Detecting of Maritime Anomalies and Security Issues Using AIS Data

UNIVERSITY OF TURKU
Department of Future Technologies
Master of Science in Technology Thesis
December 2019
Pradip Neupane

Supervisors:
Petra Virjonen
Paavo Nevalainen

UNIVERSITY OF TURKU
Department of Future Technologies

Pradip Neupane: Detecting of Maritime Anomalies and Security Issues Using AIS Data

Master of Science in Technology Thesis (55 pages)
December 2019

The maritime environment has suffered from high marine traffic. High marine traffic has caused many problems and risks in the maritime and its surrounding environments. The primary objective of this thesis is to analyze historical AIS data and do research on finding the behavior of vessels that are operated in the Baltic Sea. The thesis is focused on detecting anomalies from the ships in three main topics: missing AIS data, speed variation in vessels, and the encounters between different vessels.

The experiments on finding anomalies in ships were done by using the real AIS data. The AIS data was extracted between 4 -16 June 2019 from "The Finnish Transport Infrastructure Agency". To make experiments more specific to location-wise, a proper measurement boundary was created in the Baltic Sea. The geometry size of the measurement area was created in such a way that no major ports or islands were within that area.

From the experiments, it was found that approximately 8% of the total ships operated in the Baltic sea had missing AIS data at some stages. It was also found that ships' speed frequently changes while moving. The practical part of the thesis indicates that there were some anomalies presented in the ships. Three main possible reasons for having these types of odd behaviors shown in vessels are unreliability of AIS data, technical issues, and intentional causes. Further research can be extended in this field by addressing the limitation and the problems that are mentioned in this thesis.

Keywords: AIS data, marine, outliers, data analysis, visualization, data processing

Table of Contents

Abbreviations and Acronyms

AIS	Autonomous Identification System
IMO	International Maritime Organization
SOLAS	International Convention for the Safety of Life at Sea
MMSI	Maritime Mobile Service Identity
OOW	Officer of the Watch
COG	Course Over Ground
SOG	Speed Over Ground
ROT	Rate of Turn
MQTT	Message Queuing Telemetry Transport
GPS	Global Positioning System
BIIT	Built-in Integrity Test
LOA	Length Overall
CC	Creative Commons
API	Application Programming Interface

1 Introduction

Today, data is collected in many forms, e.g. customer feedbacks, sensor equipment, buying and selling of items in a supermarket, credit card purchases, vehicles' motions, etc. One of the common issues in data is that some of the recorded data in a dataset may not confirm the expected patterns like other data present in a dataset. An outlier is a data point that separates from other data points by appearing in an abnormal position. Before signaled out some data points as outliers, it is necessary to distinguish what are the standard data points [1]. For example, spending 200 dollars per day in a holiday season may look ordinary, but it might not look normal if that occurs outside of holiday seasons. Detecting anomaly present in a dataset is crucial because it may vastly change the overall outcome of the data insights [2].

Marine traffic has been increased dramatically and significantly in recent years. Because of high traffic in the maritime environment, some of the vessels may show unexpected behaviors in terms of speed, changing routes, and courses. These types of unexpected behaviors demonstrated by the ships are called anomalies. A lot of research and technologies are used to detect anomalies. However, one of the common problems in detecting anomalies issues in the maritime environment is that these researches also may contain high false alarms [3].

Outliers in a dataset may lead to some severe problems like errors, and structural defects. For example, an unusual traffic pattern in a computer means it may have been hacked, and it may send sensitive information stored in that computer to the criminals, anomalies transaction in a credit card means credit card may have been stolen, and it could likely be used in buying illegal activities [4]. Anomaly detection can help in minimizing fraud, money laundering, and it can provide instant notifications to prevent further damages. This kind of detection technology can be beneficial and useful in sensitive industries and organizations such as banking, hospitals, science facilities, transport, etc.

The history of detecting anomaly goes to as early to the 19th century, and since then, it has been studied and developed further [4]. Since many of the technologies have been developed to detect abnormal behaviors in data. Some technologies are developed for specific applications, for example, detecting unusual patterns in internet traffics, voltage anomalies in IoT devices, credit card fraud detection, finding anomalies in medical data, etc. [4][5]. Whereas, some of the technologies developed for abnormal detections are more generic, and it can be applied in different areas. Since, last few years, machine learning algorithms have been used in finding outliers in a dataset. The machine learning algorithms can help to detect anomalies in a dataset with minimum efforts. However, not all data in a dataset is appropriately labeled and structured. It needs some data processing and data cleaning steps before applying machine learning algorithms. In a supervised machine learning algorithm, it requires label data. Whereas, an unsupervised machine learning algorithm uses unlabeled data. Some of the standard machine learning algorithms that are used in anomaly detection are mentioned below [4].

- **Logistic regression:** Logistic regression is a predictive analysis that is commonly used to show a connection between one dependent variable (binary) to one or more ratio-level independent variables.
- **K-Nearest Neighbors:** K-nearest neighbors (KNN) is a machine learning algorithm that is used for regression and classification problems. KNN algorithm works by finding the K closest data points. For classification in KNN, the predicted output is the mode of the labels of the K nearest neighbors. Whereas, in regression, it can be mean or median or some fit of the k-nearest data points.
- **The Bayesian Network:** Bayesian network is a statistical model that uses the probabilistic graphical model from a set of variables to their conditional dependencies. These types of algorithms are popular in doing anomaly detection, prediction tasks, time-series analysis, etc.

A successfully implemented machine learning algorithm helps in instant flagging and detecting unexpected behavior in a dataset. However, the performance and detection abilities of machine learning algorithms depend on the skewness of a

training dataset. Skewness is the measure of an asymmetrical distribution of values in a bell curve where the majority of values are shifted either towards the right or left [6]. If the skewness is higher in a training dataset, it will decrease the overall performance on detecting anomaly of algorithms [7].

1.1 Background

Today, harbor and sea are becoming busier because a large number of vessels are operating in the sea with different speeds, shapes, and sizes. Between 1992 to 2012, there were dramatic increases in cargo ships in the Indian Ocean and the Chinese Sea. According to the Journal "Geophysical Research Letters," published in October 2014, the average ships' traffic has increased to 300% in the Arabian Sea and the Indian Ocean [8]. In the Baltic Sea area alone, marine traffic had increased by 19.16 % from 2000 to 2018 [9]. The impact of these crowded traffics of marine vessels is significant, and it is causing an environmental imbalance in maritime environment. The sea routes are becoming congested, and the risk of getting collisions with each other has increased significantly [10]. For the safety of maritime, the International Maritime Organization (IMO) has set strict guidelines since 2000. According to IMO guidelines, all vessels over 300 tons of weights must fit the Automatic Identification System (AIS) devices in their vessels [11]. These AIS systems transmit real-time data related to its vessels, including information like Vessel's Maritime Mobile Service Identity (MMSI) number, its current position, its speed, etc. The detail about the IMO guidelines in the AIS device is discussed in chapter 2 of this thesis.

The unusual behaviors of the marine vessels are no longer a surprising result. It is hard to decide why the vessels are showing unusual activities without doing a proper assessment. The maritime environment is a sensitive area, and if any unusual behavior of a vessel is discovered, then it needs to be further analyzed because it could be a security threat to all surrounding marine environments. Some of the problems in marine vessels are collisions, striking to shore or island or big rocks, criminal activities like smuggling, piracy, low visibilities, hijacking, spoofing, etc. According to the report published by the International Maritime

Bureau, the total numbers of attacks in marine vessels were 201 in 2018 [12]. The nature of these attacks was related to piracy, hijacking, and armed robbery. The majority of these attacks occurred in Nigeria, Indonesia, Malaysia, Ghana, Bangladesh, etc. The hijacking, piracy and armed robbery are less common in the European water. In 2009, there was a report of the suspected hijacking of a marine vessel in the Baltic Sea [13]. AIS data can help to mitigate many of these problems effectively.

One of the simple examples is that with the help of AIS data, other nearby vessels can be warned in case of any danger. However, there have been problems that some of these vessels are faking and transmitting AIS data, particularly for illegal activities. These illegal activities are increasing mainly in coastal waters and shore regions [14]. So, it is essential for a coastal guard, military personnel, etc. to perform special surveillance on the shores regions and coastal water. With the help of such surveillance and proper threat assessment, AIS data can protect the maritime environment from many issues.

There have been done several researches to mitigate some of the problems mentioned above. Lane et al. [14] have presented overviews about how to detect and tackle anomaly behavior in vessels. In the paper, they were attempting to identify the movement of the ships and how well the ships follow their designated routes [14]. If a vessel follows a different route or shows unusual behavior unexpectedly, then it is abnormal for that vessel. Further analysis can be carried out whether such a vessel is a security risk for its surroundings. Shen et al. [15] have illustrated on preventing the collision between multiple ships using a deep Q-learning algorithm. Higher traffic in marine is increasing the risk of getting a collision between ships. Even though a vessel is fitted with radars, technologies, AIS devices, etc. there is still a possibility of collision because of human errors and technical issues. In the paper, they have mentioned that around 89 -96 % of the collision types occurred in the sea area are caused by human errors [15]. With the help of AIS data, it is possible to make an automated navigational system. The primary purposes of these navigational systems are to provide directions, to give

surrounding information, and to notify with a warning within that area [15]. However, one of the critical things that need to be understood is that AIS data cannot replace the existing navigational features of the vessels. It can only provide and act as a support to navigational systems.

1.2 Research Goals and Objective of Thesis

There have been numerous researches in anomaly detection. Some of the examples are credit card fraud detection, collision detection in marine, unusual behavior detection in an airplane, etc. Many of these researches are focused on mitigating the issues which are faced by industries. In this thesis, it aims to analyze AIS data and do research on finding the behaviors of vessels by analyzing historical data. For the thesis purpose, all the marine data were taken only from class A marine vessels from the Baltic Sea. The data used in the thesis is collected from "The Finnish Transport Infrastructure Agency" [16]. The data used in this thesis is real-time data of the vessels operated in the Baltic sea area.

A small phone interview was taken with Vice Admiral Isto Mattila from the Finnish Coastal Guard service to gather information related to the problems and issues in the maritime environment. Some of the security issues that may occur in the vessels are mentioned below [17].

- By assigning valid static information like the vessel's name, flag, type of ships, ship's dimension, etc. to the non-authorized ship, such ship may gain access as authorized ship to move in the restricted water. This process is called ships spoofing. By using this type of attack, the unauthorized vessel can move freely in the water and can cause serious security risks.
- It is possible to manipulate AIS data by faking the information before the officials capture and analyzed it. There are black boxes that can be bought with the ready factory settings and which allow faking the position. This faking information can be used to lure targeted vessels in doing wrong maneuvers.
- By faking positions, two different types of vessels, e.g. fishing boat and a regular cargo ship meet to swap the illegal materials.

The interview with Mr. Mattila was a starting point, and it helped to broaden the research areas in the following topics. Below are the questions that this thesis is attempting to answer.

1. **Missing AIS Data:** Some of the vessels have missing AIS data and it is inconsistent and random. Why there is missing AIS data? How often these missing data are caused by technical reasons and intentional reasons?
2. **Speed Variation in Vessels:** Marine vessels have inconsistent speed while moving in the sea. In some cases, the speed of the marine vessels is even close to 0 m/s in the middle of the sea. Is the inconsistent speed variation in vessels caused by technical reasons or intentional reasons?
3. **Vessels Encounters**: How often two different ships encounter each other in the middle of the sea when both of their speed is close to 0.01 m/s? Are there any abnormalities in these encounters?

1.3 Structure of the Thesis

This thesis is focused on researching of detecting anomalies from the ships in the Baltic sea. For the research purpose, AIS data was extracted from "The Finnish Transport Infrastructure Agency". The data was collected using a web socket API by using Google Cloud. The thesis is divided into five main chapters. The introduction chapter is focused on providing the overall idea about the thesis goals and objectives. This chapter also explains background information and literature overviews regarding some scientific works that have been done in the field of anomalies detection.

Chapter 2 focuses on the AIS and its guidelines and data processing steps. This chapter also provides a detailed overview of how AIS data was extracted from the ships in the Baltic area, what are the techniques that were used to extract AIS data, and what were the data processing steps. At the end of this chapter, it also shows some basic overview of exploratory analysis of extracted AIS data, the spread area of the data, and its reliability. Chapter 3 and 4 are focused on the practical part of the thesis. Chapter 3 explains the detailed overview of the implementation of

research and explains the methods that were used for detecting anomalies and security issues in ships by using the AIS data. Chapter 3 also explains briefly about tools and technologies that are used in finding the anomaly patterns if they existed from the ships' data. Chapter 4 is focused on evaluating the results from chapter 3. It attempts to explain results from all possible aspects of the ships and the AIS data.

Chapter 5 is the conclusion of the thesis. This chapter attempts to summary of the critical findings of the results of this thesis. This chapter will also include some of the essential points and overview of the thesis and tries to explain if the initial research questions were answered. It also includes some of the problems and challenges that were faced in this thesis. Finally, it will provide suggestions that could be studied and researched in the future in this field.

2 AIS Data, Extraction and Processing

This chapter is subdivided into what is AIS data, and what are guidelines related to AIS data, how AIS data was collected and processed, reliability of AIS data, and basic AIS data visualizations. At the end of this chapter, it will explain briefly related to the feature extraction process in the AIS data.

2.1 AIS and Its Guidelines

This sub-topic is focused on three main areas: introduction to the AIS data, what are its guidelines, and the nature of data transmitted. At the end of this sub-topic, there will be a short introduction related to ship and its dimensions.

2.1.1 What is AIS

AIS is a fully automated system that tracks locations, positions, and related information of marine vessels. AIS was initially developed by the International Maritime Organization (IMO) as a standard technique to avoid collisions between the vessels. However, this technology was quickly improved, and it was used in the rescue operations and other areas of maritime. Since December 2004, all passengers, as well as commercial vessels, were required to have AIS technology fitted in their vessels. However, it is vital to understand that some marine vessels may not be fitted with AIS technology. Fishing boats, warships, leisure ships, etc. are not required to have this technology and officer of the watch (OOW) must be aware of these types of vessels. Since 2014, the European Commission (EU) made strict guidelines that all fishing boats of length above 15 meters should have AIS devices. From January 2014 to August 2014, around 75% of the fishing vessels operated in the EU were fitted with AIS devices [18].

2.1.2 Operational Aspects of AIS

The primary objectives of AIS data are to enhance the safety of life in the sea area by increasing navigational efficiency and to protect the maritime environment. Besides these, AIS helps to identify marine vessels, exchange information between them, assists in rescue operations, identify the surrounding situations, etc. AIS transponder units broadcast static and dynamic data automatically related to

vessels. These AIS transponder units can communicate with different vessels and can exchange information related to each other situations automatically. This process will benefit the mariner to know about the current situations of their routes and can make appropriate decisions within a reasonable time in case of sudden disturbances in the area.

AIS technology transmits two types of data: static and dynamic. It is discussed in detail in an upcoming sub-chapter. Static information contains ship MMSI (Maritime Mobile Service Identity), IMO number, ships length and width, type of ship, location of positioning-fixing antenna. Whereas, dynamic AIS data contains the ship's locations and its positional time. The information related to AIS data is shown in Universal Time Coordinated (UTC). Vessel information like destination, draught, estimated time of arrival, type of cargo, short-safety messages, etc. are categorized into specific voyage related data. These voyages related information is entered manually by the ship's captain or responsible authorized person. AIS system does not have built-in features to validate the transmitted data. It is responsible for a coastal officer to check and validate the accuracy of the transmitted data regularly.

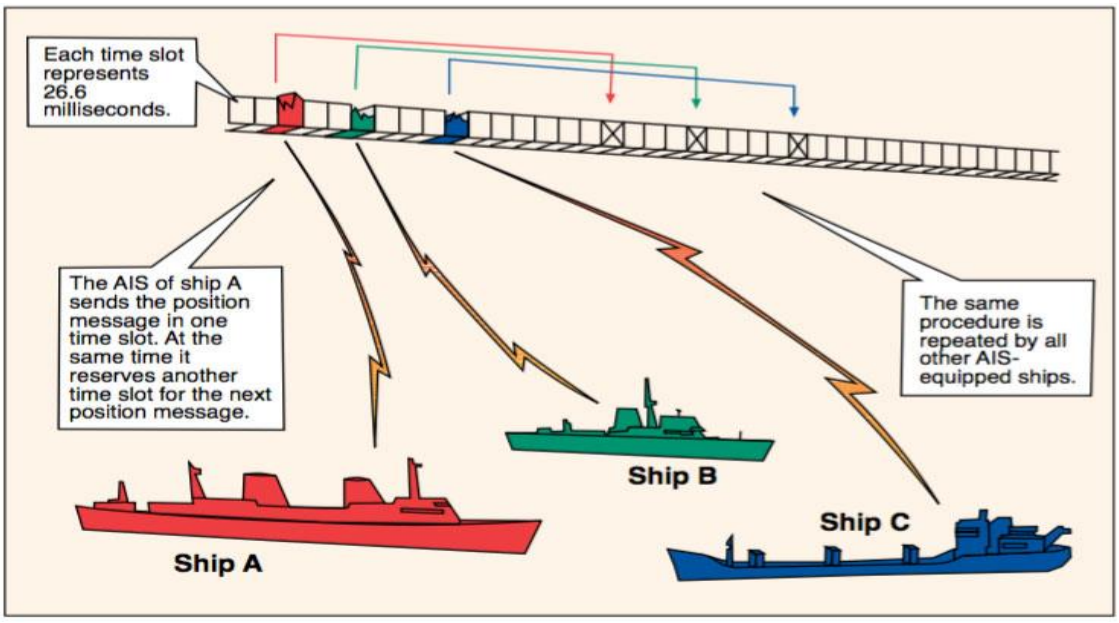


Fig 1: An example of AIS System Overview. Public domain [19]

2.1.3 Guidelines and Usages of AIS Data

It is essential to understand proper AIS guidelines, and users should interpret these guidelines before using this technology. According to the IMO, the following are some of the guidelines and regulations related to AIS technology [20].

1. Every marine vessel of over 300 tonnages is compelled to use AIS fitted system. The exceptional vessels may include fishing boats, governmental ships, warships, private leisure ships, etc. which are not compelled to use this system
2. The period of transmission of data depends on the speed of the vessels. For example, if the speed is between 0-10 knots, then it must transmit its data every 10 seconds. Whereas, if the speed is over 23 knots, then it must transmit in every 2 seconds.
3. There are two types of AIS equipment available, Class A and Class B. Class A equipment should be relevant and fully designed according to IMO guidelines. However, Class B equipment can be fitted to those ships where the International Convention for the Safety of Life at Sea (SOLAS) and IMO guidelines may not be required (for example, fishing boats). Class A equipment gets the priority of transmitting and receiving data over Class B equipment.
4. The transmitting of AIS data is not allowed to intervene, and it must be automatic and continuous. In case of sudden failure of transmitting AIS data, OOW should be notified. Shore stations can update the information of a ship by polling the ship. In any case, shore stations can only increase the reporting of information.
5. If any data is entered manually, the confidentiality should be validated according to the international guidelines and agreements.
6. Unless there are any security concerns and safety of the ship is in danger, the AIS system should be in operational mode.
7. At least once in a month or once in a voyage, OOW should check and validate the AIS transmitted data for each vessel.

8. In case of malfunction of the AIS system, an alarm is triggered by built-in AIS technology, BIIT (built-in integrity test). In such cases, AIS should be shut down, and it must stop transmitting the data.
9. It is essential to understand that AIS is not a navigational service. It provides additional information to the navigation system, but it cannot replace other navigational systems in a vessel.

2.1.4 Nature of Data Transmitted in AIS

AIS devices can receive and transmits the information between the ships and shore stations automatically. Typically, there are three types of data that are exchanged and transmitted via AIS devices: static, dynamic, and voyage. These data types are mentioned below in the next three tables [21].

A. Static Data Types in AIS

The below table provides detail information regarding the static types that are involved in the AIS system.

Table 1: Static data types in the AIS system

Information	**Descriptions**
MMSI	A unique ID for a vessel
Call Sign / Name	An international radio call sign that is assigned to the vessels during a vessel's registration
IMO number	Unique ID that is assigned during ship registration, and it is referenced to the shipowner or the company.
Vessel's Dimensions	Vessel's dimensions based on the nearest position of the AIS Station on the vessel, and it is measured in meters.
Type	Different types of ships and they are categorized based on the type of the cargos they transport and their sizes. Some of the different types of marine vessels are cargo, tanker, passenger, reserved, etc.
Positioning of antenna	Location of antenna in a vessel

B. Dynamic Data Types in AIS

In dynamic AIS data, it contains information like the position of a ship, current timestamp, the heading of a ship and its navigation status and other information. The following table provides an overview of the dynamic data type that is transmitted by the AIS devices.

Table 2: Dynamic data in the AIS system.

Information	**Descriptions**
Location Coordinates	Current position of a vessel in latitude and longitude.
Timestamp	Current positioning time of a vessel in UTC in milliseconds. This is calculated since 00:00:00 Thursday, 1 January 1970.
COG (Course over ground)	Degree relative to true north
SOG (Speed over Ground)	Speed of a vessel ranging from 0 – 102 knots
Heading	Degree of heading of a vessel from 0 – 359
Navigational Status	These are manually entered information by OOW
ROT (Rate of Turn)	Information related to right or left from 0 – 720 degree in minute

C. Voyage Data Type in AIS

This is the third data types that are transmitted by AIS devices. In this data type, the AIS data contains information related to the vessel such as vessel' draught, the nature of cargo that vessel is carrying, the destination port of the vessel, etc. The following table contains detail information related to transmitted voyage data types in AIS devices.

Table 3: Voyage data in the AIS system.

Information	**Descriptions**
Vessel's Draught	This is entered manually before starting the journey by a vessel's authorized personnel. It is represented in between 0.1 – 25 meter(s)
Type of Cargo	This is entered manually, and it can be categorized e.g. - Dangerous goods - Harmful substances - Marine pollutants
Destination and ETA	This is entered manually by a vessel's authorized personnel. It gives information related to the destination of a vessel and estimated time arrival.
Route Plan	This is entered manually before starting the voyage and gives information regarding the route plan.

D. Short-Safety Message

These are short messages which are entered manually by authorized personnel, and these short messages are used for communicating between the ships and shore stations. These kinds of short messages are used in warning the current danger situations, give information regarding specific routes, weather conditions, etc. [22]. In these types of messages, a maximum of 158 characters are allowed to broadcast between different vessels and the shore stations.

2.1.5 Dimension of Ship

A marine vessel is a three-dimensional object. It has measurements of width, depth, and length. In transmitted AIS data, the ship's dimensions are entered using the overall length and width of a ship. In AIS data, these dimensions are mentioned as referencePointA, referencePointB, referencePointC and referencePointD. While entering the internal and external reference points, the values of a ship's

dimension (A+B and C+D) should be identical. Below is a short description related to the ship's width, length, and depth.

1. **Width:** Ship's width is also called beam of the ship. It is measured in feet and inches. In the 2-dimensional figure below, the width is shown using the name beam (B).
2. **Length:** It is an essential dimension of the ship, and length is directly related to the speed, resistance, and friction of the ship. Like width, it is also measured in feet and inches. In the figure below, the length of a ship is shown by using term LOA (Length Overall) and LBP (Length Between Perpendiculars).
3. **Depth:** The depth is measured from the bottom of the ship to the side of any deck that is used as a reference point.

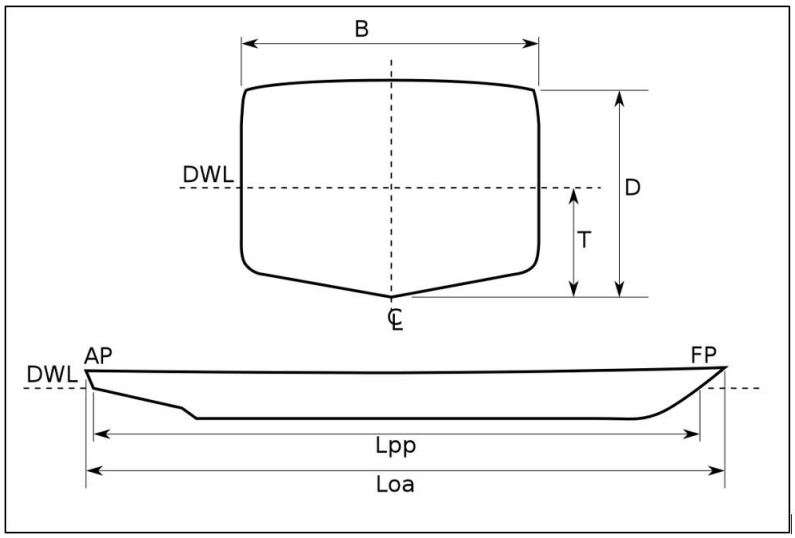


Fig 2: An example of a 2-dimensional figure of a ship. Licensed by: CC BY-SA 3.0 [23].

2.2 AIS Data Extraction

The primary purpose of this chapter is to demonstrate how AIS data was extracted and what kind of technologies and methods were used to extract AIS data. In this sub-topic, it is also going to be discussed about the reliability of the AIS data. Below contains a brief introduction related to the method and technologies to extract web-based data. The reason for discussing related to web-based data is because the AIS data extraction was done using web-based technology.

2.2.1 Extraction of Web-Based Data

In a simple term, web-based data extraction means the extraction of data from web sources. This kind of extraction may range from scrapping web pages, using web API (Application Programming Interface), HTML and XML sources, etc. The extraction of data from a web source is more straightforward and takes less human effort than doing manually. However, one of the challenges of extracting web-based data is that the extracted data may not be appropriately structured and maintainable [24]. For example, extracting data using standard web API may be structured appropriately. Whereas, extracting data using the web-scraping method can cause lots of problems in the data processing stages. Some of the challenges that can occur while doing data extraction using web sources are mentioned below [24].

- The extraction of data from web sources may reduce human errors and cost overhead. However, not all data are well structured, and it may cause many problems while cleaning the data.
- Implementing web automation to extract data may need IT experts. Highly automated technology may increase the performance of the extraction of data but does not guarantee the accuracy of data. So, it is essential to understand what are the alternative automated extraction techniques that could be used to achieve better accuracy of data.
- These types of technologies can process large volumes of data in a relatively short time. These types of requirements are needed typically in the stock market or competitive business value. So, connectivity could be significant issues and may lead to a large number of missing values while using this type of technology.
- Managing privacy and confidentiality of data is a huge problem. While doing data extraction from the web sources, it is crucial to follow guidelines and provide secure data extraction whenever needed.

2.2.2 Ensuring the Reliability of AIS Data

This sub-topic is focused on the reliability of the AIS Data and the reliability of a source of AIS data that was used in this thesis.

A. Overview of Reliability of AIS Data

AIS tracking devices help to manage thousands of marine vessels each day, and without any doubt, there are some anomalies presented there. The reliability of AIS data is always on doubt, and not even reputable tracking organizations like marinetraffic.com can provide 100% accurate AIS tracking data. It is believed that up to 2% of the data collected by reputable tracking companies contain doubtful and inconsistent AIS data [22]. This type of uncertainty data is mostly presented from non-commercial vessels like small leisure ships. The algorithms are successful between 75 – 90% in correcting and finding the reliability of the AIS data. The following are some of the leading causes that are creating problems in maintaining the reliability of AIS data [25].

1. **Intentional Interferences:** AIS data is vulnerable to security issues. The hacker can use several techniques like hijacking, spoofing, etc. and these security problems can alter the content of AIS data. The algorithms that detect these kinds of attacks are not 100% accurate. One of the solutions for this type of problem is to use the digital signature method while transmitting and receiving AIS data.
2. **Human Error:** This is one of the greatest threats to the reliability of AIS data. Human errors are creating problems in maintaining the accuracy of the AIS data. There are some manually entered attributes like destination, port, estimated time, vessel's name, IMO number, etc. in AIS data. While entering this type of attribute manually, there is a high risk of human errors, for example, spelling mistakes, writing styles, etc.
3. **Faulty AIS Device:** The reliability of AIS data can also be affected by a technical problem like faulty AIS devices. The harsh weather, e.g. lightning can damage the devices. If the devices are defective, it may not be able to broadcast and receive full transmitted data.

Below figure 3 gives an example of the reliability of extracted AIS data from the Baltic area. The figure on the left side shows that there has been no AIS data received for up to 65 km between 2019-06-12 9:42:00 and 2019-06-12 18:42:00 for mmsi 244224000. The interesting fact is that the location where AIS data was

missing is close to the Finnish and Swedish shores. The data was lost for approximately 3 hours. The figure on the right side shows the data points, which have consecutive data gaps of at least 3 hours. The graph shows that the gaps of over 3 hours are more frequent near the edge of the sea and are less frequent in the middle of the sea. More than 5% of ships had these types of data points at some stages in that extracted AIS dataset. Even though, there is missing data in that area, having a missing data for marine vessels is not considered good because IMO strict guidelines won't allow it.

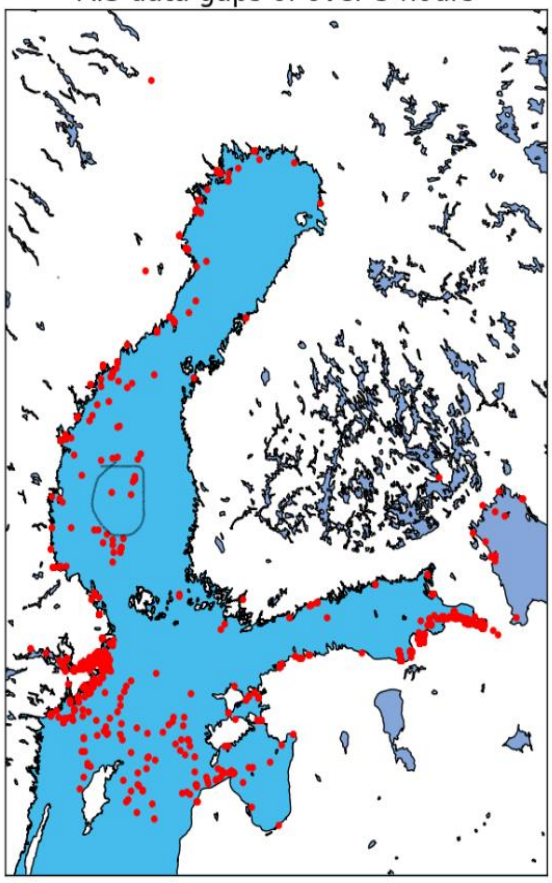
With the exception of extreme cases like the security risks and danger to the surrounding, every vessel over 300 tons must transmit data continuously while in the sea. However, there were some expected missing data in the Baltic area if the missing data gaps are over 5 hours and if it occurs near the border of the Baltic sea. The reasons behind these expected missing data are discussed in upcoming topics. The reason for AIS missing data can be categorized into two parts: technical reasons or intentional reasons. For technical reasons, the missing data can be caused by faulty devices, limited range, etc. Whereas, in intentional cases, the AIS devices are switched on and off to disrupt the AIS data. These two different causes of missing data in AIS devices are discussed detail in chapters 3 and 4.

In another scenario, it was found that two different ships had the same location coordinates at the same time. The location coordinates and time difference for two different ships had an exact match. This is not possible. However, transmitted locations are provided using GPS (Global Positioning System) system, and the GPS cannot provide 100% accuracy of the location of objects. Based on the capabilities of the device, the maximum accuracy of GPS can be around (+-) 5 meters [26]. From these types of examples, it can be seen that AIS data is inconsistent for many reasons. So, further actions need to be taken by the authority to make AIS data more reliable and trustworthy.

AIS data Missing up to 65 KM

Missing data up to 65 km

AIS data gaps of over 3 hours



Missing data gaps of at least 3 hours

Fig 3: Examples of AIS data having higher missing gaps in the Baltic Area.

B. Overview of Reliability of Source of AIS Data

The source of AIS data used in the thesis is from https://www.digitraffic.fi/en/marine-traffic/. The website was started in 2002 as a joint project between the VTT Technical Research Center of Finland and Aalto University. In the initial stage, it was providing only road data. However, the data service was expanded to rail and marine area when "The Finnish Transport Infrastructure Agency" started to manage the website. All types of data, including marine, are hosted in the cloud and can be accessed via either by using web-socket or by using standard web API. Both standard web API and web-socket API provides real-time data. Data provided by Digitraffic is open-source data and is free to use. The data is available in the Creative Commons 4.0 license [27]. The

following are some of the licensing terms and privacy policies regarding the use of its open data [28].

- Users can distribute, modify, and can use the data commercially. However, users should provide and give proper credits and references to the Finnish Transport Infrastructure Agency and should reasonably use data.
- Users cannot apply additional license terms and conditions that legally restrict from other people to use original licenses while using their services.

2.2.3 Technologies Used in AIS Data Extraction Process

Digitraffic offers two different types of data extraction process; standard web API and web socket. Initially, it was decided to use its standard API because it was easier to use, and data was already available in proper JSON format. However, one of the significant drawbacks of using standard API from Digitraffic was, there was a significant problem related to missing data. At the moment, it does not provide a history of locations via its standard API. It's standard API only returns the latest location per each vessel. So, using its standard API was not a better option. So, it was decided to use its web socket API, which also returns real-time marine data. Its web-socket API uses the MQTT protocol.

MQTT protocol is an extremely lightweight message transport system that allows a user to subscribe and publish the messages [29]. There are mainly two different methods and technology involved in the data extraction process. The first step was to develop a data extraction program, and the second step was to host and run a data extraction program in a server so that it can operate automatically without external interference.

A. Web Automation Program

To extract data from web socket API, a data extraction program was developed using Python programming language. The main library used in Python program language was paho-mqtt, version 1.4.0 [30]. The reason behind using this library is that Digitraffic's web socket API transmits data using the MQTT protocol. It was

developed a Python program which can extract data via console. The benefit of using a console program over the browser is that it gives more flexible and does not need browsers to render its incoming data. This web socket API was transmitting around 2Mb data per second, and the Python program was able to extract and save those data automatically and continuously. To extract data from web socket using the pahoClient library, it needs three main callback functions; on_connect, on_message, and on_subscribe. More information about how to use these callback functions, how to set username and password can be found in its official documentation [30]. Below is the small code snippet which was used in the web automation program while extracting AIS data.

```
clientId = str (uuid. uuid4())
topic = 'vessels/#'
Client= pahoClient.Client(clientId, transport='websockets')
Client.username_pw_set (username, password)
Client.on_connect = on_connect
Client.on_message = on_message
Client.on_subscribe = on_subscribe
Client.connect(server, port)
Client.subscribe(topic, 1)
Client.loop_forever ()
```

B. Data Extraction Process

The AIS data extraction program was hosted in the Google cloud. The data extractor program was difficult to run in a local machine because of the limited resources available. The application was extracting real-time data using web socket API. Thus, it needed a robust machine with a continuous internet connection so that incoming data can be extracted and saved without any technical difficulties. Below are the steps that were taken to run the AIS data extraction application in the Google cloud.

- An ubuntu virtual machine instance was created using suitable hardware specifications.

- After a virtual instance was ready and started, a data extraction Python program file was uploaded to a virtual machine, and then all necessary Python packages were either updated or installed. The data extraction Python program was running in the Python3 environment.
- The next step is to run a Python script continuously. Basic syntax to run Python in a Linux environment is by typing "`Python3 Pythonfile.py`". To make sure that the Python program process stays alive and runs in the background continuously, the syntax "`nohup Python3 Pythonfile.py &`" was used.
- Finally, real-time data was automatically saved in a disk, and these saved data can be accessed and downloaded from the Google cloud whenever it is needed.

2.3 AIS Data Processing

There is a common saying in data science, "properly structured data is better than fancy algorithm". If the data is not cleaned and processed correctly, the results can be different. This sub-chapter is focused on what are the methods and steps for doing data processing and data manipulation. There will be further brief discussion regarding steps that need to be considered while doing time-series data processing: resampling, and up-sampling of the data. At the end of this sub-chapter, there will be a brief introduction related to the feature extraction of AIS data.

2.3.1 AIS Data Cleaning and Processing

After extracting and saving data for two weeks, the original size of data had reached over 18 Gb. This is a massive size of data, and the data cleaning process was tedious and time-consuming. The extracted AIS data was available in an unstructured JSON format. Some of the common problems faced while attempting to read JSON files were, some data had no proper syntax, and some of the data had illegal characters. So, it was impossible to read data directly using the JSON library without cleaning data first. Below are some of the methods that were used for data cleaning and manipulation in extracted AIS data.

A. Regular Expression Approach

This is one of the most common and powerful approaches. In this approach, unwanted data, including strings, numbers, chars, etc. can be removed and replaced by identifying specific patterns or criteria using the regular expression [31]. Regular expression is supported by all major programming languages like C#, Java, Python, Perl, etc. In a regular expression, there are two different types of characters; metacharacters (e.g., `*`), which is also called wildcard and literals (e.g., 1, a, b,). Regular expression may look simple and easy to process. However, writing a complex regular expression can be surprisingly tricky. For example, let us suppose we have data text like this "`We would like to have 2 orange and 3 bananas`". If you to want to extract only numeric data, e.g., 2 and 3 from the above text. In this case, we can write a simple regular expression "\d+" which represents the numeric value and extract it. However, if we need to extract an email address from the text, then it quickly becomes a complex regular expression; `"[A-Z0-9._%+-]+@[A-Z0-9.-]+\.[A-Z]{2,}"`.

The extracted AIS data also had a lot of unwanted characters like ">", "//", unwanted numbers, timestamp problems, incomplete JSON, and many other issues. Python regular expression was used for removing and replacing these unwanted characters. Regular expression in Python can be imported using `re` module and the functions like `re.match()`, `re.findall()`, `re.search()` and `re.sub()` were used during data cleaning steps [31].

B. Removing Duplicate Observations

Duplicate observations can occur in data in many ways, e.g. merging two different datasets from the same source, web-scraping, receiving data using client-servers, etc. If there are many duplicate observations in data, then it may give unpredictable results because of the artificial bias. After doing an initial cleaning of AIS data using a regular expression-based approach, data was readable using the Python JSON library. However, data was not properly structured, and there were

duplicate observations for some vessels. The following steps were taken to remove the duplicate observations from AIS data.

1. AIS data was converted from JSON to CSV format using the pandas library in Python. The reason for this conversion was because CSV format is a more readable form than JSON, XML, etc.
2. The next step was to group the data in ascending order based on MMSI and timestamp. MMSI is a unique identification number for each vessel, and timestamp (in milliseconds) is the recorded time for that vessel in a given position.
3. Data is defined as duplicate if more than one observation is present in a dataset where MMSI and timestamp are precisely equal.
4. With the help of a pandas library in Python, all duplicate observations for vessels were removed. Because of the real-time data extraction using web-API, there were fewer duplicate observations. Less than 1% of total observations were marked as duplicates. Even though duplicate observations were rare, the majority of vessels had duplicate observations at some stages in that 2-weeks' time.

C. Removing Outliers

The outlier means the data is drastically different from the majority of the data. It is possible to have multiple outliers in a single dataset. Some of the common outlier types are explained briefly below [32].

- **Contextual Outliers:** In Contextual outliers, the data point values change significantly from other data points in that context. In another sense, the same data points which were considered outliers may not be outliers in another context. The simple example can be a time-series dataset where it contains seasonal and non-seasonal value.
- **Point Outliers:** It is called point outlier if data point value is far from the outside of the majority of the data points in the dataset. There were examples related to this outlier in the AIS dataset. For example, there were location coordinates like this; (59.366, 21.903), (59.369, 29.903). There are

three decimal points in the coordinates, and suddenly the data was coming as one decimal point. When some of the receiving data points are suddenly in the form of one decimal point (59.3, 21.9), it creates a significant gap between them. The latest received data may change the direction of the routes of the ships, and it shows a significant distance between two points. These types of data points made it look like the ship has traveled extremely fast within a short time.

- **Collective Outliers:** If a small part of the dataset deviates collectively from the entire dataset, then it is called collective outliers. In this type of outliers, individual data points are not identified as outliers. One of the simplest examples of collective outliers can be a dataset that contains seasonal trends in a time-series dataset.

Some of the common causes of these outliers mentioned above in a dataset are; measurement problems, unexpected errors, seasonal peaks, etc. Because of outliers, it may generate different patterns and may give unexpected data insights. So, it is crucial for a person who is dealing with data to discover outliers and do necessary and analyze it. One of the topics of this thesis was to detect outliers in AIS Data. It is discussed in detail in chapter 3 and chapter 4. However, the outlier in this data processing context is those types of data that do not contribute to the detection of anomalies behaviors in ships.

The dataset contains all the vessels' data, which were in the Baltic sea area, and there were some problems in it. One of the common issues was that all the vessels were not moving in the sea. For example, vessels with MMSI 209508000, 210426000, 210474000, 212709000 have not moved at all or have moved extremely slow and have covered very small distances between the extracted time-period. These types of data are unwanted, and it may cause some issues while doing data analysis. So, during a data processing stage, a total of 363 unique ships were found either not moving at all or moving extremely slow for two weeks. Thus, a total of 363 vessel data was removed as unwanted data from a dataset.

D. Handling Missing Data

Missing data is one of the significant problems in machine learning tasks. It can vastly affect the outcomes and may give false interpretations of the insights of data. It is not considered a good habit to remove or discard data if they contain missing values. This may result in having insufficient data and may not interpret real data insights. Some of the conventional statistical methods to deal with missing data are Mean Imputation, Hot-Deck Imputation, Multiple Imputation (MI) [33]. There are also ways to deal with missing data using machine learning techniques. Imputing missing values using a machine learning algorithm is a prediction task. AIS data was extracted by using web socket API, and there were missing values. In the AIS data, there were some missing gaps between two different consecutive time periods for a vessel. However, not all missing gaps found in AIS data were caused by intentional reasons.

I. Expected Missing Data:

There were some expected missing gaps between two consecutive readings for some vessels. This is because the data was extracted only within the Baltic sea region. When some ships enter the Baltic Sea area, then its data is recorded, and when it goes outside of the Baltic Sea region, the data is no longer recorded. Moreover, if the same ship enters again in the Baltic region, then its data is recorded. Because of this reason, some ships may have a higher expected missing data gap.

II. Unexpected Missing Data:

These types of missing data can happen because of several reasons. For example, while extracting data via a web socket, it needed a 24/7 internet connection. Even though the data was extracted using Google Cloud, it may be still possible that some data may have been lost. Another technical reason can be because of faulty AIS devices. Some of the reasons for missing data can be purely intentional too. The topic regarding missing and non-missing data is discussed in detail in chapters 3 and 4.

2.3.2 Fundamental Exploratory Analysis of AIS Data

This sub-topic is focused on the fundamental exploratory analysis of AIS data. During the pre-processing stage, it was done basic exploratory analysis to know about the data structures, how spread the data is, and how the data is correlated with each other. The below figure shows the plotting of all extracted AIS data that was collected from the Baltic Area for two weeks. In the below figure, it has been marked as a red circle if there are data gaps of more than 1 hour for a vessel between two consecutive locations. The green circle indicates that the consecutive data gaps are less than 1 hour for a vessel. The below figure shows the spread of AIS data in the Baltic Area. It can be seen in plotting maps that the data has spread among all four countries and their shores regions: Finland, Sweden, Estonia, and Russia.

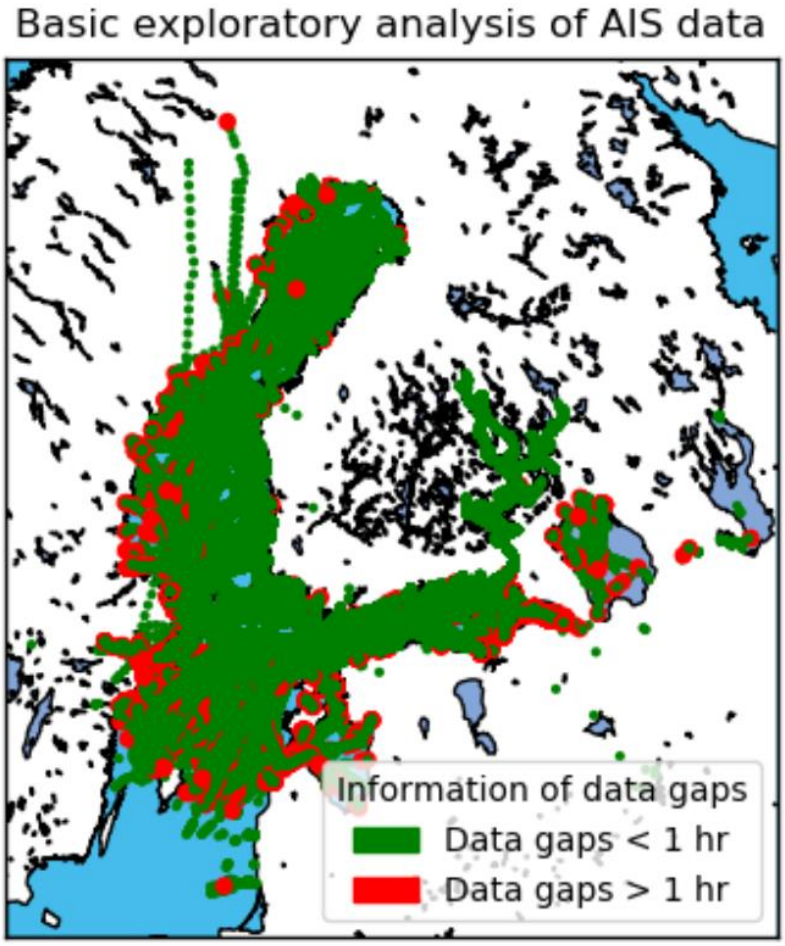


Fig 4: Basic Exploratory Analysis of AIS data between 4th – 16th June 2019

To represent how often the data was received and collected in Finnish shores, a table was created. In the below table, a different range of periods (in minutes and hours) was created. The range of time in the table is from 0 minutes to 300 hours. For example, if a time difference between two consecutive received data points for a vessel is 4 minutes, then this falls within the range of [0, 5] minutes, and it will be counted as +1 for this range of data points. It was difficult to visualize using the graph for this table because over 99% of data were within between [0, 5] minutes range. Because of this high dominant of the first range values, the histogram visualization did not provide excellent results. For this reason, the following table was created, which represents the time range, total percentage of data in that time range, and total number of vessels that have at least one data point within that time range.

Table 4: Information on extracted AIS data points at different intervals of time.

Time Difference Between Two consecutive readings for the same vessels	**% of total data points in that time range**	**Out of 1686 vessels, a total number of vessels that have at least 1 data point in that time range**
[0, 5] minutes	99.62184609	1674
[5, 10] minutes	0.263774957	1512
[10, 15] minutes	0.036457934	1239
[15, 30] minutes	0.035604785	1233
[30, 60] minutes	0.017404256	1075
[1.0, 2.0] hours	0.008626292	853
[2.0, 4.0] hours	0.005045772	669
[4.0, 8.0] hours	0.003510102	571
[8.0, 15.0] hours	0.002824874	497
[15.0, 30.0] hours	0.00247278	498
[30.0, 50.0] hours	0.001164617	315
[50.0, 100.0] hours	0.000807107	244
[100.0, 300.0] hours	0.00046043	164

2.3.3 Data Interpolation and Feature Extraction

This topic briefly explains the interpolation strategy and feature extraction process in the AIS data. Interpolation in the AIS data can help to recover the lost AIS data. So, the interpolation can be helpful for shore stations and other ships that are exchanging information to each other. The alone linear interpolation technique may not be sufficient for revealing the movement of a vessel. Nguyen et al. [34] have proposed three different combined interpolation techniques in AIS data; linear, cubic hermitian, and identification mechanism for the missing data. They have claimed that these interpolation techniques are promising and can be used in real-time for ships. However, in the thesis, a simple linear interpolation technique was used to find the missing data. This technique was used only for the smaller consecutive gaps of less than 15 minutes for a vessel. The code snippets in Python that was used for the linear interpolation and resampling is below.

```
def interpolate_aisData(AIsDataFileName):
     df = pd.read_csv(AIsDataFileName)

     df.index = pd.to_datetime(df.index, unit='ms')

     df.loc[(df['mmsi'] != df['mmsi'].shift()) |
(df['Gaps(Hrs)'].shift() > 0.25), 'Temp_MMSI'] = 1

     df['Temp_MMSI'] = df['Temp_MMSI'].cumsum().ffill()

     df1 = (df.groupby('Temp_MMSI', axis=0)
            [['mmsi', 'latitude', 'longitude']]
            .resample('3min')
            .mean()
            .groupby(level=0)
            .apply(lambda x:
x.interpolate(method='linear')).reset_index().drop('Temp_MM
SI', 1))

     ---------- more code goes here

     # returns the interpolated AIS data
```

The reason for using a linear interpolation technique, in this case, is because all data points were not interpolated. One of the critical research parts of this thesis was to research on missing data. The interpolation technique used in this thesis

was to make the data reading uniform by resampling into a fixed interval of time of 3 minutes. As the interpolation was done only for a smaller time gaps of less than 15 minutes, the error from the linear interpolation was minimum. By assuming the average speed of a ship is 15 knots, the maximum it can travel in 15 minutes is around 6.8 km.

For each vessel, the AIS data contains mmsi, cog, heading, navstat, posAccuracy, raim, rot, sog, utc_second, timestamp, longitude, latitude. However, not all features columns were not needed in this thesis practical part. For the practical part of the thesis, the selected columns of data were mmsi, timestamp, latitude, and longitude. The rest of the data points and data columns were excluded.

3 Design and Implementation

This chapter is focused on the practical part of this thesis, and it attempts to answer what are considered abnormal and what are considered normal behavior for the ships. Chapter 3 is also focused on what are the technologies and algorithms that were used in finding abnormal behavior of the ships. Chapter 3 is divided into three main research areas: finding missing data, abnormalities in ship movement, and ship's encounters. In ship's encounter, it attempts to find if any two different ships have been closer to each other for a distance of less than 500 meters when the speed has been less than or equal to 0.01 m/s for at least 3 minutes.

3.1 Background of Research

The collected original AIS data sizes from "Finnish Transport Infrastructure Agency" were over 18 GB in 14 days. After doing pre-processing, data cleaning, and removing all unnecessary data, there were approximately 300 Mb data left from 14 days. Using whole data at once for research purposes was an extremely time-consuming and tedious process. So, it was decided to do research on smaller areas in the Baltic sea.

To make the research area more concrete and specific location-wise, a proper measurement boundary was created in the Baltic Sea. The geometry size of the measurement area was created in such a way that no major ports or seashore lies within that area. The measurement area was checked and validated with the help of Google maps and marinetraffic.com. As the measurement area did not contain significant islands or ports, the illegal activities that may occur, e.g. throwing drugs to and from the ships and throwing illegal materials to the islands were excluded from the research. With the help of figure 4, the measurement area was selected in such a way that the selected area is within the proper range of receiving and transmitting AIS data from the Finnish shores. The below figure and table show the polygon-shaped measurement area and its coordinates of latitude and longitude.

Table 5: Showing the measurement area coordinates

Latitude	**Longitude**
58.57	19.67
59.32	19.31
59.70	20.64
59.50	22.25
59.13	21.87
58.76	21.15

The figure below shows how the measurement area was created in Google maps using coordinates from Table 5.

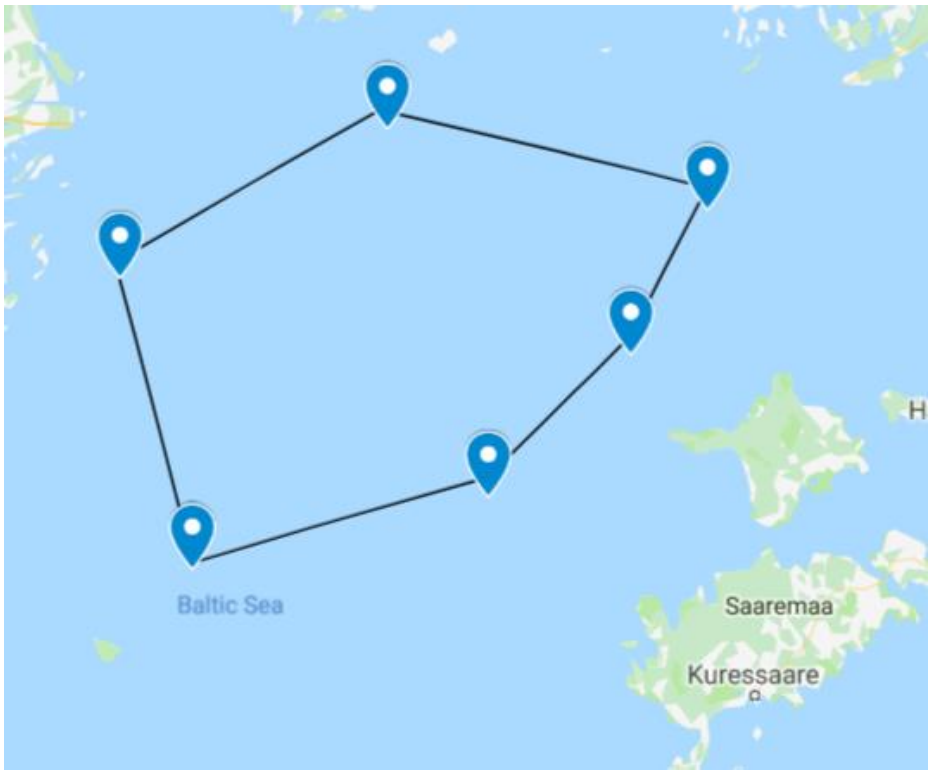


Fig 5: Polygon shape measurement area showing in the Baltic Sea.

The exact replicate of the above-shown polygon figure was created in the Python program with the help of a shapely library [35]. This measurement area was created using coordinates that are shown in Table 5. The following contains a small code snippet in Python for creating a polygon-shaped measurement area.

```
def GenerateMeasurementBoundary (longitude, latitude):
    polygon = Polygon (
        [ (19.67, 58.57),
            (19.31, 59.32),
            (20.64, 59.70),
            (22.25, 59.50),
            (21.87, 59.13),
            (21.15, 58.76)
        ])
    point_instance = Point ((longitude, latitude))
    a = polygon.contains(point_instance)
    val = np.where(a, 0, np.nan)
    return pd.Series([val])
```

The following figure shows the ships' routes in the measurement area. A total of 777 unique ships were passed through the measurement area in 14 days.

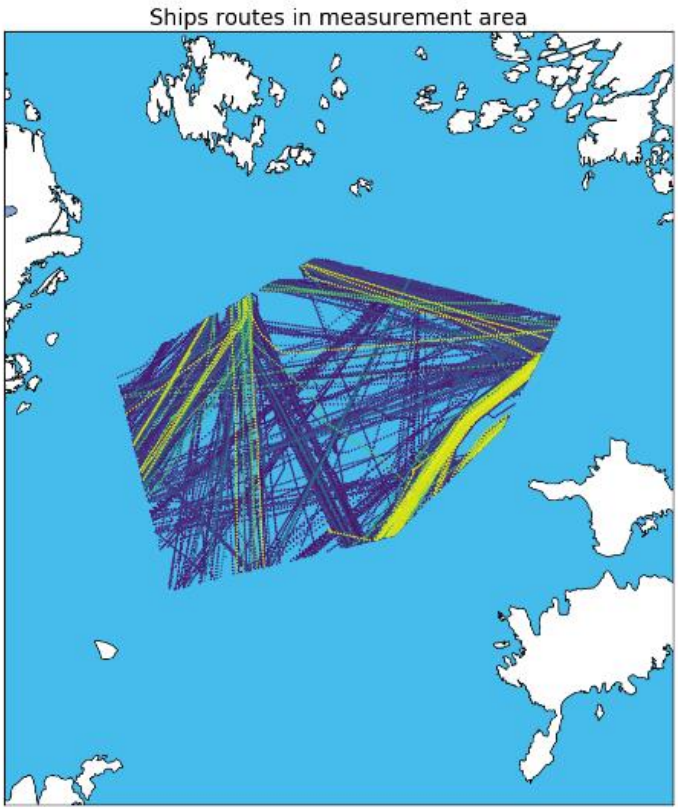


Fig 6: Ships routes in the measurement area for the period of 4th – 16th June 2019.

A. Haversine Formula for Distance Calculation in Latitude and Longitude

Since AIS data uses spherical Mercator coordinates: longitude and latitude, it can use spherical trigonometry to compute its surface distance. A haversine formula can be used to calculate the distance between two spherical Mercator coordinates. To calculate the angle between any two given points in the spherical geometry, the formula is given as follows: [36]

Equation 1: $hav\left(\frac{d}{r}\right) = hav(\varphi_2 - \varphi_1) + cos(\varphi_1)\, cos(\varphi_2)\, hav\,(\lambda_2 - \lambda_1)$

In the above equation 1, hav is a haversine function, and it is defined as below.

Equation 2: $hav(\theta) = sin^2\left(\frac{\theta}{2}\right) = \frac{1-cos(\theta)}{2}$.

In equation 1, d is the distance between any two given points in the spherical geometry. The parameter r is defined as the radius of the sphere (in this case, it is the Earth). φ1 and φ2 are the latitude1 and latitude2 of two given points in a sphere, and these points are measured in radians. λ1 and λ2 are the longitude1 and longitude2 of two given points in a sphere, and these points are also measured in radians.

The parameter d from the above equation 1 can be solved by using the inverse haversine functions or by using the inverse of a sine function. The final equation of calculating the distance is given as follow:

Equation 3: $d = 2r\, arcsin\left(\sqrt{sin^2\left(\frac{\varphi_2 - \varphi_1}{2}\right) + cos(\varphi_1)\, cos(\varphi_2)\, sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right)$.

The haversine function computation applies assuming a perfect sphere. However, this is not the case with the Earth shape because it has more complex ellipsoid and may require some complex formulas [37]. However, for the thesis purpose, the haversine formula for calculating distance between two different locations coordinates for a ship was adequate.

3.2 Research on AIS Missing Data

It is the first practical part of this thesis, and it attempts to answer if there was any AIS missing data in the ships. The reason for this research is to find how often

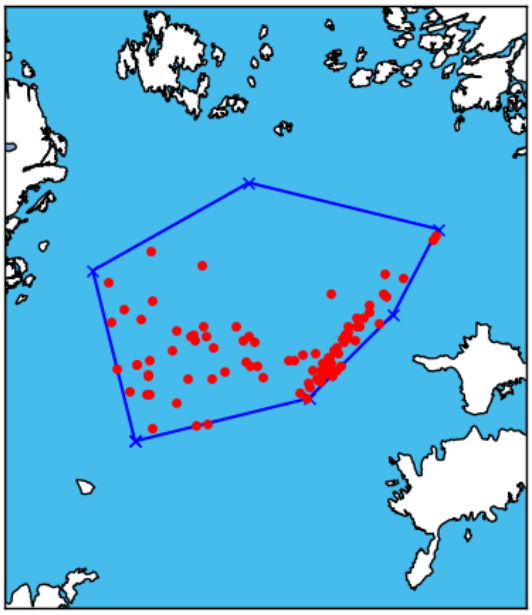
there was an occurrence of missing data when the ships transmit their locations' by using AIS devices. The time difference of missing data on two different consecutive locations for the same ship may vary a lot. In this topic, it will be focused on trying to find the patterns on missing data and analyze whether those missing data were frequent, or it happened on rare occasions. This topic also tries to analyze and create an open discussion on whether those missing data on ships should be considered as accidental or intentional. The following are some of the steps that were taken while trying to find missing data in the AIS dataset.

- From the measurement area, as shown in figure 5, each corner of that geometry shape of the measurement area was recorded as latitude and longitude coordinates.
- The next step was to filter and extract only those types of ships' data, which had routes in the measurement area. Only those ships which had gone through the measurement area were included in the research, and the rest of the data was excluded.
- After doing an analysis of missing and non-missing data of the ships from the measurement area, the graph was plotted using a basemap library in Python [38]. The benefit of plotting in a basemap library in Python is that the graph can be visualized and presented in the real geography of the world map. The graphs of missing and non-missing data for ships are visualized on the next page.

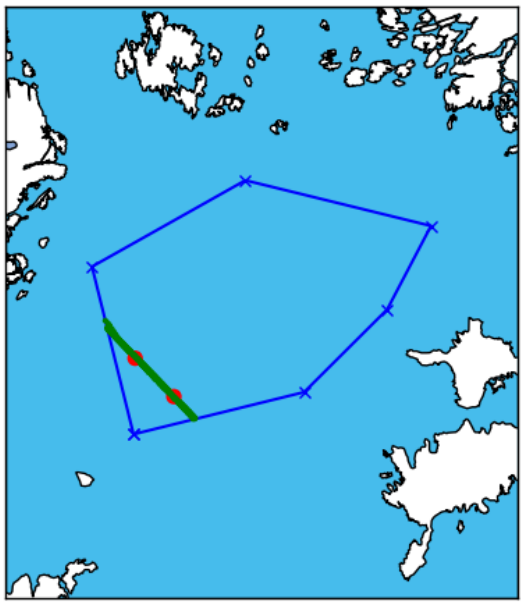
A. Finding Missing Data from AIS

It was found that there were some frequent occurrences in missing data in the measurement area. Out of 777 total ships, 63 ships had missing data from 15 minutes to 4 hours. If the ship had more than 4 hours of missing data gaps, then it was not included as missing data for that ship. This was done to prevent false missing data errors because some of the ships may have gone outside of the Baltic area. In that case, AIS data is not recorded in Finnish shores. The following are some of the example figures that show different situations of missing and non-missing data in the measurement area. The red dots in the below figure indicate

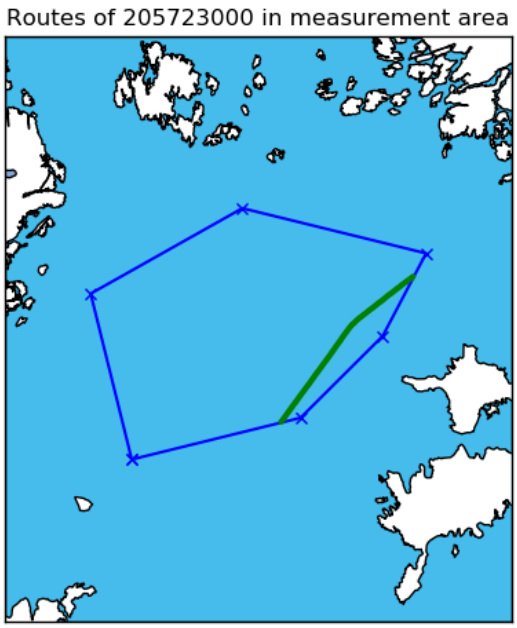
the consecutive data gaps of over 15 minutes. Whereas, the green dots indicate the consecutive gaps between two different points is less than 15 minutes.



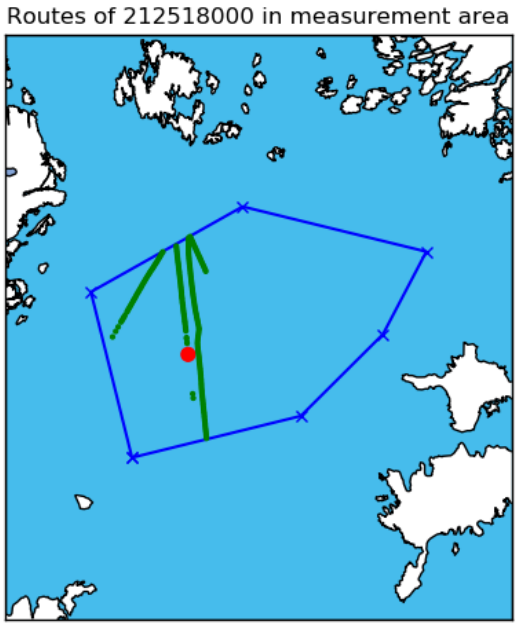
Overview of missing data



Follows the same route but some missing data in random position



Follows same route & no missing data



Random routes & missing data data

Fig 7: Example of different types of missing data in the measurement area

As it has been already mentioned that a total of 63 ships had missing data, the above graph visualizes some of the missing and non-missing data situations in the measurement area. One of the figures shows all occurrences of missing data in those two weeks' time. In one of the graphs from figure 7, there is an example of a ship where it had some missing data randomly, and it was following the same route in those 14 days. Another ship had some missing data, and it was moving in a random route. The graph also shows an example of a ship where there was no missing data in those two weeks' time. Further evaluation and discussion of results from this topic are presented in chapter 4 of this thesis.

3.3 Research on Ship Movement

This is another practical part of the thesis, and it attempts to find how the speed of the ships varies in the sea. The primary purpose of this research is to find such ships where speed goes to very low or stops in the middle of the sea area. If there exist such ships, then this thesis tries to find what are the types of those ships and how often do they change their speed while sailing in the sea. At the end of this experiment, it tries to attempt to answer why there are such types of movements in the sea. For this research, the same measurement area was used that was created using location coordinates, as shown in figure 5. The following steps were taken to analyze and research on ships' movement.

- AIS data was checked with the boundary of the measurement area, just like the same process while analyzing and finding the missing data. By doing so, only those type of AIS data was taken into consideration which passes through the measurement area.
- After that, the distance was calculated between two different consecutive location coordinates for each ship. To calculate the distance between two different coordinates for each ship, the haversine formula was used.
- To find how the ship is moving within that measurement area, each ship's speed was calculated. To get a better accuracy result of the speed, it was calculated using the formula of centered finite difference [39]. The unit of speed was meter/second. The following figure shows how the finite difference is calculated.

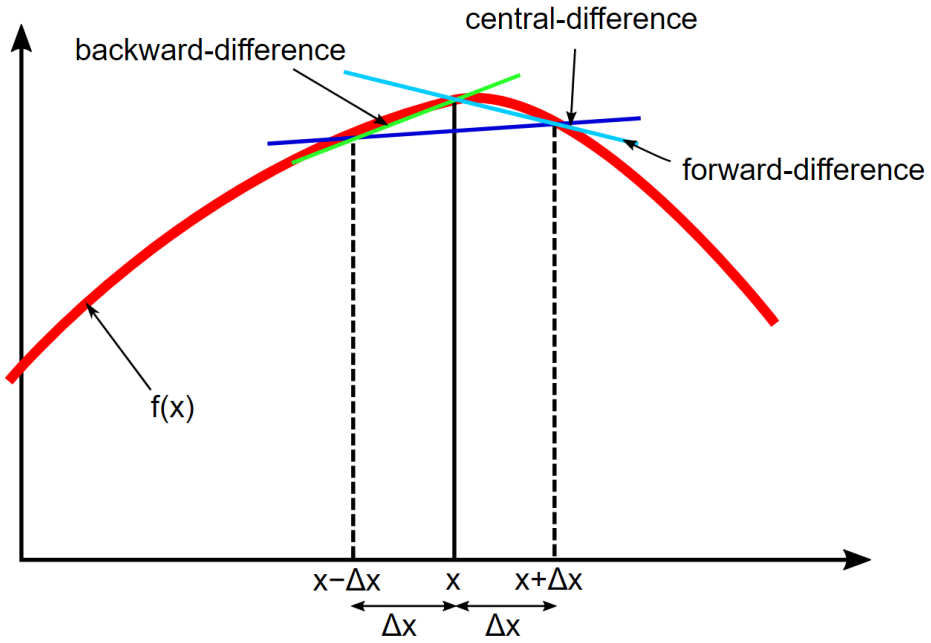


Fig 8: Three types of finite difference. Licensed by: CC BY-SA 4.0 [40]

- If there are time gaps of over 4 hours between two different consecutive locations for the same ship, then the speed between these consecutive data points was excluded. This was done to prevent errors, e.g. the same ships that may have been entered in a measurement area on different days. In that case, the speed would be low as zero because of large timestamp, and less distance traveled, which would result in false speed.

A. Finding of Ship Movement

Out of 777 ships, 254 ships had a speed of less than 0.01 meters/second on different occasions. Many of the ships had a speed of less than 0.01 m/s for a short period, and these types of speed were frequent. One interesting fact was that, out of these 254 ships, some of them had very low speed for more than 20 minutes. As an example, the next two graphs visualize how the ship's movement variation is happening while moving in the Baltic Sea. The data presented in the below graph contains all of the 14 days data for that ship's mmsi 244010945. The detail about these types of movements is discussed and evaluated in chapter 4 of this thesis.

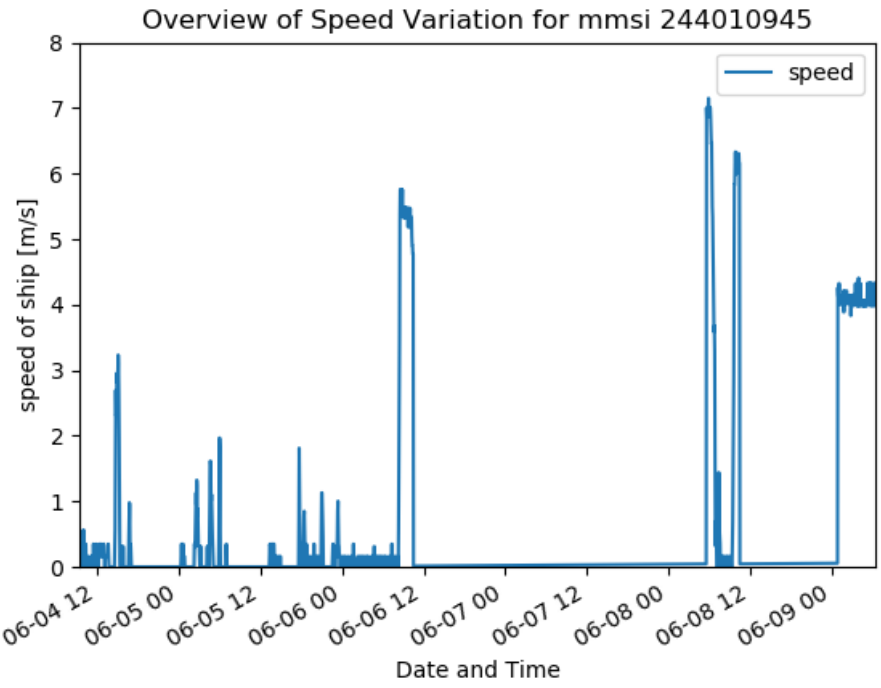


Fig 9: Graph showing the speed variations for mmsi 244010945 in the measurement area (for example, the date: 06-04 12 means 4th June at 12 pm).

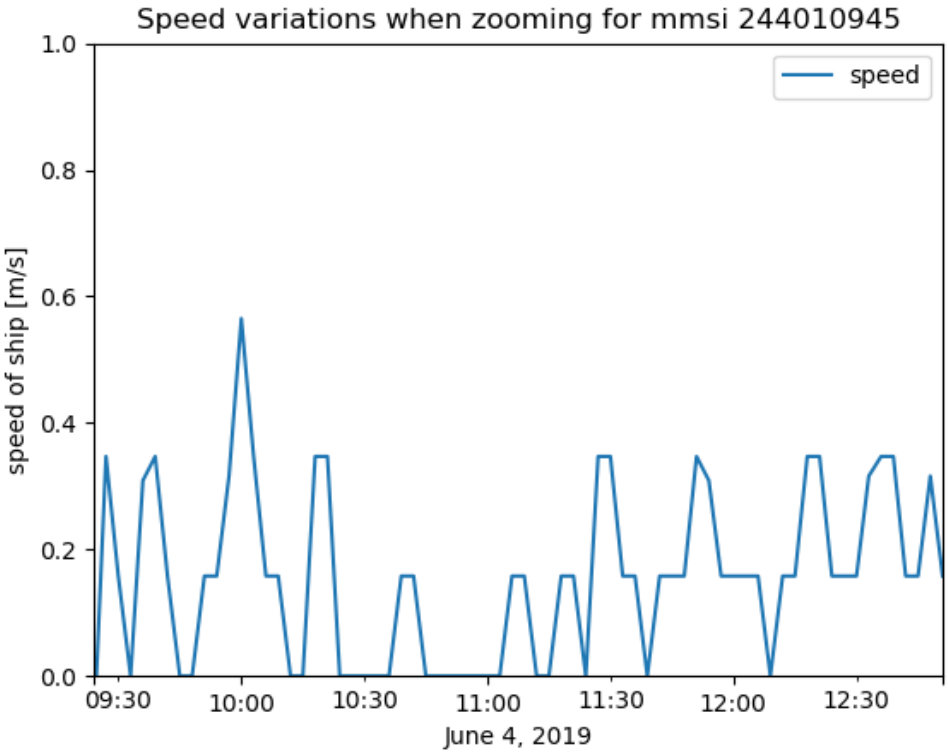


Fig 10: Speed variation while zooming smaller portion for mmsi 244010945

3.4 Research on Two Different Ships Encounters

This is the final part of research in the thesis, and the focus of this research is to find if there exist such ships whose speeds are close to 0.01 m/s, and the distance between these two different ships is less than 500 meters for a period over 3 minutes. The reason for choosing speed close to 0.01 m/s is to ensure that ships were not moving when having encounters with each other. This excludes those type of encounters which may have existed when two different ships were passing each other. For this experiment also, only those ships were included, which had gone through the measurement area and if their speed is equal to or less than 0.01 m/s. The algorithm was created in Python to find the closest two ships. The steps taken to find the encounters between the ships are the following:

- First, the distance between two different ships' positions was calculated using the haversine formula. As an example, the distance was calculated from latitude and longitude of mmsi 244010945 to latitude and longitude of mmsi 309572000.
- The next step was to check if the calculated distance between the two ships is less than 500 meters. If the distance between the two ships were less than 500 meters, then only their respective time difference between them of their encounter locations was calculated.
- If the time difference of the encounter position between two ships was less than 5 minutes and the distance between them was less than 500 meters, then this type of data was extracted and analyzed.

Below code snippet shows how the encounters between two different ships were calculated. The function takes three parameters: AIS data file (when the ship's speed is equal to less than 0.01 m/s), and mmsi of two different ships that may have been in the encounters with each other. The function returns a file that may contain encounters data (if existed between those two ships). The full source code is available in the link mentioned in the appendix of this thesis.

```
def find_encounters (AISDataWithCloseToZeroSpeed,
enounterShip1Mmsi, encounterShip2Mmsi):

    df = pd.read_csv(AISDataWithCloseToZeroSpeed)
    df.drop(['Gaps(hrs)', 'speed'], axis=1, inplace=True)
    target = df.loc[df['mmsi'] == enounterShip1Mmsi]

    target.reset_index(drop=True, inplace=True)
    outsider1 = df.loc[df['mmsi'] == encounterShip2Mmsi]
    outsider1.reset_index(drop=True, inplace=True)

    home_latlon = target[['latitude', 'longitude']]
    outsider1_latlon = outsider1[['latitude', 'longitude']]

    outsider1_timestamp = outsider1[['timestamp']]

     distanceCalc = cdist(home_latlon, outsider1_latlon,
     metric=haversine)

    result_calc = pd.DataFrame(distanceCalc,
    columns=np.arange(695))

    time_sub_ = pd.DataFrame(
                        np.where(result_calc < 0.5,
                abs(target.timestamp.values[:, None] -
                outsider1.timestamp.values), np.nan),
                index=result_calc.index,
                 columns=outsider1_timestamp.values.ravel())

    ---------------------------- more code goes here
    ---------------------------- More code goes here

    # returns the file that may contain the encounter data
(if encounter exists between two different ships)
```

A. Finding if Two Ships Have Met Each Other for Longer Time

To illustrate some of the interesting encounters between different ships in a more detailed way, the following table is created. The table contains 7 major encounters that were recorded in the measurement area. Besides these major encounters, there were presented also some minor encounters in the measurement area. However, some of these minor encounters were excluded from the table and from the research because their encounter criteria didn't fit with the limit that was applied while calculating the encounters between different ships. The table below

contains mmsi of the encounter ships, their duration of the encounters, and the closest distance between them at the time of encounters.

Table 6: Examples of encounters of ships from 4th – 16th June 2019.

Ship MMSI and Ship Type	**Ship MMSI and Ship Type**	**Total duration of encounters**	**Closest distance of encounters**
244010945; Cargo	244060802; Cargo	~ 70 minutes	~ 115 m
249110000; Tanker (Encounter 1)	244010945; Cargo	~ 7 hours and 10 minutes	~ 113 m
24911000; Tanker (Encounter 2)	244010945; Cargo	~ 50 minutes	~ 125 m
24911000; Tanker (Encounter 3)	244010945; Cargo	~ 1 hour and 20 minutes	~ 114 m
309572000; Offshore supply ship	244010945; Cargo	~ 5 hours and 45 minutes	~ 407 m
257038470; Offshore Supply Ship	249903000; Offshore Supply Ship	~ 16 minutes	~ 20 m
311070200; Offshore Supply Ship	257038470; Offshore Supply Ship	~ 3 minutes	~ 470 m

Not all of the encounters shown in the table were odd. The encounters between offshore supply ships were normal and it is discussed in detail in chapter 4. From the table above, it can be seen that there exist even multiple encounters between the same ships. The encounters between mmsi 244010945 and 249110000 had occurred up to 3 times for a period of 4th – 16th June. Another interesting encounter was between ship's mmsi; 309572000 and 244010945. According to the data, both of these two ships had stopped close to each of around 407 meters for more than 5.5 hours. This incident occurred between 2019-06-04 18:21:00 and 2019-06-05 00:09:00 at coordinates 59.3, 21.544 to 59.302, 21.55. Further

discussion of these types of results is evaluated in chapter 4. The figure below shows an example of the ships encounters between ship's mmsi; 244060802 and 244010945 in the measurement area

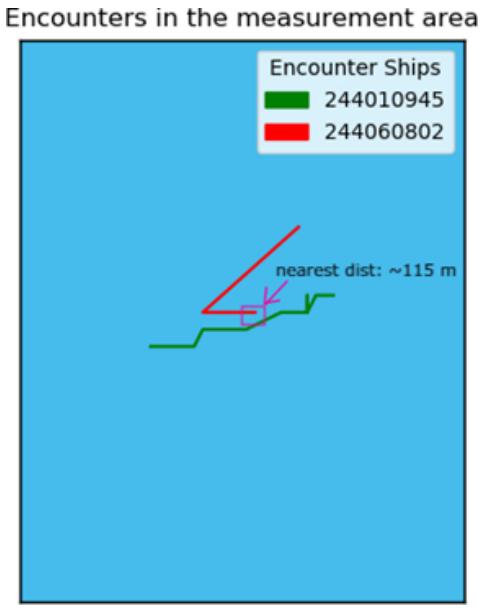


Fig 11: Ships routes and their encounters in measurement area on the same day (closest distance was approx. 115 meters)

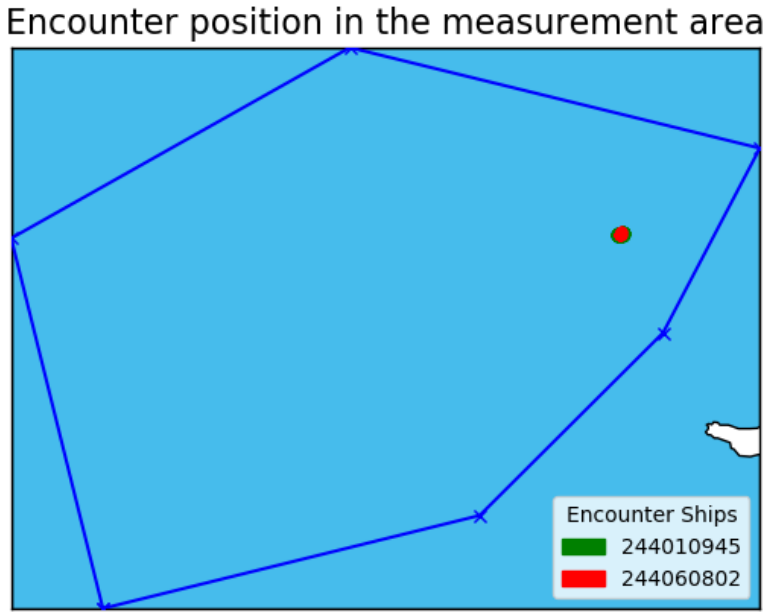


Fig 12: Encounters position for the ships in the measurement area on the same day

4 Results and Evaluations

The primary purpose of this chapter is to evaluate and discuss the results from the experiment, which was done in chapter 3. In chapter 3, there were three different scenarios where the experiment was done by using AIS data. This chapter attempts to elaborate and evaluate those results from chapter 3 and attempt to explain why there were such results and what could be the possible reason for such outcomes. All experiment cases were based on the 14 days data. The results that are evaluated in this chapter are based on the measurement area that was created in chapter 3.

4.1 Evaluation in Missing Data

Out of a total of 777 ships, only 63 ships had missing data for more than 15 minutes in the measurement area for the two weeks. Some of the interesting results obtained from the experiment were that some ship missing data occurred in the same location but not frequently in all those 14 days. While, in another case, the missing data occurrence was less frequently, and that ships were not following the same routes. In another scenario, the frequency of missing data was higher at one location. There were 63 ships and it is difficult to conclude whether that higher frequency of missing data was from one vessel or from multiple vessels without doing further research. In the majority of the ships' cases, there was no missing data.

A. Possible Cause for AIS Missing Data

According to the IMO guidelines, every vessel over 300 tons must have AIS fitted devices and ship must transmit its AIS data continuously. Unless the transmitted data causes risks to the surroundings or is in danger of being exploited, the ship must transmit data. Whether the missing data occurs in the middle of sea or shore regions, it is not considered good to have missing data because the strict IMO guidelines won't allow it. However, there is missing data in the dataset and there could be many possible reasons for it. One of the common causes of AIS missing data is the source of the data itself and the reliability of the AIS data. However, the

reliability of the AIS data and its source has been already discussed in detail in chapter 2. So, in this discussion, it will attempt to evaluate other issues that may have caused the missing data. The cause of missing data can be divided into two main categories, the drawback of AIS and the intentional reasons. Both of these scenarios for the missing data are discussed in detail below.

I. Technical Problems

AIS device is made of a small transponder that is fitted to the vessel, and it uses short wave radio signals to transmit their location data. These data can be received by vessels and to the shore stations which have AIS device fitted. However, there are some technical limitations and issues that may disrupt while transmitting and receiving AIS data. These technical issues and limitations can disrupt AIS broadcasting, and it may result in loss of the AIS data.

The following are some of the limitations and technical issues that may cause for missing data.

1. **Limited Range:** AIS has a limited range of broadcasting data to other vessels and shore stations. On average, AIS devices can transmit their data within the range of 20 – 30 miles [41]. So, if this distance increases, then there may be possible that AIS data may not be received and transmitted properly.
2. **Crowded Vessels Traffic:** AIS devices use a time slot system to transmit their data using a radio frequency. The time slots system ensures that each vessel gets its chance to transmit data. If there are congestions of vessels, then data signals may be lost because of their interference in their time slots [42].
3. **Weather Conditions:** Harsh weather conditions can likely damage the AIS antenna; for example, lightning strikes. In such a situation, it will be unable to transmit data [42].

II. Intentional Issues

Some of the vessels may have removed or switched off their AIS devices to avoid unwanted attention or to conduct illegal activities. This kind of intentional manipulation in AIS devices is done to disrupt and jammed the AIS reception. AIS devices can be easily switched on and off while doing the illicit operations, and this will case disrupt in transmission, and it will be recorded as missing data [43]. There are some extensive researches done on how to detect intentional AIS on-off switching [43]. This research paper uses different kinds of machine learning algorithms approaches like Support Vector Machine (SVM) and the Averaged RSSI Rasters (ARR) to detect such intentional issues which are frequently occurred in marine vessels [43].

4.2 Evaluation in Ship Movement

The purpose of this experiment was to find if there exist such ships where speed varies a lot while moving in the sea. From the experiment, it was found that out of 777 ships, 254 ships had a speed of equal to or less than 0.01 m/s on different occasions within the measurement area. In some of the cases, ships' speed had even changed to 0 m/s for an extended period.

Even though it may consider normal to have some speed variation and this kind of behavior is beneficial for the maritime environment, it raises suspicions if a ship's speed is close to 0 m/s for an extended period. It is difficult to make concrete decisions regarding these issues without doing further research. One of the suspicious questions is why this kind of behavior exists in ships in the middle of the sea. With the help of marinetraffic.com, there was not found any major islands or ports that these ships could have stopped. The evaluation for such behavior is discussed below in two parts: technical reasons and possible illegal activities.

Some of the technical reasons for having such odd behavior in ships are mentioned below.

- **Weather Conditions:** One of the reasons for speed variation can be caused by harsh weather. Terrible weather can affect the movement of ships.

Probably, these ships may have been stopped or change their speed close to 0 for a while to avoid such harsh weather. But some ships had 0 m/s for a longer period. So, this doesn't explain the proper reason for having speed close to 0 m/s for longer period unless there was harsh weather for longer period of time.

- **Economic and Environmental Reasons:** Frequent reductions of the speed of the ships help in saving the costs, especially when trying to optimize the timing of the port approach. Reducing speed helps to reduce on emitting of carbon dioxide (co_2). This kind of behavior is beneficial for the maritime environment and many vessels follow this behavior in the sea to protect the maritime environment [44].
- **Faulty AIS Devices:** Another reason could be that the ships were not stopped in the sea. However, because of the faulty AIS device fitted in the vessel, the AIS device was transmitting the previous location data but not the current location data. As the experiment was done based on the received historical data, the faulty AIS devices made it look like the ships had stopped on numerous occasions. However, in reality, these ships were moving.

Possible Illegal Activities

Another possible reason is that these ships may have stopped or reduced their speed to close to 0 m/s because they might have engaged in some illegal activities there. The data that was used in the experiment does not contain the data from the fishing boats. So, it is hard to determine whether there were any fishing boats nearby. However, during the interview with Admiral Isto Mattila, he had mentioned that there exists a high possibility of smuggling between cargo ships and fishing boats. There may be possibilities that some of these ships which had stopped, were probably doing some illegal activities with the help of fishing boats [45].

However, this extracted 14 days of historical AIS data is not enough to accuse any individual ships of the involvement in illegal activities because of their odd speeds.

The thesis is simply giving this type of suggestion as one of the possibilities and it cannot accuse any particular individual ships on their involvement in illegal activities. It is the responsibility of authorized personnel to investigate and do proper assessments on these issues. Only the authorized personnel can make a proper decision on whether any of these ships were involved in illegal activities or not.

4.3 Evaluation in Encounters of Ships

The primary purpose of this experiment was to find how often two different ships are close to each other when the speed is low (close to 0.01 m/s) and how long these encounters lasted. After applying speed limitation and the nearest distance to considered as the encounter, the algorithm was able to find 7 major encounters in the measurement area. Some of the encounters between ships were for short durations. Whereas, in some cases, the encounters between two different ships occurred multiple times for a longer period.

Table 6 provides an example of a detailed overview of the encounters between different ships, the total duration of the encounters, the minimum encounter distances between them. Some of the encounters found between ships were normal. The encounters that occurred between supply shore ships were normal. The primary purposes of supply offshore ships are to serve in different operation areas such as oil exploration, helping in construction, etc. in the sea [46]. These offshore supply ships carry personnel crews and swaps supplies and necessary equipment with other ships. While doing these activities by offshore ships, there are possibilities of encounters with other ships and these are normal encounters.

However, it may consider abnormal to have an encounter between two different commercial ships in the middle of the sea. One of the odd results was multiple encounters between a tanker and a cargo ship. It is difficult to make a concrete decision whether this type of behavior is normal or odd by analyzing only historical AIS data. However, in this case, all these multiple encounters occurred in just 1 day in 2 weeks and this is an odd behavior. Probably, there was a valid reason that may be resulted in these multiple encounters.

However, the analysis of historical data is not enough to provide whether there was a valid reason or there were involved in illegal activities to have these encounters. The thesis is unable to provide a proper reason for the exact causes of these incidents from the historical data. It is the responsibility of authorities to carry out further research and find the probable cause of such behaviors.

Based on the research, some of the causes of encounters between two different types of commercial ships are mentioned below.

- Authorities may have asked both of the ships to stop there for the inspection. When both of the ships were stopped and being inspected, it looked like there were suspicious encounters between them.
- Other possibilities could be that their next arrival port is busy, and there was no place available for them to park their ships. In that case, these ships were waiting there to find their place in the port.
- The third possibility can be that these ships were waiting for their cargo. This type of cargo can be from fishing boats or other ships. However, if this had happened, then loading unloading in the middle of the sea would raise the possibility of illegal activities. The AIS data does not contain data from fishing boats to validate this argument. In this case, the only option is to assume all aspects of the scenarios. The thesis can merely give this type of argument as a probable cause. The thesis cannot accuse any specific individual ships for their involvement in illegal activities from historical data. It is the responsibility of authorized personnel to investigate it.
- Other possibilities can be smuggling, doing illegal activities, and swapping illicit materials between different ships. However, for these types of accusations, historical data alone is not sufficient. Only from Historical data, it is not enough to make a concrete decision whether there was any encounter incident that was involved in illegal activities. Again, it's a responsibility of authorities to act on this kind of issue and make a proper assessment.

5 Conclusion

There are some researches done to detect the anomalies in the ships in the maritime environment. However, there have been little researches done in detecting the anomalies that are based in the Baltic sea area by using real AIS data. The primary purpose of the thesis was attempting to answer three different primary research questions: missing AIS data, abnormalities in ship's speed, and encounters between different ships. With some difficulties and limitations, the thesis was able to answers all three initial research questions.

The research part of the thesis was completed by using AIS data that was extracted from the Baltic sea for 14 days. From the experiment, it was found that there were some abnormalities in ships in the Baltic region. Approximately 8% of the total ships operated in the Baltic sea had missing AIS data at some stages. In some cases, the missing data were frequent and regular. Even though strict IMO guidelines won't allow for any moving vessels not to transmit their data, there may be other reasons such as technical and intentional to have missing data. From the research, it was also found that the ship's speed varies a lot while moving. Approximately 32 % of the total number of ships' speed had changed up to 0.01 m/s in those 14 days. This kind of variation in speed is beneficial for the maritime environment and it may look normal. However, it may not be true that all of the ships are changing their speeds for the beneficial of the maritime environment. There could be other technical reasons and possibilities of illegal activities. In the third experiment, the research was focused on finding abnormalities in the encounters of the ships. Though, there were 7 major encounters found between different ships, not all of them were in a suspicious nature. The encounters between supply shore ships were normal. Supply offshore ships help in operation areas such as oil exploration, carrying crews, construction, etc. However, one of the suspicious and exciting results was multiple encounters between a tanker and a cargo ship. It is difficult to make a concrete decision without doing further analysis. But, in 14 days, all multiple encounters occurred in just 1 day and this is odd behavior.

All three research in the thesis indicates that there were some odd behaviors in the ships. It is difficult to make a specific conclusion from only historical data for such behaviors without doing further analysis. Ensuring the reliability of AIS data is challenging. In some cases, there is a possibility of having abnormal behaviors in the ships because of not reliable AIS data. However, it cannot be ignored that there are security issues in the maritime environment and authorized personnel should take proper action to mitigate these issues.

A. Possible Future Work and Research Areas

In my opinion, one of the unexpected and exciting results found in this thesis was the abnormalities of the encounters of the different ships. There were some limitations and difficulties while doing the practical part of this thesis. Some of the limitations of this thesis were: all three experiments were done within a small area of the Baltic Sea, the experiments did not cover any abnormalities that may happen in islands, the experiment could not establish the abnormal activities that may have occurred between fishing boats and the ships. Also, there was speed limitation and duration of the encounters while finding encounters between different ships. One of the biggest challenges was the unpredictable movement of the ships. There were 777 ships in the measurement area and the majority of these ships were behaving differently each day. This caused a major problem while trying to find anomalies.

The finding presented from this thesis could be a starting point to do further research on finding the anomalies in the Baltic Sea by widening the areas and other technical aspects. There has been already research on finding the prediction of the ship movement in the Baltic sea area [47]. Further research can be extended to finding the prediction of the ship movement and relate this with abnormalities with ships' behavior. There were some limitations and challenges in the thesis and future research can broaden by addressing those problems mentioned in the thesis.

6 References

[1] Itl.nist.gov. (n.d.). 7.1.6. What are outliers in the data? [online] Available at: https://www.itl.nist.gov/div898/handbook/prc/section1/prc16.htm [Accessed 6 Sep. 2019].

[2] Bansal, R., Gaur, N. and Singh, S. (2016). Outlier Detection: Applications and techniques in Data Mining. 2016 6th International Conference - Cloud System and Big Data Engineering (Confluence) pp. 373–377. Doi: 10.1109/CONFLUENCE.2016.7508146

[3] Toloue, K. and Jahan, M. (2018). Anomalous behavior detection of marine vessels based on Hidden Markov Model. 2018 6th Iranian Joint Congress on Fuzzy and Intelligent Systems (CFIS), pp.10–12. Doi: 10.1109/CFIS.2018.8336611

[4] Chandola, V., Banerjee, A. and Kumar, V. (2009). Anomaly detection. ACM Computing Surveys, 41(3). Doi: 10.1145/1541880.1541882

[5] Gaspar, J., Catumbela, E., Marques, B. and Freitas, A. (2011). A Systematic Review of Outliers Detection Techniques in Medical Data - Preliminary Study. Proceedings of the International Conference on Health Informatics.

[6] Mathsisfun.com. (n.d.). Skewed Data. [online] Available at: https://www.mathsisfun.com/data/skewness.html [Accessed 7 Aug. 2019].

[7] Dhankhad, S., Mohammed, E. and Far, B. (2018). Supervised Machine Learning Algorithms for Credit Card Fraudulent Transaction Detection: A Comparative Study. 2018 IEEE International Conference on Information Reuse and Integration (IRI). Doi: 10.1109/IRI.2018.00025

[8] Tournadre, J. (2014). Anthropogenic pressure on the open ocean: The growth of ship traffic revealed by altimeter data analysis. Geophysical Research Letters, 41(22), pp.7924–7932. Doi:10.1002/2014GL061786

[9] Traficom. (2019). Ulkomaan meriliikenteen tilastot. [online] Available at: https://www.traficom.fi/fi/tilastot/ulkomaan-meriliikenteen-tilastot [Accessed 3 Oct. 2019].

[10] Daranda, A. (2016). A NEURAL NETWORK APPROACH TO PREDICT MARINE TRAFFIC. [online] Pdfs.semanticscholar.org. Available at: https://pdfs.semanticscholar.org/2b06/9a96284380f5bd414f59e5a2593b7add699f.pdf [Accessed 7 Jul. 2019].

[11] Imo.org. (n.d.). Automatic Identification Systems (AIS). [online] Available at: http://www.imo.org/en/OurWork/Safety/Navigation/Pages/AIS.aspx [Accessed 26 Aug. 2019].

[12] Icc-ccs.org. (2019. Privacy and Armed Robbery Against Ships. [online] Available at: https://www.icc-ccs.org/reports/2018_Annual_IMB_Piracy_Report.pdf [Accessed 4 Oct. 2019].

[13] Yle Uutiset. (2012). Suspected Hijacking of Ship in Baltic Sea of Swedish Coast. [online] Available at: https://yle.fi/uutiset/osasto/news/suspected_hijacking_of_ship_in_baltic_sea_of_swedish_coast/5293036 [Accessed 15 Oct. 2019].

[14] Lane, R., Nevell, D., Hayward, S. and Beaney, T. (2010. Maritime anomaly detection and threat assessment. 2010 13th International Conference on Information Fusion, pp.1–8. Doi: 10.1109/ICIF.2010.5711998

[15] Shen, H., Hashimoto, H., Matsuda, A., Taniguchi, Y., Terada, D. and Guo, C. (2019. Automatic collision avoidance of multiple ships based on deep Q-learning. Applied Ocean Research, 86, pp.268–288. Doi: 10.1016/j.apor.2019.02.020.

[16] Digitraffic.fi. (n.d.. Service Overview | Digitraffic - Traffic Management Finland. [online] Available at: https://www.digitraffic.fi/en/service-overview/ [Accessed 16 Jun. 2019].

[17] Balduzzi, M., Pasta, A. and Wilhoit, K. (2014. A security evaluation of AIS automated identification system. Proceedings of the 30th Annual Computer Security Applications Conference on - ACSAC '14, pp.436–445. Doi: 10.1145/2664243.2664257

[18] Natale, F., Gibin, M., Alessandrini, A., Vespe, M. and Paulrud, A. (2015. Mapping Fishing Effort through AIS Data. PLOS ONE, 10(6, p.e0130746. Doi: 10. e0130746. 10.1371/journal.pone.0130746

[19] How AIS works. (2007). [image] Available at: https://en.wikipedia.org/wiki/File:AIS-USCG-Overview.jpg [Accessed 21 Nov. 2019].

[20] Navcen.uscg.gov. (2015). [online] Available at: https://www.navcen.uscg.gov/pdf/ais/references/IMO_A1106_29_Revised_guidelines.pdf [Accessed 24 Jul. 2019].

[21] Solasv.mcga.gov.uk. (n.d.). Solas Chapter V - Annex 17 - Automatic Identification Systems (AIS). [online] Available at: http://solasv.mcga.gov.uk/Annexes/Annex17.htm [Accessed 30 Jul. 2019].

[22] K.E. Tunaley, J. (2013). Utility of Various AIS Messages for Maritime Awareness. [online] London-research-and-development.com. Available at: http://www.london-research-and-development.com/Utility-of-Various-AIS-Messages-for-Maritime-Awareness.pdf [Accessed 23 Aug. 2019].

[23] Ship main dimensions. (2006). [image] Available at: https://commons.wikimedia.org/wiki/File:Ship_main_dimensions.svg [Accessed 16 Nov. 2019].
[24] Ferrara, E., De Meo, P., Fiumara, G. and Baumgartner, R. (2014). Web data extraction, applications and techniques: A survey. Knowledge-Based Systems, 70, pp.301–323. Doi: 10.1016/j.knosys.2014.07.007
[25] Gallois.be. (2015). Ensuring the reliability of AIS data. [online] Available at: https://www.gallois.be/ggmagazine_2015/gg_03_03_2015_106.pdf [Accessed 12 Aug. 2019].
[26] Gps.gov. (2017). GPS.gov: GPS Accuracy. [online] Available at: https://www.gps.gov/systems/gps/performance/accuracy/ [Accessed 30 Aug. 2019].
[27] Vayla.fi. (2015). Terms of use - Finnish Transport Agency. [online] Available at: https://vayla.fi/web/en/open-data/terms-of-use [Accessed 20 Jun. 2019].
[28] Creativecommons.org. (n.d.). Creative Commons — Attribution 4.0 International — CC BY 4.0. [online] Available at: https://creativecommons.org/licenses/by/4.0/ [Accessed 2 Aug. 2019].
[29] Mqtt.org. (n.d.). MQTT. [online] Available at: http://mqtt.org/ [Accessed 5 Aug. 2019].
[30] PyPI. (2019). paho-mqtt. [online] Available at: https://pypi.org/project/paho-mqtt/ [Accessed 3 Sep. 2019].
[31] Docs.python.org. (2019). re — Regular expression operations — Python 3.7.4 documentation. [online] Available at: https://docs.Python.org/3/library/re.html [Accessed 6 Aug. 2019].
[32] Bansal, R., Gaur, N. and Singh, S. (2016). Outlier Detection: Applications and techniques in Data Mining. 2016 6th International Conference - Cloud System and Big Data Engineering (Confluence), pp.375–377. Doi: 10.1109/CONFLUENCE.2016.7508146
[33] Jerez, J., Molina, I., García-Laencina, P., Alba, E., Ribelles, N., Martín, M. and Franco, L. (2010). Missing data imputation using statistical and machine learning methods in a real breast cancer problem. Artificial Intelligence in Medicine, 50(2), pp.105–115. Doi: 10.1016/j.artmed.2010.05.002.
[34] Nguyen, V., Im, N. and Lee, S. (2015). The Interpolation Method for the missing AIS Data of Ship. Journal of Navigation and Port Research, 39(5), pp.377–384. Doi: 10.5394/KINPR.2015.39.5.377.

[35] Shapely.readthedocs.io. (2018). The Shapely User Manual — Shapely 1.6 documentation. [online] Available at: https://shapely.readthedocs.io/en/stable/manual.html [Accessed 12 Jul. 2019].

[36] Van Brummelen, G. and Robert, G. (2013). Heavenly mathematics. Princeton, N.J.: Princeton University Press.

[37] Snyder, John P. (1987). Map projections: A working manual. U.S. Government Printing Office.

[38] Matplotlib.org. (2019). Welcome to the Matplotlib Basemap Toolkit documentation — Basemap Matplotlib Toolkit 1.2.1 documentation. [online] Available at: https://matplotlib.org/basemap/index.html [Accessed 27 Jun. 2019].

[39] Rapp, B. (2017). Finite Difference Method. Microfluidics: Modelling, Mechanics and Mathematics, pp.623–631. Doi: 10.1016/B978-1-4557-3141-1.50030-7.

[40] 3 types of the finite difference method. (2017). [image] Available at: https:// en.wikipedia.org/wiki/Finite_difference#/media/File:Finite_difference_method.svg [Accessed 15 Sep. 2019].

[41] King, A. (2018). Seven things you should know about AIS. [online] MarineTraffic Blog. Available at: https://www.marinetraffic.com/blog/seven-things-know-ais/ [Accessed 3 Sep. 2019].

[42] Maritimeintelligence.informa.com. (2017). Understanding AIS | Maritime Intelligence. [online] Available at: https://maritimeintelligence.informa.com/resources/product-content/understanding-the-automatic-identification-system [Accessed 2 Oct. 2019].

[43] Mazzarella, F., Vespe, M., Alessandrini, A., Tarchi, D., Aulicino, G. and Vollero, A. (2017). A novel anomaly detection approach to identify intentional AIS on-off switching. Expert Systems with Applications, 78, pp.110–123. Doi: 10.1016/j.eswa.2017.02.011.

[44] Seas-at-risk.org. (2019). Seas at Risk - Reduced ship speeds make economic as well as climate sense. [online] Available at: https://seas-at-risk.org/18-shipping/953-reduced-ship-speeds-make-economic-as-well-as-climate-sense.html [Accessed 10 Sep. 2019]. **[45]** Reed, J. (2016). The fishermen convicted of £53m drug deal. [online] BBC News. Available at: https://www.bbc.com/news/uk-36785298 [Accessed 2 Sep. 2019].

[46] Marineinsight.com. (2019). [online] Available at: https://www.marineinsight.com/types-of-ships/what-are-offshore-vessels/ [Accessed 18 Oct. 2019].

[47] Virjonen, P., Nevalainen, P., Pahikkala, T. and Heikkonen, J. (2018). Ship Movement Prediction Using k-NN Method. 2018 Baltic Geodetic Congress (BGC Geomatics), pp.304–309. Doi: 10.1109/BGC-Geomatics.2018.00064

7 Appendix

Sample Source code related to the practical part of this thesis are updated in https://github.com/pradipneupane/master_thesis