YRKESHÖGSKOLAN
NOVIA

Quality Assessment of Maritime AIS Data

Aremu Jeremiah Anuoluwapo

Master Thesis

Thesis for a Master's (UAS)-degree

Automation Technology

Vaasa, 2023

MASTER'S THESIS

Author: Aremu Jeremiah Anoluwapo

Degree Programme and Place of Study: Automation Technology, Vaasa.

Specialization: Intelligent systems

Supervisor(s): Ray Pörn: Novia, Johan Westö: Institute of Intelligent Systems

Title: Quality Assessment of Maritime AIS Data

Date: 05.05.2023 Number of pages: 49 Appendices: 0

Abstract

This thesis includes a quality assessment investigation of Automatic Identification System (AIS) data retrieved through the ARPA project data platform from Digitraffic. Automatic Identification System (AIS) data is essential in improving the global shipping industry's safety, efficiency, environmental performance, and operations. The dataset includes location, navigational, and static data from thousands of ships from the Baltic Sea geographical region.

The research examines the literature on AIS data quality through which an assessment concept was constructed. This mixed-method approach combines qualitative and quantitative data analysis techniques to identify the elements that influence the data's quality and develop strategies for measuring it. The investigation focuses on four critical aspects of data quality: accuracy, completeness, consistency, and timeliness.

The findings show that AIS technology, communication protocols, ambient conditions, and human variables all impact the quality of marine AIS data. Therefore, to address these issues, the dissertation provides a set of quality indicators and data validation procedures. The effectiveness of the quality assessment procedures in identifying AIS data quality concerns was demonstrated.

According to the study, ongoing monitoring and improvement of AIS data quality is still required to improve marine safety and decision-making, ultimately making it ideal for autonomous shipping where the data is needed with a high degree of integrity.

Language: English

Key Words: data quality, Data analysis, AIS data, AIS transmission, Autonomous ship

---

Contents

List of Figures

List of Tables

List of abbreviations

IMO	International Maritime Organization
ARPA	Applied Research Platform for Autonomous Systems
AIS	Automatic Identification System
SOLAS	Safety of Life at Sea
IALA	International Association of Marine Aids to Navigation and Lighthouse Authorities
VTS	Vessel Traffic Services
AtoN	Aid to Navigation
SAR	Search and Rescue
MASS	Maritime Autonomous Surface Ship
EMSA	European Maritime Safety Agency
ROT	Rate of Turn
SOG	Speed Over Ground
COG	Course Over Ground
HDG	Heading
MMSI	Maritime Mobile Service Identity
GPS	Global Positioning System
FCS	Frequency Check Sequence
GMSK	Gaussian Minimum Shift Keying
VHF	Very High Frequency
NMEA	National Marine Electronic Association
SART	Search and Rescue Transponders
TDMA	Time Division Multiple Access
SOTMA	Self-Organizing Time Division Multiple Access
ITDMA	Incremental Time Division Multiple Access
RATDMA	Random Access Time Division Multiple Access
FATDMA	Fixed Access Time Division Multiple Access
CSTDMA	Carrier-Sense Time Division Multiple Access
GPS	Global Positioning System
DSC	Digital Selective Calling
MCAR	Missing Completely at Random

MAR	Missing at Random
MNAR	Missing not at Random
ITU	International Telecommunication Union
ISO	International Organization for Standardization
ASCII	American Standard Code II
HTTP	Hyper-Text Transfer Protocol

1. Introduction

One of the International Maritime Organization's significant achievements (IMO) is the Safety of Life at Sea (SOLAS) convention. It was adopted on June 17, 1960. But it came into force on May 26, 1965. By December 2002, in another conference, governments contracted to SOLAS concluded that; mandatorily, from December 2004, ships with more than 300 gross tonnage, cargo ships of more than 500 gross tonnage, and all passenger ships, irrespective of their size, must have AIS fitted aboard [3].

AIS is a data exchange scheme that enables the transmission of ship data continuously and at regular intervals, providing a detailed and exhaustive dataset of individual ships [1]. Although, it may have been developed originally with maritime safety and security in mind. However, due to the growing application of artificial intelligence and machine learning techniques in contemporary maritime innovations, its usage has since extended to other areas of marine operation. For instance, AIS-based algorithms can analyze real-time AIS data, utilizing it to track vessel movement, optimize routes and monitor vessel performance, bringing about considerable improvement in the operational efficiency of vessels. Moreover, resource management is not left out, as insights generated from AIS data help optimize resource allocation, manage fuel consumption, and organize cargo loading and crew scheduling. Therefore, AIS data utilization in administering maritime resources results in significant cost savings and reduced environmental impact of shipping.

Researchers have increasingly taken advantage of the abundance of data that AIS makes available as one of the fundamental building blocks for developing technologies that will increase efficiency in core maritime operational domains. Examples of such fields include Vessel Traffic Services (VTS), Aid to Navigation (AtoN), Search and Rescue (SAR), and Maritime Autonomous Surface Ships (MASS). MASS, for example, are ships that can operate without human intervention to varying degrees of autonomy. Thus, the IMO identified and provided a regulatory framework for four degrees of independence in autonomous ships: ships with automated processes and decision support, remotely controlled ships with seafarers on board, remotely controlled ships without seafarers onboard, and fully autonomous ships [2]. One critical peculiarity of these variants of vessels is that they can make some forms of intelligent decisions. Generally, intelligent

systems exhibit aspects of human intelligence, such as learning via extrapolated reasoning (data-to-knowledge). Furthermore, the learning efficiency of these smart systems can also improve over time and space. Therefore, through knowledge derived from data via a learning process, such systems can respond appropriately to changes in their immediate environmental conditions [2]. Figure 1. for instance, shows the functional block diagram for MASS as presented by the European Maritime Safety Agency (EMSA), highlighting the AIS data as one of the critical inputs MASS requires for condition detection.

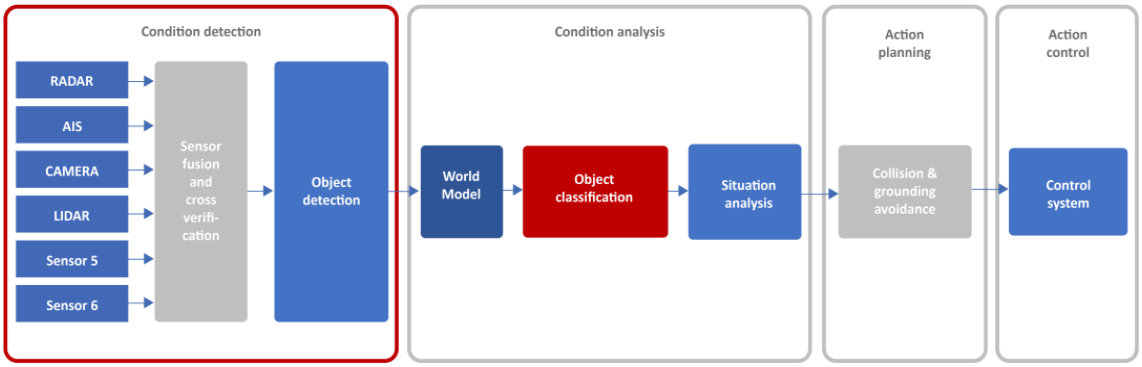


Figure 1. Functional block diagram of MASS [3]

Furthermore, as is the case for systems that operate based on data-driven forms of control, the integrity of the data they process is paramount as it is responsible for the accuracy of their decision-making. Moreso, some specific AIS data utilization areas have almost no margin for error. Hence, the actions and reactions of such AIS data-reliant autonomous systems to changes in their environmental variations must be highly accurate. To this end, ensuring the quality of the AIS data used in the various data-dependent maritime operations is critical, as the occurrence of functional errors can be of enormous safety and economic consequence. Regarding safety repercussions, reliance on inaccurate AIS data can result in collisions, groundings, or other disastrous incidents, resulting in loss of life, environmental harm, and property damage. On the other hand, the economic effects are reduced income or profit margin due to delays, increased expenses, and revenue loss. Although the accuracy and integrity of AIS data can be ensured by regular testing and condition monitoring of the AIS equipment and sensors responsible for generating the data,m a more precise way to detect problems or abnormalities in the AIS is to assess the data it produces.

This thesis attempts to investigate the quality of AIS data by evaluating the consistency and accuracy of the information it contains. Analyzing the quality of AIS data will provide an overview of the current state of its journey of becoming an all-inclusive source of data for maritime operational activities since AIS data serves as the primary driver of autonomous maritime operations. Furthermore, this work will assess if it is fit for use, especially in delicate aspects of the maritime domain, such as navigation (which involves anti-collision and path-finding algorithms), where the data must be of high integrity owing to the fatalities associated with operational errors.

Additionally, analyzing the AIS data can provide insight into the antenna or transceiver faults, and GPS signal loss, enabling timely corrective actions before safety or operational problems occur. Similarly, trends in the data, after the information in the data has been analyzed for accuracy, will provide important intuition into the traffic patterns in the Baltic Sea.

2. Background of study

AIS is a technology that is widely used in the marine industry for vessel tracking, navigation, and collision avoidance. It gives real-time data on vessel positions, speeds, and course over ground aiding maritime safety and operational efficiency. However, the usefulness of the AIS is contingent on the quality of the AIS data. Environmental circumstances, faulty devices, human errors, and intentional spoofing can all impact maritime AIS data. Inadequate AIS data quality can result in inaccurate vessel locations, missing or incorrect data, and improper vessel routes, which can endanger navigation safety or impede successful marine operations.

Data quality analysis is essential in ensuring maritime AIS data's accuracy and reliability. It involves identifying and addressing issues such as missing data, data point inconsistency, and errors in whichever form. However, while there has been extensive research on the quality of AIS data analysis using traditional data analytic methods, this study seeks to add to the body of knowledge on validating the integrity of AIS data with specific reference to the Baltic Sea. Aside from investigating the quality issues in the data, this study will propose a framework for analyzing and addressing the problems. The research will involve collecting and analyzing real-world AIS data. The results of this research will contribute to understanding data quality issues in the maritime AIS and provide practical guidance for developing and implementing an effective data quality assessment.

AIS data has immensely aided research in navigation-related spheres and other areas, such as trade flow estimation and emission estimation. For instance, in their study, Perez H et al. estimated vessel emissions along the inland river near the Texas coast using AIS data, Geographic Information System (GIS), Lloyd’s, and the American Bureau of Shipping Register of Ships. They matched each vessel's fuel and engine characteristics, combining it with ship tracking data from the AIS. Emission factors were then applied to quantify criteria and hazardous air pollutant emissions from these vessels [1]. Furthermore, a comprehensive literature review in 2019 categorized the contemporary AIS research publications into seven application-based divisions: AIS data mining, navigation safety,

ship behavior analysis, environmental evaluation, trade analysis, arctic shipping, and, lastly, ship and port performance [4]. There is indeed a promise of more to come in the application of AIS data owing to the ongoing digitalization in the maritime sector.

Hitherto, AIS has poor security features. The system is currently limited in terms of security as it is not encrypted, posing an integrity concern for its users. For example, unlawful mariners can quickly sabotage the AIS system by deliberately updating erroneous data to hide illegal activities. Similarly, other actors merely input false datapoint unintentionally, while others capitalize on the AIS's security inadequacies to engage in spoofing to mislead undiscerning mariners [5].

Researchers have conducted a couple of studies to analyze the quality of AIS data. For instance 2014, Last et al. undertook an application-specific integrity analysis of AIS data regarding vessel movement prediction with data gathered from the German North Sea coast for two months [6]. Also, Harati-Mokhtari et al., in their research, bared the various errors inherent in AIS messages. As findings, the authors discussed errors in the static data that may have originated from wrongly entered static information at the AIS commissioning period and false voyage information due to wrong data entry by the crew when the vessels were in operation [7]. Generally, most studies in the quality analysis of AIS data analyzed the data based on its intended purpose. However, some examined the integrity level of the data from a specific geographical area. Hence, this study is based on the oceanic space of the Baltic Sea.

Other studies assessed the quality by comparing the integrity level of the data from different AIS data service providers. Unfortunately, previous AIS data quality assessment research findings do not indicate comprehensive studies that carried out sanity checks on AIS data from vessels operating in the Baltic Sea. Since data quality verification remains an obligatory preliminary task for any application that uses data, this study will advance a thoughtful approach to a general validation of AIS data even beyond its intended usage. Therefore, this work will identify aspects of the data that are updated manually from the ones that are updated automatically with measurements from sensors connected to AIS, thereby observing the peculiarities of the errors in both. This will involve analyzing the accuracy and completeness of the data to ensure that it is suitable for use in various AI-based applications, as poor-quality AIS data can lead to poor judgments or incorrect decisions.

2.1 Automatic identification system

The information transmitted from ships equipped with AIS transponders is called AIS data. AIS data can be sent from other AIS transponder-fitted sources, including buoys, and received by other vessels, AIS base stations, or satellite-based receivers. Messages contained in the data include position, speed, course, vessel types, etc. Currently, this data supports a variety of purposes, including research making it indispensable in the maritime industry [8]

2.2 Types of AIS data

Depending on the type of vessel or voyage-related information contained in the data, which in turn is related to how the data points are measured and updated, AIS data are of three types which are:

- Dynamic AIS data

Dynamic AIS data refers to information about the position and movement of marine vessels. The primary positional information is updated as longitude and latitude coordinates. At the same time, it contains data related to vessel displacement, which includes information such as speed over ground (SOG), course over ground (COG), heading (HDG), and navigational status. Usually, sensors linked with the AIS system are responsible for continuously measuring and updating these data every few seconds at intervals set in the IMO standard. Dynamic AIS messages are also known as positional messages.

- Static AIS data

Static AIS data usually do not change (at least not often) but are saved during the initial

installation of the AIS equipment. It comprises information related to the characteristics of the vessel. Examples of these data include Maritimes Mobile Service Identity (MMSI) number, IMO number, call sign, ship name, ship type, and global positioning system (GPS) antenna location; the antenna location is essential for determining the ship's dimension.

- Voyage-related AIS data

Voyage-related AIS data are information concerning the current voyage that the ship is undergoing, and they are also manually updated. Typically, the recommended updating interval is 6 minutes. Nevertheless, there are exceptions as AIS data are immediately transmitted when a field in the voyage-related dataset changes. Examples of these include destination, expected time of arrival, and draught.

2.3 AIS transmission

The transmission of AIS messages remains an operational requirement for maritime and inland waterway navigation. The messages are transmitted by vessel-fitted transponders, through either of two designated VHF channels, A (87B) or B (88B), with an operating frequency of 161.975 MHz or 162.025 MHz, respectively. Messages transmitted on these channels are encapsulated in time slots, and each channel comprises of 2250 time slots per 60 seconds (i.e., 4500 time slots across both channels every 60 seconds). Therefore, a single time slot is equivalent to 26.67 ms. AIS transmission is based on a 9600-bit/s GMSK (Gaussian Minimum-Shift Keying) modulation. The bandwidth AIS uses depends on the territory's location or authority. On the high seas, the bandwidth is 25 kHz; but in territorial waters, it can be either 25 kHz or 12.5 kHz. The total length of a default packet corresponding to 1 slot is 256 bits, out of which the actual positional message can occupy a maximum of 168 bits; other bits are reserved for the training sequence, start flag, Frequency Check Sequence (FCS), end flags, and buffer.

AIS data packet is sent from left to right in the order shown in Figure 2. The training sequence synchronizes the VHF receiver. The size of the AIS messages can exceed 256 bits if it includes static, voyage-related messages, and binary broadcast messages.

Therefore, requiring more than one time slot out of a maximum of five time slots to transmit the complete message, as the data section containing the position message is only up to 168 bits long, as seen in Figure 2. [9]

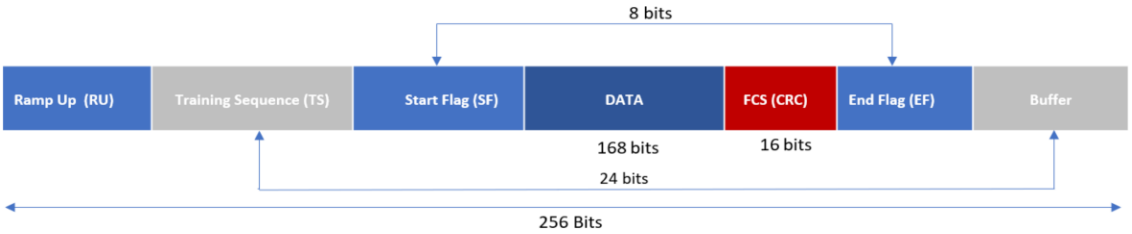


Figure 2. Bit allocation in a time slot [9]

Furthermore, due to the rising number of ships signing up for the AIS, the network has witnessed a rapid expansion over the years. This expansion spurred changes and improvements in how the data is transmitted to maintain reliability, availability, and integrity. Terrestrial AIS was the first transmission technology used. However, it was soon complemented with satellite AIS because the need to have AIS coverage beyond port areas and coastal regions alone arose, but terrestrial AIS is unfortunately limited in terms of coverage. In fact, a new AIS data transmission scheme, VDES (Vessel Data Exchange System), is undergoing final testing and will soon be launched. The new AIS variant offers a more secure and efficient method of data transmission. The following section will address the characteristics of terrestrial and satellite AIS.

2.4 Terrestrial AIS transmission

Terrestrial AIS was the first model of AIS data exchange. Its mode of operation is such that messages are transmitted at periodic intervals from vessels within a limited coverage area. At the same time, the signals are intercepted by an antenna at the transmission base station, creating a network that makes it possible to track the location of vessels at the various ports or along coastal routes where the AIS base stations are installed. Depending on the antenna's elevation at the transmission base station, terrestrial AIS covers 40-60 nautical miles. Additionally, the reach of the terrestrial AIS also depends on other factors, such as weather and environmental obstacles. Figure 3. shows the global coverage extent of the terrestrial AIS [10].

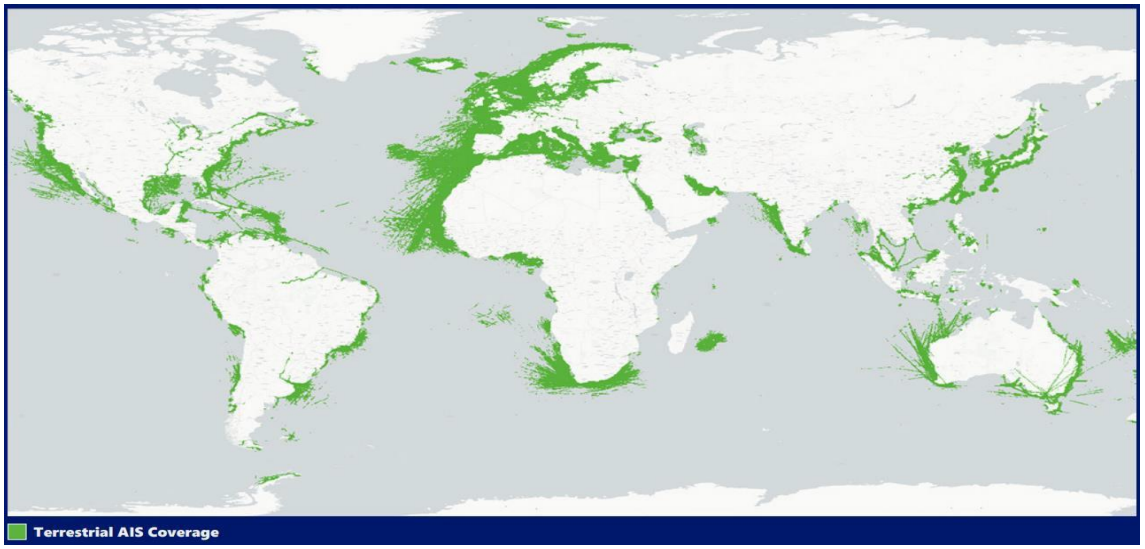


Figure 3. Terrestrial AIS coverage [10].

2.5 Satellite AIS transmission

On the other hand, Satellite AIS is not a stand-alone AIS transmission scheme; it operates in tandem with terrestrial AIS, mainly to extend its coverage. Satellite AIS allows vessels to be reliably detected and tracked efficiently beyond coastal routes since mariners became interested in monitoring ships on the high sea, far away from territorial waters in any part of the world. Small satellite constellations with relatively low orbit altitudes between 600 – 1000 km help ensure worldwide coverage. Low-orbit satellites are employed due to the AIS transponders' limited transmitting capacity [11]. Therefore, these days with the complementary operation of terrestrial and satellite AIS, AIS service providers can receive signals from AIS-fitted vessels beyond coastal regions anywhere in the world. Figure 4. shows how satellite AIS has enhanced the AIS's reach and coverage.

However, satellite AIS is not without its challenges, as it is prone to frequency offsets caused by the doppler effect because of the speed of travel of the satellite [12]. The AIS signal can also suffer attenuation depending on the satellite's altitude [13]. Data collision is also an issue in satellite AIS. It can occur between the time it takes the signal to travel from the ship to the satellite, subject to the position of the vessel and the coverage area of the satellite's antenna [14]. Finally, when AIS messages are received, they are initially stored in the satellites' storage facility until a connection is established between that satellite and a ground station before downloading all the messages saved since the last

satellite-to-ground station connection. This phenomenon is responsible for latency issues with satellite AIS. The satellite constellation and the robustness of the ground station infrastructure determine the severity of the latency challenge encountered in satellite AIS. Nonetheless, having several ground stations suitably located to ensure a continuous satellite-to-ground station connection helps to mitigate latency problems.

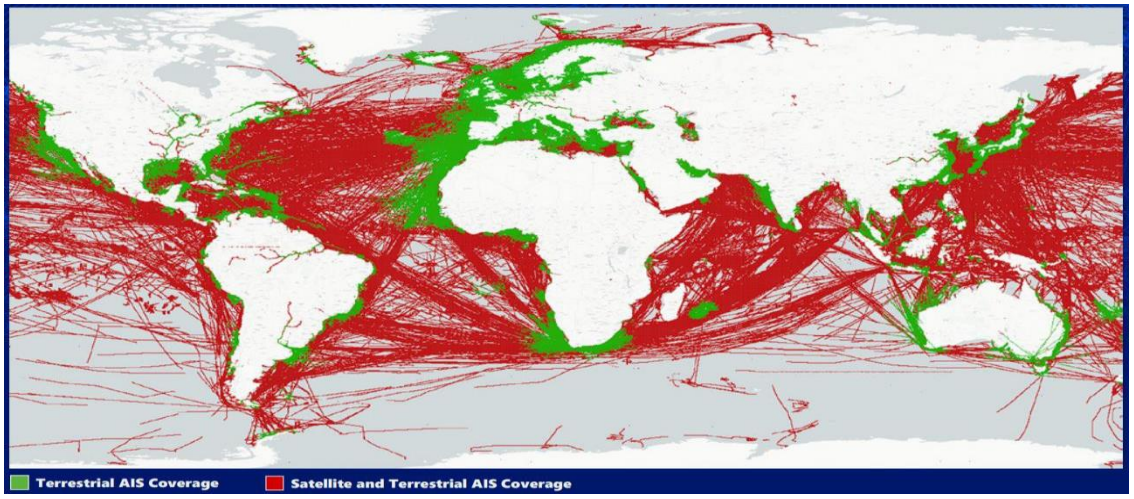


Figure 4. Satellite and terrestrial AIS coverage [10]

2.6 AIS data transmission protocols

AIS data is transmitted over two possible Very High Frequency (VHF) channels. The data carriers over the transmitting medium are time slots. Additionally, both channels have an equal number of time slots. But, depending on the size of the AIS message to be sent, a varying number of slots are required to transmit it since the time slots are limited in terms of their information-carrying capacity [9]. The limited data handling capacity of time slots makes the AIS susceptible to overloading, especially when several vessels crowded around a coastal area send messages simultaneously. But the AIS is designed in such a way that in locations densely populated with vessels, the system remediates the possibility of data collision due to overcrowding by allocating the time slots based on a logic that prioritizes closers vessels over those farther away, allowing the closer ones to transmit first.

To efficiently transmit AIS messages continuously and automatically as specified in the IMO standard, five different kinds of communication protocols are in use, and these

include; Self Organizing Time Division Multiple Access (SOTDMA), Incremental Time Division Multiple Access (ITDMA), Random Access Time Division Multiple Access (RATDMA), Fixed Time Division Multiple Access (FATDMA) and Carrier Sensing Time Division Multiple Access (CSTDMA).

2.6.1 Self-organizing time division multiple access (SOTDMA)

SOTDMA, a pre-announced time-dependent multiple access transmission protocol, is a commonly used transmission protocol for sending AIS data. Using a time slot map, SOTDMA can detect which time slot is being used and by what station, thereby avoiding it when it intends to transmit. But, when it engages a time slot for transmission, it also announces the next slot it plans to use, thereby reducing incidences of message collision as stations continue to avoid time slots known to be used or preserved by other stations.

However, transmission stations in motion come across other stations with new sets of time slots and may change their time slot depending on the ones other stations engage. Consequently, due to this constant re-organization of time slot allocation, the SOTDMA transmission protocol is described as self-organizing over time and area. In addition, a Modified- SODTMA exists, which operates on a simpler TDMA access scheme, used in a transmit-only device, primarily for emergency beacon applications such as search and rescue transceivers (SART). It randomly picks a time slot for transmission at an interval of 8 minutes; during this time interval, a burst of 8 messages per minute is transmitted to guarantee successful transmission, irrespective of any reception-inhibiting condition.

2.6.2 Random access time division multiple access (RATDMA)

When RATDMA transmission is in use, other stations on the channel do not detect the time slots carrying data. Therefore, due to the characteristics above, it is used mainly for initial entry into the radio link network. Also, because it randomly allocates time slots, it is suited for transmitting messages requiring frequent updates. For instance, when vessels are changing course [9].

2.6.3 Incremental time division multiple access (ITDMA)

ITDMA transmission protocol uses a GPS time-based reference shared by all stations using the protocol to determine the start time of each TDMA slot precisely. Using a kind of "slot status log," the transceivers randomly select a time slot not used by other stations for its future use. In the ITDMA transmission scheme, all stations announce their current transmission slot. It is mainly used when the need arises to temporarily change data's updating interval or to pre-broadcast occasional safety-related information. ITDMA is always used to support the SOTDMA protocol and usually does not operate alone [9].

2.6.4 Fixed access time division multiple access (FATDMA)

As is the case for ITDMA, the FATDMA protocol also shares a common GPS-based time reference for the determination start time of a time slot. But in this case, the station's transmitting time slots are fixed as they are allocated during installation, making it a manually managed ITDMA. Stations that use this communication protocol broadcast a datalink management message to inform other stations of the slots allocated to it and disallow other stations within the range from using it. The use of FATDMA is highly controlled because it can adversely impact the AIS network's dynamic behavior. They are used by AIS base stations or navigation stations [15].

2.6.5 carrier-sensing time division multiple access (CSTDMA)

CSTDMA transmission protocol is used mainly by Class B AIS stations. Generally, AIS is divided into class A or B based on the kind of transponders used, as described immediately in the subsequent section. It allows for the use of low-end transceivers that equally support SOTDMA transmission. However, this interoperability is such that it prioritizes SOTDMA. CSTDMA does not operate on GPS timing like the others; its timing is derived from class A or an AIS base station transmissions with the range of the receiver. The selection of a slot for data transmission is determined by analyzing the background noise level on the radio channel to determine whether a particular time slot is in use. At transmission instant, a TDMA is selected randomly, and the signal strength at the start is measured. If the signal strength considerably exceeds that of the background noise level,

then such slot is considered to be in use, and so data transmission is postponed; otherwise, transmission proceeds [15].

2.7 Class A and Class B automatic identification system

Based on the features of an AIS transponder, AIS is subdivided into two types which are class A and class B. There are also certain peculiarities in the cost and scope of an AIS under this classification. IMO compliance requirements for class A transponders include characteristics such as a transmitting power of 12.5 watts using the SOTDMA protocol and a Digital Selective Calling (DSC) receiver operating at a frequency of 156.525MHz. In addition, External GPS, heading, and rate of turn indicators characterize a class A AIS. Lastly, class A AIS must be able to transmit and receive safety-related messages. On the other hand, class B transponders do not necessarily have to comply with IMO standards. They operate using a CSTDMA transmission protocol with a transmitting power of 2 watts. Class B DSC receiver operates with a lower frequency than class A and transmission of safety-related messages is optional in this case but can be installed if desired [16].

The SOTMA protocol-based class A transponders enjoy transmission priority over class B; hence, its messages are shown to other ships and stations in the area ahead of class B. Therefore, the extent to which class B messages are transmitted would depend on how preoccupied the channels are with class A transmission. Operationally, class A AIS is usually installed on bigger ships, while class B is mainly installed on recreational boats. Table 1 below shows the differences between class A and B AIS transponders.

Table 1. Difference between class A and B transponder features [17].

	Class A AIS (SOLAS Compliant)	**Class B AIS**
Transmit power	12.5 watt (nominal), 2 watt (low power)	2 watts
Unique communication access scheme	SOTDMA (Self Organizing amongst Class A	CSTDMA (Carrier-Sense peculiar to Class A)
Frequency Range	156.025 – 162.025 MHz @ 12.5/25 KHz, DSC (156.525MHz) is required	156.025 – 162.025MHz @ 25KHz, DSC (156.525MHz) and 12.5 KHz are optional

Miscellaneous	External GPS, Heading, and Rate of Turn indicators are Required	Heading is optional
Safety text messaging	Transmits and receives	Transmit is optional and only pre-configured

2.8 AIS standard messages

There are 27 different AIS message types, and they are described in the AIS technical standard ITU-R M.1371-pg 91 - 92. They are categorized based on their datalink functions, such as message acknowledgment, interrogation, assignments, or management commands, and of course, based on the information they contain.

Briefly highlighting the important ones, AIS messages 1, 2, and 3 represent both positional messages scheduled and transmitted as a response to interrogation. Number 4 is for the base station report, the location, UTC, and slot number. Message 5 is for static and voyage-related message information for Class A vessels. 6, 7, and 8 are for Binary addressed messages, Binary acknowledgment, and Binary broadcast messages, respectively. AIS message 18 gives an account of the position for class B vessels, while message type 24 gives the static reports for class B vessels. Finally, message type 27 is called the long-range AIS; hence it provides scheduled position reports designed for satellite detection [18].

2.9 NMEA messages

The National Marine Electronics Association (NMEA) developed a standard for data exchange between marine electronic devices. NMEA messages encode AIS data into a format that can be transmitted over the VHF radio [19].

By enabling equipment interconnection and interchangeability, the NMEA interface reduces misunderstanding and confusion among producers of electronic devices, saving buyers the trouble of equipment compatibility. It standardizes electrical signal requirements, data transmission, timing, and sentence format for a 4800-baud serial data bus by providing an adequate one-way bus connection between marine and navigation electronic equipment. NMEA protocol supports one-way communication; only between a single "talker" and a multiple "listener," where a talker refers to a device

that sends data to other devices. In contrast, the listener is any device that receives data from another device operating within the NMEA standard [20].

AIS data are encoded as NMEA sentences, and the common ones begin with any of these two prefixes: "! AIVDM" or "! AIVDO". After transmission, they are decoded at the receiving end to extract the information contained. Messages with the prefix "! AIVDM" represents data from other vessels, while those with "! AIVDO" indicates data from your vessel. NMEA sentences are made up of words separated by a comma. A typical sample of an encoded AIS message is presented below:

*! AIVDM,1, 1, ,A,15086n001TJ3KutH8ar@<h;106Hh,0*5D*

The description of special characters in an NMEA AIS sentence is highlighted in Table 2. When the data for a field is unavailable, the space corresponding to that field is left blank. Nevertheless, It retains the comma delimiter as observed in the 5th character of the NMEA message sample presented above.

Table 2. NMEA sentence field description [20]

NMEA sentence field	Description
<*5D >	Data integrity checksum computed over the entire sentence except for the exclamation mark
<LF>	Line feed, end delimiter
!	Start of encapsulation sentence delimiter
$	Start Delimiter
*	Checksum
,	Field delimiter
\	TAG block delimiter
^	Code delimiter for HEX representation of ISO/IEC 8859-1 (ASCII) characters
~	Reserved

Field 1 – the first field after the exclamation mark, as earlier stated, indicates if the data was received from other ships or your own ship.

Field 2 – represents the number of fragments that make up the message since a single NMEA 0183 sentence has a maximum of 82 characters, out of which the actual positional messages can take up to a maximum of 51 characters. Therefore, AIS messages with sentences greater than the maximum are encoded in multiple sentences, but fields 2, 3, and 4 are not repeated in the adjoining sentence(s)

Field 3 – This field indicates the number of a particular sentence within the multiple sentences that make up the AIS message, but it is 1 in this case since it is a single sentence.

Field 4 – Missing in the example above because it is a single sentence but stands for the sequential message I.D. for messages with multiple sentences.

Field 5 – (A), in this case, is a representation of the radio channel code and can either be A for class A or B for class B AIS messages, occasionally 1 or 2 can be encountered, and that would readily mean 1 was used in place of A and 2 in place B.

Field 6 –an essential part of the encoded message as it is the data payload. A decoder is required to extract this information as it cannot be viewed by mere inspection.

Field 7- can take values from 0 to 5 (0 in this case). It indicates the number of vacant bits to be filled to push the data payload bit count to 6

Field 8 – the word after the asterisk (5D in the example given above) is the NMEA checksum for data-probity of the entire sentence except for "!" or "$," as the case may be.

AIS message decoder-script is then used to decipher the information in the AIS-encoded NMEA sentence. The decoders are commonly developed with C or python programming language. For instance, pyais is a popular python library for decoding NMEA messages. NMEA message decoders must be able to accept coded data from an AIS receiver or the internet and decode all 27 AIS messages, including binary messages, which are occasionally used to broadcast safety messages and save the data in a proper file format [20].

2.10 AIS data quality

Assessing the quality of a dataset has both qualitative and quantitative sides as well as a subjective and objective dimension. The objective component can either be task-dependent or task-independent. The task-dependent aspect refers to factors such as organizational business rules or government regulation metrics against which the integrity of such data is weighed. The task-independent metrics, on the other hand, are a function of the quality condition of the data, irrespective of the precise knowledge of its intended use; they are termed objective because it applies to all kinds of data [21]. The subjective aspect of data quality assessment involves the requirement of the stakeholders ranging from the collectors, the custodian, and the consumers of the data itself [22].

In his study in 1999, English described the subjectivity of data quality. He claimed that the best way to evaluate data quality is to consider what data quality generally means and to establish what quality means for a particular dataset. Therefore, it is common practice to determine the quality of a data collection based on fit-for-use conditions such as "whether it is error-free?", "if it meets consumer's expectation?", "Does the data conform to some predefined standard?" and so on. The following sections describe some of the critical data quality metrics used in this research, while Table 3., briefly explains other commonly used ones [23].

Table 3. Quality Measurement metrics and definition [20].

Data quality metric	**Definition**
Accessibility	The extent to which data is available or easily and quickly retrievable.
An appropriate amount of data	The extent to which the volume of data is appropriate for the task at hand.
Believability	The extent to which a dataset can be regarded as authentic and credible.
Ease of manipulation	The extent to which data is easy to manipulate and apply to different tasks

Interpretability	The extent to which data is in appropriate languages, symbols, units, and the definition is clear.
Objectivity	The extent to which data is unbiased, unprejudiced, and impartial.
Relevancy	The extent to which data is applicable and helpful for the task at hand.
Reputation	The extent to which data is highly regarded in terms of its source or content.
Timeliness	The extent to which the data is sufficiently up-to-date for the task at hand.
Understandability	The extent to which data can be easily comprehended.
Value-added	The extent to which data is beneficial and provides advantages from its use.

2.10.1 Completeness

Completeness in data quality assessment involves checking for missing data point(s) in a dataset. When they exist, it is essential to investigate why, as this plays a significant role in determining the best course of action for tackling the cause as well as the best way to fill them if they cannot be ignored. Yet, completeness can be misunderstood if the measured value is not correctly analyzed. For instance, a missing record of the number of children in a family observation with no kids can be erroneously adjudged as missing data. Missing data is often measured by obtaining the percentage ratio of incomplete to complete entries in a dataset [24].

2.10.2 Accuracy

Data accuracy is one of the most critical data quality measurement metrics; it measures how correctly the value of data represents the real-world quantity or scenario for which it was intended. For some data applications, decisions based on inaccurate data can be very damning as they result in wrong judgment. Therefore, the accuracy is often measured via the verification process through cross-validation against an authentic

reference and sometimes by carrying out any possible confirmatory tests [24].

2.10.3 Conformance

Conformance measures how well the data complies with predefined standards, rules, or requirements. The quality criteria can be internal or external to the organization, including industry-specific regulations, data governance policies, or quality control procedures. It is a percentage measure representing how much a data point value matches the reference standard.

For instance, IALA guide 1082, annex C, page 25, highlights the range of values for some position messages and their default values. Default values are placeholder entries used where the actual observation to be measured is unavailable. Table 4 shows some of the standard fields in positional AIS messages, their range of values, and default values. This information will later be utilized as a yardstick in this study to validate the AIS data. Furthermore, IALA also standardized the reporting intervals of AIS messages in IALA guide 1082, annex C, page 17, as seen in Table 5. Based on these standard intervals, transponders are configured to transmit AIS data with specified exceptions depending on the vessel's prevalent course over ground (COG) or speed over ground (SOG).

Table 4. Range and default values for position messages

Data Field	**Unit**	**Range**	**N.A. default value**
longitude	[°]	[± 180]	181
latitude	[°]	[± 90]	91
rate of turn (ROT)	[°/m]	[± 127]	128
speed over ground (SOG)	[kn]	[0.1022]	1023
course over ground (COG)	[°]	[0.3599]	3600
heading (HDG)	[°]	[0.359]	511
position accuracy (ACC)	[_]	[true, false]	false

Table 5. AIS Message Transmission intervals

Ship's Dynamic Conditions	**Normal reporting intervals**
Ships at anchor or moored and not moving faster than 3 knots	3 minutes

Ship at anchor or moored moving faster than 3 knots	10 seconds
Ship 0-14 knots	10 seconds
Ship 0-14 knots and changing course	3 1/3 seconds
Ship 14 – 23 knots	6 seconds
Ship 14 – 23 knots and changing course	2 seconds
Ships > 23 knots	2 seconds
Ship > 23 Knots and changing course	2 seconds

2.11 Missing Data

Investigating missing data is one of the fundamental operations performed when cleaning or exploring a data collection. Regarding AIS data, some known sources of missing data include data collision during transmission or an unavailable or malfunctioning sensor for positional messages. On the other hand, it is often a case of outright omission or failure to update this information for static and voyage-related messages.

Generally, the methods for handling missing data depend on factors such as the amount of data lost, whether the missing entries strongly correlate with other variables in the dataset, and the missing data's importance to its intended application. The above-stated conditions determine the logic for populating or deleting the missing entries. Since they are unrecorded observations in a dataset, during data pre-processing and cleaning, the magnitude of the missing data plays a part in determining the handling procedure. Usually, when the percentage of missing data is low, they are often removed alongside the variable linked to them. However, inputting entries based on reasonable presumptions offers a better solution in cases where they are in high numbers.

Overall, before taking any action about dealing with missing data, it is vital to know the reason why it is missing. Missing data are of three categories based on the pattern or randomness of their occurrence, and these include; data missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Datapoints are considered MCAR if the missing variables are unrelated to any other observable variables, particularly within the same data collection. In contrast, the MNR

type is determined based on its relationship with other variables within the dataset. Finally, for MNAR, the missing data is related to why they are missing; hence they are referred to as nonignorable [25].

3. Aims and objectives of the study

The quality of AIS data can be affected by various factors, such as technical errors, intentional or unintentional interferences, and sometimes fraudulent activities. Therefore, assessing AIS data quality is crucial for ensuring reliable and accurate information about vessel movement. Conversely, poor-quality AIS data can lead to incorrect situation awareness, misinterpretation of vessel intentions, delayed or inappropriate responses to emergencies, and increased risks of accidents. Therefore, conducting a thorough AIS data quality assessment can advance a systematic method for identifying and analyzing error, bias, and their sources. The research goes further to propose measures for improving its quality which will, in turn, improve maritime operations, safety, and security.

The AIS data used for this research is a week-long data collected through a third-party AIS data service provider limited to vessels' maritime activities in the Baltic Sea area. The relatively large dataset contains only position, static and voyage-related messages. This study would attempt to establish a procedure for identifying gaps in AIS data quality, then put forward techniques for filtering and cleaning some of its most important variables, thereby improving the data's accuracy and completeness.

Furthermore, this research aims to develop a quality metric for evaluating the AIS data's accuracy, completeness, timeliness, and consistency. The findings will help identify patterns of errors and biases and their causes, providing insights for researchers, practitioners, and decision-makers in the maritime industry and contributing to the ongoing efforts to enhance the safety and efficiency of marine transportation. The research will also provide recommendations for future research in the field of AIS quality assessment and quality assurance.

4. Methodology

The methodology chapter outlines the methods and techniques used to conduct the study. This chapter explains the process of data collection, preparation, and analysis. It also gives the reader a comprehensive understanding of the study's data quality assessment methods and how they were applied to the data. Specifically, in evaluating the quality of the AIS data, this research carried it out from the point of view of the automatically updated and manually inputted component of the AIS data. First and foremost, an overview of the AIS data used for the research is presented, and the oceanic area it covers. Also, missing data, default values, and outliers were investigated as data preprocessing aspects of the quality analysis.

Furthermore, exploratory data analysis was conducted using descriptive statistical plots, where vessel positions and trajectories were visualized with spatiotemporal maps. Key fields in the positional, static, and voyage-related messages were tested for accuracy, with the overall essence of determining the integrity of the data, especially in terms of its conformance with IMO specifications.

4.1 Data overview

This thesis section informs on the AIS data source, the collection method, and a detailed description of its many features. Most importantly, it covers the twenty-six fields that make up the position, static and voyage-related messages.

4.1.1 Data source

The AIS data to be analyzed in this dissertation was collected from the ARPA project's data platform—the data platform, in principle, warehouses data from Digi traffic through the Finnish Transport Infrastructure Agency. Digi traffic provides maritime data such as port calls, vessel locations, sea state estimation, disturbance in waterway traffic, or AtoN faults. However, the ARPA project data Platform only stores positional message data and its adjoining static and voyage-related data. In addition, the AIS data used in this work is a one-week time series data from November 23, 2022, and November 30, 2022.

4.1.2 Data description

The AIS data was retrieved by sending an HTTP POST request to the data platform, which returns a tape archive file (tar file) containing both the position and static messages. Next, the data returned was ingested into the IDE for exploration in Apache Parquet file format. Parquet files offer an efficient data encoding and compression scheme for handling large data. The retrieved positional and static messages comprise the fields shown in Tables 6 and 7.

Table 6. Position messages field of the AIS data

Position messages		
S/no	**Field name**	**The field name in the database**
1	MMSI	mmsi
2	Longitude	longitude
3	Latitude	latitude
4	Speed over ground	sog
5	Course over ground	cog
6	Navigational status	navstat
7	Rate of turn	rot
8	Position accuracy	posacc
9	Receiver autonomous integrity monitoring	raim
10	Heading	heading
11	External timestamp	timestampexternal

Table 7. Static and voyage-related message data fields.

Static and voyage-related messages		
S/no	**Field name**	**The Field name in the database**
1	MMSI	mmsi
2	Vessel name	name
3	Vessel type	shiptype
4	Callsign	callsign
5	Destination	destination
6	Postype	postype
7	External timestamp	timestamp
8	IMO number	imo
9	Draught	draught
10	Expected time of arrival	eta
11	Reference point A-D	Referencepoint A-D

4.2 Data exploratory tools used

Data exploration tools are an essential part of any data quality investigation. These tools help to efficiently explore the dataset, discover any flaws or abnormalities, and accentuate important trends and insights. This section will go through different data exploration tools that were used for the analysis. Harnessing efficient tools for data exploratory analysis is critical; for data preparation not to become overly tedious and time-consuming [27]. Moreover, about 80% of the time spent on a data analysis task is spent on data preparation [28]. Therefore, choosing the right sets of data analysis tools can considerably reduce the workload and improve the efficiency of exploring and gaining insight from big data such as AIS data [29].

The data cleaning and preparation for this study borders on finding outliers in the data, type conversion for all the fields, spotting missing and default values, and investigating gaps in the time series. To carry out the data cleaning and preparation afore-mentioned, open-source tools were selected over commercial ones for easy reproducibility. Furthermore, easy-to-use application tools such as Excel and SPSS cannot handle data beyond specific sizes. For this reason, python scripting language, which is more versatile in handling a wide range of programming tasks, was selected for the analysis in this research. Specifically, the primary Python libraries used at this stage include Pandas and NumPy. In addition, Jupyter Notebook was chosen as the Integrated development environment (IDE) for the analysis because it is a web-based open-source IDE that supports the creation of live codes and provides computational outputs, visualizations, and supplementary texts in one document.

This section also explains the basis for selecting the various data analysis tools used at different stages of the research. For example, for descriptive statistics, the main features of the dataset are examined by intuitively understanding the data distribution, relationships, and patterns with suitable visualization tools. However, the quality gaps would only be flagged since the most appropriate methods for resolving them would usually depend on the intended purpose of use. Therefore, pandas, numpy, and matplotlib were the Python libraries chosen for this project's descriptive statistics and visual analysis. Although other visualization packages, such as Plotly and Seaborn, are

available and capable of performing similar tasks, with better aesthetics and fewer lines of code.

Nevertheless, the selection of matplotlib was based on the advantage of having a more comprehensive degree of freedom in controlling all aspects of the figures. In addition, AIS data also contains spatiotemporal information regarding the geographical coordinates of vessels' different positions and where they have been over time. Therefore, out of several commonly used geospatial analysis tools, which include cartopy, folium, geoviews, geopandas, Ipyleaflet, etc., cartopy was preferred as it was built on matplotlib. Overall, the statistical visualizations were principally produced with matplotlib and cartopy, with the information presented graphically as boxplots, scatterplots, histograms, bar charts, pie charts, and spatiotemporal maps.

4.3 AIS data quality assessment.

Quality assessment is a crucial aspect of data pre-processing. This section examines the automatically updated position messages by assessing the reasonability of their entries. This study investigates the quality of the position messages by confirming if the entries' range of values conforms with the IALA specification in Table 8. On the other hand, the integrity of the manually updated fields (mainly the static and voyage-related data) was not left out. However, it is limited to navigational status and ship dimension. This analysis treated navigational status as manually updated data because, despite being one of the positional messages, the vessel operators update this information manually [9]. The ship dimension was also considered for probing because of its importance as a critical feature in developing anti-collision and vessel maneuvering algorithms. These algorithms use ship dimension and relative position to determine a safe distance that ships should keep during an encounter or to predict future trajectories of the vessels to detect potential collisions ahead. For instance, in their study, "Beam Search Algorithm for Ship anti-collision trajectory planning," J. Karbowska-Chilinska et al. presented a detailed explanation and simulation of the beam anti-collision algorithm, with ship dimension as one of its critical parameters [30]. Another importance of evaluating the integrity of the manually updated data is that it forms an impression of the compliance level of the vessel crews in updating AIS information in the geographical area considered in the study. The

navigational status, for example, is a vital feature that can improve the accuracy of collision avoidance algorithms. For instance, one article that details navigational status's importance in collision avoidance is Hsiao et al.'s research in 2013 titled "An Analysis of the navigational status in collision avoidance," where the Arthurs argued that traditional collision avoidance often does not consider the navigational status of vessels, unlike other features such as intended route, speed over ground and purpose of the voyage. The study further proposed an algorithm that considers the navigational status of both own and target vessels to improve the effectiveness of the collision avoidance model [31].

4.3.1 Automatically updated AIS data quality assessment

Table 8 highlights IMO's range of values for position data fields. This standard range of values can help to verify if the position message data are correctly populated by investigating whether they fall within the specified range of values.

The position message fields are primarily continuous data, except for the position accuracy, which is categorical. Therefore, an efficient way to investigate the quality of the entries in this aspect of the AIS data is to Inspect if the positional message entries fall within the standard range of values. An effective way to achieve this range of value audit for the continuous fields is by invoking the pandas ".describe( )" method on each feature. This function summarizes each field's statistical overview by evaluating the central tendencies, variance, and range. On the other hand, for categorical features, it is possible to determine if the entries all fall within the standard range by getting the number of unique categories using the pandas "value_counts( )" built-in function, which returns a pandas series object, corresponding to the frequency distribution of the datapoint within that field and thereafter the maximum and minimum value can be identified.

Table 8. Range of automatically updated AIS data

Data field	Unit	Range
Longitude	[°]	± 180
Latitude	[°]	± 90

ROT	[°/m]	± 127
SOG	[kn]	0 - 102.2
COG	[°]	0 - 359.9
HDG	[°]	0 - 359
Position accuracy	[_]	True/ false

4.3.2 Manually updated AIS data quality assessment

This section highlights the method employed in assessing the quality of the manually updated aspect of the AIS data, which in this case are static and voyage-related data. They are constituted by fields such as the vessel name, IMO number, MMSI, ship dimension, call sign, ship type, destination, expected time of arrival, and navigational status. However, since the quality assessment consideration borders on the data's suitability for autonomous vessels related applications, data field related to the characteristics of the ship or information on the voyage it is embarking on, such as MMSI, callsign, IMO number, expected time of arrival were not assessed. But, data point quality checks were carried out for the ship dimension and navigational status. Furthermore, the method for validating the different navigational status and ship dimensions was proposed. Finally, I analyzed the frequency of distribution and highlighted insightful patterns in the entries.

- Navigational status

The accurate and timely navigational status update fosters maritime sector safety, especially on the high sea. It also assists vessels in ascertaining the activity and active status of an oncoming ship with which it is in proximity. It is a numeric code with possible values between 0-15, and each digit corresponds to a unique piece of motion-related information about the current activity of the ship. For instance, a navigational status of 1 signifies that the vessel is underway and using engine. Meanwhile, 15 implies that it is undefined (this is a default value indicating that the navigational status is unavailable). Since navigational status is updated manually by the crew [9], a study of the degree of correctness of the navigational data points will provide insight into how reliable the manually updated components of the AIS data are.

The percentage distribution of the different navigational statuses in the data will be

identified in exploring and validating the navigational status data field. In addition, the correlations between the navigational status and the SOG in the automatically updated field will also help to validate whether the navigational status inputs are reasonable. For example, considering a navigational status entry of 5, which connotes that a vessel is moored, the corresponding speed over ground entry is expected to be zero or approximately close. Therefore, based on the correlation between the navigational status and the corresponding SOG the navigational status entries can be validated. Presented in Table 9 are quick validation methods for navigational status data points.

Table 9. Possible validation method for selected ship navigational status

Navigational status code	Description	Validation methodology
0	Underway using engine	- SOG values should be above zero Knots, - Coordinate plots on a map should be polylines connecting the different positions.
1	Anchor	- SOG values are expected to be zero Knots. - Plots of the coordinates on a map expected to stationary points
2	Not under command	- SOG values can be zero or above zero. - ROT is fixed as vessels cannot maneuver. - Plots of the coordinates on a map may be stationary points or trajectories.
3	Restricted maneuverability	- SOG values can be zero or above zero. - ROT is fixed as vessels cannot maneuver. - Plots of the coordinates on a map may be stationary points or trajectories.
4	Constrained by her draught	- SOG can be zero or above zero. - COG is expected to be relatively fixed. - The coordinate plot on a map will be a polyline connecting the different positions.
5	Moored	- SOG values are expected to be zero Knot or more depending on the mooring circumstance. - Plots of the coordinates on a map expected to be stationary points
6	Aground	- SOG is expected to be zero. - COG and ROT fixed. - Plots of the coordinates on a map expected to stationary points

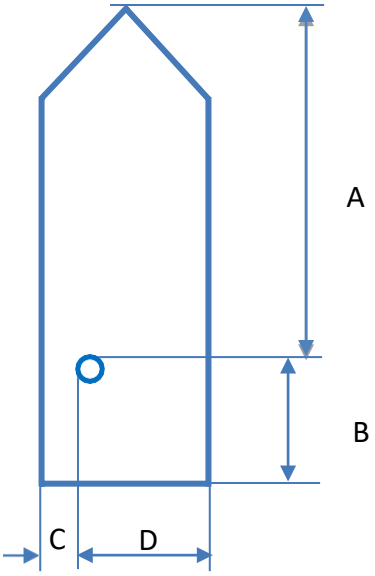
7	Engaged in Fishing	-	SOG should depict that vessels are sometimes in motion and sometimes not. - The Vessel's stationary period maybe be specific to particular geographical areas when plotted on a map.
8	Under way sailing	-	N. A
9	Reserved for future amendment of Nav. Status	-	N. A
10	Reserved for amendment of navigational status for ships carrying a hazardous substance	-	N. A
11	Reserved for future use	-	N. A
12	Reserved for future. use	-	N. A
13	Reserved for future. use	-	N. A
14	AIS-SART is active	-	N. A
15	not defined (default)	-	navigational status recorded as 15

- Ship dimension

The dimension of ships is critical in specific applications of AIS data. For instance, the quality of the result obtained by applying the projection method in the determination of collision candidates in collision estimation depends on the precision of the AIS data's position, heading information, and the reliability of the ship dimensions [26].

As observed from the static and voyage-related messages fields presented in Table 7, the dimension of vessels was given as four separate measurements with GPS antenna position as the reference point. Reference A is the distance of the GPS antenna connected to the AIS from the bow, reference B from the stern, reference C from the port, and reference D from the starboard. Three possible scenarios exist to determine the vessel dimension from the reference position given in the AIS data [9]. Firstly, if all reference measurement A-D are zero, it signifies that both the antenna reference position and the vessel dimension is unknown. Secondly, suppose the measurements

of reference A (distance from GPS antenna to bow) and reference C (distance from GPS antenna to port) are both zero. The length and beam of such vessels are identical. Therefore, the size of the ship will be the distance between the GPS antenna location to the stern or starboard. Thirdly, the GPS antenna may be placed at the port side corner of a rectangular bow, although the bow or port value must be set to one. However, this is a rare case. Figure 5 shows how the different combinations of reference measurements can be used to evaluate the size of a ship. This will be applied to the AIS data to verify whether they give the vessel accurate dimensions. Furthermore, the distribution of the length of the ships contained in the data will be plotted as the dimension of the ships is one of the major considerations in planning and allocating maritime facilities.



	Distance in meters
A	0-511; 511 = 511m or greater
B	0-511; 511 =511m or greater
C	0-63; 63 = 63m or greater
D	0-63; 63 = 63m or greater

The reference point of the reported position is not available, but the ship's dimensions are available: A = C = 0 and B = 0, and D ≠ 0.

Neither reference of the reported position nor dimensions of the ship are available: A = B = C = D = 0 (Default)

Figure 5. Ship dimension n diagram [9]

- Ship type

This thesis will investigate the accuracy and reasonability of the ship-type records because it is one of the static AIS messages, which means they are manually imputed. Ship type classification is a critical parameter in research on maritime surveillance and vessel behavior [32]. It is also a crucial factor influencing the planning and allocation of maritime resources.

Ship types are designated 2-digit codes between 0 and 99 used to represent the various types of vessels in AIS data. The first digit indicates the general category of the ship, whether it is a cargo, passenger, tanker, tugboat, or pilot ship. The second digit provides additional information about the subdivision within the initial general category.

The strategy that will be used to carry out the exploratory data analysis of the AIS data's ship types would be to compute the summary statistics of the ship-type feature. More specifically, the number of unique ship variants present in the AIS data, the frequency, and the distribution of the ship type will be estimated. The results will then be presented using suitable plots.

4.3.3 Default values

Analyzing the default values in AIS data can help identify quality issues and anomalies in AIS. Default values are placeholder entries used to represent unavailable data points. These values for essential fields in the AIS data are presented in Table 10. A function is developed to filter the AIS data and identify fields containing default values. I will further analyze the filtered data to identify patterns or anomalies in their occurrence. Furthermore, in this section, I will investigate whether their occurrence is peculiar to only components of the data that are manually updated or the automatically updated ones.

Additionally, default values in AIS data can indicate transmission inefficiency. Hence, as part of the default value investigation, the study will check whether their occurrence is peculiar to certain vessels, which may indicate faulty sensors on such vessels, or find out whether coordinates on the Sea corresponding to rows where default values are present,

is confined to a specific geographical location. A localized default value occurrence may suggest areas on the Sea that are not reliably covered by the AIS, leading to data loss.

Again, general pandas' built-in functions will be used to explore the occurrence trend mentioned above. On the other hand, a map generated with cartopy would be used to display the geographical spread of those default values. Overall, analyzing the default values can help identify areas where the accuracy and usefulness of the AIS data can be improved for a range of applications, including vessel tracking, maritime safety, and environmental monitoring.

Table 10. show the default values for AIS data fields.

AIS data field	**Default values**
Navigational status	15
Rate of turn	-128
Speed over ground	102.3
Positional accuracy	0
Longitude	181
Latitude	91
Course over ground	3600
Heading	511
Raim	0
IMO	0
Ship-type	0
ETA	0

4.3.4 Missing data

Missing data can occur in data for various reasons, such as technical failure, equipment malfunction, or human errors during data handling. It is crucial to address these missing data to avoid biases and errors in analysis. The presence of missing data can be detected with python built-in function ".isna ( ).sum( )". The function searches through the data column-wise and returns the missing data per field. The occurrence of missing values in

AIS data is often due to human error since unavailable data due to transmission problems may have been automatically replaced with default values. However, missing data can exist in static data if the information is not updated for some reason or due to an unintentional introduction during data processing.

To visualize the missing data, "missingno" is a python library that can be used for such purposes. It provides a variety of visualizations for identifying patterns in missing data, but for this thesis, the "msno.matrix ( )" function is preferred because it presents the missing data in matrix form, where each row in the matrix represents a variable in the data, and each bar corresponds to a column. In addition, white blocks in the matrix represent the missing data and its location within the dataframe.

4.3.5 Data transmission intervals

An audit of the time intervals between successive AIS messages can indicate the completeness of the data. The time intervals are analyzed by comparing them with the standard time intervals specified by IALA; this offers insight into the completeness of the data or highlights the pattern of lost messages in the data. Table 5 presents the standard time intervals for positional messages. However, for static and voyage-related messages, the regulatory body specifies a time interval of 6 minutes (360 seconds). Non-standard intervals are also commonplace in AIS data because of exceptions in position messages when the speed or course changes or in static data when any of the static data entries has just changed. In addition, breaks in the time series coinciding for both position and static data might indicate periods when the AIS system breaks down or is turned off. Also, time intervals that are a multiple of the standard ones may suggest that due to collision, messages did not update at the expected time but at a later time when time slots become available [6]. The time interval between the transmission of AIS messages will be analyzed by subtracting the timestamp of the previous AIS message transmission from that of the current one for individual vessels. I randomly selected a ship with one of the highest numbers of AIS data sent for analysis by plotting a histogram that displays the time interval distribution for the position and static data of the vessel.

5. Results and discussion

This aspect of the study presents the result of my assessment of the quality of AIS data from the Baltic Sea based on the techniques described in the methodology. It also shows insight gained from the exploration, statistical analysis, and visualization of the data. The result highlights and provides insight into the strengths and limitations of the data. Furthermore, the section will discuss the implication of the key findings for stakeholders ranging from maritime authorities, port operators, shipping companies, and researchers, contributing to advancing the knowledge on this vital subject.

5.1 Data overview

The AIS data used in this study was collected from November 23 to November 30, 2022, from the ARPA project data platform. The data returned is a pair of positional (pos.) and corresponding static and voyage-related (meta) data. The position data has 19.8 million records and 12 fields, while the static and voyage-related messages contain 107 thousand records and 14 features. The details of each variable were presented in Tables 6 and 7 for both the positional and static data. Usually, the positional messages transmission interval is between 2 – 10 seconds depending on the course or speed, but the static messages are updated every 6 minutes. Therefore, the data has more position messages than static and voyage-related messages. For example, the bar chart in Figure 7A indicates that the data's position messages are 18 times more than the static message. Hence, the significant disparity in the number of messages is due to the much lower reporting intervals of the positional messages against static and voyage-related messages.

The range of the longitude and latitude fields was evaluated to ascertain the area covered by the data. The longitudinal coordinate was between 31.42° to 16.4° while that of latitude was a range of 65.80° to 57.29°. To further validate the geographical spread of the data, all longitude and latitude coordinates of all the positional messages transmitted were plotted on a map. The trajectories of the various ships are the polylines connecting the different positions where the vessels have been over time, while the stationary positions are depicted by spots on the same map, as shown in Figure 7B. The data from the map covers parts of the Baltic Sea along Finland, Åland Island, Estonia, and Sweden.

Additionally, the frequency of transmission of positional messages is vital when developing vessel motion models [6]. The standard reporting intervals of 2, 3, 6, and 10 seconds were also evident in the data, although it would be more noticeable on plotting the difference in timestamp for a single vessel. The average time interval of zero seconds, which is not one of the standard time intervals, occurred because the data represents several vessels that have transmitted their position information concurrently, and they were also saved at the same instant in time. The predominantly zero time intervals between successive transmission is a result of plotting the time difference for all the vessels at once; hence different vessels transmitting at the same time would record a time difference of zero. Therefore to verify the presence of the IMO standard transmission intervals, the difference in time between transmissions for a one-passenger ship was plotted in Figure 8. The figure shows the standard transmission intervals of 2 seconds, 3 seconds, and 6 seconds and their multiples. Other non-standard time intervals observed maybe when the ships changed course or speed. However, the time difference of about 19 and 60 seconds appears to the outliers, which was observed to have coincided for both the positional and static messages. This is suggestive of a breakdown in the AIS, as seen in Figure 7C. Moreso, the boxplot in Figure 7C displays the average time interval of transmission for the sorted positional and static messages of the ships. There is a need to check for continuity in the timestamps of the data to be sure there are no inconsistencies or gaps in its timestamp. For this reason, I evaluated the average frequency of the positional and static messages received hourly, as presented in the histogram in Figure 7D. The figure indicates an even distribution of the messages reported hourly; since the messages received are almost even. This means there are no significant time-outs in data transmission.

Furthermore, one thousand three hundred and ninety unique vessels sent positional messages, 1229 sent static messages, and only 1227 ships sent both messages. One hundred sixty-three vessels sent only positional messages without reporting static and voyage-related data. Only two 163 ships were responsible for 94.1% of the messages sent. On the other hand, two ships sent only static and voyage-related messages without positional data, of which one ship was responsible for 99% of the total static AIS messages sent; this implies that the AIS-connected sensors are not working, or it has been deactivated. Pilot vessels, tugboats, and tankers represent 37% of ships without static and voyage-related messages. Possibly, those categories of vessels seldom send

static and voyage-related data. The pie charts in Figure 6 show the MMSI numbers of the ships that, despite sending the highest number of position or static messages, as the case may be, are missing the other message type. For instance, the pie chart to the left shows two vessels that sent the most position messages without static and voyage-related messages. The other shows the two vessels that transmitted static and voyage-related messages without position data.

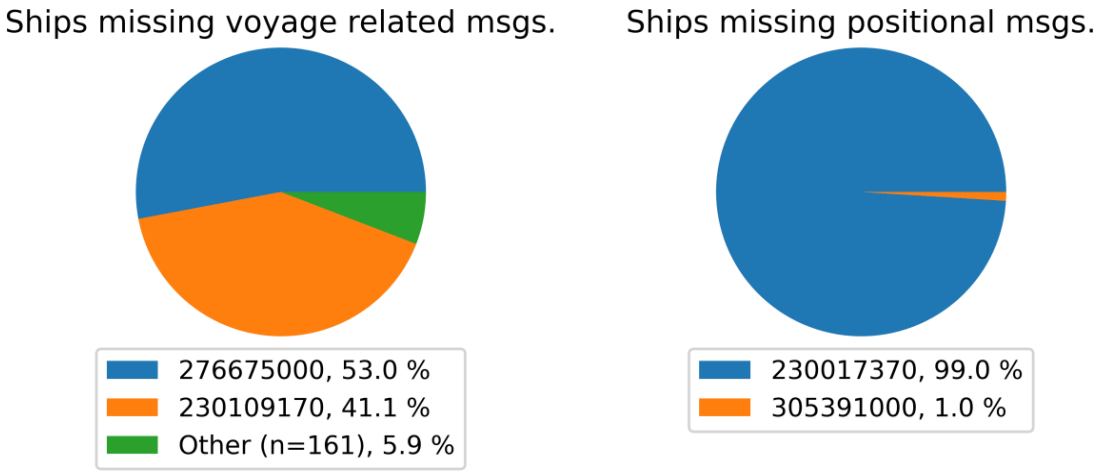


Figure 6. MMSI of ships missing positional or static and voyage-related AIS messages.

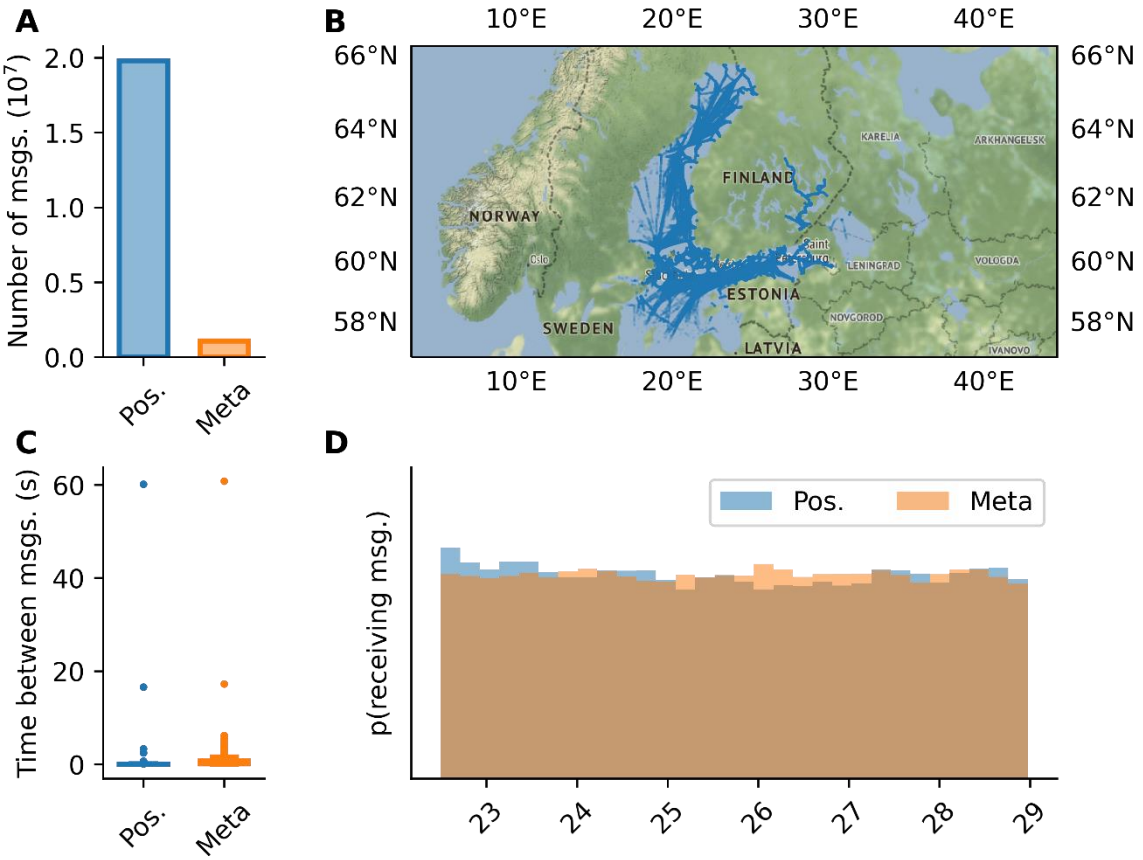


Figure 7. Data description figures

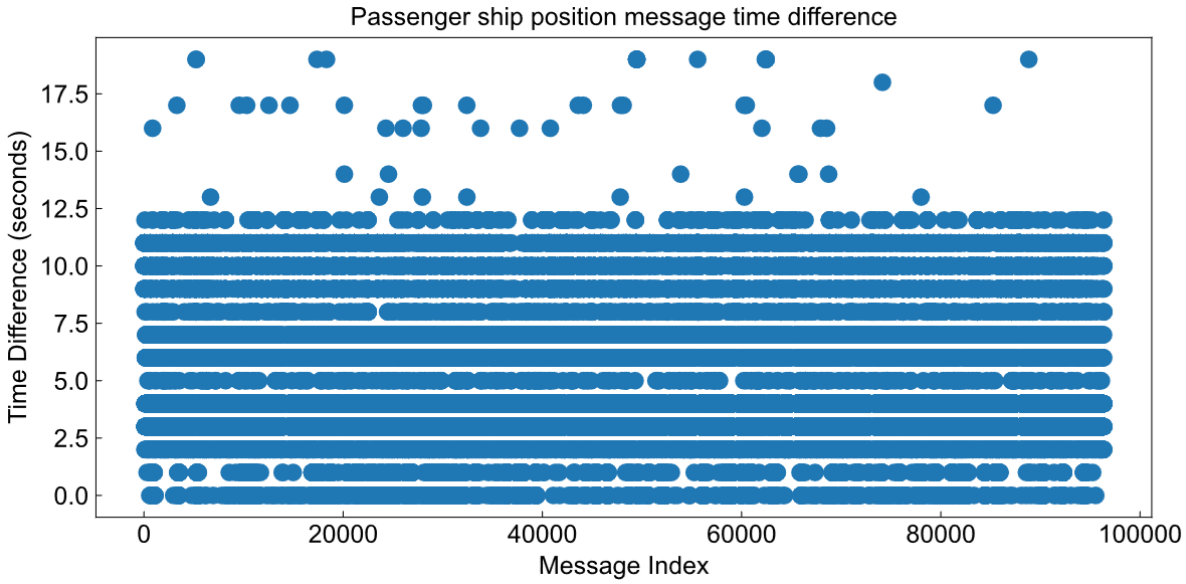


Figure 8 Position message transmission interval for a passenger ship

5.2 Exploratory analysis of automatically updated AIS data (Position messages)

The position messages are referred to as the automatically updated AIS data in this study because they are measurements read off AIS sensors without human interference. Therefore, checking the rationality of this data set gives an account of the working condition of the sensors connected to the AIS.

Table 6. highlights the 12 fields that make up the position messages. A quick way to investigate the entries is to compare them with the IALA's standard range of values presented in Table 4. To carry out the analysis, I called the ". describe( )" pandas built-in function, which returned the statistical summary of the data collection as shown in Table 11. The longitude and latitude data are all within the expected range of 180° and 90°, respectively. This is also consistent with Figure 7B, a map showing the geographical area the data covers. Similarly, the minimum and maximum entry for longitude is 16.40°and 31.42°, respectively, and 56.72° and 65.79° for latitude.

Also, from the statistical summary in Table 11, the average speed is 5.47 knots, but the maximum speed of 209 knots is suspicious. Therefore, probing further into the entries with these suspiciously high speeds, 174 vessels of 14 different ship types were observed to have reported speed values above the possible 102 knots. These even include Cargo and passenger ships. On the contrary, these categories of vessels cannot voyage at such

speed; therefore, these entries are erroneous. However, these affect only 14,898 rows out of 19 million positional messages, constituting only 0.08% of the whole position messages.

As the IALA standard specifies, the COG entries should be 0 to 359.9°, as captured in Table 11. All COG entries are within the range. The mean COG value is 188.92°, and the maximum cog reading of 360° conforms with the IMO range of values for the field. The navigational status code ranges between 0 and 15, and its statistical summary, as seen in Table 11, satisfies this standard range. The ROT entries are reported in degrees per minute, indicating the degree to which the vessel turns when the data is recorded. ROT is typically between -126 to +126; the two extreme values of -128 and +127 are default values for indicating non-available ROT data points. The statistical summary of the ROT, as seen in Table 11, means that the minimum and maximum values are the default value in both extremes. This finding is possible as the ROT compass is only mandatory for all vessels above 150 gross tonnage and all passenger ships, irrespective of size. Hence the default ROT could be for vessels without a ROT compass installed. Finally, the heading is the vessel's orientation concerning the true north. The value of this data point is typically between 0 and 359.9. However, when the heading information is unavailable, it is represented by a default value 511. Similarly, from Table 11, the minimum entry is 0, while the maximum is 511.

Table 11. Statistical summary of position message fields

Field	Count	Mean	Std	Min	25%	50%	75%	Max
longitude	1980019	23.42	2.77	16.40	21.28	23.03	25.18	31.42
latitude	1980019	60.71	1.57	56.72	59.78	60.17	60.60	65.79
sog	1980019	5.47	7.14	0.00	0.00	0.20	10.60	209.00
cog	1980019	188.92	108.35	0.00	84.9	203.20	275.8	360.00
navstat	1980019	1.52	3.96	0.00	0.00	0.00	0.00	15.00
rot	1980019	-15.88	56.03	-128	0.00	0.00	0.00	127.00
heading	1980019	213.84	145.41	0.00	90.00	198.00	296.00	511

5.3 Exploratory analysis of manually updated AIS data (static messages)

The main essence of analyzing the rationality of the manually updated static messages is that it estimates the accuracy of the operators in updating these data points. Therefore, it becomes possible to determine the reliability and trustworthiness of the information provided by the operators. However, this research will mainly consider ship dimensions, ship-types, and navigational status. The selection of the ship size for analysis is premised on the fact that it is an essential parameter in the development of anti-collision models. The navigational status entries will also be validated in this section since it is manually updated even though it is one of the positional messages.

5.3.1 Ship types

One aspect of AIS data is the identification of different types of vessels, which is essential for understanding the nature of maritime traffic. Ship type refers to the general category of vessels based on their design and operational characteristics. The IMO has over 20 established standard classifications for ship-types. In the AIS data used for this study, the ship type is a two-digit number that identifies the vessel's general category. For instance, passenger ships have ship-type numbers of 60, while it is 80 for tankers. Based on the IMO standard classification of ship types, the pie chart in Figures 9 and 10 describes the distribution of the ship types represented in the AIS data in terms of the total number of ships and the total number of positions messages sent.

Furthermore, Figure 7 presents the distribution of the ship type based on the total number of ships operating in the region. Cargo ships are the most common, constituting 38.7% of all the vessels, followed by Tankers representing 15.9%, then passenger vessels constituting 11.1%. Also, the ship-type quota with respect to the total number of positional messages sent was determined. In this regard, cargo ships sent most messages, constituting 24.8% of the overall messages; 21.2% of the messages were from passenger ships and 15.9% from pilot vessels.

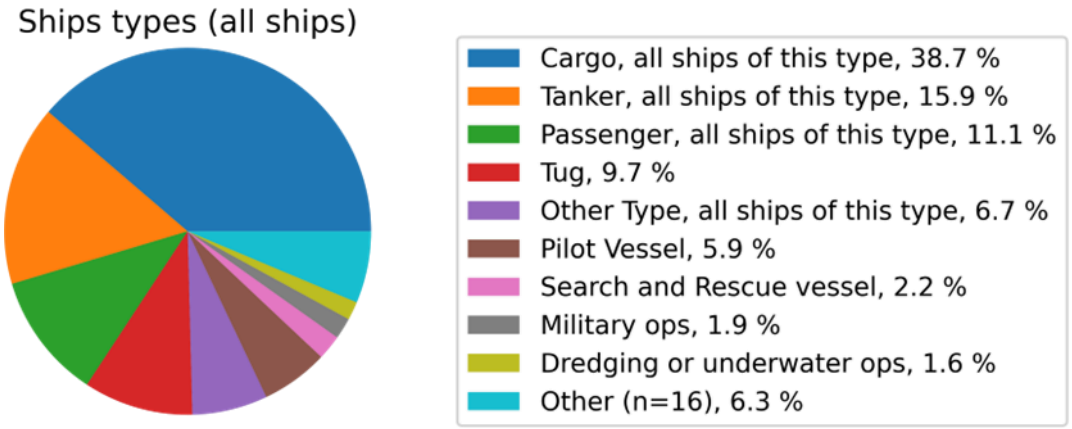


Figure 9. ship type distribution based on the total number of ships.

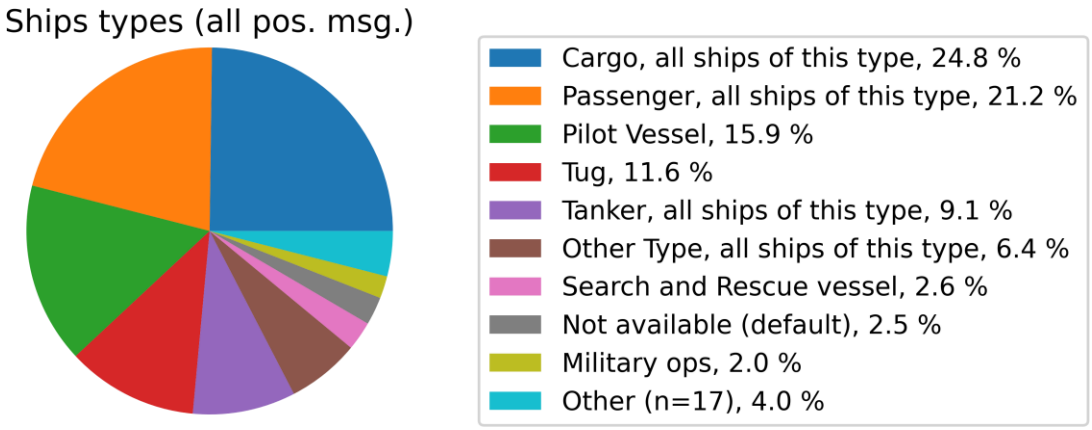


Figure 10. Percentage distribution based on the total number of positional messages.

5.3.2 Navigational status

Generally, specific parts of the static and voyage-related messages are updated manually during the voyage, others in the beginning during the commissioning of the transceivers. Conversely, positional data are read off sensors connected to the AIS. Hence, the data is updated automatically. However, navigational status is an exception because it is manually updated despite being one of the positional messages. The navigational status distribution contained in the AIS data is presented in Figure 9.

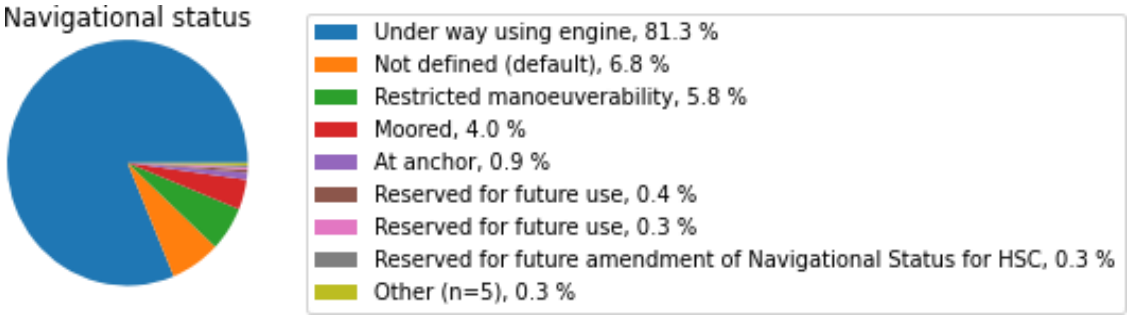


Figure 11. Navigational status distribution

Additionally, sanity checks of navigational status alongside the manually updated AIS data fields help to assess the compliance level of mariners in logging them due to the very high correlation between the navigational status and the SOG. The SOG measurement can be the basis for validating the navigational status entries. For instance, if the navigational status is zero, meaning the vessel is underway using engine, the SOG is expected to be above zero knots. In contrast, for a navigational status of 1, which implies the ship is at anchor, the SOG values are expected to be predominantly zero during this period. As shown in Figure 10, the SOG data points were divided into two categories: SOG values above 1 knot and SOG values below 1 knot for all messages with a navigational status of zero (the vessel is underway using engine). The 1 knot speed threshold was selected to give tolerance to the vessels' speed measuring device because ships might not be still at anchor. It was observed that 45.4% of the SOG values were zero for messages where the navigational status was reported as 0, which should not be. To further investigate, the position and movement of the various vessels were plotted on a map to the right also in Figure 10; the vessels' trajectories are presented as the blue polylines while stationary points are the orange spots for a navigational status of zero. Having up to 45.4% of the total messages with navigational status of 0, reporting a SOG below 1 knot or zero knot does not capture the actual dynamic status of the vessel during these periods. Therefore, the navigational status data is unreliable due to this high degree of error.

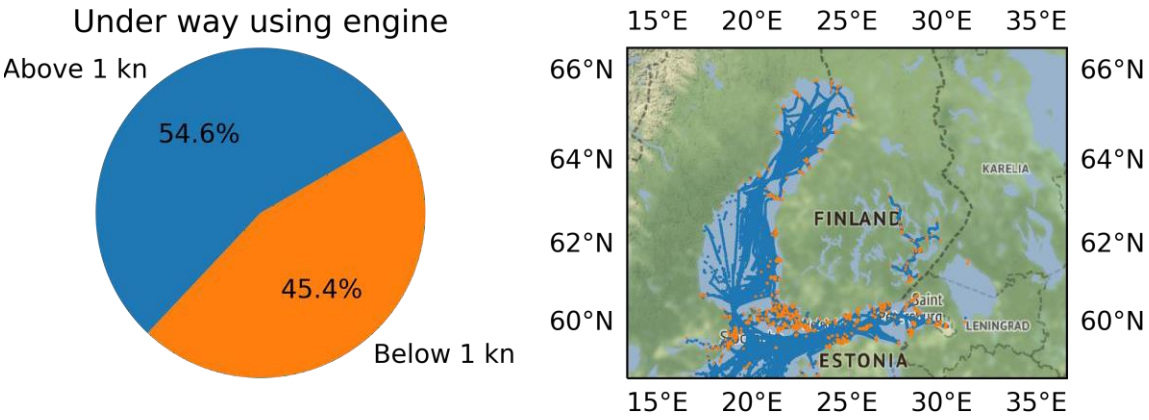


Figure 12. Navigational status accuracy check

5.3.3 Ship dimension

GPS has been the primary technology used to determine the location of vessels. The position of the GPS antenna is critical to evaluating a ship's size. The data shows ship dimensions as reference A, reference B, reference C, and reference D, where Reference A to D are distance measurements from the location of the GPS antenna. The IALA guideline for determining the ship dimension from the reference distances, as presented in Figure 5, was verified in the AIS data as follows: If the reference A to D measurement is zero for a given vessel, this indicates that those entries are unavailable since zero is the default value for ship dimension. Consequently, we cannot determine the vessel's size and reported position from the data. Of 1229 ships that sent static and voyage-related messages, 32 updated all zero-reference measurements at least once; this represents 2.6% of the vessels.

Furthermore, when the magnitude of references A and C is equal to zero, but B and D are not, for such a vessel, the reference point is unavailable; however, the ship dimension is available. From the data, only a single ship falls within this category representing just 0.08% of the total number of vessels in the data. The remaining 1196 vessels have their references position available from A to B, which implies that the data can give an accurate account of their sizes.

The addition of references A and B represents the length of a vessel, as seen in Table 5. Therefore, the size distribution of the vessels within the data was plotted and displayed in Figure 13. The size was planned as a function of the ship's length. Information about

vessel size is essential for maritime infrastructure planning and allocation.

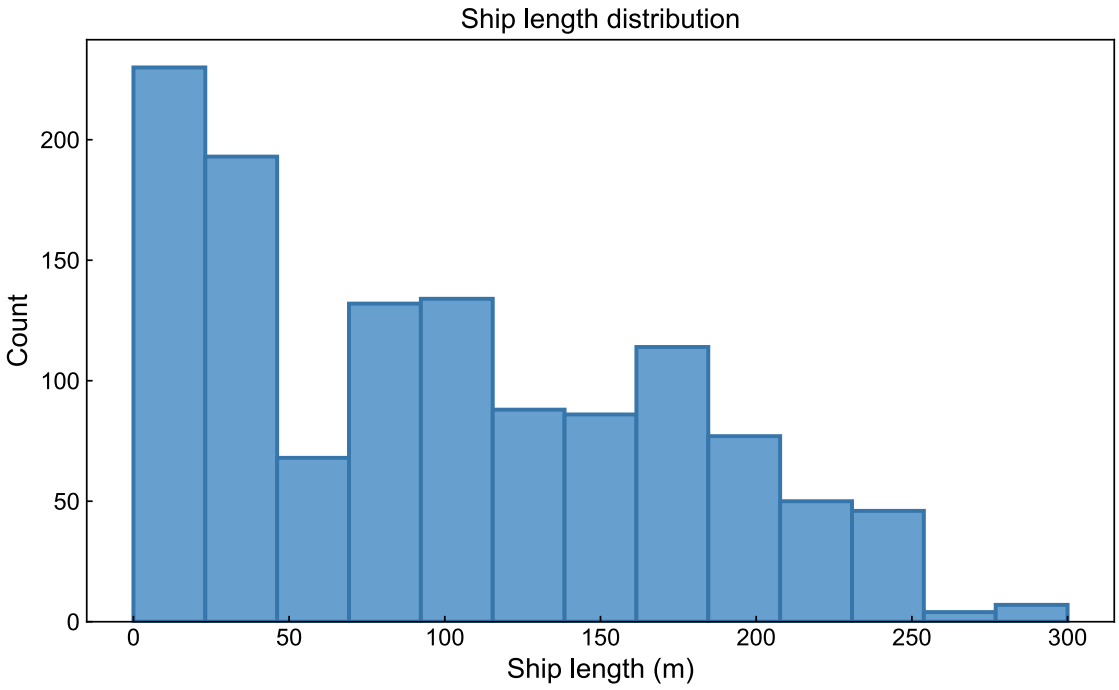


Figure 13 Distrubution of ship length for vessels in the static data

5.4 Default values

Another critical aspect of data quality analysis for AIS data is the assessment of default values. Default values can occur in various variables of the AIS dataset, especially in the position messages, since they are automatically updated. Table 10 highlights the default values for some position message fields, such as speed, heading, rate of turn, course over ground, longitude, and latitude. To evaluate the distribution of default value occurrence in the data, we leverage the existing ITU standard default value allocation for the AIS data fields presented in Table 4 by checking if they are contained in the data—pandas built-in function was employed to collect rows with those assigned default digits. I further checked if the occurrence was peculiar to certain vessels or if it cut across all ships in the AIS data. Therefore, the number of affected vessels was also determined and presented in Table 12. The table shows that the longitude and latitude fields do not have default values. The heading and COG fields have default values; however, this may not only be due to data unavailability because heading and COG are sometimes set to default when the vessel is stationary; therefore, to separate default values that are due to transmission problems, the navigational status or SOG field can be used to remove stationary ships. Table 12 shows the general distribution of the default value occurrence in some critical

AIS data fields per the number of messages and vessels affected. For instance, about 700 thousand records constituting 3.69% of the entire messages, have default COG entry, affecting 12.3% of the total vessels. The heading has a significantly high number of default values. Upon excluding rows whose navigational status indicated that the ship was stationary, 2,262,955 messages still had a default heading entry, affecting 181 vessels. The pie chart in Figure 5.7 shows the ship-type distribution of the 181 vessels with default heading datapoint, and its occurrence cuts across all the different types of ships.

Table 12. percentage default values in positional messages.

Data field	Number of messages with a default value	Percentage distribution over the total messages sent	The number of vessels affected	Percentage distribution over the number of ships
Speed over ground	12218	0.06%	171	12.3%
Course over ground	730681	3.69%	370	26.62%
Rate of turn	2443299	12.3%	264	18.99%
heading	2382270	12.03%	261	18.77%

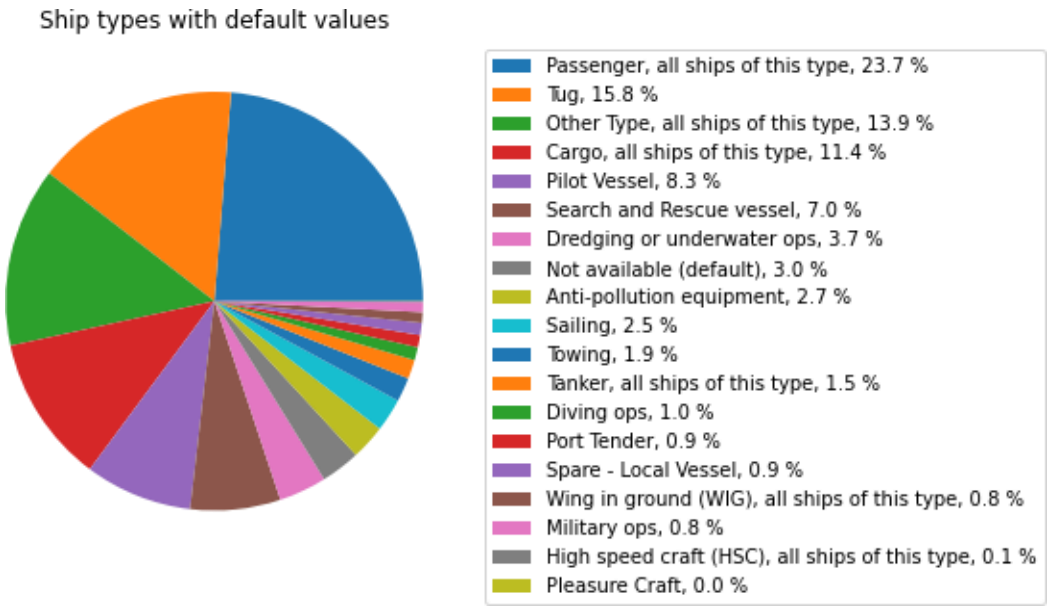


Figure 14. Ship type of vessels with default values.

5.5 Missing data

Missing data can occur for various reasons in AIS data, including technical issues with the AIS system, data transmission problems, or AIS transponders being switched off intentionally by the vessel operators. Where there are guarantees that their occurrence is not due to transmission hiccups, it may help to identify areas with poor network coverage. Moreover, depending on the missing field, it may pose a safety and security concern. For instance, missing data from vessels in a high-traffic area or during adverse weather conditions can increase collision risks or other safety incidents due to the bias introduced in situation awareness evaluation. The "isna ( )" built-in function in pandas was used for scanning missing data in the positional messages, but the position messages have no missing data. However, missing data exists in the static and voyage-related, precisely in the callsign and destination field. This affects 3269 (0.3%) of the total callsign entries and 67749 (6.28%) of the entire destination records in the static and voyage-related data.

To visualize entries missing in the AIS data, the "msno.matrix( )" function created the matrix visualization of the missing data for the static data, as seen in Figure 12. The plot shows the proportion of the missing data in each field. The white blocks in the call sign and destination bars indicate the missing entries' index-wise location.

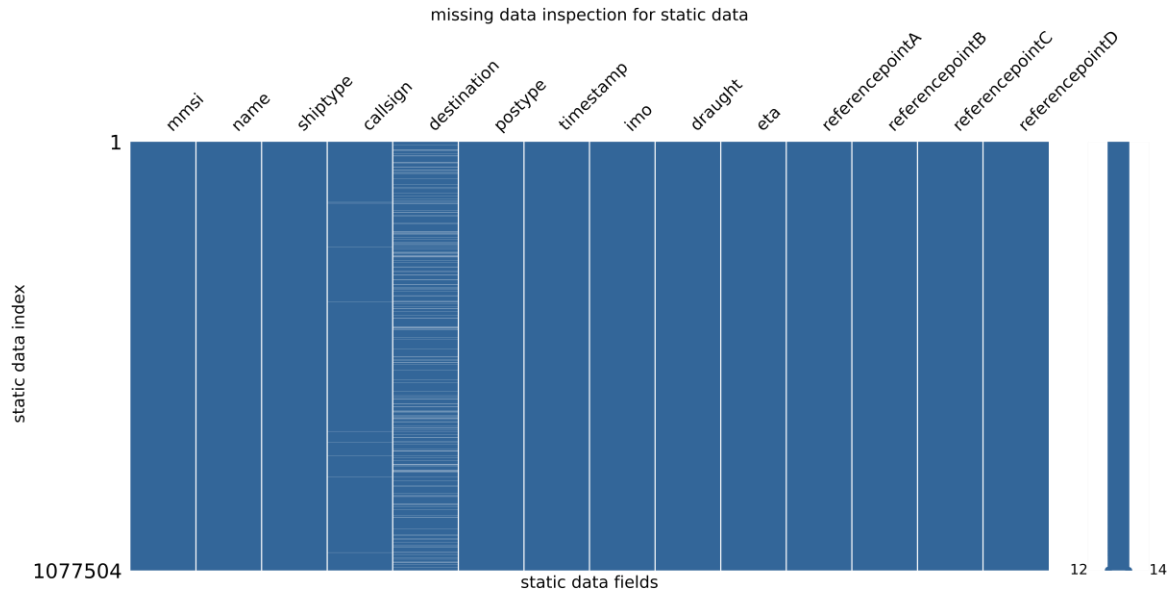


Figure 15. Bar chart showing the missing data index location for static data

6. Limitations and Future Research

This study was conducted using relatively small-sized data. The data covers a period of one week for ease of processing. However, more extensive data collected over a much more extended period may provide deeper insights since statistical analysis tends to be more accurate with a larger sample size. Additionally, a more extensive data size offers better representation as it would cover a larger range of scenarios and events, thereby increasing confidence in the analysis.

Secondly, the AIS data used in this study was pre-processed, as stated by the third-party AIS data service provider (Digitraffic). Consequent to this pre-preprocessing and the fact that adequate information was not provided on the exact pre-processing interventions carried out on the data, the analysis may have been impacted due to the introduction of unintended bias or information loss. Future research may be carried out with raw AIS data for improved confidence in the analysis and the possibility of more significant findings.

7. Conclusion

This research analyzed the quality of AIS data using exploratory data analysis and visualization to examine the trustworthiness of the data from vessels operating across the Baltic Sea. The study is critical because factors such as the quality of sensors, transmission efficiency, environmental effects, and data processing procedures can affect the data's accuracy, completeness, consistency, and timeliness. Therefore, determining a clear and effective way of identifying and dealing with errors in the data is crucial. Furthermore, it is essential because the AIS data continues to be a significant enabler of innovation and efficiency in maritime operations and decision-making.

This study identified quality issues that may pose challenges regarding its fit-for-use in developing AI-based applications in the maritime sector. This concept helped to inform the overall picture for the analysis since the quality analysis of AIS data is subjective. Hence, its quality investigation approach would often depend on the intended use of the

data. Although the quality analysis of AIS data has received appreciable research attention over the years notwithstanding, analysis specific to the Baltic Sea area is still lacking. Additionally, quality assessment and assurance is a continuous process, more so that newer studies may highlight improvements in the data or reveal quality gaps that need to be addressed.

Another study approach employed was to analyze the AIS data from the point of view of the manually or automatically updated data component. This methodology helps to critically assess the performance of the sensors reporting the positional messages, which were categorized as the automatically updated component of the data in this study. Furthermore, this dissertation offers insight into the level of precision with which the vessel operation stakeholders manually updated static and voyage-related data. The research revealed the presence of missing data in the callsign and destination field of the voyage-related data, affecting 11 and 111 ships, respectively. Also, the navigational status has up to 45.4% incorrect inputs for a navigational status of zero when the vessel is underway using an engine. Additionally, the SOG contains out-of-range speed values above the maximum range of 102 knots up to a maximum of 206 knots which is unreasonable, mainly because the ships affected had cargo and passenger ships in the mix, and at least the fastest cargo ship has its maximum navigation speed to be around 37 Knots. Finally, the investigation of default values in the dataset showed they exist predominantly in the positional messages, with the heading and ROT recording the highest occurrence at 12.03% and 12.3% of the total messages.

In addition, the results emphasize the importance of visualization and data analysis techniques in identifying and addressing quality issues in AIS data. By leveraging these techniques, it was possible to gain deeper insights into the underlying patterns and trends of the AIS data and detect anomalies, outliers, and errors, thereby facilitating decision-making processes based on accurate and reliable information. It also highlights the need to continuously monitor and improve AIS data quality by implementing quality assurance and control measures. Such measures could include using advanced sensors, regular maintenance, equipment calibration, and the continuous comparison and study of subsequent assessments.

In conclusion, the quality of AIS data is a critical aspect of maritime operations, and the

findings of this thesis provide valuable insights into quality issues in AIS from the Baltic Sea area and the factor that influences data quality. It can also inform the development of guidelines for assessing the quality of AIS data, which can ultimately contribute to maritime activities' safety, efficiency, and sustainability.

REFERENCES

[1] H. M. Perez, R. Chang, R. Billings, and T. L. Kosub, "Automatic Identification Systems (AIS) Data Use in Marine Vessel Emission Estimation".

[2] F. Mehdipour, M. F. Rahman, and K. J. Murakami, "Intelligent wireless sensor networks (iWSNs) in cyber-physical systems," in *Cyber-Physical System Design with Sensor Networking Technologies*, Institution of Engineering and Technology, 2016, pp. 219–237. doi: 10.1049/PBCE096E_ch10.

[3] "Ship Safety Standards - Maritime Autonomous Surface Ships (MASS) - EMSA – European Maritime Safety Agency." https://www.emsa.europa.eu/mass.html (accessed September 26, 2022).

[4] D. Yang, L. Wu, S. Wang, H. Jia, and K. X. Li, "How big data enriches maritime research–a critical review of Automatic Identification System (AIS) data applications," *Transp Rev*, vol. 39, no. 6,pp. 755–773, Nov. 2019, doi:10.1080/01441647.2019.1649315.

[5] C. Iphar, A. Napoli, and C. Ray, "Data quality assessment for maritime situation awareness," in *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Copernicus GmbH, Aug. 2015, pp. 291–296. doi: 10.5194/isprsannals-II-3-W5-291-2015.

[6] P. Last, C. Bahlke, M. Hering-Bertram, and L. Linsen, "Comprehensive Analysis of Automatic Identification System (AIS) Data in Regard to Vessel Movement Prediction," 2014, doi: 10.1017/S0373463314000253.

[7] A. Harati-Mokhtari, A. Wall, P. Brooks, and J. Wang, "Automatic Identification System (AIS): Data Reliability and Human Error Implications," *The Journal of Navigation*, vol. 60, no. 3, pp. 373–389, Sep. 2007, doi: 10.1017/S0373463307004298.

[8] F. Natale, M. Gibin, A. Alessandrini, M. Vespe, and A. Paulrud, "Mapping fishing effort through AIS data," *PLoS One*, vol. 10, no. 6, Jun. 2015, doi: 10.1371/journal.pone.0130746.

[9] IALA-AISM, "universal-automatic-identification-ais-volume-1-part-1-operational-issues-1028," France: Saint Germain en Laye. Accessed: April 12, 2023. [Online]. Available: https://www.e-navigation.nl/sites/default/files/universal-automatic-identification-ais-volume-1-part-1- operational-issues-1028.pdf

[10] MarineTraffic, "Understanding AIS: Terrestrial vs Satellite AIS Tracking." Accessed: October 27, 2022. [Online]. Available:

https://business.marinetraffic.com/resources/understanding- satellite-ais-tracking

[11] M. A. Cervera, A. Ginesi, and K. Eckstein, "Satellite-based vessel Automatic Identification System: A feasibility and performance analysis," *International Journal of Satellite Communications and Networking*, vol. 29, no. 2, pp. 117–142, Mar. 2011, doi: 10.1002/SAT.957.

[12] O. Cherrak, H. Ghennioui, N. T. Moreau, and E. H. Abarkan, "Blind separation of complex- valued satellite-AIS data for marine surveillance: A spatial quadratic time-frequency domain approach," *International Journal of Electrical and Computer Engineering*, vol. 9, no. 3, pp. 1732–1741, Jun. 2019, doi: 10.11591/ijece.v9i3.pp1732-1741.

[13] "Improved satellite detection of AIS M Series Mobile, radiodetermination, amateur and related satellites services," 2009, Accessed: October 06, 2022. [Online]. Available: http://www.itu.int/ITU-R/go/patents/en

[14] Y. Z. and L. F. M. Yang, "Collision and detection performance with three overlap signal collisions in space-based ais reception," *Trust, Security and Privacy in Computing and Communications (Trust Com), 2012 IEEE 11th International Conference* , pp. 1641–1648, Jun. 2012.

[15] "AIS TDMA Access schemes," *All About AIS*, 2012. http://www.allaboutais.com/jdownloads/Access%20schemes%20technical%20downloads/ais_ tdma_access_schemes.pdf (accessed April 12, 2023).

[16] "AIS (Automatic Identification System) Overview - Shine Micro." https://www.shinemicro.com/ais-overview/ (accessed September 23, 2022).

[17] "Class A And Class B Automatic Identification System (AIS) – Ocean Time Marine." https://oceantimemarine.com/class-a-and-class-b-automatic-identification-system-ais/ (accessed February 09, 2023).

[18] "Technical characteristics for an automatic identification system using time-division multiple access in the VHF maritime mobile band M Series Mobile, radiodetermination, amateur and related satellite services," 2010. [Online]. Available: http://www.itu.int/ITU-R/go/patents/en

[19] "The quality of your AIS data feed: an evaluation framework - Spire Maritime." https://spire.com/blog/maritime/the-quality-of-your-ais-data-feed/(accessed April 07, 2023).

[20] "National Marine Electronics Association NMEA 0183 Standard for Interfacing Marine Electronic Devices COPYRIGHT© NMEA 2002 NMEA 0183-Standard For Interfacing Marine Electronic Devices NMEA 0183 Version," 2002.

[21] L. L. Pipino, Y. W. Lee, R. Y. Wang, and R. Y. Yang, “Data Quality Assessment,” vol. 45, no. 4ve, Apr. 2002, Accessed: April 12, 2023. [Online]. Available: https://dl.acm.org/doi/abs/10.1145/505248.506010

[22] R. Y. Wang, "Quality Ma A Product Perspective on Total Data," *Commun ACM*, vol. 41, no. 2, 1998.

[23] L. P. English, "Improving data warehouse and business information quality : methods for reducing costs and increasing profits," p. 518, 1999, Accessed: October 09, 2022. [Online]. Available: https://www.wiley.com/en-us/Improving+Data+Warehouse+and+Business+Information+Quality%3A+Methods+for+Reduci ng+Costs+and+Increasing+Profits-p-9780471253839

[24] Silvia Valcheva, "The key data quality metric you need for DQA," *intellspot*. https://www.intellspot.com/data-quality-metrics/ (accessed April 12, 2023).

[25] E. M. Foster and G. Y. Fang, "Alternative Methods for Handling Attrition: An Illustration Using Data From the Fast Track Evaluation", doi: 10.1177/0193841X04264662.

[26] P. Silveira, A. P. Teixeira, and C. G. Soares, "Assessment of ship collision estimation methods using AIS data," in *Maritime Technology and Engineering - Proceedings of MARTECH 2014: 2nd International Conference on Maritime Technology and Engineering,* CRC Press/Balkema 2015, pp. 195-204 doi: 10.1021/b17494-27.

[27] K.' Sean, P.' Andreas, H.' Joseph, and H.' Jeffrey, "Proceedings of the SIGCHI Conference on Human Factors in Computing Systems Wrangler: interactive visual specification of data transformation scripts," *Association for Computing Machinery*, pp. 3363–3372, May 2011.

[28] H. E. Brady, "PL22CH17_Brady ARjats.cls The Challenge of Big Data and Data Science," 2019, doi: 10.1146/annurev-polisci-090216.

[29] M. Yang, "BIG DATA: ISSUES, CHALLENGES, TOOLS," M.S. thesis, Centria university of applied sciences, Kokkola, 2020.

[30] J. Karbowska-Chilinska, J. Koszelew, K. Ostrowski, P. Kuczynski, E. Kulbiej, and P. Wolejsza, "Beam search algorithm for ship anti-collision trajectory planning," *Sensors (Switzerland)*, vol. 19, no. 24, Dec. 2019, doi: 10.3390/s19245338.

[31] P. C. , L. Y. P. , & C. C. Y. Hsiao, "An analysis of the navigational status in collision avoidance. Journal of Navigation," *Journal of Navigation*, vol. 66, no. 1, pp. 33–45, 2013.

[32] Z. Yan, X. Song, H. Zhong, L. Yang, and Y. Wang, "Ship Classification and Anomaly Detection Based on Spaceborne AIS Data Considering Behavior Characteristics," *Sensors (Basel)*, vol. 22, no. 20, Oct. 2022, doi: 10.3390/s22207713.