Hindawi
Scientific Programming
Volume 2023, Article ID 8345603, 12 pages
https://doi.org/10.1155/2023/8345603

Research Article

A Real-Time AIS Data Cleaning and Indicator Analysis Algorithm Based on Stream Computing

Taizhi Lv,[1,2] **Peiyi Tang**,[2] **and Juan Zhang**[1]

[1]*School of Information Engineering, Jiangsu Maritime Institute, Nanjing 211170, China*
[2]*Nanjing Huihai Transportation Technology Company Limited, Nanjing 210019, China*

Correspondence should be addressed to Taizhi Lv; lvtaizhi@163.com

Received 31 January 2023; Revised 24 February 2023; Accepted 2 March 2023; Published 12 April 2023

Academic Editor: Bai Yuan Ding

The data quality and real-time analysis of automatic identification system (AIS) are of great significance for water transportation safety and intelligent maritime construction. To improve the AIS data quality and analyze AIS data in real time, a real-time AIS data cleaning and indicator analysis algorithm is proposed. This algorithm performs distributed real-time data cleaning and analysis for massive AIS data based on stream computing technology. It includes data fusion, deduplication, decoding, abnormal data identification, sequencing, prediction, and statistics steps. Abnormal AIS data are repaired by linear regression, multiple trajectory tracking, caching, and other technologies. The AIS status is determined in real-time via multidimensional AIS packet loss analyses, multifactor AIS data statistics, and spatial-temporal data visualization, effectively improving the intelligence level of maritime supervision applications. The proposed algorithm has been running on a production environment, and it monitors AIS data in a certain section of the Yangtze River Basin 24 hours every day without interruption. The operation results show that the proposed algorithm can improve the quality of AIS data, addresses ship trajectory jump issues, and provides timely position updates. The real-time indicator analysis results can provide the data support for ship navigation and maritime supervision.

1. Introduction

Automatic identification system (AIS) is a digital navigation aid system that exchanges navigation information between ships and shore-based stations [1]. AIS, radar, and surveillance video are the most important perceptual data source of water transportation. They all constitute an important cornerstone of smart shipping [2, 3]. With the improvement of AIS availability, AIS applications have extended from early navigation to various fields, such as navigation behavior analysis, navigation safety, trade analysis, environmental assessment, and maritime supervision. Adland et al. predicted the global oil trade according to the sea transportation volume, which is calculated from the AIS data [4]. In reference [5], a dynamic method was combined with the emission model STEAM2 to calculate the ship pollutant emissions. Ship trajectory prediction based on spatial attributes is one of the most important research areas concerning AIS data. Liu et al. proposed a BLSTM-based deep learning network. By being embedded with the dynamic AIS data and social force concept, it guarantees high-accuracy ship trajectory prediction [6]. Murry and Perera proposed a novel dual linear autoencoder approach to extract navigation trajectories and predict navigation routes from the AIS data [7].

AIS applications such as water transportation safety and maritime supervision have the very high data quality requirements. Low-quality AIS data will not only affect water traffic management but also bring wrong results to data analysis. For example, it is hard to discern the normal trajectory when multiple ships share one maritime mobile service identity (MMSI). Due to the accuracy of ship navigation sensing equipment, the reliability of shipborne AIS equipment, the layout and capacity of AIS shore-based stations, the terrain environment, artificial electromagnetic radiation, and other factors, high-quality AIS data are often unavailable, especially in inland river environments [8, 9]. Scholars have conducted many studies on the quality of AIS data. In reference [10], integrity and completeness of AIS data were discussed. It concluded that reliable AIS data are the key in the process of ship collision avoidance. By empirically investigating the Dover Straits, Baily found that a considerable amount of AIS data is inaccurate, and many incorrect MMSI lead to ship positioning failures [11]. Banyś et al. analyzed AIS data from the Baltic Sea coast in Germany and found that more than 30% of the ship heading and rotation AIS data were unknown, resulting in many problems when AIS data are used for ship collision avoidance [12]. For specific interests, many individuals and organizations falsify AIS data, which disrupts navigation order [13]. To analyze the reliability of AIS data, Liu et al. defined ship AIS service indicators for ships, regions, and shore-based stations. The AIS performance was analyzed and visualized in a section of the Yangtze River [14]. In order to improve the quality of AIS data, scholars have made improvements from hardware equipment, signal measurement, network communication, multisource data fusion, big data technology, etc. Zhang et al. proposed a pseudorange measurement algorithm based on AIS signals, which can greatly improve the estimation accuracy under the low signal-to-noise ratio (SNR) condition [15]. In reference [16], the impact of space-based AIS antenna orientation on AIS detection performance is conducted. The orientation of AIS monopole antenna can increase the detection of AIS signals. Jaskólski et al. used simultaneous localization and mapping (SLAM) process model based on the fusion of radar and AIS data to track the ship trajectory [17]. Low-quality AIS data can be improved by data preprocessing technology. Zhao et al. preprocessed AIS data in terms of physical integrity, spatial logic integrity, and time accuracy to ensure the accuracy of AIS data analysis [18]. Siegert et al. proposed an EKF method to monitor AIS data integrity and track the ship trajectory [19]. Chen et al. proposed a novel approach to detect anomaly AIS data based on the ships' maneuverability. The cubic spline interpolation method was used to repair the trajectory after eliminating the abnormal points [20]. Sang et al. constructed a smooth AIS filtering algorithm that deletes dynamic AIS data with abnormal positions according to a preset threshold [21]. In reference [22], an AIS data preprocessing method, which combines TPNet and LSTM, is proposed. It can improve the accuracy of the ship trajectory prediction.

Most studies of AIS data cleaning and analysis focus on off-line methods which are based on the traditional cloud computing model. In this model, the data processing is based on batch processing, where a large amount of historical data are processed at one time. For maritime supervision, ship collision avoidance, water traffic management, and other applications, it is necessary to obtain the cleaned AIS data and statistical indicators in real time. The traditional off-line methods cannot adapt to these application scenarios. To clean AIS data and obtain performance statistics in real-time, a real-time AIS data cleaning and indicator analysis algorithm based on stream computing is proposed. In this framework, a piece of AIS data is processed immediately when it is received. Compared with the traditional computing methods represented by Map/Reduce framework, stream computing provides a better solution for real-time AIS data processing and can better meet the real-time requirements. The cloud receives an AIS packet and performs real-time deduplication, decoding, inaccuracy identification, repairing, and analysis. When inaccuracy data are identified, the historical data are used to repair them. The reasons underlying AIS data losses are analyzed with linear predictions and data statistics according to the AIS capacity, environment, and ship factors. The statistics of the AIS packet, ship number, AIS abnormalities, and AIS losses are determined at three levels: ship level, grid level, and system level. Real-time data cleaning and indicator analysis can effectively improve the quality of AIS data, address the ship position jump issues, and provide timely ship navigation status. The improved AIS data effectively enhance the application of AIS data in ship tracking, intelligent search and rescue, and water traffic organization. Three approaches to increase the performance of real-time AIS data process are introduced to the proposed algorithm.

(i) A whole AIS stream computing architecture, which implements real-time process in the whole processing from data fusion, deduplication, decoding, abnormal data identification, sequencing, prediction to statistics, is constructed

(ii) By repairing abnormal AIS data with linear regression, multiple trajectory tracking, caching, it can effectively improve the quality of AIS data

(iii) By multidimensional AIS packet loss analyses, multifactor AIS data statistics, and spatial-temporal data visualization, it can effectively improve the intelligence level of shipping

2. Cause Analysis of AIS Data Abnormality

Low-quality AIS data seriously affect water transportation management. AIS abnormalities are classified as data inaccuracy and data loss in this paper.

2.1. Data Inaccuracy. Data inaccuracy indicates an inconsistency between the received AIS data and the actual shipping status.

2.1.1. Inaccuracy of Ship Basic Information. Since the basic ship information is input manually by crews, incorrect data or data that are not compliant with the specifications may be input. In practical applications, large inaccuracies in destination, length, and type of ships occur.

2.1.2. Inaccuracy of Ship Navigation Status. For the positioning errors caused by GPS (global position system) or BDS (BeiDou navigation satellite system) and navigation data errors caused by INS (inertial navigation system), the ship navigation information of the AIS data may be inconsistent with the actual ship navigation status.

2.1.3. MMSI Jumping. MMSI is a nine-digit code that uniquely identifies AIS equipment. Due to human errors, equipment failures, pseudo shore-based stations, and other reasons, the same MMSI may be used by different ships. This can cause jumping phenomena in AIS trajectories and seriously affect water transportation safety.

2.2. Data Loss. The AIS data transmission interval is determined according to the AIS equipment type and speed and steering rate of the ship. The data losses occur when AIS shore-based stations do not receive the corresponding AIS data within the specified time.

2.2.1. Signal Transmission Attenuation. AIS is an automatic reporting system based on very high frequency (VHF). Transmission distance, terrain on both sides of channels, adjacent buildings, weather conditions, etc., all affect signal transmission. Transmission attenuation of the radio waves may weaken the AIS signals. Noise may be amplified when AIS shore-based stations modulate and amplify the received signal. This amplification can cause the data errors. Data with errors are discarded for the lack of a fault-tolerant mechanism in AIS [23].

2.2.2. Capacity of Shore-Based Station. The capacity of an AIS shore-based station indicates the maximum amount of shipborne AIS equipment with which the shore-based station can communicate. Since the AIS reporting rate depends on the AIS equipment type, shipping speed, and steering angle, the system capacity is dynamic [24]. When the required amount of data transmitted is greater than the capacity, timeslot multiplexing occurs. Timeslot multiplexing may result in identification communication or blind communication issues when the shore-based station receives the AIS data. Multiple shipborne AIS equipment may send messages at the same timeslot. This causes the shore-based station to receive only one message correctly, or all messages cannot be received correctly [25].

2.2.3. Own Factors of Shipborne AIS Equipment. The transmit power, power supply stability, manual shutdown, and other factors may cause AIS data loss. The AIS transmit power is generally 12.5 W. When the power becomes weaker and the ship is far from the AIS shore-based station, the station may not receive the AIS data. If there is no built-in stabilized power supply in the shipborne AIS equipment, there may be insufficient current, resulting in transmission failure of the AIS signal [26].

2.2.4. Noise Jamming. Many cities are located along waterways, and artificial electromagnetic radiation sources interfere AIS signal transmission, affecting the reliability of AIS data [27]. Mirror interference and skywave interference are two kinds of interference that can reduce the signal transmission accuracy.

3. Algorithm Architecture

As shown in Figure 1, the proposed algorithm includes a data acquisition layer, data aggregation layer, data cleaning layer, data analysis layer, data storage layer, and data visualization layer.

The data acquisition layer is implemented by some AIS acquisition clients which acquire the AIS data from shore-based stations through computer networks. The acquired AIS data are transmitted to the data aggregation layer. Based on high throughput, low latency, high reliability, high concurrency, and fault tolerance [28], Kafka is used to aggregate the acquired AIS data and send them to the next layer for computing. The data cleaning layer realizes AIS data deduplication, decoding, and inaccurate data identification. As an excellent stream computing framework, Flink is used to realize real-time computing. The computing results are transmitted to the data analysis layer and the data storage layer. The data storage layer includes the Redis cluster, MySQL cluster, and Doris cluster. By the multilevel storage structure, the layer can support distributed storage and has high availability [29]. The data analysis layer calculates real-time statistics at different levels according to the various indicators. The visualization layer adopts the front-end and back-end separation technology. The back end uses Spring Boot framework to realize business logic processing, and the front end uses Vue, Echarts, and MapBox to visualize data.

4. Rea-Time AIS Data Cleaning and Indicator Analysis

4.1. Processing Flow. As shown in Figure 2, the proposed algorithm is divided into two Flink cluster processing levels. The first-level Flink cluster completes the data cleaning and ship trajectory prediction. The processing results are then sent to the second-level Flink and the data storage layer via the message queue. The second-level Flink cluster computes the capacity and analyzes the AIS packet loss and AIS service performance according to ship, grid, and shore-based station levels.

4.2. AIS Data Cleaning

4.2.1. Data Deduplication. For the reception areas overlap of different shore-based stations, broadcast AIS packets are received by multiple AIS shore-based stations [30]. Thus, the AIS data must be deduplicated to reduce the computational burden for the next computing. An AIS packet does not contain a complete timestamp and contains only UTC seconds. Thus, two packets may be equal when the MMSI, longitude, latitude, speed, etc., are all the same. In most cases, it takes less than 1 minute for an AIS packet to be sent from a ship to the cloud. It is extremely unlikely to receive two identical packets within one minute, and the generation time difference between two packets is greater than one minute.

The time to live (TTL) feature of Redis is used to identify duplicated AIS packets within 1 minute, and duplicated data

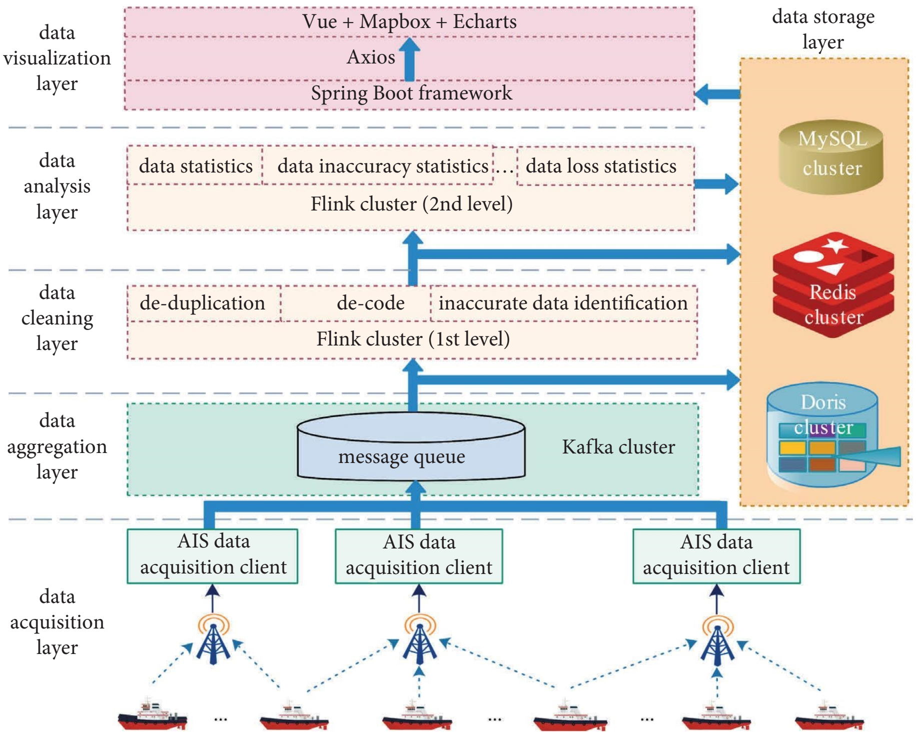

Figure 1: The framework of the proposed algorithm.

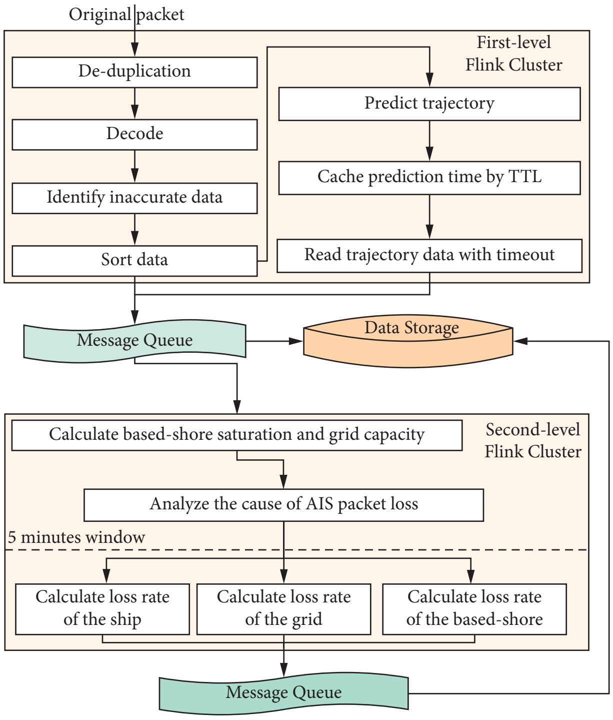
Original packet
De-duplication
Decode
Identify inaccurate data
Sort data
First-level Flink Cluster
Predict trajectory
Cache prediction time by TTL
Read trajectory data with timeout
Message Queue
Data Storage
Calculate based-shore saturation and grid capacity
Second-level Flink Cluster
Analyze the cause of AIS packet loss
5 minutes window
Calculate loss rate of the ship
Calculate loss rate of the grid
Calculate loss rate of the based-shore
Message Queue

Figure 2: The processing flow.

are filtered out by the filter operator of Flink. The TTL feature ensures that the stored data are valid only for a certain period of time. When the life cycle ends, the stored data become invalid [31]. Thus, if the new AIS packet into the Flink cluster already exists in the Redis cache, the packet is repeated, and it is filtered. If the packet does not exist in the Redis cache, the packet is stored in the Redis cache while the TTL is set to 1 minute.

4.2.2. Data Decoding. To improve the transmission efficiency, the AIS packet is transmitted by compression coding. The map operator of Flink is used to decode the original packet in real time. The map operator converts each AIS packet to the plain text. As shown in Figure 3, a complete original AIS packet consists of 7 fields separated by commas.

Data decoding converts the encapsulated message from the original AIS packet into the plain text according to the VDM protocol. The parsing process first determines whether the cyclic redundancy check (CRC) is consistent with the packet. If the CRC is consistent, the corresponding start byte is determined according to the corresponding message type, the specified bit width is maintained, and the data are converted and merged to determine the required information.

4.2.3. Inaccurate Data Identification and Repairing. This step identifies and repairs inaccurate AIS data. The main inaccuracies include position abnormalities, speed and heading abnormalities, multiple trajectory abnormalities, and receiving sequence abnormalities. The repaired data are marked and stored in the data storage layer before being sent to the next Flink cluster for next operations.

(1) Position Abnormality. If the values of the longitude and latitude are not within the normal range, the AIS packet is marked with abnormal position. Abnormal longitude and

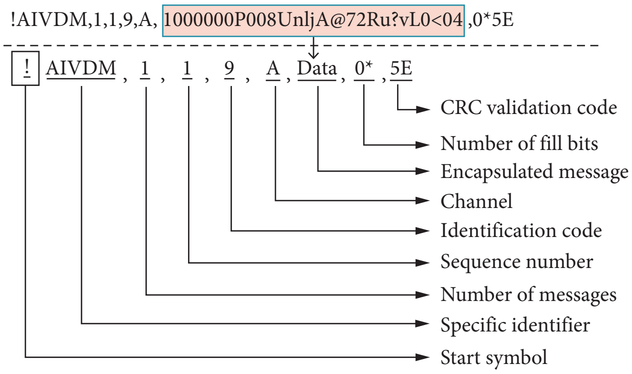

Figure 3: The original AIS packet format.

latitude are repaired according to the ship trajectory. If the time difference between adjacent packets is less than 10 minutes, the current packet is repaired according to the ship trajectory. If the time difference exceeds 10 minutes, the packet is not repaired and is marked as invalid. If the speed and heading differences between adjacent AIS packets are within the threshold, the average value is taken as the speed and heading. The Mercator algorithm which is a difference calculation method based on Mercator chart projections with equal angles and constant straight direction lines [32] is used to calculate the ship trajectory.

The longitude and latitude in the previous AIS packet are denoted as long_{t-1} and lat_{t-1}, the speed of the current AIS packet is denoted as speed_t, the heading angle is denoted as HDG_t, t_{diff} denotes the time difference between the two adjacent packets, and S_t denotes the distance between the adjacent packets. The current longitude long_t and latitude lat_t are calculated as follows:

$$\mathrm{lat}_t = \mathrm{la}t_{t-1} + S_t \times \cos(\mathrm{HDG}_t), \tag{1}$$

$$\mathrm{long}_t = \mathrm{long}_{t-1} + \mathrm{DMP} \times \tan(\mathrm{HDG}_t), \tag{2}$$

$$S_t = \mathrm{speed} \times t_{\mathrm{diff}}, \tag{3}$$

$$\mathrm{DMP} = \mathrm{MP}(\mathrm{lat}_t) - \mathrm{MP}(lat_{t-1}), \tag{4}$$

where MP is the meridian arc length of the projection of the unit latitude in the Mercator map vertical coordinate.

$$\mathrm{MP}(\varphi) = 7915.70447 \times \lg\left(\tan\left(\frac{\pi}{4}+\frac{\varphi}{2}\right) \times {\frac{1-e\times\sin(\varphi)}{1+e\times\sin(\varphi)}}^{\frac{e}{2}}\right). \tag{5}$$

If the speed and heading differences between the two adjacent AIS packets exceed the threshold, the current longitude and latitude of the ship are iteratively derived by evenly inserting $n-2$ points. The first node is the position of the previous AIS packet, and the n-th node is the position of the current AIS packet. Assuming that the changes of acceleration and steering angle of the ship are uniform, the speed and heading angel of the ship at the i-th node can be calculated as follows:

$$\mathrm{speed}_i = \mathrm{speed}_{t-1} + \frac{(\mathrm{speed}_t - \mathrm{speed}_{t-1}) * i}{n-1}, \tag{6}$$

$$\mathrm{HDG}_i = \mathrm{HDG}_{t-1} + \frac{(\mathrm{HDG}_t - \mathrm{HDG}_{t-1}) * i}{n-1}. \tag{7}$$

The longitude and latitude of each point are successively derived according to equations (1)–(5).

(2) Speed and Heading Abnormality. If the speed and heading of the ship are not within the normal range, the AIS packet is marked with abnormal speed and heading. The ship positions of the adjacent AIS packets are used to repair the current ship speed and heading.

$$\mathrm{HDG}_t = \arctan\left(\frac{\mathrm{long}_t - \mathrm{long}_{t-1}}{\mathrm{DMP}}\right), \tag{8}$$

$$\mathrm{speed}_t = \frac{(\mathrm{lat}_t - \mathrm{lat}_{t-1}) \times \sec(\mathrm{HDG}_t)}{t_{\mathrm{diff}}}, \tag{9}$$

when the latitudes from the two adjacent AIS packets are the same, the ship is moving along a parallel. If long_t is greater than long_{t-1}, then $\mathrm{HDG}_t = 90$; otherwise, $\mathrm{HDG}_t = 270$.

(3) Multiple Trajectory Abnormality. The MMSI is manually input into a shipborne AIS equipment. Thus, input errors may cause two or more ships to use the same MMSI. As a result, the trajectory jumps between these ships. It results ship identification and tracking errors. Multiple trajectory abnormalities identification and repairing process are shown in Figure 4.

Whether a ship is newly into the Flink cluster is determined by the MMSI information stored in the Redis cluster. For a new MMSI, the ship information is stored in the Redis cluster, and a new trajectory is added for the novel MMSI. For ships whose MMSI exists in the Redis cluster, the position of the current AIS packet is matched to the existing trajectories. The distance between the position of the new AIS packet and the last position of each trajectory is calculated. If the distance is greater than the maximum possible navigation distance, the new position does not match the given trajectory. If there is no matching trajectory, a new trajectory is added for the MMSI. If there is a matching trajectory, the matching trajectory is updated.

(4) Receiving Sequence Abnormalities. Since AIS packets are acquired from different shore-based stations, the packet receiving order is uncertain. The packet that is sent first may be received last. If the sequence is not appropriately ordered, the ship trajectory may jump back and forth. To sort the AIS packets, AIS packets are not stored directly to the data storage layer and sent to the next step for statistical analyses. Instead, the AIS packets are cached, and all AIS packets acquired in the time window are sorted.

The time window is set to 1 minute, and the AIS data in this window are sorted. The AIS data in the window are denoted as $x_1, x_2, \ldots, x_{n-1}$, and the newly received AIS data are denoted as x_n. x_n is compared with x_i $(i = n-1, n-2, \ldots 1)$ and inserted into

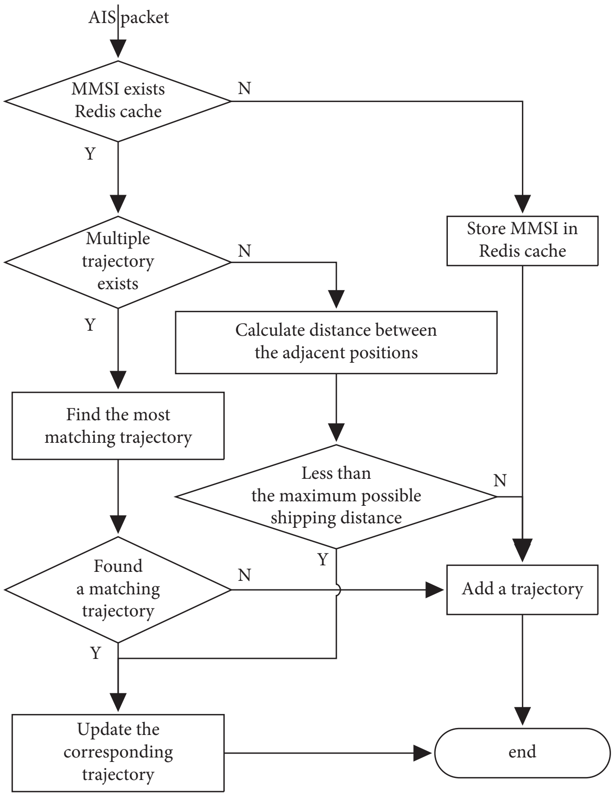
AIS packet
MMSI exists Redis cache
N
Y
Store MMSI in Redis cache
Multiple trajectory exists
N
Y
Calculate distance between the adjacent positions
Find the most matching trajectory
Less than the maximum possible shipping distance
N
Y
Found a matching trajectory
N
Add a trajectory
Y
Update the corresponding trajectory
end

Figure 4: Multiple trajectory identification flow.

the appropriate position. The direction angle between x_n and x_i is calculated. If the angle is consistency with the heading, x_n is inserted after x_i, and the sorting is finished. If the angle is inconsistency with the heading, then x_n is compared with x_{i-1} until all data in the cache have been compared.

4.3. AIS Data Prediction.

4.3. AIS Data Prediction. The transmission time of AIS packets depends on the ship navigation status and water traffic environment. To analyze the reliability of AIS data, it is necessary to predict the received time and ship status of the next AIS packet. This step not only realizes trajectory filling when the AIS packet is lost but also provides data for AIS reliability statistics.

4.3.1. Receiving Time Prediction for the Future AIS Packet. The AIS transmission frequency depends on the message type, shipborne AIS equipment type, and navigation status. AIS static data are transmitted every 6 minutes, and dynamic data depend on the ship speed and deviation angle. For dynamic data, the transmission intervals of class A equipment are 2 seconds, 3.33 seconds, 6 seconds, 10 seconds, and 180 seconds. The transmission intervals of class B equipment are 5 seconds, 15 seconds, 30 seconds, and 180 seconds. The received time of the next AIS packet is calculated as follows:

$$\mathrm{Tpredict} = T + \Delta T, \tag{10}$$

where T_{predict} is the predicted received time of the next packet, T is the received time of the current packet, and ΔT is the predicted interval time. The invalid time of the cache storage is set to twice the predicted AIS transmission interval, i.e., $2\Delta T$.

4.3.2. Processing of the Invalid Data in the Redis Cache. Before the predicted AIS data expire in the Redis cache, the data and validation time are updated according to the newly received AIS data. Invalid AIS data indicate that there is an AIS data loss. To ensure continuous ship trajectories, missing trajectories must be filled.

Most ships move at constant speeds. Thus, a linear system can be assumed. The Redis cache stores the navigation data received within ten minutes, and the ship speed and heading stored in the Redis cache are denoted as speed_{t1}, $\mathrm{speed}_{t2}, \ldots, \mathrm{speed}_{\mathrm{tn}}$ and $\mathrm{heading}_{t1}$, $\mathrm{heading}_{t2}, \ldots, \mathrm{heading}_{\mathrm{tn}}$. The acceleration and steering angle between adjacent AIS packets can be calculated. The current acceleration and steering angle $\mathrm{acce}_{\mathrm{pd}}$ and $\mathrm{turn}_{\mathrm{pd}}$ are predicted by linear regression. The ship speed and heading are calculated according to the current acceleration and steering angle and previous speed and heading, as follows:

$$\mathrm{speed}_{\mathrm{pd}} = \mathrm{speed}_{tn} + \mathrm{acce}_{\mathrm{pd}}, \tag{11}$$

$$\mathrm{HDG}_{\mathrm{pd}} = \mathrm{HDG}_{tn} + \mathrm{turn}_{\mathrm{pd}}. \tag{12}$$

The longitude and latitude of the ship are calculated according to equations (1)–(5). The predicted ship position, speed, and heading are sent to the Kafka message queue for the next calculation. The predicated data are continuously stored in the Redis cache, and the number of consecutive packet losses is determined.

4.4. AIS Data Statistics. In the data statistics step, a 5-minute time window is used to perform relevant statistics operations.

4.4.1. AIS and Ship Quantity Statistics. Data statistics for the grids, shore-based stations, and the system are executed, respectively. For grids, the number of AIS packets and ship density are calculated. For shore-based stations, the number of covered ships and actual capacity are calculated. For the system, the number of AIS packets and the AIS transmission time interval are calculated.

(1) Data Statistics for Grids. To calculate the grid density, a certain time point must be selected. The middle time point of the window is used as the basis of the calculations. The position at the middle time is calculated according to the positions at the adjacent times. If there are packets before and after the middle time, the two AIS packets closest to the middle time are used to calculate the ship position.

$$\begin{aligned} pos_{\mathrm{middle}} = \mathrm{pos}_{\mathrm{before}} \times \frac{t_{\mathrm{middle}} - t_{\mathrm{before}}}{t_{\mathrm{after}} - t_{\mathrm{before}}} + pos_{\mathrm{after}} \\ \times \frac{t_{\mathrm{after}} - t_{\mathrm{middle}}}{t_{\mathrm{after}} - t_{\mathrm{before}}}, \end{aligned} \tag{13}$$

where $\text{pos}_{\text{middle}}$ is the ship position at the middle time, t_{middle} is the middle time, t_{after} and t_{before} are the adjacent receiving times after and before the middle time, respectively, and $\text{pos}_{\text{before}}$ and $\text{pos}_{\text{after}}$ are the corresponding positions.

If there are packets only before or only after the middle time, the ship position at the middle time can be deduced according to position of the closest AIS packet according to equations (1)–(5). The number of ships in the grid is determined according to the grid where the ships are located.

(2) Data Statistics for Shore-Based Stations. The number of ships covered by the shore-based station is calculated. Similar to the data statistics for grids, the middle time in the window is used as the basis for the calculations. The distance between the ship and the shore-based station is calculated as follows:

$$\begin{aligned} d = R \times \arccos\left(\cos\left(\text{lat}_b\right) \times \cos\left(\text{lat}_s\right) \times \cos\left(\text{long}_b - \text{long}_s\right) \right.\\ \left. + \sin\left(\text{lat}_b\right) \times \sin\left(\text{lat}_s\right)\right), \end{aligned} \tag{14}$$

where d is the distance between the ship and the shore-based station, R is the radius of the earth, lat_s and long_s are the latitude and longitude of the ship, and lat_b and long_b are the latitude and longitude of the shore-based station.

If the distance is less than the radius covered by the shore-based station, the ship is covered by the shore-based station, and the number of covered ships is accumulated. The AIS divides one minute into 2250 time slots and transmits a complete position report message in each time slot, which indicates the upper bound that the system can sustain, that is, the system capacity. The actual capacity is the time slot number that the system actually needs. Since the number of time slots required by each ship per minute differs, the time slots occupied by different ships are accumulated to determine the actual capacity of the shore-based station. The number of ships in a certain transmission time interval is denoted as M_i ($i = 1, 2, \ldots, 8$), and the transmission time interval is denoted as T_i ($i = 1, 2, \ldots, 8$) seconds. The actual capacity of the shore-based station is calculated as follows:

$$N = \sum_{i=1}^{8} \frac{60}{T_i} \times M_i. \tag{15}$$

The ratio between the actual capacity and the maximum capacity of the shore-based station is calculated. If the ratio is greater than 1, the shore-base station is overloaded. The number of time slot multiplexes increases as the actual capacity of the shore-based station increases. When the ratio reaches 500%, the time slot multiplexes are close to 100%, and almost every time slot is reused. The reuse of a time slot leads to blind communication and identification communication issues, eventually resulting in AIS data loss.

(3) Data Statistics for the System. The total number of AIS packets that the system receives is calculated. The actual number of ships in different transmission time intervals is also calculated.

4.4.2. Inaccurate AIS Data Statistics. Since inaccurate data are mainly caused by the ship itself, the inaccurate AIS data of each ship in each time window are calculated, including inaccurate ship positions, inaccurate navigation status, and MMSI reuse.

(1) Inaccurate Position. The number of AIS packets with inaccurate position is obtained by accumulating the AIS packets with inaccurate positions. The inaccurate position rate is calculated as follows:

$$R_{\text{pe}} = \frac{N_{\text{pe}}}{N}, \tag{16}$$

where N_{pe} is the number of AIS packets with inaccurate positions, and N is the total number of AIS packets.

(2) Inaccurate Navigation Status. The number of AIS packets with inaccurate navigation status is obtained by accumulating the AIS packets with inaccurate speed and heading. The inaccurate ship navigation status rate is calculated as follows:

$$R_{\text{nv}} = \frac{N_{\text{nv}}}{N}, \tag{17}$$

where N_{nv} is the number of AIS packets with inaccurate navigation status.

(3) Reusing MMSI. The statistics of reusing MMSI are calculated according to the number of ships with multiple trajectories.

4.4.3. AIS Loss Analysis. AIS packet losses can be divided into three categories: shore-based station factor, environmental factor, and ship factor.

(1) Shore-Based Station Factor. The cause of packet loss is analyzed according to the actual capacity of the shore-based station that is nearest to the ship. When the ratio between the actual capacity and the maximum capacity of the shore-based station is less than 0.8, essentially no time slot reuse occurs, and the shore-based station factor is excluded. If the ratio is greater than 0.8, the reason for AIS packet loss is determined according to the distance between the ship and the shore-based station as shown in Table 1.

(2) Environmental Factor. If the shore-based station factor can be excluded, the number of AIS packets in the grid with data loss is

Table 1: Analysis of packet losses caused by stations.

Distance	Region	Analysis of packet loss
0–0.3R	Protection zone	Shore-based station factor are excluded
0.3–0.5R	Identification zone	If the number within 0.5R exceeds the maximum capacity, identification communication issues occur. Packet loss is attributed to shore-based station factor
0.5-R	Cut-over zone	Blind communication occurs due to time slot reuse, and the packet loss is attributed to shore-based station factor

analyzed. If the packet loss ratio in the grid exceeds 10%, the data loss is attributed to the environmental factor.

(3) Ship Factor. A ship all has lost AIS packet in the current grid and other grids. This packet loss is attributed to the ship itself.

4.4.4. AIS Loss Statistics

(1) Statistics of Packet Loss Rate for Shore-Based Stations Factor. The packet loss rate of shore-based stations at the time window is calculated as follows:

$$R_{\text{base}} = \frac{Ne_{\text{base}}}{N_{\text{base}}}, \tag{18}$$

where N_{base} is the number of packets which should be received by the shore-based station, and Ne_{base} is the number of lost packets due to shore-based station factor.

(2) Statistics of Packet Loss Rate for Grid Factor. The packet loss rate of a grid at the time window is calculated as follows:

$$R_{\text{grid}} = \frac{Ne_{\text{grid}}}{N_{\text{grid}}}, \tag{19}$$

where N_{grid} is the number of packets which should be received in the grid, and Ne_{grid} is the number of lost packets due to the grid environmental factor.

(3) Statistics of Packet Loss Rate for Ship Factor. The packet loss rate of the ships is calculated as follows:

$$R_{\text{ship}} = \frac{Ne_{\text{ship}}}{N_{\text{ship}}}, \tag{20}$$

where N_{ship} is the number of AIS packets for the ship, and Ne_{ship} is the number of lost packets due to the ship-based factor.

5. System Operation Analysis

The proposed algorithm has been deployed in a maritime supervision system. It acquires, cleans, counts, and visualizes AIS data from 13 AIS shore-based stations, which locates in a section of the Yangtze River Basin, in real time. It processes about 12000 dynamic AIS packets every minute and runs 24 hours every day without interruption. In reference [14], AIS data performance indicators of five days from five shore-based stations are off-line counted, and the calculation time is not given. It is impossible to determine whether it can meet the requirements of real-time AIS data processing. In reference [18], 110623842 dynamic AIS data are systematically preprocessed, which improves the data quality. The experiments are based on off-line batch processing, and it is impossible to determine whether it can implement real-time computation. Compared with the algorithms in reference [14, 18], the proposed algorithm implements a real-time calculation in an actual production environment, which not only realizes data cleaning in real time but also realizes indicator analysis in real time. It can meet the requirements of maritime real-time supervision. The cleaned data can also provide computing basis for ship collision avoidance and trajectory tracking.

5.1. AIS Real-Time Statistics Overview. The real-time AIS statistical analysis results are displayed by the data visualization layer. Figure 5 shows the real-time AIS statistical map of a section of the Yangtze River. The number of ships, the average number of AIS packets received per second, the number of lost packets per second, and the number of packets with inaccurate AIS data per second can be seen. The relevant statistics results can be viewed by clicking AIS statistics, ship statistics, and abnormal AIS statistics. The base map is a satellite map that is implemented by the MapBox tool. The back end obtains the ship position data in real time from the data storage layer and transmits the data to the front end in GeoJson format. The front end visualizes the dynamic ship position with a green dot.

The AIS statistics at the minute and hour levels are shown in Figures 6 and 7, respectively. Figure 6 shows the number of AIS packets received per minute in 1 hour. Between 10 : 32 and 11 : 31 on a certain day, the average number of AIS packets received per minute is approximately 12000. The lowest number of packets received is 11469, and the highest is 130009. The number of packets received per minute does not change considerably throughout the hour.

Figure 7 shows the number of AIS packets received per hour over approximately 24 hours on a certain day. Fewer AIS packets are received at night than during the day, which is consistent with actual ship navigation conditions.

5.2. Single Ship Indicator Analysis. Taking a day in 2022 as an example, the packet loss rate of the single ship is counted, and the five ships with the highest and lowest packet loss rates are identified. Tables 2 and 3 show AIS analyses for the ships with the highest and lowest packet loss rates, respectively.

Various ships have considerably different packet loss rates, ranging from 0.181% to 46.39%. From a data perspective, the AIS data of ships with higher packet loss rates

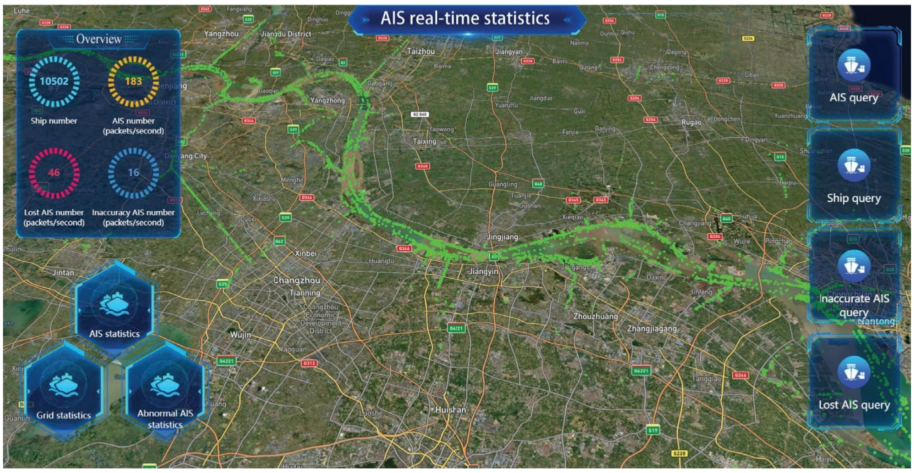

Figure 5: AIS real-time statistics overview.

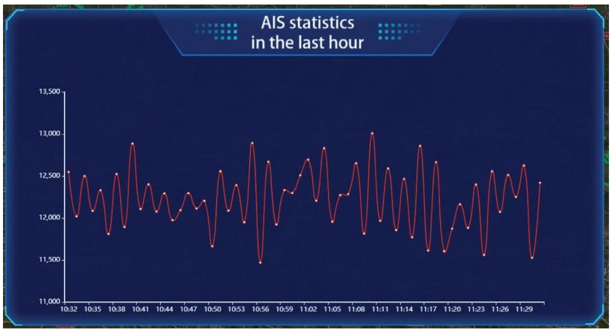

Figure 6: AIS statistics in the last hour.

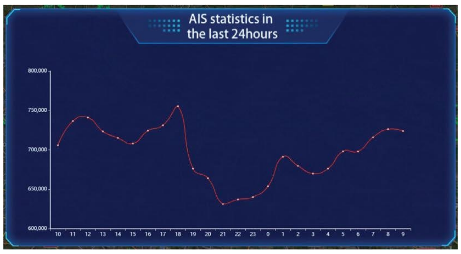

Figure 7: AIS statistics in the last 24 hours.

Table 2: Five ships with the highest packet loss rate.

MMSI	Loss rate (%)	Inaccuracy rate (%)
413***388	46.39	2.31
413***327	42.85	1.46
413***068	41.66	1.58
413***285	41.26	4.35
413***256	41.26	1.45

are more inaccurate, and this relationship should be studied further.

For lost AIS packets, the traditional means is to utilize the position of the last AIS packet or to exclude the ship. Water traffic differs from land traffic, and the ship navigation is more stable. Thus, the trajectory within a certain period of time can be repaired by the shipping status. AIS repair is helpful in maritime applications for tracking complete trajectories and can update ship positions in real time. Figure 8 shows the comparison between unrepaired and repaired trajectories from the ship with a MMSI of 413***388.

Table 3: Five ships with the lowest packet loss rate.

MMSI	Loss rate (%)	Inaccuracy rate (%)
413***880	0.128	0.45
413***656	0.151	0.36
413***528	0.159	0.15
413***358	0.166	0.65
412***550	0.181	0.11

Figure 8(a) shows the ship trajectories based on the original AIS data. For the loss of AIS packets, the ship trajectories are not smooth enough, and some trajectories even pass over the land. The packet loss also makes the maritime administrators unable to track the ship position in real time. Figure 8(b) shows the repaired trajectories. It can be seen that the repaired trajectories are smoother. By trajectory repairing, it can also meet the requirements of real-time scdslffhip tracking.

Incorrect settings and stolen MMSI may result in muladstiple ships using the same MMSI. Ships with the same MMSI often appear in different areas. The MMSI reuse seriously affects the navigation safety of ships and the development of the shipping industry. If multiple trajectories with the same MMSI are not identified, the ship trajectory jumps between different areas, and maritime administrators cannot accurately track the ship trajectories. Figure 9 shows a multitrajectory map of two ships with the same MMSI. It can be seen that the same MMSI appears in two different areas in a short time, and the two areas are hundreds of kilometers apart. There must be a case of MMSI being reused.

The two trajectories with the same MMSI are recognized by the multiple trajectory abnormality identification step. The separated trajectories are shown in Figure 10. The separated trajectories are smoother and more consistent

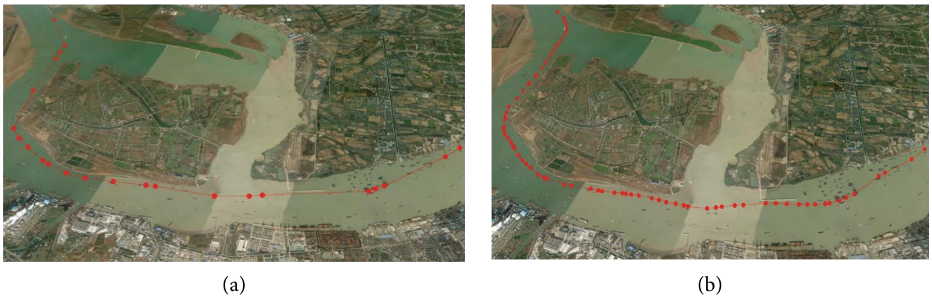

Figure 8: Comparison between the unrepaired and repaired trajectories: (a) the trajectory before AIS repairing and (b) the trajectory after AIS repairing.

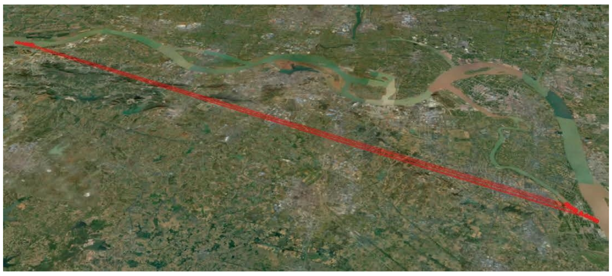

Figure 9: The ship trajectories jump.

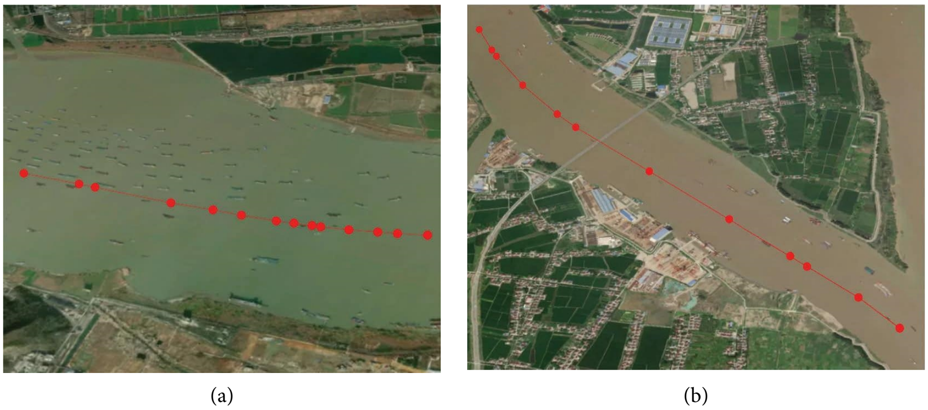

Figure 10: Multitrajectories identification and separation: (a) the first trajectory and (b) the second trajectory.

with the actual navigation situation. Multiple trajectory abnormalities are a considerable risk that may transmitted to maritime data centers. The data center can integrate CCTV to intelligently identify ships with stolen MMSI.

5.3. Grid Indicator Analysis. The grid indicators, which include the statistics of the ship number and the statistics of the AIS lost number in grids, are shown by heatmap. Heatmap is a representation method of data distribution by different colors. It can help maritime administrator to grasp the ship distribution, the AIS quality, and the water traffic flow in real time.

The statistics of the ship number are to count the ship number in each grid. The latest grid statistical results are stored in the Redis cache, and the historical statistical results are stored in the Doris cluster. Each grid statistic result includes the statistical time, longitude and latitude of the grid center, and ship number. Figure 11 shows the ship density heatmap at a certain time. The darker the color is, the higher the density is. Red has the highest density, followed by blue, and green has the lowest density. The red area is concentrated the downstream, demonstrating that the downstream ships are more densely distributed. The statistics of the proposed algorithm are based on the ship number, without considering the size of ships. In the future research, the density will be calculated according to the total ship number and the total ship size in each grid. It will better reflect the characteristics of water traffic.

The statistics of the AIS lost number are to count the AIS packet lost number in each grid. The storage means is same as the statistics of the ship number. Each grid statistic result

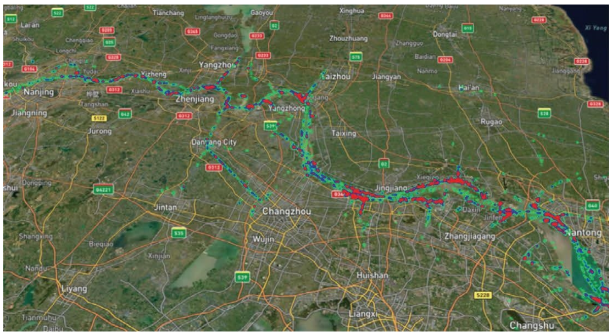

FIGURE 11: The heatmap of the ship density.

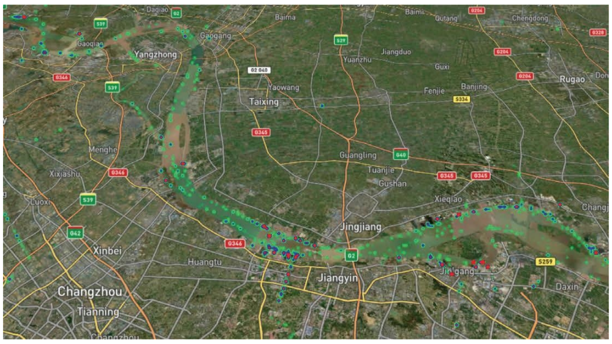

FIGURE 12: The density heatmap of AIS packet lost.

includes the statistical time, longitude and latitude of grid center, and AIS packet lost number. Figure 12 shows an AIS packet lost density heatmap. Serious packet loss occurs in some grids which are represented by red. This may be due to the influence of the shore-based station layout and regional geographical environment. The analysis results can help the maritime administrator grasp the AIS quality in real time. It also can be used as a reference for improving AIS shore-based station layouts.

6. Conclusion

To improve the quality of AIS data and analyze AIS data in real-time, a real-time AIS cleaning and indicator analysis algorithm based on stream computing is proposed. It includes data acquisition, data aggregation, data cleaning, data analysis, data storage, and data visualization steps. It is verified with AIS data from a certain area of the Yangtze River. Real-time deduplication is used to effectively improve the quality of the AIS data. The trajectory and navigation status are repaired, effectively improving the ship tracking. By identifying multiple trajectories, the problem of MMSI reuse is addressed, and the results provide a support for maritime supervision applications. Moreover, the ship packet loss rate and regional packet loss rate are analyzed. Thus, the causes of the ship packet loss could be effectively analyzed, providing a reference for maritime management applications and AIS shore-based station layouts.

The proposed algorithm cleans and analyzes AIS data in real time, effectively improving the quality of AIS data and providing a reference for maritime intelligent management applications. However, several aspects need to be studied further. (1) The algorithm analyzes dynamic AIS data but not static AIS data. Static AIS data are of great significance to ship traffic organization applications. Since static AIS data are entered manually, it is difficult to evaluate the reliability of these data. (2) AIS data can be used to identify abnormal ship behaviors based on real-time machine learning algorithms. Machine learning algorithms are widely used in AIS data processing and ship behavior analysis. Most traditional machine learning algorithms are based on off-line processing. The prerequisite for real-time analysis of ship abnormal behavior based on machine learning is to be able to complete real-time prediction of ship abnormal behavior and real-time model update by integrating new AIS data. In future studies, a real-time abnormality identification model that considers ship spacing, ship entry and exit, ship trajectory and other abnormalities should be developed. This model should use stream computing and a space-time clustering algorithm to meet the needs of real-time online warning in maritime supervision applications.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was financially supported by the Natural Science Foundation of the Jiangsu Higher Education Institutions of China (21KJB580007), the China Postdoctoral Science Foundation (2019M651844), the Jiangsu Postdoctoral Science Foundation (2018K035C), the Young Academic Leaders for QingLan Project of the Jiangsu Higher Education Institutions of China, the Excellent Scientific and Technological Innovation Team of the Jiangsu Higher Education Institutions of China (Maritime Big Data Team), and the Excellent Teaching Team for QingLan Project of the Jiangsu Higher Education Institutions of China (Big Data Technology Teaching Team with Shipping Characteristic).

References

[1] J. F. Zhang, X. J. Ren, H. H. Li, and Z. Yang, "Incorporation of deep kernel convolution into density clustering for shipping AIS data denoising and reconstruction," *Journal of Marine Science and Engineering*, vol. 10, no. 9, pp. 1319–1324, 2022.

[2] X. Q. Chen, J. Ling, S. Z. Wang, Y. Yang, L. Luo, and Y. Yan, "Ship detection from coastal surveillance videos via an ensemble Canny-Gaussian-morphology framework," *Journal of Navigation*, vol. 74, no. 6, pp. 1252–1266, 2021.

[3] R. W. Liu, Y. Guo, J. T. Nie et al., "Intelligent edge-enabled efficient multi-source data fusion for autonomous surface vehicles in maritime internet of things," *IEEE Transactions on Green Communications and Networking*, vol. 6, no. 3, pp. 1574–1587, 2022.

[4] R. Adland, H. Y. Jia, and S. P. Strandenes, “Are AIS-based trade volume estimates reliable? The case of crude oil exports,” *Maritime Policy and Management*, vol. 44, no. 5, pp. 657–665, 2017.

[5] G. N. Xiao, T. Wang, X. Q. Chen, and L. Zhou, “Evaluation of ship pollutant emissions in the ports of los angeles and long beach,” *Journal of Marine Science and Engineering*, vol. 10, no. 9, pp. 1206–1224, 2022.

[6] R. W. Liu, J. Nie, S. Garg, Z. Xiong, Y. Zhang, and M. S. Hossain, “Data-driven trajectory quality improvement for promoting intelligent vessel traffic services in 6g-enabled maritime IOT systems,” *IEEE Internet of Things Journal*, vol. 8, no. 7, pp. 5374–5385, 2021.

[7] B. Murray and L. P. Perera, “A dual linear autoencoder approach for vessel trajectory prediction using historical AIS data,” *Ocean Engineering*, vol. 209, Article ID 107478, 2020.

[8] W. He, J. Y. Lei, X. M. Chu, S. Xie, C. Zhong, and Z. Li, “A visual analysis approach to understand and explore quality problems of AIS data,” *Journal of Marine Science and Engineering*, vol. 9, no. 2, pp. 1–8, Article ID 198, 2021.

[9] D. Yang, L. Wu, S. Wang, H. Y. Jia, and K. X. Li, “How big data enriches maritime research–a critical review of Automatic Identification System (AIS) data applications,” *Transport Reviews*, vol. 39, no. 6, pp. 755–773, 2019.

[10] A. Felski, K. Jaskólski, and P. Banyś, “Comprehensive assessment of automatic identification system (AIS) data application to anti-collision manoeuvring,” *Journal of Navigation*, vol. 68, no. 4, pp. 697–717, 2015.

[11] N. J. Bailey, “Training, technology and AIS: looking beyond the box,” in *Proceedings of the Seafarers International Research Centre’s Fourth International Symposium*, pp. 108–128, Cardiff, UK, July 2005.

[12] P. Banyś, T. Noack, and S. Gewies, “Assessment of AIS vessel position report under the aspect of data reliability,” *Annual of Navigation*, vol. 19, no. 1, pp. 5–16, 2012.

[13] P. Kelly, “A novel technique to identify AIS transmissions from vessels which attempt to obscure their position by switching their AIS transponder from normal transmit power mode to low transmit power mode,” *Expert Systems with Applications*, vol. 202, pp. 117205–117212, 2022.

[14] L. Liu, Z. L. Jiang, and D. Y. Zhang, “Research on inland AIS service performance index and platform design,” *Port & Waterway Engineering*, vol. 75, no. 5, pp. 152–158, 2019.

[15] J. B. Zhang, S. F. Zhang, and J. P. Wang, “Pseudorange measurement method based on AIS signals,” *Sensors*, vol. 17, no. 5, pp. 1–20, Article ID 1183, 2017.

[16] W. Hasbi, M. Mukhayadi, M. Mukhayadi, and U. Renner, “The impact of space-based AIS antenna orientation on in-orbit AIS detection performance,” *Applied Sciences*, vol. 9, no. 16, pp. 3319–19, 2019.

[17] K. Jaskólski, L. Marchel, A. Felski, M. Jaskólski, and M. Specht, “Automatic identification system (AIS) dynamic data integrity monitoring and trajectory tracking based on the simultaneous localization and mapping (SLAM) process model,” *Sensors*, vol. 21, no. 24, pp. 1–19, Article ID 8430, 2021.

[18] L. B. Zhao, G. Y. Shi, and J. X. Yang, “Ship trajectories preprocessing based on AIS data,” *Journal of Navigation*, vol. 71, no. 5, pp. 1210–1230, 2018.

[19] G. Siegert, P. Banyś, and C. S. Martínez, “EKF based trajectory tracking and integrity monitoring of AIS data,” in *Proceedings of the 2016 IEEE/ION Position, Location and Navigation Symposium (PLANS)*, pp. 887–897, Savannah, GA, USA, April 2016.

[20] S. G. Chen, Y. K. Huang, and W. Lu, “Anomaly detection and restoration for ais raw data,” *Wireless Communications and Mobile Computing*, vol. 2022, Article ID 5954483, 11 pages, 2022.

[21] L. Sang, A. Wall, Z. Mao, X. Yan, and J. Wang, “A novel method for restoring the trajectory of the inland waterway ship by using ais data,” *Ocean Engineering*, vol. 110, pp. 183–194, 2015.

[22] D. W. Gao, Y. Zhu, J. Zhang, Y. He, K. Yan, and B. Yan, “A novel MP-LSTM method for ship trajectory prediction based on AIS data,” *Ocean Engineering*, vol. 228, Article ID 108956, pp. 1–16, 2021.

[23] F. Ma, X. P. Yan, and X. P. Chu, “Correlation between signal failure and field strength in automatic identify system,” *Journal of Dalian Maritime University*, vol. 37, no. 3, pp. 111–114, 2011.

[24] C. Liu, M. Z. Cao, and F. Han, “A model for fuzzy data correlation of AIS and radar,” *International Journal on Smart Sensing and Intelligent Systems*, vol. 5, no. 4, pp. 843–858, 2012.

[25] C. Liu, “Study of shore-based AIS network link capacity,” in *Proceedings of the 2010 2nd International Conference on Signal Processing Systems*, pp. 263–267, Dalian, China, July 2010.

[26] Q. Hu, X. Zhang, and S. Zhang, “A remote automatic test system with high precision for AIS performance,” in *Proceedings of the 2016 International Conference on Applied Mathematics, Simulation and Modelling*, pp. 223–227, Beijing, China, May 2016.

[27] H. Kuschel, “VHF/UHF radar. Part 2: operational aspects and applications,” *Electronics & Communication Engineering Journal*, vol. 14, no. 3, pp. 101–111, 2002.

[28] Y. Du, M. Chowdhury, and M. Rahman, “A distributed message delivery infrastructure for connected vehicle technology applications,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 99, pp. 1–15, 2017.

[29] T. Z. Lv, J. Zhang, Y. Y. Chen, and C. Y. He, “A real-time AIS data computing platform based on Flink,” in *Proceedings of the 2021 IEEE 3rd International Conference on Frontiers Technology of Information and Computer (ICFTIC)*, pp. 378–381, Greenville, SC, USA, November 2021.

[30] B. P. Zen, “The concept of big data analysis for maritime information on Indonesian waters using K-Means algorithm,” *INISTA: Journal of Informatics, Information System, Software Engineering and Applications*, vol. 3, no. 2, pp. 43–52, 2021.

[31] P. Petrov, G. Dimitrov, and O. Bychkov, “Real time big data analysis by using Apache Kudu and NoSQL Redis in web applications,” *Izvestia Journal of the Union of Scientists-Varna. Economic Sciences Series*, vol. 9, no. 1, pp. 26–34, 2020.

[32] Y. Z. Hao, P. J. Zheng, and Z. Han, “Automatic generation of water route based on AIS big data and ECDIS,” *Arabian Journal of Geosciences*, vol. 14, no. 6, pp. 533–538, 2021.