THE JOURNAL OF NAVIGATION (2018), **71**, 1210–1230. 
doi:10.1017/S0373463318000188

# Ship Trajectories Pre-processing Based on AIS Data

Liangbin Zhao, Guoyou Shi and Jiaxuan Yang

(*Navigation College, Dalian Maritime University, Dalian, China*)
(E-mail: vszlb@126.com)

Data derived from the Automatic Identification System (AIS) plays a key role in water traffic data mining. However, there are various errors regarding time and space. To improve availability, AIS data quality dimensions are presented for detecting errors of AIS tracks including physical integrity, spatial logical integrity and time accuracy. After systematic summary and analysis, algorithms for error pre-processing are proposed. Track comparison maps and traffic density maps for different types of ships are derived to verify applicability based on the AIS data from the Chinese Zhoushan Islands from January to February 2015. The results indicate that the algorithms can effectively improve the quality of AIS trajectories.

KEYWORDS

1. AIS Trajectories. 2. AIS error. 3. Data quality. 4. Pre-processing.

Submitted: 11 July 2017. Accepted: 12 March 2018. First published online: 22 April 2018.

1. INTRODUCTION. With the fast development of networking, data storage, and data collection capacity, "Big Data" is now rapidly expanding in all science and engineering domains. The techniques of data mining may give useful information for optimising industrial structures and improving production efficiency.

Receiving Automatic Identification System (AIS) messages from satellites is becoming increasingly common (Pallotta et al., 2013) and the establishment of AIS terminals at shore-based network centres makes it possible to collect large-scale AIS data. Mature information technology and hardware can provide unprecedented computing capability for processing data and consequently, many maritime researchers are investigating AIS data mining.

There are two main research areas concerning AIS data: analysis of ship trajectories based on spatiotemporal attributes and statistical analysis based on other attributes. Analysis of ship trajectories is used for diverse purposes such as mapping shipping density for maritime situational awareness (Wu et al., 2016; MMO, 2013; Shelmerdine, 2015; Vettor and Soares, 2015; Fiorini et al., 2016), characterising marine traffic patterns (De Souza et al., 2016; Tsou, 2010; Iperen, 2015; Aarsæther and Moan, 2009; Chen et al., 2015; Altan

and Otay, 2017; Breithaupt et al., 2017; Wang et al., 2017), anomaly detection (Pallotta et al., 2013; Zhang et al., 2016; Zhen et al., 2017; Ristic et al., 2008) and risk analysis (Zhou et al., 2013; Mou et al., 2010; Silveira et al., 2013; Mazaheri et al., 2015). An issue that these researchers have to deal with is cleaning raw AIS data, which is not easy, because of many erroneous AIS messages.

Bailey (2005) conducted research into reliability and completeness of AIS information in the Vessel Traffic Service (VTS) area of The Dover Strait. Results showed that more than 50% of ship destination information was incorrect, and 5% of messages contained a false Maritime Mobile Service Identity (MMSI) and course. Statistical research of AIS quality was made by Harati-Mokhtari et al. (2007). This study listed problems found based on a few months of data from Liverpool VTS and the AISLive company. For instance, more than one station broadcasting the same MMSI number creates discrepancies. Details show that information such as MMSI number, vessel type, position, are not reliable. An analysis by Banyś et al. (2012) focused on Heading (HDG) and Rate Of Turn (ROT) parameters. It was found that the AIS system is prone to receiving incomplete data broadcast by vessels' transmitters.

The quality of AIS data is a subject of interest for many researchers (Felski et al., 2015; Wawruch, 2017; Jaskólski, 2017; Peters and Hammond, 2011), but published research on pre-processing raw data to improve quality is limited. Shelmerdine (2015) took the development of a vessel database as the key to managing AIS data and for quality control. All fields were checked for obvious outliers. If it was not possible to correct an outlier, it was removed. The common method to filter inaccurate single position points is the gating of position, velocity and course (Sang et al., 2015; MMO, 2013). To solve the problem of sharing MMSI numbers, a method of elimination was applied by MMO (2013) and Pallotta et al. (2013). Mazzarella et al. (2014) proposed a nearest neighbour approach to assign AIS messages to the right tracks, but there was no detailed experimental method, performance or results. Wu et al. (2016) created a simple algorithm to calculate the likelihood of an association between an AIS message and each candidate vessel. It is used for processing massive data at a global scale but it cannot be applied in a small region where AIS messages are sampled at a high rate because the algorithm is unable to handle an association in the case where there are at least three consecutive abnormal track points. A similar method with velocity gating was proposed by Greidanus et al. (2016). Given that none of these techniques is universally applicable, it is necessary to propose a method with general applicability. The accuracy of time is the key to kinematic gating (Greidanus et al., 2016; Wu et al., 2016), but there are few studies on this. Most researchers remove data with the wrong MMSI and only focus on obvious mistakes. They seldom take account of the problem of temporal and spatial attributes in the dataset, such as time delay, boundary problems and other influencing factors.

The quality of the dataset is not only a key to the comprehensiveness of analysis but also the essential factor in avoiding misleading results. The lack of a systematic pre-processing method limits studies using AIS data mining. To improve the quality, this paper uses an overall error detection based on a month of AIS data from Zhoushan Islands and proposes methods for resolving various errors that were found.

The remainder of the paper is structured as follows. In Section 2, we summarise the errors in various quality dimensions of AIS data. In Section 3, the pre-processing methods are introduced. Section 4 shows the results of pre-processing AIS trajectories, and we conclude the paper in Section 5.

https://doi.org/10.1017/S0373463318000188 Published online by Cambridge University Press

2. AIS DATA MATERIALS AND ERRORS. All ships of 300 gross tonnes and above displacement engaged on international voyages, cargo ships of 500 gross tonnes and above displacement not engaged in international voyages and all passenger ships regardless of size are required to fit an AIS transceiver (IMO, 2003).

There are 27 different types of AIS messages which can be broadcast. The content of AIS messages varies widely according to the type (Iphar et al., 2015). This paper mainly focuses on position reports (message 1, message 18) and static information reports (message 5, message 19). Of the fields in these reports, seven fields (MMSI, longitude, latitude, Speed over Ground (SOG), hour, minute, and second) are considered essential for representing AIS trajectory and two fields (type, length) supply additional information for the corresponding trajectory.

Our data was collected from the AIS data centre of the Chinese Maritime Authority in Ningbo. The research area is shown in Figure 1. The data is from January to February 2015. The quantity of raw data in the research area is given in Table 1. Figure 2 shows the raw AIS track data of the first week apart from the abnormal tracks (see Figure 4) which would obscure the other tracks in the figure.

There are many errors in the raw AIS track dataset. We conducted a study of AIS data quality, and the findings are summarised below, organised by AIS quality dimension.


Figure 1. Research area.

Table 1. The quantity of raw data in the research area.

| Class | The number of position reports | The number of static information reports | The number of ships | The number of ships with additional information |
|---|---|---|---|---|
| A | 87,434,689 | 33,591,924 | 5,665 | 5,385 |
| B | 23,189,153 | 7,050,398 | 15,919 | 4,080 |
| Total | 110,623,842 | 40,642,322 | 21,584 | 9,465 |

https://doi.org/10.1017/S0373463318000188 Published online by Cambridge University Press

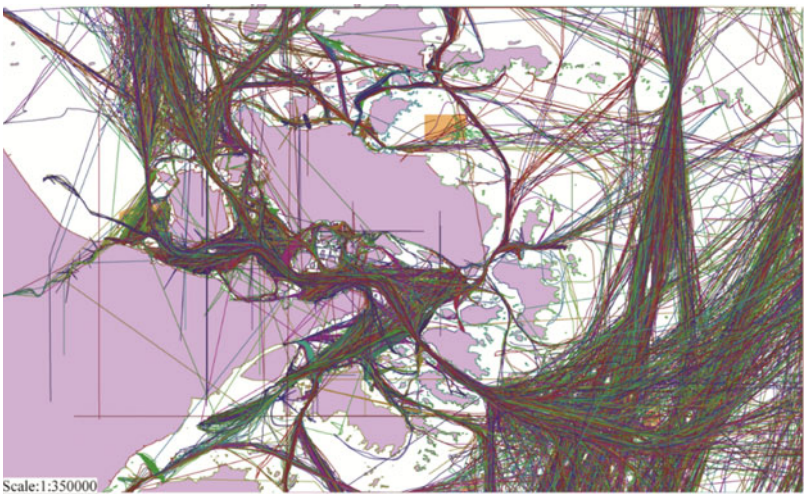



Figure 2. Distribution map of raw AIS track data from 1–7 January 2015.

Table 2. Statistical result of unreliable messages.

| Field | Invalid value | The number of messages | Proportion |
|---|---|---|---|
| SOG (kn) | [102·3, ∞) | 37,460 | 0·03% |
| Second (s) | [60, ∞) | 2,841,159 | 2·57% |
| Min (minute) | [60, ∞) | 36,937 | 0·03% |
| Hour (hour) | [24, ∞) | 36,878 | 0·03% |

Table 3. Statistical result of ships with unreliable additional information.

| Field | The number of ships | Proportion |
|---|---|---|
| Type | 795 | 8·40% |
| Length | 517 | 5·46% |

2.1. *Physical integrity.* Physical integrity is a measure of the degree of validity of an individual track and AIS message, which includes reliability of the AIS message and completeness of the track.

2.1.1. *Reliability of AIS message.* Reliability means the general coherence of the AIS message with respect to ITU recommendations (ITU, 2010). Our experiment focuses on fields including MMSI, SOG, Longitude, Latitude, Second, Minute, Hour, Length of ship and Type of ship. For example, 4·25% of total AIS type 1 messages that we collected from AIS base stations had latitude or longitude values larger than the maximum allowable values of 90˚ and 180˚. We also found errors in other fields, and the proportion of messages that contain these errors and ships with unreliable additional information in the research area are presented in Tables 2 and 3.

2.1.2. *Completeness of track.* The completeness of track is defined by two standards. If the number of track points in a track is too small, or there is no corresponding static and

https://doi.org/10.1017/S0373463318000188 Published online by Cambridge University Press

Table 4. Random errors in a track segment (MMSI: 413303240)

| No. | Time | Longitude (East) | Latitude (North) | Notes |
|---|---|---|---|---|
| 1 | 2015/1/16 21:00:47 | 122·2481 | 29·9288 | |
| 2 | 2015/1/16 21:00:51 | 122·2481 | 29·7604 | fluctuation |
| 3 | 2015/1/16 21:02:50 | 122·2481 | 29·9288 | |
| 4 | 2015/1/16 21:06:55 | 122·2481 | 29·9288 | |
| 5 | 2015/1/16 21:07:01 | 122·2481 | 29·9288 | |
| 6 | 2015/1/16 21:07:37 | 122·2481 | 29·7621 | fluctuation |

voyage related information report (Table 3), the track will be considered as a track that lacks completeness, because it cannot represent track features of interest.

2.2. *Spatial logical integrity.* Spatial logical integrity means the extent of the correctness of time-space relationship between the messages in a sequential set of trajectory points, which includes accuracy, consistency, and relevance.

2.2.1. *Accuracy of track.* Considering an attribute a of an entity e, its standard value is v. The accuracy of a v' value would be the degree of closeness of v' with respect to v. In the case of an AIS trajectory, we take the number of logical track points in a trajectory as the measurement of the degree of accuracy. If all the track points are logical, the accuracy is maximal. However, in general, logical points form the majority of the whole track. Consequently, we determine the accuracy of the track based on whether illogical points are present.

We found that there are random errors of position in a proportion of AIS tracks; their values of latitude or longitude changed illogically, as shown in Table 4. There are illogical positions in the chart of the track, as illustrated in Figure 3(a). In addition, there are also consecutive outliers in the ship movement (Figure 3(b)) that can be recognised by kinematic gating. The results show that 12·36% of ships generate such errors.

2.2.2. *Consistency of track.* In an AIS trajectory, the source of the different track points must agree. The MMSI number is a unique number given to every vessel for identification. It is the sole means of discrimination between AIS ship stations, which is usually used to extract ships' trajectories. However, because of improper use, deliberate or accidental, it is observed that the same MMSI number can be used by different ships. If those data are collected in the same observation period, the ship may jump between multiple positions on the chart. Trajectories of ships with the same MMSI are illustrated in Figure 4.

In our research, there are two types of position reports (class A and class B) from different types of shipborne AIS devices, which are obviously from different ships. However, we found 127 MMSI numbers in both types of reports, such as 413000000, 88888888, and 413999990.

2.2.3. *Relevance of track.* Relevance can measure the relational degree between two data objects. In the case of AIS trajectory, the problem of the time relationship between track points in a track is inevitable. Not all of the track data are completely preserved in the research area. For example, ships may cross the borderline repeatedly in the period, and this results in a loss of data. The tracks that lack relevance are illustrated in Figure 5, which includes the illogical tracks crossing the land and along the borderline.

2.3. *Accuracy of time.* AIS messages collected from AIS receiving stations are usually marked with an external time stamp, which is called recorded time. In the process of generating an AIS message, a communication time stamp is coded into the sentence, which

https://doi.org/10.1017/S0373463318000188 Published online by Cambridge University Press

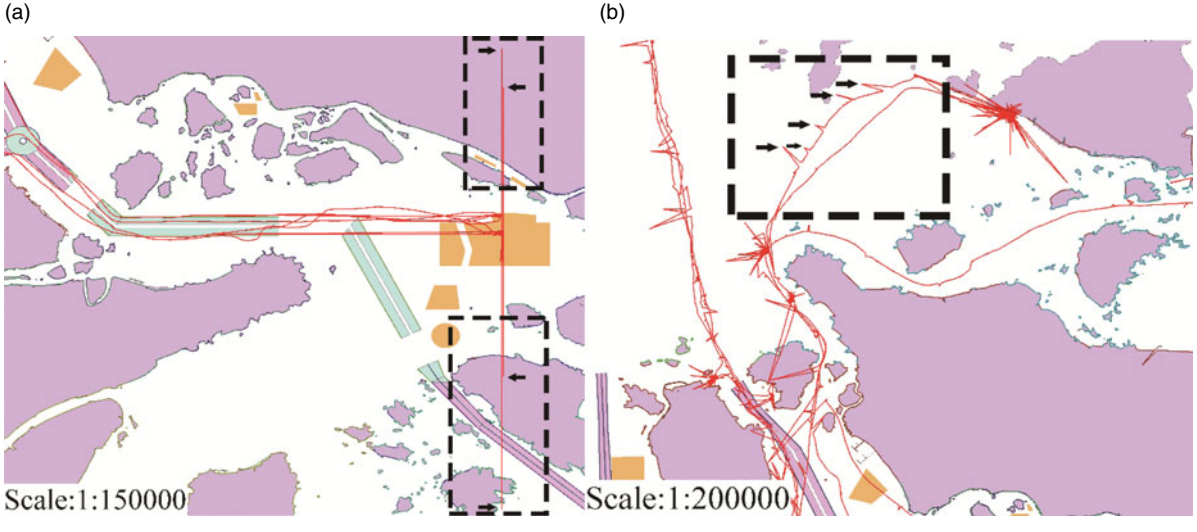



Figure 3. Random errors of position in AIS tracks.


Figure 4. Trajectories of multiple ships which share the same MMSI. (MMSI: 413000000)

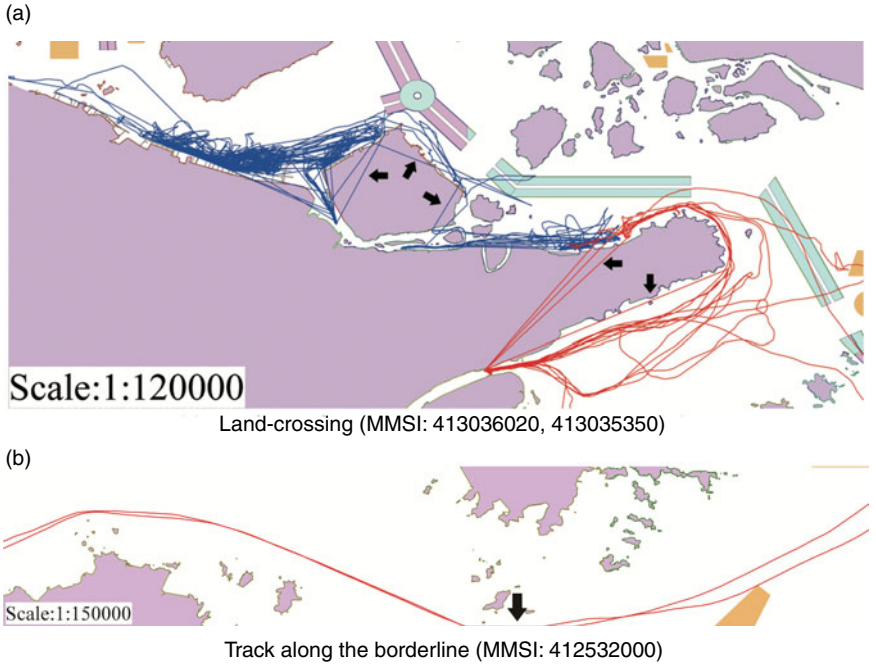



Figure 5. Illogical track segments which are generated by two irrelevant track points.

https://doi.org/10.1017/S0373463318000188 Published online by Cambridge University Press

Table 5. Statistical results of all the deviations.

| Deviation | Frequency | Proportion |
|---|---|---|
| less than −2 | 600,412 | 2·00% |
| −2 | 2,846,703 | 9·48% |
| −1 | 9,999,484 | 33·31% |
| 0 | 8,155,762 | 27·17% |
| 1 | 4,790,758 | 15·96% |
| 2 | 1,827,850 | 6·09% |
| 3 | 627,036 | 2·09% |
| larger than 3 | 1,172,581 | 3·91% |

Table 6. Large deviations of consecutive track points (MMSI: 413468760).

| | Recorded time | | | Generated time | | | |
|---|---|---|---|---|---|---|---|
| No. | Hour | Minute | Second | Hour | Minute | Second | deviation |
| 1 | 7 | 18 | 46 | 5 | 31 | 27 | 6,439 |
| 2 | 7 | 21 | 5 | 5 | 33 | 47 | 6,438 |
| 3 | 7 | 23 | 45 | 5 | 36 | 26 | 6,439 |
| 4 | 7 | 25 | 54 | 5 | 38 | 35 | 6,439 |
| 5 | 7 | 26 | 14 | 5 | 38 | 55 | 6,439 |
| 6 | 7 | 27 | 34 | 5 | 40 | 15 | 6,439 |

is called generated time. It is observed that the time data of AIS messages have errors. Greidanus et al. (2016) compared AIS messages of the same ship received by different stations, and they found that these external time stamps can be inconsistent due to clock offsets and instabilities. Their results show that the quality of time information is related to the receiving station; the time of generation cannot be recorded precisely. This result is verified in our research.

A portion of position reports contain information of generated time including second, hour and minute (ITU, 2010). In our research, 15·66% of class A position reports contain such information. However, the proportion for class B is only 1·15%. We determine the time accuracy of a whole track based on a sequential set of all the deviations between the recorded time (in seconds) and the generated time (in seconds). The deviation is calculated as follows.

$$\begin{aligned} deviation = &(hour_{recorded} * 3600 + min_{recorded} * 60 + second_{recorded}) \\ &- (hour_{generated} * 3600 + min_{generated} * 60 + second_{generated}) \end{aligned} \quad (1)$$

The statistical results of all the deviations is presented in Table 5. 94·09% of deviations are in the range of [−2, 3], which means that most records of time are almost the same as generated time. However, there are also large deviations which cannot be ignored in our research. Results show that the large deviations in some ranges are relatively more frequent, which means there are systematic errors in the time dataset. As illustrated in Table 6, the recorded time and generated time of those track points in the case of one exemplar MMSI demonstrate a more or less consistent deviation of 6,439s. These systematic errors may be caused by signal delay.

https://doi.org/10.1017/S0373463318000188 Published online by Cambridge University Press

Table 7. Abnormal change around 00:00 (MMSI: 412036180).

| | Recorded time | | | Generated time | | | |
|---|---|---|---|---|---|---|---|
| No. | Hour | Minute | Second | Hour | Minute | Second | deviation |
| 1 | 23 | 58 | 28 | 23 | 57 | 5 | 83 |
| 2 | 23 | 59 | 59 | 23 | 58 | 35 | 84 |
| 3 | 0 | 0 | 0 | 23 | 58 | 35 | −86,315 |
| 4 | 0 | 1 | 19 | 23 | 59 | 56 | −86,317 |
| 5 | 0 | 2 | 31 | 0 | 1 | 6 | 85 |
| 6 | 0 | 3 | 40 | 0 | 2 | 15 | 85 |

Table 8. Generated times that turned into a zero value (MMSI: 413442770).

| | Recorded time | | | Generated time | | | |
|---|---|---|---|---|---|---|---|
| No. | Hour | Minute | Second | Hour | Minute | Second | deviation |
| 1 | 19 | 17 | 43 | 19 | 17 | 43 | 0 |
| 2 | 19 | 18 | 52 | 19 | 18 | 52 | 0 |
| 3 | 19 | 19 | 23 | 0 | 19 | 23 | 68,400 |
| 4 | 19 | 20 | 1 | 19 | 20 | 2 | −1 |
| 5 | 19 | 22 | 13 | 19 | 22 | 13 | 0 |
| 6 | 19 | 23 | 42 | 19 | 23 | 43 | −1 |

Table 9. Fields becoming zero consecutively (MMSI: 229733000).

| | Recorded time | | | Generated time | | | |
|---|---|---|---|---|---|---|---|
| No. | Hour | Minute | Second | Hour | Minute | Second | deviation |
| 1 | 8 | 33 | 27 | 8 | 33 | 25 | 2 |
| 2 | 8 | 34 | 54 | 0 | 0 | 0 | 30,894 |
| 3 | 8 | 38 | 6 | 0 | 0 | 0 | 31,086 |
| 4 | 8 | 38 | 34 | 8 | 38 | 33 | 1 |
| 5 | 8 | 38 | 44 | 0 | 0 | 0 | 31,124 |
| 6 | 8 | 39 | 27 | 8 | 39 | 25 | 2 |

The deviation itself has errors in calculation. Firstly, because the time information only contains the values of hour, minute and second, an abnormal change may appear around 00:00 hours when the calculation is made based on time values that belong to different days, as illustrated in Table 7. Secondly, some fields in generated time may suddenly become zero (Table 8), and this phenomenon may appear consecutively (Table 9). Thirdly, tracks of different deviations may be linked together (Table 10). These random errors may be caused by information missing in the process of generating AIS messages.

3. PRE-PROCESSING METHOD. Data quality is a measure of the extent to which a database accurately represents the essential properties of the intended application (Brodie, 1980). To improve the quality of AIS trajectories for better data mining studies, this paper proposes the following methods, organised by AIS quality dimension.

3.1. *Physical integrity.* The method in our research for improving physical integrity is to discard AIS messages that are not compatible with the ITU standard and tracks that lack completeness.

https://doi.org/10.1017/S0373463318000188 Published online by Cambridge University Press

Table 10. Different deviations in the same period (MMSI: 413410960).

| No. | Recorded time | | | Generated time | | | deviation |
|---|---|---|---|---|---|---|---|
| | Hour | Minute | Second | Hour | Minute | Second | |
| 1 | 7 | 2 | 37 | 7 | 2 | 36 | 1 |
| 2 | 7 | 3 | 48 | 7 | 3 | 46 | 2 |
| 3 | 7 | 3 | 57 | 7 | 3 | 56 | 1 |
| 4 | 7 | 10 | 19 | 7 | 8 | 15 | 124 |
| 5 | 7 | 10 | 30 | 7 | 8 | 25 | 125 |
| 6 | 7 | 11 | 39 | 7 | 9 | 35 | 124 |

As described in Section 2.1.2, we judge the completeness of a track based on the number of track points and whether there is corresponding additional information. It is easy to collect large amounts of AIS data over a short time, because the AIS data sampling rate is very high (2 s −10 s). However, there are reasons which may cause a shortage of track points, such as signal loss and a brief stay in the research area. We believe that short tracks whose number of track points is less than 100 cannot characterise the movement of a ship. The threshold of completeness in pre-processing space data is empirically set as 100. That is to say, the track will be discarded if the number of track points is less than 100.

Additional information is included in the static and voyage-related information report (message 5) and extended class B equipment position report (message 19). The update rate of those reports is once every 6 min, so not every position report has a timestamp close to the additional information. After retrieving a record by the MMSI and timestamp of the initial and last track point, if we cannot find all the valid information that we need, the track will be discarded.

3.2. *Spatial logical integrity.* To solve the problems in Section 2.2 such as abnormal individual points, tracks that lack relevance and sharing of MMSIs, a method which can quickly process large amounts of historical AIS data is proposed, see Algorithm 1. The method processes the time data first, and then the space data, which consists of three parts, as shown in Figure 6. The first part is the partition, and a breakpoint is found to split the track into sub-tracks (see Algorithm 2 and the partition part in Figure 6). The breakpoint is determined based on the thresholds of space (speed gate: 15 knots) and time (10 min), which are set empirically. A sub-track can be an individual point or a track segment. The second part is the association of sub-tracks. All the sub-tracks will be judged by the threshold of space based on the last track point and the first track point of all the following sub-tracks (see Algorithm 3 and the association part in Figure 6. The big arrow between two sub-tracks means they can be associated). If the judgment condition is met, the sub-tracks can be associated with each other. The judgment of the new sub-track will continue until the individual association is done. The third part is filtering tracks that lack completeness, as described in Section 3.1 (see filtering part in Figure 6(a), the track that lacks completeness is marked with a note 'outlier').

3.3. *Accuracy of time.* To solve the problem in Section 2.3, a method of correcting time data is proposed based on the deviation values between the recorded time and generated time of AIS messages, see Algorithm 4. D is a set of track points, and there are three steps including obtaining deviation results, cleaning deviation data and getting a value for correction.

https://doi.org/10.1017/S0373463318000188 Published online by Cambridge University Press

Algorithm 1. Space Data Pre-processing.

**Require:** track points list $D = [Pt_0, \ldots, Pt_i]$, *threshold_time* = 10 min, *threshold_completeness* = 100, *threshold_space* = 15 *knots*

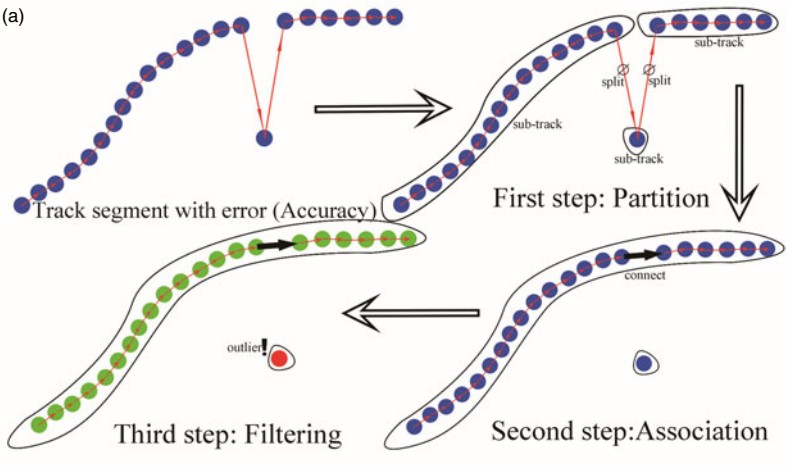
```
// cleaning time data
tracks_time = Partition(D, threshold_time)
// filtering of physical integrity
for each subtrack in tracks_time do
    if the number of Pt in subtrack is less than threshold_completeness then
        Remove subtrack from tracks_time
    end if
end for
// cleaning space data
for each subtrack in tracks_time do
    subtracks_space = Partition(subtrack, threshold_space)
    tracks_space = Association(subtracks_space, threshold_space,
    threshold_completeness)
    Add tracks_space into the queue Output
end for
return Output
```

Firstly, the time is converted to seconds for calculating deviation using Equation (1), and the calculation of the offset is made to remove the error caused by crossing 00:00 hours. The data whose fields of generated time are all zero and where the absolute value of deviation is larger than five will be discarded. Then a pre-processing method consisting of partition (Algorithm 2) and association (Algorithm 3) is used to remove the errors in deviation data. (The threshold of completeness in pre-processing time data is set empirically as five). Finally, the corrected value will be determined by the two statistical values of the deviation dataset which are mean and range. The range can measure the stability of the deviations between recorded time and generated time. The smaller the value of the range, the more stable the track's deviations are, which means the average of the deviations can more accurately measure the time accuracy of an entire track.

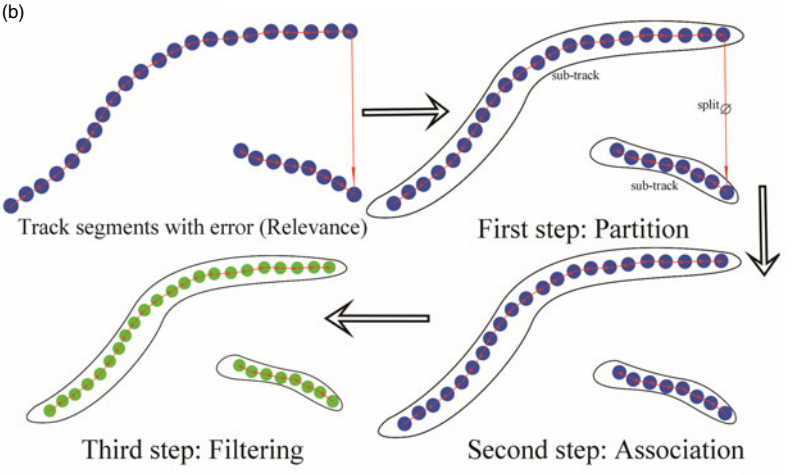
Most deviations are in the range of $[-2, 3]$, so the threshold of deviation in Algorithm 4 is set as 5 seconds. The statistical results of the range of deviation based on the track data after Algorithm 1 is shown in Table 11. 96% of ranges are less than or equal to five, so we also consider 5 seconds as the threshold of range in the deviation dataset. If the range is less than or equal to five, the negative value of the mean of deviations will be considered as the corrected value.

However, there are local changes in the density distribution of deviations in some tracks, so the value of the mean may become relatively invalid in those tracks whose range is larger than five, as illustrated in Figure 7 (units on the x-axis are seconds). In such cases, the corrected value will be discussed according to the mean value of deviations. Most deviations are in the range $[-2, 3]$ so if the mean remains in this range, the local change will be ignored. However, if the mean is out of the range of $[-2, 3]$, the minus value that has the highest probability in the set of deviations will be considered as the corrected value based

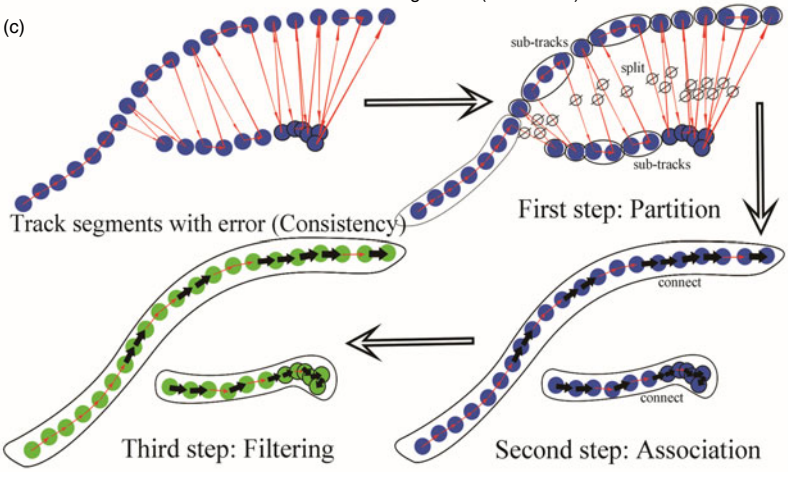
https://doi.org/10.1017/S0373463318000188 Published online by Cambridge University Press




Process of removing errors (Accuracy)




Process of removing errors (Relevance)




Process of space data pre-processing

Figure 6. Process of space data pre-processing.

https://doi.org/10.1017/S0373463318000188 Published online by Cambridge University Press

Algorithm 2. Partition.

**Require:** track points list $D = [Pt_0, \ldots, Pt_i]$, *threshold_partition*

**for each** *Pt* in *D* **do**
  **if** *Pt* has previous *track point* **then**
    Calculate the *difference_value* based on *Pt* and the previous *track point*
// *difference_value* is depend on data type. Space data: Difference between the recorded *speed* of *Pt* and calculated speed based on positions and time, Time data: Interval, The time deviation: Difference of time deviations.
    **if** *difference_value* > *threshold_partition* **then**
      Add the serial number of *Pt* into the queue *Breakpts*
    **end if**
  **end if**
**end for**
Split the *D* into subqueue *subtracks* based on *Breakpts*
Add all the *subtracks* into the queue *C*
**return** *C*

Algorithm 3. Association.

**Require:** Track list $T = [Track_0, \ldots, Track_i]$, *threshold_association*, *threshold_completeness*

**while** (*T*) **do**
  Let *M* be the first *track* in *T*
  Add *M* into the queue *boat*
  Remove *M* from *T*
  **if** (*T*) **then**
    **for each** *track* in *T* **do**
      Calculate *association_value* based on the first *point* in *track* and the last *point* in *boat*
// *association_value* depends on data type. Space data: Difference between the recorded *speed* of the first *point* in *track* and calculated speed based on positions and time, The time deviation: Difference of the time deviations.
      **if** *association_value* <= *threshold_association* **then**
        Add *track* into *boat*
        Remove *track* from *T*
      **end if**
    **end for**
// filtering of physical integrity
    **if** the number of *point* in *boat* is larger than *threshold_completeness* **then**
      Update the *identifier* of all the *points* in *boat*
      Add *boat* into the queue *C*
    **end if**
    Let *boat* be an empty queue
  **end if**
**end while**
**return** *C*

https://doi.org/10.1017/S0373463318000188 Published online by Cambridge University Press

Algorithm 4. Time correction.

**Require:** track points list $D = [Pt_0, \ldots, Pt_i]$, *threshold_deviation* = 5, *threshold_completeness* = 5, *threshold_random_error* = 5, *threshold_range* = 5, *MSH*(Minus number of seconds of half day) = −43200, *SH*(Number of seconds of half day) = 43200, *offset*(Number of seconds of whole day) = 86400

```
// get deviation results
for each Pt in D do
    if Pt has the generated time then
      R_time and G_time is the number of seconds of recorded time and generated
      time
      Pt.deviation = R_time − G_time
      // offset
      if Pt.deviation < MSH then
        Pt.deviation = Pt.deviation + offset
      end if
      if Pt.deviation >= SH then
        Pt.deviation = Pt.deviation − offset
      end if
// delete random error
      if all the fields of generated time of Pt are zero and |Pt.deviation| >
      threshold_ random_error then
        Pt is marked as error
      else
        Add Pt into the queue set_deviation
      end if
    end if
end for
// cleaning deviation data
subsets_deviation = Partition (set_deviation, threshold_ deviation)
sets_deviation = Association (subsets_deviation, threshold_deviation,
threshold_completeness)
// getting corrected value
for each set in sets_deviation do
    Calculate mean and range of Pt.deviation in set
      if range <= threshold_range then
        Add the minus value of mean into corrected_value_set
      elseif mean is in the range of [−2, 3] then
        Add the minus value of mean into corrected_value_set
      else
        Find the value_kde that has the highest probability in set based on kernel
        density estimation
        Add the minus value of value_kde into corrected_value_set
      end if
end for
return corrected_value_set
```

https://doi.org/10.1017/S0373463318000188 Published online by Cambridge University Press

Table 11. Statistical results of range.

| Range | Frequency | Proportion |
|---|---|---|
| 0 | 222 | 0·40% |
| 1 | 5,078 | 8·80% |
| 2 | 20,409 | 35·50% |
| 3 | 20,329 | 35·40% |
| 4 | 6,707 | 11·70% |
| 5 | 2,418 | 4·20% |
| larger than 5 | 2,299 | 4·00% |

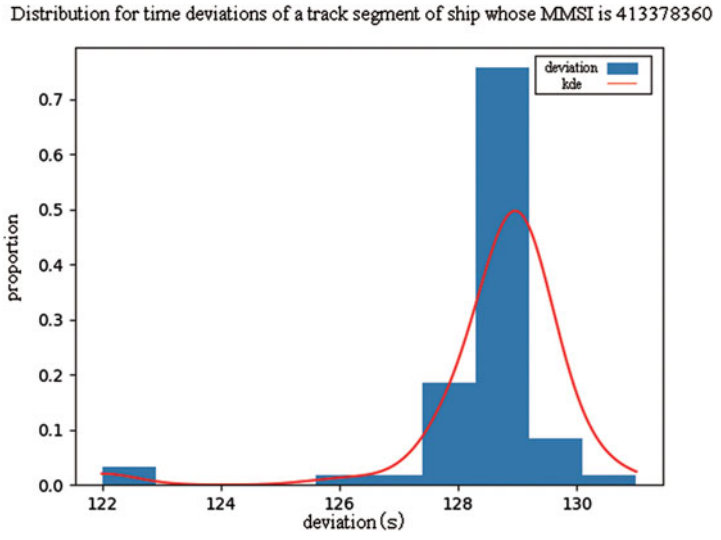



Figure 7. A distribution of deviations whose range is larger than five.

on kernel density estimation, as illustrated in Figure 7 (KDE curve). The set of deviations is considered as $X = \{x_1, x_2 \ldots x_n\}$. The probability can be calculated as follows:

$$f(x) = \frac{1}{n\eta\sqrt{2\pi}} \sum_{i=1}^{n} \exp\left(-\frac{(x - x_i)^2}{2\eta^2}\right) \tag{2}$$

$$\eta = \left(\frac{4}{3n}\right)^{\frac{1}{5}} \sigma \tag{3}$$

where $\sigma$ is the sample variance and the kernel function is a Gaussian kernel function.

It is worth mentioning that, in our period of observation, there were 269,493 messages whose number of seconds of generated time is larger than or equal to 60, that were not compatible with the ITU standard and discarded at the beginning of pre-processing. However, the proportion of these errors in the messages that contain generated time is very small (1·9%), and most tracks still have sufficient valid messages to calculate the deviation. Consequently, those errors will not influence the method.

https://doi.org/10.1017/S0373463318000188 Published online by Cambridge University Press

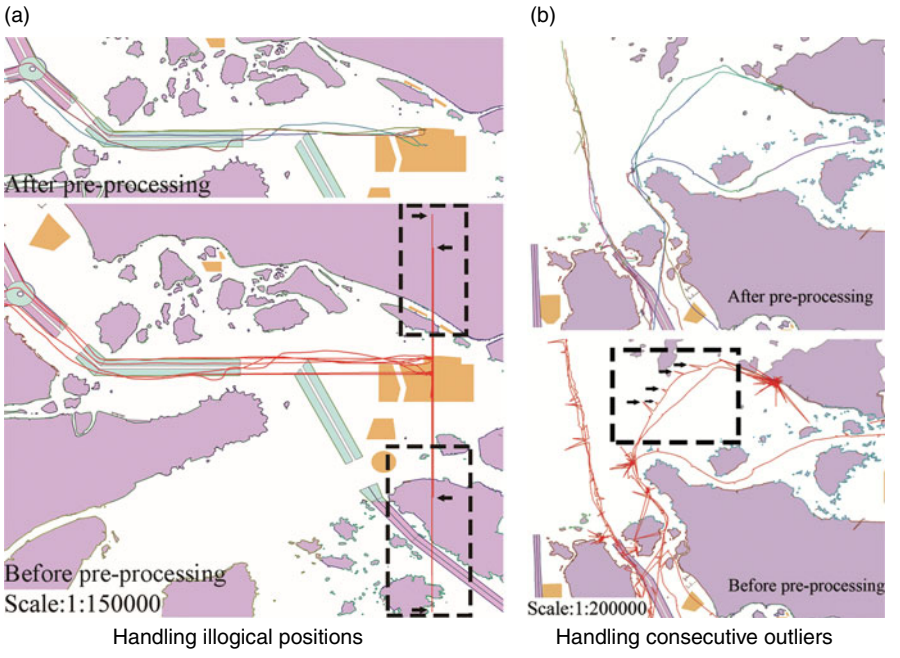



Figure 8. Pre-processing effect in terms of track accuracy

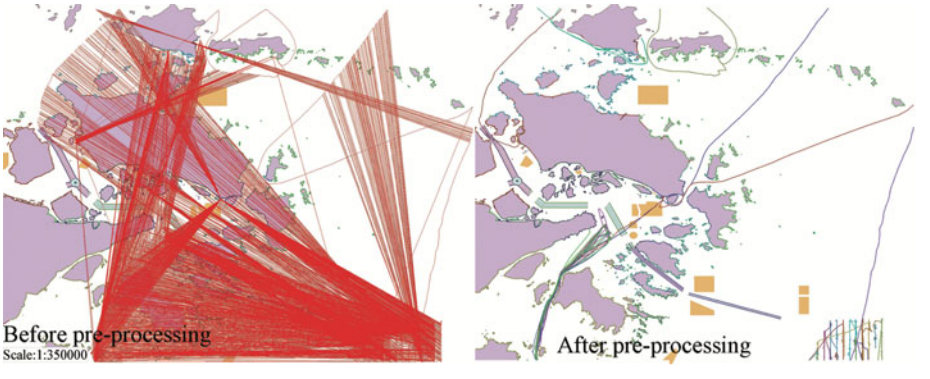



Figure 9. Pre-processing effect in terms of track consistency.

4. RESULTS. In our research, there were 110,623,842 position report messages from 21,585 distinct MMSIs. After pre-processing, 106,245 track segments were output. 1·27% of the track points are discarded in space data pre-processing, and only 0·42% of the track points were discarded in the time correction. The detailed results are shown from two aspects: space and time.

4.1. *Results of space.* Based on the method proposed in Section 3.2, the distribution maps of AIS trajectories are shown in Figures 8–10.

Compared with the tracks in Figure 3, it is obvious that the abnormal track points have been discarded. The quality of track accuracy is improved (Figure 8).

https://doi.org/10.1017/S0373463318000188 Published online by Cambridge University Press

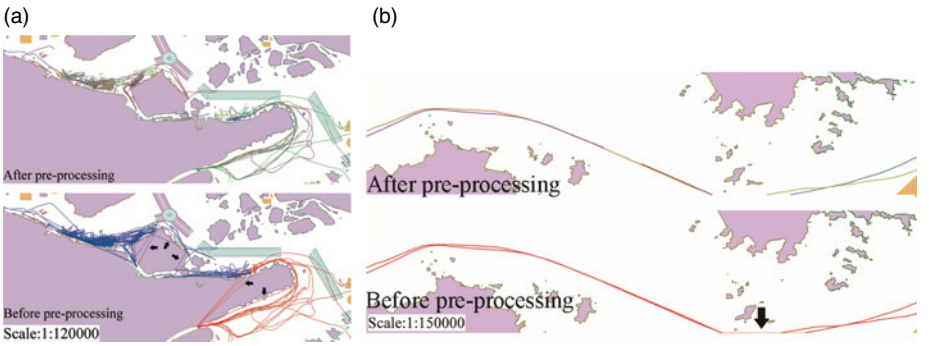



Figure 10. Pre-processing effect in terms of track relevance.

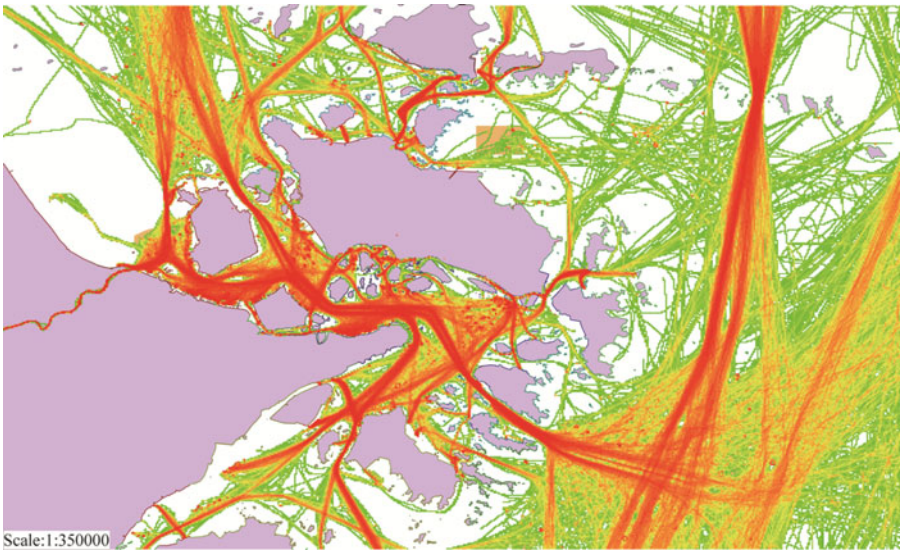



Figure 11. Density map of AIS track data after pre-processing (1–7 January 2015).

Compared with the tracks in Figure 4, chaotic tracks of ships with the same MMSI have been recognised and split into several regular tracks with different colours (Figure 9). The quality of track consistency is improved.

Compared with the tracks in Figure 5, it is obvious that the problems of crossing land and unreal tracks along the borderline have been eliminated (Figure 10). The quality of track relevance is improved. The density maps of processed tracks are shown at resolutions of 0·2 $km^2$.

Figure 11 illustrates the density of AIS trajectories collected from 1–7 January 2015. All the popular traffic routes are recognised and marked in red. Also, the distribution of traffic volume is distinctly shown. The areas where ships often stop, such as anchorages, are highlighted in the form of a red spot.

https://doi.org/10.1017/S0373463318000188 Published online by Cambridge University Press

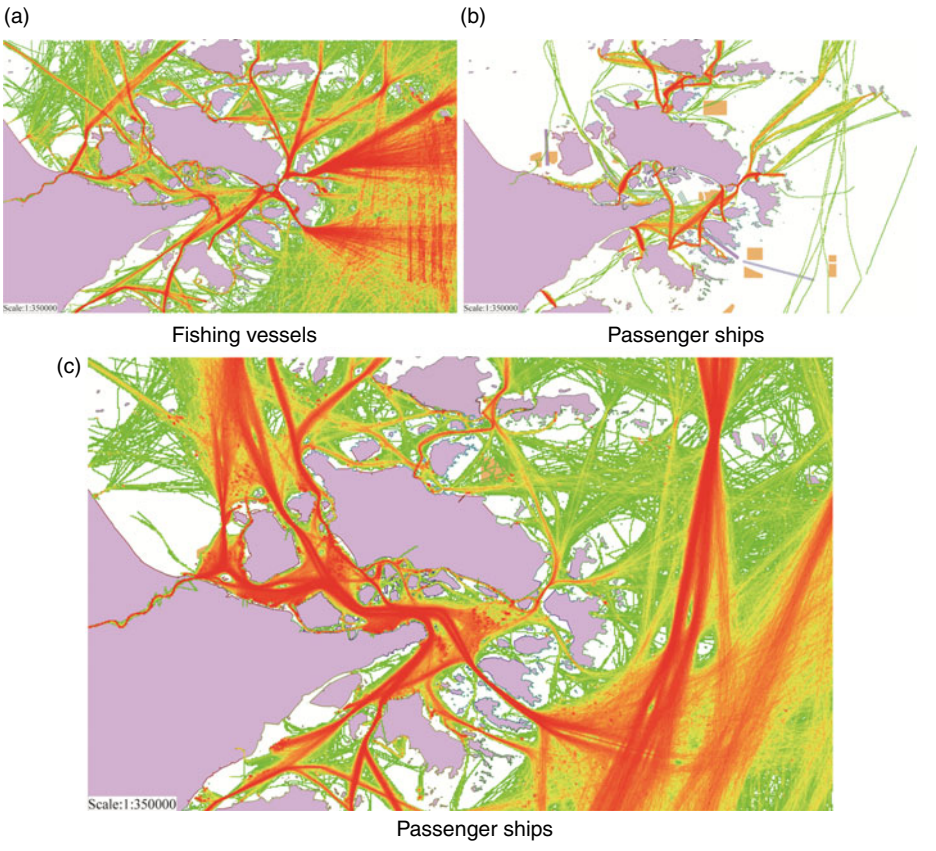


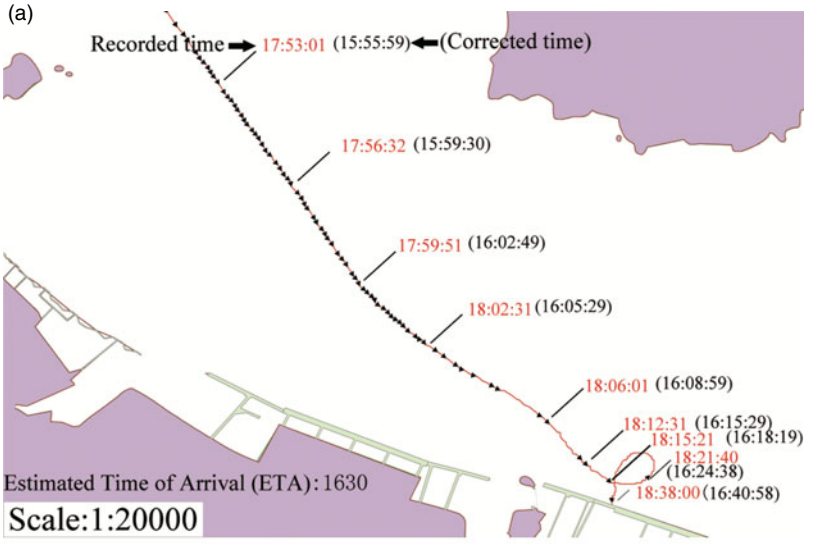

Figure 12. Density map of different types of ships based on AIS track data after pre-processing (January 2015).

Based on AIS trajectories collected from January to February 2015, different types of density maps are shown. There are relatively few popular routes for fishing vessels. However, it is observed that there are radial track patterns between port areas and the open sea (Figure 12(a)) because the operating area of each fishing vessel and each time may be different. In contrast, the lines of passenger ships are fixed (Figure 12(b)), so most tracks are distributed in their prescribed routes. Figure 12(c) is the density map for cargo ships. Those tracks are distributed widely, and it is easy to determine popular routes considering economy and security.

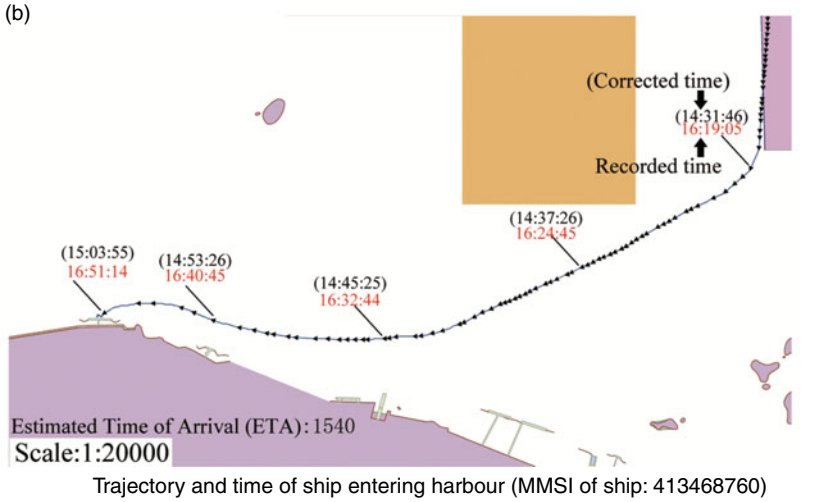
4.2. *Results of time.* Our statistical results from pre-processing time data shows that there are 146 track segments whose mean time deviations are greater than 1 hour. Correction effects on two tracks of ships entering harbour on 25 January 2015 are illustrated in Figure 13.

The Estimated Time of Arrival (ETA) of those two ships are 16:30 and 15:40 (25 January), respectively (see Figure 13). However, recorded time of track points near the destination are already 18:38 and 16:51 because of the delay in receiving AIS messages. Based on the method in Section 3.3, we determined the corrected value ($-7,022$ seconds and $-6,439$ seconds). The corrected time is the sum of corrected value and recorded time,

https://doi.org/10.1017/S0373463318000188 Published online by Cambridge University Press




Trajectory and time of ship entering harbour (MMSI of ship: 412378840)




Trajectory and time of ship entering harbour (MMSI of ship: 413468760)

Figure 13. Pre-processing effect in terms of time accuracy.

which is shown in brackets (Figure 13). Obviously, corrected time of arrival is closer to ETA, so the corrected time is more precise.

5. CONCLUSION. As discussed in this paper, AIS data mining plays a key role in characterising the patterns of traffic on water, planning of aids to navigation, risk assessment in maritime supervision and management of sea areas. However, there are unavoidable errors in the dataset because of improper use, signal transmission problems and the AIS system itself. High-quality AIS track information is essential for acquiring valuable knowledge. For example, in the application of maritime accident investigation and obtaining digital evidence, the time of AIS track points is directly related to the encounter of vessels.

https://doi.org/10.1017/S0373463318000188 Published online by Cambridge University Press

In this paper, the errors in AIS track data sets are analysed systematically and summarised by quality dimensions including physical integrity, spatial logical integrity, and accuracy of time. An algorithm with general applicability is proposed for improving the quality of spatial data, and a method based on time deviation is proposed for improving the accuracy of time which has rarely been reported in previous works. Distribution maps of ship tracks and density maps of different types of ships are shown based on one month of AIS data in the Zhoushan Islands area and the methods are validated by comparison.

The method cannot improve the quality of static and voyage related messages because it relies heavily on AIS users. It is observed that there are records of receiving time that are earlier than generated time, which is not researched in this paper. Future study will focus on more detailed application for further verification, and other pre-processing methods such as trajectory simplification.

## FINANCIAL SUPPORT

This work was partly supported by "National Natural Science Foundation of China" (grant number: 51579025) and "Natural Science Foundation of Liaoning Province" (grant number: 201602084).

## REFERENCES

Aarsæther, K.G. and Moan, T. (2009). Estimating navigation patterns from AIS. *The Journal of Navigation*, **62**(4), 587–607.

Altan, Y.C. and Otay, E.N. (2017). Maritime Traffic Analysis of the Strait of Istanbul based on AIS data. *The Journal of Navigation*, **70**(6), 1367–1382.

Bailey, N. (2005). Training, technology and ais: looking beyond the box. *Proceedings of the Seafarers International Research Centre's 4th International Symposium Cardiff University*, Cardiff, 108–128.

Banyś, P., Noack, T. and Gewies, S. (2012). Assessment of AIS vessel position report under the aspect of data reliability. *Annual of Navigation*, **19**(1), 5–16.

Breithaupt, S.A., Copping, A., Tagestad, J. and Whiting, J. (2017). Maritime Route Delineation using AIS Data from the Atlantic Coast of the US. *The Journal of Navigation*, **70**(2), 379–394.

Brodie, M.L. (1980). Data quality in information systems. *Information & Management*, **3**(6), 245–258.

Chen, J., Lu, F. and Peng, G. (2015). A quantitative approach for delineating principal fairways of ship passages through a strait. *Ocean Engineering*, **103**(103), 188–197.

De Souza, E.N., Boerder, K., Matwin, S. and Worm, B. (2016). Improving fishing pattern detection from satellite AIS using data mining and machine learning. *Plos One,* **11**(7), e0158248.

Felski, A., Jaskolski, K. and Banyś, P. (2015). Comprehensive assessment of Automatic Identification System (AIS) data application to anti-collision manoeuvring. *Journal of Navigation*, **68**(4), 697–717.

Fiorini, M., Capata, A. and Bloisi, D.D. (2016). AIS Data Visualization for Maritime Spatial Planning (MSP). *International Journal of e-Navigation and Maritime Economy*, **5**, 45–60.

Greidanus, H., Alvarez, M., Eriksen, T. and Gammieri, V. (2016). Completeness and Accuracy of a Wide-Area Maritime Situational Picture based on Automatic Ship Reporting Systems. *The Journal of Navigation*, **69**(1), 156–168.

Harati-Mokhtari, A., Wall, A., Brooks, P. and Wang, J. (2007). Automatic Identification System (AIS): data reliability and human error implications. *The Journal of Navigation*, **60**(3), 373–389.

International Maritime Organization (IMO). (2003). International Convention for the Safety of Life at Sea (SOLAS).

International Telecommunications Union (ITU). (2010). Technical characteristics for an automatic identification system using time-division multiple access in the VHF maritime mobile band, *Recommendation ITU-R M. 1371–4*.

Iperen, W.H. (2015). Classifying ship encounters to monitor traffic safety on the North Sea from AIS data. *TransNav: International Journal on Marine Navigation and Safety of Sea Transportation*, **9**(1), 51–58.

Iphar, C., Napoli, A. and Ray, C. (2015). Detection of false AIS messages for the improvement of maritime situational awareness. *Oceans'2015*. Oct 2015, Washington, DC, United States, 1–7.

Jaskólski, K. (2017). Two-dimensional coordinate estimation for missing Automatic Identification System (AIS) signals based on the discrete Kalman filter algorithm and Universal Transverse Mercator (UTM) projection. *52 Scientific Journals of the Maritime University of Szczecin*, **52**, 82–89.

Marine Management Organization (MMO). (2013). *Spatial trends in shipping activity*. Marine Management Organization.

Mazaheri, A., Montewka, J., Kotilainen, P., Sormunen, O.V.E. and Kujala, P. (2015). Assessing grounding frequency using ship traffic and waterway complexity. *The Journal of Navigation*, **68**(1), 89–106.

Mazzarella, F., Vespe, M., Damalas, D. and Osio, G. (2014). Discovering vessel activities at sea using AIS data: Mapping of fishing footprints. *International Conference on Information Fusion*, Salamanca, Spain, 1–7.

Mou, J.M., Van Der Tak, C. and Ligteringen, H. (2010). Study on collision avoidance in busy waterways by using AIS data. *Ocean Engineering*, **37**(5), 483–490.

Pallotta, G., Vespe, M. and Bryan, K. (2013). Vessel pattern knowledge discovery from AIS data: a framework for anomaly detection and route prediction. *Entropy*, **15**(6)**,** 2218–2245.

Peters, D.J. and Hammond, T.R. (2011). Interpolation between AIS reports: probabilistic inferences over vessel path space. *The Journal of Navigation*, **64**(4), 595–607.

Ristic, B., Scala, B.L., Morelande, M. and Gordon, N. (2008). Statistical analysis of motion patterns in AIS Data: Anomaly detection and motion prediction. *IEEE International Conference on Information Fusion*, 1–7.

Sang, L.Z., Wall, A., Mao, Z., Yan, X.P. and Wang, J. (2015). A novel method for restoring the trajectory of the inland waterway ship by using AIS data. *Ocean Engineering*, **110**, 183–194.

Shelmerdine, R.L. (2015). Teasing out the detail: how our understanding of marine AIS data can better inform industries, developments, and planning. *Marine Policy*, **54**, 17–25.

Silveira, P.A.M., Teixeira, A.P. and Soares, C.G. (2013). Use of AIS data to characterise marine traffic patterns and ship collision risk off the coast of Portugal. *The Journal of Navigation*, **66**(6), 879–898.

Tsou, M.C. (2010). Discovering knowledge from AIS database for application in VTS. *The Journal of Navigation*, **63**(3), 449–469.

Vettor, R. and Soares, C.G. (2015). Detection and analysis of the main routes of voluntary observing ships in the North Atlantic. *The Journal of Navigation*, **68**(2), 397–410.

Wang, J., Zhu, C., Zhou, Y. and Zhang, W. (2017). Vessel Spatio-temporal Knowledge Discovery with AIS Trajectories Using Co-clustering. *The Journal of Navigation*, **70**(6), 1383–1400.

Wawruch, R. (2017). Ability to test shipboard automatic identification system instability and inaccuracy on simulation devices. *Scientific Journals of the Maritime University of Szczecin*, **52**, 128–134.

Wu L., Xu, Y., Wang, Q., Wang, F. and Xu, Z. (2016). Mapping global shipping density from AIS data. *Journal of Navigation*, **70**(1), 67–81.

Zhang, W., Goerlandt, F., Kujala, P. and Wang, Y. (2016). An advanced method for detecting possible near miss ship collisions from AIS data. *Ocean Engineering*, **124**(1), 141–156.

Zhen, R., Jin, Y., Hu, Q., Shao, Z. and Nikitakos, N. (2017). Maritime Anomaly Detection within Coastal Waters Based on Vessel Trajectory Clustering and Naïve Bayes Classifier. *The Journal of Navigation*, **70**(3), 648–670.

Zhou, M., Chen, J., Ge, Q. and Huang, X. (2013). AIS data based identification of systematic collision risk for maritime intelligent transport system. *IEEE International Conference on Communications*, 6158–6162.