http://dx.doi.org/10.5755/j02.eie.27661 ELEKTRONIKA IR ELEKTROTECHNIKA, ISSN 1392-1215, VOL. 27, NO. 1, 2021

A Novel Fitting Model for Practical AIS Abnormal Data Repair in Inland River

Wei He[1, 2, 3], Xinlong Liu[1, 2, 3, *], Xiumin Chu[1, 2, 4], Zhiyuan Wang[1, 2], Pawel Fracz[5], Zhixiong Li[6]
[1]*School of Physics and Electronic Information Engineering, Minjiang University,*
Fuzhou 350108, China
[2]*Fujian Engineering Research Center of Safety Control for Ship Intelligent Navigation,*
Minjiang University,
Fuzhou 350108, China
[3]*Fujian Provincial Key Laboratory of Information Processing and Intelligent Control,*
Minjiang University,
Fuzhou 350108, China
[4]*National Engineering Research Center for Water Transport Safety, Wuhan University of Technology,*
Wuhan 430063, China
[5]*Department of Manufacturing Engineering and Automation Products, Opole University of Technology,*
45758 Opole, Poland
[6]*Yonsei Frontier Lab, Yonsei University,*
50 Yonsei-ro, Seodaemun-gu, Seoul 03722, Republic of Korea
liuxl@mju.edu.cn

***Abstract*—Affected by the environment of inland waterway, an Automatic Identification System (AIS) collects lots of abnormal data, which significantly reduces the inland river navigation performance using AIS data. To this end, this paper aims to restore the AIS data by repairing the lost data points. By analysing enormous abnormal AIS data, the abnormal data were firstly divided into three types, i.e., the erroneous data, short-time lost data, and long-time lost data. Then, a cubic spline interpolation method was employed to deal with the erroneous data and short-time lost data. Meanwhile, a least square support vector machine method was utilized to repair the long-time lost data. Finally, field experiments were carried out to validate the applicability of the proposed method, and it is shown that the fitting model can repair the AIS data with an accuracy of more than 90 %.**

***Index Terms*—Inland waterway; AIS data; Abnormal data; Data repair; Least squares support vector machine.**

I. INTRODUCTION

Intelligentization is the goal of inland river navigation, and acquiring inland river vessel traffic flow information is an important support to realize the goal of inland river intelligent shipping. The peculiarities of obtaining complete information on the ship traffic flow have attracted increasing attention from all walks of life. Automatic Identification System (AIS) is the main way to obtain real-time ship dynamic information in inland rivers. Compared with the coastal AIS network, the inland AIS network has its own characteristics. On the one hand, inland river navigation environment is generally in the area with dense water network and the main road of the river, and the channel is curved and the terrain is complex. The dense area of water network is mainly plain, and there are many hydraulic structures such as Bridges, so there is some shadowing phenomenon. At the same time, some inland waterway trunk lines are mostly located in mountainous areas, so it is difficult for AIS base station to achieve the coverage of the whole river. On the other hand, because the AIS does not have a complete information verification mechanism, there are many abnormal data in the actual application of AIS. Therefore, according to the characteristics of abnormal data of AIS and specific repair requirements, the repair process of AIS abnormal data in inland river environment is investigated, and a novel repair model is developed. Experimental evaluation demonstrates that the proposed repair model is able to improve the completeness of AIS data for intelligent shipping.

II. RELATED WORK

The automatic identification system is composed of shore-based and shipborne equipment. It is a new type of modern digital navigation aids systems. AIS adopts a series of information technologies, including data communication technology, information display technology, computer technology, network technology, etc. Its main application fields include ship collision avoidance, maritime management, and enhanced Vessel Traffic Services (VTS), etc.

In the aspect of AIS data transmission, the characteristics of inland river AIS communication link and the packet error rate of the receiver are mainly studied. In [1], the

Manuscript received 13 September, 2020; accepted 18 January, 2021.
This research was supported under the National Key Research Plan under Grant No. 2018YFB1600404, by Fujian Province Natural Science Foundation under Grant No. 2020J01860, and under University Industry Cooperation Program by the Department of Science and Technology of Fujian Province under Grant No. 2019H6018.

relationship between AIS message length and packet error rate is analysed, and the corresponding packet error rate prediction model is given. On this basis, the limit capacity of AIS base station to send short messages to ships under crowded waterway is proposed, and the consistent effect is obtained in the actual verification. Chu, Liu, Ma, Liu, and Zhong [2] optimized the parameters of Okumura-Hata model by using linear regression method, and the signal field distribution characteristics of AIS communication system in mountainous channel were studied by using Okumura-Hata model. Hu, Cao, Gao, Xu, and Song [3] aimed at solving the problem of AIS false alarm rate, and others studied a false alarm rate measurement system based on very high-frequency (VHF) fault diagnosis.

In terms of AIS message parsing, many scholars have conducted researches from the perspectives of improving the real-time and efficiency of AIS message parsing and how to make full use of AIS messages. Goudossis and Katsikas [4] introduce the Identity-Based Public Cryptography and Symmetric Cryptography to enhance the security properties of the AIS. Li and Yang [5] introduced the static and dynamic information of the AIS in detail in the process of studying the AIS message information. They studied the analytical method of the AIS message and completed the analysis of the AIS data with the actual data. In order to reduce trajectory data storage space, Sun, Chen, Piao, and Zhang [6] added the sliding window to the classical SPM (scan-pick-move) algorithm to better compress vessel trajectory data regarding compression efficiency. Wang and Fang [7] used a large number of original AIS data, and others analysed the characteristics of AIS message with secret code and proposed an AIS information analysis method based on a secret code.

In the aspect of AIS data analysis, relevant scholars mainly studied the trajectory information in AIS packets. Gaglione *et al.* [8] proposed a Bayesian based method to integrate AIS data and oceanographic high-frequency surface-wave (HFSW) radars for multi-target tracking. Sang, Wall, Mao, Yan, and Wang [9] took advantage of the cubic spline interpolation algorithm to repair the ship's trajectory. In the verification and comparison of the actual trajectory in the bridge section, it is found that the trajectory repair effect is better. Zhang [10] thought that the characteristics of longitude and latitude variation of ship track point are analysed, and it is pointed out that the method of cubic exponential smoothing is more suitable for ship track prediction. Yan *et al.* [11] proposed a ship trip semantic object (STSO) based on method to extract ship traffic routes at sea using ship history AIS data. Iphar, Napoli, and Ray [12] introduces the quality dimensions of data that shall be used in a quality assessment of AIS messages. Wu *et al.* [13] summarized several types of AIS track anomalies based on the in-depth analysis of a large number of AIS data, and ship track anomalies are automatically detected according to the characteristics of each type. Chu *et al.* [14]–[16] aimed at solving the problems such as loss or error of ship AIS data. The preliminary repair or prediction of AIS data was realized by using the Piecewise cubic Hermite interpolation, and the back-propagation (BP) neural network training set and test set were established to carry out single point and continuous multi-point AIS data prediction. Zhong, Jiang, Chu, and Liu [17] thought that recursive neural network is used to restore the AIS data of inland river ships and solve the problem of missing AIS repair.

In conclusion, because of the complexity of the AIS system, related scholars have usually carried on the thorough research only for certain aspects, such as AIS data transmission, AIS message parsing, short-term abnormal AIS data restoration algorithm, and the application of AIS data. Domestic and foreign scholars have put forward method, and the model can better solve the corresponding problems [18]. The system has repaired of all kinds of abnormal AIS data caused by the limitations of the environment, but the equipment and the AIS system have not been in-depth. Therefore, this paper will first define abnormal AIS data in inland river environment, then propose restoration methods and evaluation models of various abnormal AIS data systems, and finally carry out empirical research.

III. DEFINITION OF ABNORMAL AIS DATA

The basic definition of abnormal data refers to system failure, data loss, and data integrity destruction. In terms of inland river traffic flow, AIS system is an important way to obtain real-time ship traffic flow data in the inland river. However, due to the fact that inland waterway is dominated by natural waterway, which passes through mountainous areas, basins, and other regions with complex topography in the form of ribbon, and with the increasing number and tonnage of inland river vessels, the performance of AIS system of inland river has a great decline compared with that of open sea, which is mainly reflected in the following aspects:

- Reduction of the effective coverage of the AIS system. As the AIS signal is transmitted in a straight line by radio waves, it belongs to the VHF transmission mode, which is prone to interference in the transmission process, and its effective coverage is limited. However, there is much interference, such as coastal mountains, bridges, and urban high-rise buildings in inland waterways, which are easy to cause the attenuation of AIS signals and the reduction of effective coverage range, and eventually lead to errors or even data loss in the original AIS data.
- AIS channel capacity is insufficient. The AIS design capacity cannot meet the communication needs of increasingly crowded inland rivers.
- Abnormal AIS equipment causes error information. Because the shipborne AIS in the inland river comes from different manufacturers, and different manufacturers have different technological and technical levels, there is a problem of poor performance of positioning equipment in the inland river application, and a large number of wrong location data appear.

Therefore, the abnormal data of AIS are defined as follows: due to the influence of the abnormalities of the AIS equipment itself, the transmission characteristics of VHF and environmental factors, the collected AIS data are lost and wrong, which affects the completeness of the AIS data. According to the performance characteristics of abnormal data, the abnormal data of AIS are divided into three categories: AIS error data, AIS loss data in a short period of time, and AIS loss data in a long period of time.

A. AIS Error Data

The longitude, latitude, speed, and course of AIS dynamic data have certain value range under inland river conditions. In the main channel of the Yangtze river (e.g., between longitude 90 ° ~ 122 °E and latitude 24 ° ~ 35 °N, and heading 0 ° ~ 360 °), the ships' longitude and latitude values will change at different sailing speeds. Figs. 1–3 show the ships' velocity profile of the Wuqiao Bridge at the Yangtze River, where the most of the velocity values exceed the normal range. As a result, these AIS data can be regarded as error data.

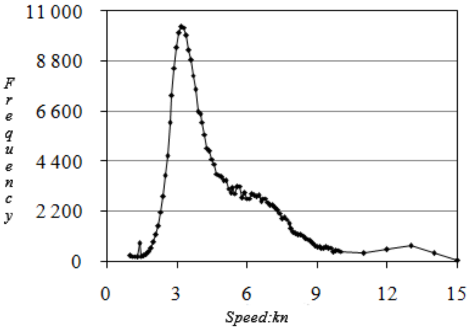

Fig. 1. Velocity distribution of ships in Wuqiao Bridge section of the Yangtze River.

B. AIS Data Loss in Short Time and Data Loss in Long Time

According to Table I and Fig. 1, it can be seen that there are at least 5 data in 1 min for ships with type A shipborne AIS and at least 2 data in 1 min for ships with type B shipborne AIS. After statistics of 6,829 original AIS data (including 3,975 for Type A shipborne AIS and 2,854 for type B shipborne AIS), it is found that the number of data pieces for type A shipborne AIS in 1 min is mainly 1 ~ 5, with an average of 3.3 bars/min. The number of data bars for type B shipborne AIS in 1 min is mainly 1 ~ 2, with an average of 1.6 bars/min. The number of messages per minute distributed for type A and Type B shipborne AIS is shown in Fig. 2 and Fig. 3. The message loss rule is shown in Table II. At the same time, it is found by statistics that the loss time length of type A shipborne AIS is mainly concentrated within 3 minutes, accounting for 75 %. The length of time for data loss of Type B shipborne AIS is mainly concentrated within 5 minutes, accounting for 84 %. Therefore, considering the accuracy of data repair and the time spent, the short-time data loss of type A shipborne AIS is defined as the case where the data number is less than 3 within 1 min and there is no data within 3 min, and the long-time data loss is defined as no data beyond 3 min. Similarly, the short-time data loss of type B shipborne AIS is defined as the situation where the data number is less than 2 within 1 min and there is no data within 5 min, and the long-time data loss is defined as no data over 5 min.

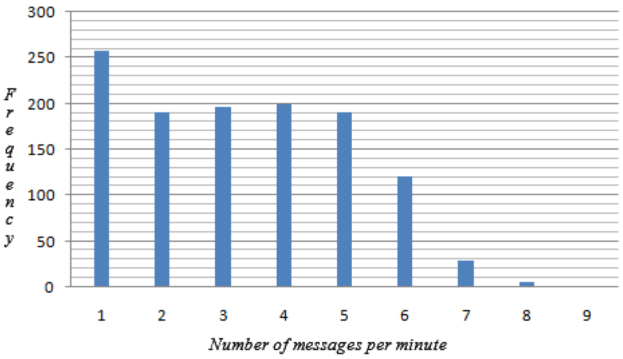

Fig. 2. Number of messages per minute on Type A shipborne AIS.

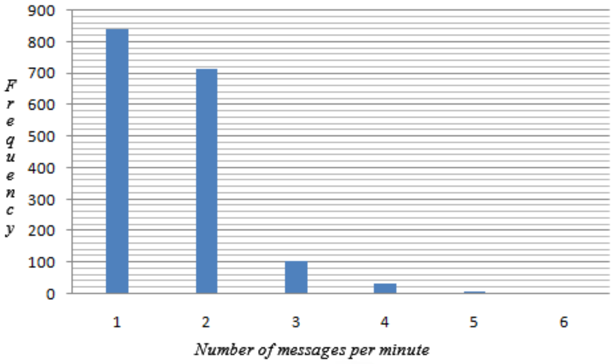

Fig. 3. The number of messages per minute of Type B shipborne AIS.

TABLE I. TYPICAL AIS ERROR DATA.

MMSI	ARCHIVE TIME	LON	LAT	SPEED/kn	COURSE	Note
100902071	2015/7/22 12:35:36	447.3924	223.6960	18.9	329.6	Latitude and longitude data are out of the normal range; Course data is out of normal range; The speed data is out of the normal range.
413768303	2015/8/11 2:37:00	394.3598	111.8481	0.5	0	
413769758	2015/8/12 18:06:15	273.0147	145.9764	14.4	27.7	
41371803	2015/6/24 10:04:48	263.6981	115.3988	2.5	52	
900028013	2015/8/8 17:18:43	394.3598	111.8481	2.5	0	
413593470	2015/8/14 6:33:10	114.3338	30.6354	0	369.5	
413593470	2015/8/15 13:31:33	114.3338	30.61419	0	409.5	
100900709	2015/8/14 15:13:05	114.1844	30.47083	31.3	92.2	
413000037	2015/8/11 20:25:06	114.3359	30.58347	25.6	1	

*Note: *MMSI - Maritime Mobile Service Identify; LON - longitude; LAT – latitude.*

TABLE II. RULES OF AIS DATA LOSS TIME.

Type A berth								
Lost time/min	1–2	2–3	3–4	4–5	5–6	6–7	7–10	> 10
Frequency	80	26	11	6	7	2	1	9
Type B building berth								
Lost time/min	1–2	2–3	3–4	4–5	5–6	6–7	7–10	> 10
Frequency	19	73	21	18	5	8	2	44

IV. RESTORATION PRINCIPLE OF ABNORMAL AIS DATA IN INLAND RIVER ENVIRONMENT

A. Abnormal AIS Data Repair Model Framework in Inland River Environment

The purpose of abnormal AIS data repair is to obtain relatively complete inland river AIS data, and the specific result is to obtain the correct value of the target ship's heading, speed, longitude and latitude at the corresponding time point. According to definition of error data, short-time loss data, and long-time loss data, the cubic spline interpolation method is used to repair the error data and short-time loss data, while the Least Square support vector machine (LSSVM) is used to repair the long-time lost data. The overall framework for abnormal AIS data repair is shown in Fig. 4.

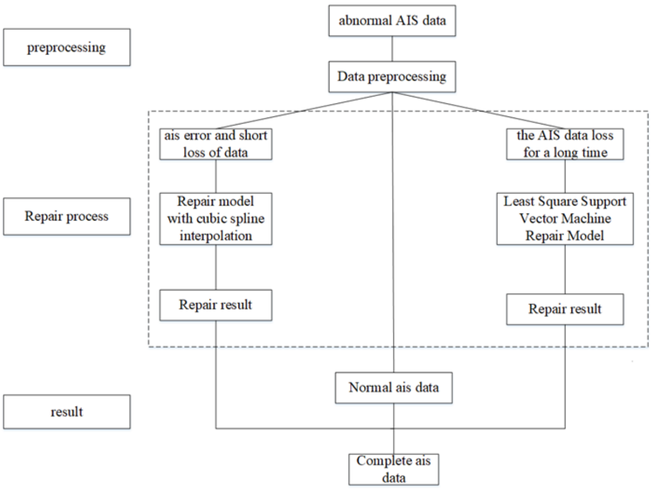
abnormal AIS data
preprocessing
Data preprocessing
ais error and short loss of data
the AIS data loss for a long time
Repair process
Repair model with cubic spline interpolation
Least Square Support Vector Machine Repair Model
Repair result
Repair result
Normal ais data
result
Complete ais data

Fig. 4. Abnormal AIS data repair model framework.

It is necessary to determine the repair time of abnormal data before repairing AIS data. Figure 5 shows the velocity distribution of the cargo ship and passenger ship in Wuqiao Bridge section of the Yangtze River Waterway, respectively. It can be obtained that the velocity of the cargo ship and passenger ship is basically 0–14 knots. According to the information update frequency of class A and B AIS, when the ship speed is 0–14 knots, the message transmission frequency of class A berth is 10 s, and that of class B berth is 30 s.

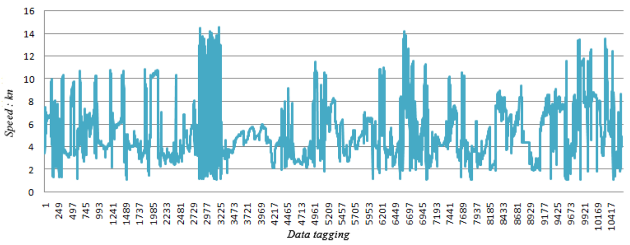

Fig. 5. Velocity distribution of freighter in Wuqiao Section of Yangtze River Waterway.

B. Cubic Spline Interpolation Model for AIS Data Repair

For AIS error and data loss in a short time, this paper will adopt cubic spline interpolation method to quickly restore, and its calculation model is as follows.

If there is a set of AIS data, then within the longitude range, the cubic spline interpolation function is:

$$s(x)=\sum_{i=1}^{m-1} s_i\left(x_i,y_i\right). \tag{1}$$

Among them: $i=1,2,\cdots,m-1s_i\left(x_iy_i\right)$ is a cubic polynomial function in the interval $\left[x_i,x_{i+1}\right]$.

If the piecewise cubic track function is, there are three unknowns in the function

$$si(xi,yi)=a_i\left(x-x_i\right)^3+b_i\left(x-x_i\right)^2+c_i\left(x-x_i\right)+y_i. \tag{2}$$

There is a total unknown in the three tracks of each segment $m-13(m-1)$. If you get an unknown, you get the curve track function $3(m-1)s(x)$.

For the cubic spline function, every point on the function is continuous, so the position point of AIS is continuous on the two piecewise cubic spline functions near it, i.e., the position of AIS also satisfies the cubic spline interpolation function in the previous section, namely

$$\begin{gathered}y_{i+1}=a_i\left(x_{i+1}-x_i\right)^3+b_i\left(x_{i+1}-x_i\right)^2+\\+c_i\left(x_{i+1}-x_i\right)+y_i,\quad i=1,2,\cdots,m-1.\end{gathered} \tag{3}$$

At the same time, as the intersection point of two adjacent piecewise cubic spline interpolation, the first derivative and the second derivative of the AIS position point at the intersection point are also continuous $s'(x)s''(x)$. To satisfy the condition that the first derivative and the second derivative are continuous, then:

$$\begin{cases}s_i^{'}\left(x_i\right)=s_{i+1}^{'}\left(x_i\right),\\ s_i^{''}\left(x_i\right)=s_{i+1}^{''}\left(x_i\right),\end{cases} i=1,2,\cdots,m-2. \tag{4}$$

There are also continuous first derivative and second derivative at two end points of the curve track. If the curve is connected to the straight track at the end point and the second derivative of the straight track is 0, then there are boundary conditions

$$s_1^{''}=0. \tag{5}$$

If the curve is connected with the circular track at the end point, there are boundary conditions

$$s_2^{''}=\frac{R_i^2}{\left(q_i-y_i\right)^2}. \tag{6}$$

Substituting (3), (4), (5), and (6) into (1), the curve function can be solved.

C. LSSVM Regression Algorithm for AIS Data Repair

The LSSVM is a new support vector machine (SVM) proposed by Suyken J.A.K. LSSVM uses the least square linear system as the loss function and replaces the quadratic programming method adopted by traditional support vector machine, simplifies the complexity of calculation, and improves the operation speed. Therefore, this paper adopts the LSSVM to repair the long-lost AIS data. The calculation process is as follows.

Given a set of N training samples, k = 1, 2, ... $\{x_k, y_k\},\ N.$ Each training sample includes n-dimensional input $x_k \in R^n$ and one-dimensional output $y_k \in R.$ According to statistical theory, the regression prediction problem can be described as the following optimization problem:

$$\begin{cases} \min\limits_{\omega,b,e} J_P(\omega,e) = \frac{1}{2}\omega^T\omega + \gamma\frac{1}{2}\sum\limits_{k=1}^{N} e_k^2, \\ y_k\left[\omega^T\varphi(x_k)+b\right] = 1 - e_k,\ k = 1, ..., N, \end{cases} \tag{7}$$

where $\varphi(g): R^n \to R^m$ is the kernel space mapping function, $\omega \in R^m$ is the weight vector, $e_k \in R$ is the error variable, b is the deviation quantity, and γ is the positive regularized parameter. The kernel function $\varphi(g)$ can map the sample in the original space to a vector in the high dimensional eigenspace and solve the problem of linear inseparability. Generally, Lagrange method is used to solve this optimization problem

$$\Gamma(\omega,b,e;\alpha) = J_P(\omega,e) - \\ -\sum_{k=1}^{N}\alpha_k\left\{y_k\left[\omega^T\varphi(x_k)+b\right]-1+e_k\right\}, \tag{8}$$

where the Lagrang multiplier Γ is available by:

$$\begin{cases} \frac{\delta\Gamma}{\delta\omega} = 0 \to \omega = \sum\limits_{k=1}^{N}\alpha_k\varphi(x_k), \\ \frac{\delta\Gamma}{\delta b} = 0 \to \sum\limits_{k=1}^{N}\alpha_k = 0, \\ \frac{\delta\Gamma}{\delta e_k} = 0 \to \alpha_k = \gamma e_k, \\ \frac{\delta\Gamma}{\delta\alpha_k} = 0 \to \omega^T\varphi(x_k)+b+e_k-y_k = 0. \end{cases} \tag{9}$$

This transforms the optimization problem into a linear solution problem, i.e., to solve the following problems:

$$\begin{pmatrix} 0 & 1^T \\ 1 & K+\frac{1}{\gamma}I \end{pmatrix}\begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix}, \tag{10}$$

where 1 is the matrix with 1 for each element, which meets the Mercer criterion, I is the identity matrix, and $K = K(x_i, x_j) = \left(\varphi(x_i)\times\varphi(x_j)\right)_{i,j=1}^{n}.$ The solution to equation (10) is the following regression estimation function

$$g(x) = \sum_{i=1}^{n}\alpha_i\left(\varphi(x_i)\times\varphi(x)\right)+b. \tag{11}$$

D. AIS Data Repair Evaluation Model

The evaluation index of the model is an important part of the model, which is used to detect the pros and cons of the repair results of the model AIS data, i.e., whether the repair results of the model are within the acceptable range. In this paper, the square of the correlation coefficients and mean square error (MSE) are used as evaluation indexes of AIS abnormal data repair model. MSE is the expected value of the squared difference between the corrected value and the true value of the data, and the correlation coefficient is a statistical indicator to reflect the degree of correlation between variables MSE and R2. The calculation formula is as follows.

Mean square error

$$MSE = \frac{1}{l}\sum_{i=1}^{l}(f(x_i)-y_i)^2. \tag{12}$$

Correlation coefficient

$$\mathrm{R2} = \frac{(l\sum\limits_{i=1}^{l}f(x_i)y_i - \sum\limits_{i=1}^{l}f(x_i)y_i)^2}{(l\sum\limits_{i=1}^{l}f(x_i)^2-(\sum\limits_{i=1}^{l}f(x_i))^2)(l\sum\limits_{i=1}^{l}y_i^2-(\sum\limits_{i=1}^{l}y_i)^2)}. \tag{13}$$

In the evaluation index: the closer the repair value is to the true value, the smaller the MEAN square error (MSE) will be. It also indicates that the repair effect of the repair model is better and the generalization ability of the repair model is stronger. On the contrary, if the MSE is smaller, the repair effect and generalization ability of the model will be worse. As for the correlation coefficient R2, it indicates the degree of correlation between two variables. Usually, R2 is within 0 to 1 and R2 = 1 means perfect fit.

V. EMPIRICAL RESTORATION OF ABNORMAL AIS DATA IN INLAND RIVER ENVIRONMENT

A. Empirical Data Collection

Wuhan Bridge section is located in the middle reaches of the Yangtze River channel and is a typical bridge section. The original AIS data in this paper mainly come from the collection points of Baishazhou Bridge, Wuhan Yangtze River Bridge, and Tianxingzhou Bridge in this section (Fig. 6). Generally, ships sail all year round and their routes are fixed. Therefore, the time of passing through these three collection points is relatively regular and similar historical data can be easily obtained. To find similar historical data in convenience, this article used the AIS data from Changhang freight 0316 ship. This test ship collected each month's AIS data, including type A and Type B. The acquisition receiving point of empirical data is shown in Figs. 3–5.

In the repair of AIS data, it is necessary to pre-process the original AIS data. The specific process is as follows:

1. Judge the type of shipborne AIS of the ship in this article;
2. If the data point LON > 180 or latitude LAT > 90 or COURSE > 359.9 in the data of the AIS data sample is deleted and marked as the AIS error data point, it needs to

be fixed.

3. If the number of data per 1 min in the AIS data sample is less than the standard number of data (type A shipborne AIS is 3, type B shipborne AIS is 2), it is marked as the data point of AIS loss in a short time and needs to be repaired.

4. If there is no data in type A shipborne AIS data sample for from 1 to 3 minutes continuously or from 1 to 5 minutes in type B shipborne AIS data sample, it is marked as AIS lost data in a short time and needs to be repaired.

5. If there is no data in the type A shipborne AIS data sample for more than 3 min or in the type B shipborne AIS data sample for more than 5 min, it shall be marked as AIS data loss for A long time and needs to be repaired.

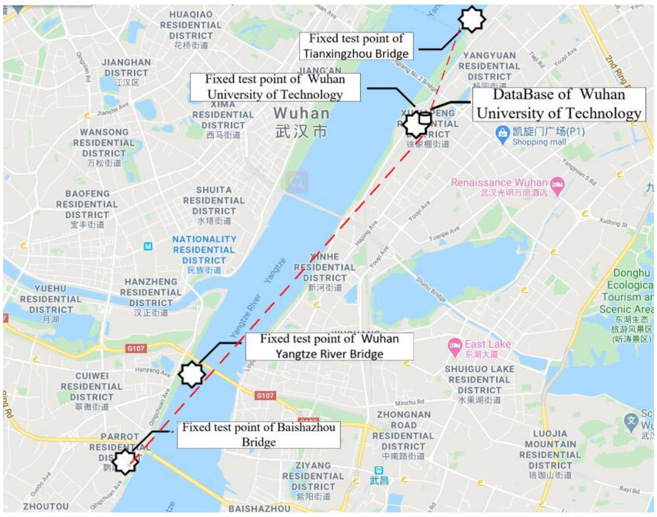
Fixed test point of Tianxingzhou Bridge
Fixed test point of Wuhan University of Technology
DataBase of Wuhan University of Technology
Fixed test point of Wuhan Yangtze River Bridge
Fixed test point of Baishazhou Bridge
Wuhan
武汉市
Yangtze River
HUAQIAO RESIDENTIAL DISTRICT
JIANGHAN DISTRICT
XIMA RESIDENTIAL DISTRICT
YANGYUAN RESIDENTIAL DISTRICT
WANSONG RESIDENTIAL DISTRICT
BAOFENG RESIDENTIAL DISTRICT
SHUITA RESIDENTIAL DISTRICT
NATIONALITY RESIDENTIAL DISTRICT
XINHE RESIDENTIAL DISTRICT
YUEHU RESIDENTIAL DISTRICT
HANZHENG RESIDENTIAL DISTRICT
CUIWEI RESIDENTIAL DISTRICT
SHUIGUO LAKE RESIDENTIAL DISTRICT
ZHONGNAN ROAD RESIDENTIAL DISTRICT
ZIYANG RESIDENTIAL DISTRICT
LUOJIA MOUNTAIN RESIDENTIAL DISTRICT
PARROT RESIDENTIAL DISTRICT
Renaissance Wuhan
East Lake
Donghu Ecological Tourism and Scenic Area
ZHOUTOU
BAISHAZHOU
G107

Fig. 6. Velocity distribution of freighter in Wuqiao Section of Yangtze River Waterway.

B. AIS Error Data Repair

The idea of AIS error data repair is to use interpolation method to find the interpolation after removing the error data. In order to facilitate the comparison between the repair value and the corresponding real value, this paper artificially sets some error data as the original data of data repair (see Table III). The article 11 in the table is typical error data.

Specific repair process:

1. Data pre-processing: 11 pieces of data except time in Table III were removed and marked as error data points.

2. Data input samples. Its sample is X = {1, 2, ..., 11}, Y = {LONi, or LATi, or SPEEDi, Or COURSEi}, I = 1, ..., 11.

3. Cubic spline interpolation repair.

4. Get the repair value (see Figs. 7–10).

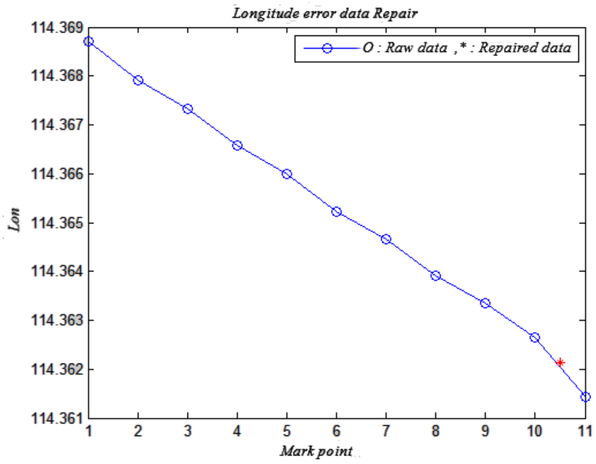

Fig. 7. Longitude error data Repair.

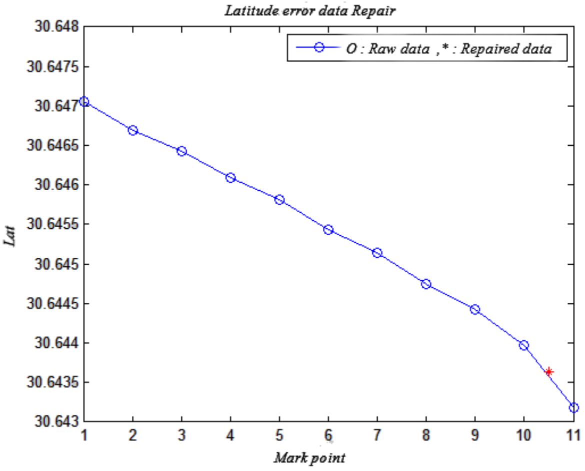

Fig. 8. Latitude error data Repair.

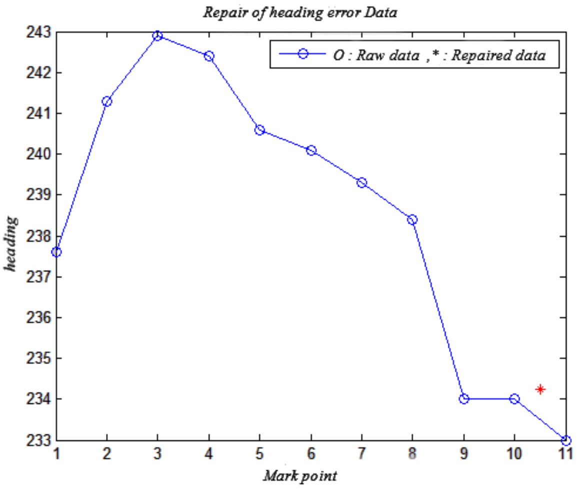

Fig. 9. Repair of heading error Data.

TABLE III. RAW DATA - ERROR DATA (DATED MAY 28, 2018).

The serial number	Time	Longitude (LON)	Latitude (LAT)	Speed (SPEED)	Course (COURSE)
1	15:30:12	114.368703	30.647057	4.8	237.6
2	15:30:46	114.367905	30.646690	4.8	241.3
3	15:31:12	114.367335	30.646433	4.8	242.9
4	15:31:46	114.366572	30.646088	4.8	242.4
5	15:32:13	114.365990	30.645817	4.8	240.6
6	15:32:45	114.365225	30.645435	4.8	240.1
7	15:33:12	114.36465	30.645145	4.8	239.3
8	15:33:45	114.363903	30.644753	4.8	238.4
9	15:34:13	114.363358	30.644428	4.8	234
10	15:34:46	114.362653	30.643977	4.8	234
11	15:35:16	180.362052	90.643587	34	360
12	15:35:47	114.361440	30.643187	4.8	233

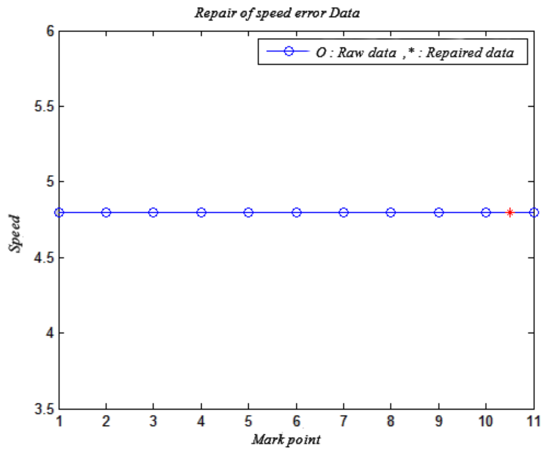

Fig. 10. Repair of speed error data.

Analysis of results: In the above repair, the pair of real and repair values is shown in Table IV.

TABLE IV. ANALYSIS OF ERROR DATA REPAIR RESULTS.

Data types	Longitude	Latitude	Speed	Course
The real value	114.362 052	30.643 587	4.9	234
Repair value	114.362 157	30.643 621	4.8	234.3
Accuracy, %	99.96	99.99	97.96	99.87

After a large number of data verification, the results show that the repair accuracy is more than 93 %. After the error data is repaired, the original data can be replaced with the repaired data.

C. *Repair of AIS Data Loss in a Short Time*

The idea of AIS short-time lost data repair is also to use interpolation method. Different from the wrong data, the short-time lost data often needs to obtain several consecutive interpolation points. Therefore, in order to obtain higher repair accuracy, the input sample is larger than the sample of the wrong data repair. Table V shows the raw data in the case of a short period of data loss. At least 4 data were lost between 08:04:11 and 08:06:52 in Table V.

Specific repair process:

1. Data pre-processing: mark from 8:04:11 to 08:06:52 in Table V as data loss in a short time.
2. Data input samples. Its sample is X ={1, 2, ..., 16}, Y = {LONi, or LATi, or SPEEDi, or COURSEi}, I = 1, ..., 16.
3. Cubic spline interpolation repair.
4. Get the repair value (see Figs. 11–14).
5. As for the time data of the missing point, it does not need to be very accurate, so it can be obtained by means of average value.

TABLE V. RAW DATA - DATA LOST FOR A SHORT TIME (DATE: AUGUST 26, 2017).

The serial number	Time	Longitude (LON)	Latitude (LAT)	Speed (SPEED)	Course (COURSE)
1	08:00:01	114.408 088	30.662 027	5.1	246.5
2	08:00:41	114.407 110	30.661 582	5	238.7
3	08:01:12	114.406 440	30.661 192	5	236.7
4	08:01:52	114.405 552	30.660 672	5	239.1
5	08:02:12	114.405 075	30.660 442	4.9	240.3
6	08:02:42	114.404 392	30.660 108	4.8	240
7	08:03:11	114.403 710	30.659 780	4.8	241.9
8	08:03:42	114.403 032	30.659 462	4.8	242.3
9	08:04:11	114.402 345	30.659 152	4.7	243.4
10	08:06:52	114.398 595	30.658 060	4.4	257.2
11	08:07:11	114.398 128	30.657 968	4.4	258.6
12	08:07:41	114.397 428	30.657 845	4.4	259
13	08:08:11	114.396 718	30.657 713	4.5	258.3
14	08:08:52	114.395 760	30.657 533	4.5	257.5
15	08:09:12	114.395 263	30.657 437	4.5	256.6
16	08:09:51	114.394 310	30.657 247	4.6	255.3

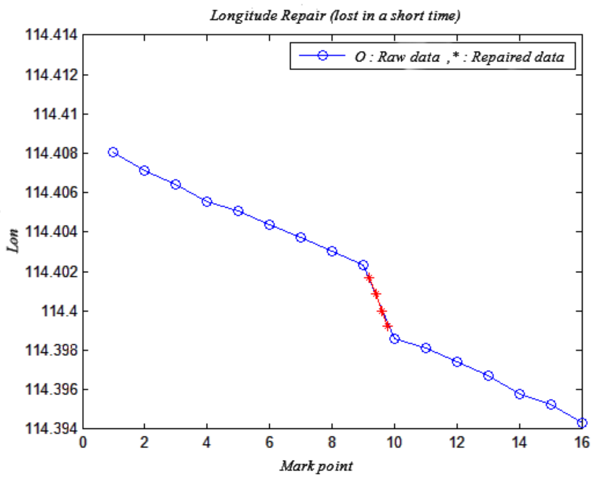

Fig. 11. Longitude Repair (lost in a short time).

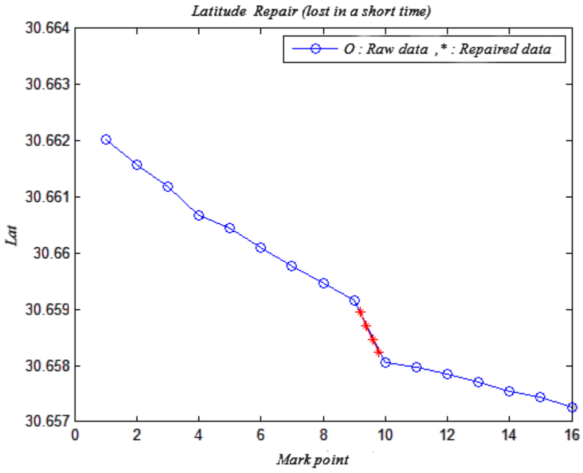

Fig. 12. Latitude repair (lost in a short time); (Repair values: 114.401 721, 114.400 934 (Repair values: 30.659 012, 30.658 724, 114.400 012, 114.399 214) 30.658 532, 30.658 245).

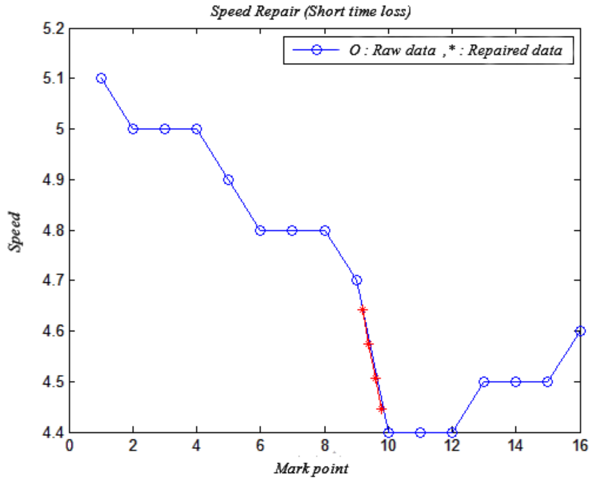

Fig. 13. Speed Repair (short-time loss).

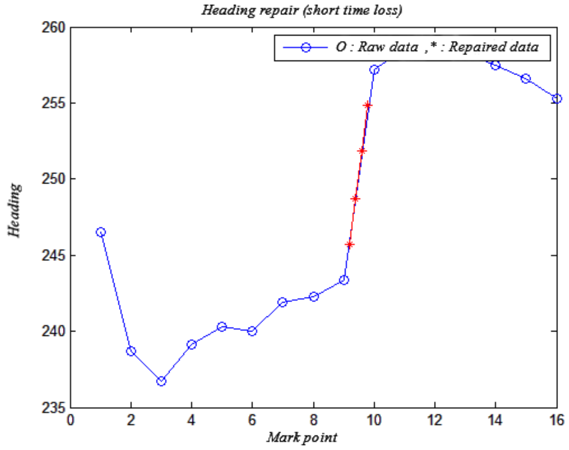

Fig. 14. Heading repair (short-time loss); (Repair values: 4.6, 4.6, (Repair values: 245.7, 248.7, 4.5, 4.4) 251.9, 254.9).

TABLE VI. COMPARISON OF SHORT-TERM LOST DATA REPAIR RESULTS.

A data item	Time	Longitude	Latitude	Speed	Course
The real value	08:04:52	114.401 408	30.658 800	4.6	251.2
Repair value	08:04:51	114.401 721	30.659 012	4.6	245.7
The real value	08:05:11	114.400 935	30.658 660	4.6	252.5
Repair value	08:05:11	114.400 934	30.658 724	4.6	248.7
The real value	08:05:52	114.400 005	30.658 388	4.5	253.2
Repair value	08:05:32	114.400 012	30.658 532	4.5	251.9
The real value	08:06:11	114.399 532	30.658 258	4.5	249.9
Repair value	08:06:12	114.399 214	30.658 245	4.4	254.9

According to the analysis of results, it can be seen from Table VI that the repair value of the data lost in a short time is close to the real value. After calculation, the accuracy of the repair value is above 90 %, and the time point of the repair data is not much different from the real time point. After the fix value is obtained, the fix value can be inserted into the fix marker in the raw data.

D. AIS Loss Data Repair for a Long Time

The repair idea of AIS lost data for a long time is to use LSSVM regression fitting method and similar historical data to get the repair value. The core idea of using historical data to repair lost data assumes of functional relationship between lost data and historical data. So, the key to repair lost data is to find a function that can describe the relationship between lost data and historical data. Therefore, the problem of using historical data to repair lost data is essentially the same as that of function fitting. The loss data of AIS data is selected for repair. According to the time series variation rule of ship traffic flow, the AIS data of a ship are inevitably related to the AIS data in a similar navigation environment, which is because ships will adopt similar sailing mode in a similar environment.

The key to using LSSVM model to repair long-lost AIS data is to adopt appropriate parameter optimization method and utilize the historical data closest to the data to be repaired for training and testing. In Section IV, similar historical data are defined and compared with common parameter optimization methods. In this paper, the parameter optimization method of LSSVM model is particle swarm optimization (PSO) algorithm. In order to verify the result of data repair by comparing the repaired value with the real value, the following number of the data to be repaired is selected artificially.

Data to be repaired:

1. A section of data passed by the data receiving point on May 16, 2018 during the descending process of Long Line 0316. Starting time: from 2018-05-16 10:33:45 to 2018-05-16 10:50:10; longitude range: from 114.369, 315 to 114.397677; latitude range: from 30.603 428 to 30.628655.

2. A section of data passed by the data receiving point of Long Haul Terminal 0316 on September 24, 2017 during its uplink. Starting from 2017-09-24 11:02:15 to 2017-09-24 11:30:46, it ranges from 114.500 593 to 114.4666 323 in longitude and from 30.685 915 to 30.678 190 in latitude.

The corresponding training data are:

1. In the original AIS database, the data containing the latitude and longitude ranges of the data to be repaired are similar to the downstream data on May 16, 2018.

2. In the original AIS database, the data containing the latitude and longitude ranges of the data to be repaired are similar to the upstream data on September 24, 2017, which are 2017-08-26 and 2017-07-22.

The specific repair process is:

1. Data pre-processing: According to the definition of AIS lost data for a long time in Section II, find the data to be repaired.

2. Model parameter optimization: standard PSO algorithm is adopted for optimization.

3. LSSVM model training.

4. Repair data.

5. Calculate evaluation indexes.

Figures from 15 to 18, respectively, show the repair results of up-linking lost data on May 16, 2018, in which the blue circle represents the true value and the red star represents the predicted value. In order to improve the accuracy of the prediction results, the selected training samples are all larger than the latitude and longitude range of the data to be repaired, so the repair data are only a part of the predicted value.

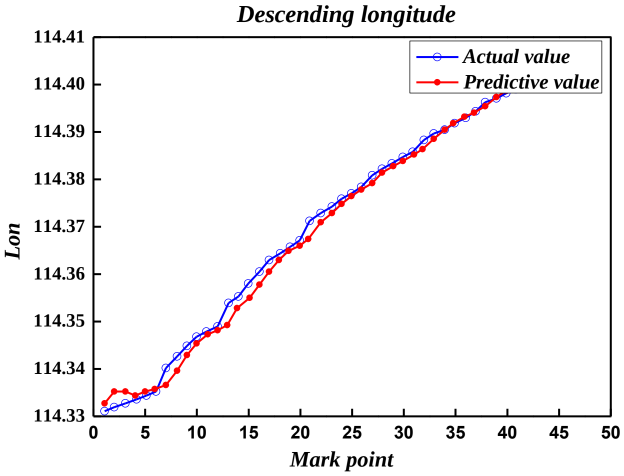

Fig. 15. Heading repair (long-time loss).

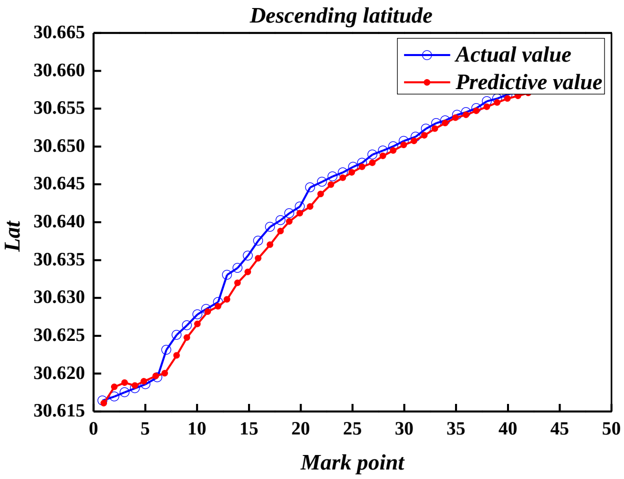

Fig. 16. Heading repair (long-time loss).

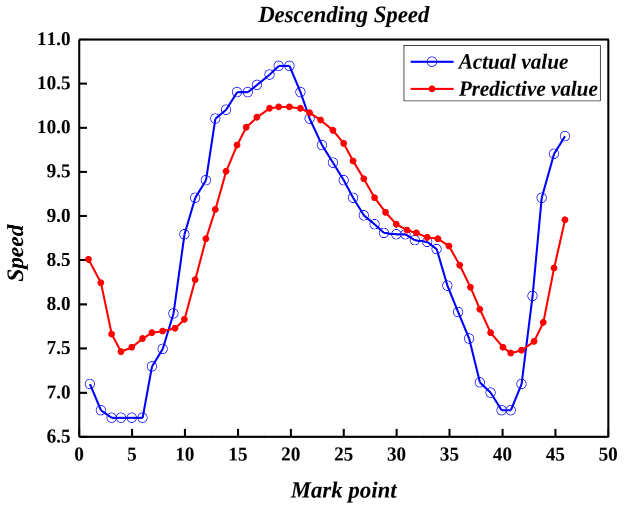

Fig. 17. Heading repair (long-time loss).

Figures from 19 to 22 show the repair results of the lost data on September 24, 2017, respectively, and the repair value is only a part of the predicted value. R2 is correlation coefficient, MSE is mean square error, and both of them are evaluation indexes, where R2, the closer to 1, while the closer MSE is to 0, the better the repair.

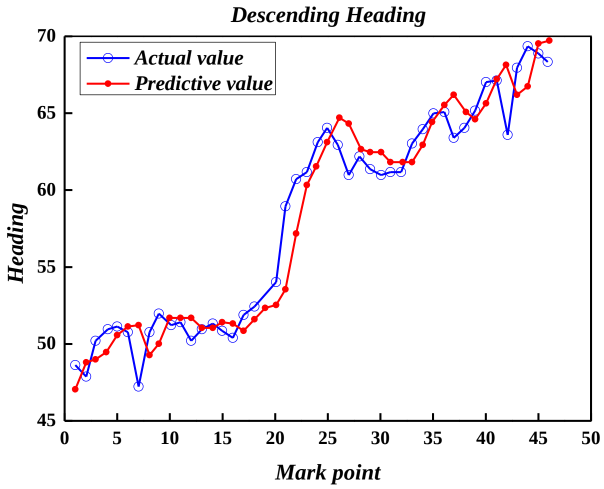

Fig. 18. Heading repair (long-time loss).

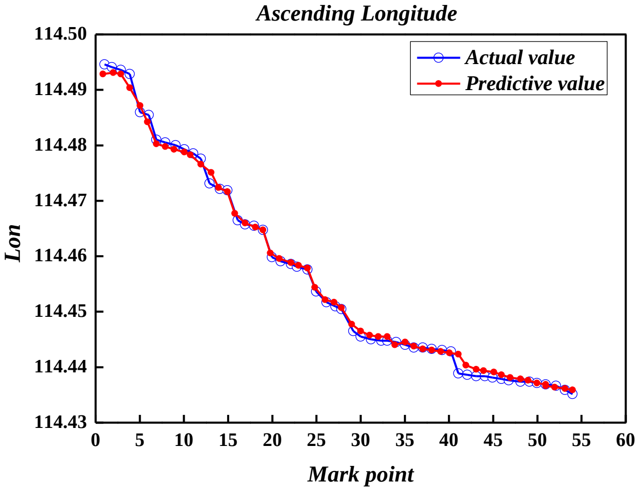

Fig. 19. Repair results of ascending longitude (R2 = 0.9976, MSE = 9.6398E-004).

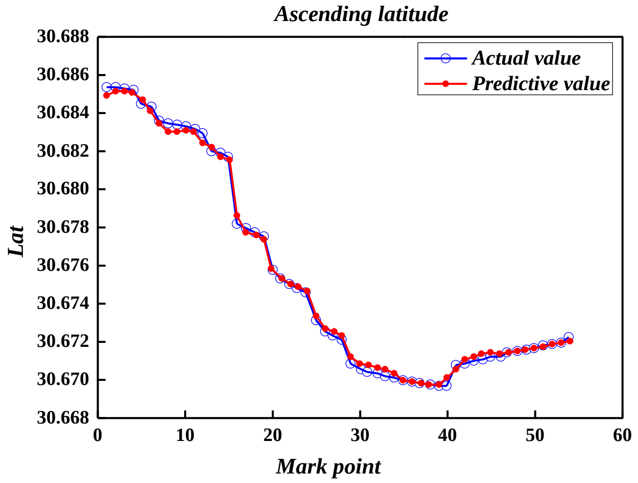

Fig. 20. Upline latitude repair results (R2 = 0.9982, MSE = 9.0608E-004).

Results Analysis: According to the evaluation indexes of the repair results in Figs. 15–22, the repair effect is good, which can meet the requirements of the repair of AIS lost data for a long time. Table VII is a part of the result of collation after restoration of the uplink lost data on September 24, 2017, in which the time points are the same as those in the short time data and obtained by the average method.

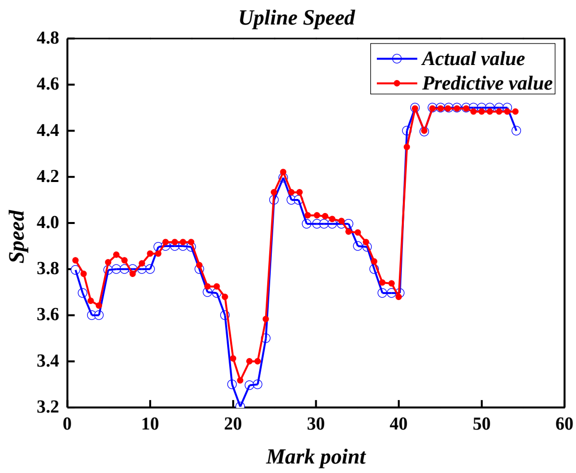

Fig. 21. Repair results of upline speed (R2 = 0.9824, the MSE = 0.0018).

Recently, the intelligent algorithms [19]–[22], including the Fuzzy model, recurrent neural network (RNN), and Random Forest (RF), have been widely used for complex data mining. These methods exhibited powerful ability for short time data mining and uncertainty estimation; however, for long time data, they may be not superior to the proposed method in this work. In order to evaluate the long-time data repair, Table VIII and Table IX compare the repair results, where the Fuzzy, RNN, and RF are directly applied to the long-time data without pre-process because for different algorithms a suitable pre-processing method is not always available. As can be seen in tables VIII and IX, the proposed method provides better accuracy over the other three intelligent algorithms for long-time data repair. The reason is probably that the long-time data are pre-processed to tailor the LSSVM. As a result, the proposed method is more suitable for long-time data repair.

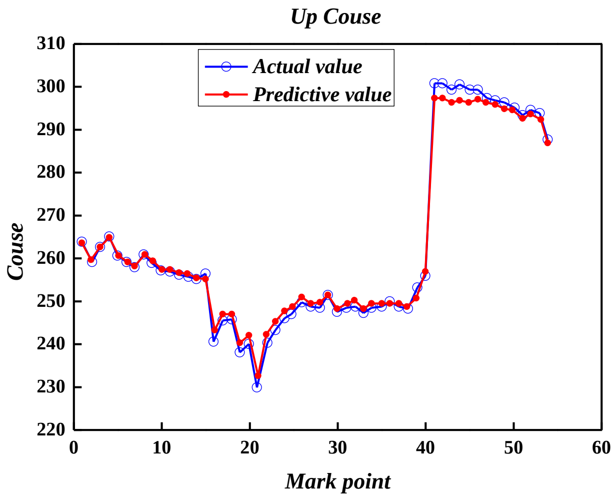

Fig. 22. Upcourse repair result (R2 = 0.9945, MSE = 0.0020).

TABLE VII. PARTIAL RESULTS OF LONG-TERM LOST DATA REPAIR (UPLINKS).

A data item	Time	Longitude (LON)	Latitude (LAT)	Speed (SPEED)	Course (COURSE)
The real value	11:02:15	114.499 784	30.685 891	3.6	266.4
Repair value	11:02:15	114.499 992	30.685 887	3.6	267.6
The real value	11:02:46	114.498 467	30.685 834	3.7	266.8
Repair value	11:02:45	114.498 456	30.685 812	3.8	266.4
The real value	11:03:15	114.498 325	30.685 745	3.7	266.9
Repair value	11:03:15	114.498 336	30.685 798	3.6	267.3
The real value	11:03:46	114.497 734	30.685 742	3.7	264.6
Repair value	11:03:45	114.497 772	30.685 756	3.7	263.7
The real value	11:04:15	114.497 172	30.685 714	3.7	262.3
Repair value	11:04:15	114.497 163	30.685 702	3.8	261.5
The real value	11:04:46	114.496 713	30.685 692	3.8	261.4
Repair value	11:04:45	114.496 702	30.685 687	3.8	261.8
The real value	11:04:15	114.496 143	30.685 671	3.8	263.8
Repair value	11:05:15	114.496 137	30.685 659	3.8	264.6
The real value	11:05:46	114.495 734	30.685 653	3.8	262.6
Repair value	11:05:45	114.495 702	30.685 637	3.8	261.8
The real value	11:06:15	114.495 112	30.685 565	3.8	261.6
Repair value	11:06:15	114.495 092	30.685 553	3.8	261

TABLE VIII. ACCURACY (MSE) OF DIFFERENT METHODS.

	Fuzzy	RNN	RF	Proposed method
Ascending longitude	9.8103E-4	9.7355E-4	9.7054E-4	9.6398E-4
Upline latitude	9.1064E-4	9.0802E-4	9.0736E-4	9.0608E-4
Upline speed	0.0023	0.0021	0.0021	0.0018
Upcourse	0.0027	0.0023	0.0025	0.0020

TABLE IX. ACCURACY (R2) OF DIFFERENT METHODS.

	Fuzzy	RNN	RF	Proposed method
Ascending longitude	0.9714	0.9763	0.9825	0.9976
Upline latitude	0.9871	0.9974	0.9956	0.9982
Upline speed	0.9768	0.9815	0.9801	0.9824
Upcourse	0.9905	0.9938	0.9936	0.9945

E. Discussions

Although this paper has done some research on the restoration of abnormal AIS data in the inland river and achieved certain results, due to the limited time, no more detailed or in-depth research has been conducted. The content of this paper can be further studied and improved in the following aspects:

1. The classification of abnormal AIS data is based on statistical analysis, which may affect the classification result because the original AIS data is not large enough. Enough data can be used to classify AIS abnormal data in a more detailed way.
2. The performance of LSSVM prediction model is closely related to the selection of parameters in the model, but there is no general method to realize the optimization of model parameters. In this paper, by analysing the results of different parameter optimization methods, PSO algorithm is adopted to optimize the parameters of the model. In order to further improve the accuracy of the repair model, the optimization of LSSVM prediction model parameters can be considered for in-depth study.
3. The problem is about the time point of data repair. In this paper, the mean value method is adopted to obtain the repair time point, which is not completely consistent with the real value. An in-depth analysis of the repair time points of abnormal AIS data can be considered.

VI. Conclusions

Repairing abnormal AIS data is important to obtain complete AIS data for intelligent ship navigation in inland waters. This paper introduces the LSSVM and cubic spline interpolation to solve the restoration of abnormal AIS data in the inland river environment. The main conclusions can be drawn as follows:

1. The abnormal AIS data can be divided into three categories: AIS error data, short-time AIS loss data, and long-time AIS loss data.
2. Different repair models were established for different types of abnormal AIS data. The LSSVM was adopted to repair AIS data error and short-time lost data, while the cubic spline interpolation was for long time AIS loss data repair.
3. Real-world AIS data were used to verify the proposed repair models. The repair results demonstrate that the abnormal AIS data can be repaired with high accuracy, and the proposed method outperforms popular intelligent models.

Conflict of Interest

The authors declare that they have no conflicts of interest.

References

[1] F. Ma, X.-m. Chu, X.-p. Yan, “Short message characteristics of AIS base station”, *Journal of Traffic and Transportation Engineering*, vol. 6, pp. 111–118, 2012. DOI: CNKI:SUN:JYGC.0.2012-06-021.

[2] X.-m. Chu, T. Liu, F. Ma, X.-l. Liu, and M. Zhong, “Field intensity distribution characteristics of AIS signal in mountainous channel”, *Journal of Traffic and Transportation Engineering*, vol. 6, pp. 117–126, 2014. DOI: 10.3969/j.issn.1671-1637.2014.06.015.

[3] Q. Hu, J. Cao, G. Gao, L. Xu, and M. Song, “Study of an evaluation model for AIS receiver sensitivity measurements”, *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 4, pp. 1118–1126, 2020. DOI: 10.1109/TIM.2019.2910341.

[4] A. Goudossis and S. K. Katsikas, “Towards a secure automatic identification system (AIS)”, *J. Mar. Sci. Technol.*, vol. 24, pp. 410–423, 2019. DOI: 10.1007/s00773-018-0561-3.

[5] L. Li and Ch. Yang, “Multithreaded AIS information processing and storage”, *Ship Electronic Engineering*, vol. 3, pp. 175–177, 2007. DOI: 10.3969/j.issn.1627-9730.2007.03.052.

[6] Sh. Sun, Y. Chen, Z. Piao, and J. Zhang, “Vessel AIS trajectory online compression based on scan-pick-move algorithm added sliding window”, *IEEE Access*, vol. 8, pp. 109350–109359, 2020. DOI: 10.1109/ACCESS.2020.3001934.

[7] B. Y. Wang and X. Fang, “AIS message analysis and implementation based on bit fields”, *Navigation Technology*, vol. 2, pp. 41–44, 2014. DOI: CNKI:SUN:HHJS.0.2014-02-025.

[8] D. Gaglione, G. Soldi, F. Meyer *et al.*, “Bayesian information fusion and multitarget tracking for maritime situational awareness”, *IET Radar, Sonar & Navigation*, vol. 14, no. 12, pp. 1845–1857, 2020. DOI: 10.1049/iet-rsn.2019.0508.

[9] L.-zh. Sang, A. Wall, Zh. Mao, X.-p. Yan, and J. Wang, “A novel method for restoring the trajectory of the inland waterway ship by using AIS data”, *Ocean Engineering*, vol. 110, part A, pp. 183–194, 2015. DOI: 10.1016/j.oceaneng. 2015.10.021.

[10] W. Zhang, “Research on early warning system of ship navigation risk in Qiaoqu district, M.S. thesis, Wuhan University of Technology, Wuhan, China, 2013. DOI: 10.7666/d.Y2504804.

[11] Zh. Yan, Y. Xiao, L. Cheng *et al.*, “Exploring AIS data for intelligent maritime routes extraction”, *Applied Ocean Research*, vol. 101, p. 102271, 2020. DOI: 10.1016/j.apor.2020.102271.

[12] C. Iphar, A. Napoli, and C. Ray, “Detection of false AIS messages for the improvement of maritime situational awareness”, in *Proc. of Oceans 2015-MTS/IEEE Washington*, 2015, pp. 1–7. DOI: 10.23919/OCEANS.2015.7401841.

[13] J. H. Wu, C. Wu, W. Liu *et al.*, “Automatic detection and repair algorithm for abnormal vessel AIS tracks”, *China Maritime Navigation*, vol. 40, pp. 8–12, 2017. DOI: 101. 10.3969/j.issn.1000-4653.2017.01.003

[14] L. Liu, Zh. Jiang, X. Chu, Ch. Zhong, and D. Zhang, “The ship automatic identification system data restoration and prediction algorithm research”, *Journal of Harbin Engineering University*, vol. 40, pp. 1072–1077, 2019. DOI: 10.11990/jheu.201803011.

[15] J. Li, X. Chu, X. Liu *et al.*, “Restoration method of missing track of inland river ships”, *Journal of Harbin Engineering University*, vol. 40, pp. 67–73, 2019. DOI: 10.11990/jheu.201708038.

[16] J. Li, X. Chu, X. Liu, Sh. Xie, and W. He, “An approach for restoring the lost trajectories of vessels in inland waterways”, *Journal of Harbin Engineering University*, vol. 40, pp. 37–41, 2019. DOI: 10.11990/jheu.201708038.

[17] Ch. Zhong, Zh. Jiang, X. Chu, and L. Liu, “Inland ship trajectory restoration by recurrent neural network”, *The Journal of Navigation*, vol. 72, no. 6, pp. 1–19, 2019. DOI: 10.1017/S0373463319000316.

[18] L. Liu, Ch.-zh. Wu, D.-f. Chu *et al.*, “Research on ship trajectory repair based on VONDRAK filtering and cubic interpolation”, *Traffic Information and Safety*, vol. 33, pp. 100–105, 2015. DOI: 10.3963/j.issn1674-4861.2015.04.016.

[19] R. Naseem, Z. Shaukat, M. Irfan *et al.*, “Empirical assessment of machine learning techniques for software requirements risk prediction”, *Electronics*, vol. 10, no. 2, p. 168, 2021. DOI: 10.3390/electronics10020168.

[20] A. Glowacz, “Fault diagnosis of electric impact drills using thermal imaging”, *Measurement*, vol. 171, p. 108815, 2021. DOI: 10.1016/j.measurement.2020.108815.

[21] D. Andriukaitis, A. Laucka, A. Valinevicius, M. Zilys, V. Markevicius, D. Navikas, R. Sotner, J. Petrzela, J. Jerabek, N. Herencsar, and D. Klimenta, “Research of the Operator’s Advisory System Based on Fuzzy Logic for Pelletizing Equipment”, *Symmetry*, vol. 11, no. 11, p. 1396, Nov. 2019 DOI: 10.3390/sym11111396.

[22] O. AlShorman, M. Irfan, N. Saad *et al.*, “A review of artificial intelligence methods for condition monitoring and fault diagnosis of rolling element bearings for induction motor”, *Shock and Vibration*, vol. 2020, article ID 8843759, 2020. DOI: 10.1155/2020/8843759.