AIS Data Quality Assessment

Authors:
Prof dr. Albert Veenstra
Ir Rogier Harmelink

Rotterdam, December 2025

MEL Centre for Maritime Economics and Logistics; Working Paper 25/001

MEL Center for Maritime Economics & Logistics

Abstract

In this working paper, we present, for the first time, a quantitative measurement approach for the quality of AIS data. AIS data is ubiquitous, and many scientific contributions observe problems with the quality of that data. They often offer solutions. However, hardly any work exists that quantifies the extent of the quality problem. The main sources for AIS data, as well as some companies that develop information products based on AIS data, provide data that is cleaned, processed, and, more often than not, parsed into specific time intervals. This 'cleaned' data does not allow us to verify some of the important quality dimensions of AIS data. For our analysis, we have collected and recoded our own data to get access to completely unprocessed, truly real-time, data.

Our data quality methodology generates a number of relevant insights. First of all, we find that the quality problems related to the AIS information that comes directly from the ship's systems are relatively limited. At the same time, we observe severe quality problems in the manoeuvring information in the AIS data. We also analysed the ships' adherence to the reporting frequency requirements. We find that AIS messages are sent out more or less according to the AIS guidelines, but we could not confirm the absolute fulfilment of the requirement to send messages more frequently due to a course. We also find that manually entered data, such as destination, do not adhere to any prescribed standard.

Finally, we identify the use of default values in the AIS system as one of the main sources for data quality disturbances. These default values prevent empty data fields from occurring, but they do result, to a relatively high degree, in faulty or unusable entries in the data.

Our work is relevant for the large volume of AIS data studies. As long as these studies use the locational elements in the AIS messages, the data, and therefore the results of these studies is relatively reliable. As soon as other elements of the AIS data, such as speed, manoeuvring and rate of turn are used, more caution is advisable.

Table of Contents

1. Introduction

In recent years, studies involving tracking ships with data has taken enormous strides. The requirement for ships in international voyages above 300 gross tonnage (GT) to have an automated identification system (AIS) present aboard ships has been formally discussed in the International Maritime Organization (IMO) since 2000 and was adopted as part of Chapter V in the Safety of Life at Sea (SOLAS) convention in 2004. The performance standards for AIS equipment, such as the frequency of messages and the required content of static and dynamic information, have been around since 1998. Much of the required technology was developed in the 1990s with the express aim to avoid collisions and improve navigational safety (Yang et al., 2019). The data from this analogue technology is now ubiquitous, and is used extensively, by among others, maritime researchers.

The infrastructure to collect the AIS messages from ships has developed along two lines. AIS transponders, piggybacking on the relatively short-range VHF radio wave communication, now has a range of up to 40 nautical miles. In addition, collective data sharing arrangements have developed, where everyone who collects data with an antenna can contribute to a data pool and obtain access to all other data. An example is AISHub (aishub.net), associated with the Vesselfinder platform (www.vesselfinder.com). This AIS antenna network comes with a coverage restriction because there are (coastal) areas of the world with very few or no antennae. The high seas are not covered at all.

The second source of AIS data emerged from 2008 onwards, when an AIS satellite network was developed, and AIS data was also collected through satellites. This network aims to solve the coverage problem discussed above. An example is the German company Fleetmon (www.fleetmon.com, currently integrating with MarineTraffic and Keplr), established in 2010 and explicitly integrated satellite and ground station data. The extended coverage is illustrated in a map in Figure 1 below.

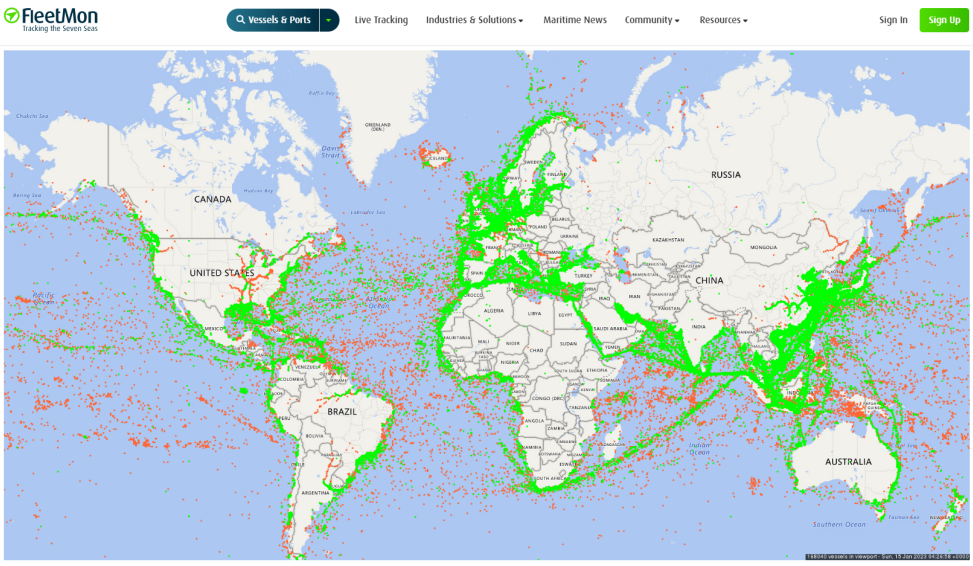

Figure 1: AIS coverage map based on Fleetmon
Source: fleetmon.com/global-vessel.coverage

Due to this data ubiquity, it is unsurprising that the body of research based on AIS data has also grown fast. At the same time, practitioners, researchers and AIS data specialists have observed significant problems with the quality of AIS data. We also know that data volume and bandwidth challenges are impacting the real-time nature of the data, and significant volumes of the data stream contain no data or zeroes (see for early analyses on this Harati-Mokhtari et al., 2007 and Qu et al, 2011 and, more recently, McFadden et al., 2019).

What is relatively unknown, however, is what the magnitude of the quality problem for AIS data is. This paper aims to contribute two results to the current AIS body of literature. First, we aim to suggest a comprehensive measurement approach for AIS data quality. This will offer researchers and practitioners a more objective way to evaluate the AIS data they are working with. Second, we aim to provide actual data quality measurements for a specific set of AIS data we collected ourselves. This should contribute to the understanding that AIS data users have to record and justify their data clean-up efforts, as well as that of practitioners who have developed real-world applications that rely on AIS data.

The remainder of this paper is organised as follows. We first briefly review the literature that addresses AIS data quality. Following this, we start by developing our data quality measurement methodology. As part of that methodology, we also provide a brief background on the technical nature of AIS data. After that, we present measurements on several data quality dimensions based on data specifically collected for this paper. We finish this paper with recommendations for users of AIS data and further research.

2. Literature on AIS data quality assessment

We reviewed papers that explicitly address elements of AIS data quality in the last decade (since 2015). In our search, we used the key word AIS, in combination with words that can be associated with data quality: just data quality, quality problems, but also, vulnerability, perils, (abnormal) data repair, false, errors, gaps, data integrity, data restoration, data reconstruction, data denoising, data pre-processing. We included one paper on anomaly detection (Wolsing et al 2021), because it is a review paper. Most of the time, these key words appeared in the title of the paper. In a few cases, however, the abstract would also quickly reveal specific attention to one or more AIS data quality aspects.

Very early discussions on AIS data quality can be found in Harati-Mokhtari et al (2007). This paper stands at the time of introduction of AIS in the maritime world, and evaluates the balance between AIS' contribution to solving safety problems and AIS causing problems of its own. This paper does provide measurements of errors, for instance for navigation status: about 30% of the vessels displayed wrong information, based on their data collection. Other early sources, such as Shelmerdine (2015) and Iphar et al (2015) elaborately classify quality problems but provide no analysis or measurement.

We identified some 24 papers that deal in some detail with data quality issues of AIS data, either by identifying sources for quality problems, or by measuring the magnitude of these problems. In our search, we focused on papers published in journals only. Some IEEE conference papers, book chapters and student theses that address AIS data quality are thus not considered in our overview below.

Several of the papers provide a (usually brief) discussion on types of errors in AIS data (He et al. 2021a, He et al. 2021b, Lei et al. (2021), Yang et al. (2021), Lee et al. (2019), Chen et al. (2020, 2022), Zhang et al. (2022), Meyers et al. (2022), Mieczyńska & Czarnowski (2021), El Mekkaoui et al (2022): invalid data, errors, values missing, abnormal values, duplicate records, locational outliers. Several papers focus on a specific data error: Lei et al. (2021) and Mieczyńska & Czarnowski (2021) focus on the MMSI nr, Yang et al. (2021) and Huang et al (2025) on destination data, Meyers et al. (2022) on static ship information (length, beam, draught and type), Mekkaoui et al. (2022) on spatial outliers, Zhao a et al. (2018) on accuracy of vessel tracks, and, finally, Lei et al (2021) as well as Serra-Sogas et al (2021) focus on coverage problems related to inland shipping and small recreational crafts, respectively. In some cases, the data quality problems are attributed to the technological infrastructure and equipment: He et al. (2021b), Mieczynska & Czarnowski (2021), Androjna et al. (2021), or weather (Liang et al, 2024).

As a solution, comparing the AIS data to some other data source is suggested: either visual inspection, radar, GPS, or aerial survey (He et al. (2021/1), Jaskolski et al. (2021), Serra-Sogas (2021)).

The majority of papers discuss AIS data quality in order to set a context for their proposed solution: automated data cleaning, clustering, consistency verification, trajectory

reconstruction, signal fusing and splitting, and interpolation. Emmens et al (2017) as well as Wolsing et al (2022) provide an overview of AIS data problems without offering new solutions of their own. Strozyna et al (2020) present a generic data quality measurement approach, which they then apply to various open AIS sources. Their analysis shows considerable quality differences between sources, ranging from Marinetraffic to Fleetmon.

A few papers provide an approach to actual measure data quality dimensions. Iphar et al (2015) provides a method, but no actual measures. Emmens et al (2017) also provide measures, on missing data and noise. Jaskolski et al (2021) provides some measures on position inaccuracy, through experimental simulations. Kiersztyn et al (2024) provides some measures on missing values for ship identifiers (MMSI, name, callsign, and IMO nr). They find a combined score of 54% missing data. Meyers et al (2022) , assessing the static ship related elements in AIS messages (length, beam and so on), find significant missing data, but also observe that this ratio is declining over time.

Our literature review reveals a broad consensus that AIS data has data quality problems. However, there is no consistent insight into the quantitative aspects of data quality measures across the AIS data source. Even thought some papers make some effort to provide measures for the extent of the data quality problem they aim to address, a consistent overview is lacking. We do not know *how often* the MMSI number is missing or incorrect or how often location data is irregular. We do know that almost all elements of the AIS data messages can be wrong (Mekkaoui et al., 2022).

3. Data quality model

There is extensive literature on data quality assessment, data quality management and data quality management methodologies. We do not intend to contribute to this ongoing line of research, and therefore, we apply a standard data quality assessment approach to maritime data. We follow the line of thinking of Zhang et al. (2019), who describe a data quality assessment in four steps:

a) Data analysis: insight into the basic structure and content of the data
b) Data quality requirements analysis: user expectations vis-à-vis data quality
c) Identification of critical areas: identification of sources of the data to be analysed
d) Process modelling: the approach to producing the data to be analysed
e) Measurement of quality: the selection of the quality dimensions and corresponding metrics

Zhang et al. (ibid) distinguish between objective and subjective quality measurement. In this paper, we focus on objective (i.e. based on quantitative metrics) quality measurement. We will expand on these five steps below.

a. Data analysis

Much material is available for a technical introduction to AIS data. We refer the reader to the basic regulation in SOLAS chapter V regulation 19 and the related IMO resolutions MSC.74(69) dd 12 May 1998 and A 29/Res.1106 dd 14 December 2015, which specify the technical details and communication frequency requirements.

An AIS system can generate 27 message types. In this paper, we will focus exclusively on the 'standard' vessel position report (messages type 1 and 3 combined with message type 5) and the data quality of these messages. We also focus on the commercial AIS technology and requirements (the so-called AIS Class A).

A generic vessel position report contains four main data clusters (IMO, 2015):

1. Static data: the ship's maritime mobile service identity (MMSI) and other vessel details;
2. Dynamic data: ship's position (GPS coordinated), position time stamp, course over ground, speed over ground, heading, navigational status and rate of turn;
3. Voyage-related data: draught, hazardous cargo type, destination and estimated time of arrival (ETA), route plan (waypoints)
4. Safety-related data: free text messages that can be sent to all receivers in range or a specific addressee.

Such a report is an amalgamation of the main content of messages 1/3 and 5. Messages 1 and 3 are location messages. Message 3 is semantically the same as message 1 but is a response to an interrogation. These messages contain location, course and speed. Message 5 is the voyage message, which contains information such as destination, ETA and ship particulars such as IMO number. The key to link the messages together is the MMSI number.

Messages 1 and 3 require broadcasting depending on the navigational status and speed. Anchored or moored ships do not have to send AIS messages as frequently as fast-moving ships. The specific message broadcasting requirements are reproduced in Table 1.

Table 1: Class A AIS equipment reporting frequencies

Navigational situation	**Reporting frequency**
Ship at anchor or moored and not moving faster than 3 knots	3 min
Ship at anchor or moored and moving faster than 3 knots	10 s
Ship 0-14 knots	10 s
Ship 0-14 knots and changing course	3 1/3 s
Ship 14-23 knots	6 s
Ship 14-23 knots and changing course	2 s
Ship >23 knots	2 s
Ship >23 knots and changing course	2 s

Source: IMO (2015); *'min' stands for minutes, and 's' for seconds.*

Message 5, as well as other safety and voyage-related messages are communicated every 6 minutes or as requested. As a result, there are (many) more messages 1 and 3 than 5.

Observe that the data source for the first three data clusters differs. The *static data* is fixed upon installation of the AIS equipment on board. However, a ship can exchange AIS equipment, giving the ship another MMSI number. The *dynamic data* will most often come from the navigational system on the ship. This is essentially the output from a sensor system where data is recorded automatically. The *voyage-related data* in message 5 is more often than not entered by the ship's crew manually since this data is not automatically recorded and may require some estimation or calculation (for instance the ETA). The regulations require this data to be updated every six minutes or amended as required (IMO, 2015, p. 6).

b. Data quality requirements analysis

To understand data requirements for AIS data, we have performed a literature review on papers using AIS data to assess data quality requirements. We carry out this review in two steps. First, we have reviewed all existing journal papers for a year (2021) to obtain an overview of applications using AIS data. Second, we provide a more detailed discussion of the papers that consider AIS data quality specifically.

As a first step, we identified all papers on AIS data in 2021. All in all, by using the combination of keywords 'AIS' and 'maritime' and/or 'shipping', we found 116 individual papers. The reference list for these 116 papers is available separately. This list contains only journal papers, and no book chapters or publicly available conference presentations. In our analysis, we are interested in the distribution of topics for these papers and how they deal with the quality of AIS data.

We applied an inductive thematic coding approach (see, for instance, Rivas 2012) in two steps to sort the papers into application categories. Our first-level topics were:

- AIS: papers on the usefulness of AIS data in maritime research,

- Biology: applications of AIS data to identify problems in marine biology, such as fishery monitoring,
- Bunkering: using AIS data to calculate bunkering statistics,
- Connectivity: network connectivity studies using AIS data,
- Cyber problems: security of the AIS data infrastructure,
- Data analytics: deriving ship classification from AIS satellite observations,
- Data integration/compression: working with large (AIS) data streams or sets,
- Engineering: signal conversion or integration
- Maritime operations: using AIS data to calculate ETAs, destination predictions and so on,
- Maritime Communication: performance analysis in maritime communication,
- Navigation: risk analysis and safety improvements,
- Oceanography: determining traffic densities or currents in the sea,
- Traffic analysis: collision analysis, congestion analysis, route analysis and environmental impact,

Of the 116 papers, the four most common themes are traffic analysis (62 papers), discussion on AIS data (13), data integration/compression (8) and maritime operations (8).

A further breakdown of the most common subthemes is provided in Table 2.

Table 2: Subtheme breakdown

Traffic analysis: 62	**AIS: 13**
Trajectory prediction: 24	Data quality: 8
Collision analysis: 10	Literature reviews: 3
Environmental impact: 4	A further 2 individual topics
Covid: 3	
Vessel behaviour patterns: 3	
Congestion: 2	
A further 16 individual topics	

The second-level label in Table 2 (lefthand-side), 'Trajectory prediction', includes the sub-labels: trajectory reconstruction, route prediction, and destination prediction.

From the combined the literature analysis in this step, we identify the main requirement for the data quality of AIS data: the data should be suitable for performing some form of trajectory reconstruction analysis. This means that the basic combination of ship identification, location data, and navigation information (speed, heading, course, destination, manoeuvring) should be trustworthy.

c. Identification of the data source

Generally, AIS data can be obtained from an AIS data hub. However, given that reporting AIS data does require decoding, as described above, it is important to look for a source of data

that is as untampered as possible. We found such a source by collecting our AIS data with a dedicated antenna.

We collected our data by plugging directly into the Dutch AIS infrastructure of the National Digital Infrastructure Authority (in Dutch: RDI, Rijksdienst voor Digitale Infrastructuur). We obtained data from their receiver at Hook of Holland (just north of the Port of Rotterdam) on three consecutive days in May/June 2023. Under normal circumstances, this receiver has a reception range of about 70 kilometres. We used open-source software from Arundale (www.arundale.com)[1] to capture the AIS messages 1, 3 and 5 payloads into a CSV file. Arundale appends the two parts of message 5 (message 5 requires two message spaces in the VHF channel) into one line in our CSV file, but then simply stores the undecoded message payloads. We decoded the message payload ourselves using the standard AIS documentation.

We recognise that we conduct our analysis on a sample of all available AIS data. This is a sample is in time (about 48 hours) and in space (a circle of about 70 kilometres around the Port of Rotterdam). On the other hand, this is a busy region, with a broad range of ships and maritime activity. We consider it therefore a rich sample. A heatmap of our observations is provided in Figure 2.

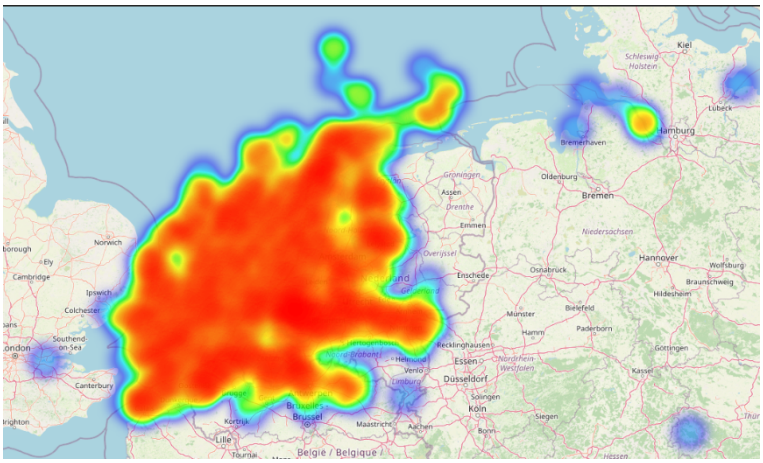

Figure 2: Heat map of AIS message observations 31 May – 2 June 2023

Collecting data from the receiver gave us some 4 million (mln) messages over 50 hours. Since the period at the end of May/beginning of June had very fair weather, we captured many messages in a range wider than the usual 70-kilometre range of the antenna. Figure 2 shows the heatmap of our messages around Hook of Holland. Most of the messages are received in a rough circle around Rotterdam. We also observe quite a lot of river traffic. There are strange observations as well. For instance, zeroes in the location information will result in observations that seem to be in London (Greenwich, in fact). Observations in Hamburg or Bremen make sense, although they should have been out of range. Observations in non-water areas in the middle of Germany, do not make sense. This visual inspection already confirms that the location information in AIS data cannot always be trusted

[1] https://arundaleais.github.io/docs/ais/ais_decoder_v3_downloads.html

d. Process modelling

From a technical perspective, the data transferred in the AIS messages is a so-called AIVDM / AIVDO data packet. This data packet is part of the National Marine Electronics Association (NMEA) standard. This packet is a 'sentence' that contains the main payload of a fixed bit length and some message particulars before and after. Upon capturing the data with an antenna, additional data, such as a timestamp, is added. A payload could look like this:

13aNa@0P02PCTNRMdM;:W?wp2p22

Such a string of characters is called an ASCII-encoded bit vector. This encoding means that each character, even the punctuation marks, represents six bits. The payload needs to be decoded in a particular way to obtain the full data string. For this, see any manual on dealing with AIVDM/AIVDO protocols. The resulting six-bit strings then represent numbers. For each fixed-length field in the message structure, these numbers either have direct meaning (for instance, the MMSI number) or can be associated with specific information for that field via another table (for instance, for navigational status). Most AIS data pools will handle this decoding process before making the AIS data available.

Since this decoding process already reveals data quality problems, for instance, when the payload does not allow complete decoding of the MMSI number, it is unclear how the various data collection pools for AIS handle these problems. If they simply delete all unsuccessful decoding cases, quantitative quality measurement may be underestimated.

e. Measurement of quality

The data quality measurement literature often points to Batini et al. (2009), who proposed a comprehensive data quality (CDQ) methodology (see, for instance, Zhang et al., 2019 and Krasikov & Legner, 2023). Batini et al. (ibid) also discuss the consensus concerning data quality dimensions. These are:

1. Accuracy; this refers to the correctness or correspondence between recorded and real-world values.
2. Completeness; this refers to the degree to which all required data elements are present.
3. Consistency; this refers to (the absence of) violation of logic, semantics or other rules in the data.
4. Timeliness; this relates to time-related dimensions of the data: currency describes how long ago the data was collected, 'volatility' describes the period the data is valid in the real world, and timeliness itself describes the duration between a real event and its recording with data.

These four dimensions appear in many earlier works, for instance, Fox et al. (1994). We find, however, that these four dimensions do not capture the full extent of data quality for AIS data. We will resort to some other sources for additional dimensions.

One useful source from the AIS data domain is Iphar et al (2015), who mention 7 quality dimensions: accuracy, completeness, consistency, currentness (i.e. timeliness), precision, reliability, and integrity. Professional data management companies provide additional views on data quality management. In a whitepaper from data management service provider Simplity (describing their Accurity data quality management platform[2]), the dimensions 'uniqueness' and 'validity' are added. The former represents the presence or absence of duplicates in the data, and the latter describes the adherence to formats and data types. The dimension validity also captures the last three dimensions in Iphar et al (2015) - precision, reliability, and integrity. 5We will therefore work in the remainder with the following six quality dimensions: (1) accuracy, (2) completeness, (3) consistency, (4) timeliness, (5) uniqueness and (6) validity. We thus add uniqueness to the dimensions proposed by Iphar et al (2015), and capture three of their dimensions under the umbrella of 'validity'. These six data quality dimensions are also found in other professional sources, such as Gawande (2022) from iCEDQ[3]. He proposes the ratio of inaccurate to total number of records as a generic measure.

Interestingly, the scientific literature does not discuss the problem of duplicates. Batini (2009) only states that it affects all four quality dimensions. In our particular research problem, duplicates can be a significant problem since many users are drawing data from a pool filled with different ground stations, some of which are close together. These will, therefore, capture largely the same set of ships. The solution could be as simple as eliminating all exact duplicates, but this raises the question: Are there 'near duplicates' that should also be identified and eliminated? We, therefore, consider duplicates, or more properly 'uniqueness' as an additional dimension of our data quality model.

Validity is a more practical. To reduce the problem of measurement and standardisation, many professional data collectors, such as national statistics organisations, use codes and standardised data values. The standardisation of ship types in shipping, or the Harmonised System for goods classification in trade, are examples in the maritime and trade domains. This problem is relevant for our particular case since our data contains a subcluster of data that is entered manually. This data includes the data field 'Destination' in message 5. IMO (2015) contains only limited directions on what standard to use (in principle UN LOCODE), and thus, variation in this field seems unavoidable.

In addition, we include the so-called default values under this quality dimension, 'validity'. These are values that are automatically imputed (by the AIS equipment) if there is no data to report. Default values are usually 'obviously wrong' values. For speed, for instance, the default value in the AIS systems is 102,3 Kts. In quite some cases, however, this default value is a number that is read as a value in the reception infrastructure. In applications such as ETA estimation, this will result in wrong outcomes.

[2] https://www.accurity.ai/whitepaper/how-to-establish-a-data-quality-management-framework/
[3] https://icedq.com/6-data-quality-dimensions

4. Objective AIS data quality measurement

Above, we have identified six data quality dimensions: accuracy, completeness, consistency, timeliness, uniqueness and validity. We can identify simple tests to assess the magnitude of the problem in each dimension. Table 3 contains our research plan, with comments on the feasibility of the assessment. We have ordered the table based on the feasibility of the tests.

Table 3: Research plan by data quality dimension

Data quality dimension	Quality test	Feasibility comment
Uniqueness	Identify the number of duplicates.	We will concentrate on exact duplicates.
Completeness	Identify complete messages based on the prescribed length of the payload.	This amounts to counting the length of the message payload.
Validity	Assess the degree to which manual data uses standards and assess the use of default values.	This amounts to counting the use of the specific default value occurrences, as well as – for the destination – cataloguing the most common destinations. We also present occurrence ratios.
Timeliness	Verify if the frequency requirements are met (cf Table 1).	Here, we look at the time interval between two consecutive messages, given the condition in Table 1 (speed).
Consistency	Assess a consistent sequence of recorded positions.	This requires a consistent business rule that includes the information on speed and the calculation of the distance between two locations.
Accuracy	Verify recorded position with actual position.	With our historical data, this is not possible.

The content of Table 3 results in several observations:

- We cannot investigate all dimensions of data quality equally. Accuracy is a dimension we have to exclude. There is some interesting literature on this already, however. See, for instance, Jankowski et al. (2021), who have compared AIS location data with radar observations and found considerable inconsistencies. Androjna et al (2021) study spoofing of AIS locations, which represents a deliberate attempt to distort the actual locations of a maritime object.
- Some quality assessments are simple counting exercises: uniqueness, completeness, validity and timeliness,
- Some quality assessments, such as consistency, require calculation based on the data and a business rule.

5. Data quality assessment

In this section, we present our data quality measurements. In Table 4 below, we first record our expectations for the findings.

Table 4: AIS data quality research steps

Data quality dimensions	Research activity	Comment	Expected outcome
Uniqueness	We will present an inventory of duplicates.	We will consider exact duplicates.	Since we collect data from a single antenna, we do not expect to find duplicates.
Completeness	We will look at incomplete messages by assessing the length of the message payload and the possibility of linking message 5 content to message 1/3 content.	The basic technical requirement is a payload of 168 bits.	We expect the length of the payloads to follow the basic AIS technical requirements.
Validity	We will present an analysis of default values in all possible data fields We will present a specific analysis of the variation of destination values in message 5.	Validity may vary with the type of data elements.	We expect to find some default values. However, the occurrence should be minimal: <0,1% of the data.
Timeliness	We will assess the ships' compliance level with the frequency requirements in the IMO (2015) resolution.	We need to correct for ships entering and leaving our reception range, where not all messages may have been captured; we employ a geofencing approach for this.	We expect all ships to comply with the frequency requirements of IMO.
Consistency	We will evaluate the logical consistency of subsequent AIS messages.	We restrict ourselves to the linking of speed, locations, and distance.	Given that this is a common application of AIS data and many solutions have been proposed, we expect to find considerable inconsistency.
Accuracy	We cannot verify the accuracy of the AIS data with outside data sources.	We will present an analysis combining some aspects of accuracy in our consistency analysis.	There is no research activity for this quality dimension.

Note that consistency is a way to infer something about accuracy. For consistency, we attempt to count the number of cases where our consistency test fails. While we do not know which data element was inaccurate in that case, these observations could also be recorded as an

identified inaccuracy. To fully assess accuracy, we need additional information to verify the correctness of the data. Since we do not have access to other observations of ships, for instance, by radar or through visual observation, we cannot carry out such a test on accuracy. At the same time, our consistency test could be seen as a combined test on consistency and accuracy.

As a final remark, the order in which we assess the different quality dimensions is relevant. Corrections we have to apply as a result of earlier quality checks (removing incomplete data, for instance), are included in later tests.

6. Quality measurement

The number of observations in our data source can be found in Table 5.

Table 5: Numbers of AIS messages

Date	Total number of messages	Message type 1	Message type 3	Message type 5
31 May 2023	823.127	640.702	153.693	28.732
1 June 2023	2.088.215	1.640.969	373.938	73.308
2 June 2023	1.137.269	901.515	194.985	40.769
Total	4.048.611	3.183.186	722.616	142.809

Observe that we have many more messages 1 and 3 than messages 5, as we have already mentioned above.

Uniqueness
For uniqueness, we look at the occurrence of duplicates. We do not expect to find any duplicates. Looking at the information in the payload, we can identify 34.868 (0,89%) exact duplicates for messages 1 and 3 and 139.757 (97,9%) for message 5.

We thus observe that even for our data collection with a single antenna, we find almost 1% duplicates. It is extra strange that there are duplicates since the AIS system contains a repeat indicator in the messages, which should prevent an exact duplicate of the message if it is sent out multiple times. However, the occurrence can still be considered to be relatively low.

The score for message 5 is a different story. Here, we find many duplicates. Since this message contains much relatively unchanging voyage information, this could still make sense: destination and ETA do not often change during a voyage. Only if the ship enters a port will the ETA be adjusted. At the same time, we are observing ships around the Port of Rotterdam, where we expect ships to at least make navigational adjustments because of the traffic they will encounter. So, a duplicate percentage of 98% is very high. We ended up with just 3.052 unique, or usable, messages of type 5 out of 142.809. Note that this has repercussions for any analysis that attempts to verify destination or ETA prediction in the data based on some algorithm. While the data for message 5 seems abundant, its real statistical informational content is limited if more than 95% of this data consists of duplicates.

Completeness
Under the completeness dimension, we look at missing information. First, we consider the payload. All messages have a prescribed length. For messages 1 and 3, this is 168 bits. For message 5, this is more complicated, and we will explain this in more detail below.

Reviewing the payload length for messages 1 and 3, one can infer that a message may contain an error if the standard length is not found. In our 3,9 mln messages, we find 27 incomplete payloads and 48 that contain characters that cannot be decoded. This amount of incomplete and incorrect payloads is minor (<0,001%). We have to remove these few observations from

our dataset since we cannot correctly decode these messages. All other payloads in messages 1 and 3 have the prescribed length of 168 bits, which aligns with expectations (Table 4).

For message 5, the length of the payload varies. It should be 424 bits, but 426 bits is the dominant length in our data set. Part of the reason is that this message contains manually entered data, among which is the ship's destination. This destination needs to be typed in by a navigation officer on the bridge, which may result in differences in the text string. The difference in payload length is foreseen in the IMO documentation, where most decoding manuals include statements such as 'Robust decoders should ignore trailing garbage and deal gracefully with a slightly truncated destination field' (Raymond, 2023). If we count our message 5 with shorter length, we identify 8 incomplete payloads on a total of 3.052 unique messages (0,002%). This we also consider to be minor.

Another way of looking at completeness is to observe our data content for all ships indexed by MMSI. Our data set has 4.182 unique MMSIs. Of these, 2.040 do not have a type 5 message: we only have message types 1 and 3 for these MMSIs. For these 2.040 ships, therefore, we do not have more information than the MMSI: no IMO number, no type information, no destination or ETA. Of course, additional information on ship type could be gathered through an outside source, such as ship register information. However, few academics have direct access to ship register data. There is a group of 602 MMSIs for which we have a single type 5 message and 1.540 MMSIs for which we have multiple type 5 messages.

We do not find many 'empty' fields in the data, due to the AIS system design: message payloads with a fixed length and empty fields are 'filled in' by the equipment, either with zeroes or default values. This is a feature of the robust design of the AIS system. In several cases, zero values have meaning: for heading, speed and course over ground. Also, default values may be numeric, even though they do not have 'meaning'. We look at the occurrence of the default values under the data quality dimension 'validity'.

Validity

For validity, we first look at the use of default values. The AIS system has many default values to make it robust for broadcasting at sea, as well as for adaptations and future development. In addition, the AIS system has reserved values for future use. The current most common values, for instance, for *navigation status* are: 0 (underway using engine) and 1 (at anchor), as well as 5 (moored), 7 (fishing) and 8 (sailing). Additional there is: 9 = 'reserved for future amendment of navigational status for ships carrying DG, HS, or MP, or IMO hazard or pollutant category C (HSC)', 10 = 'reserved for future amendment of navigational status for ships carrying DG, HS or MP, or IMO hazard or pollutant category A (WIG)'; 11-14 = 'reserved for regional or future use', 15 = ' undefined (default)'.

We report the number of messages that use additional and default values for navigation status in Table 6.

Table 6: Message count default values

Navigation status	Number of messages	Occurrence
9	23	0,00%
10	143	0,00%
11-14	2.072	0,05%
15	193.396	4,9%

The occurrence is based on the original size of our data set

While the observations for statuses 9, 10, and 11-14 are again low, the navigational status 15 (undefined) is used substantially. If only a single message is available, it is not clear what these ships are engaged in.

The AIS manual shows that speed over ground has a default value of 102,3. This number is not a feasible speed for any ship. It is a number, however, and if researchers are not paying attention, it will end up in average or maximum speed calculations. At the very least, this will make average speed calculations unreliable. We found this default value in 14.840, or 0,4%, of the messages.

In Table 7, we report a complete overview of the default values in the AIS messages 1/3 and 5.

Table 7: Overview of default values and message counts

Message field	**Message**	**Default value**	**Number of messages**	**Percentage of messages**
Rate of Turn	1&3	128	1.830.488	47,3%
Speed over Ground	1&3	102,3	14.840	0,4%
Longitude	1&3	181	14.638	0,4%
Latitude	1&3	91	14.635	0,4%
Course over Ground	1&3	3600	690.175	17,8%
True Heading	1&3	511	1.839.892	47,5%
Maneuver Indicator	1&3	0	2.608.288	67,4%
Navigational Status	1&3	various	195.634	5,0%
Ship Type	5	various	90	3,0%
ETA month	5	0 &15	578	19,0%
ETA day	5	0	577	19,0%
ETA hour	5	24 & 31	384	12,6%
ETA minute	5	60 & 63	383	12,6%

Percentages are based, for the messages 1/3 on 3.905.802 and for message 5 on 3.052 total number of messages.

Note that there are wildly different scores, from very low percentages to staggeringly high percentages. The highest percentage is obtained for the Maneuver Indicator. The value options here are: 0 (default), 1 (not engaged in special maneuver) and 2 (engaged in special maneuver). So, the transponder should send out value 1 most of the time. There is apparently

no genuine interest in this indicator showing the correct information. The actual data received is 'wrong' two-thirds of the time.

The variables of rate of turn, course over ground, and true heading are all navigational information fields in messages 1 and 3. One would expect these to be filled from the ship's systems. We interpret the small error percentages as a confirmation of this. Still, an estimate of 0,4% applied to some 100.000 ships in the world merchant fleet indicates that several hundred ships are consistently sending out wrong location and speed information.

Finally, note the considerable number of messages (seen against the much smaller volume of unique type 5 messages) in which the ETA time stamp contains default values. The inaccuracy for month and day information is higher than for hour and minute information. Given the importance of ETA information for other parties, such as port authorities and agents, using default values in ETA's is worrying. This use of default values confirms the general lack of quality in ETA data from AIS that we have observed elsewhere (Veenstra & Harmelink, 2021).

As part of our analysis of data validity, we look at the level of standardization of the destination field in type 5 messages. The destination is a free text field of 20 positions, so some variety is expected. The AIS guidelines say that UN/LOCODE and ERI terminal codes should be used[4] for the destination field. We know that manual data entry results in data quality problems (Counsell et al., 2007)

Based on 3.052 unique messages, we present the destination values with the highest frequency. Note that we collected messages more or less in a 75-150 km radius around Rotterdam. Therefore, a considerable reference to the Port of Rotterdam as a destination is expected. Note further that, the 20-character space is sometimes filled with a default character, decoded as '@'.

Note that the consequence of a free text field and manual entry is that the destination information in AIS systems is highly varied. The variation is apparent from Table 8: nobody apparently uses the ERI terminal coding standard, even though a specific port or terminal destination is added, in many cases in Dutch. The occurrences reported in Table 8 represent about one-third of the available message type 5 data (3.052 unique messages).

In addition, due to the way the AIS system works with default values, which may or may not be decoded, the destination may contain spaces, the '@'-character, or other text, which makes this data hard to work with. The location references are imprecise and may very well be inaccurate. We observed, for instance, 'NLRTM<>GBHRW' and 'ROTT HARW VV@@@@@' (or Rotterdam to and from Harwich) for the ferries in the Port of Rotterdam in our data. These examples are not a destination but a route. If these ships are sailing this leg daily, it is perhaps understandable that they enter the information this way, but it is not compliant with the IMO

[4] The ERI (Electronic Reporting International) is a maritime based coding system with a quite accurate reference system for terminal locations, which includes fairway indicators, terminal identifiers as well as hectometer references.

AIS guidelines. The consequence of this quality problem with destination data is a considerable amount of literature on destination prediction (see, e.g., Yang et al., 2021).

Table 8: Destination value varieties

Destination value	Count	Cumulative count	Comment
Rotterdam@..@	349	349	Different variants, with and without spaces
'@@@@@@@@@@@@-@@@@@@@@@@'	241	590	The @ character results if the field is empty or filled with zeroes
NLRTM@...@	169	759	This is the LOCODE for Rotterdam
NL RTM@...@	85	844	
BEANR	33	877	This is the LOCODE for Antwerp
ANTWERPEN@...@	22	899	
ANTWERP	18	917	
DORDRECHT@..@	18	935	
ROTTERDAM BOTLEK@@@@	16	951	This is a reference to the port + terminal (no locode was used, however)
DINTELHAVEN@..@	16	967	
VLAARDINGEN@..@	15	982	
ROTTERDAM PRINSES AM	15	997	
BE ANR	15	1.013	
SCHEVENINGEN	13	1.026	
AMSTERDAM	13	1.039	
ROTTERDAM 3E PETROHA	12	1.051	
ROTTERDAM 2E PETROHA	11	1.062	
EUROPOORT@..@	10	1.072	

'@..@' indicates that the data field was filled up to 20 characters with either empty spaces or zeroes.

Timeliness

The AIS regulation has requirements for the frequency of messages; see Table 1. This can be measures by looking at the time intervals between consecutive messages, given the trigger 'speed' in the first message. Since we have observations from a single antenna, we want to control for radio frequency distortion that may exist at the boundary of our reception area. Incidental messages may be lost due to poor reception, which will influence our calculations. To manage this, we drew a circular geofence around our antenna in Hook of Holland (see Figure 3). Within this circle, we minimise the influence of the distortion.

We need to look at multiple messages *from the same* ships. In our data set, we find 4.177 unique MMSIs. Of these, 262 and 283 ships have empty or a single message of type 1 and 3, respectively. For the other cases, we look at the time interval between all occurring combinations of two messages.

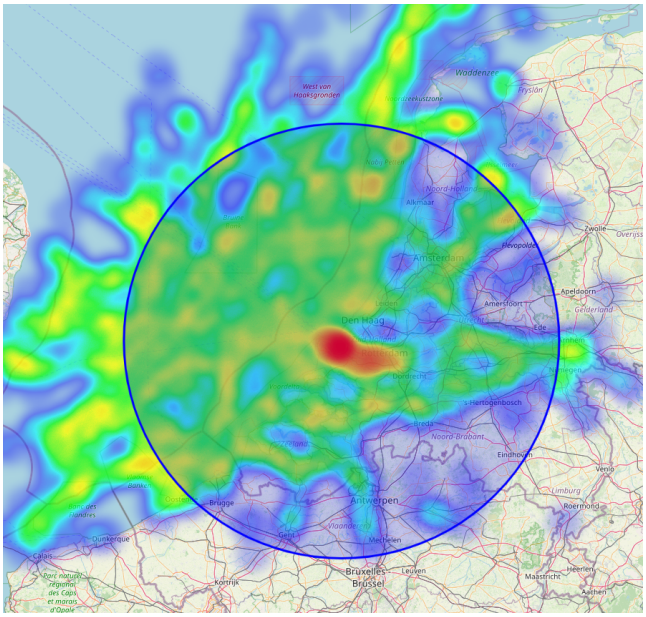

Figure 3: Observed messages and geofence circle

In Table 9, we provide various measures for the timeliness analysis across message multiples and ships. From previous research (Veenstra & Harmelink, 2021), we know there may be incidental but significant outliers that might impact *average* time interval calculations. Therefore, we also calculated median time intervals. Finally, we calculate a message frequency indicator not based on a summary statistic: the number of messages between 0 and 20 seconds.

Note from the message numbers in Table 9 (columns 3 and 4) that compliance with the time interval guidelines is rather poor. Observe specifically the cases of changing course at relatively low speeds (0-14 knots and 14-23 knots). These are typically speeds for huge ships (13-14 kts is the standard tankers and bulk carriers; 18 kts is a fairly standard speed for container vessels) and ships entering ports.

A few of the observed average frequencies are close to the prescribed values. This holds for the 3-minute interval for ships moored or anchored and for ships sailing faster than 23 knots. Based on the median time intervals, even more observed time intervals align with the IMO guidelines.

There are also deviations, however. For ships anchored, we observe a considerable difference between the average and the median value. The average here is an average of short and longer time intervals. In most cases, the ships communicate more frequently than required. This could result from a relatively large volume of type 3 messages, which result from active interrogation from other ships. This is to be expected, in an anchorage area. We also observe that the prescribed frequency for ships sailing between 0 and 14 knots has a low overall correctness percentage (38,9% and 7,2% when changing course).

Table 9: Timeliness results

Speed category	Interval expected	Count correct	Count incorrect	% correct	Average Speed (course change)	Average time (m:ss)	Median Time (m:ss)	Count messages in 0-20s
Anchor/moored < 3 knots	3 m	70.210	11.526	85,9%	0,2 k	3:43	0:19	53,8%
Anchor/moored > 3 knots	10 s	10.688	5.435	66,3%	7,4 k	1:25	0:09	79,0%
Ship 0-14 knots	10 s	627.165	986.329	38,9%	0,4 k	1:19	0:11	58,6%
Ship 0-14 knots changing course	3 1/3 s	140.396	1.822.009	7,2%	4,3 k (-0,140)	1:24	0:11	65,0%
Ship 14-23 knots	6 s	8.178	4.739	63,3%	16,2 k	0:12	0:06	86,9%
Ship 14-23 knots changing course	2 s	17.303	74.377	18,9%	17,5 k (4,136)	1:06	0:06	84,3%
Ship >23 knots	2 s	2.545	677	79,0%	27,0 k	0:02	0:02	91,7%
Ship >23 knots; changing course	2 s	25.603	26.535	49,1%	27,0 k (8,657)	0:16 s	0:03	92,9%

's' and 'ss' stand for second, 'm' stands for a minute, and 'k' stands for knots.

Another observation is that the increased message frequency for ships 'changing course' is not found in the data. Instead, ships changing course will send messages less frequently (based on the calculated average time intervals). We wonder if this is due to the poor data on the maneuver indicator in the type 1 and 3 messages, which seems to be unused. On the other hand, our indicator that looks at the number of messages within 20 seconds does find a slight increase due to the ship changing course for two of the three-speed categories. For speed between 14-23 knots, however, this indicator shows fewer messages when changing course. This is non-compliant with AIS standards.

Overall, many ships are broadly compliant with the frequency requirements, but there is also considerable non-compliance in message frequency. In addition, we draw attention to the frequency requirements for the ship's changing course. Here, we observe structural non-compliance with the formal frequency requirements.

Consistency and accuracy

We have already indicated that we cannot perform a full accuracy test on our data. We can, however, observe inaccuracy that results from inconsistent observations. This is a matter of degree: if we find relatively minor deviations in location, we take these to result from inconsistency in the recording of locations of speed. However, if these deviations become large, this is no longer an inconsistency but a data inaccuracy. In other words, if the ship makes big, unexplainable jumps, this is more of an accuracy problem than a consistency problem.

To perform the consistency check on our data, we need consecutive messages with valid speed and location data. We also take into account a timeliness threshold of 60 seconds: messages should not be more than 1 minute apart. For ships anchored, this means we are stricter than the regulations. In all other cases, we are more lenient. We use the well-known Haversine formula to calculate distances between two geolocations at sea.

We compare our calculated distance between two points with the sailing distance based on the first location and speed. We do this only for messages that are entirely within our geofence (see Figure 3). This results in 176.431 message sequences we can work with, where the sequences can be as short as 2 messages or as long as 26.483 messages. This latter case corresponds to a ship that is sailing relatively fast through our geofence area and sends out messages every 3 seconds or so.

As was mentioned earlier, we combine the data quality dimensions for consistency and accuracy in our test. We evaluate the results as follows:

- Differences between the two distances below 10 m: consistent and accurate,
- Differences between 10 and 35 m: data is not consistent, but still accurate,
- Differences above 35 m: data is inaccurate.

The 10-metre criterion is based on the generic inaccuracy of positioning equipment. The 35-metre criterion is the mean value of all our difference measures above 10 metres + 10%.

Our measures are reported in Table 10.

Table 10: consistency and accuracy results

Distance differences	Message count	Percentage (%)
Below 10 m	4.102.207	97,4
Between 10 m and 35 m	96.449	2,3
Above 35 m	11.665	0,3

'm' stands for metre

We observe some degree of inconsistency in the data. Nevertheless, at 2,3%, we do not consider this very significant. This result is good news for all the colleagues who work with the location data in AIS. Consecutive messages are largely consistent regarding the relationship between location, distance, and speed. This is in line with our results reported under validity.

We have also looked at the consistent performance of individual ships. We find that a median value of 2 inconsistencies is found per ship. There are also strange occurrences where this triplet of data (location, speed and distance) is far off the mark. The relative importance of this problem is small, but that does not mean it does not occur. In our 4 mln messages, we have close to 12.000 cases in which the data shows unexpected outcomes. The reasons could be various: the speed data is incorrect, the location information is incorrect, or both. And, since we found the median value of inaccuracies per ship equals 2, every ship occasionally has this type of disturbance in its data.

7. Concluding remarks

In this paper, we have presented an approach to *quantify* data quality problems for AIS data. We have collected our own AIS data to present data quality measures on the rawest, purest, possible data set. While this gives us much control over the data wrangling, we acknowledge that our data is only a small and regionally restricted sample. We present a structured data quality measurement methodology to obtain quantitative insights into AIS data quality. We have used a data quality measurement framework with six dimensions: uniqueness, completeness, validity, timeliness, consistency and accuracy

We will confront our initial expectations with the outcomes we have presented in Table 4.

Table 11: AIS data quality assessment outcome

Data quality dimensions	**Expected outcome**	**Research results**
Uniqueness	Since we collect data from a single antenna, we do not expect to find duplicates.	We find < 1% duplicates in messages 1 and 3, and 97,9% duplicates in message 5. The latter message contains a lot of static information, such as destination and ETA.
Completeness	We expect the length of the payloads to follow the basic AIS technical requirements.	We find very minor numbers of incomplete messages of types 1 and 3, as well as 5: < 0,001%. We find that we have 4,182 unique MMSIs, of which 2.040 cases without type 5 messages (48,7%), 602 cases with only 1 type 5 message (14,4%) and 1.540 cases with multiple type 5 messages (36,8%).
Validity	We expect to find some default values. However, the occurrence should be minimal: <0,1% of the data.	We find that (illogical or impossible) default values occur very frequently. For full results, see Table 7. Highest percentages are found for Rate of Turn (default 128 in 47,3% of cases), True Heading (default 511; 47,5%) and Navigation Status (default 0; 67,4%). In message 5, ETA information may contain default values in up to 19% of the data. We also observe significant non-standard entries in manual field such as destination. Here non-compliance with prescribed standards is almost 100%.
Timeliness	We expect all ships to comply with the frequency requirements of IMO.	Here we find mixed results: Ships anchored and sailing certain speeds send out messages according to the AIS standard. However, the prescribed increase in reporting is not observed for any of the three speed categories. (In fact, in one of them, the frequency decreases).
Consistency	Given that this is a common application of AIS data and many solutions have been proposed, we expect to find considerable inconsistency.	According to our thresholds, we find considerable consistency in the combination of locations and speed: 97,3%. We do find that every ship in our data set has one or more consistency violations, however.
Accuracy	There is no research activity for this quality dimension.	We did not present results for this criterion.

In summary, for the dimensions of **uniqueness** and **completeness**, the problems are relatively small. We attribute this to the AIS system design, which prevents missing data but introduces a considerable amount of default values and our data collection approach through a single antenna (albeit a very large one).

This has repercussions for our third measure, **validity**, however. Here, we find that default values in the data are a persistent problem that varies considerably with the specific data element. Most dynamic data originating from ship's systems have relatively small default data problems (smaller than 1%). However, these could still potentially significantly impact research results (e.g., calculated average speeds). Manoeuvring information, such as the special manoeuvre indicator and rate of turn, exhibit considerable default data problems. Here, we observe 67% and 47% default values in our data, which makes the data uninformative on these data elements. We also observe the use of default data in reporting ETA information in at least 12-19% of the date structure of ETAs. Finally, the free text value option in the destination field results in a wide variety of input and little adherence to standards.

Under **timeliness**, we observe that most ships try to adhere to the message frequency standards. On the other hand, our analysis shows that ships are largely non-compliant with the frequency requirements for ships changing course. We conjecture that this may be linked to the poor quality of other manoeuvring related data in the AIS messages, such as the Rate of Turn and the Manoeuvre Indicator.

For our fifth dimension, **consistency**, we find that the data is largely consistent but that every single ship in our sample has some cases of inconsistent or even inaccurate data. We did not present results for the sixth measure, accuracy.

Our overall conclusion is that the AIS system generates data that is useful for analysis and practical application, but the system's design, and a lack of supervision on the quality of the data, result in potential flaws. System design issues relate to the broad use of default values and the option to allow free text data entry for specific voyage-related data fields. In addition, we find considerable problems with the information that AIS data conveys about the ships' manoeuvring. Many data elements related to this (manoeuvre indicator, rate of turn) are unreliable, and the requirement to increase the frequency of messages for ships changing course also suffers. This is a serious problem for a system that was designed to support the safety of navigation.

Our conclusions support much of our colleagues' route reconstruction, destination prediction and ETA calculation work. We hope that our quantitative assessment of data quality issues assists in a more detailed discussion of data clean-up activities in future papers. In addition, our analysis also points to concerns about the use of AIS data as a basis for VTS systems. There are severe inaccuracies, especially in the manoeuvring information in the data, that need to be addressed to obtain reliable traffic situational awareness.

For further research, we recommend that our, or a similar, data quality program is carried out in the AIS data pools that exist in the world and reported on transparently. Many researchers

and practitioners obtain their data from these sources, and they should expect a data quality report that helps them assess the reliability of their data. Finally, we advise the International Maritime Organization (IMO) and the International Association of Marine Aids to Navigation and Lighthouse Authorities (IALA) to consider updating the AIS specifications to mitigate risks related to data quality found in this research.

References

Androjna, A., Perkovič, M., Pavic, I., & Mišković, J. (2021). AIS data vulnerability indicated by a spoofing case-study. *Applied Sciences 11*(11), 5015.

Batini, C., Cappiello, C., Francalanci, C., and Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM Comput. Surv. 41*(3), Article 16 (July 2009), 52 pages

Chen, X., Ling, J., Yang, Y., Zheng, H., Xiong, P., Postolache, O., & Xiong, Y. (2020). Ship trajectory reconstruction from AIS sensory data via data quality control and prediction. *Mathematical Problems in Engineering*, 1-9.

Chen, X., Chen, H., Xu, X., Luo, L., & Biancardo, S. A. (2022). Ship tracking for maritime traffic management via a data quality control supported framework. *Multimedia Tools and Applications 81*(5), 7239-7252.

Counsell, S., Loizou, G., & Najjar, R. (2007). Quality of manual data collection in Java software: an empirical investigation. *Empirical Software Engineering 12*(3), 275-293.

El Mekkaoui, S., Berrado, A., & Benabbou, L. (2022). Automatic Identification System Data Quality: Outliers Detection Case. Paper presented at the International Conference on Industrial Engineering and operations Management, Istanbul, 7-10 March 2022.

Emmens, T., Amrit, C., Abdi, A., & Ghosh, M. (2021). The promises and perils of Automatic Identification System data. *Expert Systems with Applications*, *178*, 114975.

Fox, C., Levitin, A., & Redman, T. (1994). The notion of data and its quality dimensions. *Information processing & management 30*(1), 9-19.

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

He, W., Lei, J., Chu, X., Xie, S., Zhong, C., & Li, Z. (2021a). A visual analysis approach to understand and explore quality problems of AIS data. *Journal of Marine Science and Engineering 9*(2), 198.

He, W., Liu, X., Chu, X., Wang, Z., Fracz, P., & Li, Z. (2021b). A Novel Fitting Model for Practical AIS Abnormal Data Repair in Inland River. *Elektronika ir Elektrotechnika 27*(1), 60-70.

Huang, L., Wang, J., Huang, Y., Zhu, M., Wen, Y., & Zhou, Y. (2025). Probabilistic prediction of ship destinations based on traffic pattern awareness in maritime networks. *Ocean Engineering*, *316*, 119933.

IMO (2015). Revised Guidelines for the Onboard Operational Use of Shipborne Automatic Identification Systems (AIS). Resolution A.1106(29), adopted 2 December 2015.

Iphar, C., Napoli, A., & Ray, C. (2015, September). Data quality assessment for maritime situation awareness. In *ISSDQ 2015-The 9th International Symposium on Spatial Data Quality* (Vol. 2).

Jankowski, D., Lamm, A., & Hahn, A. (2021). Determination of AIS Position Accuracy and Evaluation of Reconstruction Methods for Maritime Observation Data. *IFAC-PapersOnLine 54*(16), 97-104.

Jaskólski, K., Marchel, Ł., Felski, A., Jaskólski, M., & Specht, M. (2021). Automatic Identification System (AIS) Dynamic Data Integrity Monitoring and Trajectory Tracking Based on the Simultaneous Localization and Mapping (SLAM) Process Model. *Sensors 21*(24), 8430.

Krasikov, P. & Legner, C. (2023). A Method to Screen, Assess and Prepare Open Data for Use. ACM *Journal of Data and Information Quality 15*(4), 1-25.

Kiersztyn, A., Czerwiński, D., Czermański, E., Oniszczuk-Jastrząbek, A., Smoliński, K., Miazek, P., ... & MIŚKIEWICZ10, R. (2024). Data integrity analysis on the example of AIS database. *Scientific Papers of Silesian University of Technology. Organization & Management/Zeszyty Naukowe Politechniki Slaskiej. Seria Organizacji i Zarzadzanie*, (208).

Lei, J., Chu, X., & He, W. (2021). Trajectory data restoring: A way of visual analysis of vessel identity base on optics. *Journal of Web Engineering*, 413-430.

Lee, E., Mokashi, A. J., Moon, S. Y., & Kim, G. (2019). The maturity of automatic identification systems (AIS) and its implications for innovation. *Journal of Marine Science and Engineering 7*(9), 287.

Liang, M., Su, J., Liu, R. W., & Lam, J. S. L. (2024). AISClean: AIS data-driven vessel trajectory reconstruction under uncertain conditions. *Ocean Engineering*, *306*, 117987.

McFadden, D., Lennon, R., & O'Raw, J. (2019, September). AIS Transmission Data Quality: Identification of Attack Vectors. In *2019 International Symposium ELMAR* (pp. 187-190). IEEE.

Mieczyńska, M., & Czarnowski, I. (2021). DBSCAN algorithm for AIS data reconstruction. *Procedia Computer Science*, *192*, 2512-2521.

Meyers, S. D., Azevedo, L., & Luther, M. E. (2021). A Scopus-based bibliometric study of maritime research involving the Automatic Identification System. *Transportation research interdisciplinary perspectives 10*, 100387.

Meyers, S. D., Yilmaz, Y., & Luther, M. E. (2022). Some methods for addressing errors in static AIS data records. *Ocean Engineering*, *264*, 112367.

Qu, X., Meng, Q., & Suyi, L. (2011). Ship collision risk assessment for the Singapore Strait. *Accident Analysis & Prevention*, *43*(6), 2030-2036.

Raymond, E.S. (2023). AIVDM/AIVDO protocol decoding, version 1.58, 24 June 2023. https://gpsd.gitlab.io/gpsd/AIVDM.html#_introduction

Rivas, C., 2012. Coding and analysing qualitative data. *Researching society and culture*, *3*(2012), pp.367-392.

Serra-Sogas, N., O'Hara, P. D., Pearce, K., Smallshaw, L., & Canessa, R. (2021). Using aerial surveys to fill gaps in AIS vessel traffic data to inform threat assessments, vessel management and planning. *Marine Policy 133*, 104765.

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.

Stróżyna, M., Filipiak, D., & Węcel, K. (2020). Data quality assessment–a use case from the maritime domain. In *International Conference on Business Information Systems* (pp. 5-20). Cham: Springer International Publishing.

Veenstra, A. and Harmelink, R. (2021). On the quality of ship arrival predictions. *Maritime Economics & Logistics*, pp. 1-19.

Wolsing, K., Roepert, L., Bauer, J., & Wehrle, K. (2022). Anomaly detection in maritime AIS tracks: A review of recent approaches. *Journal of Marine Science and Engineering*, *10*(1), 112.

Yang, D., Wu, L., & Wang, S. (2021). Can we trust the AIS destination port information for bulk ships?–Implications for shipping policy and practice. *Transportation Research Part E: Logistics and Transportation Review 149*, 102308.

Yang, D., Wu, L., Wang, S., Jia, H., & Li, K. X. (2019). How big data enriches maritime research–a critical review of Automatic Identification System (AIS) data applications. *Transport reviews*, *39*(6), 755-773.

Zhang, J., Ren, X., Li, H., & Yang, Z. (2022). Incorporation of deep kernel convolution into density clustering for shipping AIS data denoising and reconstruction. *Journal of Marine Science and Engineering 10*(9), 1319.

Zhang, R., Indulska, M. & Sadiq, S. (2019). Discovering Data Quality Problems. *Bus Inf Syst Eng 61*(5), 575-593.

Zhao, L., Shi, G., & Yang, J. (2018). Ship trajectories pre-processing based on AIS data. *The Journal of Navigation*, *71*(5), 1210-1230.