

PAPER • OPEN ACCESS

Trajectory Outlier Detection Algorithm for ship AIS Data based on Dynamic Differential Threshold

To cite this article: Sun Shuang *et al* 2020 *J. Phys.: Conf. Ser.* **1437** 012013

View the [article online](#) for updates and enhancements.

You may also like

- [Quantitative Performance of the Mopex Multi-Frame Outlier-Detection Algorithm](#)
Russ Laher, Andrew Grant and Fan Fang
- [An outlier detection-based method for artifact removal of few-channel EEGs](#)
He Chen, Hao Zhang, Chuancai Liu et al.
- [Locally adaptive decision in detection of clustered microcalcifications in mammograms](#)
Maria V Sainz de Cea, Robert M Nishikawa and Yongyi Yang



244th Electrochemical Society Meeting

October 8 – 12, 2023 • Gothenburg, Sweden

50 symposia in electrochemistry & solid state science



Deadline Extended!
Last chance to submit!

New deadline:
April 21
submit your abstract!

Trajectory Outlier Detection Algorithm for ship AIS Data based on Dynamic Differential Threshold

Sun Shuang*, Chen Yan, Zhang Jinsong

School of Maritime Economics and Management, Dalian Maritime University, Dalian, China

*470582807@qq.com, +86 18104704599

ABSTRACT. Trajectory outlier detection is one of the most important branches of data mining topics. Most existing outlier detection algorithms only utilized location of trajectory points and neglected some important factors such as speed, acceleration, and corner. In this paper, we propose a trajectory outlier detection algorithm based on dynamic differential threshold, called TODDT. Considering the influence of target motion information in a period of time, TODDT improves the traditional distance-based algorithm by combining dynamic differential threshold into outlier detection to discover more meaningful trajectory outliers. The experiments with real trajectory data sets in ship AIS (Automatic Identification System) show that TODDT algorithm performs efficiently and effectively when applied to the problem of trajectory outlier detection and the quality of trajectory data is greatly improved.

1. INTRODUCTION

With the popularity of ship-borne AIS equipment, the full coverage of AIS base station and the maturity of data management technology, ship AIS data have entered the era of big data. Ship Automatic Identification System (AIS) consists of shore-based facilities and ship-borne equipment [1]. It is a digital navigational aid system which integrates network technology, modern communication technology, computer technology and electronic information display technology. The core of AIS is to broadcast the GPS information of ships to other ships in the vicinity by VHF digital communication, so that the information can be shared among ships, and the automatic identification and monitoring of ships can be realized [2-4]. Due to the limitation of positioning technology and the influence of external interference factors (such as equipment failure and human operation error), there are a lot of outliers in trajectory data. The existence of these outliers seriously reduces the quality of trajectory data and results in inaccurate results of subsequent trajectory knowledge discovery, such as trajectory compression, trajectory clustering, path planning, outlier pattern detection and track association analysis.

Hawkins formally defined outliers in 1980: outliers are those data objects that deviate significantly from the vast majority of other data in the data set, causing people to suspect that they are generated by completely different mechanisms. Outlier detection algorithms have been widely studied at home and abroad, which can be divided into the following four categories:

(1) Statistical-based methods. Most of these methods are developed from anomaly detection methods for different distributions (such as t test, Dixon test, Grubbs test, Nair test, etc.). Usually users use distribution to fit data sets. Assuming that there is a distribution or probability model (such as normal distribution or Poisson distribution) in the given data set, and then the data inconsistent with



the model or the edge of the model distribution are determined as abnormal data. These methods are based on standard statistics and have a solid theoretical foundation.

(2) Distance-based methods. The basic idea of these methods is that if an object is far away from most other objects, then the object is abnormal. The advantage of these methods is that it is easier to determine the meaningful proximity measure of a data set than to determine its statistical distribution. It combines the distribution-based idea and overcomes the main shortcomings of the distribution-based method.

(3) Clustering-based methods. The basic idea of these methods is to cluster the input data with clustering algorithm, scan all the objects in the clustering results, and evaluate the degree of clustering. If an object is not strong enough to belong to any cluster, it is called an outlier based on clustering. For prototype-based clustering, the distance from the object to its cluster center can be used to measure the degree to which the object belongs to the cluster.

(4) Density-based methods. When a data set contains multiple distributions or data sets are mixed by different density subsets, whether the data is abnormal depends not only on the distance between the data set and the surrounding data, but also on the data density in the neighborhood.

(5) Methods based on extremum theory. The main idea of the methods based on extreme value theory is to judge whether the statistic is significantly higher than a critical value. In this way, the selection and evaluation criteria of critical value become the key.

In this paper, a trajectory outlier detection algorithm based on dynamic differential threshold (TODDT) is proposed. The algorithm considers the influence of ship motion state and noise in data in adjacent period, and dynamically calculates the speed threshold and differential threshold of current segment data. By analyzing the actual experimental data, this paper also introduces the limitation of the time length of the segment data and the number of points to select the segments. TODDT can automatically obtain adaptive and reasonable thresholds according to the motion information of the ship's whole trajectory and trajectory segment. The experimental results show that TODDT algorithm can detect outliers in ship trajectory data in real time.

2. DATA PREPROCESSING

Preprocessing is a basic step that performs at the beginning and it aims at improving quality of trajectory data and generating sub-trajectories [5-6]. Ship AIS provides ship static data, ship dynamic data and ship range data. Ship static data include ship name, call sign, MMSI, IMO, ship type, captain and ship width; ship dynamic data include longitude, latitude, ship head, track and speed; ship range data include ship status, draft and destination. Among them, ship dynamic data is the main source of ship speed and position longitude and latitude data. In this paper, the ship speed data is taken as the research object, and the algorithm of outlier detection of ship AIS trajectory is studied.

(1) Trajectory separation

The specific process of ship trajectory separation is divided into two steps: separating data from different ships and separating data from different trajectories of the same ship. MMSI is the only identification of a ship, which can be used as a basis for separating different ships. The ship's trajectory is discontinuous in different time periods, and different trajectories of the same ship can be separated according to AIS data timestamp.

(2) Time interval standardization

AIS data transmission time interval has a standard specification; however, a large number of AIS data acquisition time interval does not conform to the standard.

(3) Data cleaning

Longitudinal and latitudinal data of ship AIS are derived from GPS. In relative positioning of GPS, errors may occur due to atmospheric delay, multipath effect and diffraction. Cleaning data aims to discard impossible locations or trajectories exploiting specific constraints,

3. DYNAMIC DIFFERENTIAL THRESHOLD ALGORITHM

The ship track in ship AIS data consists of a series of spatio-temporal data points (i.e. longitude, latitude and time), which are usually obtained by various positioning methods [7-9]. Due to the difference of positioning errors between different data sources, environmental interference and human errors, there are a large number of abnormal points in ship trajectory data that do not conform to the law of target motion [10-11]. The original track of a ship is shown in Figure 1-2. There are many abnormal points in the data which deviate from the track seriously. Detection of abnormal points in the track is the most important step before the application of data processing.

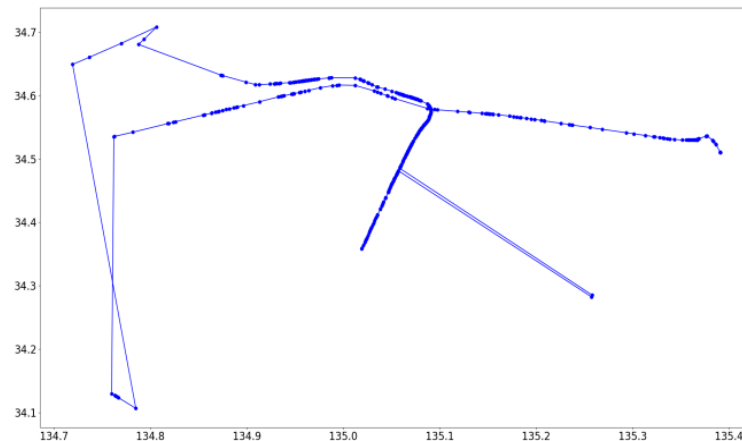


Figure 1. Original ship trajectory1.

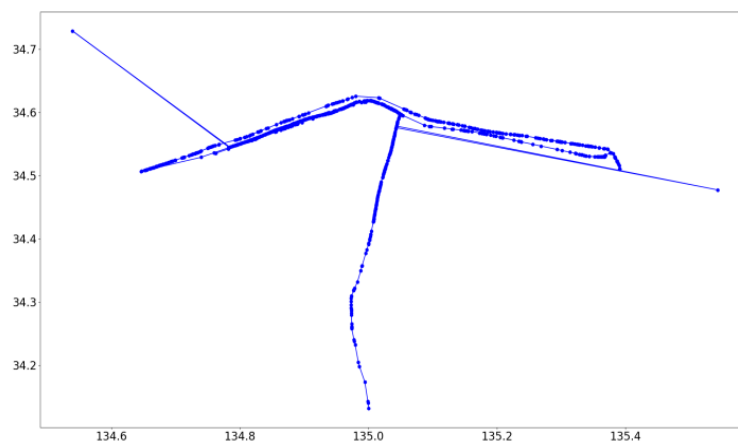


Figure 2. Original ship trajectory2.

The core idea of the differential dynamic threshold algorithm proposed in this paper is that because ships are large inertial systems, especially large ships have great inertia and time lag in the marine environment, it is impossible to make sudden changes in unit time, that is, the data of speed and position will not change abruptly in a large range. The algorithm dynamically calculates the speed threshold and differential threshold of current segment data. When both indexes are satisfied, they are normal points, and if one of them is not satisfied, they are determined as outliers. In the case of navigation data with large data background, dynamic setting of calculation threshold can reduce the false detection rate, and dual thresholds of speed threshold and differential threshold can improve the reliability of the results.

3.1 Local Threshold Calculation

Local thresholds are divided into velocity thresholds and differential thresholds. Considering that the motion state of the point to be detected is closely related to the motion state of the adjacent track, the

algorithm processes the data to be detected dynamically in segments. Through the analysis of the trajectory data and referring to the parameter settings in [12], the time length of the segment data and the limitation of the number of points are set in Table 1. When the conditions are satisfied, the piecewise threshold is selected which is calculated from the data in the segment, otherwise, the global threshold which is calculated from the whole trajectory information is selected. This strategy can effectively detect the outliers whose speed suddenly increases after the target travels at a steady speed for a certain period of track.

Table 1. Conditions of segment data

Condition name	Value	Condition name	Value
Upper bound of speed threshold	40kn	Lower bound of speed threshold	15kn
Upper bound of differential threshold	0	Lower bound of differential threshold	15
Global speed multiplier	5	Piecewise speed multiplier	7
Global difference multiplier	5	Piecewise difference multiplier	7
Minimum points of segment	7	Maximum points of segment	31
Minimum duration of segment	20s	Maximum duration of segment	5 min

Based on the 3σ criteria of mathematical statistics, this paper defines the threshold as mean plus a set multiple multiplied by standard deviation. The formulas for calculating global and piecewise velocity thresholds are designed as follows:

$$v_{gth} = \max(V_{th\min}, \min(V_{th\max}, \bar{v}_g + M_{vg} \times \sigma_{vg})) \quad (1)$$

$$v_{pth} = \bar{v}_p + M_{vp} \times \sigma_{vp} \quad (2)$$

Among them, $V_{th\min}$, $V_{th\max}$ are lower and upper bounds of speed threshold respectively. M_{vg} is the global speed multiplier and M_{vl} is the local speed multiplier. \bar{v}_g , σ_{vg} are the mean and standard deviation of the global velocity obtained from all points on the trajectory, \bar{v}_p , σ_{vp} are the mean and standard deviation of the local velocity obtained from the interior points of the segment. Because of the influence of outliers, the mean value calculated by global information may far exceed the target's motion ability, the upper and lower bounds are added to ensure the validity of the global threshold.

Differential threshold is to use the differential value of a discrete data set to judge the change rule of the data, so as to judge the point of abnormal change. For continuous function, the first-order differential can be obtained by formula (3) at the two adjacent points.

$$d_i = (v_i - v_{i-1}) / (t_i - t_{i-1}) \quad (3)$$

In the formula, d_i is the differential value of discrete AIS data, v is the speed and t is the time after digitization. Sometimes when the speed threshold is satisfied, the differential threshold will not be satisfied because of sampling noise and data transmission, so it is necessary to satisfy the two threshold conditions at the same time to be judged as the normal point. The formulas for calculating the differential threshold are as follows:

$$d_{gth} = \max(A_{th\min}, \min(A_{th\max}, \bar{d}_g + M_{dg} \times \sigma_{dg})) \quad (4)$$

$$d_{pth} = \bar{d}_p + M_{dp} \times \sigma_{dp} \quad (5)$$

Among them, $A_{th\min}$, $A_{th\max}$ are the lower and upper bounds of the differential threshold, M_{dg} is the global difference multiple, M_{dp} is the piecewise difference multiple, σ_{dg} is the mean and standard deviation of the global difference of all points on the trajectory, and \bar{d}_p , σ_{dp} are the mean and standard deviation of the piecewise difference of all points on the trajectory.

3.2 Algorithm Flow

Before a trajectory point is detected, the Limitations of Time Length and points of segmented data should be set. If there exist points of a trajectory whose speed and difference are larger than $v_{gth}, v_{pth}, d_{gth}, d_{pth}$, the detected point would be considered as an outlier. The pseudo code of TODMF can be described as follows:

Algorithm: TODDT

Input: $T_i = \{P_1, P_2, \dots, P_i, \dots, P_n\}, V, A$

Output: O_r

01: Initialize $VT[n], DT[n]$ to store speed threshold and differential threshold, n is the number of points

02: For i=1 to n

03: For j = 1 to n

04: if $7 < j-i < 31$

05: if $20s < t_{P_j} - t_{P_i} < 5min$

06: For each v_{P_i}

07: compute $v_{gth}, v_{pth}, d_{gth}, d_{pth}$

08: if $v_{P_i} > v_{gth}, v_{pth} \ \& \ d_{P_i} > d_{gth}, d_{pth}$

09: add P_i to O_r

10: j++;

11: i++;

4. RESULT

This paper uses a real ship trajectory data set of a port to verify the effectiveness of the algorithm. The data set records the real-time tracks of 165 ships around the port in one month, including more than 3000 tracks and millions of trajectory points. The results of the algorithm implemented by Python are shown in Figure 3.

```

ship:431003062 piecewise speed and difference threshold are :0.729127 0.015633
ship:431003062 piecewise outlier is : 2018/3/1 4:22:20 4.2
ship:431003062 global speed and difference threshold are :31.050578 0.500000
ship:431003062 global outlier is: 2018/4/24 2:12:48 102.3
ship:431003062 global speed and difference threshold are :32.000000 0.500000
ship:431003062 global outlier is: 2018/4/24 2:53:18 102.3
ship:431003062 global speed and difference threshold are :32.000000 0.500000
ship:431003062 global outlier is: 2018/4/24 3:18:20 102.3
ship:431003062 global speed and difference threshold are :32.000000 0.500000
ship:431003062 global outlier is: 2018/4/24 3:24:0 102.3
ship:431003062 global speed and difference threshold are :32.000000 0.500000
ship:431003062 global outlier is: 2018/4/24 3:42:38 102.3
ship:431008952 piecewise speed and difference threshold are :0.933333 0.031851
ship:431008952 piecewise outlier is : 2018/3/1 5:36:44 7.1

```

Figure 3. Original ship trajectory.

Dynamic differential threshold algorithm detects three types of outliers on ship data sets, including isolated outliers, continuous outliers and obvious identifying outliers. The visualization results are shown in Figure 4-6. The outliers are marked by triangles. Figure 4 shows the isolated outliers detected. The algorithm can effectively detect the outliers of velocity fluctuation in stationary trajectory by local threshold. In Figure 5, the algorithm identifies four successive abnormal points, which deviate from the ship's track obviously, and the velocity of the four points differs greatly from that of the other points. Figure 6 shows that the left track is significantly different from the right track. The two tracks are actually two ships, but they are identified as a moving object in the process of data collection, and the anomaly is caused by misidentification.

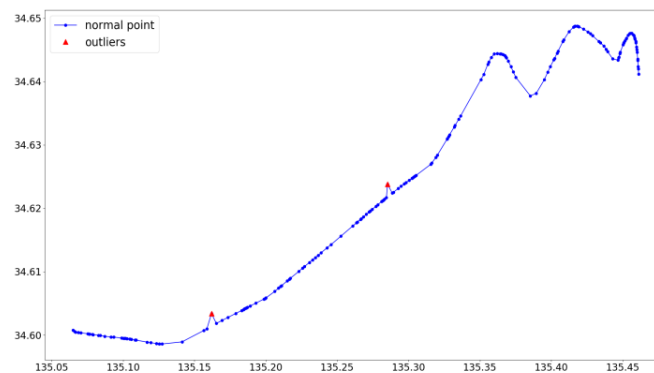


Figure 4. Isolated outliers.

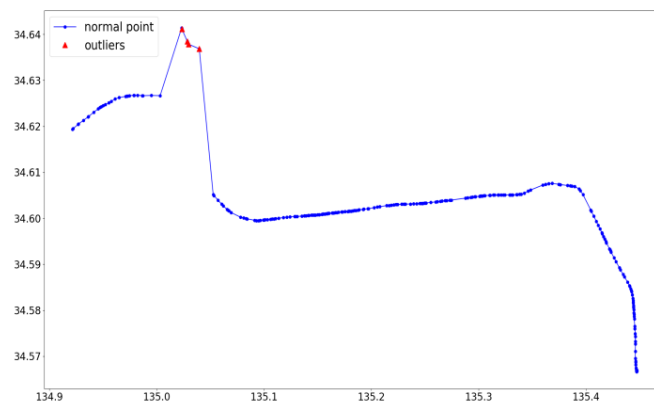


Figure 5. Continuous outliers.

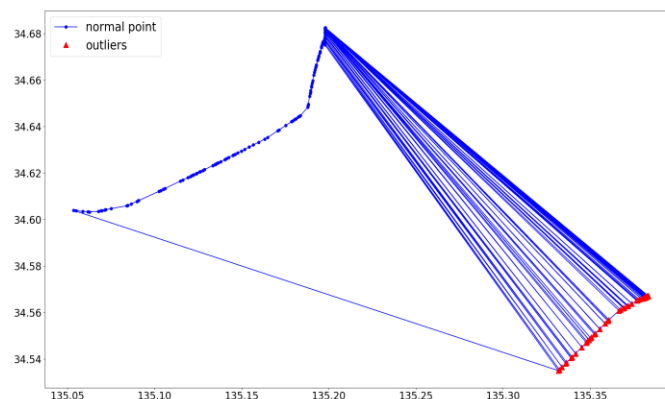


Figure 6. Obvious identifying outliers.

In order to show the effectiveness of TODDT algorithm intuitively, Figure 7 shows the trajectory comparison before detection and after eliminating outliers in Figure 1. It can be clearly seen from the

graph that the detected trajectory is smoother and accords with the ship's motion. The algorithm effectively detects the unreasonable outliers which deviate from the original trajectory. Only after abnormal points are detected can ship trajectory data be used for subsequent processing and knowledge mining.

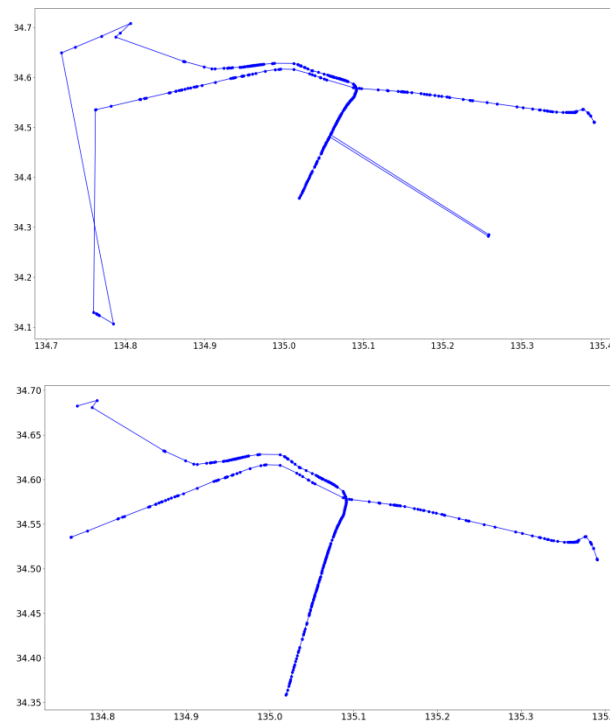


Figure 7. Trajectory comparison after eliminating outliers.

The experimental results show that the TODDT algorithm detects 763 outliers, while the constant velocity threshold method detects 633 outliers, including some erroneous normal points. TODDT algorithm can get dynamic thresholds in real time according to different data. Compared with the constant velocity threshold method, it can get more accurate detection results. It is very suitable for real-time detection of outliers in ship trajectory data.

5. SUMMARY

A trajectory outlier detection algorithm TODDT based on dynamic differential threshold is proposed in this paper. According to the data of ship's whole trajectory and trajectory segment, the dynamic threshold can be obtained by using the segment threshold, difference and conditional rules, and the real-time anomaly detection of ship's trajectory data can be realized. The experimental results show that the proposed algorithm can effectively detect unreasonable trajectory points which deviate from the trajectory and have large velocity fluctuations, and greatly improve the quality of the trajectory data, thus providing guarantee for subsequent data processing and knowledge discovery. In future, on the basis of this algorithm, the features of the trajectory data will be extracted, and the machine learning classification method will be introduced to realize the automatic classification of whether the ship data points are abnormal or not.

ACKNOWLEDGMENTS

The authors are very grateful to the editors and reviewers for their valuable comments and suggestions. This work was supported by the National Social Science Foundation of China (Grant No. L15CGL031) and the Fundamental Research Funds for the Central Universities (No.3132019353).

REFERENCES

- [1] Harati-Mokhtari A, Wall A, Brooks P, et al. Automatic Identification System (AIS): data reliability and human error implications. *The Journal of Navigation*, 2007, 60(3): 373-389.
- [2] Mou J M, Van Der Tak C, Ligteringen H. Study on collision avoidance in busy waterways by using AIS data. *Ocean Engineering*, 2010, 37(5-6): 483-490.
- [3] Hansen M G, Jensen T K, Lehn-Schiøler T, et al. Empirical ship domain based on AIS data. *The Journal of Navigation*, 2013, 66(6): 931-940.
- [4] Xiao F, Ligteringen H, Van Gulijk C, et al. Comparison study on AIS data of ship traffic behavior. *Ocean Engineering*, 2015, 95: 84-93.
- [5] Yu Q, Luo Y, Chen C and Wang X. 2018. Trajectory outlier detection approach based on common slices sub-sequence. *Applied Intelligence*, 48,9(2018):2661-2680.
- [6] Han J. *Data Mining: Concepts and Techniques*. 2005. Morgan Kaufmann Publishers Inc.
- [7] Liu Z, Pi D, Jiang J. Density-based trajectory outlier detection algorithm. *Journal of Systems Engineering and Electronics*, 2013, 24(2): 335-340.
- [8] Zheng Y. Trajectory data mining: an overview. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2015, 6(3): 29.
- [9] Wiegand P, Pell R, Comas E. Simultaneous variable selection and outlier detection using a robust genetic algorithm. *Chemometrics and Intelligent Laboratory Systems*, 2009, 98(2): 108-114.
- [10] Yuan G, Xia S, Zhang L, et al. Trajectory outlier detection algorithm based on structural features. *Journal of Computational Information Systems*, 2011, 7(11): 4137-4144.
- [11] Knorr E M, Ng R T, Tucakov V. Distance-based outliers: algorithms and applications. *The VLDB Journal—The International Journal on Very Large Data Bases*, 2000, 8(3-4): 237-253.
- [12] Zhao-Rong H, Guang-Luan X U, Ting-Lei H, et al. Vessel Trajectory Outlier Detection Algorithm Based on Adaptive Threshold[J]. *Computer and Modernization*, 2018.