Journal of *Marine Science and Engineering*

MDPI

Article

An Anomaly Detection Method for AIS Trajectory Based on Kinematic Interpolation

Shaoqing Guo [1,2], Junmin Mou [1,2], Linying Chen [1,2] and Pengfei Chen [1,2,*]

[1] School of Navigation, Wuhan University of Technology, Wuhan 430063, China; gsqnbsr@whut.edu.cn (S.G.); moujm@whut.edu.cn (J.M.); linyingchen@whut.edu.cn (L.C.)
[2] Hubei Key Laboratory of Inland Shipping Technology, Wuhan University of Technology, Wuhan 430063, China
* Correspondence: chenpf@whut.edu.cn

Abstract: With the enormous amount of information provided by the ship Automatic Identification System (AIS), AIS is now playing a significant role in maritime transport system-related research and development. Many kinds of research and industrial applications are based on the ship trajectory extracted from raw AIS data. However, due to the issues of equipment, the transmission environment, and human factors, the raw AIS data inevitably contain abnormal messages, which have hindered the utilization of such information in practice. Thus, in this paper, an anomaly detection method that focuses on AIS trajectory is proposed, making comprehensive use of the kinematic information of the ship in the AIS data. The method employs three steps to obtain non-error AIS trajectories: (1) data preprocessing, (2) kinematic estimation, and (3) error clustering. It should be noted that steps (2) and (3) are involved in an iterative process to determine all of the abnormal data. A case study is then conducted to test the proposed method on real-world AIS data, followed by a comparison between the proposed method and the rule-based anomaly detection method. As the processed trajectories show fewer abnormal features, the results indicate that the method improves performance and can accurately detect as much abnormal data as possible.

Keywords: AIS; ship trajectory; anomaly detection; kinematic interpolation; clustering

Citation: Guo, S.; Mou, J.; Chen, L.; Chen, P. An Anomaly Detection Method for AIS Trajectory Based on Kinematic Interpolation. *J. Mar. Sci. Eng.* **2021**, *9*, 609. https://doi.org/10.3390/jmse9060609

Academic Editor: Alessandro Ridolfi

Received: 30 April 2021
Accepted: 30 May 2021
Published: 1 June 2021

1. Introduction

Recently, the demand for AIS (Automatic Identification System) data for research and development in the maritime transport discipline has been continuously increasing [1]. AIS enables ships to communicate with each other or contact satellites and base stations using various information such as a Maritime Mobile Identification Number (MMSI), longitude, latitude, speed over ground (SOG), course over ground (COG), headings, etc. [2]. Researchers have found that AIS can function as a big data source not only for maritime safety but also for other kinds of research, such as traffic analysis, transport economy, emissions, etc. [3].

Although AIS has played a significant role in maritime transport-related research with its enormous amount of data, data error is likely to occur due to various factors [4,5]. Those incorrect data lead to inaccurate conclusions in trajectory analysis, which is crucial for further applications. Therefore, it is of great concern to detect and remove those abnormal data to improve the quality of AIS data.

To date, researchers have made great efforts to identify abnormal data in AIS through different methods. According to [6,7], the abnormal data detection methods for ship trajectory can be divided into two types: knowledge-driven and data-driven. In general, those methods that correspond to the knowledge-driven approaches can be regarded as rule-based methods. The simplest way to conduct anomaly detection for AIS data is to use a predefined data range to determine and exclude the outliers [8], which is efficient but has relatively poor performance in terms of its accuracy and reliability. In [9], the authors

considered the geometric shape of a ship trajectory and proposed a vector-based method to detect anomalies. Besides, since AIS data can describe the kinematic characteristic of a ship, based on the nearly-constant velocity (NCV), an Ornstein–Uhlenbeck model has been proposed to detect whether a ship deviates from a planned route [10]. Apart from the single factors such as the shape characteristic of ship trajectories, factors such as position, speed, acceleration, rate of turn, etc. can be used to discover abnormal data [11,12] following some pre-defined criteria for the anomaly. However, the drawbacks of this kind of approach are obvious: the definition of the rules relies on human knowledge and the characteristics of the regional data, and thus it is difficult for this approach to provide a generalized anomaly detection method.

Different from the knowledge-driven approaches, the data-driven approach focuses on learning the ship behaviors from the trajectory data to generate motion patterns. Those behaviors that differ from the extracted patterns are considered anomalies, the associated data will then be detected. Based on the way to extract motion patterns, the data-driven approaches can be further divided into three kinds: Similarity-Based, Supervised-Learning-Based, and Unsupervised-Learning-Based. For Similarity-based methods, the similarity between the trajectories is applied as an alternative to determining the anomaly of the trajectory by comparing it with all labeled trajectories. A common challenge in similarity calculation is the unequal length of different trajectories, and thus a varies of methods have been proposed to overcome the issue [13]. A method based on Hausdorff distance is proposed by [14] to compare the similarity of multi-dimensional trajectories and detect the abnormal in them. Reference [15] proposed a method of asynchronous trajectory matching based on piecewise space-time constraints (PTSCTM) to reconstruct and discover the anomalies of ships, in which the Euclidean distance and time distance are used to find similar trajectories points.

However, when the scale of the dataset increases, using similarity becomes impractical as all the labeled trajectories have to be considered whenever a new trajectory is included. With the rapid development of artificial intelligence, machine learning is widely used to analyze and learn patterns from data. The supervised-learning-based method learns the mapping relationship between trajectory data and motion patterns that have been utilized in the anomaly detection of a ship's AIS trajectory. A model using the Received Signal Strength Indicator (RSSI) and One-Class Support Vector Machine (SVM) is proposed in [16] to detect anomalies in AIS and further identify intentional AIS on-off transitions. In [17], the authors used hierarchical and k-medoids to learn typical ship navigation patterns and adopted the Naive Bayes classifier to detect anomalous ship behavior.

Different from supervised learning, in many situations, when given a set of inputs, the output is not specifically defined. When the motion patterns of the trajectory are not well defined, the Unsupervised-Learning-Based approaches play an important role. Reference [18] the combined topic model with a generic algorithm to calculate the anomaly probability of a new trajectory. An unsupervised model called Traffic Route Extraction and Anomaly Detection (TREAD) is establish in [19] to automatically learn maritime traffic patterns. To reduce the training time, the water area is partitioned in [20] to establish a training framework based on Adaptive Kernel Density Estimation (AKDE). The combination of Supervised-Learning-Based and Unsupervised-Learning-Based methods results in a hybrid approach. This complementary method is often used as a predictor to discover anomalies. Reference [21] applied Density-Based Spatial Clustering of Applications with Noise (DBSCAN) to obtain traffic patterns which will be then used to train a Recurrent Neural Network (RNN) to predict the trajectory for anomaly detection.

In general, the data-driven approach is simple to conduct as it does not require much knowledge from experts to build the model. However, these methods need a large amount of training data to establish a sophisticated model, and the quality of the training data directly influences the performance of the anomaly detection. To overcome the drawbacks of knowledge-driven and data-driven approaches and to establish a model that can efficiently identify the abnormal data in AIS data sets, this paper provides another

perspective on anomaly detection; i.e. identifying anomalies of ship trajectory data based on the kinematic characteristics of the ship—its speed and course information. By comparing the estimated data according to the motion characteristics of the ship and the original data, the anomalies in the trajectory data can be identified using the clustering technique. Compared with the existing methods, the main contributions of this work are as follows: (1) we make full use of longitude, latitude, SOG, and COG information in AIS to estimate the reference AIS data based on the kinematic characteristics of the ship; (2) we introduce a clustering method to identify abnormal data by comparing the original data and the kinematic-based estimation. With this design, the proposed method can be of great practical value as it does not require a large amount of expert experience or training data with high quality. It should be noted that, although missing data also correspond to abnormal data, as they do not exist in the obtained dataset and can be restored by other reconstruction methods [22], they are not considered in this paper. With the improved effectiveness of the anomaly detection of AIS data, the proposed method would facilitate AIS-based research and development in the maritime transportation industry, such as for maritime traffic management, Maritime Autonomous Surface Ship (MASS), autonomous collision avoidance, etc.

The rest of this paper is structured as follows: Section 2 presents the methodology used in the research, followed by details of the designed models in Section 3. A case study including several types of trajectories is presented in Section 4, together with further discussion on the performance of the proposed method, with a comparison between the rule-based anomaly detection approach. Section 5 concludes the research and presents possible directions for improvement.

2. Methodology

The raw AIS data contain different types of abnormal data due to various influences such as the capacity of the regional system and malfunctions in equipment such as the Global Navigation Satellite System (GNSS), etc. These anomalous data points can be found in the AIS data in the form of information that is significantly different from the characteristics of the data sets. In this section, we illustrate the main methods used in data preprocessing and anomaly detection. The whole framework is shown in Figure 1.

2.1. Data Preprocessing

The raw AIS data consist of chaotic messages from different ships with both dynamic data and static data. To extract trajectory data, we first decode the raw AIS data according to the regulation in [23]. A further step is applied in this research to eliminate the messages with obvious illegal values, such as an incorrect MMSI number, obviously incorrect longitude or latitude value, invalid kinematic parameters, etc. Tables 1 and 2 show examples of decoded AIS data from a trajectory dataset. Please note that, in [23], the SOG with, a value of 102.3, and the true heading, with a value of 511, indicate they are not available. If a message only has a value of 511 for true heading but normal values for other fields, the message will not be flagged as abnormal as it can still provide sufficient research value.

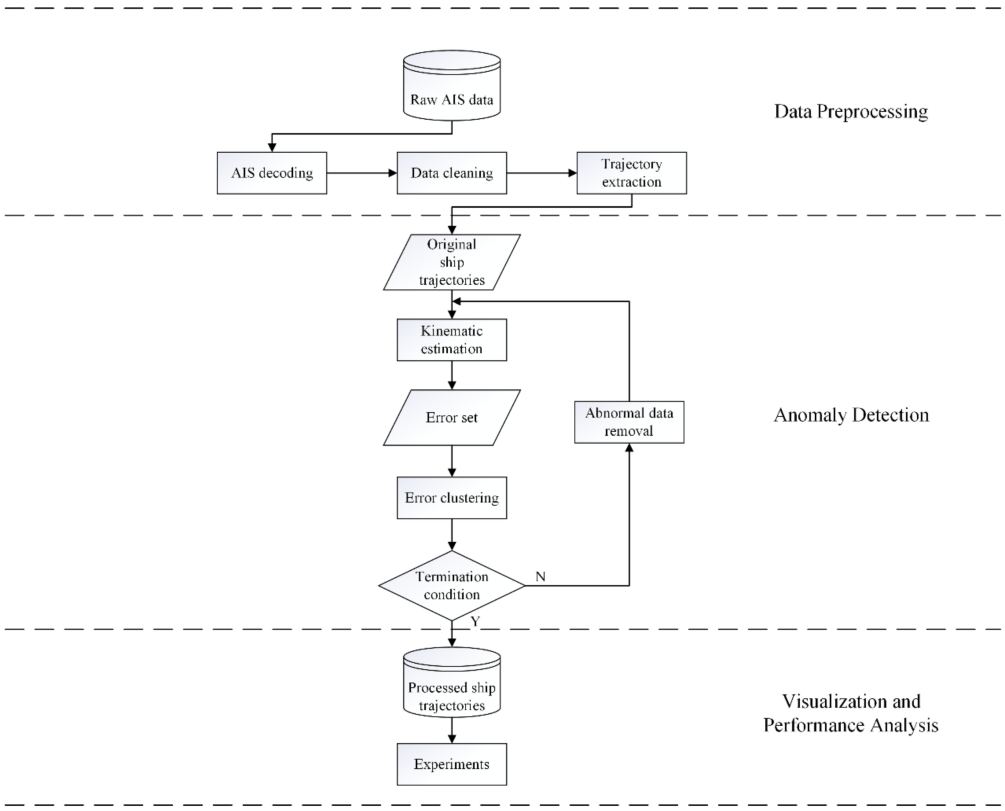


Figure 1. Research methodology.

Table 1. Example of correct trajectory data.

MMSI	Timestamp (UTC)	Longitude (°)	Latitude (°)	SOG (Kt)	COG (°)	True Heading (°)
412XXX410	1539561613	121.9508	30.4141	4.6	59.0	70
412XXX410	1539585463	121.9511	30.4142	4.6	59.5	71
412XXX410	1539585473	121.9513	30.4143	4.6	60.0	70
412XXX410	1539585492	121.9517	30.4145	4.6	59.9	68
412XXX410	1539585503	121.9519	30.4146	4.6	59.1	68
412XXX410	1539585533	121.9526	30.4150	4.6	58.0	71
412XXX410	1539585533	121.9530	30.4152	4.6	59.0	72

Table 2. Example of eliminated trajectory data (outliers).

MMSI	Timestamp (UTC)	Longitude (°)	Latitude (°)	SOG (Kt)	COG (°)	True Heading (°)
123456	1539562037	181.0	91.0	102.3	360.0	313
413XXX180	1539620379	181.0	91.0	102.3	360.0	511
413XXX460	1539620396	122.2476	29.1889	102.3	360.0	209

2.2. Anomaly Detection

According to [24], abnormal data can be determined to be data that deviate significantly from other members of the sample. In this research, a trajectory point is an abnormal point when it disobeys the motion trend of the ship; i.e., a point is determined to be abnormal if the kinematic information of the point does not follow the moving trend of the ship. In the literature, the position is one of the frequently utilized factors to determine an anomaly in AIS data [25,26]. In practice, the trajectory is accompanied by varied kinematic information such as speed and direction. For a data point that is normal in terms of position features, the speed and direction information might be abnormal. For this kind

of abnormal data, it would be insufficient to determine an anomaly simply based on the position information.

In the data interpolation research, to overcome the lack of kinematic information used in the data interpolation process, a concise and effective method called kinematic interpolation has been presented in [27]. By establishing an acceleration function of the moving object in the considered trajectory segment, the location and speed of this object can be calculated at any time using the following kinematic equation:

$$\begin{aligned} x(t) &= x(t_i) + \int_{t_i}^{t} v(t)dt, \\ v(t) &= v(t_i) + \int_{t_i}^{t} a(t)dt. \end{aligned} \tag{1}$$

Given two adjacent points p and q, t_i and t_j denote their timestamps. $x(t)$ and $v(t)$ are the moving object's position and speed at time t ($t_i < t < t_j$) with acceleration $a(t)$.

Although we proposed the estimation of the possible location of an unknown point, we have found that kinematic interpolation can also be used to perform anomaly detection by analyzing the ship trajectory data while considering the moving trend of the ship. Figure 2 shows the principle of using kinematic interpolation to estimate an unknown point. An estimation point can be obtained with an interpolation process considering the kinematic information of the ship. If the data point follows the moving trend, the error between the estimated point and original data should be small and vice versa. With this principle, a kinematic-based data anomaly detection method is proposed in our work.

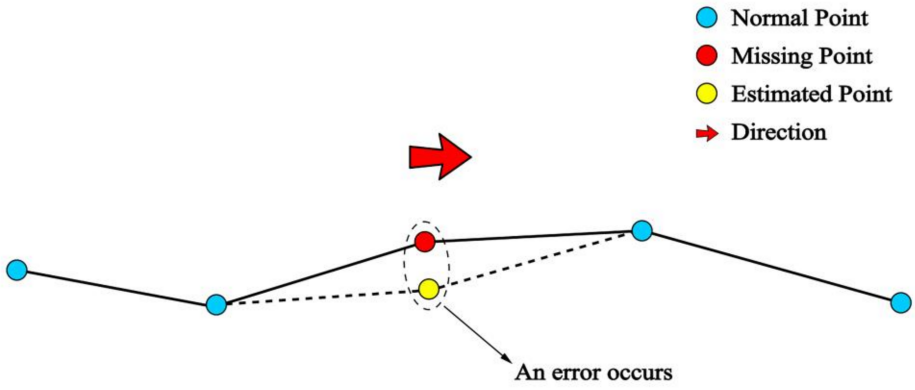


Figure 2. Kinematic interpolation.

Forming part of the widely used unsupervised anomaly detection techniques, clustering-based methods can identify anomalies without prior knowledge and are suitable for multiple data types [28]. One of the most popular clustering methods, K-means clustering, has been applied to detect abnormal data [29]. The procedure of K-means is as follows: (1) select initial clustering centers for a given number of clusters, (2) divide each cluster object into the cluster with the minimum distance to the cluster center, (3) update the cluster centers, and (2) repeat until the cluster centers no longer change. However, drawbacks of K-means are obvious [30]: (i) the number of clusters should be determined first, (ii) there is no efficient and universal method for identifying the initial partitions, and (iii) the method is sensitive to noise and outliers. With the simplicity and effectiveness of K-means clustering in anomaly detection, in this research, we proposed a modified K-means clustering technique as a procedure for identifying abnormal trajectory data with the integration of a kinematic-based approach. The details of the model are presented in Section 3.

3. Model Design

The objective of this research is to propose a new anomaly detection method for ship trajectory data from the perspective of the kinematic characteristics of ships. To achieve this objective, three major components need to be developed, which are as follows: (1) a

kinematic estimation model for AIS data, (2) an anomaly detection model based on the kinematic estimation model, and (3) an iterative detection model based on loop detection with a termination condition. This section describes the models utilized in this method in detail.

3.1. Kinematic Estimation

The kinematic estimation of AIS data is conducted to estimate the data points of a ship trajectory according to their kinematic characteristics, such as velocity and course, etc. The objective of this procedure is to provide a reference point to determine if the trajectory data follow the kinematic characteristics of the ship. To conduct this operation, the definition of the trajectory should be first introduced.

For the ship trajectory data set T, the definition for the i_{th} trajectory Tr_i in T is described as follows: $Tr_i = (MMSI_i, Pts_i)$, where $MMSI_i$ indicates the name of the ship to which Tr_i belongs, and Pts_i is the trajectory points set of the trajectory Tr_i. For the j_{th} point $p_{i,j}$ in Pts_i, $p_{i,j} = (t_{i,j}, lon_{i,j}, lat_{i,j}, sog_{i,j}, cog_{i,j})$, where $t_{i,j}, lon_{i,j}, lat_{i,j}, sog_{i,j}, cog_{i,j}$ represent the timestamp, longitude, latitude, SOG, and COG of $p_{i,j}$, respectively. $|Pts_i|$ denotes the number of points in Pts_i. The details of the anomaly detection of the AIS trajectory are elaborated in the following sections.

To attain our objective, we introduced a sliding window method to the trajectory to calculate each estimated point using the kinematic interpolation method. The principle of the method can be seen in Figure 3. The size of the window was set to 3 (w = 3), containing two endpoints and a mid-point.

As the window slides, an estimated point corresponding to each trajectory point is produced with the kinematic interpolation method. Please note that in Equation (1), we mainly use three kinds of kinematic information—position, velocity, and acceleration—at certain times. However, AIS trajectory data do not contain acceleration information. The solution to this is to introduce the linear motion characteristic suggested in [27] based on the frequency characteristics of ship AIS trajectory data, which is shown in Equation (2):

$$a(t) = b + m(t - t_i) \tag{2}$$

where $a(t)$ is the acceleration of a ship at time t; t is the time of the estimated point; t_i is the time of the start point; and b and m are two parameters to be determined with the data. It should be noted that Equations (1) and (2) only consider the moving object in one dimension. For a trajectory point $p_{i,j}$, its positions in longitude and latitude directions are known as $lon_{i,j}$ and $lat_{i,j}$, and the velocity can also be determined based on Equation (3):

$$\begin{aligned} v_{i.j}^{lon} &= sog_{i,j} \times \sin(cog_{i,j}) \times \tfrac{1852}{3600} \\ v_{i.j}^{lat} &= sog_{i,j} \times \cos(cog_{i,j}) \times \tfrac{1852}{3600} \end{aligned} \tag{3}$$

To simplify the following description, we only illustrate the formulas in one dimension; the other dimension can be determined in the same manner. Equation (1) can be rewritten as Equation (4) with the integration of Equation (2):

$$\begin{aligned} x(t) &= x(t_i) + v(t_i)(t - t_i) + \tfrac{1}{2}b(t - t_i)^2 + \tfrac{1}{6}m(t - t_i)^3 \\ v(t) &= v(t_i) + b(t - t_i) + \tfrac{1}{2}m(t - t_i)^2 \end{aligned} \tag{4}$$

If the start point and end point are $p_{i,p}$ and $p_{i,q}$, the parameters b and m can then be solved by substituting their positions and velocities into Equation (4), which is shown in Equation (5):

$$\begin{aligned} b &= \frac{6(x(t_{i,q}) - x(t_{i,p}))}{(t_{i,q} - t_{i,p})^2} - \frac{2(2v(t_{i,p}) + v(t_{i,q}))}{(t_{i,q} - t_{i,p})} \\ m &= \frac{6(v(t_{i,p}) + v(t_{i,q}))}{(t_{i,q} - t_{i,p})^2} - \frac{12(x(t_{i,q}) - x(t_{i,p}))}{(t_{i,q} - t_{i,p})^3} \end{aligned} \tag{5}$$

Once b and m are solved, the object's status at any given time $t(t_{i,p} < t < t_{i,q})$ can be estimated by Equation (4).

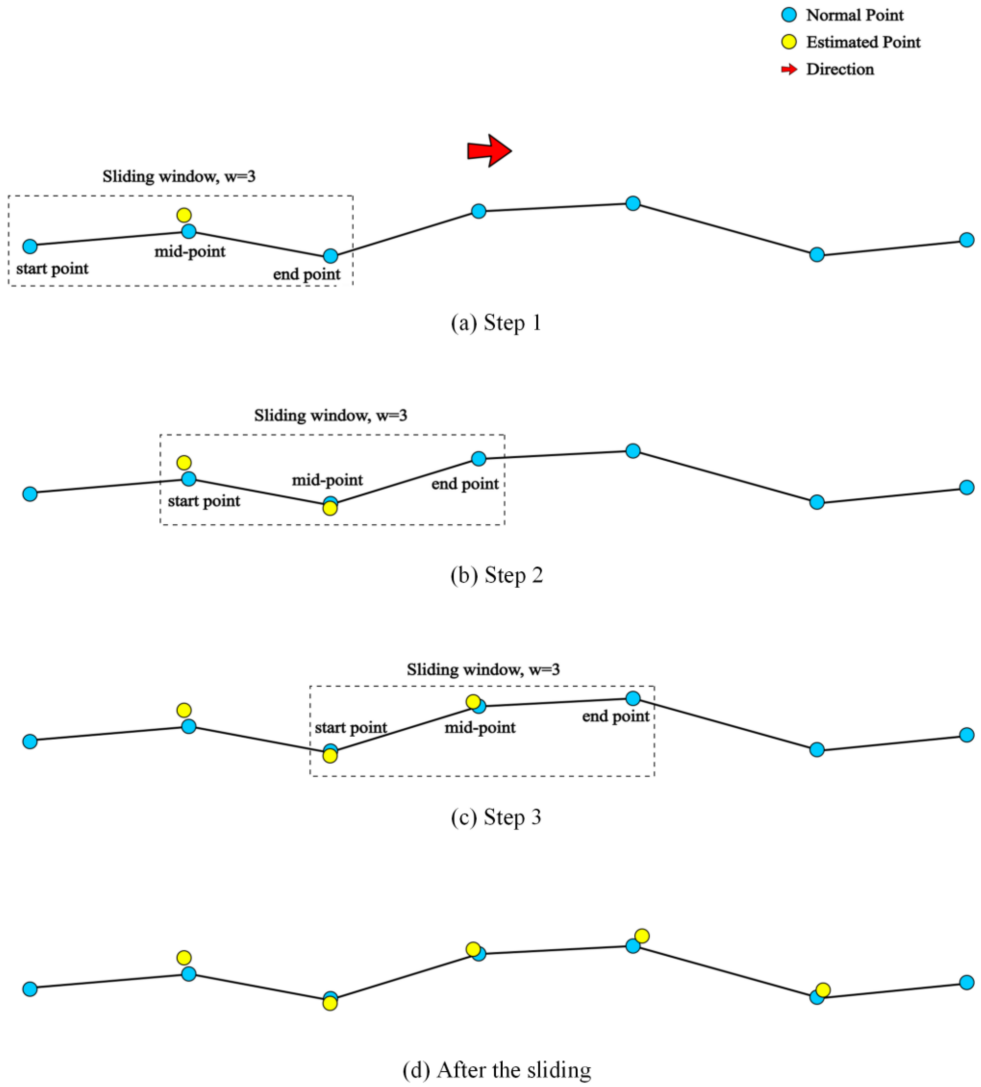


Figure 3. Principle of kinematic estimation for anomaly detection.

3.2. Anomaly Detection of AIS Data based on Error Clustering

As mentioned above, the principle of this research is to identify an anomaly in ship AIS data considering the kinematic characteristics of the data. To achieve this objective, the kinematic estimation is first applied to provide a reference for the verification of the data, based on which the error between the estimated data points and AIS data are analyzed to determine the anomaly in the data sets. The details are presented below.

3.2.1. Error Calculation and Error Weight

When the sliding window process is complete, one can see from Figure 3d that, except for two endpoints, each datapoint has a corresponding estimated point. If there is no error in the trajectory data, the estimated point should be close to the known points, but when there is any error in position, velocity, or direction, the estimated point will be far away from the known point. The principle of the error estimation process for the trajectory points is shown in Figure 4.

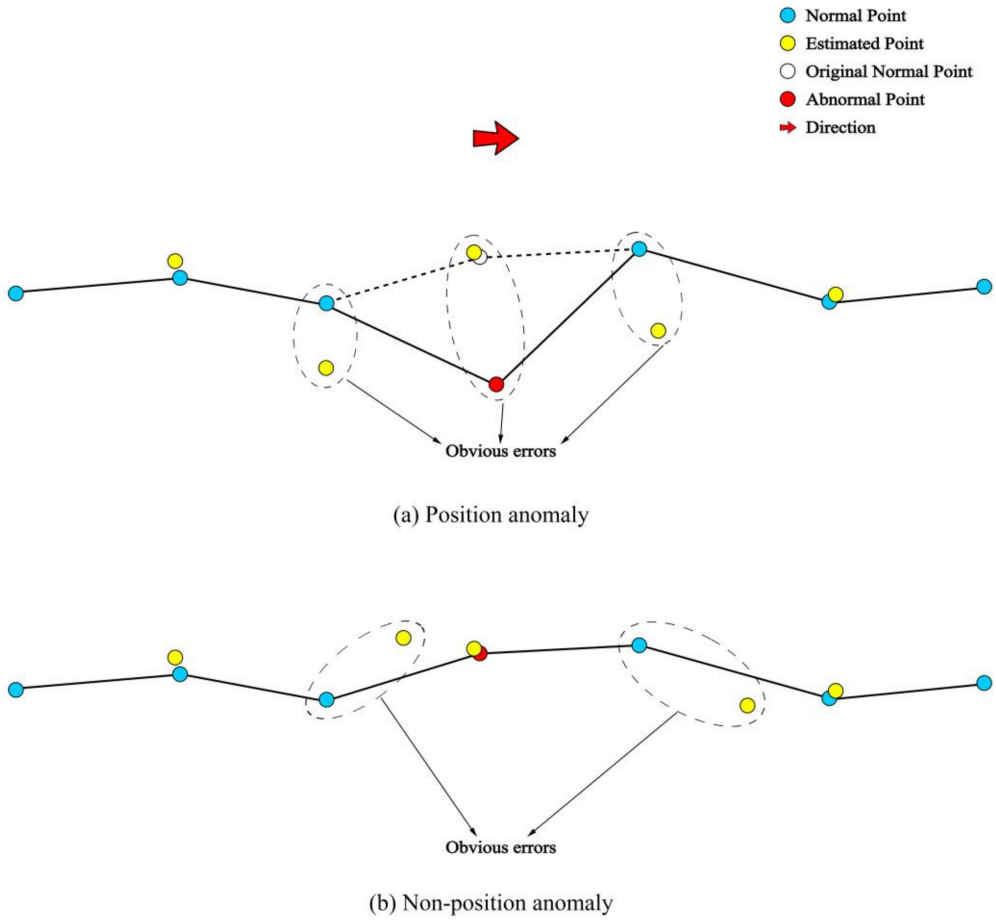


Figure 4. Continuous error estimation and determination for a ship trajectory.

In Figure 4a, one can see that when a position anomaly occurs, three continuous estimated points show obvious errors. As shown in Figure 4b, the abnormal data containing velocity errors are difficult to identify from the positioning perspective, as their position information is correct. To identify this anomaly in its velocity perspective, the error between a trajectory point $p_{i,j}$ and its accompanied estimated point $e_{i,j}$ should be estimated with the integration of velocity information following Equation (6):

$$\begin{aligned} dp_{i,j} &= dist(p_{i,j}, e_{i,j}) \\ dv_{i,j} &= \left| v_{i,j} - v'_{i,j} \right| \end{aligned} \tag{6}$$

where $dist(p_{i,j}, e_{i,j})$ denotes the distance between $p_{i,j}$ and $e_{i,j}$ using the Mercator projection method, and $dv_{i,j}$ denotes their velocity difference.

Since $p_{i,j}$ can be used up to three times to calculate the relevant estimated points during the sliding window process, we introduce an error set $Er_{i,j}$ to denote relevant errors of $p_{i,j}$.

$$Er_{i,j} = \left\{ (dp_{i,k}, dv_{i,k}) \middle| k \in N^*, j-1 \le k \le j+1, k \le |Pts_i| \right\} \tag{7}$$

with this design, the error weight $w_{i,j}$ for the data point $p_{i,j}$ can be calculated as shown in Equation (8):

$$w_{i,j} = \left(\frac{\sum\limits_k dp_{i,k}}{|Er_{i,j}|}, \frac{\sum\limits_k dv_{i,k}}{|Er_{i,j}|} \right) \tag{8}$$

where k has the same setting as that in $Er_{i,j}$, and $|Er_{i,j}|$ is the number of elements in $Er_{i,j}$, which indicates how many times $p_{i,j}$ has been used to calculate the errors. Thus, each point has an error weight value. The error weight set can be denoted as Equation (9) shows:

$$W_i = \{w_{i,j} | j \in N^*, j \leq |Pts_i|\} \tag{9}$$

As mentioned above, the errors in the position and velocity of the AIS data are all considered in the anomaly detection model. Using this design, $w_{i,j}$ contains two kinds of values, which is the errors in position and velocity. Their dimensions are different, which are m and m/s. To obtain an accurate result for anomaly detection, first, we need to eliminate the influences of dimension and order of magnitude. Therefore, a standardization process is applied to W_i:

$$\begin{aligned} ws_{i,j} &= \frac{w_{i,j} - \overline{W_i}}{\sigma} \\ Ws_i &= \{ws_{i,j} | j \in N^*, j \leq |Pts_i|\} \end{aligned} \tag{10}$$

where $\overline{W_i}$ and σ are the mean value and standard deviation of W_i, $ws_{i,j}$ is the standardization result of $w_{i,j}$, and Ws_i is the standardization weight set.

3.2.2. K-Means Clustering and Anomaly Detection

In the prevision section, an error estimation method was proposed based on the kinematic information of ship AIS data to provide a reference for the anomaly detection from the ship kinematic perspective. The next step is to propose a method to identify which data points are abnormal based on their error estimation. As a widely used clustering method, K-means can detect abnormal data effectively. In this section, we illustrate the details of utilizing an improved K-means clustering approach to identify an anomaly in an AIS data set.

First, for the number of clusters, a trajectory point $p_{i,j}$ can be used up to three times in the sliding window process to obtain the error weights of $p_{i,j}$. In this case, there are four possibilities when determining an anomaly in the data: (1) no anomaly occurs, (2) an anomaly occurs once in the error weights, (3) two anomalies occur in the error weights, and (4) three anomalies occur in all three iterations of the error calculations. Thus, the number of clusters can be determined as 4.

For the clustering of the error weights of the AIS data points, in this research, we adopt an improved method called K-means++ [31]. The principle of K-means++ is to ensure the distance between the initial clustering centers is as far as possible, which can reduce the influence of the selection of the initial clustering center on clustering results. Finally, the noise and outliers can be automatically detected with the proposed clustering method. A short example of this method is shown in the following section.

Figure 5a shows an illustration of Ws_i for a randomly chosen trajectory. By adopting the K-means clustering method on Ws_i, points in Figure 5a can be divided into four clusters, as shown in Figure 5b. If a point $p_{i,j}$ is not abnormal, then $ws_{i,j}$ should be as close to (0, 0) as possible in the coordinate system; otherwise, the further it is from (0, 0), the higher the possibility that $p_{i,j}$ could be an anomaly. One can see in Figure 5b that the red diamond points are far away from (0, 0); points with these standardization weight values have been identified as abnormal data, indicating that each of them has anomalies in all three calculations in the kinematic estimation. In this way, the most likely abnormal data can be detected, and the point set relevant to the farthest clusters to (0, 0) is defined as $A_{cluster}$.

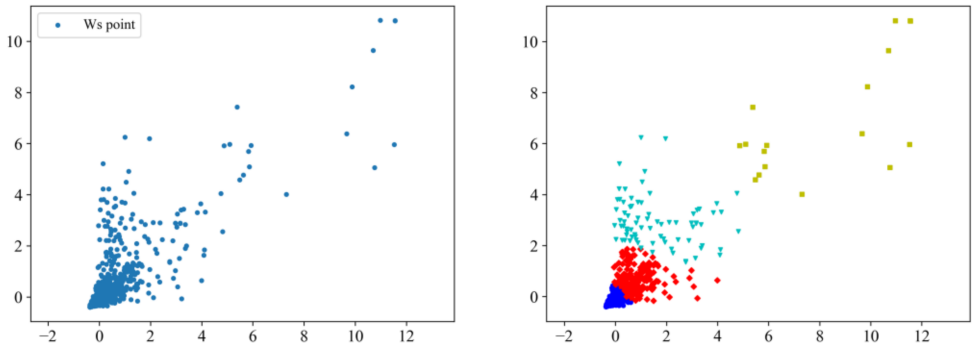


Figure 5. Example of clusters of standardization weight set.

3.3. Loop Detection and Termination Condition

Since an improved K-means clustering method is applied on Ws_i, it is possible to identify the most abnormal data in a trajectory. However, some hidden anomalies might be omitted from the detection. To detect them all, a further step is necessary. Here, we introduce a loop detection process, which is implemented to detect all abnormal points in trajectory data by the repetition of the kinematic estimation and error clustering process.

Within the data set, $w_{i,j}$ can be influenced by the neighbors of $p_{i,j}$, as $p_{i,j}$ can be normal but its neighbors abnormal. For the detection process, an abnormal data set $A_{cluster}$ is established with the clustering process. The data points in this cluster are removed in one iteration of the error clustering process. Then, the point set of Tr_i changes to ${Pts_i}^l$. Let $Pts_i^0 = Pts_i$, which indicates the initial number of points in the original trajectory data set; the definition of ${Pts_i}^l$ is shown in Equation (11):

$$ {Pts_i}^l = {Pts_i}^{l-1} - {A_{cluster}}^l \tag{11} $$

where l is the current number of loops, ${Pts_i}^{l-1}$ is the trajectory point set before the current loop's start, ${A_{cluster}}^l$ is the abnormal point set detected in the current loop, and ${Pts_i}^l$ is the trajectory point set after the current loop. When the loop detection is terminated, let $Pts_i = Pts_i^l$, which indicates the final number of points after the anomaly detection process.

The termination condition of this process is set as follows: Considering the final state of the anomaly detection, the trajectory Tr_i should have no abnormal data in Pts_i. From this perspective, we utilize the performance of the clustering as the criteria to determine when the loop can be terminated; i.e., if the clustering shows good performance with only one cluster, that would mean the abnormal data have been identified and removed from the trajectory. To describe the performance of our approach, we introduce the silhouette coefficient, which was first proposed in [32]. The silhouette coefficient is defined as follows:

$$ \begin{aligned} S(p_{i,j}) &= \frac{D_{nearest\_cluster}(p_{i,j}) - D_{own\_cluster}(p_{i,j})}{\max\left\{D_{own\_cluster}(p_{i,j}), D_{nearest\_cluster}(p_{i,j})\right\}} \\ S &= \frac{\sum_{j=1}^{|Pts_i^l|} S(p_{i,j})}{\left|Pts_i^l\right|} \end{aligned} \tag{12} $$

where $D_{own\_cluster}(p_{i,j})$ is the average distance between $p_{i,j}$ and all other points in the cluster to which $p_{i,j}$ belongs. $D_{nearest\_cluster}(p_{i,j})$ is the average distance between $p_{i,j}$ and all points in the nearest cluster. To make this easy to understand, $D_{own\_cluster}(p_{i,j})$ and $D_{nearest\_cluster}(p_{i,j})$ can be understood as the inner distance and outer distance, respectively. The average silhouette coefficient S is the mean value of $S(p_{i,j})$. The range of S is set between -1 and 1; i.e., a higher score indicates a better clustering result.

When S is lower than a certain value, denoted by S_c, this means that the clustering result is not good enough and is not appropriate to set the initial number of clusters as 4.

This indicates that those points with the most errors in kinematic estimation have been detected and removed. For the rest of the points, if they have anomalies, the worst situation for a point is that two anomalies occur in two of the three calculations; therefore, the initial number of clusters will change to 3. From this perspective, in a detection loop, when the initial number of clusters is set to be 4, 3, and 2 in turn, if $S < S_c$, this means that the detection should terminate. The determination of the criteria is illustrated in Figure 6.

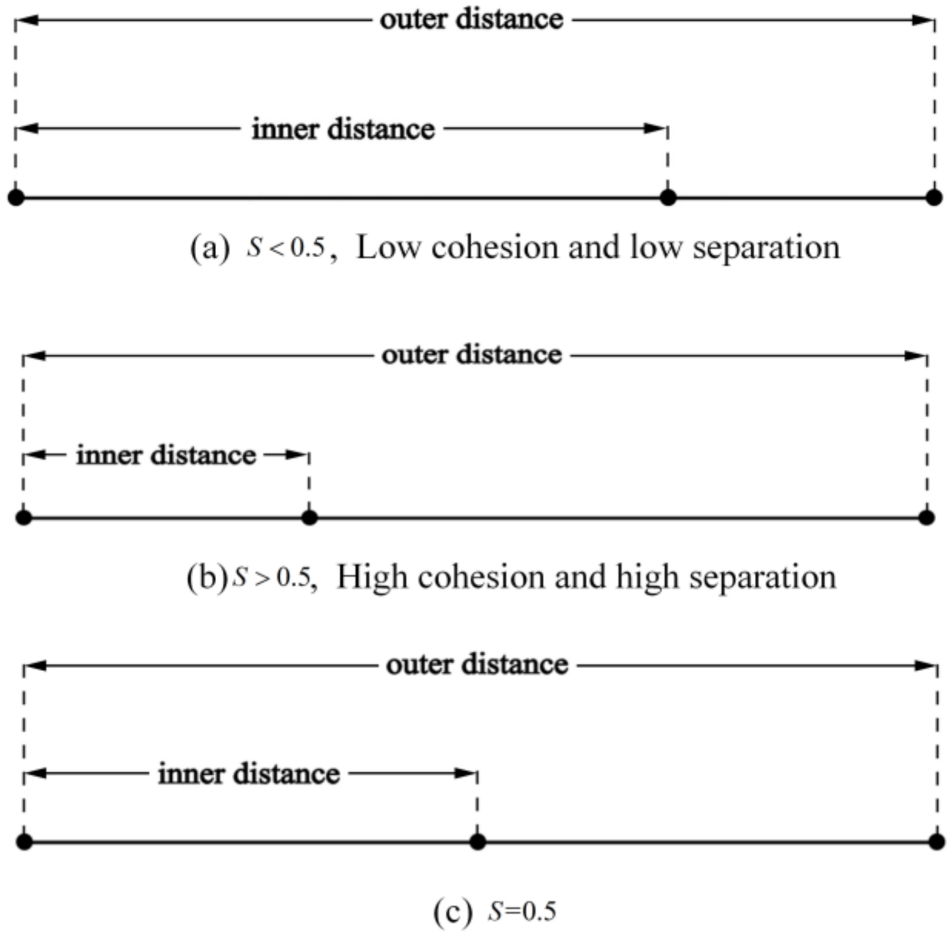


Figure 6. Different values of S.

When S is positive and lower than 0.5, the clustering result can be considered as low cohesion and low separation, as shown in Figure 6a. When S is greater than 0.5, the result is considered to show high cohesion and high separation. Thus, as a critical state, we set S_c to be 0.5. We will not discuss the situation when S is negative, because this would mean that the result is so poor that it has to be re-clustered.

Once the termination condition is set, the anomaly detection model is completed. The pseudocode of the whole anomaly detection model is shown in Algorithm 1, followed by the design of the loop detection model in Algorithm 2.

Algorithm 1. Anomaly Detection.

Inputs : Trajectory point set Pts_i.
Kinematic estimation and error calculation
FOR $p_{i,j}$ IN Pts_i:
    IF $p_{i,j}$ IS NOT endpoint:
        $e_{i,j}$= KI$\left(p_{i,j-1}, p_{i,j+1}\right)$;
        error = Error$\left(p_{i,j}, e_{i,j}\right)$;
        ADD error TO $Er_{i,j}$ AND ITS NEIGHBORS;
FOR EACH $Er_{i,j}$:
    $w_{i,j}$= avg$\left(Er_{i,j}\right)$;
    ADD $w_{i,j}$ TO W_i;
Ws_i= Standardization(W_i);
# Clustering
N = 4;
WHILE N != 1:
    clusters = Kmeans(Ws_i, cluster_num=N);
    IF S >= S_c:
        PICK OUT ${A_{cluster}}^i$ BY REFERRING TO clusters;
        FROM Pts_i REMOVE ${A_{cluster}}^i$;
        BREAK;
    ELSE:
        N = N − 1;
IF N == 1:
    loop = FALSE;
ELSE:
    loop = TRUE;
RETURN Pts_i, loop;

Algorithm 2. Loop Detection.

Inputs : Trajectory Tr_i.
${Pts_i}^l$= Pts_i;
${Pts_i}^l$, loop = AnomalyDection$\left({Pts_i}^l\right)$;
WHILE loop:
    ${Pts_i}^l$, loop = AnomalyDection$\left({Pts_i}^l\right)$;
anomaly_data = $Pts_i - {Pts_i}^l$;
$Tr_i = (MMSI_i, {Pts_i}^l)$;
RETURN Tr_i, anomaly_data;

4. Case Study

4.1. Data Description for the Case Study

To validate the effectiveness of the proposed anomaly detection method, in this section, a case study is illustrated by first applying the method on three different trajectories, where a rule-based method is also applied for comparison. The AIS data were provided by the database from the Wuhan University of Technology. The reason for choosing three different trajectories to validate the proposed method was to verify its performance for different scenarios of trajectories. The trajectories were picked from an AIS dataset from the Zhoushan area, China, on 15 October 2018 which contained some typical errors in the AIS data. The trajectories can be seen in Figure 7, and the information for each trajectory is shown in Table 3. Then, the method was applied to the whole dataset in this area to show the performance on a large trajectory set.

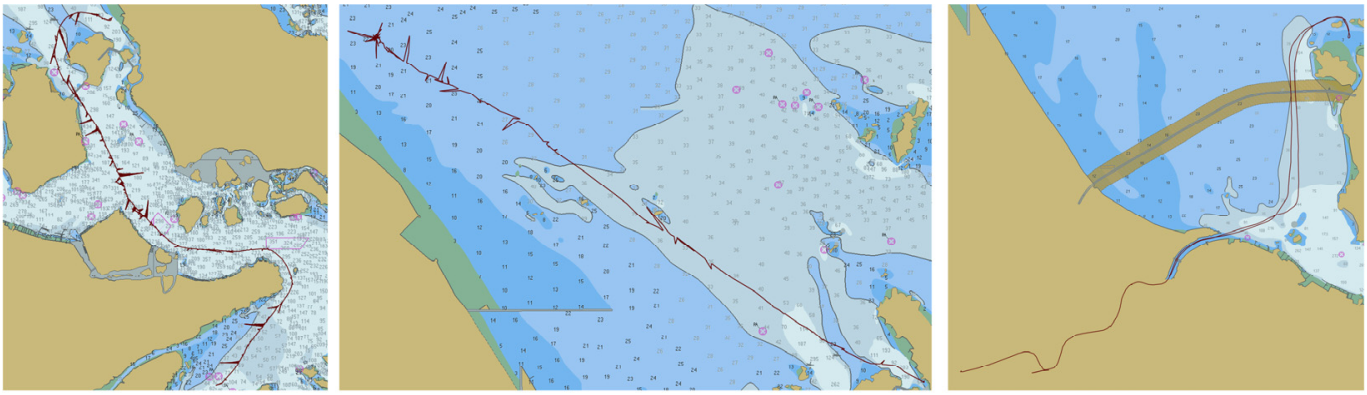


Figure 7. Trajectories Tr_1, Tr_2, and Tr_3.

Table 3. Trajectories in case study.

Trajectory	MMSI	Start Time	End Time	Number of Points
Tr_1	412XXX930	15 October 2018 12:01:04	15 October 2018 17:59:58	2935
Tr_2	413XXX210	15 October 2018 12:00:26	15 October 2018 18:00:00	1930
Tr_3	413XXX150	15 October 2018 12:00:04	15 October 2018 17:59:54	1953

4.2. Results and Comparison

4.2.1. Results of the Proposed Method

In the dataset, we found that there were many trajectories with obvious positional oscillation anomalies, such as Tr_1 and Tr_2. Tr_3 had fewer positional anomalies, but may have had kinematic anomalies such as velocity anomalies. Taking Tr_1 as an example, the first round of anomaly detection using kinematic estimation is shown in Figure 8. The blue line is the ship trajectory with the original data points, and the orange line is the trajectory obtained with the estimation points obtained with the kinematic method. By connecting all the estimated points following the time order, a sub-trajectory was obtained to better show the difference between known points and estimated points. One can see from Figure 8 that, when the points were normal, the sub-trajectory constructed with the estimated points was almost identical to the original path. On the contrary, if an anomaly occurred, the obvious difference between the estimated sub-trajectory and the original trajectory could be identified.

Figure 9 shows the clustering result during the loop detection process, where the S in each loop with the number of clusters is also shown. On this trajectory, the detection process was repeated 11 times and stopped at the 11th loop. Before the 9th loop, the standardization error set was divided into four clusters. In the 9th loop, the clustering result was 3. In the 10th loop, only two clusters remained. Finally, in the 11th loop, the clustering module separated the data into two clusters. However, the performance indicator S = 0.3384, which was lower than 0.5. The detection process therefore terminated, and the rest of the points were all considered to be normal points. Finally, we identified 1101 abnormal points, with 1834 normal points remaining. By applying the method on each trajectory, the amount of abnormal data for the cases can be seen in Table 4. The processed trajectories are shown in Figure 10 with comparisons between the original trajectories.

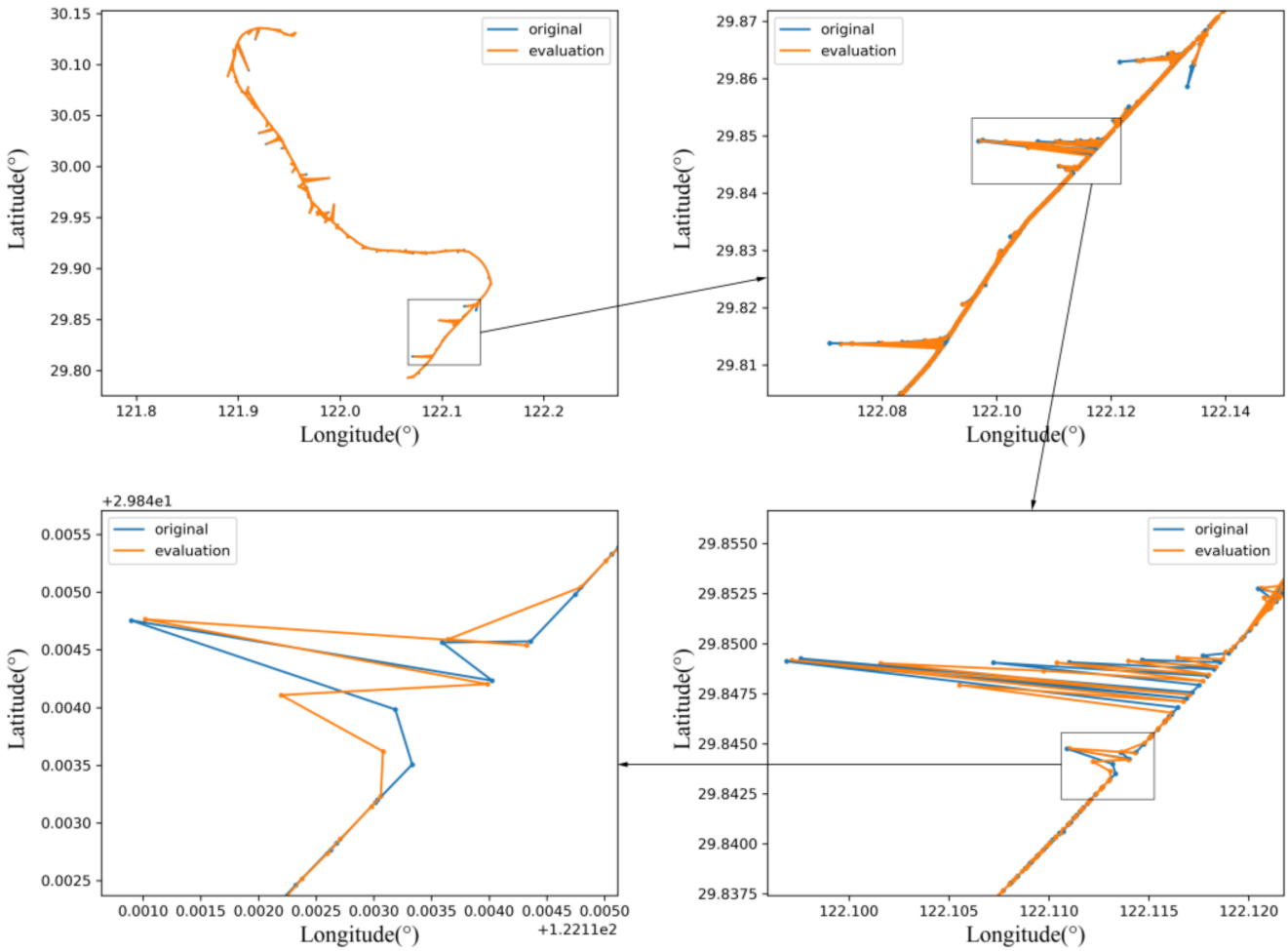


Figure 8. Kinematic estimation in the first detection round on Tr_1.

Table 4. Trajectory points in different states.

Trajectory	Number of Points before Detection	Number of Identified Abnormal Points	Number of Points after Detection
Tr_1	2935	1101	1834
Tr_2	1930	1115	815
Tr_3	1953	35	1918

Figure 10b,d,f shows that the position anomalies in Tr_1, Tr_2, and Tr_3 were successfully detected and removed. Then, the distances of the trajectories were reduced to varying degrees depending on their numbers of positional anomalies, and their average speeds were changed to normal, as shown in Table 5.

Table 5. Trajectory features.

Trajectory	Distance (n mile)		Average Speed (Kt)	
	Before	After	Before	After
Tr_1	176.94	33.24	29.58	5.56
Tr_2	74.11	33.90	12.37	5.80
Tr_3	39.01	37.67	6.50	6.28

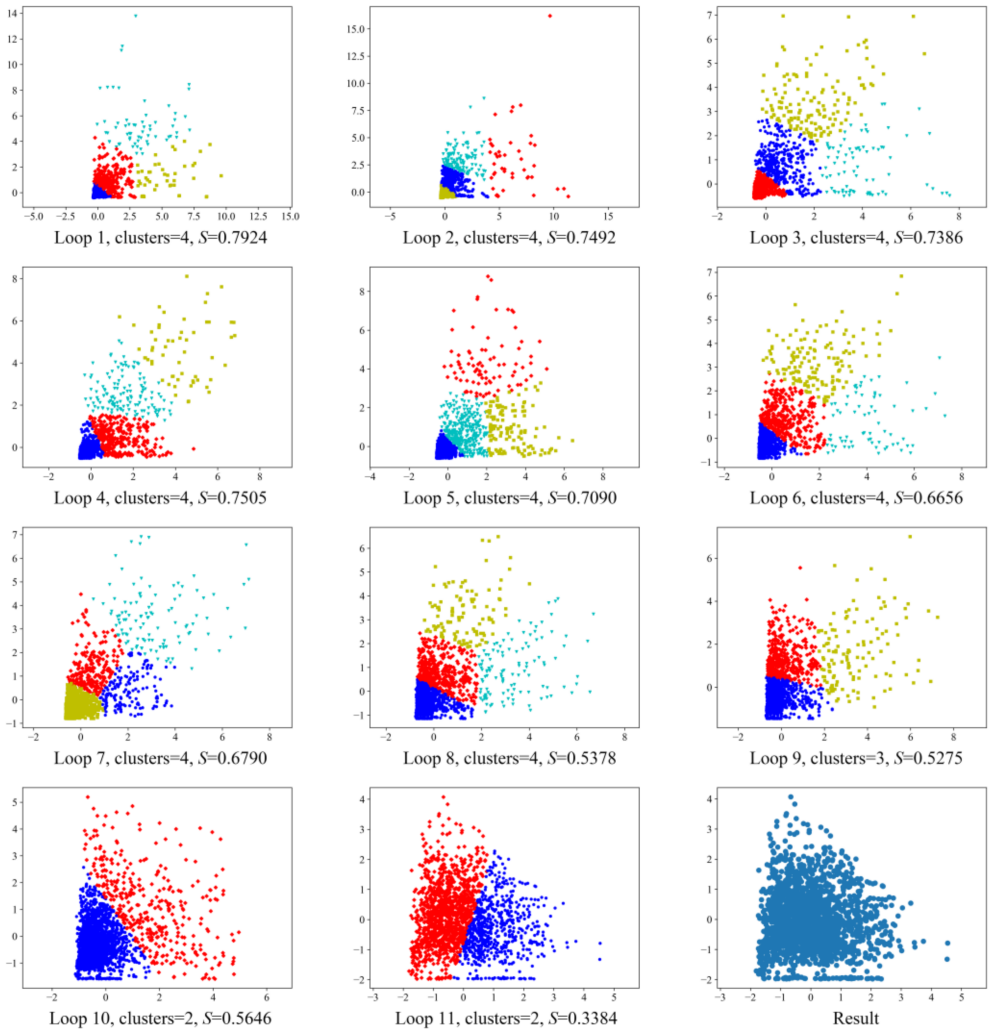


Figure 9. Error clustering on Tr_1.

In addition to positional anomalies, velocity anomalies should also be considered. The velocity change curves of each trajectory before and after the detection process are shown in Figure 11. The blue curve indicates the velocity information in the original trajectories and the orange curve indicates the velocity information in the anomaly-removed trajectories. In Figure 11a–c, the orange curves are much more stable than the blue curves, where the latter contain many sharp parts, presenting anomalies in the velocity. In Tr_2, when the ship was considered to be in a stationary state, some original data even showed that the ship was moving at a speed of 16 knots. Although the orange curve almost coincided with the blue curve in Tr_3, a few sharp parts were also detected. These results indicated that most of the velocity anomalies were detected effectively.

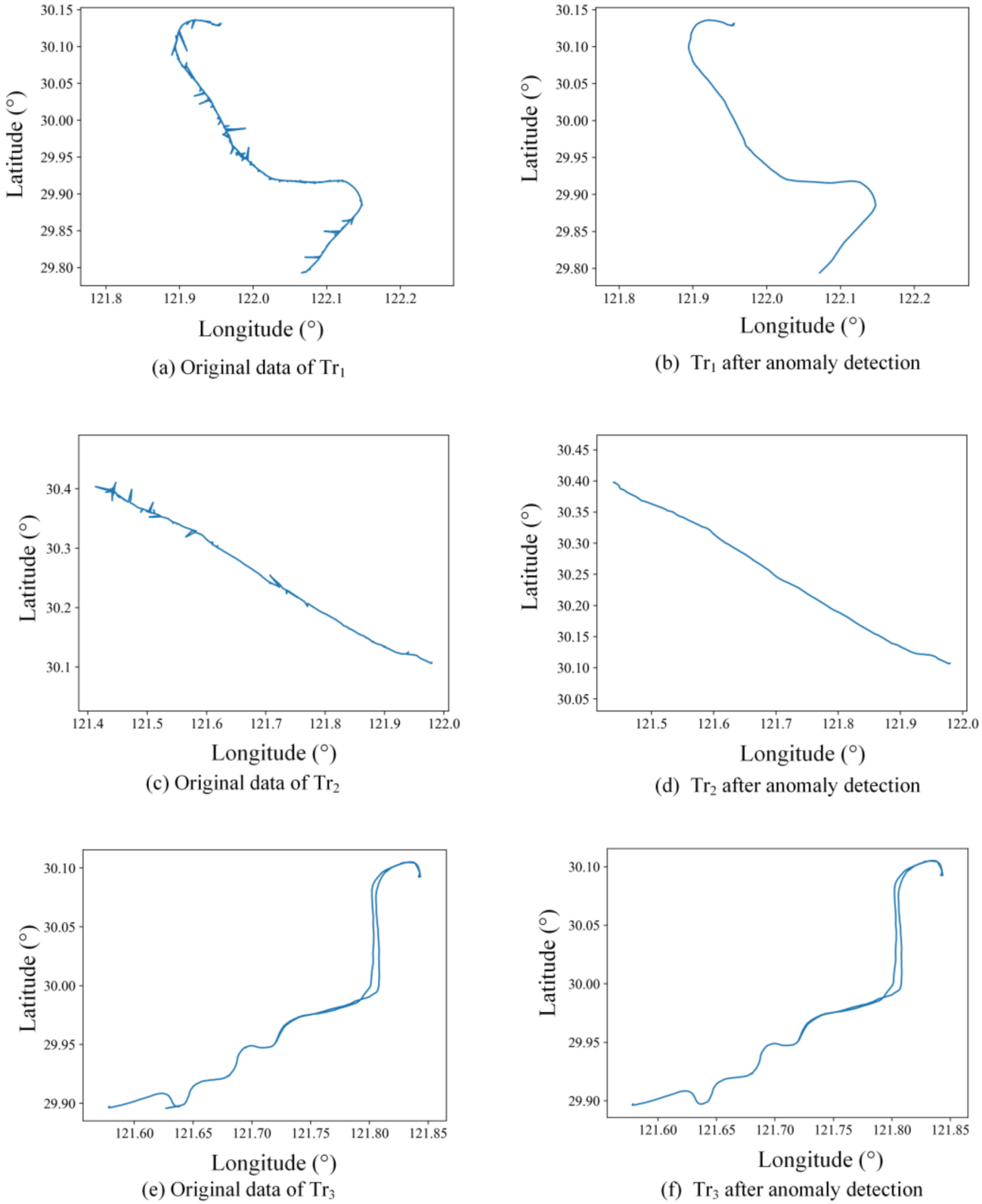


Figure 10. Comparison between the original trajectory and anomaly-removed trajectory.

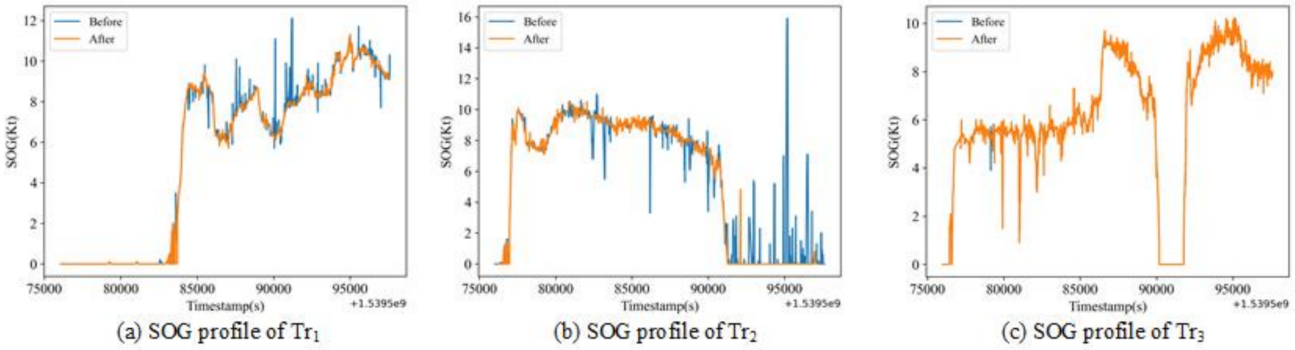


Figure 11. Comparison between the original trajectory and anomaly-removed trajectory from the speed perspective.

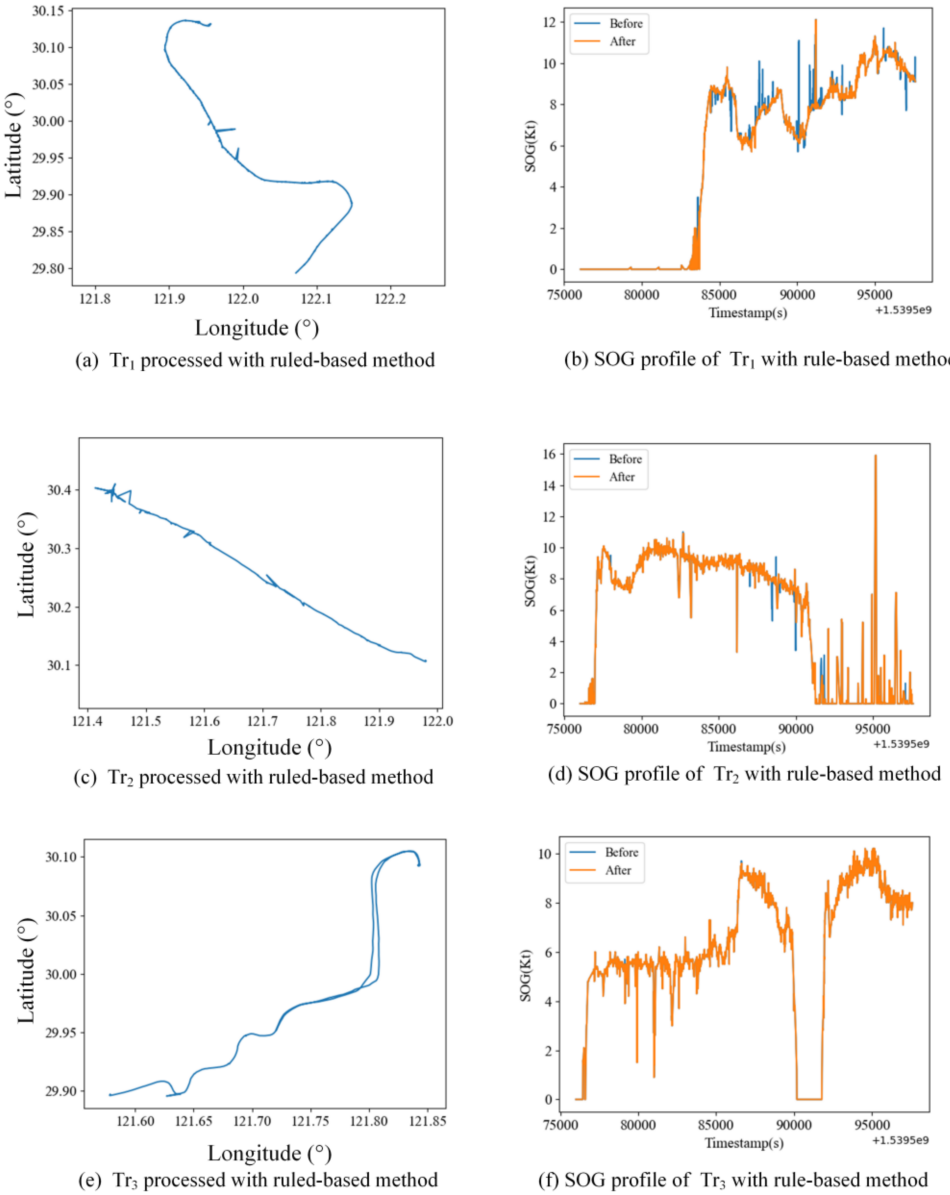
4.2.2. Comparison with the Rule-Based Detection Method

To further investigate the performance of the proposed method compared with the conventional anomaly detection approach that is frequently utilized in the pre-processing of AIS data, in this research, we conducted a comparison between the kinematic-based method proposed in our research and a typical rule-based anomaly detection algorithm such as the one utilized in [8]. Three of the same case trajectories are utilized in this section to provide clear results for the comparison of their performance in terms of the detection capability for location and velocity anomalies, respectively. A large-scale analysis of the performance of these algorithms is also conducted. The results are presented below.

Figure 12 shows the results of the anomaly detection utilizing the rule-based method. The column on the left side indicates the shape of the trajectories after the detection and removal process, where the blue line is the trajectory line. The column on the right side indicates the SOG profile of the trajectories before and after the detection and removal process, where the blue line indicates the original SOG profile obtained with the data points and the orange line indicates the SOG profile after the detection and removal process. As regards the performance of the anomaly detection process from the position perspective, compared with the results of the proposed method, it can be seen clearly from the figures that there are still many obvious position errors in the trajectory for Tr_1 and Tr_2. The difference for Tr_3 is not as significant as in the first two trajectories; the anomaly on the left-bottom corner of the trajectories was not successfully identified and removed with the rule-based approach. The same characteristics can also be seen in the SOG profile of the trajectories. As can also be seen from Figure 12b,d,f, the orange lines still contain various peaks in velocity and do not show a difference from the original SOG profile. To better indicate the performance of the method of the rule-based approach, the description of the data before and after the anomaly detection is also shown in Table 6. Compared with Table 4, the rule-based method identified fewer abnormal data points in the trajectories. Considering the analysis of the results of the rule-based anomaly detection method, the proposed method showed better performance in both the position and speed perspectives.

Table 6. Trajectory points in different states in the rule-based method.

Trajectory	Number of Points before Detection	Number of Identified Abnormal Points	Number of Points after Detection
Tr_1	2935	371	2564
Tr_2	1930	64	1866
Tr_3	1953	13	1940



(a) Tr_1 processed with ruled-based method
(b) SOG profile of Tr_1 with rule-based method
(c) Tr_2 processed with ruled-based method
(d) SOG profile of Tr_2 with rule-based method
(e) Tr_3 processed with ruled-based method
(f) SOG profile of Tr_3 with rule-based method

Figure 12. Results of the rule-based method.

In the aforementioned sections, the performance of the two methods is compared in a detailed manner using the three typical cases. To further investigate their performance on the large-scale data set, a ship trajectory data set containing 865,595 points was processed for anomaly detection and removal. With the proposed kinematic-based method, 137,421 abnormal points were successfully detected and removed from the original dataset, while 3911 abnormal points were detected and removed with the rule-based method, which further proved that the proposed kinematic-based anomaly detection has better performance in data anomaly detection for ship AIS data.

5. Conclusions

AIS has played a significant role in the research and development of the maritime traffic industry. However, anomalies and errors in the data have impinged on the data quality and therefore posed challenges to researchers and data scientists in facilitating this process. Therefore, as a fundamental step for the utilization and application of AIS data, it is of great significance to identify and remove anomalies and improve the data quality. In this research, a novel anomaly detection method for AIS trajectories has been proposed by

integrating the kinematic information in AIS and a clustering-based method to identify anomalies in AIS data considering the kinematic characteristics of the ship.

The abnormal data in the ship AIS data set are detected by using kinematic interpolation. Using the knowledge of known trajectory points collected from raw AIS data, kinematic interpolation is used to estimate the possible errors of the original trajectory data. After the kinematic estimation and error calculation processes, the possibility of an anomaly in each point is measured with an error weight. An improved K-means clustering method is then applied to identify abnormal data by clustering the error weights of the data points. Furthermore, to achieve comprehensive detection, the error detection and clustering process is further integrated with a loop design by utilizing the silhouette coefficient as a termination condition to evaluate the performance of the clustering.

To validate the effectiveness of the presented method on different scenarios of data anomalies, a case study associated with three trajectories was conducted. From the results with kinematic estimation, one can see that the proposed method was able to successfully identify the position and velocity anomaly at the same time, showing better performance than the conventional method that only considers the problem from a position perspective. The repeated clustering process enabled the proposed method to identify all the anomalies in the trajectories and improve the data quality as much as possible. The comparison between the proposed method and the rule-based anomaly detection method indicates that, for both detailed analysis and application on large-scale data sets, the proposed kinematic-based method can identify more anomalies in both positions and speed in a data set than previous approaches.

Author Contributions: S.G.: Conceptualization, Methodology, Formal analysis, Investigation, Visualization, Writing—Original manuscript; J.M.: Methodology, Resources, Supervision, Funding acquisition; L.C.: Methodology, Visualization, Funding acquisition, Writing—Review; P.C.: Conceptualization, Methodology, Supervision, Writing—Review and Editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research is financially sponsored by the National Natural Science Foundation of China (Grant No. 52001242) and the Fundamental Research Funds for the Central Universities (WUT: 2021IVA049, 2021IVA051).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data available on request due to restrictions of privacy.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Yang, D.; Wu, L.; Wang, S.; Jia, H.; Li, K.X. How big data enriches maritime research—A critical review of Automatic Identification System (AIS) data applications. *Transp. Rev.* **2019**, *39*, 755–773. [CrossRef]
2. Wei, Z.; Xie, X.; Zhang, X. AIS trajectory simplification algorithm considering ship behaviours. *Ocean. Eng.* **2020**, *216*. [CrossRef]
3. Svanberg, M.; Santén, V.; Hörteborn, A.; Holm, H.; Finnsgård, C. AIS in maritime research. *Mar. Policy* **2019**, *106*. [CrossRef]
4. Iphar, C.; Ray, C.; Napoli, A. Data integrity assessment for maritime anomaly detection. *Expert Syst. Appl.* **2020**, *147*. [CrossRef]
5. He, W.; Lei, J.; Chu, X.; Xie, S.; Zhong, C.; Li, Z. A Visual Analysis Approach to Understand and Explore Quality Problems of AIS Data. *J. Mar. Sci. Eng.* **2021**, *9*, 198. [CrossRef]
6. Riveiro, M.; Pallotta, G.; Vespe, M. Maritime anomaly detection: A review. *WIREs Data Min. Knowl. Discov.* **2018**, *8*. [CrossRef]
7. Praczyk, T. Ship trajectory anomaly detection. *Intell. Data Anal.* **2019**, *23*, 1021–1040. [CrossRef]
8. Chen, X.; Ling, J.; Yang, Y.; Zheng, H.; Xiong, P.; Postolache, O.; Xiong, Y. Ship Trajectory Reconstruction from AIS Sensory Data via Data Quality Control and Prediction. *Math. Probl. Eng.* **2020**, *2020*, 1–9. [CrossRef]
9. Zhang, X.H.; He, Y.X.; Tang, R.H.; Mou, J.M.; Gong, S. A Novel Method for Reconstruct Ship Trajectory Using Raw AIS Data. In Proceedings of the 2018 3rd Ieee International Conference on Intelligent Transportation Engineering (ICITE), Singapore, 3–5 September 2018; pp. 192–198.
10. d'Afflisio, E.; Braca, P.; Millefiori, L.M.; Willett, P. Maritime Anomaly Detection based on Mean-Reverting Stochastic Processes Applied to a Real-World Scenario. In Proceedings of the 2018 21st International Conference on Information Fusion (Fusion), Cambridge, UK, 10–13 July 2018; pp. 1171–1177.

11. Sang, L.-z.; Wall, A.; Mao, Z.; Yan, X.-P.; Wang, J. A novel method for restoring the trajectory of the inland waterway ship by using AIS data. *Ocean Eng.* **2015**, *110*, 183–194. [CrossRef]
12. Zhang, L.; Meng, Q.; Xiao, Z.; Fu, X. A novel ship trajectory reconstruction approach using AIS data. *Ocean Eng.* **2018**, *159*, 165–174. [CrossRef]
13. Morris, B.T.; Trivedi, M.M. A Survey of Vision-Based Trajectory Learning and Analysis for Surveillance. *IEEE Trans. Circuits Syst. Video Technol.* **2008**, *18*, 1114–1127. [CrossRef]
14. Laxhammar, R.; Falkman, G. Sequential Conformal Anomaly Detection in Trajectories based on Hausdorff Distance. In Proceedings of the 14th International Conference on Information Fusion, Chicago, IL, USA, 5–8 July 2011.
15. Liu, C.; Wang, J.; Liu, A.; Cai, Y.; Ai, B. An Asynchronous Trajectory Matching Method Based on Piecewise Space-Time Constraints. *IEEE Access* **2020**, *8*, 224712–224728. [CrossRef]
16. Mazzarella, F.; Vespe, M.; Alessandrini, A.; Tarchi, D.; Aulicino, G.; Vollero, A. A novel anomaly detection approach to identify intentional AIS on-off switching. *Expert Syst. Appl.* **2017**, *78*, 110–123. [CrossRef]
17. 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]
18. Zhou, Y.F.; Wright, J.; Maskell, S. A Generic Anomaly Detection Approach Applied to Mixture-of-unigrams and Maritime Surveillance Data. In Proceedings of the 2019 Symposium on Sensor Data Fusion: Trends, Solutions, Applications (SDF 2019), Bonn, Germany, 15–17 October 2019.
19. 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]
20. Sidibé, A.; Shu, G.; Ma, Y.; Wanqi, W. Big Data Framework for Abnormal Vessel Trajectories Detection using Adaptive Kernel Density Estimation. In Proceedings of the 2nd International Conference on Big Data Research-ICBDR 2018, Weihai, China, 27–29 October 2018; pp. 43–46.
21. Zhao, L.; Shi, G. Maritime Anomaly Detection using Density-based Clustering and Recurrent Neural Network. *J. Navig.* **2019**, *72*, 894–916. [CrossRef]
22. Jie, X.; Chaozhong, W.; Zhijun, C.; Xiaoxuan, C. A novel estimation algorithm for interpolating ship motion. In Proceedings of the 2017 4th International Conference on Transportation Information and Safety (ICTIS), Banff, AB, Canada, 8–10 August 2017; pp. 557–562.
23. ITU. Technical Classifications of Recommendations ITU-M. 1371-4. Technical Characteristics for an Automatic Identification System Using Time-Division Multiple Access in the VHF Maritime Mobile Band. Available online: https://www.itu.int/rec/R-REC-M.1371/en (accessed on 28 February 2021).
24. Grubbs, F.E. Procedures for Detecting Outlying Observations in Samples. *Technometrics* **1969**, *11*, 1–21. [CrossRef]
25. Kumar, D.; Bezdek, J.C.; Rajasegarar, S.; Leckie, C.; Palaniswami, M. A visual-numeric approach to clustering and anomaly detection for trajectory data. *Vis. Comput.* **2015**, *33*, 265–281. [CrossRef]
26. Soleimani, B.H.; De Souza, E.N.; Hilliard, C.; Matwin, S. Anomaly Detection in Maritime Data Based on Geometrical Analysis of Trajectories. In Proceedings of the 2015 18th International Conference on Information Fusion (Fusion), Washington, DC, USA, 6–9 July 2015; pp. 1100–1105.
27. Long, J.A. Kinematic interpolation of movement data. *Int. J. Geogr. Inf. Sci.* **2015**, *30*, 854–868. [CrossRef]
28. Shi, P.; Zhao, Z.; Zhong, H.; Shen, H.; Ding, L. An improved agglomerative hierarchical clustering anomaly detection method for scientific data. *Concurr. Comput. Pract. Exp.* **2020**, *33*. [CrossRef]
29. Wang, Z.; Zhou, Y.H.; Li, G.M. Anomaly Detection by Using Streaming K-Means and Batch K-Means. In Proceedings of the 2020 5th Ieee International Conference on Big Data Analytics (IEEE ICBDA 2020), Xiamen, China, 8–11 May 2020; pp. 11–17.
30. Saxena, A.; Prasad, M.; Gupta, A.; Bharill, N.; Patel, O.P.; Tiwari, A.; Er, M.J.; Ding, W.; Lin, C.-T. A review of clustering techniques and developments. *Neurocomputing* **2017**, *267*, 664–681. [CrossRef]
31. Rukmi, A.M.; Iqbal, I.M. Using k-means++ algorithm for researchers clustering. In Proceedings of the International Conference on Mathematics: Pure, Applied and Computation: Empowering Engineering Using Mathematics, Surabaya, Indonesia, 23 November 2016.
32. Rousseeuw, P.J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53–65. [CrossRef]