This model is also useful in an outlier behavior detection.

The research paper [25] offers a detailed survey of visual analytics for vessel trajectory data. The authors discuss a variety of methods, including map-based visualization, timeline-based visualization, and interactive visualization.

The survey [4] delves into the growing focus on semantically rich trajectories in movement data analysis, covering concepts, management issues, and techniques for constructing, enriching, and mining trajectories, with attention to emerging privacy challenges.

The paper [26] comprehensively reviews various approaches for vessel trajectory predicting, including clustering algorithms and machine learning algorithms. It also discusses the challenges and future research directions, such as the uncertainty in the data, the dynamic environment, and the computational complexity.

Among the identified challenges, which are subjects of ongoing research and require additional attention, three are worthy of specific mention: navigating dynamic maritime environments poses a substantial challenge in accurately identifying vessel paths (I); ensuring stability, explainability, and managing the computational cost of the model add further complexity (II); finally, addressing the need for flexibility, scalability, and practical applicability is crucial for a comprehensive solution in the field of vessel path identification (III). Motivated by these challenges, we aim to develop a framework that focuses on vessel path identification and potentially tackling such challenges faced in maritime transportation.

III. METHODOLOGY

This section covers the theoretical background and description of our proposed framework's underlying methodology. The framework of vessel path identification is depicted in Figure 2.

1) Problem Formulation

The equations (1-3) serve as a mathematical representation to describe the clustering of vessel paths.

It is worth mentioning that the clustering process is conducted sequentially, point by point, while the labeling of path clusters is performed for the entire voyage. Therefore, each voyage has a single label of path cluster.

$$Voyages \in Path\ Clusters \tag{1}$$

$$Voyages = \{ts_1[p_1, \ldots, p_n], \ldots, ts_j[p_1, \ldots, p_n]\} \tag{2}$$

$$Path\ Cluster\ Set = \{cluster_1, \ldots, cluster_k\} \tag{3}$$

where:
Voyages: a collection of time series, each representing a voyage of the vessel taken following a given path, with a predicted cluster.
ts_j: a time series corresponding to voyage j, i.e., a sequence of n data points, where each data point p represents the vessel position and is defined by a pair of coordinates, namely latitude and longitude.
n: the number of time steps (duration) of each voyage, which can differ from one voyage to another.
j: a total number of voyages.
Path Cluster Set: a set of k clusters into which the path of voyages are being clustered.

Remark: In the definition of k clusters, it is important to clarify that this study is for labeling or identifying predefined fairways for vessels navigating in confined waters rather than open sea, resulting in a predetermined number of path clusters, k.

2) Distance-Based Method

The similarity between two paths is measured by the average nearest neighbor distance (ANND), as shown in Eq. (4).

$$ANND(i,j) = \frac{1}{n_i}\sum_{k=1}^{n_i} Distance(P_i^k, NN(P_j^k)) \tag{4}$$

where:
$ANND(i,j)$: is the average nearest neighbor distance between path i and path j, present in the distance matrix at row i and column j. It is a symmetric, meaning that $ANND(i,j)$ is the same as $ANND(j,i)$
$Distance(P_i^k, NN(P_j^k))$: The distance between the k^{th} point in path i, denoted as P_i^k, and its corresponding nearest neighbor point in path j, indicated as $NN(P_j^k)$. n_i is the total number of points in path i.
The measure *Distance* is an Euclidean distance. However, for longer curved routes, Haversine or Great-circle distance would be more suitable.

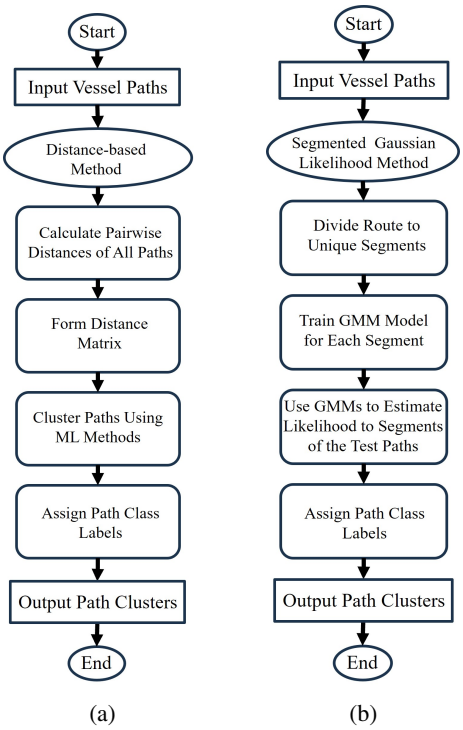


FIGURE 2: Framework of vessel path identification. (a) Flowchart of distance-based method. (b) Flowchart of segmented Gaussian likelihood method.

The ANND, as expressed in Eq. (4), is computed by averaging the distances between each point in one path and its nearest neighbor in the other path.
Then, the similarity value (i.e., ANND) of this pair of paths is stored as an element in the distance matrix.
A lower ANND indicates that the paths within a cluster are more similar. The distance matrix will have dimensions $(m \times m)$, where m is the number of paths.
For instance, the computed distance matrix for a set of 12 paths is illustrated in Figure 3.

After the construction of the distance matrix, the machine learning (ML) technique is applied

to cluster the paths based on their corresponding values in this distance matrix. The ML techniques that we used are k-means, Gaussian Mixture Model (GMM), and hierarchical clustering.

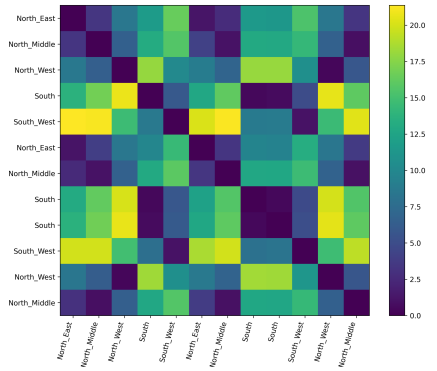


FIGURE 3: Heatmap of the distance matrix for 12 paths of Cinderella II vessel.

3) Segmented Gaussian Likelihood Method

In addition to identifying the vessel's path, to better understand how the vessel changes its paths, we employ Gaussian distributions on several distinct segments of the vessel route. This technique can be summarized as follows:

- Utilize a training dataset comprising vessel position information that should adequately represent all potential paths of the vessel route.
- Divide the route into different distinct segments.
- Train a single GMM model for each segment to find the Gaussian distributions of all segments of the route.
- Estimate likelihoods of the segments by using the trained GMM models with their corresponding segments of each vessel voyage in the test dataset.
- Label the path clusters based on the estimated likelihood at the unique segments of the route.

IV. CASE STUDY

In this section, we describe the case study, including the data collection, preprocessing, and analysis.

A. DATA COLLECTION

In this study, we utilized datasets collected from two passenger ships, named Cinderella II and Buro, operating in Sweden. The vessels are shown in Figure 4, additional information about vessels Cinderella II and Buro can be found in [27].

Cinderella II operates in Stockholm archipelago, east of Sweden. The data of Cinderella II spans over five months (July to November 2022). It comprises information on 124 voyages of this vessel, connecting the two main ports of Vaxholm in the east and Sodra in the west.

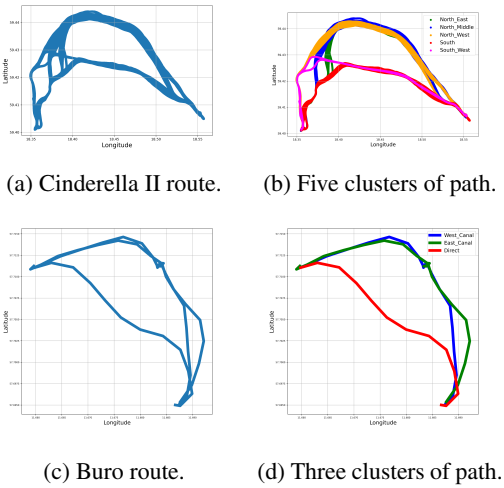
While Buro works in Gothenburg, west of Sweden. It dataset has been gathered over a period of 15 months (between January 2020 and March 2021). The data of Buro has 1755 voyages, between two main ports, Groto in the south and Ockero in the northwest.

(a) Cinderella II. (b) Buro.

FIGURE 4: Images of two vessels that are used in the case study [27].

B. DATA PREPARATION AND ANALYSIS

In our approach, we emphasize the significance of data representation. As a result, we group the path points based on their timestamps with a resolution of one second and store these grouped path points with distinctive Voyage IDs.
The routes of Cinderella II and Buro, along their chosen paths, are depicted in Figure 5. Path cluster distribution for both vessels are illustrated by the histograms in Figure 6.



(a) Cinderella II route. (b) Five clusters of path.

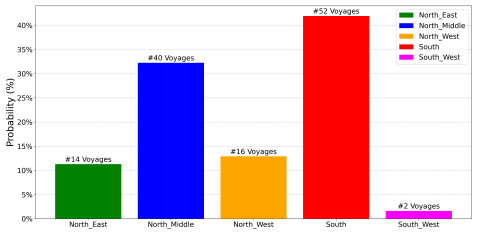
(c) Buro route. (d) Three clusters of path.

FIGURE 5: Routes of the two vessels beside their all possible path clusters, which will be identified by applying proposed framework.
Note: The colored lines on right side are correct path clusters or ground truth for both vessels.

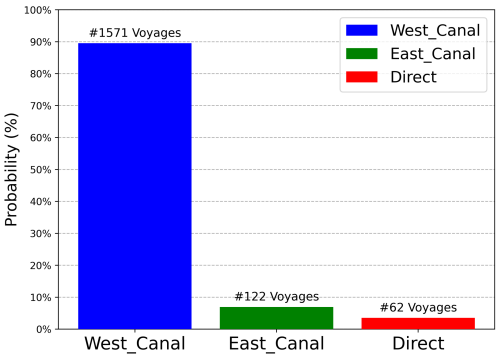
Afterward, these paths are ready to be processed by the path clustering approach to determine the overall path cluster. In order to exploit the resulting cluster information, statistical analysis is conducted to determine differences in vessel paths concerning fuel, time, distance, and speed. The histograms representing these quantities across various path clusters can be found in Figure 7. Notably, when the vessel traverses the shorter southern paths, it employs slower speeds, effectively reducing fuel consumption without significantly impacting travel time.

V. RESULTS AND DISCUSSION

In this section, we present the results of our spatial clustering approach for vessel path identification and discuss the implications of these findings. The approach involved the utilization of position information and various clustering techniques, specifically k-means, hierarchical clustering, and Gaussian distributions clustering, with the dataset containing 124 voyages.



(a) Distribution of Cinderella II voyages across the five path clusters.



(b) Distribution of Buro voyages across the three path clusters.

FIGURE 6: Distribution of voyages if both vessels across the the path clusters.

A. EVALUATION OF PATH CLUSTERING

The results of vessel path identification are evaluated through visual inspection and tabulation using metrics such as confusion matrix, precision, recall, and F1-score [28].

The hits and messes of path clustering are presented by the confusion matrix. For our results of path clustering, the confusion matrix is a one-vs-one type matrix. Then, the confusion matrix is converted into a one-vs-all type matrix (binary-class confusion matrix) as shown in Eq. (5), for calculating class-wise metrics like precision, recall, and F1-score.

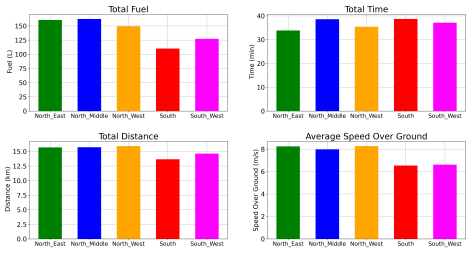


FIGURE 7: Average fuel and time, distance, and speed of five path clusters of Cinderella II vessel.

	Pred. Pos.	Pred. Neg.
Act. Pos.	TP	FN
Act. Neg.	FP	TN

(5)

where True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN) are determined by comparing the predicted (Pred.) and actual (Act.) path clusters.

The confusion matrix transformation involves considering one class as positive at a time, while combining all other classes as negative. This process is repeated iteratively for each class, resulting in multiple binary-class confusion matrices.

The following performance metrics were used:

- Precision: the ratio of true positives to the total number of predicted positives.
- Recall: the ratio of true positives to the total number of actual positives in the data.
- F1-score: the harmonic mean of precision and recall.

The equations for precision, recall, and F1-score are shown in Eqs. (6), (7), and (8).

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

$$F1\text{-}score = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)} \tag{8}$$

B. DISCUSSION

1) Results of Cinderella II Vessel

Table 1 shows the results of applying k-means or Gaussian Mixture Model (GMM) models to identify the vessel paths from the distance matrix in the distance-based method of the path identification approach. Notably, the paths with clusters of North-West, South, and South-West achieved an F1-score of 1.0, indicating that the approach correctly identified all the paths of these clusters.

In contrast, the North-East and North-Middle paths exhibited lower F1 scores compared to other clusters. The path cluster of North-Middle is the most challenging path to identify since six such paths have been clustered as North-East, as can be seen by comparing Figures 8 and 9, which are the visualization for all the paths, color-marked based on their ground truth clusters. Figure 10 illustrates the probability distribution of mis-clustered paths with respect to latitude and longitude coordinates. It is obvious that these paths have nearly identical coordinates, which makes them challenging paths to cluster with k-means or GMM. This suggests that there is still room for improvement by using other ML clustering methods.

Table 2 presents the results of employing hierarchical clustering to the distance matrix in the distance-based method for clustering the vessel paths. In hierarchical clustering, there is a parameter called "Dendrogram Cut-off threshold," and its value should be selected depending on the number of path clusters. Hence, as illustrated in Figure 11, this parameter is denoted by the Y-axis as a clustering height, and its value is set to 100 for clustering the vessel path into five clusters.

Remarkably, all path clusters achieved an F1-score of 1, indicating that hierarchical clustering successfully identified all paths with high accuracy from the distance matrix using the distance-based method. This suggests that the choice of ML clustering technique with the distance matrix can influence the accuracy of path identification. Table 2 displays the outcomes of clustering, now

by applying the segmented likelihood Gaussian method. This method achieved perfect precision, recall, and F1-score for all path clusters.
Figures 12, 13, and 14 present visualizations for the segmented Gaussian likelihood method.

The accuracy in results by hierarchical and segmented Gaussian likelihood clustering for path clusters indicates the efficacy of the developed approach of spatial clustering for vessel path identification.

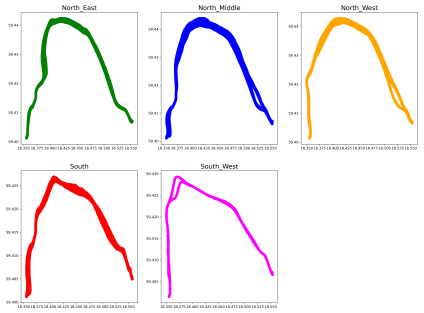


FIGURE 8: The observations of the five clusters of path for Cinderella II vessel. Note: These are the correct path clusters or ground truth.

2) Results of Buro Vessel

In this part the results from applying the framework of both hierarchical clustering and segmented likelihood Gaussian clustering of Buro vessel paths to three clusters. Following a similar evaluation procedure as in the case of Cinderella II vessel.

As it can be seen, the results of likelihood Gaussian clustering also achieved F1-score of 1, as in Cinderella vessel. But for the hierarchical clustering in Buro vessel case has F1-scores of 0.957 and 0.996 for clustering East_Canal and West_Canal paths respectively. This results can be considered remarkable when take into account that these two paths, East_Canal and West_Canal, are challenging to be clustered, since they have several paths that are slightly different form each other. Figure 19 shows the heatmap of the distance matrix of 12 sample paths for Buro vessel. Notably, the East_Canal and West_Canal paths exhibit high similarity, which make not easy clustering task.

TABLE 1: Result of implementing of both k-means and GMM clustering of the paths of Cinderella II vessel to five clusters.

(a) Precision, Recall, and F1-score

Paths	Precision	Recall	F1-score
North-East (NE)	0.7	1	0.824
North-Middle (NM)	1	0.85	0.919
North-West (NW)	1	1	1
South (S)	1	1	1
South-West (SW)	1	1	1

(b) Confusion Matrix

Actual	Predicted					Total
	NE	NM	NW	S	SW	
NE	14	0	0	0	0	14
NM	6	34	0	0	0	40
NW	0	0	16	0	0	16
S	0	0	0	52	0	52
SW	0	0	0	0	2	2
Total	20	34	16	52	2	124

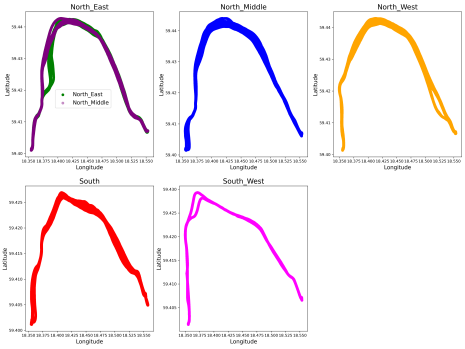


FIGURE 9: Results of both k-means and GMM clustering to five path clusters for Cinderella II. The quantitative results for correct and incorrect paths into the five clusters are shown by the confusion matrix in Table 1 (b).

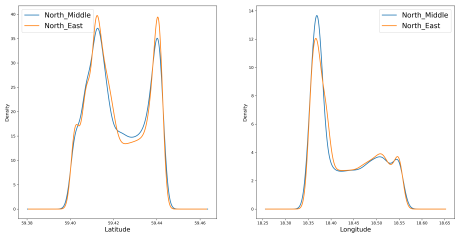


FIGURE 10: Probability distribution of location coordinates for mis-clustered paths of Cinderella II by both k-means and GMM.

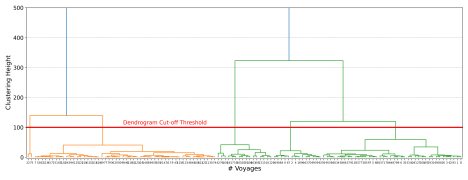


FIGURE 11: Results of hierarchical clustering the paths of Cinderella vessel to five clusters.

Choosing a proper value of the agglomerative threshold for the hierarchical clustering to get the three clusters, as shown in Figure 15, with an agglomerative threshold = 4.7.

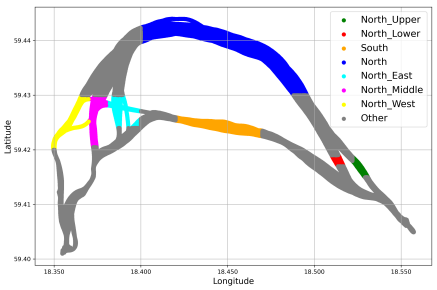


FIGURE 12: Distribution of the route of Cinderella II into eight segments.

TABLE 2: Result of implementing both hierarchical clustering and segmented likelihood Gaussian clustering of paths for Cinderella II vessel into five clusters.

(a) Precision, Recall, and F1-score

Paths	Precision	Recall	F1-score
North-East (NE)	1	1	1
North-Middle (NM)	1	1	1
North-West (NW)	1	1	1
South (S)	1	1	1
South-West (SW)	1	1	1

(b) Confusion Matrix

Actual	Predicted					Total
	NE	NM	NW	S	SW	
NE	14	0	0	0	0	14
NM	0	40	0	0	0	40
NW	0	0	16	0	0	16
S	0	0	0	52	0	52
SW	0	0	0	0	2	2
Total	14	40	16	52	2	124

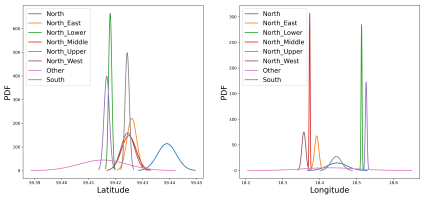


FIGURE 13: Probability distributions of location coordinates for the eight segments of the route of Cinderella II.

3) Results of Implementing TRACLUS Algorithm

For benchmarking our framework, we employed TRACLUS, a well-established algorithm for trajectory clustering [5], to provide a baseline for evaluating the performance of our framework.

Figure 20 displays the implementation of TRACLUS clustering approach on the case study of Cinderella II vessel.

The quantitative outcomes of employing TRACLUS algorithm on both vessels, Cinderella II and Buro, are detailed in Table 6 and Table 5, respectively.

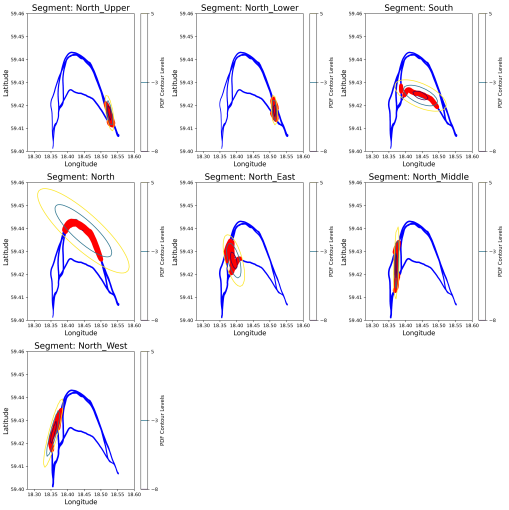


FIGURE 14: Gaussian distributions for seven segments of Cinderella II vessel route.

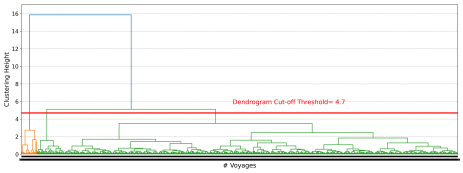


FIGURE 15: Results of hierarchical clustering to three path clusters of Buro vessel.

TABLE 3: Result of Buro vessel by implementing hierarchical clustering to three path clusters

(a) Precision, Recall, and F1-score

Paths	Precision	Recall	F1-score
Direct	1	1	1
East_Canal	0.946	0.861	0.901
West_Canal	0.989	0.996	0.993

(b) Confusion Matrix

Actual	Predicted			Total
	Direct	E-C	W-C	
Direct	62	0	0	62
E-C	0	105	17	122
W-C	0	6	1565	1571
Total	62	111	1582	1755

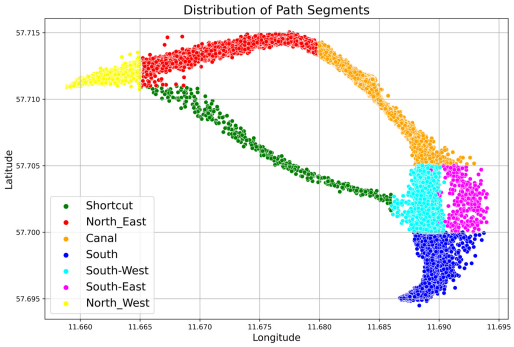


FIGURE 16: Distribution of the route of Buro vessel into seven segments.

By implementing the TRACLUS as a benchmark model for our case studies, it exhibits relatively lower performance.
TRACLUS utilizes DBSCAN to group similar trajectories together, a method that inherently identifies clusters based on density-connected sets [29]. The inherent characteristics of the DBSCAN clustering model come into play, particularly in scenarios where trajectory density varies significantly with complex navigational patterns are present. In addition, the sensitivity of TRACLUS to threshold parameters further complicates its performance, as it requires careful tuning to suit specific datasets. This sensitivity can lead to difficulties in reproducibility and practical application.

These limitations emphasize the necessity of exploring alternative methodologies that better align with the rigorous demands of the maritime industry.

VI. CONCLUSION

The proposed approach is able to identify the vessel paths with partially defined or unknown paths. In the distance-based method, the hierarchical clustering used in the approach outperforms k-means and GMM clustering techniques.

The approach of hierarchical clustering includes a user-customizable parameter, a cut-off

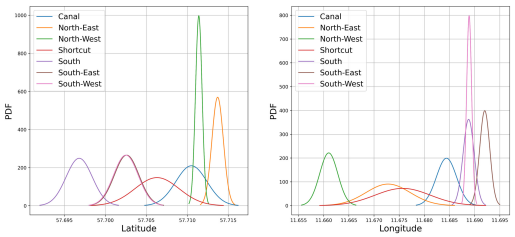


FIGURE 17: Probability distributions of location coordinates for the seven segments of the route of Buro vessel.

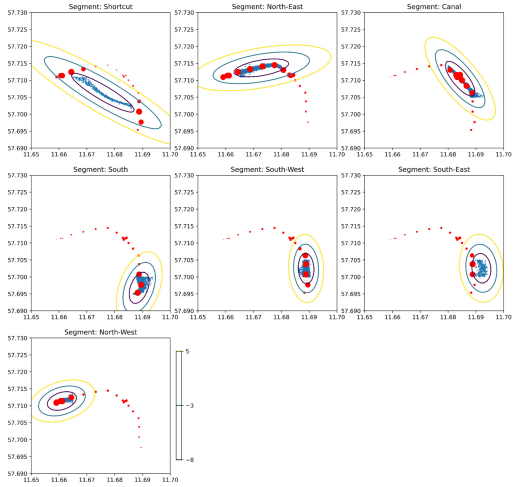


FIGURE 18: Gaussian distribution for seven segments of the route of Buro vessel.

threshold, which allows desired control for the number of path clusters, enhancing the flexibility and adaptability of the proposed approach.

In the distance-based method, adopting ANND as a measure of similarity makes path clustering less affected by noise or outliers and provides a more intuitive interpretation of path similarity, ultimately enhancing the robustness and interpretability of our approach.

The segmented Gaussian likelihood method is particularly useful for identifying and analyzing the vessel path alterations at different segments of the vessel route.

The proposed approach is computationally efficient and has the potential to be a valuable tool for planning vessel paths. Accurate path identification can contribute to safer and more efficient maritime transportation practices, aiding in route planning, collision avoidance, and navigation optimization.

Nevertheless, the framework has some potential limitations, such as the segmented Gaussian likelihood method exhibiting sensitivity to segment definition, which could affect its salable performance, particularly in complex mar-

TABLE 4: Result of Buro vessel by implementing segmented likelihood Gaussian clustering to three path clusters

(a) Precision, Recall, and F1-score

Paths	Precision	Recall	F1-score
Direct	1	1	1
East_Canal	1	1	1
West_Canal	1	1	1

(b) Confusion Matrix

Actual	Predicted			Total
	Direct	E-C	W-C	
Direct	62	0	0	62
E-C	0	122	0	122
W-C	0	0	1571	1571
Total	62	122	1571	1755

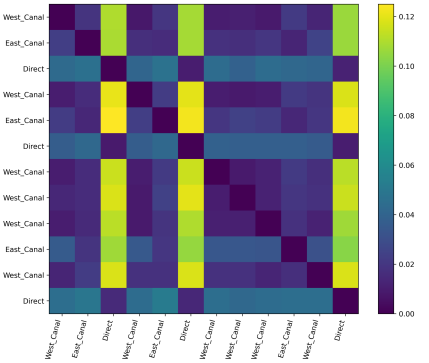
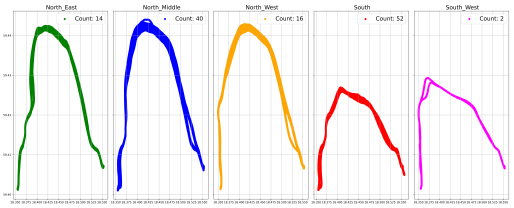
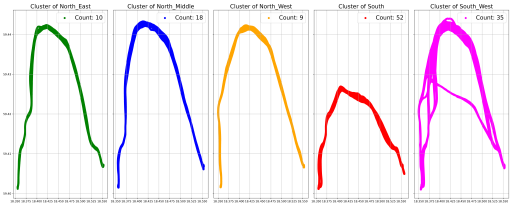


FIGURE 19: Heatmap of the distance matrix for 12 paths of Buro vessel.



(a) The observations of five path clusters of Cinderella II, ground truth.



(b) The result of TRACLUS for five path clusters of Cinderella II.

FIGURE 20: Clustering path clusters of Cinderella II to five clusters by TRACLUS algorithm.

TABLE 5: Result of implementing TRACLUS for clustering the paths of Buro vessel to three path clusters

(a) Precision, Recall, and F1-score

Paths	Precision	Recall	F1-score
Direct	1	1	1
East_Canal	0	0	0
West_Canal	0.928	0.999	0.962

(b) Confusion Matrix

Actual	Predicted			Total
	Direct	E-C	W-C	
Direct	62	0	0	62
E-C	0	0	122	122
W-C	0	6	1565	1571
Total	62	6	1687	1755

itime scenarios. Moreover, while the study case demonstrates that the framework is computationally efficient, it is essential to discuss any potential scalability issues, especially when dealing with large datasets, since the computational efficiency may vary depending on the dataset size and the nature of the paths.

TABLE 6: Result of implementing of TRACLUS for clustering the paths of Cinderella II vessel to five clusters.

(a) Precision, Recall, and F1-score

Paths	Precision	Recall	F1-score
North-East (NE)	1	0.714	0.833
North-Middle (NM)	1	0.450	0.621
North-West (NW)	1	0.562	0.720
South (S)	1	1	1
South-West (SW)	0.057	1	0.108

(b) Confusion Matrix

Actual	Predicted					Total
	NE	NM	NW	S	SW	
NE	10	0	0	0	4	14
NM	0	18	0	0	22	40
NW	0	0	9	0	7	16
S	0	0	0	52	0	52
SW	0	0	0	0	2	2
Total	10	18	9	52	35	124

Our framework is specifically designed to exceed the rigorous requirements of the maritime industry, surpassing typical clustering approaches like TRACLUS.

Further work could explore the scalability and real-world applicability of the proposed clustering approach, as well as its integration with related systems of maritime transportation.

In future studies, we aim to upgrade the framework by incorporating additional information such as time, speed, direction, and pertinent environmental factors for clustering trajectories in open sea navigation. This will enhance the applicability and robustness of our approach for various real-world maritime operations, including optimizing route planning, risk assessment, and navigation management.

ACKNOWLEDGMENT

This research project is funded by Sweden's innovation agency (Vinnova).
We also wish to thank the diverse group at the Center for Applied Intelligent Systems Research

(CAISR), Halmstad University, for helpful discussions.

SUPPLEMENTARY MATERIALS

The source codes that are implemented on Python 3.9.7 to produce the results are available at: https://github.com/MohamedAbuella/Path_Clustering

REFERENCES

[1] E. Tu, G. Zhang, L. Rachmawati, E. Rajabally, and G.-B. Huang, "Exploiting ais data for intelligent maritime navigation: A comprehensive survey from data to methodology," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 5, pp. 1559–1582, 2017.

[2] Y. Zheng, "Trajectory data mining: an overview," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 6, no. 3, pp. 1–41, 2015.

[3] L. M. Petry, C. A. Ferrero, L. O. Alvares, C. Renso, and V. Bogorny, "Towards semantic-aware multiple-aspect trajectory similarity measuring," *Transactions in GIS*, vol. 23, no. 5, pp. 960–975, 2019.

[4] C. Parent, S. Spaccapietra, C. Renso, G. Andrienko, N. Andrienko, V. Bogorny, M. L. Damiani, A. Gkoulalas-Divanis, J. Macedo, N. Pelekis *et al.*, "Semantic trajectories modeling and analysis," *ACM Computing Surveys (CSUR)*, vol. 45, no. 4, pp. 1–32, 2013.

[5] J.-G. Lee, J. Han, and K.-Y. Whang, "Trajectory clustering: a partition-and-group framework," in *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, 2007, pp. 593–604.

[6] Y. Zheng and X. Zhou, *Computing with spatial trajectories*. Springer Science & Business Media, 2011.

[7] R. Yan, S. Wang, L. Zhen, and G. Laporte, "Emerging approaches applied to maritime transport research: Past and future," *Communications in Transportation Research*, vol. 1, p. 100011, 2021.

[8] G. Yuan, P. Sun, J. Zhao, D. Li, and C. Wang, "A review of moving object trajectory clustering algorithms," *Artificial Intelligence Review*, vol. 47, pp. 123–144, 2017.

[9] G. Pallotta, M. Vespe, and K. Bryan, "Vessel pattern knowledge discovery from ais data: A framework for anomaly detection and route prediction," *Entropy*, vol. 15, no. 6, pp. 2218–2245, 2013.

[10] Z. Yan, Y. Xiao, L. Cheng, R. He, X. Ruan, X. Zhou, M. Li, and R. Bin, "Exploring ais data for intelligent maritime routes extraction," *Applied Ocean Research*, vol. 101, p. 102271, 2020.

[11] L. Zhao and G. Shi, "A novel similarity measure for clustering vessel trajectories based on dynamic time warping," *The Journal of Navigation*, vol. 72, no. 2, pp. 290–306, 2019.

[12] G. K. D. De Vries and M. Van Someren, "Machine learning for vessel trajectories using compression, alignments and domain knowledge," *Expert Systems with Applications*, vol. 39, no. 18, pp. 13 426–13 439, 2012.

[13] S. Capobianco, L. M. Millefiori, N. Forti, P. Braca, and P. Willett, "Deep learning methods for vessel trajectory prediction based on recurrent neural networks," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 57, no. 6, pp. 4329–4346, 2021.

[14] H. Li and Z. Yang, "Incorporation of ais data-based machine learning into unsupervised route planning for maritime autonomous surface ships," *Transportation Research Part E: Logistics and Transportation Review*, vol. 176, p. 103171, 2023.

[15] H. Li, H. Jiao, and Z. Yang, "Ais data-driven ship trajectory prediction modelling and analysis based on machine learning and deep learning methods," *Transportation Research Part E: Logistics and Transportation Review*, vol. 175, p. 103152, 2023.

[16] N. Zygouras, A. Troupiotis-Kapeliaris, and D. Zissis, "Envclus*: Extracting common pathways for effective vessel trajectory forecasting," *IEEE Access*, 2024.

[17] C. Huang, X. Qi, J. Zheng, R. Zhu, and J. Shen, "A maritime traffic route extraction method based on density-based spatial clustering of applications with noise for multi-dimensional data," *Ocean Engineering*, vol. 268, p. 113036, 2023.

[18] L. Wang, P. Chen, L. Chen, and J. Mou, "Ship ais trajectory clustering: An hdbscan-based approach," *Journal of Marine Science and Engineering*, vol. 9, no. 6, p. 566, 2021.

[19] L. Eljabu, M. Etemad, and S. Matwin, "Spatial clustering model of vessel trajectory to extract sailing routes based on ais data," *International Journal of Computer and Systems Engineering*, vol. 16, no. 10, pp. 491–501, 2022.

[20] ——, "Spatial clustering method of historical ais data for maritime traffic routes extraction," in *2022 IEEE International Conference on Big Data (Big Data)*. IEEE, 2022, pp. 893–902.

[21] X. Han, C. Armenakis, and M. Jadidi, "Modeling vessel behaviours by clustering ais data using optimized dbscan," *Sustainability*, vol. 13, no. 15, p. 8162, 2021.

[22] F. Farahnakian, F. Nicolas, F. Farahnakian, P. Nevalainen, J. Sheikh, J. Heikkonen, and C. Raduly-Baka, "A comprehensive study of clustering-based techniques for detecting abnormal vessel behavior," *Remote Sensing*, vol. 15, no. 6, p. 1477, 2023.

[23] A. Moavinis, A. Gounaris, and I. Constantinou, "Detection of anomalous trajectories for vehicle traffic data," in *Proceedings of the Workshops of the EDBT/ICDT 2023 Joint Conference, Ioannina, Greece*, 2023.

[24] I. Varlamis, I. Kontopoulos, K. Tserpes, M. Etemad, A. Soares, and S. Matwin, "Building navigation networks from multi-vessel trajectory data," *GeoInformatica*, vol. 25, pp. 69–97, 2021.

[25] H. Liu, X. Chen, Y. Wang, B. Zhang, Y. Chen, Y. Zhao, and F. Zhou, "Visualization and visual analysis of vessel trajectory data: A survey," *Visual Informatics*, vol. 5, no. 4, pp. 1–10, 2021.

[26] X. Zhang, X. Fu, Z. Xiao, H. Xu, and Z. Qin, "Vessel trajectory prediction in maritime transportation: Current approaches and beyond," *IEEE Transactions on Intelligent Transportation Systems*, 2022.

[27] "Marine Traffic," [Online]. Available: https://www.marinetraffic.com/en/ais/details/ships/shipid:322946/mmsi:265609540/imo:8619821/vessel:CINDERELLA_II, Nov. 2023.

[28] R. Yan, S. Wang, and C. Peng, "An artificial intelligence model considering data imbalance for ship selection in port state control based on detention probabilities," *Journal of Computational Science*, vol. 48, p. 101257, 2021.

[29] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *kdd*, vol. 96, no. 34, 1996, pp. 226–231.

MOHAMED ABUELLA received his M.S. and PhD degrees in Electrical and Computer Engineering from Southern Illinois University at Carbondale and University of North Carolina at Charlotte, in 2012 and 2018 respectively. He is a postdoctoral researcher at Halmstad University since 2022. His research interests include energy analytics and AI for sustainability.

M. AMINE ATOUI obtained his Ph.D. degree from LARIS, Polytech' Angers, France, in 2015. Currently, he is affiliated with the Center for Applied Intelligent Systems Research at Halmstad University, Sweden. His research interests encompass probabilistic and explainable Machine Learning, causal and Bayesian Inference, transmission/communication, and automatic control.

SLAWOMIR NOWACZYK is a Professor in Machine Learning at the Center for Applied Intelligent Systems Research, Halmstad University, Sweden. He received his MSc degree from Poznan University of Technology in 2002 and his PhD from the Lund University of Technology in 2008. During the last decades, his research has focused on machine learning, knowledge representation, and self-organising systems. The majority of his work concerns industrial data streams, often with predictive maintenance as the goal. Given that accurate and relevant labels are usually impossible to obtain in such settings, Slawomir's contributions primarily take advantage of proxy labels, such as transfer learning and multi-task learning, or concern semi-supervised and unsupervised modelling. He is a board member of the Swedish AI Society and a research leader for the School of Information Technology at Halmstad University. Slawomir has led multiple research projects on applying Artificial Intelligence and Machine Learning in different domains, such as transport and automotive, energy, smart cities, and healthcare. In most cases, this research was done in collaboration with industry and public administration organisations – inspired by practical challenges and leading to tangible results and deployed solutions.

SIMON JOHANSSON is a MSc graduate of Chalmers University of Technology's program in Engineering Mathematics and Computational Science in 2020, currently works in Cetasol, a marine company specialising in CO2 reduction and energy optimisation for vessels. His practical application of computational methods and dedication to environmental sustainability align with his role, contributing to global efforts to mitigate climate change. Simon's commitment to advancing eco-friendly solutions in the marine industry reflects a seamless integration of academic excellence and real-world impact.

ETHAN FAGHANI is the CEO and founder of Cetasol. Before Cetasol, Ethan was Chief Engineer of Automation and AI at Volvo Penta. Ethan has experience working with cutting-edge technologies in other transportation segments in both big enterprises and his own founded startup. Ethan obtained his Ph.D. in mechatronics from UBC and Innovation and Entrepreneurship from Stanford Business School.