Proceedings of the 35th European Safety and Reliability & the 33rd Society for Risk Analysis Europe Conference
Edited by Eirik Bjorheim Abrahamsen, Terje Aven, Frederic Bouder, Roger Flage, Marja Ylönen
 Published by Research Publishing, Singapore.
doi: 10.3850/978-981-94-3281-3_ESREL-SRA-E2025-P6110-cd

A Surrogate Ship Trajectory Construction Method for Efficient Similarity Measurement in AIS Data Clustering Analysis

Shaoqing Guo *

** Corresponding Author, Research group on Safe and Efficient Marine and Ship Systems, Marine and Arctic Technology, Department of Mechanical Engineering, Aalto University, Finland; Kotka Maritime Research Centre, Finland. Email: shaoqing.guo@aalto.fi*

Victor Bolbot

Research group on Safe and Efficient Marine and Ship Systems, Marine and Arctic Technology, Department of Mechanical Engineering, Aalto University, Finland; Kotka Maritime Research Centre, Finland. Email: victor.bolbot@aalto.fi

Osiris A. Valdez Banda

Research group on Safe and Efficient Marine and Ship Systems, Marine and Arctic Technology, Department of Mechanical Engineering, Aalto University, Finland; Kotka Maritime Research Centre, Finland. Email: osiris.valdez.banda@aalto.fi

Since the advent of Automatic Identification System (AIS) has opened opportunities for shipping data to be disseminated worldwide, trajectory clustering has seen increasing applications in maritime traffic pattern recognition, trajectory prediction, anomaly detection, and route planning. Trajectory similarity measurement is a central concept in ship trajectory clustering, where the majority of computational time is spent on similarity calculations. However, the exponentially growing volume of AIS messages has posed significant challenges to efficient processing, with popular trajectory simplification methods such as Douglas-Peucker (DP) algorithm showing limited effectiveness in improving trajectory similarity calculations. In this study, we propose a novel surrogate ship trajectory construction (SurTraC) method to reduce the complexity of similarity calculations, where the Geohash gridding technique is employed to aggregate spatially adjacent points. The method can generate an alternative sparse trajectory that uniformly and precisely represents the original one. A case study using one-week AIS data from Gulf of Finland indicates that SurTraC can effectively simplify the trajectory dataset while maintaining the entirety of the features. Compared to the DP-based methods proposed in previous research, a discussion from the perspectives of trajectory simplification, similarity measurement, and clustering demonstrates that SurTraC can significantly accelerate similarity measurement without compromising clustering performance.

Keywords: SurTraC, Surrogate ship trajectory, Trajectory simplification, Similarity measurement, Maritime big data, AIS, Geohash, Clustering, DBSCAN, Gulf of Finland

1. Introduction

In recent years, with the worldwide dissemination of Automatic Identification System (AIS) data, trajectory clustering has emerged as a powerful tool to identify recurring patterns and group similar ship trajectories (Bai et al. 2023). This technique has facilitated numerous maritime traffic pattern recognition applications, including trajectory prediction (Chen et al. 2024), anomaly detection (Guo et al. 2021a), and route planning (Yan et al. 2023), significantly enhancing navigational risk assessment and decision-making processes (Guo et al. 2023).

Assessing ship trajectory similarity is central to trajectory clustering, as accurate measurement identifies similar features and reveals repetitive ship behaviors (Zhao and Shi 2019). Given that a ship trajectory inherently exhibits typical spatiotemporal characteristics, similarity measurement is often performed using spatial distance as a primary metric (Yang et al. 2022). To comprehensively evaluate trajectory similarities, advanced metrics in addition to

spatial distance such as speed and course differences, and trajectory length are also evidenced in (Zhang et al. 2021). Since this study focuses on investigating the computational efficiency of similarity measurement, we mainly adopt spatial distance as a metric for subsequent analysis.

The increasing volume of AIS data has introduced significant computational challenges due to the extensive number of pairwise point calculations involved (Taha and Hanbury 2015). To mitigate this concern, researchers have primarily focused on applying trajectory simplification methods to reduce the number of input points, where Douglas-Peucker (DP) algorithm is the most widely used (Zhang et al. 2018). Although DP algorithm has shown strong compression capability in previous studies, the substantial data volume and its continuous growth still result in inefficient similarity measurement (Guo, Bolbot, and Valdez Banda 2024). In addition, the uneven spatial distribution of the trajectory points after DP compression may also lead to inaccurate similarity measurement for the methods relying on pairwise point distance, such as Hausdorff distance (Huang and Guan 2024). Accordingly, overcoming these limitations remains a critical yet challenging study for precise and efficient similarity measurement.

This paper proposes a novel surrogate trajectory construction method named SurTraC for efficient similarity measurement in clustering analysis. This method applies Geohash gridding technique to effectively balance a high compression rate with the preservation of the overall trajectory features, simultaneously achieving a uniform distribution of simplified trajectory points. To further validate the effectiveness of the proposed method, a case study using AIS data from Gulf of Finland was conducted. By utilizing Hausdorff distance and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) method, we evaluated the processing efficiency and accuracy of our method by comparing it with the results obtained from the original dataset, as well as the compressed trajectory dataset by DP-based algorithms.

The remainder of the paper is organized as follows: Section 2 elaborates on the methodology of this study. Section 3 presents a case study to illustrate the surrogate trajectory construction results, followed by a thorough discussion in Section 4 to analyze the findings and their implications. Finally, Section 5 summarizes the key contributions of the study and outlines a potential direction for future improvement.

2. Methodology

Unlike traditional DP algorithm that adopts a selective reduction strategy to preserve key feature points (Douglas and Peucker 1973), the method proposed in this study employs a representative aggregation strategy to simplify the trajectory while maintaining its overall features. The flowchart of SurTraC is shown in Fig 1. The process begins with Geohash encoding, which converts trajectory points into Geohash strings for efficient spatial indexing. Then, a bucketing process is applied to group points with the same Geohash strings. Next, centroid point calculation is performed for each bucket, obtaining a representative point to embody the overall features of this bucket. Finally, by connecting centroid points sequentially, the surrogate trajectory is constructed.

2.1. *Geohash Encoding*

Geohash is a hierarchical spatial data structure that encodes geographic coordinates into a compact base32 string representation, facilitating efficient spatial querying and indexing (Li et al. 2024). It consists of two steps: spatial subdivision and binary encoding.

As depicted in Fig 2, the spatial subdivision iteratively divides the Earth's surface into progressively smaller cells by alternately bisecting longitude and latitude ranges. For illustration, this

figure does not consider the spherical nature of the Earth. Each resulting cell inherits a unique binary identifier based on its position relative to the parent cell. This process shares fundamental principles with the Z-order curve, as both approaches preserve spatial locality through a similar bit-interleaving mechanism (Wightman et al. 2022).

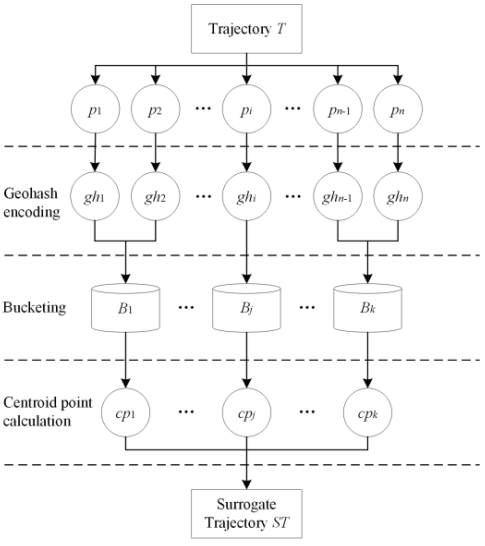


Fig 1. Flowchart of the proposed method.

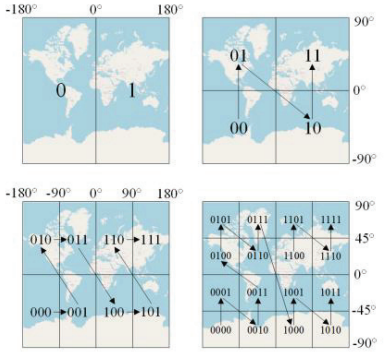


Fig 2. Spatial subdivision.

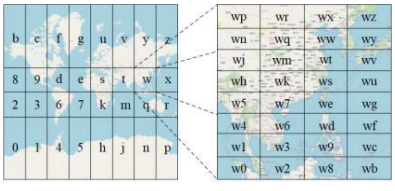
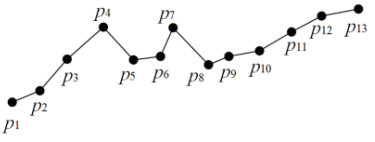
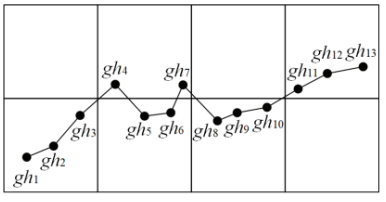


Fig 3. Binary encoding.

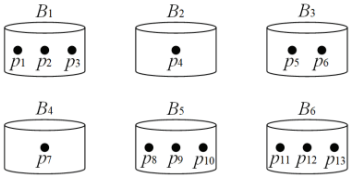
The binary encoding process then groups the interleaved binary sequences into blocks of 5 bits. Each 5-bit block is mapped to a corresponding base32 character, generating the final Geohash string. As shown in Fig 3, where the spherical nature of the Earth is considered, an increase in precision results in a longer Geohash string, with each additional character reducing the size of the encoded geographic area by a factor of 32. To conclude, Geohash precision defines the number of characters in the Geohash string and consequently determines the size and range of the grid it represents.



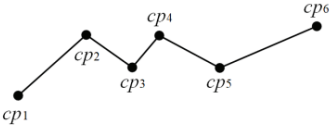
(a) Original trajectory



(b) Geohash encoding



(c) Bucketing



(d) Surrogate trajectory

Fig 4. Example of surrogate trajectory derivation.

In this study, a ship trajectory is denoted by $T = \{p_1, p_2, ..., p_n\}$, where each trajectory point is defined as $p_i = (t_i, x_i, y_i, s_i, c_i)$. Here, t_i denotes the timestamp of the point, x_i and y_i represent

longitude and latitude respectively, while s_i and c_i indicate the speed over ground (SOG) and course over ground (COG). As shown in Fig 4(a) and Fig 4(b), by applying Geohash encoding under a certain precision, each trajectory point is mapped to a Geohash string.

2.2. *Bucketing*

Recognizing that AIS trajectory data inherently follows a high sampling frequency, several consecutive points may share the same Geohash code. Hence, the bucketing process aims to group those points within the same geographical area and assign them to a bucket. The process is described as follows:

First a bucket $B_1 = \{p_1\}$ is created. By traversing the trajectory, if the current point p_{cur} has the same Geohash string as the previous point p_{pre}, which means $gh_{cur} = gh_{pre}$, then p_{cur} is put into the same bucket; otherwise, a new bucket $B_j = \{p_{cur}\}$ is created. This process is repeated until all trajectory points have been visited. An example of bucketing is illustrated in Fig 4(c).

2.3. *Centroid Point Calculation*

Previous research has suggested that redundant information in ship trajectories ought to be removed for efficient similarity measurements (Huang et al. 2023). Since each bucket represents several points that are located closely to each other, provided that the Geohash precision is appropriately set, we could assume that they demonstrate similar ship dynamic features. Therefore, a centroid point calculation process is designed to find the representative point of each bucket, reducing the number of points that need to be processed.

Taking bucket $B_i = \{p_{i1}, p_{i2}, ..., p_{i|B_i|}\}$ as an example, its centroid point is described as:

$$cp_i = \left(\frac{1}{n}\sum_{j=1}^{|B_i|} t_{ij}, \frac{1}{n}\sum_{j=1}^{|B_i|} x_{ij}, \frac{1}{n}\sum_{j=1}^{|B_i|} y_{ij}, \frac{1}{n}\sum_{j=1}^{|B_i|} s_{ij}, \frac{1}{n}\sum_{j=1}^{|B_i|} c_{ij} \right) \quad (1)$$

where $|B_i|$ is the number of points in B_i. The arithmetic mean is appropriate for AIS trajectory data owing to their high density and frequent reporting, allowing the centroid point to effectively represent the trajectory by leveraging the similarity of nearby points. As shown in Fig 4(d), by sequentially connecting the centroid points, the surrogate trajectory is formed. This approach significantly reduces the number of trajectory points, thereby simplifying the representation of the trajectory while preserving its essential spatiotemporal features.

3. Case Study

The AIS data for case study was collected from Gulf of Finland over a one-week period, spanning from June 1 to June 7, 2022. The experiments were conducted using Python 3.8 on a Windows 10 computer with an Intel Core i7-12700KF processor, 32 GB RAM, and a 64-bit operating system.

3.1. *Data Preprocessing*

Raw AIS data is essentially a collection of disordered ship position reports rather than sequential trajectory data (Guo et al. 2021b). Additionally, it inevitably contains anomalies affected by the transmission environment. Therefore, we employed the method proposed by Guo et al. (2021b) to clean and reorganize the trajectory dataset in dictionary order of (MMSI, timestamp). Finally, 2068 trajectories with 211686 points were obtained, whereas the visualization of the trajectory dataset is provided in Fig 5.

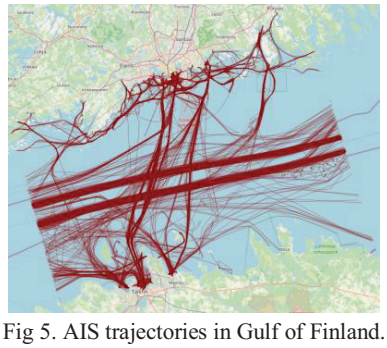


Fig 5. AIS trajectories in Gulf of Finland.

3.2. *Precision Determination*

Using a centroid point to represent a group of points assumes an appropriate Geohash precision.

According to the purpose of improving similarity measurement computation, we analyzed the correlation between Geohash precision and compression rate based on a one-day subset to determine the optimal precision for this study.

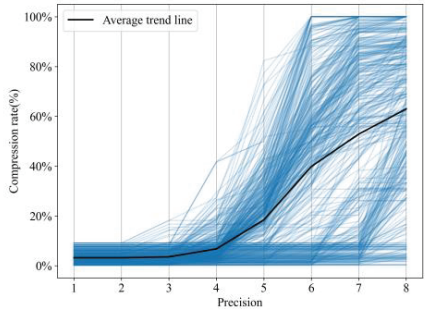


Fig 6. Correlation between precision and compression rate.

As shown in Fig 6, when the precision reaches 4, the compression rate decreases significantly with further increases in precision. This is because higher precision results in progressively smaller grid sizes, leading to a sharp reduction in the number of trajectory points within each grid. This phenomenon indicates that higher precision rapidly approaches the spatial granularity of AIS trajectory data, thereby losing its ability to compress redundant points effectively. Consequently, a Geohash precision of 4 was selected for the experiments in this study.

3.3. *Surrogate Trajectories*

By setting the Geohash precision to 4, Fig 7 displays an example of ship XXX471XXX. Due to confidentiality reasons, part of the MMSI digits have been concealed. The ship departed from Kilpilahden Öljysatama port in Sköldvik and sailed southwestward. The original trajectory contains 150 points, while the surrogate trajectory reduces the number to 8. Despite some inherent distortions, the substitute trajectory successfully captures the essential features of the original one.

Moreover, Fig 8 presents the entire trajectory dataset after the surrogating, with point numbers decreasing from 211686 to 8382, reaching a compression rate of 96.04%. Owing to a large number of trajectory points being aggregated into centroid points, the visualization of the surrogate trajectories reveals a predominantly linear feature. Nevertheless, the overall trend is captured, as shown in comparison with Fig 5.

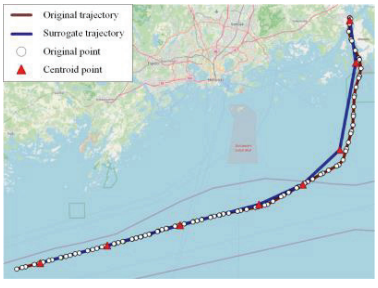


Fig 7. Example of ship XXX471XXX.

Fig 8. Surrogate trajectories.

4. Discussion

In order to validate the performance of the proposed method, the DP algorithm, which is among the most prevalent trajectory simplification methods in ship trajectory clustering analysis, was applied for a comparison experiment. First, to examine the performance of DP algorithm under a similar compression rate, a threshold of 1000 m was set for the traditional DP algorithm. Then, as suggested by previous studies (Guo, Bolbot, and Valdez Banda 2024), an advanced version of DP algorithm employs an adaptive threshold setting based on 0.8 times the ship length. Hence, we adhered to these settings in the validation experiments and denoted these two comparative methods by DP and ADP. The

experiment was divided into three components: trajectory simplification, similarity measurement, and clustering.

4.1. *Trajectory Simplification*

Trajectory simplification is typically the first step in clustering analysis. As listed in Table 1, SurTraC spent only 0.69 s to compress 96.04% of the trajectory points. DP method was controlled to have the same compression rate yet requiring 1.64 s to simplify the trajectory dataset. As for ADP, since more details were preserved, slower processing speed and lower compression rate were observed. In this case, SurTraC appears as the most efficient method.

Table 1. Trajectory simplification results

Method	Compression rate	Running time (s)
SurTraC	96.04%	0.69
DP	96.04%	1.64
ADP	85.09%	2.89

4.2. *Similarity Measurement*

Conducting similarity measurement on a simplified trajectory dataset can substantially shorten the computation time. Hausdorff distance is a widely used metric for clustering analysis in many studies owing to its robustness in assessing the similarity between sets (Liu, Yang, et al. 2024). It is also important to consider both the distance between two trajectories and their relative positions to other trajectories. Therefore, the following similarity function as recommended by Wang et al. (2021) was employed to calculate the similarity matrix:

$$sm_{ij} = \begin{cases} 1, & i = j \\ e^{-\frac{d_{ij}^2}{2\sigma_i\sigma_j}}, & i \neq j \end{cases} \quad (2)$$

Here, sm_{ij} is the element in the similarity matrix, indicating the similarity between i th and j th trajectories. d_{ij} is the Harsdorff distance between the two trajectories, with σ_i and σ_j denoting the mean Hausdorff distances of them to all other trajectories respectively.

Flattening the matrix into a one-dimensional array and applying Kernel Density Estimation (KDE) allows distribution analysis of similarity values across all trajectory pairs. The results are depicted in Fig 9, where the curves for the three methods closely resemble the original one. Despite the curve of SurTraC demonstrating slight distortion due to its nature of aggregating original trajectory points, the resemblance still indicates that all the methods effectively preserved the overall distribution of similarity values as exhibited in the original matrix. As a supplement, Table 2 presents the execution time for similarity matrix computation, where the compression time is not included. It is evident that using the original trajectory dataset to measure similarities is unbearable, which spent more than 20 hours to produce the similarity matrix. The removal of 85.09% points by ADP significantly facilitated the computation, which reduced the running time to 1531.81 seconds. However, the results for SurTraC and DP proved that there was still room for improvement, where the running times were further shortened to 97.11 and 97.86 seconds. As a result, SurTraC and DP algorithms with a threshold of 1000 m can help achieve more efficient similarity measurement for big trajectory data.

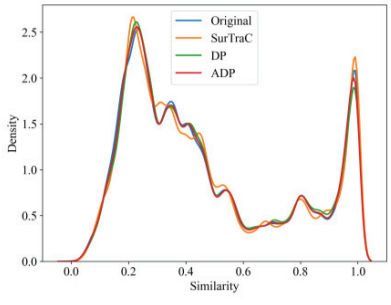


Fig 9. KDE of flattened similarity matrices.

Table 2. Similarity matrix computation times.

Method	Running time (s)	Running time (h)
Original	72621.75	20.17
SurTraC	97.11	0.03
DP	97.86	0.03
ADP	1531.81	0.43

4.3. *Clustering*

To further explore the impact of these trajectory simplification methods on clustering performance, the similarity matrices were input to DBSCAN clustering algorithm. Usually, DBSCAN can work with a distance matrix if it is precomputed. Therefore, we directly transformed similarity matrices into distance matrices for a more efficient clustering process through the following equation:

$$\delta_{ij} = 1 - sm_{ij} \tag{3}$$

where δ_{ij} is the element in the transformed distance matrix.

As can be seen in Fig 9, a prominent peak is evidenced near 0.985, indicating that many similar trajectories have pairwise similarities close to this value. Meanwhile, we are interested in frequently occurring trajectory patterns. Thus, the parameters of DBSCAN were set to eps = 0.015 and $minpts$ = 25 respectively.

The clustered trajectories as shown in Fig 10 reveal that our proposed method has negligible impact on the clustering results. The trajectories were classified into 13 clusters when original trajectories and surrogate trajectories were used for similarity measurement. With the adoption of DP and ADP, the dataset was grouped into 10 and 9 clusters respectively, resulting in the loss of several main trajectory clusters. This phenomenon can be explained by the simplifying principles of different methods. SurTraC employs Geohash gridding technique to introduce a structural simplification of trajectories. All trajectory points are mapped to designated grids for simplification, which helps similar trajectories to exhibit analogous simplified structures. In contrast, DP and ADP independently consider the characteristics of each trajectory, removing redundant points without accounting for the potential relationships between similar trajectories. As a consequence, they either oversimplify or under-simplify, leading to suboptimal clustering results or higher computational costs.

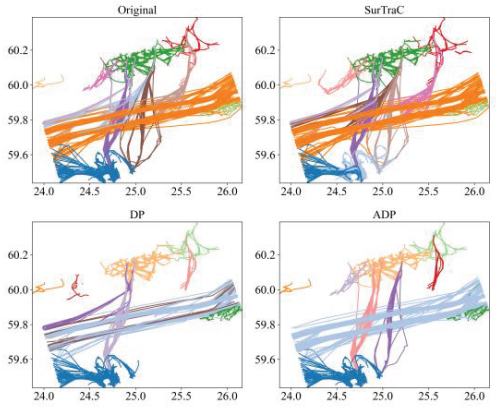


Fig 10. Clustering results.

5. Conclusions

As a pivotal role in facilitating maritime surveillance and supporting safe navigation decision-making, trajectory clustering analysis is currently facing the bottleneck problem of low computational efficiency owing to the rapid growth of AIS data volume. To address the challenge, this study proposes a novel method, SurTraC, to construct surrogate ship trajectories for efficient similarity measurement. The presented method can contribute to efficient processing in similarity measurement without compromising clustering performance. On the contrary, the experiment results prove that the widely applied DP-based algorithms struggle to simultaneously ensure efficient similarity computation and satisfactory clustering performance. This consequently underscores the substantial practical potential of our method for practical applications.

Nevertheless, the approach presents certain limitations that warrant further study. Considering that trajectories typically exhibit varying scale characteristics depending on navigational tasks and ship types, the use of fixed Geohash precision for surrogate trajectory computation may cause non-negligible distortion that impact similarity measurements. A promising future direction is to refine SurTraC method by adopting an adaptive precision selection approach.

Acknowledgement

This research was financially supported by China Scholarship Council (Grant Number: 202206955019), Safe, ClimAte Resilient Infrastructure (SAFARI) project funded by the European Union's Horizon Europe Programme, and Merenkulun säätiö. The authors want to express their gratitude to the Baltic Marine Environment Protection Commission (Helsinki Commission, HELCOM) for providing AIS data for the analyzed sea area.

References

Bai, Xiangen, Zhexin Xie, Xiaofeng Xu, and Yingjie Xiao. 2023. "An adaptive threshold fast DBSCAN algorithm with preserved trajectory feature points for vessel trajectory clustering." *Ocean Engineering* 280:114930.

Chen, Pengfei, Fengkai Yang, Junmin Mou, Linying Chen, and Mengxia Li. 2024. "Regional ship behavior and trajectory prediction for maritime traffic management: A social generative adversarial network approach." *Ocean Engineering* 299:117186.

Douglas, David H, and Thomas K Peucker. 1973. "Algorithms for the reduction of the number of points required to represent a digitized line or its caricature." *Cartographica: The International Journal for Geographic Information and Geovisualization* 10 (2):112-122.

Guo, Shaoqing, Victor Bolbot, Ahmad BahooToroody, Osiris A Valdez Banda, and Chee Loon Siow. 2023. "Identification of hazardous encounter scenarios using AIS data for collision avoidance system testing." In *Advances in the Collision and Grounding of Ships and Offshore Structures*, 43-50. CRC Press.

Guo, Shaoqing, Victor Bolbot, and Osiris Valdez Banda. 2024. "An adaptive trajectory compression and feature preservation method for maritime traffic analysis." *Ocean Engineering* 312:119189.

Guo, Shaoqing, Junmin Mou, Linying Chen, and Pengfei Chen. 2021a. "An anomaly detection method for AIS trajectory based on kinematic interpolation." *Journal of Marine Science and Engineering* 9 (6).

Guo, Shaoqing, Junmin Mou, Linying Chen, and Pengfei Chen. 2021b. "Improved kinematic interpolation for AIS trajectory reconstruction." *Ocean Engineering* 234.

Huang, Changhai, Xucun Qi, Jian Zheng, Ranchao Zhu, and Jia Shen. 2023. "A maritime traffic route extraction method based on density-based spatial clustering of applications with noise for multi-dimensional data." *Ocean Engineering* 268.

Huang, Zicong, and Keping Guan. 2024. "Ship trajectory clustering based on improved Hausdorff distance." Seventh International Conference on Traffic Engineering and Transportation System (ICTETS 2023).

Li, Yan, Bi Yu Chen, Qi Liu, and Yu Zhang. 2024. "Geohash coding-powered deep learning network for vessel trajectory prediction using clustered AIS data in maritime Internet of Things industries." *Computers Electrical Engineering* 120:109611.

Liu, Zhiyao, Haining Yang, Chenghuai Xiong, Feng Xu, Langxiong Gan, Tao Yan, and Yaqing Shu. 2024. "Research on the Optimization of Ship Trajectory Clustering Based on the OD–Hausdorff Distance." *Journal of Marine Science Engineering* 12 (8):1398.

Taha, Abdel Aziz, and Allan Hanbury. 2015. "An efficient algorithm for calculating the exact Hausdorff distance." *IEEE Transactions on Pattern Analysis Machine Intelligence* 37 (11):2153-2163.

Wang, Lianhui, Pengfei Chen, Linying Chen, and Junmin Mou. 2021. "Ship AIS trajectory clustering: An HDBSCAN-based approach." *Journal of Marine Science and Engineering* 9 (6):566.

Wightman, Pedro, Mayra Zurbaran, Augusto Salazar, and Lorena Garcia. 2022. "Hall of Mirrors: A novel strategy to address locality in geocoded-based PoI private queries." *IEEE Access* 10:61769-61783.

Yan, Zhaojin, Guanghao Yang, Rong He, Hui Yang, Hui Ci, and Ran Wang. 2023. "Ship trajectory clustering based on trajectory resampling and enhanced BIRCH algorithm." *Journal of Marine Science and Engineering* 11 (2).

Yang, Jiaxuan, Yuan Liu, Lingqi Ma, and Chengtao Ji. 2022. "Maritime traffic flow clustering analysis by density based trajectory clustering with noise." *Ocean Engineering* 249.

Zhang, Mingyang, Jakub Montewka, Teemu Manderbacka, Pentti Kujala, and Spyros Hirdaris. 2021. "A big data analytics method for the evaluation of ship-ship collision risk reflecting hydrometeorological conditions." *Reliability Engineering & System Safety* 213:107674.

Zhang, Shu-kai, Guo-you Shi, Zheng-jiang Liu, Zhi-wei Zhao, and Zhao-lin Wu. 2018. "Data-driven based automatic maritime routing from massive AIS trajectories in the face of disparity." *Ocean Engineering* 155:240-250.

Zhao, Liangbin, and Guoyou Shi. 2019. "A novel similarity measure for clustering vessel trajectories based on dynamic time warping." *The Journal of Navigation* 72 (2):290-306.