Journal of *Marine Science and Engineering*

MDPI

Article

A Contextually Supported Abnormality Detector for Maritime Trajectories

Kristoffer Vinther Olesen [1,*,†], Ahcène Boubekki [2], Michael C. Kampffmeyer [2], Robert Jenssen [2,3,4], Anders Nymark Christensen [1], Sune Hørlück [5] and Line H. Clemmensen [1]

[1] Applied Mathematics and Computer Science, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark; anym@dtu.dk (A.N.C.); lkhc@dtu.dk (L.H.C.)
[2] Machine Learning Group, UiT The Arctic University of Norway, 9019 Tromsø, Norway; ahcene.boubekki@uit.no (A.B.); michael.c.kampffmeyer@uit.no (M.C.K.); robert.jenssen@uit.no (R.J.)
[3] Pioneer Centre for AI, University of Copenhagen, 1350 Copenhagen, Denmark
[4] Norwegian Computing Center, 0373 Oslo, Norway
[5] Terma A/S, 8520 Lystrup, Denmark; snhr@terma.com
* Correspondence: kvol@dtu.dk
† Work done while corresponding author was at UiT.

Abstract: The analysis of maritime traffic patterns for safety and security purposes is increasing in importance and, hence, Vessel Traffic Service operators need efficient and contextualized tools for the detection of abnormal maritime behavior. Current models lack interpretability and contextualization of their predictions and are generally not quantitatively evaluated on a large annotated dataset comprising all expected traffic in a Region of Interest. We propose a model for the detection of abnormal maritime behaviors that provides the closest behaviors as context to the predictions. The normalcy model relies on two-step clustering, which is first computed based on the positions of the vessels and then refined based on their kinematics. We design for each step a similarity measure, which combined are able to distinguish boats cruising shipping lanes in different directions, but also vessels with more freedom, such as pilot boats. Our proposed abnormality detection model achieved, on a large annotated dataset extracted from AIS logs that we publish, an ROC-AUC of 0.79, which is on a par with State-of-the-Art deep neural networks, while being more computationally efficient and more interpretable, thanks to the contextualization offered by our two-step clustering.

Keywords: maritime surveillance; vessel traffic service; AIS; maritime traffic patterns; trajectory clustering; anomaly detection

Citation: Olesen, K.V.; Boubekk, A.; Kampffmeyer, M.; Jenssen, R.; Christensen, A.N.; Hørlück, S.; Clemmensen, L.H. A Contextually Supported Abnormality Detector for Maritime Trajectories. *J. Mar. Sci. Eng.* **2023**, *11*, 2085. https://doi.org/10.3390/jmse11112085

Academic Editors: Sebastian Feuerstack, Marko Perkovic and Lucjan Gucma

Received: 22 September 2023
Revised: 20 October 2023
Accepted: 23 October 2023
Published: 31 October 2023

1. Introduction

According to the International Maritime Organization, international shipping is currently responsible for 80% of global trade and is the most efficient and cost-effective form of long-distance transportation [1]. Despite international efforts to combat maritime piracy, it remains a serious threat to international shipping and is estimated to have a global financial impact of up to 16 billion dollars annually [2]. Accidents such as collisions or groundings can lead to the loss of lives, environmental damage, and disruption of trade routes [3,4]. Furthermore, it is estimated that one fifth of all wild-caught fish is caught illegally or not reported, endangering marine ecosystems and resulting in industry losses upwards of 23.5 billion dollars [5]. Most recently, the sabotage of the Nord Stream gas pipelines in the Baltic Sea [6] has raised the issue of territorial protection and protection of key infrastructure assets. These vulnerabilities in our society highlight the need for maritime security, safety, and threat assessment, to protect the stability of the global supply chain and key infrastructure.

Real-life maritime laws and regulations are complex. While commercial vessels such as cargo and tankers mainly follow well-defined shipping lanes with near-constant speeds, other ship types, such as fishing vessels and sailing ships, have fewer constraints and more

complex behaviors. Maritime security requires extensive knowledge of maritime traffic patterns. Research in this field has gained momentum over the past decade, thanks to the advent of the Automatic Identification System (AIS) [7]. The AIS is compulsory for all vessels that exceed a certain tonnage, and it provides hundreds of millions of messages every day on a global scale [8]. The data include static information, such as the unique identifier of the ships (MMSI), size, and dynamic information such as Global Positioning System (GPS) coordinates, speed, course, etc. The AIS forms the basis for modern maritime trajectory data collection, which allows one to model navigational characteristics and rules. The wide variety of possible maritime behaviors means that the most common method of analysis remains clustering, either of individual AIS updates [9,10] or of trajectories compared using specific similarity measures [11,12].

Dense and critical maritime areas are constantly monitored, but the ever-increasing traffic and amount of data call for automatic decision support for Vessel Traffic Service (VTS) operators. In practice, whether or not an event is abnormal is a combination of several factors: the location, the speed, the course, the type of vessel, and the time of day/week/year, etc. For instance, in specific locations, pilot boats steam between harbors and commercial traffic in the shipping lanes, but in other locations or for different ship types, this type of behavior may be highly unexpected. Similarly, it can be expected that diving vessels perform frequent starts or stops to support divers in the water, but if this behavior occurs near major shipping lanes, it can be considered abnormal unless permission has been granted. Automatic detection of abnormal maritime behaviors is thus a difficult, ill-defined problem that requires the disentanglement of multiple possible explanatory factors, of which location, kinematic behavior, and type of ship are the most important.

Previous works have shown that shipping lanes can be identified using only a positional clustering of vessels [9,11,13]. However, to disentangle traffic not constrained to major shipping lanes, the exact route is less important than the local kinematic behavior—that is, changes in speed and course [10,14]. Recently, deep neural network models have been suggested for abnormality detection of multiple ship types [13,15]. These models classify as anomalies the trajectories for which they fail to predict the future position or to make an adequate reconstruction. The main drawback here is the lack of interpretability as to why the networks fail to predict/reconstruct the correct trajectory. According to Riveiro et al. and Stach et al. [16,17], abnormality detection models should offer a large degree of interpretability, to accommodate any skepticism VTS operators may have, and they should promote human–machine interaction [16,17]. At the same time, recent research on dynamical decision making indicates that automated systems for decision making tend to perform poorly, as human operators simply copy the decision of the automated system [18]. This causes a degradation of operator experience and makes human operators less likely to take over manual control when needed. Furthermore, many previous studies simplify their data, to focus on a restricted Region of Interest (ROI) and on a single aspect of maritime traffic—for instance, the behavior of commercial merchant traffic [19] or port entry/exit ways [11,12,20]. However, such restrictions hinder the evaluation of the practical viability of the methods. Stach et al. [17] further highlighted the need for standardized datasets annotated with maritime abnormalities, to bridge the gap between research and practical implementation in VTS operations.

A precise definition of what constitutes a maritime abnormality is difficult to state. To this day, however, many maritime surveillance operations are conducted manually by military or law enforcement. Thus, the reasons for flagging a behavior as abnormal are often classified information, and the specific type of behavior that interests operators is not fully known. For this reason, obtaining a large list of annotated maritime trajectories for training or evaluation is often impossible. Previous works [21,22] have used simulated or self-annotated labels based on extreme values. However, extreme values might not define abnormal trajectories of operational interest for surveillance operators, who are ultimately interested in deterring illegal or dangerous activities, such as the earlier examples. As for unsupervised methods, the most common way to evaluate them is through qualitative

examples [15,23,24]. This form of verification illustrates the potential type of abnormal behavior that could be flagged. However, it completely negates the issue of false negatives, which in a military or law enforcement operation may be of greater importance. Ideally, abnormality detection algorithms would be evaluated on datasets with known behavior of operational interest to operators. These datasets should be annotated by subject matter experts, and the annotations should reflect the degree to which the operators find the behavior suspicious or otherwise abnormal.

In this paper, we aim to overcome two gaps in the current research on vessel traffic abnormality detection: the lack of interpretability and contextualization of the predictions and the size and quality of the datasets used for evaluation. First, we introduce an abnormality detection algorithm that provides an explanation of the closest expected/normal behaviors. The normalcy model is learned using a two-step clustering method that disentangles positional and kinematic behavior. The training is performed on historical AIS data and consists of two stages: in the first step, a clustering is learned based on the positional data of the trajectories; in the second step, each positional cluster is refined on the basis of the kinematics of its trajectories—that is, speed and course. The final clustering of the whole data is thus a summary of the typical behavioral patterns in the area. Concurrently, a Local Outlier Factor (LOF) [25] is trained on each positional cluster, also based on the kinematic data, to detect trajectories with abnormal speed and course sequences. In a practical scenario, a new trajectory is first assigned to one of the positional clusters. If flagged as abnormal, using an LOF, it is brought to the attention of an operator who can further assess the situation, using the kinematic clustering of the positional cluster as support.

Next, we present and use for the evaluation two large hand-annotated AIS traffic datasets for abnormal maritime behavior detection that we have created, based on contextual knowledge of the environment and news events. These datasets contain more than 30,000 trajectories from 11 types of vessels and are expected to be of operational interest to operators. Labeled abnormalities cover a full day and include a collision accident, Search and Rescue activity, and deviating commercial traffic. The collision accident is by itself an important test case, but may also serve as a proxy evaluation method for the detection of rendezvous situations that are of interest in finding smuggling events. Similarly, Search and Rescue activity is always of interest, especially as similar behavioral patterns may be seen in more nefarious activities like smuggling and illegal mapping of the seabed or seafloor infrastructure [26]. For the sake of reproducibility and to foster research on methods suitable to a real-world scale, we provide public access to the datasets. Further details are given in Section 4.1.

We summarize our contributions as follows:

- We present a novel method for detecting abnormal maritime trajectories based on two-step clustering, which also provides a contextual decision support tool to help a VTS operator make the final decision.
- We design positional and kinematic similarity measures that focus on different dimensions of maritime trajectories.
- We provide evidence that a multi-step clustering approach can disentangle positional and kinematic information, resulting in a better description of behavioral patterns in a large ROI.
- We provide public access to datasets of preprocessed maritime trajectories in regions of Danish waters, including annotations during a Search and Rescue event.

The paper is organized as follows. In Section 2, we provide an overview of the related work within trajectory clustering. In Section 3, we present our proposed two-step clustering. In Section 4, we give a detailed description of the maritime traffic datasets that we publish and that we also use to show the ability of our proposed method to disentangle maritime traffic patterns, as well as to detect real-life abnormal trajectories from a ship collision. Finally, we present our conclusion in Section 5.

2. Related Work

Clustering of maritime trajectories has been widely employed to extract traffic patterns and find abnormal trajectories. The type of behavior discovered by clustering spatio-temporal trajectories depends heavily on the chosen similarity measure. Laxhammar et al. [27] suggested using the maximum synchronous Euclidean distance between each pair of coinciding points along two trajectories of the same length. The requirements of equal trajectory length and synchronous comparison can be relaxed by using either the Hausdorff distance [12,19,28], Dynamic Time Warping (DTW) [11,29], or the Longest Common Subsequence (LCSS) [24] to measure trajectory similarity. The Hausdorff distance is independent of the time component, which can make trajectories following the same route in opposite directions indistinguishable. In addition, from the clusters reported in [19], we note that the Hausdorff distance may assign a large similarity to significantly different trajectories. Additionally, both of these methods have quadratic time complexity, and several works, therefore, suggest a compression using the Douglas–Peucker (DP) algorithm [30]. Klaas et al. [29] proposed a two-stage DP algorithm: first, reducing the trajectory based on the speed time series and, secondly, based on the position. This two-stage approach was found to better retain periods of acceleration, such as stops.

Several different clustering algorithms have been applied to clustering of trajectories. Methods such as K-means [29] and K-medoids [28] have been utilized in collaboration with different similarity measures. However, density-based clustering techniques have long been the predominant approach to data mining within maritime trajectory analysis. Pallotta et al. [31] proposed the widely used TREAD method, to cluster trajectories into traffic routes, which can then be used for anomaly detection and trajectory prediction. TREAD is a point-based method that extracts the coordinates of new entries, exits, and stops within the ROI. These points are clustered using DBSCAN [32], to form waypoints in which ships enter, exit, or stop within the ROI. A route between waypoints is then formed whenever a certain number of transitions between them have been observed. Several works [11,12,19,24] have combined the idea of a similarity measure and density-based clustering. First, trajectories are simplified using the DP algorithm. The similarities are then computed using the Hausdorff distance, DTW, or LCSS, before being clustered by DBSCAN. Wang et al. [19] considered a hierarchical search over the hyperparameters of DBSCAN, which allowed for groups with different densities, and helped to find clusters in sparsely populated geographical regions.

Recently, trajectory similarities based on deep learning have been suggested. Murray et al. [13] clustered the latent encodings of a Recurrent Variational Autoencoder (RVAE) trained for trajectory reconstruction using hierarchical DBSCAN and found clusters corresponding to the major shipping lanes. The clusters were then used to train neural networks to predict the future position. Luo et al. [33] proposed a graph-based trajectory contrastive learning framework. A Graph Neural Network encoder was trained, using contrastive learning with five different trajectory augmentations. The similarity of two trajectories could then be computed by their distances in the latent space. The method was evaluated by downsampling random trajectories from the training set as test trajectories. The proposed similarity measure was found to perform better than traditional trajectory distance measures.

The abovementioned approaches only considered the positional input, yielding clusters that mostly corresponded to the primary shipping lanes. Zhen et al. [28] introduced the difference of the average course in their similarity measure, and Liu et al. [10] extended the DBSCAN clustering model, to consider not only the geographical distance of the coordinates, but also the difference in speed and course. This allowed them to distinguish between shipping lanes in opposite directions and to find speed differences within the main shipping lanes. However, the work was limited to small geographical areas and a limited number of ship types. Li et al. [23] suggested a similar extension to the DBCSAN algorithm but split the speed and course differences into two different clustering models.

Knowledge about maritime traffic patterns is useful for detecting abnormal activity. Widyantara et al. [24] directly reported outliers from the DBSCAN clustering, but several

clustering methods have been extended with a detection step. Often, this step includes knowledge about the kinematic behavior. Pallotta et al. [14] proposed a two-stage anomaly detection scheme, using the routes extracted by TREAD. First, using only the positional inputs, a trajectory was associated with a route. Afterwards, kinematic outliers were found, by comparing the speed and course to the average behavior of the route. Liu et al. [10] proposed to divide the clusters into smaller geographical regions and compute the average kinematic values for each split. These values would then be used to detect abnormalities [20]. Zhao et al. and Li et al. [23,34] used normal trajectories determined from DBSCAN clusters to train deep neural networks for trajectory prediction. Abnormalities were then detected, based on the prediction error.

Recently, an abnormality detection model based purely on deep learning has been suggested. Hu et al. [21] suggested an ensemble of a Variational LSTM AutoEncoder and a Graph Variational AutoEncoder. Each ensemble member was trained to reconstruct the input trajectory, and the reconstruction errors were then combined, to make a final binary prediction of the abnormality. Liu et al. [22] self-annotated training data based on extreme position, speed, or course values and trained a deep neural network to classify abnormalities. Nguyen et al. [15] suggested a Variational Recurrent Neural Network (VRNN) for the detection of abnormalities based on trajectory reconstructions. In this work, Nguyen et al. also suggested an A-Contrario detection methodology, which was supposed to account for regional differences in reconstruction accuracy. Although the results reported using VRNN looked promising, our feedback from VTS operators mentioned the lack of explainability as a key limitation for operational use. The lack of explainability of decision support tools has been identified as a key issue for the automated detection of abnormal maritime behavior in surveys by Riveiro et al. and Stach et al. [16,17].

Table 1 summarizes the normalcy models and the limitations of these normalcy models utilized by the previous research discussed. We present a novel abnormality detection algorithm based on a positional clustering followed by a kinematic clustering of historical maritime trajectories. We rely on an efficient positional similarity measure, which allows us to process a large, complex dataset of maritime trajectories representative of real-life traffic in a reasonable time. The abnormality detection is made with respect to the kinematic of the vessel, for which we design an alternative similarity measure based on DTW. The latter is able to distinguish behaviors within the same positional clustering, giving VTS operators a clear summary of normal behavior when assessing the suggestion of our abnormality detector.

Table 1. Summary of the normalcy models and the limitations of these normalcy models utilized by the previous research.

Normalcy Model	Limitation of Normalcy Model	Works
Clustering of individual updates	Applied on restricted datasets Lack description of kinematic behavior	[20] [9,14]
Clustering of trajectory similarities	Applied on restricted datasets Lack description of kinematic behavior	[12,19,28] [12,19,24,27,29,33]
Deep learning methods	Interpretability	[15,21–23,33,34]

3. Methodology

In this section, we first discuss similarity measures for trajectories and then introduce our abnormality detection algorithm.

3.1. Notations

An AIS trajectory A of length $T_A \in \mathbb{N}$ is a four-dimensional time series $A = (a_1, \ldots, a_{T_A})$, where $a_t = (\text{lon}_t, \text{lat}_t, s_t, c_t)$, with each dimension representing, respectively, the longitude, latitude, speed, and course of the vessel as recorded in its AIS message at time t. For

legibility, the timestamp t is used indiscernibly as an index of a variable, such that $A(t) = \mathrm{a}_t$, or $s(t) = s_t$.

Throughout the section, we consider two AIS trajectories A and B of time duration T_A and T_B, and two timestamps $t \in \{0, \ldots, T_A\}$ and $\tau \in \{0, \ldots, T_B\}$. Also, we assume that the trajectories are regularly sampled without missing data. The function d is a generic distance on $\mathbb{R}$ or $\mathbb{R}^2$, depending on the context.

3.2. Similarity Measures

In the following, we discuss three commonly used trajectory similarity measures: the Hausdorff distance, the average Haversine distance, and Dynamic Time Warping (DTW). We define these similarity measures without specifying which dimensions of the time series are used (positional or kinematics), as this depends on the use case. We also propose a variant of DTW tailored to kinematic data.

3.2.1. Hausdorff

The Hausdorff distance [35] between two trajectories corresponds to the maximum smallest distance realized by any pair of points in each one of the trajectories:

$$\mathrm{Hausdorff}(A, B; d) = \max_{t \in [0, T_A - 1]} \min_{\tau \in [0, T_B - 1]} d\Big(A(t), B(\tau)\Big). \tag{1}$$

The computations require a comparison of all possible pairs of points, resulting in quadratic time complexity. Furthermore, the Hausdorff distance ignores the time component. This means that ships along parallel shipping lanes sailing in opposite directions are not distinguishable. Such a situation is studied in [19].

3.2.2. Average Haversine

The quadratic time complexity and the issues mentioned above make the Hausdorff distance unsuitable for measuring the similarity of many long and complex sequences of geographical coordinates. On the other hand, the Average Haversine distance (AH) proposed in [36] is able to compare the positional evolution of the AIS trajectories in linear time, with respect to the length of the trajectories. It is defined as a continuous distance measure, but it can be approximated using the trapezoidal rule and assuming a regular sampling:

$$\mathrm{AH}(A, B; d_H) = \sum_{t=0}^{T-1} \frac{d_{\mathrm{H}}\Big(A(t), B(t)\Big) + d_{\mathrm{H}}\Big(A(t+1), B(t+1)\Big)}{2T}, \tag{2}$$

where $T = \min(T_A, T_B)$ and d_{H} is the Haversine distance [36]. This similarity measure computes the geographical distance between the trajectory points one by one in a linear fashion until the length of the shortest trajectory is reached. This means that the measure places an increased weight on the beginning of the trajectories. Thus, we expect the measure to be able to separate trajectories based on their starting location. This is ideal in a real-time operational setting when observing new trajectories, as even short trajectories can very quickly be classified into a subset of historical trajectories with similar behavior.

3.2.3. Dynamic Time Warping

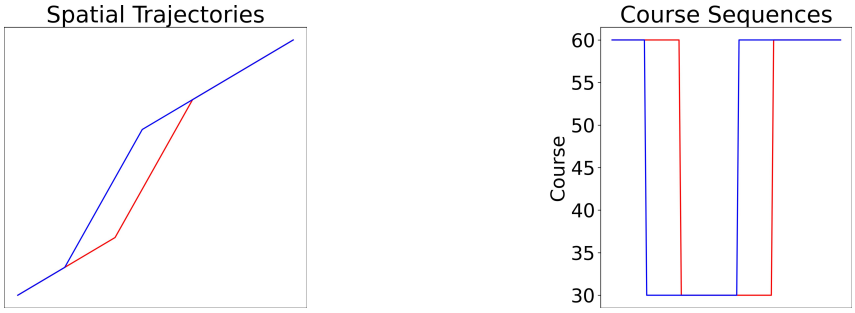
Dynamic Time Warping minimizes the pair-wise distance by re-indexing (alignment of) the data points in the trajectories, according to certain rules. It can be defined as follows:

$$\mathrm{DTW}(A, B; d) = \min_{\pi \in \Pi(T_A, T_B)} \left(\sum_{(i,j) \in \pi} d\Big(A(t_i), B(t_j)\Big) \right), \tag{3}$$

where $\Pi(T_A, T_B)$ is the set of all possible alignments that are sequenced pairs of indices $(i, j) \in [0, T_A - 1] \times [0, T_B - 1]$ satisfying three constraints: (1) the beginning and end of the time series must be matched; (2) the sequence must be monotonically increasing in i and j;

(3) all indices i and j must appear at least once. These ensure that the sequences start and end together and that each point on either sequence is mapped onto at least one point of the other sequence without these mappings crossing in time.

As DTW processes pairs of indices, it also has a quadratic time complexity. The DTW alignment may overestimate the distance of trajectories with similar behavior if this behavior is spread over a large area. For example, consider two trajectories with the same starting point and sailing along the same direction as illustrated in Figure 1. At one point, trajectory A makes a 30 degree turn and continues in this direction, moving away from trajectory B. Later, trajectory B makes a similar 30 degree turn and continues parallel to trajectory A. Both vessels return to their initial course some time later, and the trajectories terminate at the same point. As these two trajectories have the same origin and terminal location and have similar behaviors throughout the journey, we would expect the distance between them to be very small. However, their distance, calculated by DTW on the sequence of geographical coordinates, may be significant. The re-indexing procedure of DTW aligns the course changes between the trajectories. However, due to the spatial nature of the geographical coordinates, DTW calculates the geographical distance between the location where the trajectories changed course. If we instead were to use the time series of the measured angles towards true north, the DTW distance would calculate the difference of the course values. As these values are the same before and after the changes, the DTW distance between the two trajectories would be zero. Using the time series data, we remove the spatial dependence, and DTW can properly calculate the similarity of the course after aligning the changes. Therefore, DTW is a good candidate as the building block of a similarity measure for course and speed time series.



(a) DTW distance: 45.38 (b) DTW distance: 0.00

Figure 1. (**a**) Two trajectories with the same origin and terminal point and similar behaviors throughout the journey may obtain a large distance calculated by DTW; (**b**) however, the course sequences of the two trajectories may be warped perfectly onto each other and have zero distance between them calculated by DTW.

3.2.4. Kinematic DTW

Following the previous discussion, we propose a variation of DTW for kinematic data, referred to as D_{kin}. The measure is defined as the sum of the DTW of the time series of the speed and of the DTW of the course time series:

$$D_{\text{kin}}(A, B) = \text{DTW}(s_A, s_B; d_{\text{speed}}) + \text{DTW}(c_A, c_B; d_{\text{course}}), \tag{4}$$

where s_A, s_B and c_A, c_B are, respectively, the speed and course sequences of trajectories A and B. The differences in speed and course at each timestamp are measured, respectively, by d_{speed} and d_{course}, which correspond to the standardized absolute difference of the speed and the normalized angular difference in radians, respectively:

$$d_{\text{speed}}(x, y) = \frac{|x - y|}{\Sigma}, \tag{5}$$

$$d_{\text{course}}(x,y) = \frac{1}{\pi} \cdot \begin{cases} |x-y| & \text{if } |x-y| \leq \pi, \\ \pi - (|x-y| \bmod \pi) & \text{otherwise,} \end{cases} \tag{6}$$

where Σ is the standard deviation of the speed computed empirically from the speed time series s_A and s_B.

3.3. Two-Step Clustering for Abnormality Detection

Our intention was to design an abnormality detection algorithm to assist VTS operators, which may serve as a contextual decision support tool and let them make the final decision based on the contextual information provided by the algorithm itself. The reason for a trajectory to be flagged as abnormal is that it is either similar to other abnormal trajectories or that it diverges from the most similar non-abnormal trajectories. It is important to state that the notion of the behavior of a vessel is not limited to a sequence of locations, but also includes its speed and course. The similarity measure involved, to compare trajectories, thus needs to take into account both the spatial and the temporal information. Note that the assignment of kinematic clusters gives a context to the prediction of the LOF, as it shows the most similar trajectories. Yet it is not an explanation, as the kinematic clustering is not used by the detector.

We modeled this line of thought as a two-step algorithm:

1. Assign an input trajectory to a cluster, based on its positional dimensions (latitude, longitude, and time).
2. Decide on abnormality, based on the kinematic dimension (speed and course), and provide a context to the decision with the most similar trajectories.

Figure 2 shows the flow of our proposed two-step method.

(1) Positional Clustering: The first step required clustering of a historical database and a classifier, using a fast-to-compute similarity measure, to ensure the reactivity of the system. Hence, the Hausdorff and DTW were excluded. Also, both measures either distort or simply disregard the time component, which is at odds with the rationale exposed above. We chose to rely, both for clustering and classifying, on the average Haversine distance (Equation (2)), which has a linear complexity and compares synchronous positions. The clustering was a hierarchical clustering with average linkage which, once computed, allowed us to easily change the number of clusters and, thus, isolate outliers. During inference, cluster assignment was decided by a K-Nearest Neighbors (KNN) classifier with $k = 3$ trained on the clustering.

(2a) Abnormality Detection: As we did not have access to a large set of labels, we made the assumption that none of the historical trajectories were abnormal. Therefore, none of the positional clusters were considered as abnormal, and abnormality was defined as a divergence from the training set. More precisely, we defined it as a divergence with respect to the trajectories within the assigned positional cluster. As outlier detection, we employed the Local Outlier Factor (LOF) [25] and D_{kin} (Equation (4)) as a similarity measure. Before calculating D_{kin}, we compressed the trajectories, using the two-stage DP compression [29]. As the trajectory had been assigned to a cluster based on its positional information, it could not be an outlier purely based on these data. The divergence needed to be measured on the basis of another aspect of the behavior, namely the variations of the kinematics (speed and course), which could be understood as the derivative of the positional data.

The LOF compares the density of the local neighborhood of a point to that of its KNN. If the density of a point is significantly lower than its neighbors, the point is flagged as an outlier. Following the discussion in [27], we set $k = 5$ nearest neighbors for the LOF algorithm in our experiments. In practice, we did not see large changes in the number of outliers detected when varying that number. However, we recommend a low value to capture information only from the local neighborhood. The LOF also has a hyperparameter, called contamination, related to the expected percentage of outliers. As we expect only a small number of outliers, we recommend again to use small values for the hyperparameter. See Section 4.6.2 for an ablation study.

(2b) Kinematic Clustering: The explanation or context of the prediction of the LOF consists of the most similar historical trajectories. Again, with the aim of speeding up calculations, these trajectories will be extracted from a precomputed hierarchical clustering with average linkage of the trajectories of the assigned positional cluster, using D_{kin} as a similarity measure. The cluster assignment is decided by a KNN classifier with $k = 3$ using D_{kin} and trained on that kinematic clustering.

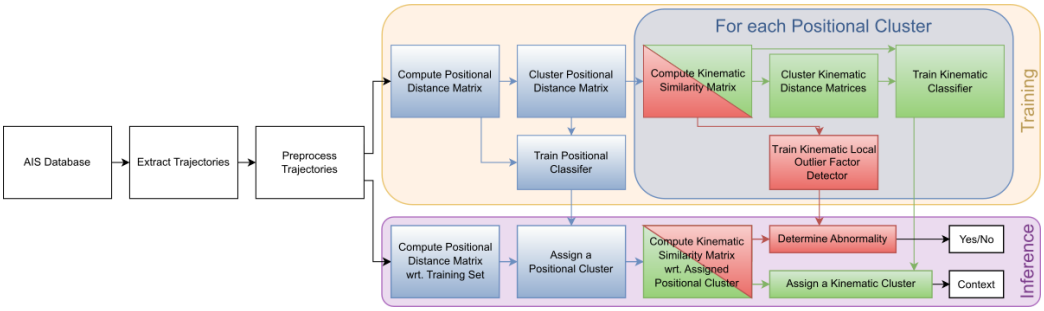


Figure 2. Flowmap of the training and inference of our proposed two-step abnormality detector. In the training phase, the positional distance matrix based on the Haversine distance is computed for the preprocessed training dataset. The matrix is then used to train a hierarchical clustering base and a KNN classifier for that clustering. For each positional cluster, the kinematic similarity matrix based on D_{kin} is computed and then used to train an LOF (red), a hierarchical clustering (green), and a KNN classifier for that clustering. In the inference phase, after preprocessing, the distances between the input trajectory and the training data are used by the positional classifier (blue) to assign a position cluster. In the second step, the kinematic similarities to the trajectories within the same cluster are computed and then used by, on the one hand, the LOF to determine abnormality (red) and, on the other hand, by the kinematic classifier to assign a kinematic cluster (green). The output is thus double: the answer of the LOF and a context, in the form of the trajectories of the kinematic cluster.

4. Experimental Results

In this section, we evaluate the choice of similarity measure and of algorithm for each positional and kinematic clustering. We also compare the final clustering to that of a single-step algorithm. Finally, we discuss the abnormality detection capabilities of our approach and compare it to State-of-the-Art neural-network-based baselines.

4.1. Datasets

For this work, we built two datasets of AIS data from Danish waters that cover large ROIs and contain various types of vessels with different priorities and expected behavior patterns. Both datasets are available for public use, to facilitate reproducibility and to give researchers the ability to evaluate their proposed models on a complex dataset representative of a real-world setting [37] (datasets available at https://data.dtu.dk/collections/AIS_Trajectories_from_Danish_Waters_for_Abnormal_Behavior_Detection/6287841. Accessed on 20 October 2023). The complete AIS data from all Danish waters are available publicly [38]; however, minor differences between the two sources may occur.

The first dataset covers a rectangular ROI covering the island of Sjælland, bounded by (54.4° N, 10.5° E) to (56.4° N, 13.5° E). The data were collected during November 2021 and contain 18.738 of trajectories from 11 different types of ships, ranging from commercial cargo and tanker ships to private sailing and fishing boats. The second dataset covers a rectangular ROI around Bornholm Island bounded by (54.5° N, 13° E) to (56° N, 16° E). The data were collected during December 2021 and contain 12.591 of trajectories from 8 different types of ships. The speed was limited to 20 m/s, and updates with higher speeds were discarded. For both datasets, if the time interval between two successive AIS messages exceeded 15 min, the trajectory was split into two contiguous trajectories. Trajectories shorter than 10 min were discarded and trajectories exceeding 12 h were divided into smaller trajectories, each between 10 min and 12 h. All trajectories were resampled every

120 s, using linear interpolation. Table 2 shows an example of the information associated with each extracted trajectory.

Table 2. Example information associated with a trajectory from a passenger ship.

MMSI	Timestamp	Latitude	Longitude	Speed	Course
211149000	2021-11-29 22:47:39	54.40	12.16	7.13	25.83
211149000	2021-11-29 22:49:39	54.41	12.17	7.12	27.30
211149000	2021-11-29 22:51:39	54.42	12.18	7.15	27.05
211149000	2021-11-29 22:53:39	54.42	12.18	7.16	27.40
⋮	⋮	⋮	⋮	⋮	⋮
211149000	2021-11-30 09:45:39	54.41	11.91	8.16	122.16
211149000	2021-11-30 09:47:39	54.40	11.90	8.09	233.83

We used the Sjælland data to evaluate the proposed two-step clustering algorithm. The entire dataset was used for training. We evaluated the proposed automated anomaly detection algorithm on the Bornholm dataset. Data from 13 December 2021 was withheld as a test set, the rest serving as a training set. On that day, a collision accident between two ships occurred, causing several abnormal trajectories. Trajectories from this day were manually labeled, resulting in 25 abnormalities out of 521 trajectories. In addition to the colliding vessels, the abnormal trajectories corresponded to commercial traffic, which had to deviate from the planned course, for Search-and-Rescue and law-enforcement vessels responding to the accident, and to any other vessel taking part in the search for the two missing sailors.

4.2. Experimental Setting

Similarity measure baselines include the Hausdorff distance and DTW, both based on Haversine distance, as suggested in [11,12]. In terms of the clustering algorithm, we compared hierarchical clustering and DBSCAN. The linkage distance threshold for hierarchical clustering was decided using the Kneedles algorithm [39], to select the number of clusters. The hyperparameters of the DBSCAN were tuned by creating candidate lists of minimum distances and samples, as suggested in [12]. The optimal value of these candidates was then determined, using the Kneedles algorithm [39]. We have provided quantitative, qualitative, and runtime analyses of the clustering. We quantitatively evaluated the clusterings, using the Silhouette score [40]. The qualitative evaluation was performed by manually gauging the similarity of the behaviors of the extracted clusters while also accounting for the purpose of the two steps. In the positional clustering, the primary purpose was to be a fast clustering of all the trajectories into groups, in which similar behaviors might be discovered. As such, we were looking for trajectories that originated in the same area and shared some common positional evolution. In the kinematic clustering, we were interested in clusters describing uniquely different maritime behavior, i.e., we wished to identify different behavioral patterns across clusters. The baselines for abnormality detection included the State-of-the-Art VRNN [15] and RVAE [13] deep learning architectures. We measured the performance of the detectors, using the area under the receiver operating characteristic (AUC). The code was implemented in Python 3.8, using standard libraries, and it ran on an Intel Xeon Processor 2660v3. Similarity calculations were parallelized across eight cores.

4.3. Positional Clustering

The positional clustering is the basis for the outlier detector and the kinetic clustering. It needs, thus, to separate well-different behaviors, e.g., by distinguishing vessels traveling along shipping lanes in different directions. We considered different combinations of distance measures (Hausdorff, DTW based on Haversine distance, and average Haversine distance computed using Equation (2)) and clustering algorithms (hierarchical, DBSCAN). Recall that our model combines the average Haversine distance with hierarchical clustering.

4.3.1. Quantitative Analysis

In Table 3, we report various statistics on the clustering of each combination of distance measure and algorithm, including hyperparameters selected using the Kneedles algorithm, the number of clusters found, the median number of members in each cluster, and, when DBSCAN was used, the percentage of outliers. The quality of the clusterings was measured in terms of silhouette score (the larger, the better). Our combination of hierarchical clustering with the average Haversine distance achieved the best score.

Table 3. Positional clustering performance, in terms of the silhouette score for various combinations of distance and clustering algorithms, along with the hyperparameters and characteristics of the clusterings. Bold denote the highest recorded silhouette score. Our model corresponds to the last line.

Distance Measure	Clustering Algorithm	Eps-Threshold	MinSamples	# Clusters	Median of # of Members	% Outliers	Silhouette Score
Hausdorff	Hierarchical	9000	-	1515	2	-	0.535
Hausdorff	DBSCAN	12,504	242	15	569	45.7	0.127
Hausdorff	DBSCAN	12,504	25	53	110	10.7	0.376
Hausdorff	DBSCAN	27,000	242	7	1446	12.6	0.265
DTW	Hierarchical	140,000	-	2862	1	-	0.349
DTW	DBSCAN	60,941	91	7	190	68.8	−0.454
Avg. Haversine	DBSCAN	1.07	261	14	485.5	57.8	−0.033
Avg. Haversine	Hierarchical	10	-	52	232	-	**0.651**

The silhouette scores show that DBSCAN generally performed worse than hierarchical clustering. One explanation could be the large number of trajectories flagged as outliers by DBSCAN. For all three similarity measures, DBSCAN considered at least 45% of the data as outliers. This is too many false positives for an automated system to be useful. Despite the better silhouette scores, hierarchical clustering with the Hausdorff distance and DTW suffered from a similar phenomenon. In fact, both combinations produced the largest number of clusters. Most of these clusters contained very few trajectories and, thus, served a similar purpose as the outliers in DBSCAN.

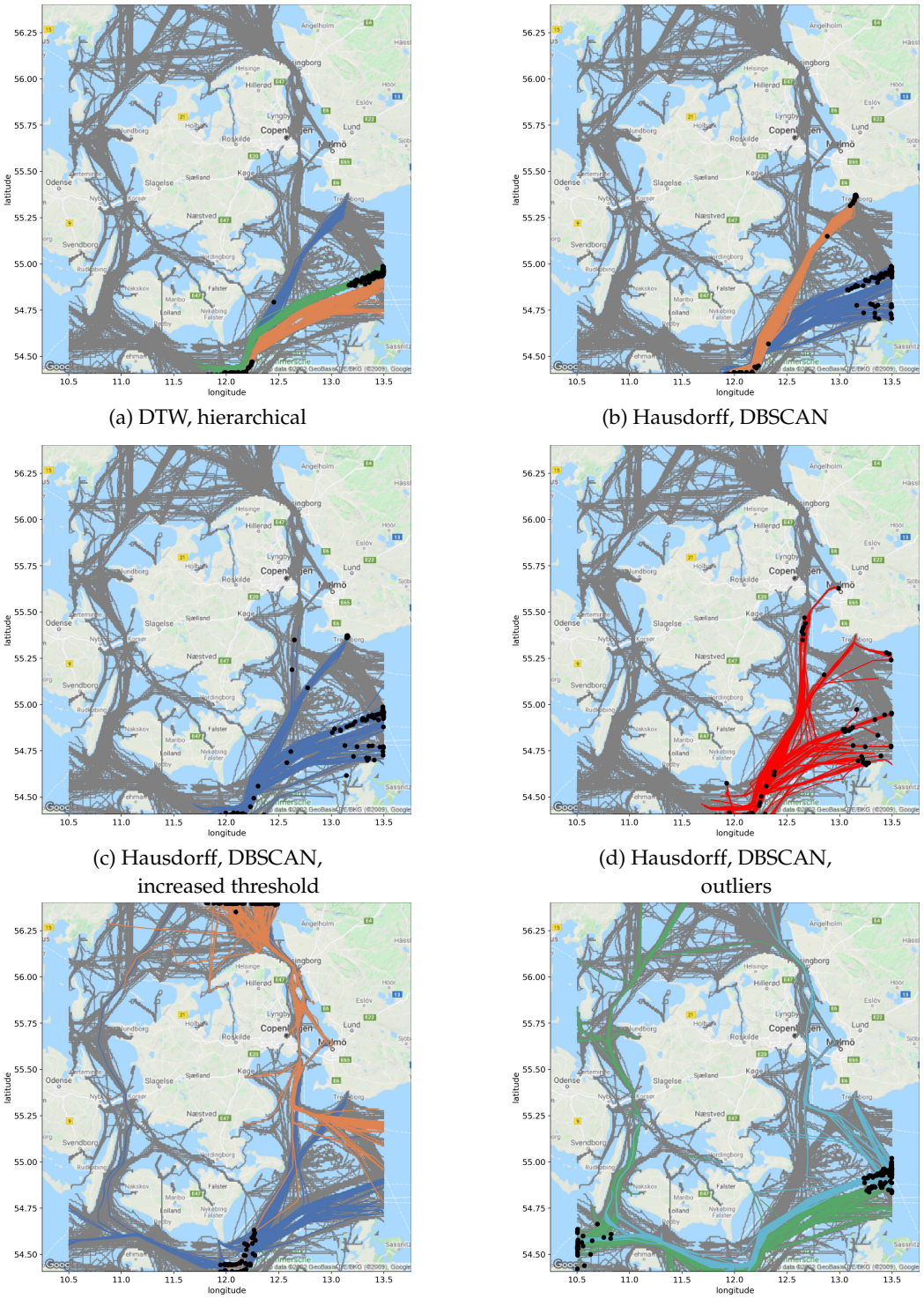
On the other hand, our proposed combination found a reasonable number of 52 clusters with a median number of trajectories per cluster of 232. These were reasonable numbers that allowed for further analysis in the second step. Note that these numbers of trajectories per cluster are comparable to the size of the full datasets used in most other works, such as [11,12].

4.3.2. Qualitative Analysis

Clustering using DTW or Hausdorff distances resulted in clusters corresponding to well-defined shipping lanes, as seen in Figure 3a,b. However, they failed to cluster two types of trajectories: small groups of trajectories on less populated maritime routes and trajectories that shared a segment along a shipping lane but did not follow it; see Figure 3d. These trajectories were marked as outliers by DBSCAN or as single-observation clusters by hierarchical clustering. Fine-tuning the hyperparameters might reduce the number of outliers by slowly admitting trajectories with similar positions into the clusters but with the risk of joining clusters of different shipping lanes, as shown in Figure 3c. This indicates that real, unfiltered trajectory recordings from a diversely populated ROI have too much randomness for these combinations of measures and algorithms to find well-separated clusters using only the latitude, longitude, and the timestamps without flagging the majority of the data as outliers.

In Figure 3e,f, we plotted the four most populated clusters in the ROI around Sjælland, using our positional clustering algorithm. Each cluster shown contained more than 1000 trajectories. We see that in each cluster, the trajectories began in the same geographical area unique to each cluster. This was expected, due to the increased attention by the average Haversine distance to the initial part of the trajectories. The discovered clusters contained

trajectories from all different shipping lanes that originated in a given area. However, this was acceptable, as we expected the second-step clustering to separate the shipping lanes based on their common kinematic behaviors and the outlier detector to catch those that did not travel steadily along the lanes.



(a) DTW, hierarchical

(b) Hausdorff, DBSCAN

(c) Hausdorff, DBSCAN, increased threshold

(d) Hausdorff, DBSCAN, outliers

(e) Avg. Haversine, hierarchical

(f) Avg. Haversine, hierarchical

Figure 3. Examples of clusters and outliers discovered in the first positional step, using different similarity measures and clustering methods. Clustered trajectories are shown in color, historical traffic is shown in gray, and trajectory origins are denoted by a black circle.

Note that the blue cluster of Figure 3e looks similar to that of Figure 3c. However, the difference is major. In Figure 3e, all the trajectories start in the same area, while in Figure 3c there are trajectories starting where others end. This is due to the fact that the Hausdorff distance compares asynchronous pairs of positions and is thus invariant to the direction of travel. This is a potential problem for the real-time classification of incomplete trajectories in the discovered clusters.

4.3.3. Runtime Analysis

In Table 4, we report average runtimes for computing the average Haversine distance, based on Equation (2), DTW, and the Hausdorff distance. The distances were implemented in Python 3.8.11 programming language, using Numpy 1.23.2. The Hausdorff and DTW distances were calculated using the trajectory_distance library (https://github.com/bguillouet/traj-dist accessed on 20 October 2023), implemented in Cython 0.29.24. With its linear time complexity, the average Haversine distance is undoubtedly the fastest to compute: it is 10 times faster than DTW and 100 times faster than the Hausdorff distance.

Table 4. Average time in seconds to compute a pair of trajectory similarities during computation of the distance matrix.

Avg. Haversine, Equation (2)	DTW	Hausdorff	Kinematic, Equation (4)
16.38 µs	107.2 µs	1084 µs	12,788 µs

4.3.4. Discussion of the Positional Clustering

The traditional distance measures DTW and Hausdorff result in many outliers when applied to a complex, unfiltered dataset that resembles trajectories expected in real-life applications. By contrast, the average Haversine distance results in clusters that contain all the different routes that originate in a location that varies between clusters. Additionally, the average Haversine distance is much faster to compute, which allows for real-time assignment of unseen trajectories into precomputed clusters. However, these clusters do contain trajectories from many different maritime routes. Therefore, simply reducing the threshold in the hierarchical clustering does not yield a more detailed clustering describing their global positional or local kinematic behavior. To refine the cluster, a different distance measure must thus be used.

4.4. Kinematic Clustering

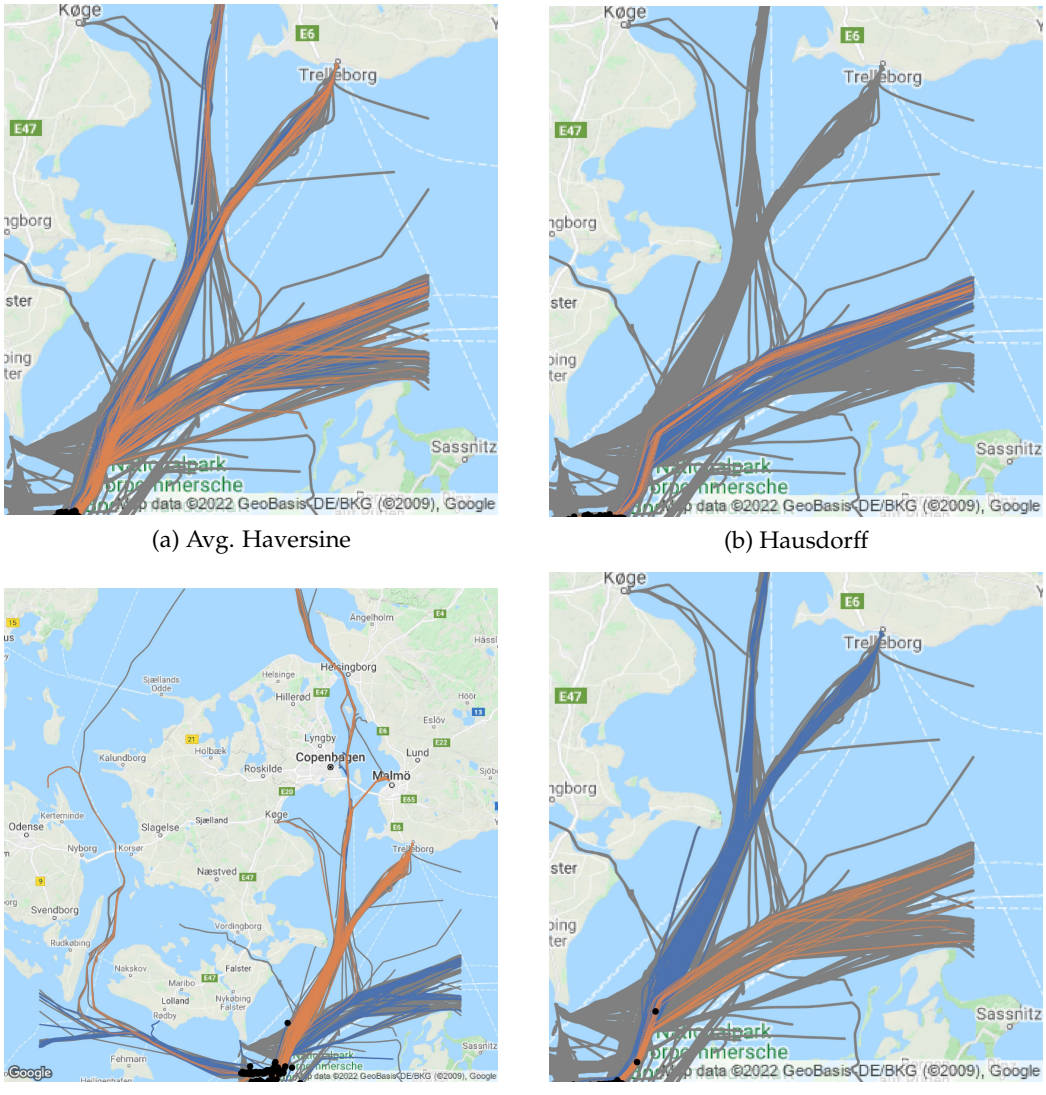
The kinematic clustering of the second step serves to provide a context to the outlier detectors prediction. It is expected to refine the gross positional clustering. Therefore, in this section, we study the refinement of the blue positional cluster shown in Figure 3e. The trajectories of this cluster originated at the southeastern edge of the ROI and split into four major shipping lanes—one going west towards the Kieler Channel, allowing passage to the Atlantic, one going north towards the North Sea, one going east towards the Baltic Sea, and one going northeast, terminating in the Swedish port of Trelleborg. In addition to these shipping lanes, the Danish port of Gedser (southern tip of the Lolland island) is a hub for pilot boats, which often have to rendezvous with larger ships passing through the Fehmarn Belt between Denmark and Germany. These pilot boats form a triangle fanning outwards east from the port of Gedser, seen in the bottom of Figure 3e.

We compared the clusters obtained using both kinematic and positional similarity measures. Regarding the positional clustering, based on the results of Section 4.3, we considered only hierarchical clustering combined with the average Haversine distance and the Hausdorff distance. The former produced better groupings, and the latter showed potential to further split clusters. As for the kinematic clustering, we tested our proposed kinematic similarity measure D_{kin}, Equation (4) with hierarchical clustering, and DBSCAN. Finally, to evaluate the benefit of basing D_{kin} on DTW, we also considered the average of the synchronous speed and course distances of Equations (5) and (6):

$$\mathrm{AK}(A,B) = \sum_{t=0}^{T-1} \frac{d_{\text{speed}}\left(s_A(t), s_B(t)\right) + d_{\text{course}}\left(c_A(t), c_B(t)\right)}{2T}, \tag{7}$$

where $T = \min(T_A, T_B)$.

We report in Table 5 the hyperparameters and statistics about each clustering. The two positional-based clusterings found fewer clusters and obtained better silhouette scores than our proposed kinematic distance measure, Equation (4). However, if we look at some of the clusters shown in Figure 4a,b, we note that these methods did not produce a more detailed clustering, in terms of the local kinematic behavior of the trajectories. This was expected, as without a refinement in local kinematic behavior a two-step clustering was not relevant, as we would have expected to find the same subclusters if we had accepted more groups in the first step when clustering the whole data.



(a) Avg. Haversine (b) Hausdorff

(c) Avg. Kinematic (d) Kinematic, Equation (4)

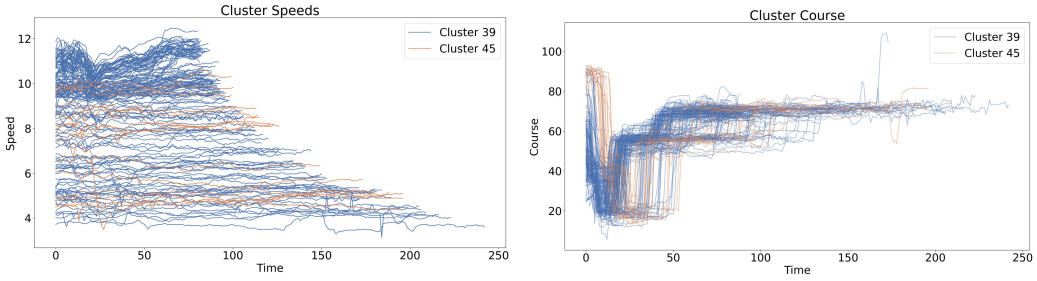
Figure 4. Second-step clusters obtained using hierarchical clustering and different similarity measures. Clustered trajectories are shown in color, trajectories in gray denote trajectories in the same positional cluster, and trajectory origins are denoted by a black circle.

Table 5. Hyperparameter values and clustering results of DBSCAN and hierarchical clustering using distances computed by Hausdorff, the average Haversine distance, Equation (2), or our proposed kinematic distance measure, Equation (4), on trajectories assigned to positional cluster 0 in the first step. Bold denote the highest recorded silhouette score.

Distance Measure	Clustering Method	Eps-Threshold	MinSamples	# Clusters	# Outliers/ Singletons	Silhouette Score
Avg. Haversine, Equation (2)	Hierarchical	0.9	-	49	12	**0.558**
Hausdorff	Hierarchical	6250	-	134	62	0.533
Avg. Kinematic	Hierarchical	1.8	-	34	0	0.126
Kinematic	DBSCAN	12.0	46	2	821	0.036
Kinematic	DBSCAN	12.0	2	21	477	−0.221
Kinematic	Hierarchical	22.5	-	221	147	0.217

4.4.1. Positional Similarity Measures

Clustering based on the average Haversine distance (Figure 4a) was unable to split the shipping lanes. We believe the focus on the initial part was the cause. The Hausdorff distance (Figure 4b) allowed the separation of the maritime routes through the ROI. We also noticed some maritime routes divided into two or more clusters, as seen in Figure 4b. Thus, we gained a more detailed clustering, in terms of describing their global positional behavior. In Figure 5, we show the speed and course of the trajectories assigned to the two clusters of Figure 4b. We see that both clusters are not clearly distinguishable, in terms of speed or course. Note that fast trajectories significantly decreased their speeds while in the shipping lane (blue trajectories with an initial speed of about 12 m/s). We would expect such trajectories to be grouped separately. Looking at the course, we see the two clusters generally had similar course changes, although they happened at different times, due to the time invariance of the Hausdorff distance. Based on the results above, we conclude that using the Hausdorff distance in the second-step clustering resulted in a more detailed clustering regarding the global positional behavior but not the local kinematic behavior.



(a) Speed (b) Course

Figure 5. Kinematic time series of the trajectories assigned to clusters obtained from the Hausdorff clustering shown in Figure 4b.

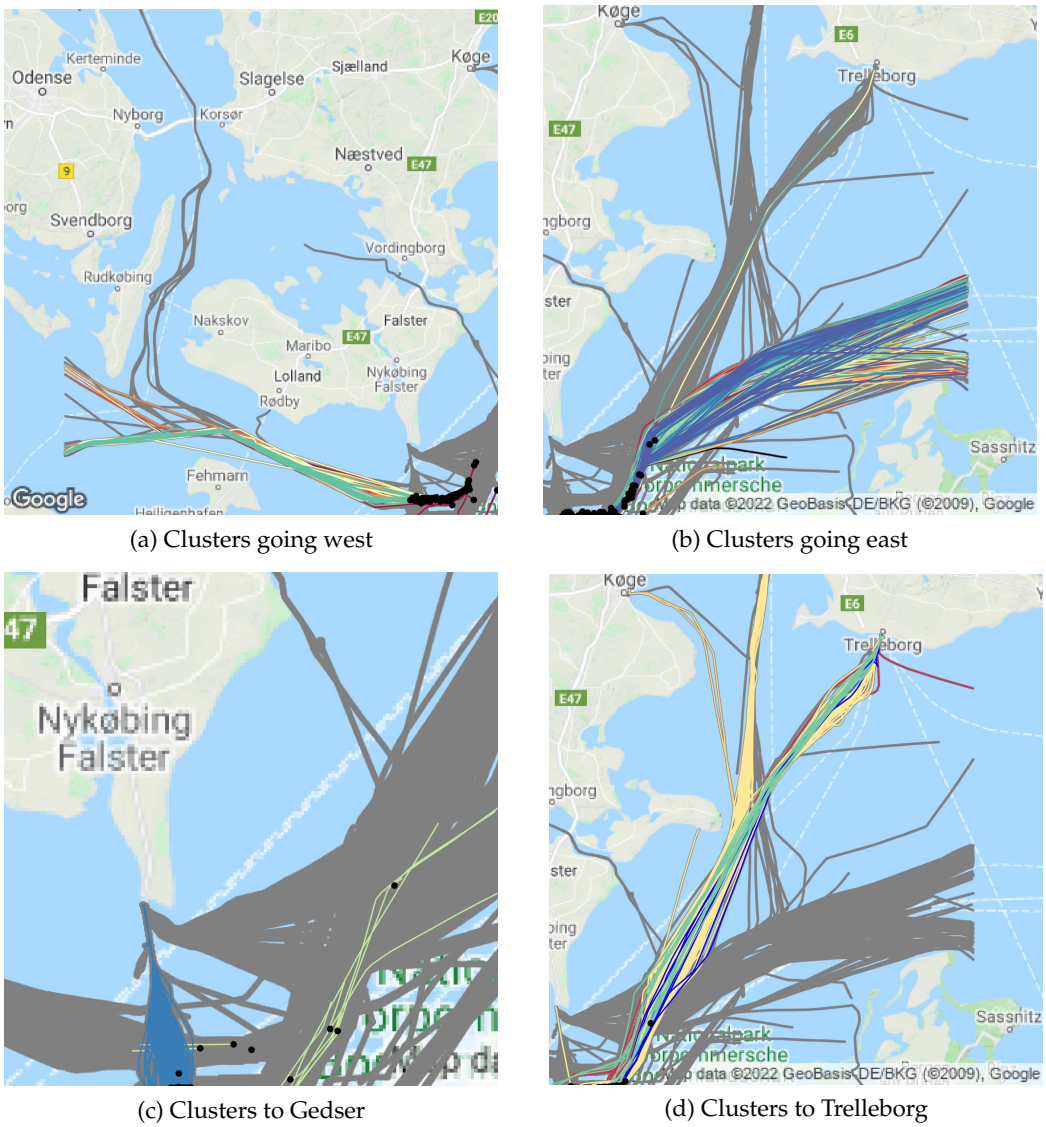
4.4.2. Kinematic Similarity Measures

We now study the clusterings obtained using kinematic-based similarity measures. Our proposed similarity measure combined with hierarchical clustering obtained a higher silhouette score than the average kinematic distance, yet the latter found fewer clusters.

We found that our choice of hierarchical clustering with average linkage was superior to the DBSCAN variants, as shown in Table 5. DBSCAN returned only two large clusters and most of the trajectories were assigned as outliers. Reducing the minimum number of samples required to define clusters increased the number of clusters to 21, but the vast majority of trajectories were assigned to the same cluster.

We see in Figure 4c that the average kinematic distance groups trajectories followed very different shipping lanes. Using Equation (4) as a similarity measure for hierarchical clustering resulted in 221 different clusters. Most of these clusters were singleton clusters

and could themselves be considered outliers. Despite their naturally similar behavior, we found that pilot boats were not clustered together, but belonged to singleton clusters that were closer to one other than to other kinematic clusters. This shows that singleton clusters could occur for normal expected behavior in sparse regions of our feature space. Clusters with more than five trajectories assigned to them are shown in Figure 6. The clusters clearly split the major maritime routes. Looking at the speed trajectories within these clusters (Figure 7a–c), we note that the clusters clearly partitioned the speed behaviors. Although less obvious, this also applied to the trajectories of Figure 6d heading towards the port of Trelleborg. In Figure 7b,d, we can distinguish five unique types of behavior: slow-speed returns south (green), fast-speed return south (red), slow-speed stops in port (orange), fast-speed stops in port (yellow), and fast-speed stops in port with a spike in speed during slowdown (blue). Note that the high-frequency course changes at low speeds in Figure 7d were due to a vessel drifting in port. These random course changes may artificially decrease the similarity between trajectories of the same behavior, but it is expected that two-stage DP compression [29] filters out the majority of these stationary periods at drift. In general, D_{kin} yielded well-separated clusters with consistent kinematic behavior.



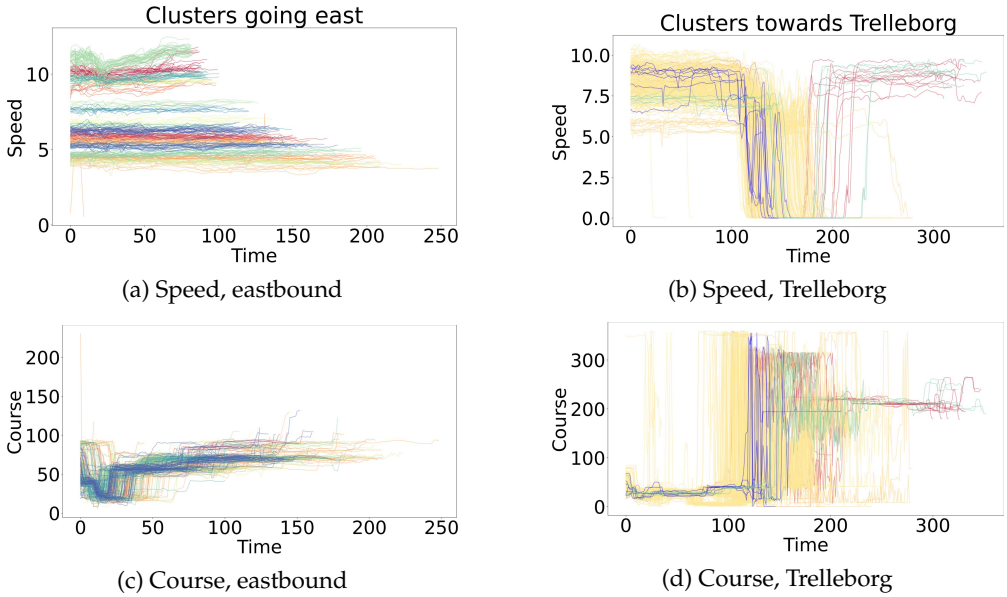
(a) Clusters going west

(b) Clusters going east

(c) Clusters to Gedser

(d) Clusters to Trelleborg

Figure 6. Second-step clusters with more than five assigned trajectories obtained using the kinematic distance matrix. Colors denote different clusters. Trajectories in gray denote trajectories in the same positional cluster.



(a) Speed, eastbound (b) Speed, Trelleborg

(c) Course, eastbound (d) Course, Trelleborg

Figure 7. Kinematic time series of all clusters following two different maritime routes; see Figure 6b–d. Different colors represent different clusters.

4.4.3. Discussion on the Kinematic Clustering

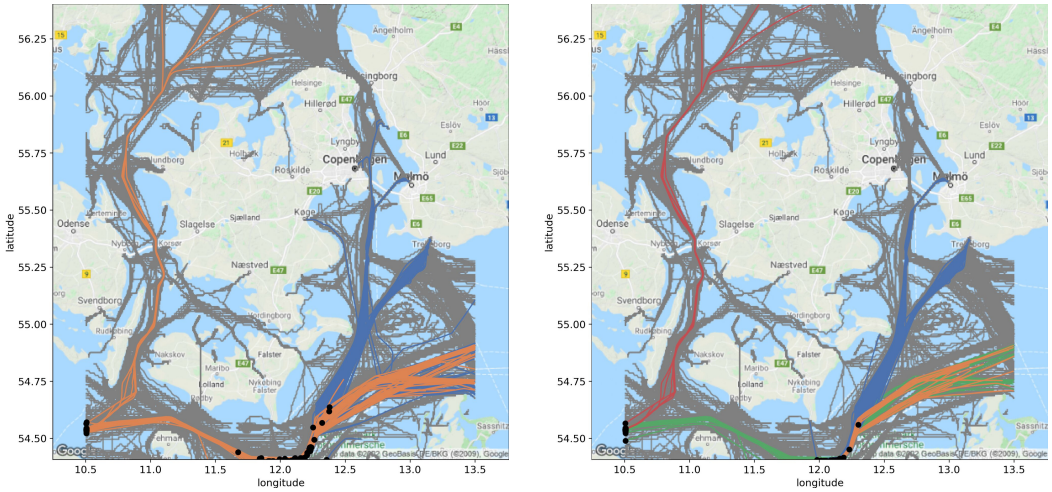
Our combination of D_{kin} and hierarchical clustering was able to refine a positional clustering from the first step. This refinement disentangled the positional and kinematic features, which resulted in subgroups with well-defined and unique kinematic behaviors, thus obtaining a more detailed description of maritime behavioral patterns. There existed an inherent trade-off between clustering kinematic behaviors that we knew to be similar but that naturally had a higher distance and clustering different behaviors that naturally were very close to one another. As discussed above, all the pilot boats were clustered into singleton clusters. But had the clustering threshold been increased, all the pilot boats would have formed a single kinematic cluster. However, increasing the threshold would have had the added downside of merging the different speed clusters of Figure 7a.

4.5. Single-Step Clustering

We now compare the clustering obtained from our two-step algorithm with one computed with a single hierarchical clustering using the sum of the average Haversine distance, Equation (2), and of D_{kin} of Equation (4) as the similarity measure. Combining positional and kinematic information into a single similarity measure may hide some variations that may be captured when the dimensions are processed separately, as in our approach. Therefore, the two-step method was expected to return more clusters.

The Kneedles algorithm for the single-step clustering obtained a silhouette score of 0.095 and found 2880 clusters. The two-step clustering approach achieved a silhouette score of 0.162 and a total of 6963 clusters when applied to the entire dataset. In both methods, the majority of the discovered clusters were singleton clusters, which we previously highlighted for the two-step algorithm. The single-step and two-step clustering approaches found 2168 and 6221 singleton clusters, respectively.

Using single-step clustering, trajectories traveling along different shipping lanes could be clustered together (Figure 8a), while our two-step algorithm disentangled them (Figure 8b). A detailed analysis reveals that the single-step approach grouped together (orange cluster of Figure 8a) trajectories with distant initial points and following different shipping lanes because the speed of the trajectories was very similar. The differences in position were compensated by similar speed behaviors. On the other hand, the two-step algorithm split these two routes into multiple clusters, as it processed positional and kinematic information separately.



(a) Single-step clusters (b) Two-step clusters

Figure 8. Trajectory clusters produced by the single-step clustering (**a**) that were split into multiple clusters, (**b**) using the the two-step algorithm. Clustered trajectories are shown in color, trajectories in gray denote trajectories in the same positional cluster, and trajectory origins are denoted by a black circle.

Our proposed two-step approach results in a better disentanglement of position and kinematic behavior. Even though our proposed similarity measure focuses on different aspects of the trajectories, treating them as a sum results in a situation where differences in the positional similarity are canceled by differences in the kinematic similarity. The better disentanglement of the two-step approach results in clusters with more well-defined and unique kinematic behaviors, thus obtaining a more detailed description of maritime behavioral patterns. Better disentanglement also means that the learned normalcy model can distinguish a higher number of possible kinematic behaviors at each location in the ROI. As such, the normalcy model is more useful for supplying contextual information to VTS operators. As discussed previously, the two-step approach has a trade-off between clustering behaviors that we know to be similar but that naturally have a larger distance and clustering different behaviors that naturally are very close to one another. Using a single-step approach seems to push this trade-off towards the latter option, automatically. Additionally, the two-step approach is computationally more efficient than the single-step approach. The computational requirements of the kinematic distance measure shown in Table 4 are large compared to the positional distance measure. Thus, computing the proposed kinematic distance measure on the entire dataset is not feasible for large datasets. Comparatively, the first positional clustering in the two-step approach functions as a filter that reduces the number of trajectory pairs for which to calculate the kinematic distance.

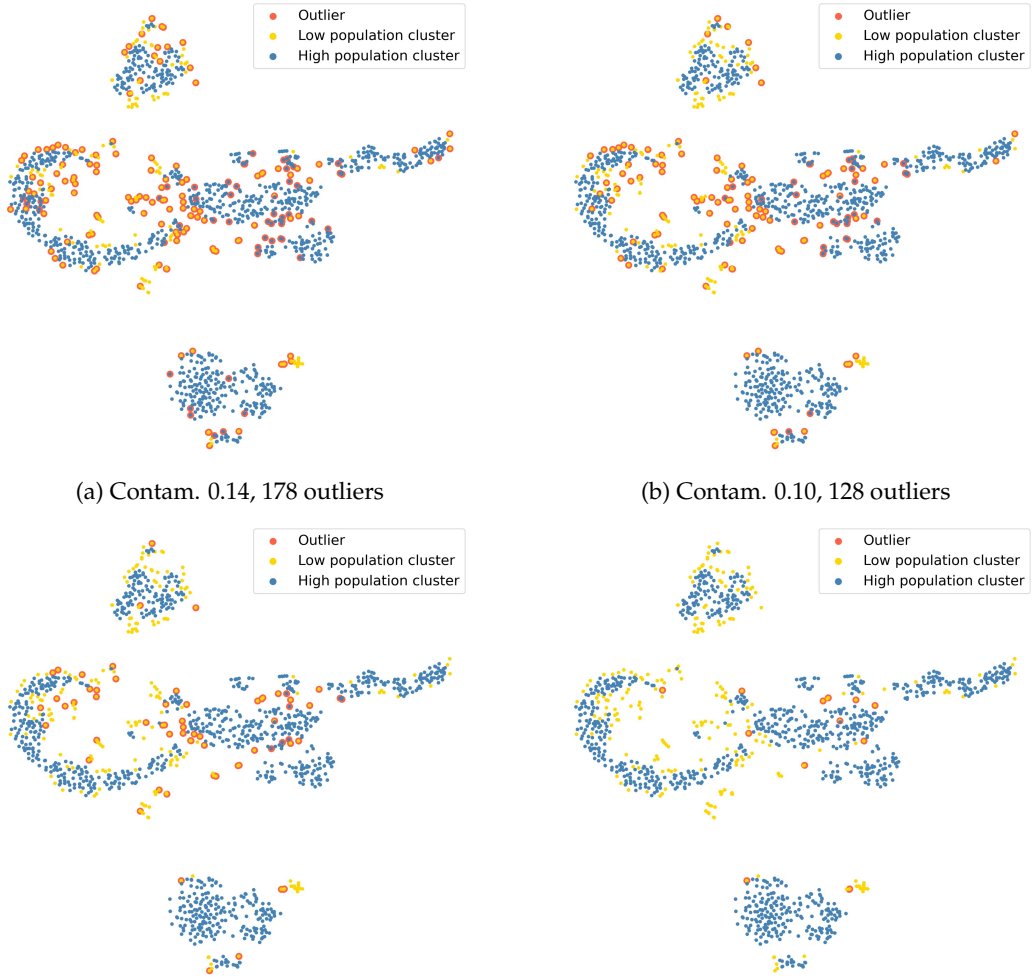
4.6. Outliers and Embedding Analysis

In this section, we use a TSNE representation of the kinematic similarity matrix of the blue positional cluster of Figure 3e, to show how the LOF and the kinematic clustering are related despite being computed independently.

4.6.1. LOF Contamination and Cluster Size

Although the LOF does not use kinematic clustering, it is computed on the same similarity matrix. In this experiment, we use a TSNE representation of the kinematic similarity matrix, to show how the kinematic clustering and the LOF handle outliers.

In Figure 9, we plotted a TSNE embedding of the kinematic distance matrix and the LOF results with different contamination levels of all trajectories of the blue positional cluster of Figure 3e. In the following, we compare four values of the contamination hyperparameter: (a) 0.14, computed as suggested in [25], (b) 0.1, (c) 0.05, and (d) 0.01. Points in clusters with at least five trajectories are colored blue and the others are in yellow. Outliers are denoted by red borders.



(a) Contam. 0.14, 178 outliers

(b) Contam. 0.10, 128 outliers

(c) Contam. 0.05, 64 outliers

(d) Contam. 0.01, 13 outliers

Figure 9. T-SNE of the kinematic distance matrix of the blue cluster of Figure 3e with varying levels of contamination in the LOF. Crosses denote detected outliers. Colors denote cluster assignments, with gray being clusters with fewer than five members. The contamination level in (a) is determined using [25].

We see that the TSNE projects the trajectories in low populated clusters in the fringe of those in highly populated. Increasing or decreasing the contamination hyperparameter causes more or less of the trajectories along the fringe to be flagged as outliers. With a contamination of 0.10, even trajectories from large clusters are flagged as abnormal. Reducing it to 0.01 leaves almost no outliers. Therefore, we recommend contamination levels of 0.05.

Note that yellow dots that are grouped together in the TSNE projection without necessarily being in the same kinematic cluster better resist the increase of the contamination hyperparameter and remain inliers. An example is the group of yellow dots on the top-right of the large blue cluster at the bottom of the representation.

4.6.2. Outliers and Embedding Analysis

In Figure 10, we plot the same TSNE representation of the kinematic similarity matrix of all trajectories in positional cluster 0, shown in blue in Figure 3e, but the colors denote here cluster assignments of the kinematic clustering. The gray dots are clusters with fewer than five trajectories. The red borders denote outliers flagged by the LOF, using a contamination level of 0.05. The squares denote embeddings of pilot boats, triangle trajectories with U-turns, and crosses trajectories with stops outside a port.

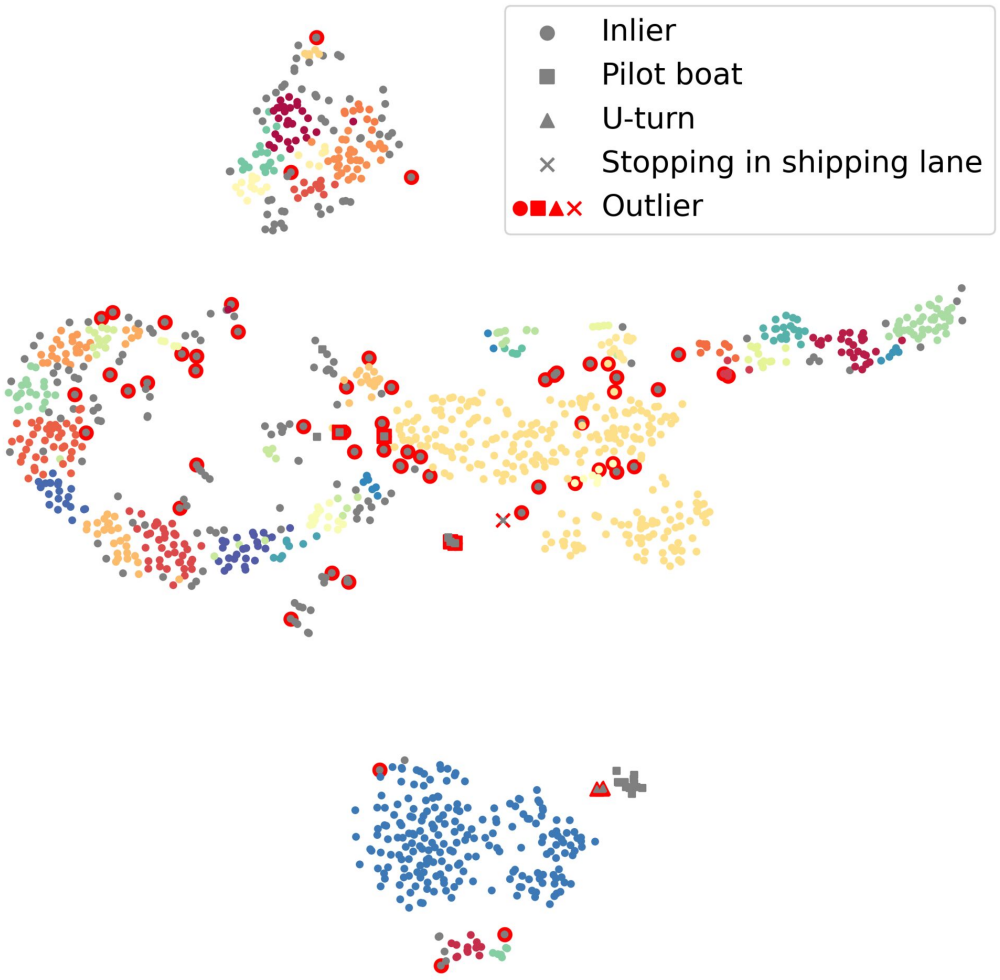
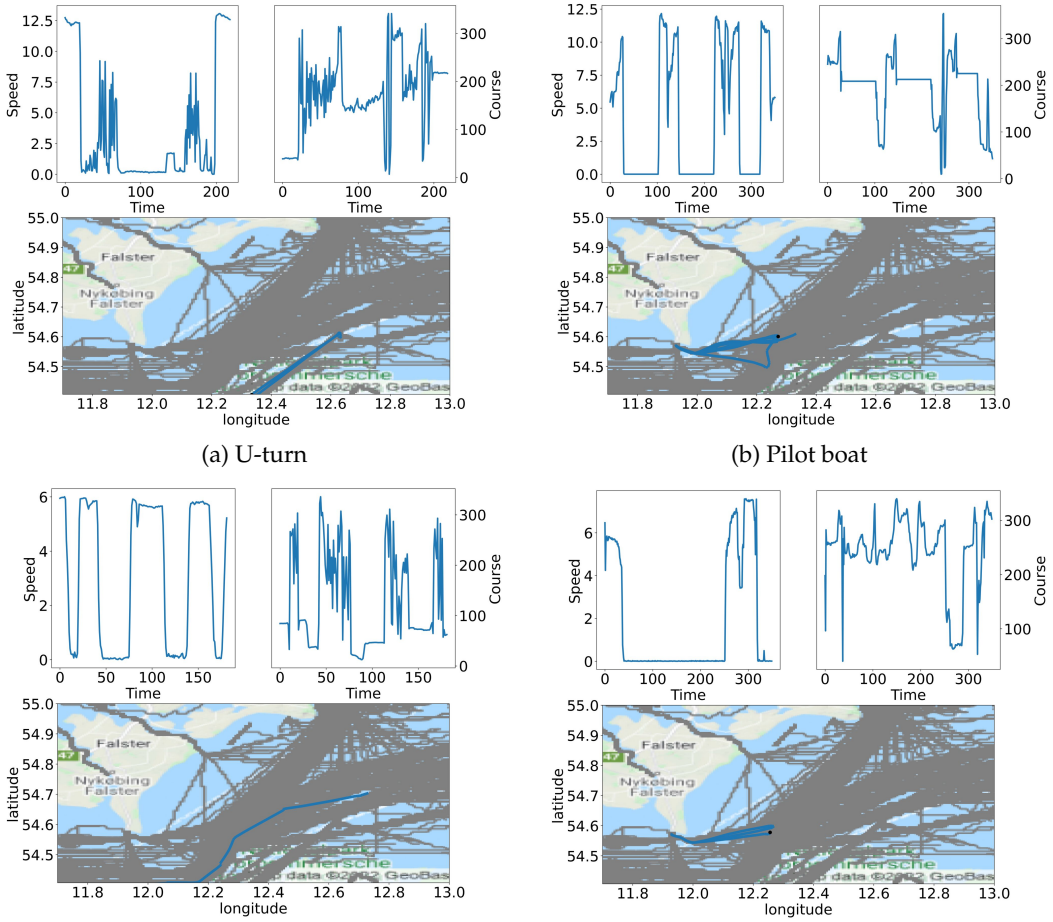


Figure 10. TSNE of the kinematic distance matrix of the blue cluster in Figure 3e. The colors denote the kinematic cluster assignment, with gray being clusters with less than five members. The red borders denote outliers flagged by the LOF. The shapes denote manually identified special trajectories.

As we mentioned previously, the majority of the closest neighbors of the pilot boats were other pilot boats. Most of the pilot boats form a small gray group near a larger cluster (blue) in the bottom of the figure. This blue cluster represents ships coming from the south that docked at the port of Gedser and return back towards the south after docking. In the same area, we find cases of ships (red, green) sailing north and docking at Trelleborg before returning south. In between the pilot boats and the large blue cluster we notice three trajectories discovered to be outliers by the LOF. These three trajectories were boats coming from the south and making a stop and a U-turn in the middle of the shipping lane, very similarly to the pilot boats. However, small differences in the course time series led the three U-turns trajectories to be flagged as abnormal. In Figures 11a,b, we show an example of a U-turn trajectory and that of a pilot boat nearby.

Note the crescent clustered in many different clusters on the left of Figure 10. These clusters correspond to the traffic following the shipping lanes to the east (Figure 6b) at different speeds. Around the right end of this crescent, we find some pilot boats (square) and a highlighted outlier corresponding to a trajectory following the shipping lanes going east with multiple sudden stops in the middle of the shipping lane (cross). This type of behavior is of interest for certain types of ships, such as research vessels and diving vessels. These two trajectories are shown in Figure 11c,d. The large yellow cluster in the center of Figure 10 corresponds to trajectories going north and stopping in the port of Trelleborg. Contrary to the previous smaller red and green clusters in the bottom of Figure 10, these

trajectories ended in the port. This means that our proposed kinematic distance measure found the start-stopping behavior to be more akin to the trajectories ending in port or pilot boats; however, not so much as to be part of their cluster, due to differences in the course. This confirms that our proposed distance measure is capable of capturing local kinematic similarities, regardless of the geographical position, and it can serve to flag trajectories with abnormal local kinematic behavior.



(a) U-turn (b) Pilot boat

(c) Multiple stops on shipping lane (d) Pilot boat

Figure 11. Trajectories of abnormal activity, such as (**a**) U-turns and (**c**) stopping in the shipping lane, found to have similarity with the behavior of pilot boats (**b**,**d**). Historical traffic is shown in gray, and trajectory origins are denoted by a black circle.

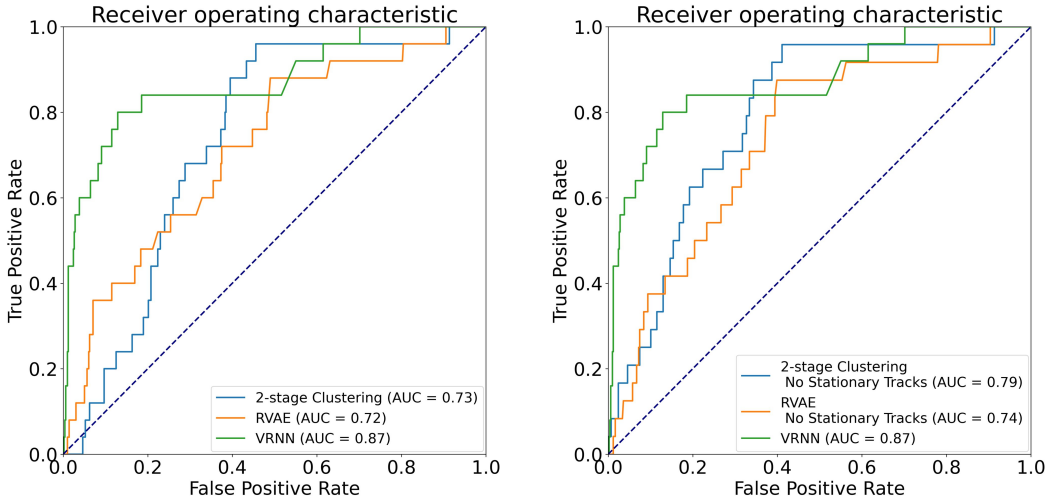
4.7. Abnormality Detection

We evaluate now the performance of our algorithm for the detection of abnormalities. We first compare it to State-of-the-Art abnormality detectors and then provide an example showing how the context provided by the kinematic clustering helps to understand the prediction of the LOF. For comparison, we also propose an interpretation of the VRNN baseline.

Throughout this section, we use the data from the Bornholm area of December 2021. All the models were trained using all the data, except that from 13 December and tested on that day. As a Search and Rescue operation took place that day, all the trajectories were annotated.

4.7.1. Anomaly Detection

We investigate the precision of our model and discuss which value of contamination parameter to use, based on the receiver operating curve from outlier detection on the 13 December data, as shown in Figure 12. We compare to the A-Contrario outlier detection method [15], using RVAE [13] and VRNN [15].



(a) All trajectories (b) Without stationary trajectories

Figure 12. ROC of the outlier detection on the Bornholm 13 December data.

We see that our method outperforms detection based on RVAE reconstruction but not VRNN reconstruction. In particular, we see that our method suffers from false positives early in the detection. Looking at these trajectories, we note that the vessels were mostly stationary in port with a highly varying course, resulting in a large directional distance to all other trajectories. By removing trajectories with an average speed of less than 0.3 m/s from the detection, we can reduce the number of early false positives to similar levels as RVAE.

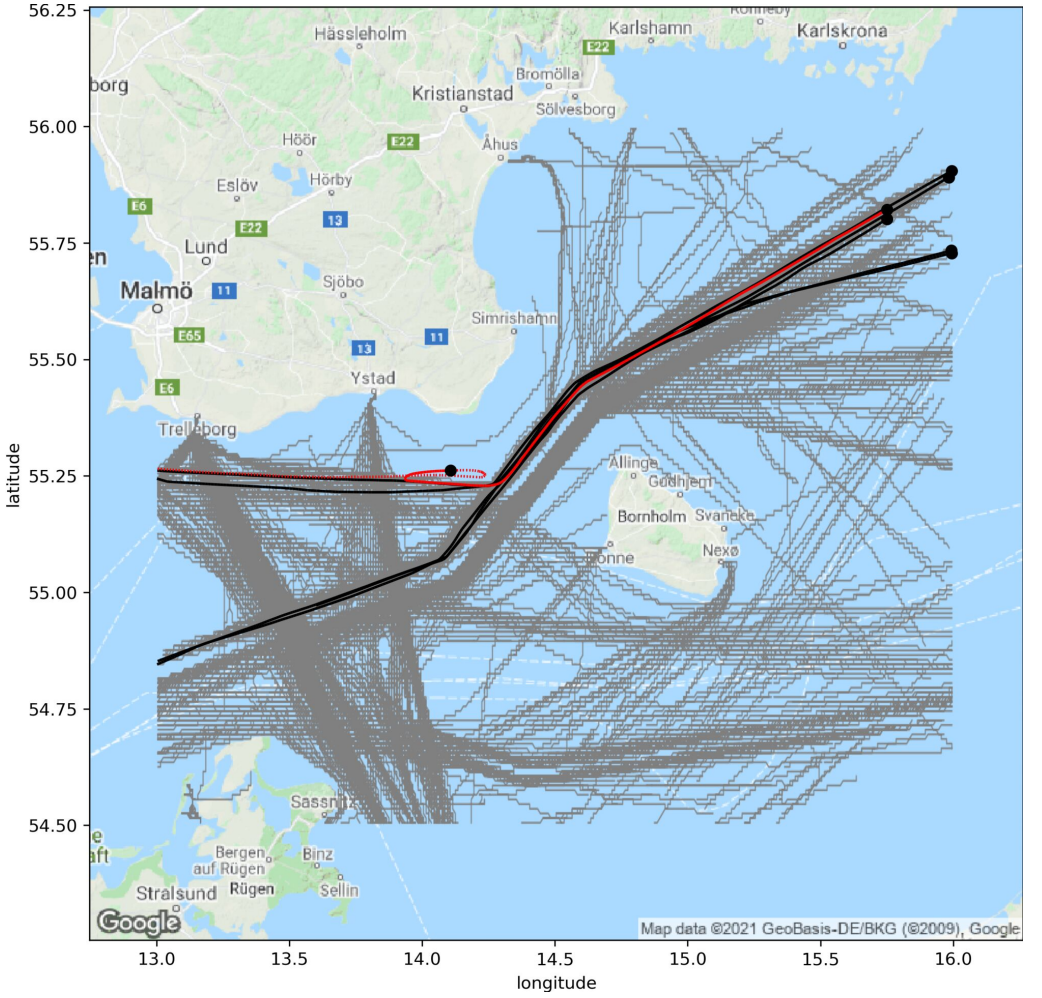
Previously, we counted outliers to infer the contamination parameter. Here, we show how the ROC curve may provide more insight into this hyperparameter. In order to detect 96% of the abnormal trajectories, a false alarm was triggered on 40% of the normal trajectories, which corresponded to a contamination rate of 0.43. A contamination level of 0.05, 0.10, and 0.14 yielded a true positive rate of about 20%, 25%, and 40%, respectively.

The detection performance of our proposed method is below that of the VRNN, but it is much faster to computer. Indeed, in our experiments, the VRNN model and A-Contrario detection had an average evaluation time of 176 seconds per trajectory. By comparison, our two-step algorithm required only 6.8 seconds per trajectory. The large processing time of the VRNN model and A-Contrario detection is a major drawback that questions the viability of both algorithms for a real-life scenario where a quick response is vital, e.g., in the case of a collision or prediction of ongoing piracy. Our algorithm spends most of the 7 s to compute the positional similarity matrix, which compares the input to all the training dataset. This operation could be dramatically sped up, using some heuristics to avoid computations with obviously different trajectories.

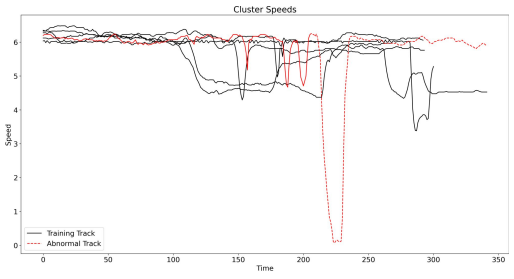
4.7.2. Anomaly Detection and Context

We present here an example to highlight how the context provided by the kinematic clustering helps to assess the prediction of the LOF. We consider the case of an abnormal trajectory making a double U-turn in the shipping lane. The trajectory was made by the sister ship of the vessel that caused the collision accident and was returning to the site of the collision, perhaps to transfer crew to the collided vessel.

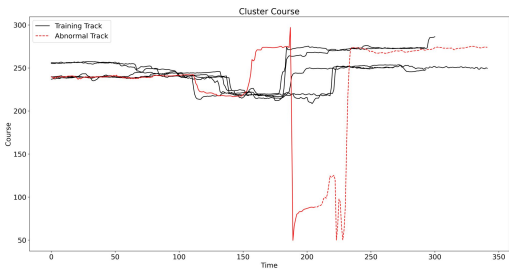
We fed the algorithm with the trajectory truncated at $t = 205$, which corresponded to when the vessel finished the first U-turn and was traveling in the opposite direction of the shipping lane. The positional trajectory (a), speed (b), and course (c) time series of that trajectory are depicted in red (plain before $t = 205$ and dotted after) in the plots of Figure 13. The five most similar trajectories in the kinetic context are depicted in black. Note that all five trajectories originated in the northeastern part of the ROI, and all five vessels had extended periods of time in which they traveled at reduced speeds.



(a) Geographical evolution



(b) Speed



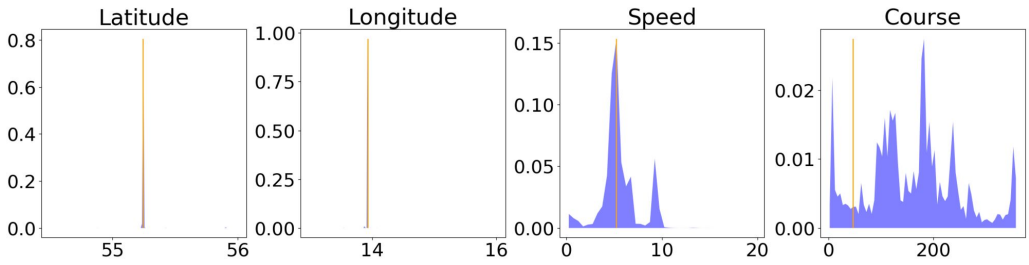
(c) Course

Figure 13. Top five most similar trajectories (black) to the abnormal trajectory in red determined by Equation (4) at time step $t = 205$. The future trajectory is shown in dashes.

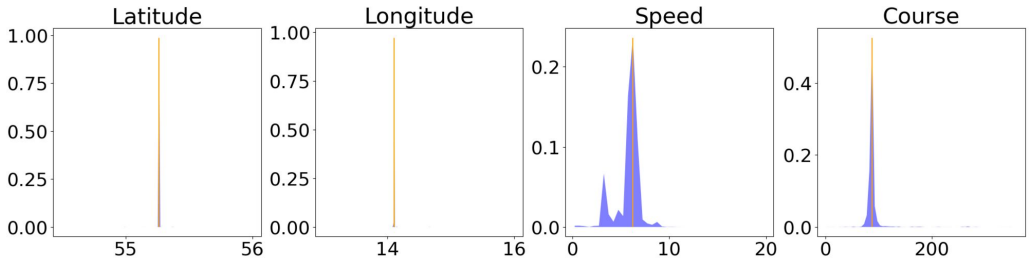
The three visualizations issued from the kinetic context provide complementary information to the VTS operator. Indeed, if the the U-turn may be overlooked in Figure 13a,b, the plot of the course is clear. Another hypothesis could be that the abrupt change of direction suggests a turn toward a harbor, but this is dismissed by the fact that the vessel is still close to the shipping lane.

For comparison, we propose an interpretation of the outputs of the VRNN, the best-performing model, to justify the anomaly of the red trajectory. Note that VRNN outputs a multivariate Bernoulli distribution, which indicates if an input can or cannot be predicted/reconstructed accurately. In Figure 14, we plot the multivariate Bernoulli distribution of the VRNN model at two time steps during and after the first U-turn. During the U-turn, the speed and course are not reconstructed accurately. After the U-turn, the

model recovers and is able to reconstruct accurately both the speed and the course. This means that at $t = 205$, the trajectory will still be flagged as abnormal because it was at $t = 189$. Thus, the operator needs to manually evaluate several previous updates, to understand the context that led the model to flag the anomaly. Another solution is to assess the situation based on the visualization of the different dimensions of the trajectory, as in Figure 13, but without the support of the black trajectories. This complicates the task, as one would expect—for example, pilot boats performing U-turns close to a shipping lane but not commercial vessels. Again, the operator needs to look for more information before making a decision.



(a) During first U-turn $t = 189$



(b) After first U-turn $t = 205$

Figure 14. Multivariate Bernoulli distribution (blue) for the VRNN output at two time steps observed during and after the first U-turn. True values are indicated by orange lines.

5. Conclusions

In this work, we have presented and made public two large hand-annotated AIS traffic datasets for abnormal maritime behavior detection that we have created. One dataset contains all traffic AIS data around Sjælland Island during November 2021. The second dataset contains all AIS trajectories around Bornholm Island during December 2021, with the full annotation for the data of 13 December 2021, with events expected to be of interest to VTS operators. In addition, we also have proposed an abnormality detection algorithm for maritime trajectories based on two-step clustering, for which we proposed a novel kinematic similarity measure based on DTW. The two separate steps allow the clustering to better focus on the kinematic behavior expected in a certain geographical position. This disentanglement of positional and kinematic features results in better descriptions of behavioral patterns and clusters with well-defined and unique kinematic behaviors. These behavioral clusters and kinematic similarity measures can be used to provide context to the VTS operator, to accept or reject the algorithm prediction. We evaluated our proposed abnormality detection method on the annotated data of 13 December 2021 of the Bornholm dataset. Although our proposed model achieved a lower area under the ROC than the VRNN model, it had a clear advantage, in terms of runtime and interpretability, over current deep learning methods. In future work, we aim to compute the proposed similarity measures with neural networks and to utilize deep models for reconstruction-based outlier detection that also align trajectories in the latent space, to preserve a certain level of interpretability.

A limitation of our method is that it has been designed for and tested mostly on near-coastal traffic with a large variety of different maritime behaviors. We speculate that

application in the open ocean with more extreme weather differences would result in separate clusters for each weather profile. We leave this question for future work.

Author Contributions: Conceptualization, K.V.O.; methodology, K.V.O. and A.B.; software, K.V.O. and A.B.; formal analysis, K.V.O.; investigation, K.V.O.; resources, S.H.; data curation, K.V.O. and S.H.; original draft preparation, K.V.O.; review and editing, A.B., M.C.K., R.J., A.N.C. and L.H.C.; visualization, K.V.O.; supervision, A.B., M.C.K., R.J., A.N.C. and L.H.C.; funding acquisition, R.J. and L.H.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was financially supported by the Danish Ministry of Defence Acquisition and Logistics Organisation, grant no. 4600005159. Visual Intelligence publications are financially supported by the Research Council of Norway, through its Centre for Research-based Innovation funding scheme (grant no. 309439), and Consortium Partners.

Data Availability Statement: The data used in this paper are made freely available at https://data.dtu.dk/collections/AIS_Trajectories_from_Danish_Waters_for_Abnormal_Behavior_Detection/6287841, (accessed on 20 October 2023).

Acknowledgments: The datasets used in this paper and the visualization tools to identify abnormal behavior were provided by Terma A/S (Lystrup, Denmark).

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

AIS	Automatic Identification System
MMSI	Maritime Mobile Service Identity
GPS	Global Positioning System
VTS	Vessel Traffic Service
LOF	Local Outlier Factor
ROI	Region of Interest
DTW	Dynamic Time Warping
LCSS	Longest Common SubSequence
DP	Douglas–Peucker algorithm
RVAE	Recurrent Variational AutoEncoder
VRNN	Variational Recurrent Neural Network
AH	Average Haversine
KNN	K-Nearest Neighbors
AUC	Area Under the receiver operating characteristic Curve

References

1. IMO. *About IMO*; International Maritime Organization: London, UK, 2020. Available online: http://www.imo.org/en/About/Pages/Default.aspx (accessed on 20 October 2023).
2. Asariotis, R.; Benamara, H.; Lavelle, J.; Premti, A. Maritime Piracy. Part I: An Overview of Trends, Costs and Trade-Related Implications . UNCTAD 2014. Available online: https://eprints.soton.ac.uk/368254/ (accessed on 20 October 2023).
3. Lebedev, A.O.; Lebedeva, M.P.; Butsanets, A.A. Could the accident of "Ever Given" have been avoided in the Suez Canal? *J. Phys. Conf. Ser.* **2021**, *2061*, 12127. [CrossRef]
4. European Maritime Safety Agency. *Annual Overview of Marine Casualties and Incidents*; Technical Report; European Maritime Safety Agency: Lisbon, Portugal, 2022.
5. Long, T.; Widjaja, S.; Wirajuda, H.; Juwana, S. Approaches to combatting illegal, unreported and unregulated fishing. *Nat. Food* **2020**, *1*, 389–391. [CrossRef]
6. Ljungqvist, M. Confirmed Sabotage at Nord Stream . (In Swedish) Available online: https://www.aklagare.se/nyheter-press/pressmeddelanden/2022/november/bekraftat-sabotage-vid-nord-stream/ (accessed on 20 October 2023).
7. International Maritime Organization (IMO). *International Convention for the Safety of Life at Sea (SOLAS), Chapter V: Safety of Navigation, Regulation 19*; International Maritime Organization (IMO): London, UK, 1998.
8. MarineTraffic. A Day in Numbers. MarineTraffic Blog. Available online: https://www.marinetraffic.com/blog/a-day-in-numbers/ (accessed on 20 October 2023).

9. Pallotta, G.; Vespe, M.; Bryan, K. Traffic knowledge discovery from AIS data. In Proceedings of the 16th International Conference on Information Fusion, IEEE, Istanbul, Turkey, 9–12 July 2013.
10. Liu, B.; De Souza, E.N.; Matwin, S.; Sydow, M. Knowledge-based clustering of ship trajectories using density-based approach. In Proceedings of the 2014 IEEE International Conference on Big Data, IEEE Big Data 2014, Washington, DC, USA, 27–30 October 2014; pp. 603–608.
11. Zhao, L.; Shi, G. A trajectory clustering method based on Douglas-Peucker compression and density for marine traffic pattern recognition. *Ocean Eng.* **2019**, *172*, 456–467. [CrossRef]
12. Yang, J.; Liu, Y.; Ma, L.; Ji, C. Maritime traffic flow clustering analysis by density based trajectory clustering with noise. *Ocean Eng.* **2022**, *249*, 111001. [CrossRef]
13. Murray, B.; Perera, L.P. An AIS-based deep learning framework for regional ship behavior prediction. *Reliab. Eng. Syst. Saf.* **2021**, *215*, 107819. [CrossRef]
14. Pallotta, G.; Jousselme, A.L. Data-driven detection and context-based classification of maritime anomalies. In Proceedings of the 2015 18th International Conference on Information Fusion (Fusion), IEEE, Washington, DC, USA, 6–9 July 2015.
15. Nguyen, D.; Vadaine, R.; Hajduch, G.; Garello, R.; Fablet, R. GeoTrackNet-A Maritime Anomaly Detector using Probabilistic Neural Network Representation of AIS Tracks and A Contrario Detection. *IEEE Trans. Intell. Transp. Syst.* **2021**, *23*, 5655–5667. [CrossRef]
16. Riveiro, M.; Pallotta, G.; Vespe, M. Maritime anomaly detection: A review. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2018**, *8*, e1266. [CrossRef]
17. Stach, T.; Kinkel, Y.; Constapel, M.; Burmeister, H.C. Maritime Anomaly Detection for Vessel Traffic Services: A Survey. *J. Mar. Sci. Eng.* **2023**, *11*, 1174. [CrossRef]
18. Endsley, M.R. From Here to Autonomy: Lessons Learned From Human—Automation Research. *Hum. Factors* **2017**, *59*, 5–27. [CrossRef]
19. Wang, L.; Chen, P.; Chen, L.; Mou, J. Ship AIS Trajectory Clustering: An HDBSCAN-Based Approach. *J. Mar. Sci. Eng.* **2021**, *9*, 566. [CrossRef]
20. Liu, B.; de Souza, E.N.; Hilliard, C.; Matwin, S. Ship movement anomaly detection using specialized distance measures. In Proceedings of the 2015 18th International Conference on Information Fusion (Fusion), IEEE, Washington, DC, USA, 6–9 July 2015.
21. Hu, J.; Kaur, K.; Lin, H.; Wang, X.; Hassan, M.M.; Razzak, I.; Hammoudeh, M. Intelligent Anomaly Detection of Trajectories for IoT Empowered Maritime Transportation Systems. *IEEE Trans. Intell. Transp. Syst.* **2022**, *24*, 2382–2391. [CrossRef]
22. Liu, H.; Liu, Y.; Li, B.; Qi, Z.; Rizvi, J.; Liu, H.; Liu, Y.; Li, B.; Qi, Z. Ship Abnormal Behavior Detection Method Based on Optimized GRU Network. *J. Mar. Sci. Eng.* **2022**, *10*, 249. [CrossRef]
23. Li, J.; Liu, J.; Zhang, X.; Li, X.; Wang, J.; Wu, Z. A Novel Hybrid Approach for Detecting Abnormal Vessel Behavior in Maritime Traffic. In Proceedings of the 2023 7th International Conference on Transportation Information and Safety (ICTIS), Xi'an, China, 4–6 August 2023; pp. 1–7.
24. Widyantara, I.M.O.; Hartawan, I.P.N.; Karyawati, A.A.I.N.E.; Er, N.I.; Artana, K.B. Automatic identification system-based trajectory clustering framework to identify vessel movement pattern. *Iaes Int. J. Artif. Intell.* **2023**, *12*, 1–11. [CrossRef]
25. Breunig, M.M.; Kriegel, H.P.; Ng, R.T.; Sander, J. LOF: Identifying Density-Based Local Outliers. In Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, SIGMOD '00, Dallas, TX, USA, 16–18 May 2000; pp. 93–104.
26. Larsen, M.S. Russian 'Ghost Ships' Are Turning the Seabed into a Future Battlefield, 2023. Available online: https://foreignpolicy.com/2023/05/02/russia-europe-denmark-spy-surveillance-ships-seabed-cables/ (accessed on 20 October 2023).
27. Laxhammar, R.; Falkman, G. Inductive conformal anomaly detection for sequential detection of anomalous sub-trajectories. *Ann. Math. Artif. Intell.* **2015**, *74*, 67–94. [CrossRef]
28. Zhen, R.; Jin, Y.; Hu, Q.; Shao, Z.; Nikitakos, N. Maritime Anomaly Detection within Coastal Waters Based on Vessel Trajectory Clustering and Naïve Bayes Classifier. *J. Navig.* **2017**, *70*, 648–670. [CrossRef]
29. Klaas, G.; De Vries, D.; Van Someren, M. Machine learning for vessel trajectories using compression, alignments and domain knowledge. *Expert Syst. Appl.* **2012**, *39*, 13426–13439.
30. Douglas, D.H.; Peucker, T.K. Algorithms for the Reduction of the Number of Points Required to Represent a Digitized Line or its Caricature. In *Classics in Cartography: Reflections on Influential Articles from Cartographica*; John Wiley & Sons, Ltd.: Hoboken, NJ, USA, 2011; pp. 15–28.
31. Pallotta, G.; Vespe, M.; Bryan, K. Vessel Pattern Knowledge Discovery from AIS Data: A Framework for Anomaly Detection and Route Prediction. *Entropy* **2013**, *15*, 2218–2245. [CrossRef]
32. Ester, M.; Kriegel, H.p.; Sander, J.; Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining, Portland, OR, USA, 2–4 August 1996.
33. Luo, S.; Zeng, W.; Sun, B. Contrastive Learning for Graph-Based Vessel Trajectory Similarity Computation. *J. Mar. Sci. Eng.* **2023**, *11*, 1840. [CrossRef]
34. Zhao, L.; Shi, G. Maritime Anomaly Detection using Density-based Clustering and Recurrent Neural Network. *J. Navig.* **2019**, *72*, 894–916. [CrossRef]
35. Shamos, M.; Preparata, F. Computational Geometry An Introduction. In *Computational Geometry an Introduction*, 1st ed.; Schneider, F., Gries, D., Eds.; Springer: New York, NY, USA, 1985; Chapter 5, p. 223.
36. Nanni, M.; Pedreschi, D. Time-focused clustering of trajectories of moving objects. *J. Intell. Inf. Syst.* **2006**, *27*, 267–289. [CrossRef]

37. Olesen, K.V.; Christensen, A.N.; Hørlück, S.; Clemmensen, L.K.H. AIS Trajectories from Danish Waters for Abnormal Behavior Detection. 2022. Available online: https://data.dtu.dk/collections/AIS_Trajectories_from_Danish_Waters_for_Abnormal_Behavior_Detection/6287841 (accessed on 20 October 2023).
38. Soefartsstyrelsen. Historical AIS Data. Available online: https://dma.dk/safety-at-sea/navigational-information/ais-data (accessed on 20 October 2023).
39. Satopaa, V.; Albrecht, J.; Irwin, D.; Raghavan, B. Finding a "Kneedle" in a Haystack: Detecting Knee Points in System Behavior. In Proceedings of the 31st International Conference on Distributed Computing Systems Workshops, Minneapolis, MI, USA, 20–24 June 2011.
40. Rousseeuw, P.J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53–65. [CrossRef]