International Journal of Innovative Computing, Information and Control
Volume **17**, Number **2**, April **2021**

ICIC International ©2021 ISSN 1349-4198
pp. 443–459

SHIP TRAJECTORY CLEANSING AND PREDICTION WITH HISTORICAL AIS DATA USING AN ENSEMBLE ANN FRAMEWORK

Yang Sun[1], Xinqiang Chen[2], Ling Jun[3], Jiansen Zhao[1,*], Qinyou Hu[1]
Xianghui Fang[2,*] and Ying Yan[4]

[1]Merchant Marine College
[3]Institute of Logistics Science and Engineering
Shanghai Maritime University
No. 1550, Haigang Avenue, Shanghai 201306, P. R. China
*Corresponding author: jszhao@shmtu.edu.cn

[2]Institute of Atmospheric Sciences
Fudan University
No. 220, Handan Road, Shanghai 200438, P. R. China
*Corresponding author: fangxh@fudan.edu.cn

[4]College of Transportation
Chang'an University
Middle-Section of Nan'er Huan Road, Xi'an 710064, P. R. China

Received October 2020; revised February 2021

Abstract. *Ship trajectory provides crucial spatial-temporal maritime traffic information, which helps maritime traffic participants enhance the safety and efficiency for the traffic control and management. In that manner, significant focuses were paid to obtain accurate ship trajectory with the help of historical Automatic Identification System (AIS) data. To that aim, we proposed an ensemble Artificial Neural Network (ANN) model to cleanse and predict ship trajectory from the AIS data (i.e., latitude, longitude, speed). The proposed framework smoothed noises in the raw AIS data via the ensemble Hampel Filter (HF) and Butterworth Filter (BF). Then, the proposed framework normalized the smoothed AIS data to equalize the time interval between neighboring AIS samples. After that, we predicted the ship trajectory with the help of ANN model. The experimental results showed that our proposed model was effective and efficient in removing the AIS data outlier, and obtained satisfied ship trajectory prediction results.*

Keywords: Historical AIS data, Ensemble artificial neural network model, Data denoising, Trajectory prediction, Sustainable maritime traffic

1. **Introduction.** Automatic Identification System (AIS) data provides both dynamic and static information to maritime traffic participants, which includes latitude, longitude, Speed over Ground (SOG), Course over Ground (COG), call sign, Maritime Mobile Service Identify (MMSI), etc. The AIS system broadcasts the ship information in an automatic self-reporting manner, and thus helps nearby ships take early initiative behaviors to avoid potential maritime traffic collisions. According to the International Convention for Safety of Life at Sea (SOLAS), the AIS is a mandatory facility for both all cargo ships (with the gross tonnage larger than 300GT) and passenger ships [1,2]. In that manner, the AIS system is very important for the maritime safety and management due to advantage of tremendous spatial-temporal information and extensive deployment [3]. More specifically, the own-ship can forecast neighboring ship movements (e.g., turning left, turning right,

DOI: 10.24507/ijicic.17.02.443

moving straight) and then varied collision-avoidance actions will be taken in advance to enhance maritime safety with the support of real-time/historical AIS data [4,5].

The AIS data quality largely relies on the on-spot traffic volume (i.e., on-site maneuvering ships) ship status (e.g., acceleration, deceleration, constant speed), AIS base station location [6]. Previous studies suggested that the AIS data may be lost when the traffic volume is large (i.e., maritime traffic jam) [7-9]. The main reason is that the to-be-transmitted AIS data exceed the ship-borne AIS receiver capacity considering that the AIS broadcast frequency is positive to the ship moving state (i.e., larger AIS transmitting frequency leading to more AIS data). Moreover, ships installed with Class A AIS equipment (e.g., passenger ships, international cargo ships) provide better AIS data quality compared to those with Class B AIS equipment (usually installed on the inland waterway ships) [10]. It is noted that the AIS data quality transmitted from coast ships is better than those from the ships sailing in open sea [11].

Previously, various studies were reported to quantitatively analyze AIS data quality [12-14]. Hu et al. proposed a novel evaluation model to measure the AIS system receiver sensitivity in terms of packet error rate [15]. Sheng and Yin employed the AIS data to explore regular shipping route patterns with steps of data preprocessing, structure similarity measurement, data clustering and trajectory extraction [16]. It is noted that little attention was paid to the AIS data noise elimination and ship trajectory restoration performance is heavily relied on the raw data quality [11]. Yan et al. extracted ship traffic routes from historical AIS data by transforming the raw ship trajectory information into ship semantic object [17]. Xu et al. proposed a novel framework to control the autonomous ships by applying the dynamic time warping model to exploit the AIS data [18]. We found that the AIS data cleaning relevant studies were mainly implemented by focusing on the latitude and longitude information correction, regardless of removing ship kinematic data outliers (e.g., ship moving speed) [19,20]. In that manner, trivial noises in latitude and longitude data may be wrongly detected as AIS details leading to biased ship trajectory exploitation results.

To sum, less attention was paid to simultaneously suppress the dynamic and static outliers from the AIS data, and thus the research findings obtained from the noisy AIS data may be questionable. In that way, data quality control procedure is quite important for implementing the AIS supported task (especially for the ship trajectory exploitation and analysis). Our study aims to cleanse and predict ship trajectory by fully exploiting spatial-temporal AIS information (i.e., latitude, longitude, speed). More specifically, we firstly denoise the raw AIS with preprocessing steps of trajectory segment separation and data cleaning. Then, we normalize the time interval for the smoothed AIS trajectory samples, which is further processed by the ship prediction module. The study helps maritime authorities and ship crew take early-warning initiatives to avoid potential collisions, and thus significantly improve maritime traffic safety. The remainder of the paper is organized as follows. We introduce the proposed framework details in Section 2. The data source and experimental results are illustrated in Section 3. Section 4 briefly concludes the study and potential future directions.

2. Methodology.

The trivial outliers will largely reduce the AIS data applicability for fulfilling the task of traffic kinematic information exploitation. To alleviate the disadvantage, our proposed framework preprocesses the raw AIS data with the steps of data segmentation, outlier removal (with Hampel filter and Butterworth filter), sample interval normalization. Note that impulsive noise is common in the AIS data samples. To address the issue, we firstly employ the Hampel filter to suppress the data outlier. Then, the Butterworth model is introduced to smooth trivial anomalies in the AIS data. After

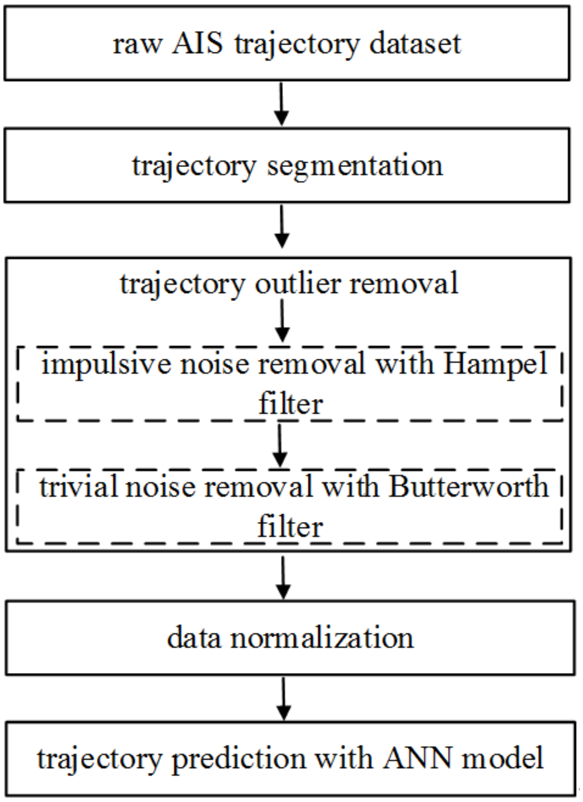

raw AIS trajectory dataset
trajectory segmentation
trajectory outlier removal
impulsive noise removal with Hampel filter
trivial noise removal with Butterworth filter
data normalization
trajectory prediction with ANN model

FIGURE 1. Schematic view for the proposed framework

that, we forecast ship trajectory variation tendency with the smoothed AIS data with the support of Artificial Neural Network (ANN). The proposed framework overview is shown in Figure 1.

2.1. **Ship trajectory segmentation.** The database stores the trajectory samples from different ships (i.e., AIS data) via the data delivering/receiving timestamp, and thus we need to cluster trajectory data for the same ship for the purpose of further trajectory analysis task. It is noted that each ship is assigned with a unique MMSI code by the International Maritime Organization (IMO), which is used in our study to identify AIS data from different ships. We remove the repetition trajectory sample following the rule in Equation (1) considering that data duplication is common (i.e., ship with different MMSI with the same latitude, longitude and timestamp). Moreover, a few time interval samples between neighboring AIS data were quite large (e.g., 2 hours), which is beyond the reasonable AIS data transmission frequency (i.e., ranging from 2 seconds to 10 minutes). Such unreasonable time gap indicates that the ship may start new voyage after manually fine-tuning the ship trajectory from electronic navigational chart. Another possible reason is that many AIS data are lost due to AIS receiver capacity limitation (i.e., huge AIS data are being forwarded simultaneously in the wireless channel). We divide the initial ship trajectory into two sub-trajectories when the neighboring time gap difference satisfies the criterion in Equation (2).

$$\begin{cases} MMSI_a \neq MMSI_b \\ Ts_a = Ts_b \\ La_a = La_b \\ Lo_a = Lo_b \end{cases} \tag{1}$$

$$\begin{cases} |Ts_1 - Ts_2| > T_1 \\ |Ts_2 - \ Ts_3| \leq T_2 \end{cases} \tag{2}$$

where $MMSI_a$, Ts_a, La_a, Lo_a are MMSI, timestamp, latitude and longitude in the AIS database for trajectory a, and $MMSI_b$, Ts_b, La_b, Lo_b are the counterparts for the trajectory b. The symbols Ts_1, Ts_2 and Ts_3 are the three neighboring timestamps from the same trajectory, while T_1 and T_2 are two thresholds.

2.2. **AIS data outlier removal with Hampel and Butterworth filter.**

2.2.1. *Impulsive AIS data outlier smoothing with Hampel filter.* Impulsive noises are considered as a type of typical data outliers in the raw AIS data, which showed significant data variation tendency in comparison to the neighbors. We employ the Hampel Filter (HF) to remove the impulsive noises by averaging the neighbors. Note that the HF is a type of decision-based filter, which removes the AIS impulsive outlier with the support of median filter and mean absolute deviation scale estimator [21]. More specifically, the HF computes the median with moving window consisting of the AIS sample and six neighbors (i.e., three per side). Note that we employ the symbol M to represent the moving window width for the purpose of generalization. The AIS data sample is considered as an outlier when the absolute difference between the AIS data sample and the median is larger than the given threshold. In that way, the AIS sample outlier is replaced by the median value and the filter response $\widetilde{S}_i$, which are calculated through Equations (3) and (4), respectively. Note that the standard deviation is estimated by multiplying a constant C and median value difference (between two AIS data samples). We set the constant C into 1.4826 by following the rule in previous study [22].

$$\widetilde{S}_i = \begin{cases} S_i \left|S_i - \bar{S}_i\right| \leq tSd_i \\ \bar{S}_i \left|S_i - \bar{S}_i\right| > tSd_i \end{cases} \tag{3}$$

$$Sd_i = C \times median_{j\in[-M,M]} \left\{\left|S_{i-j} - \bar{S}_i\right|\right\} \tag{4}$$

where the S_i is the ith AIS data sample, and the symbol $\bar{S}_i$ is median value with the given moving window width M. The symbol S_{i-j} is the $(i-j)$th AIS data sample $(i > j)$, and the Sd_i is the ith standard deviation. Note that the parameter t is a scale factor.

2.2.2. *Trivial AIS data outlier with Butterworth filter.* The output from the above step provides us impulsive-noise free AIS data. However, the trivial anomalies are quite common which can be easily determined from the HF-denoised AIS data. The Butterworth Filter (BF) removes the trivial outlier by transforming the AIS data into frequency domain, and thus suppresses the trivial outliers by setting appropriate stopband frequency threshold. More specifically, the BF removes the anomaly oscillations which shows its potential in smoothing the outliers from the input HF-denoised AIS data series. The outlier removal procedure for the BF model is implemented by mapping the raw AIS data into Z-domain with the help of the digital transfer function (see Equation (5)). We obtain the zero and pole points for the AIS data in the Z-domain with the help of transfer function. Note that the zero relevant points are employed to calculate the numerator in Equation (5), and the poles for the denominator counterparts. The BF-smoothed AIS data is determined by calculating the difference between the HF-denoised and the previous input BF-smoothed data samples (see Equation (6)). To determine the BF coefficients, we establish a similar transfer function for the purpose of obtaining the coefficients, which can be found in Equation (7). The relationship between the similar and digital transfer function is exploited with the help of Equation (8). We can easily obtain the BF coefficients when the similar transfer function equals the digital transfer function (see Equation (9)).

$$H(z) = \frac{u_0 + u_1 z^{-1} + \cdots + u_n z^{-n}}{1 + v_1 z^{-1} + \cdots + v_n z^{-n}} \tag{5}$$

$$AIS_{BF}(i) = \sum_{k=0}^{n} u_k * AIS_{HF}(i-k) - \sum_{k=1}^{n} v_k * AIS_{BF}(i-k) \tag{6}$$

$$G(q) = \frac{U_0 + U_1 q^{-1} + \cdots + U_n q^{-n}}{V_0 + V_1 q^{-1} + \cdots + V_n q^{-n}} \tag{7}$$

$$q = \frac{D\left(1 - z^{-1}\right)}{1 + z^{-1}} \tag{8}$$

$$G(q) = H(z) \tag{9}$$

where u_k and v_k ($k = 0, 1, \ldots, n$) are the coefficients for the purpose of identifying BF frequency response. The parameter n is the order for the BF model. The $AIS_{HF}(i-k)$ is the $(i-k)$th sample point for the HF-smoothed AIS data, and $AIS_{BF}(i-k)$ is the counterpart for the BF-smoothed data. The $AIS_{BF}(i)$ is the ith BF smoothing sample point, and parameter D is a constant.

2.3. **Data interval normalization.** The above-mentioned ensemble HF and BF model provides us noise-free AIS data with different time intervals for the same AIS trajectory. The main reason is that the AIS data is delivered at various frequencies under different ship travelling states (anchoring state, fast speed, etc.). Thus, it is essential to normalize the time interval before exploiting the AIS spatial-temporal patterns. Following the rule in previous studies [23], we employ the cubic spline interpolation model to normalize the time span among the AIS data samples. Given three neighboring AIS data samples AIS_1, AIS_2 and AIS_3, we store both the AIS_1 and AIS_3 samples when one of the following assumptions is satisfied (see Equations (10), (11) and (12)). The AIS_2 will be replaced by interpolation samples generated by the cubic spline interpolation model, and more details are suggested to refer to [24].

$$\begin{cases} |Dis_{12} - Dis_{23}| > Th_1 \\ Dis_{12} < Th_2 \end{cases} \tag{10}$$

$$T_{12} > Th_3 \tag{11}$$

$$|sp_1 - sp_2| > Th_4 \tag{12}$$

where Dis_{12} is the ship moving displacement between the AIS_1 and AIS_2 samples, and the T_{12} is the corresponding traveling time. The symbol sp_1 is the ship speed recorded in the AIS_1 item.

Parameter Dis_{23} is the ship travelling distance between the position AIS_2 and AIS_3, and the sp_2 is speed in the AIS_2. Note that ship speed is considered as a scalar parameter in our study (i.e., we neglect ship heading direction in Equation (12)). The Th_1, Th_2, Th_3 and Th_4 are the thresholds.

2.4. **Ship trajectory prediction with ANN model.** The ANN model is good at prediction relevant tasks (e.g., traffic flow prediction at varied time resolutions [25]), and thus the ANN is applied to forecasting the ship trajectory in our study. The ANN model is a type of bionic algorithm which exploits the intrinsic data variation pattern with neuron perception. The Back-Propagation (BP) neural network, a type of feed-forward ANN model, is introduced to implement ship trajectory prediction task with the AIS data. The BP neural network consists of the input layer, hidden layer and output layer, and the hidden layer aims to learn ship kinematic information from the input AIS samples. More specifically, the BP neural network extracts the ship travelling patterns from the input training samples with the help of transfer function. We employ the sigmoid function (see Equation (13)) as the transfer function in BP network in our study. Note that the BP model is fine-tuned by adjusting the network structure and parameter setups with the feedback results (i.e., the error between predicted and ground truth AIS data samples).

$$f\left(sig_{he}^{i}\right) = \frac{1}{\left(1 + \exp\left(-sig_{he}^{i}\right)\right)} \tag{13}$$

where sig_{he}^{i} is the state of the eth neuron of the hidden layer with the ith data sample.

2.5. **Prediction goodness measurement.** For the purpose of verifying model prediction accuracy, we quantify the difference between the ground truth and predicted AIS data with several statistical indicators. More specifically, we verify the proposed framework performance with Root Mean Square Error (RMSE), Mean Square Error (MSE), Mean Absolute Error (MAE), Symmetric Mean Absolute Percentage Error (SMAPE), Mean Absolute Percentage Error (MAPE), Average Euclidean Distance (AEUD) and Frechet Distance (FRD). We evaluate the proposed framework performance from the perspective of one- and two-dimensional data prediction accuracy, respectively. The RMSE, MSE, MAE, SMAPE, MAPE are introduced to verify the model performance on the longitude prediction task, which is applied to the latitude and speed data, too. The above-mentioned five indicators are quite popular yet typical for fulfilling the task of prediction error measurement. The two-dimensional AIS data is verified with the help of AEUD and FRD indicators. More specifically, the two statistics measure the sample distance between the predicted and ground-truth data supported by the longitude and latitude data. Note that smaller RMSE, MSE, MAE, SMAPE, MAPE, AEUD and FRD indicate better prediction accuracy, and vice versa. The formulas for calculating the RMSE, MSE, MAE, SMAPE, MAPE, AEUD and FRD indicators are shown as follows (see Equations (14) to (20)).

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(p_i - gt_i\right)^2} \tag{14}$$

$$MSE = \frac{1}{m}\sum_{i=1}^{m}\left(p_i - gt_i\right)^2 \tag{15}$$

$$MAE = \frac{1}{m}\sum_{i=1}^{m}\left|p_i - gt_i\right| \tag{16}$$

$$SMAPE = \frac{1}{m}\sum_{i=1}^{m}\frac{|p_i - gt_i|}{(|p_i| + |gt_i|)/2} \tag{17}$$

$$MAPE = \frac{1}{m}\sum_{i=1}^{m}\left|\frac{p_i - gt_i}{gt_i}\right| \tag{18}$$

$$FRD = \max_{i\in[1,m]}\sqrt{(p_i(lat) - gt_i(lat))^2 + (p_i(lon) - gt_i(lon))^2} \tag{19}$$

$$AEUD = \frac{\sum_{i=1}^{m}\sqrt{(p_i(lat) - gt_i(lat))^2 + (p_i(lon) - gt_i(lon))^2}}{m} \tag{20}$$

where m is the length of AIS trajectory (i.e., number of data samples). The symbols p_i and gt_i are the predicted and ground truth AIS data, respectively. The latitude and longitude in the p_i are represented as $p_i(lat)$ and $p_i(lon)$, respectively. The rule is applicable to the $gt_i(lat)$ and $gt_i(lon)$.

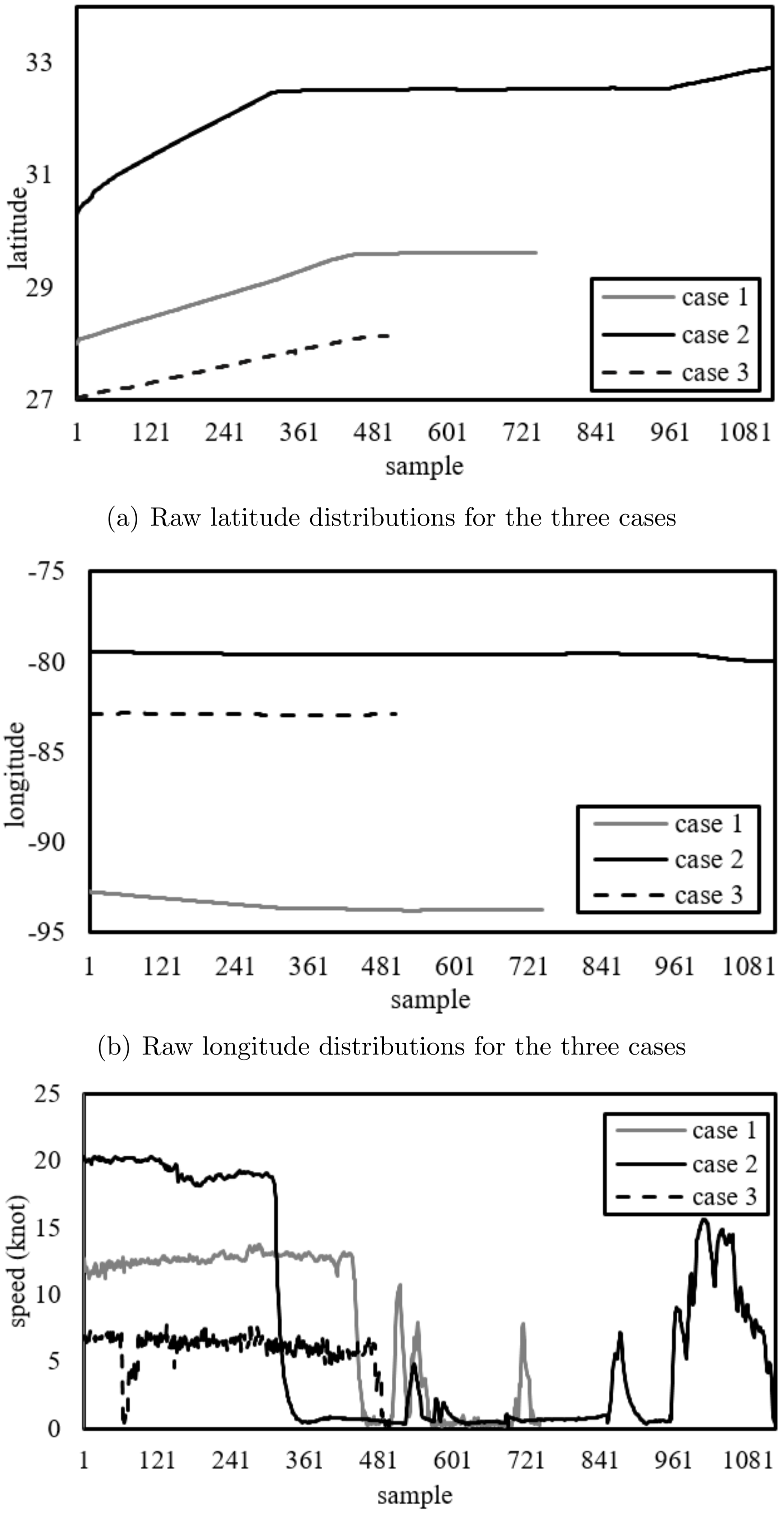
3. **Experiments and Results.**

3.1. **Data.** The National Oceanic and Atmospheric Administration and Bureau of Ocean Energy Management cooperate to publish large scale AIS data (preprocessed with data masking technique), which can be freely downloaded from the application programming interface (https://marinecadastre.gov/ais/) by request [26]. Each AIS data sample contains latitude, longitude, speed, course over ground, MMSI, coordinated universal time, timestamp, call sign, etc. We select three AIS data segments locating in the Gulf of Mexico. The latitude starts from 20°N to 35°N, and the longitude ranges from 75°W to 95°W. Following the rule in previous studies, the southbound latitude (westbound longitude) is presented as a negative number, and the northbound (eastbound) counterpart with a positive value. More specifically, the first AIS segment (with MMSI code 209289000) latitude ranges from 27°N to 30°N, and the longitude from 92°W to 94°W, and the trajectory contains 742 AIS data samples. The second and the third AIS trajectory segments (with MMSI code 212416000 and 366740340, respectively) start from 25°N to 33°N in terms of latitude, and the longitude from 79°W to 83°W. We collect 1124 AIS data samples for the case 2 and 502 points for the case 3. We focus on the ship latitude, longitude and speed cleaning and prediction with the AIS data, and the raw data are shown in Figure 2. We denote the first, second and the third AIS trajectory segments as cases 1, 2 and 3 for simplicity.

Figure 2(a) showed that the raw latitude did not contain obvious data outliers for the three cases (see the light line, dark line and dash line). However, we can still find a few impulsive outliers from the dash line (i.e., the case 3) when zooming into the AIS trajectory details. The longitude data showed quite similar distributions as those of the latitude samples (as shown in the three curves in Figure 2(b)). In that way, we did not observe obvious outliers in the longitude data. In sum, typical outliers for the latitude and longitude data can be considered as trivial oscillations. The speed distributions for the three cases showed more abnormal oscillations in comparison to the latitude and longitude counterparts (see the three curves in Figure 2(c)). Note that the timestamp for each AIS sample is labeled as order number in our study due to inconsistent timespan of the three trajectories, and the rule is applicable to following sections.

3.2. **Ship trajectory cleaning and prediction for case 1.** Ship movement in the channel is supposed to be smooth due to that sudden ship travelling decision (e.g., sharp turn) can trigger ship turning-over due to the large inertia. More specifically, ship kinematic data (latitude, longitude, speed, etc.) from AIS with significant variation under a given time span can be deemed as outliers. We analyzed our model performance on the task of ship trajectory cleaning and prediction in detail for case 1, which was further verified on case 2 and case 3, respectively. For the purpose of performance comparison, we implemented the BF, HF, HF+BF (the smoothing module in our proposed framework), Ensemble Empirical Mode Decomposition (EEMD) [27], wavelet filter [23] to smooth out the AIS trivial noises. The latitude and longitude smoothing results for the case 1 were shown in Figures 3(a) and 3(b), respectively. It is noted that different denoising methods showed quite similar performance due to that the raw latitude and longitude from the case 1 were not significantly contaminated by the noises.

The speed smoothing results for the case 1 indicated that different models obtain varied performance. The speed data smoothed by the wavelet filter showed significant dip and choppy samples (with sample number ranging from 400 to 500), which can be found in the dark curve in Figure 3(c). The main reason is that partial detail coefficients from the raw speed data series were wrongly retained as the speed data details. Note that the BF, HF, HF+BF and EEMD successfully removed such outliers from the same speed



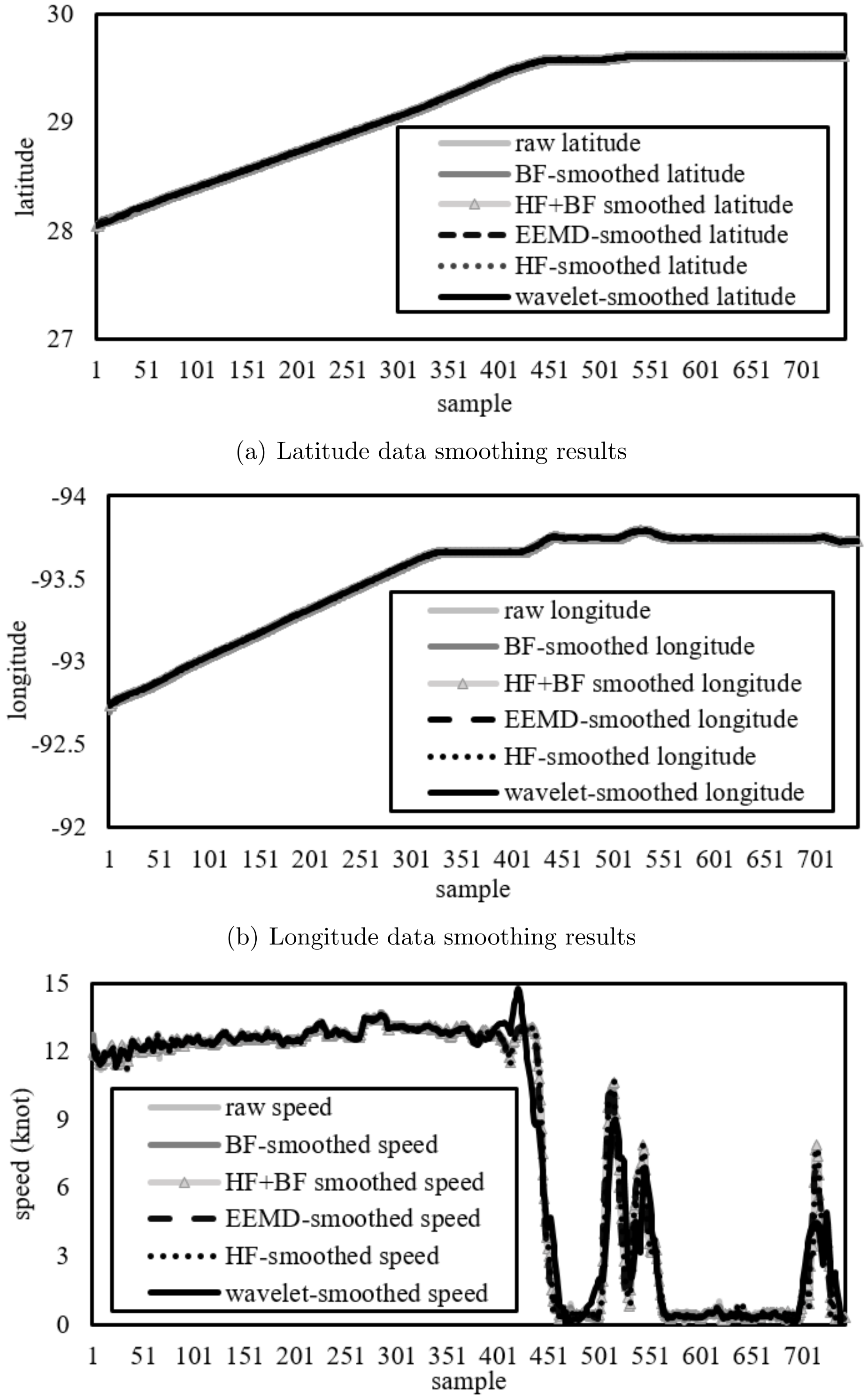
(a) Raw latitude distributions for the three cases

(b) Raw longitude distributions for the three cases

(c) Raw speed distributions for the three cases

FIGURE 2. Raw AIS data for the three trajectory segments

samples. Moreover, the BF, HF and EEMD models' smoothed speed distributions showed trivial abnormal fluctuations (with the data samples ranging from 550 to 700 shown in Figure 3(c)). However, the speed distributions obtained by our proposed framework showed satisfied performance considering that the abnormal oscillations were successfully



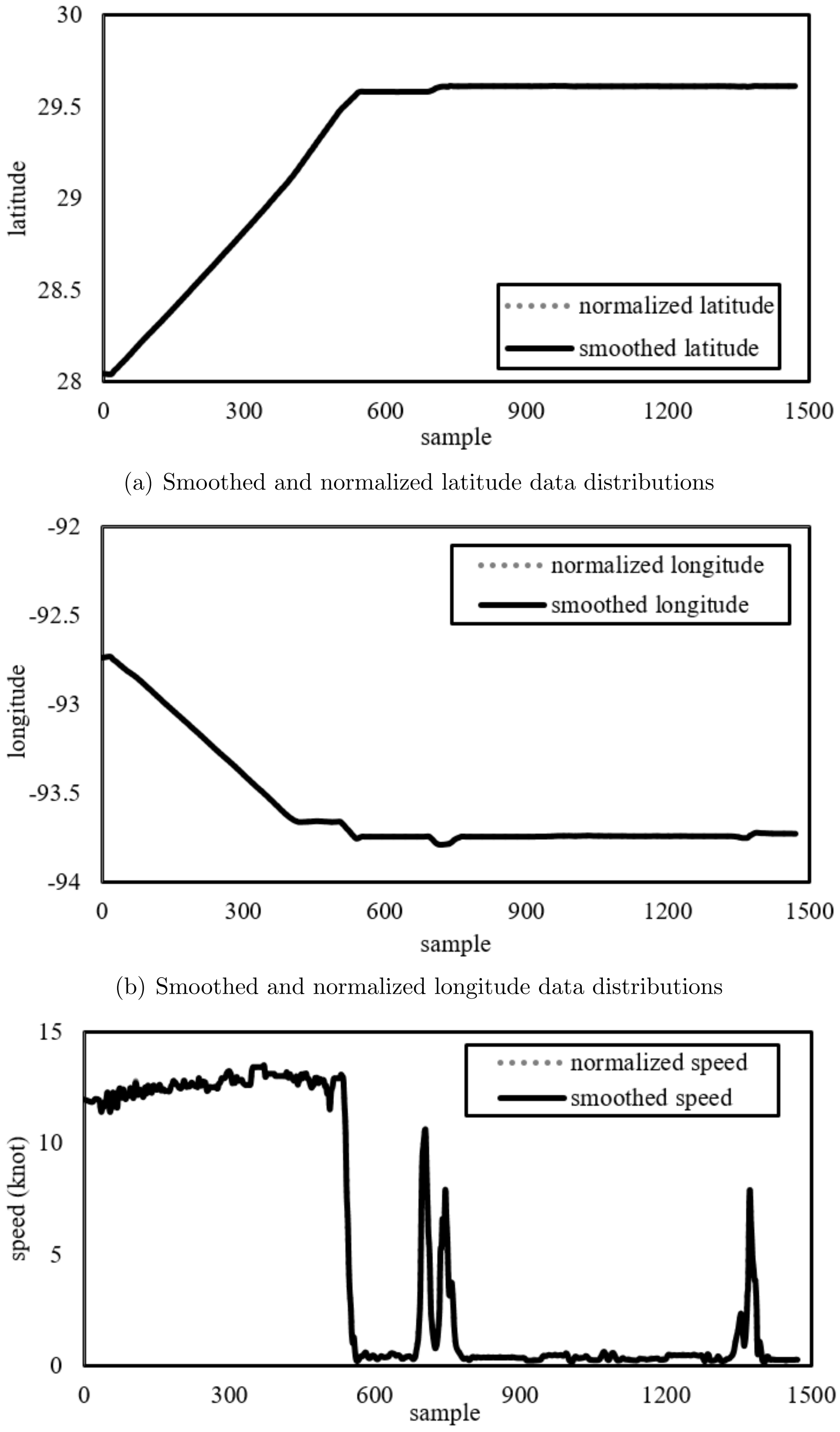
(a) Latitude data smoothing results

(b) Longitude data smoothing results

(c) Speed data smoothing results

Figure 3. AIS data smoothing results for case 1

removed. In that manner, our proposed ship trajectory reconstruction framework can handle the data outlier suppress task. Moreover, we obtained the normalized data based on the HF-BF smoothed AIS data which were shown in Figure 4. More specifically, the HF+BF smoothed latitude, longitude and speed data (which are labeled as smoothed data in Figures 4(a), 4(b) and 4(c)) were employed to generate the normalization counterparts.



(a) Smoothed and normalized latitude data distributions

(b) Smoothed and normalized longitude data distributions

(c) Smoothed and normalized speed data distributions

Figure 4. Normalized AIS data distributions for case 1

For the purpose of quantifying model performance, we provided the ship trajectory prediction accuracy in terms of latitude, longitude and speed by compared to Long Short-Term Memory (LSTM) [28] and Support Vector Machine (SVM) [29]. Note that our proposed framework is abbreviated as HBA. From the perspective of MAPE, our proposed framework (i.e., the HBA column shown in the following tables) significantly outperformed the counterparts of LSTM and SVM considering that the HBA-obtained MAPE is 8.63×10^{-7} (see Table 1). It is noticed that the SMAPE values for the three models were the same as those of MAPE due to that we only keep two digits for the fractional part. The minimal MAE is 8.07×10^{-5} (obtained by our proposed model) and the maximum value is 1.76×10^{-2} (obtained by the LSTM). Moreover, the minimal RMSE and MSE are 1.66×10^{-4} and 2.75×10^{-8}, respectively, which were both obtained by HBA model. Based on the above analysis, we can draw the conclusion that our proposed model obtained better latitude prediction accuracy compared to the LSTM and SVM models.

Table 1. Latitude prediction accuracy distributions for different models

	HBA	LSTM	SVM
MAPE	$\mathbf{8.63 \times 10^{-7}}$	5.92×10^{-4}	3.81×10^{-5}
SMAPE	$\mathbf{8.63 \times 10^{-7}}$	5.93×10^{-4}	3.81×10^{-5}
MAE	$\mathbf{8.07 \times 10^{-5}}$	1.76×10^{-2}	3.57×10^{-3}
RMSE	$\mathbf{1.66 \times 10^{-4}}$	1.76×10^{-2}	6.72×10^{-3}
MSE	$\mathbf{2.75 \times 10^{-8}}$	3.08×10^{-4}	4.52×10^{-5}

The longitude and speed prediction performance for the case 1 were shown in Tables 2 and 3, respectively. We noticed that longitude prediction performance is quite similar to that of the latitude prediction performance. More specifically, the MAPE, SMAPE, MAE, RMSE and MSE obtained by our proposed framework (i.e., HBA) are 3.65×10^{-6}, 3.65×10^{-6}, 1.05×10^{-4}, 2.60×10^{-4} and 6.77×10^{-8}, which were smaller than the counterparts of LSTM and SVM. From perspective of speed prediction accuracy, the HBA

Table 2. Longitude prediction accuracy distributions for different models

	HBA	LSTM	SVM
MAPE	$\mathbf{3.65 \times 10^{-6}}$	2.51×10^{-5}	3.64×10^{-5}
SMAPE	$\mathbf{3.65 \times 10^{-6}}$	2.51×10^{-5}	3.64×10^{-5}
MAE	$\mathbf{1.05 \times 10^{-4}}$	2.35×10^{-3}	1.07×10^{-3}
RMSE	$\mathbf{2.60 \times 10^{-4}}$	2.98×10^{-3}	1.99×10^{-3}
MSE	$\mathbf{6.77 \times 10^{-8}}$	8.85×10^{-6}	3.96×10^{-6}

Table 3. Speed prediction accuracy distributions for different models

	HBA	LSTM	SVM
MAPE	$\mathbf{2.22 \times 10^{-2}}$	3.10×10^{0}	5.21×10^{-1}
SMAPE	$\mathbf{2.21 \times 10^{-2}}$	1.69×10^{-1}	2.31×10^{-1}
MAE	$\mathbf{4.21 \times 10^{-2}}$	1.59×10^{-1}	4.48×10^{-1}
RMSE	$\mathbf{9.60 \times 10^{-2}}$	3.50×10^{-1}	1.16×10^{0}
MSE	$\mathbf{9.22 \times 10^{-2}}$	1.22×10^{-1}	1.34×10^{0}

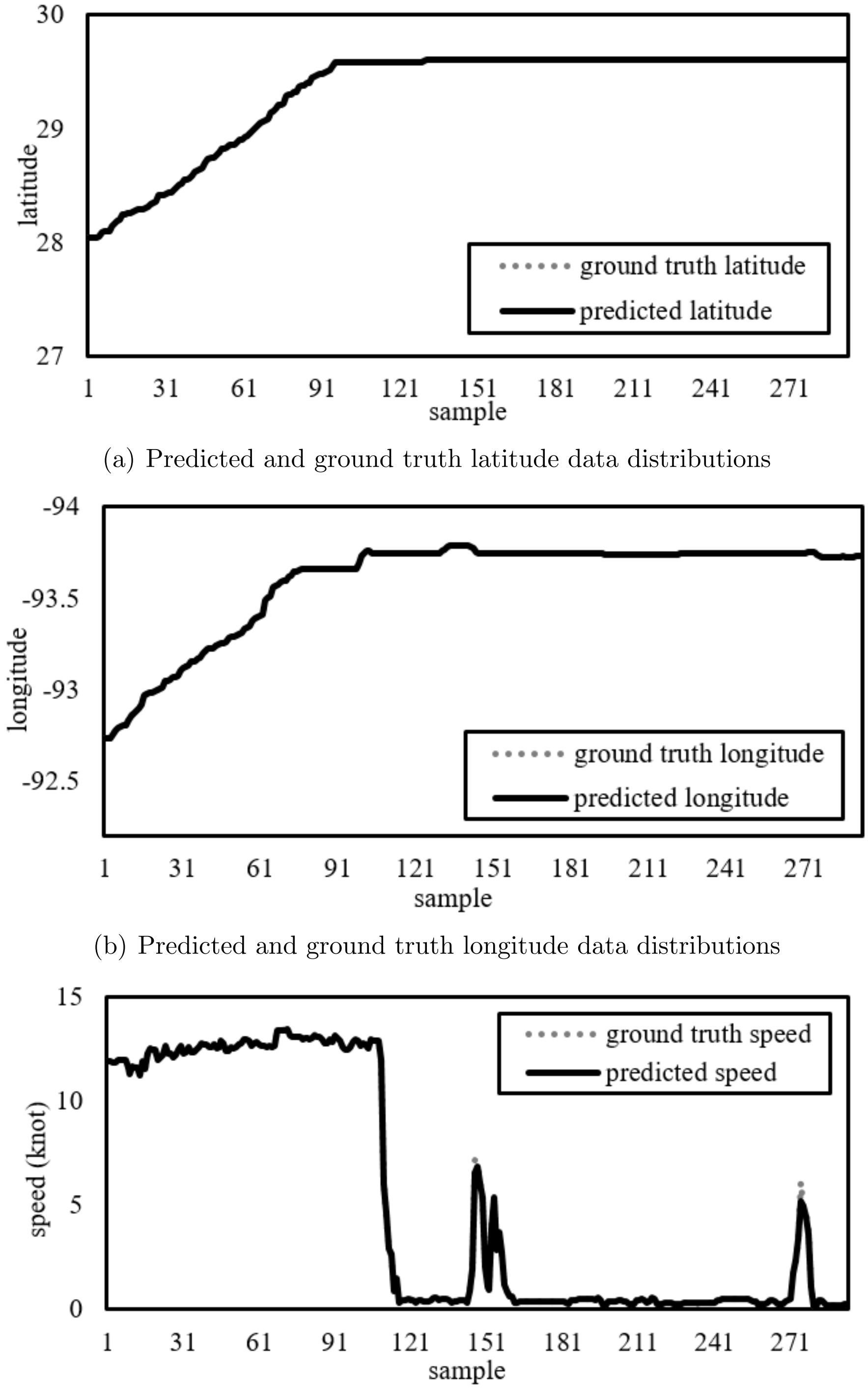
also outperformed the LSTM and SVM considering that the minimal MAPE, SMAPE, MAE, RMSE and MSE are 2.22×10^{-2}, 2.21×10^{-2}, 4.21×10^{-2}, 9.60×10^{-2} and 9.22×10^{-2}. From the aspects of longitude and speed prediction performance, we concluded that the SVM model obtained more accurate prediction results in comparison to the LSTM model. From the perspective of one-dimensional data prediction performance analysis (i.e., longitude, latitude and speed), the proposed model (i.e., HBA) outperformed the LSTM and SVM counterparts. The primary reason is that trivial outlier interference was successfully suppressed by the trajectory outlier removal procedure for the proposed HBA. In that manner, the HBA model learned intrinsic data features from the input AIS data, and thus obtained better data prediction accuracy. Both of the LSTM and SVM models extract features from the raw AIS data sample, which were contaminated by the trivial noises. More specifically, the two models failed to learn the intrinsic AIS data patterns due to trivial outlier interference.

The FRD and AEUD indicators presented accumulated ship trajectory prediction performance (see Table 4). The FRD for the proposed HBA model is 2.45×10^{-3}, and the counterparts for the LSTM and SVM are 2.07×10^{-2} and 3.28×10^{-2}. The FRD distribution showed that the LSTM and SVM prediction accuracy were both lower than that of the HBA model. The AEUD indicators distribution is consistent with that of the FRD, and the minimum AEUD is 4.57×10^{-2} (obtained by our proposed HBA model). The maximum AEUD is 5.39×10^{0} (obtained by the LSTM model), which confirmed our above-mentioned analysis. We presented the ground truth and predicted AIS data distributions which are shown in Figure 5. Note that we presented the predicted AIS data by our proposed HBA model considering that no significant difference can be observed from AIS distributions curves. Factually, our proposed model obtained good prediction accuracy due to that the ground truth and prediction curves were close to each other (see each subplot in Figure 5).

Table 4. Ship trajectory prediction accuracy distributions for different models

	FRD	AEUD
HBA	$\mathbf{2.45 \times 10^{-3}}$	$\mathbf{4.57 \times 10^{-2}}$
LSTM	2.07×10^{-2}	5.39×10^{0}
SVM	3.28×10^{-2}	1.15×10^{0}

3.3. **Ship trajectory cleaning and prediction for case 2 and case 3.** The proposed ship trajectory cleaning and prediction model was further verified on another two ship trajectory segments (i.e., case 2 and case 3). We did present speed data smoothing results in Figure 6 considering that the latitude and longitude smoothing results for different models were quite similar. It is noted that the HF (wavelet) smoothed speed data showed sharp variation from with samples ranging from 200 to 300 (350 to 400) which was observed in Figure 6(a). The raw speed data for the case 3 showed obvious oscillations, and thus the smoothing results for different models can be better identified. The black line in Figure 6(b) with sample interval from 50 to 100 indicated that speed obtained by the wavelet model was much larger than the counterparts. Moreover, the speed values for the last 50 wavelet-obtained samples were inconsistent with raw speed and the other counterparts (i.e., speed obtained by the HF, BF, HF+BF and EEMD). We consider the wavelet model provided more biased speed data considering that no obvious data outliers were observed by exploring the raw AIS data.



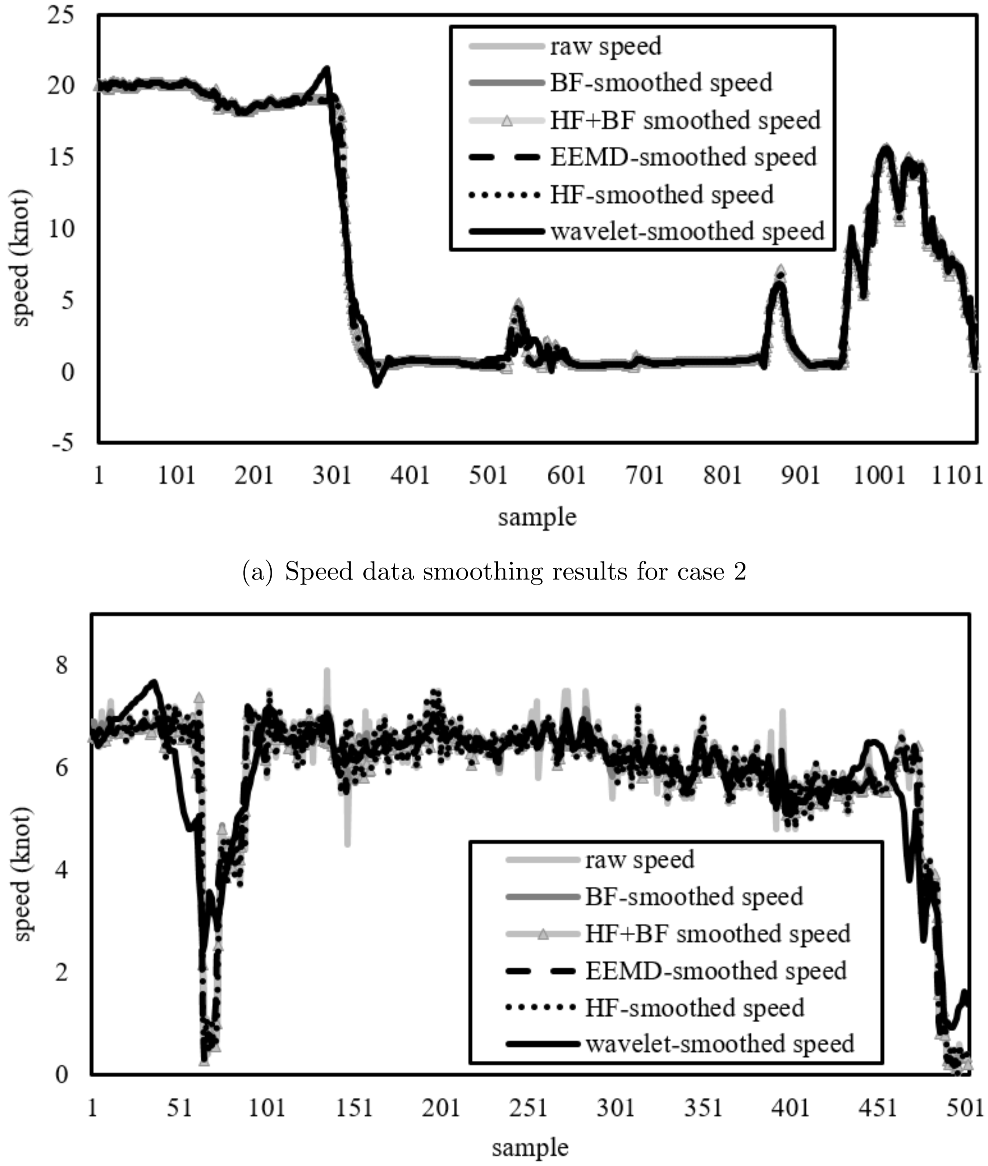
(a) Predicted and ground truth latitude data distributions

(b) Predicted and ground truth longitude data distributions

(c) Predicted and ground truth speed data distributions

Figure 5. The predicted and ground truth AIS data distributions

The latitude prediction performance for cases 2 and 3 were shown in Table 5. The MAPE, SMAPE, MAE, RMSE and MSE indicator distributions showed that HBA obtained higher prediction accuracy compared to those of the LSTM and SVM for both cases 2 and 3. For instance, the minimal MAPE for cases 2 and 3 were 8.80×10^{-8} and 6.25×10^{-7}, which were both achieved by the proposed HBA model. The longitude and speed prediction performance (see Tables 6 and 7) indicated that the proposed H-BA model outperformed the LSTM and SVM models for the cases 2 and 3 as well. For



(a) Speed data smoothing results for case 2

(b) Speed data smoothing for case 3

Figure 6. Speed data smoothing results for cases 2 and 3

Table 5. Latitude prediction accuracy distributions for case 2 and case 3

	HBA		LSTM		SVM	
	case 2	case 3	case 2	case 3	case 2	case 3
MAPE	8.80×10^{-8}	6.25×10^{-7}	8.91×10^{-4}	4.33×10^{-4}	4.99×10^{-5}	9.96×10^{-6}
SMAPE	8.80×10^{-8}	6.25×10^{-7}	8.91×10^{-4}	4.33×10^{-4}	4.99×10^{-5}	9.96×10^{-6}
MAE	7.00×10^{-5}	5.18×10^{-5}	2.91×10^{-2}	1.22×10^{-2}	3.93×10^{-3}	8.26×10^{-4}
RMSE	1.25×10^{-4}	8.13×10^{-5}	3.41×10^{-2}	1.26×10^{-2}	7.54×10^{-3}	1.64×10^{-3}
MSE	1.57×10^{-8}	6.61×10^{-9}	1.16×10^{-3}	1.57×10^{-4}	5.69×10^{-5}	2.69×10^{-6}

instance, the minimal MAPE and SMAPE for the longitude prediction task were both 6.03×10^{-6} (7.38×10^{-6}) for case 2 (case 3), and the maximal MAPE and SMAPE were 1.44×10^{-4} (6.22×10^{-5}) for case 2 (case 3). Similarly, the minimal MAE, RMSE and MSE for the longitude prediction task were obtained by the HBA model (see Table 6). The speed prediction results for the case 2 and case 3 were consistent with that of the case 1 (see Table 7), which verified the proposed framework model performance. Besides, the minimal FRD for case 2 and case 3 were 5.15×10^{-3} and 2.04×10^{-3}, respectively (see Table 8), while the minimum AEUD for the two cases were 6.08×10^{-2} and 3.30×10^{-2}. In sum, the latitude, longitude, speed and trajectory prediction performance for the cases 2 and 3 indicated that our proposed HBA model can successfully handle the ship trajectory cleaning and prediction task.

Table 6. Longitude prediction accuracy distributions for case 2 and case 3

	HBA		LSTM		SVM	
	case 2	case 3	case 2	case 3	case 2	case 3
MAPE	6.03×10^{-6}	7.38×10^{-6}	1.44×10^{-4}	6.22×10^{-5}	5.44×10^{-5}	1.94×10^{-5}
SMAPE	6.03×10^{-6}	7.38×10^{-6}	1.44×10^{-4}	6.22×10^{-5}	5.44×10^{-5}	1.94×10^{-5}
MAE	1.88×10^{-4}	2.04×10^{-4}	1.15×10^{-2}	5.16×10^{-3}	1.74×10^{-3}	5.35×10^{-4}
RMSE	5.91×10^{-4}	3.01×10^{-4}	1.63×10^{-2}	5.96×10^{-3}	3.57×10^{-3}	1.11×10^{-5}
MSE	3.50×10^{-7}	9.06×10^{-8}	2.65×10^{-4}	3.55×10^{-4}	1.27×10^{-5}	1.23×10^{-6}

Table 7. Speed prediction accuracy distributions for case 2 and case 3

	HBA		LSTM		SVM	
	case 2	case 3	case 2	case 3	case 2	case 3
MAPE	2.23×10^{-2}	3.06×10^{-2}	1.08×10^{-1}	3.32×10^{-1}	3.41×10^{-1}	1.23×10^{-1}
SMAPE	2.00×10^{-2}	3.17×10^{-2}	9.92×10^{-2}	5.35×10^{-1}	2.54×10^{-1}	1.693×10^{-1}
MAE	4.42×10^{-2}	8.85×10^{-2}	4.49×10^{-1}	4.68×10^{-1}	6.49×10^{-1}	4.10×10^{-1}
RMSE	8.82×10^{-2}	1.50×10^{-2}	7.03×10^{-1}	6.22×10^{-1}	1.24×10^{0}	7.84×10^{-1}
MSE	7.77×10^{-3}	2.24×10^{-2}	4.94×10^{-1}	3.87×10^{-1}	1.54×10^{0}	6.14×10^{-1}

Table 8. Ship trajectory prediction accuracy distributions for case 2 and case 3

	FRD		AEUD	
	case 2	case 3	case 2	case 3
HBA	5.15×10^{-3}	2.04×10^{-3}	6.08×10^{-2}	3.30×10^{-2}
LSTM	5.63×10^{-2}	1.95×10^{-2}	9.11×10^{0}	1.73×10^{0}
SVM	3.84×10^{-2}	7.27×10^{-3}	1.34×10^{0}	1.74×10^{-1}

4. **Conclusion.** The AIS data provides crucial ship trajectory information (i.e., latitude, longitude, speed, etc.) to various maritime traffic participants, and thus help them make early-warning decisions for the purpose of ensuring maritime traffic safety. However, the historical AIS data may be contaminated by unpredictable noises which reduce its usage

in tackling the task of exploiting ship trajectory spatial-temporal variation tendency. In this study, we proposed an ensemble ship trajectory cleansing and prediction framework via the steps of AIS noise removal, data normalization with HF and BF model, trajectory prediction with ANN method. Firstly, the proposed framework suppressed the noises from the raw AIS data with HF and BF models, while the former removed the impulsive noises and the latter removed the trivial outliers. Secondly, the denoised data was normalized for the purpose of obtaining AIS data samples with the same time interval. Thirdly, we predicted the ship trajectory variation tendency with the proposed framework. To evaluate the model performance, we implemented the ensemble framework on smoothing and predicting three ship trajectories. Moreover, typical smoothing models (i.e., HF, BF, EEMD and wavelet filter) and prediction methods (i.e., LSTM and SVM) were conducted on the same trajectories for the purpose of model performance comparison. The experimental results verified that our proposed framework obtained satisfied denoising and prediction performance.

The following directions can be further exploited to potentially enhance our model performance. First, we can introduce additional deep learning relevant models to determine intrinsic trivial noises in the raw AIS data series. Second, we can implement maritime situation awareness task with the support of the cleansed AIS data obtained by our framework. Last but not least, we can further testify the model performance under varied maritime environments and weather conditions.

Acknowledgments. This work was jointly supported by the National Key R&D Program of China (2019YFB1600605), National Natural Science Foundation of China (520712 00, 51978069, 52072237, 62073212), Shanghai Committee of Science and Technology, China (18DZ1206300, 18040501700, 18295801100).

REFERENCES

[1] L. Zhang et al., A novel ship trajectory reconstruction approach using AIS data, *Ocean Engineering*, vol.159, pp.165-174, 2018.

[2] H. Zhou and J. Wang, Non-coherent sequence detection scheme for satellite-based automatic identification system, *Journal of Systems Engineering and Electronics*, vol.28, no.3, pp.442-448, 2017.

[3] H. Tang et al., Detection of abnormal vessel behaviour based on probabilistic directed graph model, *Journal of Navigation*, vol.73, no.5, pp.1014-1035, 2020.

[4] Z. Wei, X. Xie and X. Zhang, AIS trajectory simplification algorithm considering ship behaviours, *Ocean Engineering*, vol.216, 2020.

[5] C. Lee and B.-D. Lee, Enhancement for automatic extraction of RoIs for bone age assessment based on deep neural networks, *ICIC Express Letters*, vol.14, no.2, pp.163-170, 2020.

[6] X. Wang and S. Zhang, Evaluation of multipath signal loss for AIS signals transmitted on the sea surface, *Ocean Engineering*, vol.146, pp.9-20, 2017.

[7] M. Gao and G.-Y. Shi, Ship collision avoidance anthropomorphic decision-making for structured learning based on AIS with Seq-CGAN, *Ocean Engineering*, vol.217, 2020.

[8] M. Gao and G.-Y. Shi, Ship-handling behavior pattern recognition using AIS sub-trajectory clustering analysis based on the T-SNE and spectral clustering algorithms, *Ocean Engineering*, vol.205, 2020.

[9] L. Wang et al., Use of AIS data for performance evaluation of ship traffic with speed control, *Ocean Engineering*, vol.204, 2020.

[10] F. Xiao et al., Comparison study on AIS data of ship traffic behavior, *Ocean Engineering*, vol.95, pp.84-93, 2015.

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

[12] D. Yang et al., 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.

[13] X. Chen et al., Robust ship tracking via multi-view learning and sparse representation, *Journal of Navigation*, vol.72, no.1, pp.176-192, 2019.

[14] Z. Liu, Z. Wu and Z. Zheng, Modelling ship density using a molecular dynamics approach, *Journal of Navigation*, vol.73, no.3, pp.628-645, 2019.

[15] Q. Hu et al., Study of an evaluation model for AIS receiver sensitivity measurements, *IEEE Trans. Instrumentation and Measurement*, vol.69, no.4, pp.1118-1126, 2020.

[16] P. Sheng and J. Yin, Extracting shipping route patterns by trajectory clustering model based on automatic identification system data, *Sustainability*, vol.10, no.7, 2018.

[17] Z. Yan et al., Exploring AIS data for intelligent maritime routes extraction, *Applied Ocean Research*, vol.101, 2020.

[18] H. Xu, H. Rong and C. G. Soares, Use of AIS data for guidance and control of path-following autonomous vessels, *Ocean Engineering*, vol.194, 2019.

[19] X. Q. Chen et al., Traffic flow prediction at varied time scales via ensemble empirical mode decomposition and artificial neural network, *Sustainability*, vol.12, no.9, 2020.

[20] X. Chen et al., Augmented ship tracking under occlusion conditions from maritime surveillance videos, *IEEE Access*, vol.8, pp.42884-42897, 2020.

[21] R. K. Pearson et al., Generalized hampel filters, *EURASIP Journal on Advances in Signal Processing*, no.1, 2016.

[22] B. Lin et al., A systematic approach for soft sensor development, *Computers & Chemical Engineering*, vol.31, nos.5-6, pp.419-425, 2007.

[23] X. Chen et al., Robust visual ship tracking with an ensemble framework via multi-view learning and wavelet filter, *Sensors (Basel)*, vol.20, no.3, 2020.

[24] H. Behjat et al., Domain-informed spline interpolation, *IEEE Trans. Signal Processing*, vol.67, no.15, pp.3909-3921, 2019.

[25] X. Chen et al., Traffic flow prediction at varied time scales via ensemble empirical mode decomposition and artificial neural network, *Sustainability*, vol.12, no.9, 2020.

[26] E. Tu et al., Exploiting AIS data for intelligent maritime navigation: A comprehensive survey from data to methodology, *IEEE Trans. Intelligent Transportation Systems*, vol.19, no.5, pp.1559-1582, 2018.

[27] X. Chen et al., Anomaly detection and cleaning of highway elevation data from Google earth using ensemble empirical mode decomposition, *Journal of Transportation Engineering, Part A: Systems*, vol.144, no.5, 2018.

[28] X. Chen et al., Sensing data supported traffic flow prediction via denoising schemes and ANN: A comparison, *IEEE Sensors Journal*, vol.20, no.23, pp.14317-14328, 2020.

[29] Y. Yang, J. Wang and Y. Yang, Exploiting rotation invariance with SVM classifier for microcalcification detection, *The 9th IEEE International Symposium on Biomedical Imaging (ISBI)*, 2012.