RESEARCH ARTICLE OPEN ACCESS

Impact of Integral Temporal Component in Density-Based Clustering of AIS

Nitin Newaliya, Vikas Siwach, Harkesh Sehrawat, Yudhvir Singh
Department of CSE, UIET, Maharishi Dayanand University, Rohtak, Haryana, India

ABSTRACT
The AIS data can be subject to data mining to ascertain various insights. One of the popular algorithms used for data analytics in the maritime domain is DBSCAN (Density-Based Spatial Clustering of Applications with Noise). Due to its inherent property of determining clusters of arbitrary sizes, it is well suited to cluster AIS (Automatic Information System) data. One category of maritime adaptations of DBSCAN involves the use of spatial attributes, along with non-spatial attributes, to determine the outputs. Temporal information has, however, not been used as an integral part of the attributes for calculating the distance. Assessment of the inclusion of the Time field along with non-spatial components of AIS data has been undertaken experimentally in this paper, and it has been seen that the performance of this adaptation is superior to other algorithms.
Keywords — Data exploration, Automatic Identification System, DBSCAN, Maritime, Clustering

I. INTRODUCTION

The seas and oceans are extensively used for transportation activities. It provides a very convenient means of connecting various places for transportation of goods and personnel. There are a large number of vessels, and, as per a UN report, there are approximately 1,05,000 vessels, which are 100 GT or more [1]. With a considerable number of vessels sailing together, there was a need to develop a system to ensure safety at sea and prevent accidents. The IMO (International Maritime Organisation) promulgated the use of a system known as the Automatic Identification System (AIS) [2]. Identity details, position, movement information and certain static information are transmitted by each ship on which this VHF-based system is installed. These AIS messages are received by other ships nearby and thus enhance the awareness of each ship about its neighbourhood.

This AIS data has an interesting secondary application in that the data could be subject to data mining to ascertain various insights. This led to considerable interest in researching various applications utilising the AIS data. Traditional data analytic algorithms have been actively used while undertaking research on AIS data. Further, these native algorithms have also been modified by innovative adaptations to enhance the analytical outputs.

One of the popular algorithms used for data analytics in the maritime domain is DBSCAN (Density-Based Spatial Clustering of Applications with Noise). Due to its inherent property of determining clusters of arbitrary sizes, it is well suited to cluster AIS data. As is expected, multiple adaptations of DBSCAN have been proposed to produce additional insights and develop applications.

One category of maritime adaptations of DBSCAN involves the use of multiple attributes while calculating the distance between points for forming clusters. Spatial attributes, along with non-spatial attributes, have been commonly used in various combinations and methodologies to determine the outputs. Temporal information has, however, not been used as an integral part of the attributes for calculating the distance in maritime applications. Certain significant attribute-based adaptations of DBSCAN in the maritime domain have been reviewed in this paper, and the influence of temporal attributes on the clustering results has been examined. The motivation is to determine if the use of temporal components can affect the results of clustering and possibly produce refined clustering. This could then be used in various maritime applications requiring such insights. Assessment of the inclusion of the Time field along with non-spatial components of AIS data has been undertaken experimentally in this study, and it has been seen that the performance of this adaptation is superior to other algorithms.

The composition of the paper is as follows: Section 2 brings out related work in this domain. Section 3 introduces the adaptation of DBSCAN being proposed in this paper, while the experimental settings are presented in Section 4. Section 5 presents the details of the experiments and compares them with other DBSCAN adaptations. Section 6 presents the conclusion and future work on this subject.

II. RELATED WORK

An easy way to comply with the conference paper formatting requirements is to use this document as a template and simply type your text into it.

A. AIS Data

The AIS system is installed onboard the ship and transmits in the VHF band of frequencies. While earlier, there were two frequencies for AIS, this has now been increased to four to account for the long-range detection of the messages by satellites [3]. Specifications of the AIS systems are brought out in [4].

The AIS data is transmitted between vessels in the form of 27 different types of messages, which can be categorised into three primary groups: the identification of the vessel, type, and dimensions are static messages, and Spatial, temporal and movement-related information to form part of the dynamic

messages. Details concerning the current journey of the vessel, such as destination, are the Voyage-related messages. An extract of the AIS message is shown in Fig. 1, while a graphical representation is shown on GIS software in Fig. 2. The data has been downloaded from MarineCadastre [5]. The rate at which the messages are transmitted has been promulgated and can be up to a message every 2 seconds.

Various data analytic techniques can be applied to AIS data to determine insights, which include Rules-based, Clustering, classification, Pattern Recognition, etc. [6], [7], [8].

MMSI	BaseDateTime	LAT	LON	SOG	COG	Heading	VesselName	IMO
235076283	04-10-2022 01:...	34.23407	-121.55728	0.51162790697...	0.96763105127...	0.65753424657...	PARAMOUNT H...	IMO9453999
538008452	06-10-2022 14:...	34.23409	-121.84982	0.42790697674...	0.73388713835...	0.51076320939...	PHILOXENIA	IMO9857250
338371000	05-10-2022 23:...	34.2341	-121.7787	0.71162790697...	0.89802348897...	0.61839530332...	POLAR ENDEAV...	IMO9193551
369567000	05-10-2022 00:...	34.23435	-121.62474	0.39534883720...	0.36006874820...	0.24266144814...	SEA RELIANCE	IMO9275878
636021005	06-10-2022 10:...	34.23446	-121.78094	0.53023255813...	0.72214265253...	0.50880626223...	MATSON MOL...	IMO9338084
369040000	04-10-2022 14:...	34.23454	-121.8798	0.46976744186...	0.94757949011...	0.65557729941...	AMERICAN EN...	IMO9759886
538008452	06-10-2022 14:...	34.23471	-121.84633	0.42325581395...	0.73302778573...	0.51076320939...	PHILOXENIA	IMO9857250

Fig. 1 Extract of AIS data (Source: MarineCadastre [5])

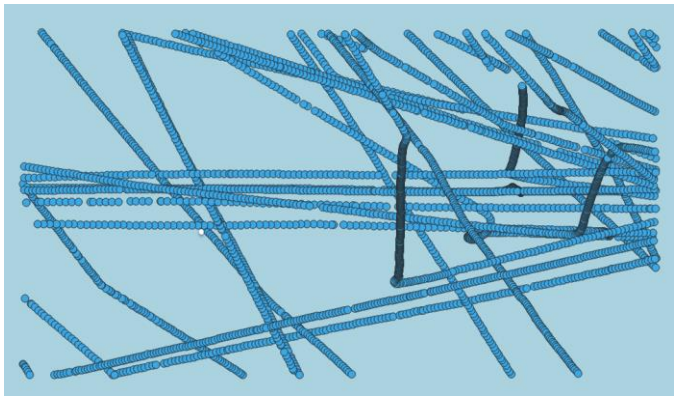

Fig. 2 Plot of AIS data on QGIS

B. DBSCAN

After being first presented in [9], DBSCAN has been extensively used in clustering data of various domains. It has the ability to form arbitrary-size clusters, and arbitrary-shaped clusters are feasible using this method. In order to cluster the data using DBSCAN, two parameters need to be selected. One is the radius of the neighbourhood around a point, also called Eps, and the other is MinPts, which states the minimum points required in a cluster. Core, Border and Noise are the three types of points defined in a DBSCAN clustering. Points having a neighbourhood comprising at least MinPts within Eps distance are called Core points. Clusters are formed with Core points that are neighbours. Border points are those points within a cluster which are not Core points; a noise point is one that is neither a Core nor a Border point. DBSCAN has been used frequently in various maritime applications, which include fishing, anomaly detection and the determination of patterns [10], [11], [12].

C. Multi-attribute Adaptations of DBSCAN in Maritime Applications

While many types of adaptations of DBSCAN have been undertaken, only those adaptations that take into account multiple attributes of the AIS data as an integral component of DBSCAN in calculating the distance in the maritime domain are being examined.

1) ***Multi-Dimensional DBSCAN:*** Multi-Dimensional DBSCAN (MD-DBSCAN) has been developed in [13] towards the extraction of traffic routes. The similarity is determined by considering Course as an additional attribute in DBSCAN along with spatial components, as shown in Equation (1).

DBSCAN Attributes = [Latitude, Longitude, Course] (1)

2) ***Mod-DBSCAN:*** Mod-DBSCAN extracts route patterns of vessels for which trajectory clustering is undertaken [14] by using speed and course as additional information (equation (2). Using coordinates to measure spatial distance, course to measure direction distance and average speed to measure speed distance, these independently calculated values are then converted into a synthetic distance, which can also be weighted, as shown in Equation (2).

DBSCAN Attributes = [Latitude, Longitude, Course, Speed] (2)

Synthetic Distance = $w_{sp}*d_{sp} + w_{cr}*d_{cr} + w_{sp}*$ (3)

where w_{sp}, w_{cr} and w_{sp} are the weights, and d_{sp}, d_{cr} and d_{sp} are the space, speed and direction weights and distances, respectively.

3) ***Optimised DBSCAN:*** Optimised DBSCAN enables the modelling of vessel behaviours by considering speed, course, and heading along with the spatial parameters [15], [16] as seen in equation (4). It also uses the Mahalanobis distance metric in lieu of the default Euclidean distance metric.

DBSCAN Attributes = [Latitude, Longitude, Course, Speed, Heading] (4)

4) ***Velocity DBSCAN***. Velocity DBSCAN tries to find out the main waypoint areas by considering the velocity changes [17] at the start and end of the change, thus forming a 4D vector as shown in equation (5).

DBSCAN Attributes = [Latitude, Longitude, Velocity direction before, Velocity direction after] (5)

5) ***DBSCAN_SD***. DBSCAN_SD enhances the original DBSCAN by considering Speed and Direction [18], [19] and stating that the points are neighbours if, apart from spatial proximity, the speed and direction difference must also be within a defined range, as shown in equation (6).

DBSCAN Attributes = [Latitude, Longitude, Direction difference, Speed difference] (6)

A variation of DBSCAN_SD is presented in [20], wherein the standard deviation is used instead of the maximum variation of direction and speed, which is shown in equation (7).

DBSCAN Attributes = [Latitude, Longitude, Direction standard deviation, Speed standard deviation] (7)

6) ***Dynamic Parameter DBSCAN***. Dynamic Parameter DBSCAN (DP-DBSCAN) [21] considers speed, course and

their variance, average and median, along with the spatial components in order to determine the similarity between trajectories, as given in equation (8).

DBSCAN Attributes = [Latitude, Longitude, speed, average speed, speed standard deviation, speed variable interval, course, average course, course standard deviation, course variable interval] (8)

D. Findings

As can be seen in the section above, the adaptions of DBSCAN for maritime applications using additional fields for calculating the distance primarily include non-spatial components. Temporal components have not been considered as an integral part of DBSCAN for distance measurements. Therefore, an assessment has been undertaken to examine the effect of the inclusion of temporal components as integral to DBSCAN and evaluate and compare the results with other implementations.

III. MODIFIED ALGORITHM

The native DBSCAN algorithm considers only the spatial components, viz. latitude and longitude, in the case of AIS data. The higher dimension maritime adaptations of DBSCAN have included speed, course, and their variations, together or in parts. In the present case, in order to examine the influence of the temporal component, Higher Dimension and Time DBSCAN (HDT-DBSCAN) is proposed in which DBSCAN has been modified to include non-spatial components, viz. speed, course and heading and Time has also been included as an integral component. Therefore, the attributes are as given in equation (8).

DBSCAN Attributes = [Latitude, Longitude, Speed, Course, Heading, Time] (8)

In this experiment, all attributes have been given equal weights. The distance metric used is Mahalanobis. The time field has been converted to Epoch for the calculations to be undertaken.

IV. EXPERIMENAL SETTINGS

A. Data

Historic AIS data hosted on the MarineCadastre website [5] has been used for the experiments. This data is from the coastal regions of the United States.

B. Data Pre-processing

Pre-processing of the data has been undertaken to ensure that the data is consistent. The data is examined for issues such as missing values, erroneous values, etc. and corrected accordingly.

C. Normalisation

This is an important step in the process to ensure optimal clustering. The normalisation of all non-spatial fields within the data is undertaken to ensure that the data is brought to a uniform scale, which will ensure no one field influences the results sub-optimally.

D. Clustering Algorithm

The DBSCAN algorithm has been modified for the experiments. The results have been compared with the native DBSCAN algorithm and with higher dimension DBSCAN (HD-DBSCAN).

E. Validation Assessment

Common performance metrics have been used to assess the results of the experiments. Both intrinsic metrics, which use the internal data, and extrinsic metrics, which use the ground truth information, have been used [22], [23]. The assessment has also been done graphically.

F. Environment for Development

The modification of the DBSCAN algorithm and evaluation with other algorithms has been undertaken in Python, while the visualisations and graphical assessments have been seen in QGIS.

V. RESULTS AND DISCUSSIONS

A. Method of Experiment

The flowchart of the clustering using the HDT-DBSCAN model is shown in Fig. 3. After preprocessing, the Time field has been converted to Epoch. Thereafter, the latitude, longitude, speed, course heading and time fields have been selected for further processing. Normalisation of the non-spatial components has been undertaken to bring them to the same scale, following which the selected data has been clustered using DBSCAN with Mahalanobis distance measure. The parameters of DBSCAN (EPS and MinPts) have been determined by trial and error. The experiments have been repeated with native DBSCAN, which considered only spatial components, and with HD-DBSCAN, which considered spatial and non-spatial components other than Time. The results of each experiment have been compared using clustering performance indices and also graphically.

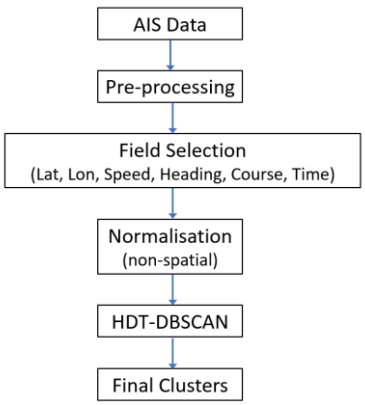

Fig. 3 Clustering Process

B. Assessment using Clustering Performance Measures

The results of clustering for each of the cases have been evaluated with intrinsic and extrinsic performance metrics. These are shown in Table 1 for Intrinsic and Table 2 for Extrinsic measures. It can be seen that the performance of HDT-DBSCAN is superior to that of DBSCAN and is enhanced over that of HD-DBSCAN.

TABLE I
INTRINSIC PERFORMANCE METRIC VALUES

Metric	Silhouette	Davies-Bouldin	Calinski-Harabasz
Native DBSCAN	-0.5719	1.83600	40.396
HD-DBSCAN	0.16838	1.60421	413.25
HDT-DBSCAN	0.34881	1.43963	587.44

TABLE II
EXTRINSIC PERFORMANCE METRIC VALUES

Metric	Homogeneity	Completeness	ARAND
Native DBSCAN	0.3205	0.5112	0.1099
HD-DBSCAN	0.9527	0.8588	0.7387
HDT-DBSCAN	0.9692	0.9065	0.7926
	AMIC	**VMEA**	**FM**
Native DBSCAN	0.3558	0.3940	0.2431
HD-DBSCAN	0.8966	0.9033	0.7560
HDT-DBSCAN	0.9332	0.9368	0.8042

C. Graphical Assessment

The clustering outputs have been observed graphically using QGIS, and these are shown in Fig. 4 for DBSCAN, Fig. 5 for HD-DBSCAN and Fig. 6 for HDT-DBSCAN. A similar assessment as the one that arrived during the performance metric examination can be seen. The clustering using HDT-DBSCAN is better than DBSCAN and improved over HD-DBSCAN.

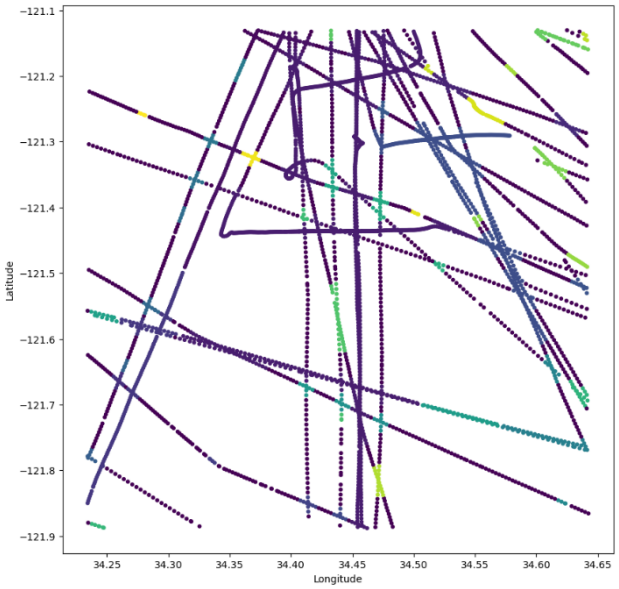

Fig. 4 DBSCAN

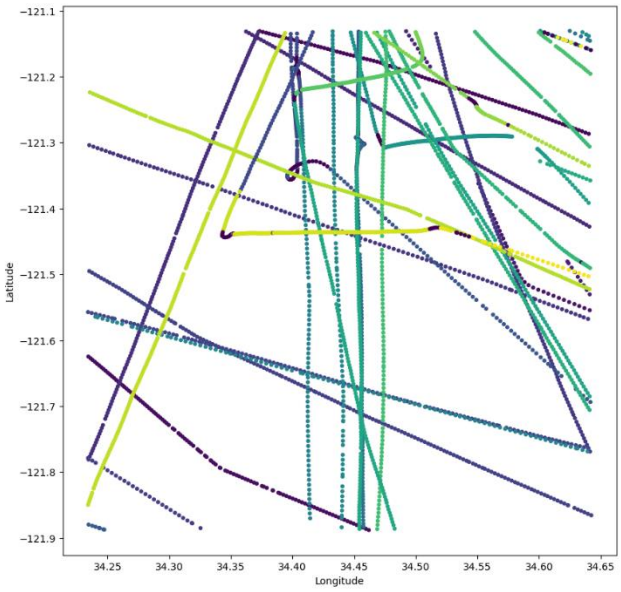

Fig. 5 HD-DBSCAN

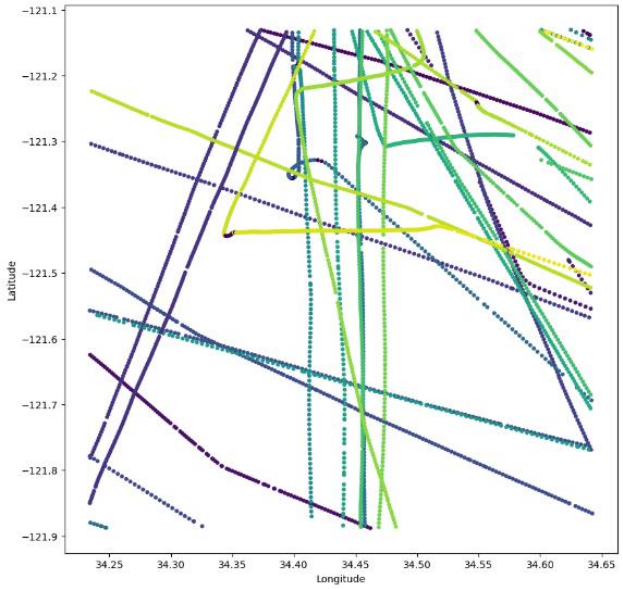

Fig. 6 HDT-DBSCAN

D. Discussions

DBSCAN is a powerful clustering algorithm and, even in its native form, can reveal useful maritime clustering insights. The various modifications of DBSCAN make use of some characteristics of the data to enhance the clustering results, which may be useful for certain applications. Generally, with such enhancement, the clustering outputs become more refined and intricate details and insights become evident. This is evident in the three clustering results using the different cluster models.

The influence of Time as an integral component of DBSCAN has been studied, and the HDT-DBSCAN model evolved. The HDT-DBSCAN, which includes non-spatial parameters, including Time, gives much more refined results. The various trajectories are also discernible to a larger extent. This is not feasible with the other two clustering algorithms. The inclusion of Time has, thus, enhanced the clustering results, improving on the other algorithms.

Determination of trajectories is of immense use in the maritime domain, and efforts to determine these reliably are actively being researched. HDT-DBSCAN is able to progress efforts in this direction. Further refinement is necessary to enable even better identification of the trajectories in an unsupervised method.

VI. CONCLUSION AND FUTURE SCOPE

Clustering maritime AIS data is a complex activity due to its peculiarities. Clustering results, however, are extremely useful in analysing the characteristics of maritime movement and can be used in many applications.

This paper examines an innovative adaptation of the DBSCAN clustering algorithm. The new adaptation, HDT-DBSCAN, takes into consideration non-spatial components of the AIS data, including Time, while undertaking clustering of AIS data. The results of this clustering show that the clustering outputs are enhanced as compared to DBSCAN and HD-DBSCAN. The influence of the Time field in refining the output of DBSCAN clustering is evident. Thus, this adaptation of DBSCAN will find use in undertaking the clustering of maritime movement data for various applications.

There is scope for examining this adaptation further by considering weighting the various non-spatial components. This will be particularly useful in case the influence of a particular parameter is to be studied. Modifications to en-able accurate determination of trajectories can also be researched.

ACKNOWLEDGMENT

AIS data from the Bureau of Ocean Energy Management (BOEM) and the National Oceanic and Atmospheric Administration (NOAA). MarineCadastre.gov. Vessel Traffic Data. Retrieved Oct 01, 2022, from marinecadastre.gov/data.

REFERENCES

[1] "Merchant fleet – UNCTAD Handbook of Statistics 2023." Accessed: Mar. 05, 2024. [Online]. Available: https://hbs.unctad.org/merchant-fleet/

[2] "AIS transponders." Accessed: Mar. 05, 2024. [Online]. Available: https://www.imo.org/en/OurWork/Safety/Pages/AIS.aspx

[3] "Use of Appendix 18 to the Radio Regulations for the maritime mobile service M Series Mobile, radiodetermination, amateur and related satellite services," 2014, Accessed: Nov. 26, 2023. [Online]. Available: http://www.itu.int/ITU-R/go/patents/en

[4] "Technical characteristics for a universal shipborne automatic identification system using time division multiple access in the VHF maritime mobile band (Question ITU-R 232/8)," 1998.

[5] "AccessAIS - MarineCadastre.gov." Accessed: Nov. 26, 2023. [Online]. Available: https://marinecadastre.gov/accessais/

[6] M. Hadzagic, M. O. St-Hilaire, S. Webb, and E. Shahbazian, "Maritime traffic data mining using R," *Proceedings of the 16th International Conference on Information Fusion, FUSION 2013*, no. November 2015, pp. 2041–2048, 2013.

[7] P. Schmitt, M. L. Bartosiak, and T. Rydbergh, "Spatiotemporal Data Analytics for the Maritime Industry," 2021, pp. 335–353. doi: 10.1007/978-3-030-50892-0_20.

[8] N. Le Guillarme and X. Lerouvreur, "Unsupervised extraction of knowledge from S-AIS data for maritime situational awareness," *Proceedings of the 16th International Conference on Information Fusion, FUSION 2013*, pp. 2025–2032, 2013.

[9] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," *Proceedings of*

the 2nd International Conference on Knowledge Discovery and Data Mining, pp. 226–231, 1996.

[10] V. F. Arguedas, F. Mazzarella, and M. Vespe, "Spatio-temporal data mining for maritime situational awareness," *MTS/IEEE OCEANS 2015 - Genova: Discovering Sustainable Ocean Energy for a New World*, 2015, doi: 10.1109/OCEANS-Genova.2015.7271544.

[11] Y. Li, Y. Zhang, and F. Zhu, "The Method of Detecting AIS Isolated Information Based on Clustering and Distance Van Li, Yingjun Zhang, Feixiang Zhu," *2016 2nd IEEE International Conference on Computer and Communications*, 2016.

[12] V. F. Arguedas, G. Pallotta, and M. Vespe, "Maritime Traffic Networks: From Historical Positioning Data to Unsupervised Maritime Traffic Monitoring," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 3, pp. 722–732, 2018, doi: 10.1109/TITS.2017.2699635.

[13] 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, no. July 2022, p. 113036, 2023, doi: 10.1016/j.oceaneng.2022.113036.

[14] P. Sheng and J. Yin, "Extracting shipping route patterns by trajectory clustering model based on Automatic Identification System data," *Sustainability (Switzerland)*, vol. 10, no. 7, 2018, doi: 10.3390/su10072327.

[15] X. Han, C. Armenakis, and M. Jadidi, "Modeling vessel behaviours by clustering ais data using optimized dbscan," *Sustainability (Switzerland)*, vol. 13, no. 15, Aug. 2021, doi: 10.3390/su13158162.

[16] X. Han, C. Armenakis, and M. Jadidi, "DBscan optimization for improving marine trajectory clustering and anomaly detection," *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, vol. 43, no. B4, pp. 455–461, 2020, doi: 10.5194/isprs-archives-XLIII-B4-2020-455-2020.

[17] P. Coscia, P. Braca, L. M. Millefiori, F. A. N. Palmieri, and P. Willett, "Multiple Ornstein–Uhlenbeck Processes for Maritime Traffic Graph Representation," *IEEE Trans Aerosp Electron Syst*, vol. 54, no. 5, 2018.

[18] B. Liu, E. N. De Souza, S. Matwin, and M. Sydow, "Knowledge-based clustering of ship trajectories using density-based approach," *Proceedings - 2014 IEEE International Conference on Big Data, IEEE Big Data 2014*, pp. 603–608, 2015, doi: 10.1109/BigData.2014.7004281.

[19] X. Wang, X. Liu, B. Liu, E. N. De Souza, and S. Matwin, "Vessel route anomaly detection with Hadoop MapReduce," *Proceedings - 2014 IEEE International Conference on Big Data, IEEE Big Data 2014*, pp. 25–30, 2014, doi: 10.1109/BigData.2014.7004464.

[20] I. Kontopoulos, I. Varlamis, and K. Tserpes, "A distributed framework for extracting maritime traffic patterns," *International Journal of Geographical Information Science*, pp. 1–26, 2020, doi: 10.1080/13658816.2020.1792914.

[21] M. Zhang *et al.*, "A method for the direct assessment of ship collision damage and flooding risk in real conditions," *Ocean Engineering*, vol. 237, 2021, doi: 10.1016/j.oceaneng.2021.109605.

[22] "2.3. Clustering — scikit-learn 1.3.2 documentation." Accessed: Nov. 27, 2023. [Online]. Available: https://scikit-learn.org/stable/modules/clustering.html#clustering-performance-evaluation

[23] "Introduction to clustering evaluation - Hyperskill." Accessed: Nov. 27, 2023. [Online]. Available: https://hyperskill.org/learn/step/28809