IDENTIFICATION OF SHIP ROUTE ANOMALIES ON AIS DATA USING DBSCAN ALGORITHM

GEMİ ROTALARINDAKİ ANOMALİLERİN DBSCAN ALGORİTMASI İLE AIS VERİSİ KULLANILARAK TESPİTİ

İHSAN BUĞRA COŞKUN

ASST. PROF. DR. BERK ANBAROĞLU

Supervisor

Submitted to

Graduate School of Science and Engineering of Hacettepe University

as a Partial Fulfillment to the Requirements

for the Award of the Degree Master of Science

in Geomatics Engineering

2022

ABSTRACT

IDENTIFICATION OF SHIP ROUTE ANOMALIES ON AIS DATA USING DBSCAN ALGORITHM

İhsan Buğra COŞKUN

Master of Science, Department of Geomatics Engineering

Supervisor: Asst. Prof. Dr. Berk ANBAROĞLU

May 2022, 60 pages

Maritime is frequently used for purposes such as trade, fishing and transportation. As the intensity of use increases, the safety of maritime transportation becomes an important issue. For this reason, the positions of the ships are constantly monitored by the AIS system. AIS data includes information such as the location of the ship, the type of ship, speed, and destination port. Although AIS data is aimed to be collected regularly, it may not always be recorded properly due to system errors and GPS errors. In the analyses made, firstly, erroneous data was removed and properly recorded data were obtained. An algorithm has been developed that estimates the journey times of passenger ships with the cleaned data. From the beginning of the journey, the ship is expected to travel the same route at the same time. For this reason, the journey is divided into 10-minute segments. All experiments are applied on segment based in whole thesis. DBSCAN algorithm was used for anomaly detection. The results obtained are called possible anomalies. Possible anomaly situations were examined with weather conditions and the effect of weather on travel times was examined. Weather data is obtained based on location via OpenWeatherMap. The data includes information such as wind speed, temperature, cloud rate. As a result, it has been observed that the weather conditions have an effect on the ship's journey times. It has been determined that 45% of possible anomaly points are due to wind speed. It is aimed that all experiments can be easily repeated by people who will do similar research. For this purpose, a Jupyter Notebook was prepared and shared as open source. All analysis were divided into 10 different functions. By following these steps, it was ensured that the analyses could be repeated.

Keywords: Anomaly Detection, Trajectory Data, AIS, DBSCAN Algorithm, Computational Reproducibility

ÖZET

GEMİ ROTALARINDAKİ ANOMALİLERİN DBSCAN ALGORİTMASI İLE AIS VERİSİ KULLANILARAK TESPİTİ

İhsan Buğra COŞKUN

Yüksek Lisans, Geomatik Mühendisliği Bölümü

Tez Danışmanı: Dr. Öğretim Üyesi Berk ANBAROĞLU

Mayıs 2022, 60 sayfa

Deniz yollarının etkin kullanımı sayesinde balıkçılık, ulaşım, turizm gibi birçok ticari amaç gerçekleştirilebilmektedir. Kullanım yoğunluğu arttıkça, deniz ulaşımının güvenliği de önemli bir konu haline gelmektedir. Bu nedenle gemilerin konumları Automatic Identification System (AIS) sistemi ile sürekli olarak takip edilmektedir. AIS verisinin içerisinde geminin konum, geminin tipi, hızı, varış noktası gibi bilgiler bulunmaktadır. AIS verisi düzenli olarak toplanması hedeflenmesine rağmen sistemsel hatalardan ve GPS hatalarından dolayı her zaman düzgün olarak kayıt edilmeyebilir. Yapılan analizlerde ilk olarak hatalı veriler ayıklanarak düzgün kayıt edilen veriler elde edilmiştir. Temizlenen veriler ile yolcu gemilerinin yolculuk sürelerini belirleyen bir yöntem geliştirilmiştir. Yolculuğun başlangıcından itibaren geminin aynı sürede benzer yolu gitmesi beklenmektedir. Bu nedenle gemilerin yolculuk süreleri 10 dakikalık bölümlere ayrılmıştır. Tez boyunca yapılan bütün deneyler bölüm bazlı olarak uygulanmıştır. Anomali tespiti için DBSCAN algoritması kullanılmıştır. Elde edilen sonuçlara olası anomali adı verilir. Olası anomali durumları hava durumu ile incelenerek yolculuk süreleri üzerinde hava durumu etkisi incelenmiştir. OpenWeatherMap aracılığı ile hava durumu verileri konum bazlı olarak elde edilir. Veri içerisinde rüzgar hızı, sıcaklık, bulut oranı gibi bilgiler bulunmaktadır. Sonuç olarak hava durumunun gemi yolculuk sürelerine etkisi olduğu gözlemlenmiştir. Olası anomali noktalarının %45'i rüzgar hızından dolayı olduğu tespiti edilmiştir. Yapılan tüm deneylerin benzer araştırmaları yapacak kişiler tarafından kolayca tekrarlanabilmesi hedeflenmiştir. Bu amaç ile bir Jupyter Notebook hazırlanarak açık kaynak olarak paylaşılmıştır. Bu aşamaları takip eden bir gönüllü ile analizlerin farklı gemi güzergahlarında başarıyla tekrarlanabildiği görülmüştür.

Keywords: Anomali tespiti, iz verisi, AIS, DBSCAN algoritması, Tekrarlanabilirlik

ACKNOWLEDGMENTS

First of all, I would like to thank my advisor, Asst. Prof. Dr. Berk Anbaroğlu, who guided me through all stages of my master's thesis. In addition, I would like to thank him for his motivational speeches and endless support.

I would like to thank my dear family, my mother Nurhayat Coşkun, my father Mehmet Fahrettin Coşkun, my bigger sister Büşra Selvi and brother-in-law Burak Selvi, my grandmother Nurhayat Coşkun, my grandfather İhsan Coşkun, my aunt Işıl Pamukoğlu and my uncle Şevki Ziya Coşkun who have always stood by me with their endless support, love, and effort.

I would like to my love Deniz Hisarkaya for her endless support and love.

I would like to thank my dear friends and colleagues Kübra Atasoy, Halil Gültekin, Yasin Metehan Ulubınar and Mert Aydoğan, Abdullah Kirli for their endless support.

I would also like to thank Oğuzhan Akgöz. He volunteered for our analysis. His analyses and feedback were very useful.

Special thanks to OpenWeatherMap contributors for share their licence for my thesis and they support me for usage the data in the thesis.

İhsan Buğra COŞKUN

May 2022, Ankara

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

AIS	Automatic Identification System
COG	Course Over Ground
DBMS	Database Management System
DBSCAN	Density Based Spatial Clustering of Applications With Noise
EMPER	European Environmental Report Maritime Transport
EMSA	European Maritime Safety Agency
GCS	Geographic Coordinate System
GIST	Generalized Search Tree
GLONASS	Global Navigation Satellite System
GMM	Gaussian Mixture Model
GNSS	Global Navigation Satellite Systems
GPS	Global Positioning System
LSTM	Long Short-Term Memory
MMSI	Maritime Mobile Service Identity
POSTGRES	PostgreSQL
UNCTAD	United Nations Conference on Trade and Development
IMO	International Maritime Organization

1. INTRODUCTION

Maritime transportation has been a traditional way of transporting people and goods, dating back to 1500 BC [1]. It is used for various purposes including tourism, transportation and trade. With the increase in population, there has been an increased demand to for maritime transportation. According to United Nations Conference on Trade and Development (UNCTAD), maritime transportation accounts for nearly 80% of global trade [2]. The safety and security of maritime transit has become a contentious issue because of this situation.

European Maritime Safety Agency (EMSA) was founded in 2002 to ensure the safety of vessels at sea. According to the European Environmental Report Maritime Transport (EMPER) report published in 2021 [3]:

- 32.09% of accidents are due to loss of control between 2014 and 2020 years,
- 17.21% are due to communication faults,
- 14.70% are due to equipment damage,
- 12.77% are caused by the collision of vessels

The lives of the people on the vessel are in danger if vessels collide. In addition, substantial financial losses occur. Automatic Identification System (AIS) is one of the technological ways to ensure the safety of ships, in which the position of a ship is collected and stored at regular intervals in addition to its speed, direction and heading.

AIS is a system which is collect information of vessels. AIS system belongs to Vessel Traffic Service (VTS). As an open source, the movements of the ships can be monitored [4]. AIS is mandatory for vessels of 300 gross tonnage and upwards engaged on international voyages and cargo ships of 500 gross tonnage and upwards non-international voyages. On the other hand, all passenger ships irrespective of size are required to have AIS [5].

There are some limitations of AIS. First, the vessels outside the specified rules may not be included in the system. Secondly, the position accuracy of AIS message as good as the accuracy of AIS data transmitted. Additionally, the received position does not always referenced to WGS84 datum [6].

The collected messages create a trajectory data for every vessel. A sample AIS message cycle is illustrated in Figure 1.

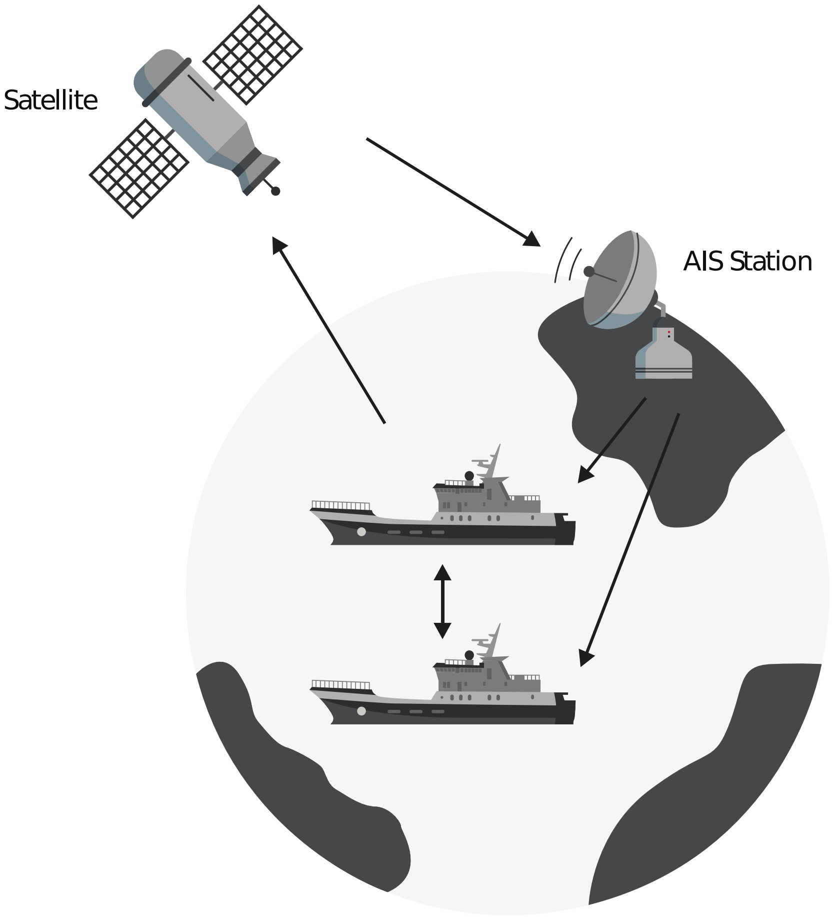

Figure 1.1. AIS message cycle

AIS data is collected with VTS. Additionally, AIS data is a dataset that gives instant information message about vessels. 600 million AIS messages are collected on average every month and it is increasing in time [7]. Every vessel has a Maritime Mobile Service Identity (MMSI) to identify the received AIS message as unique. Additionally, every message contains vessel name, International Maritime Organization (IMO) number, vessel length, type of ship as static data. Dynamic information includes, but not limited to, geographic coordinates (latitude, longitude) of vessel, message time and course over ground (COG). Voyage information include draught, type of cargo, destination, and waypoints.

A study defines anomaly as ‘Anomalies are patterns in data that do not conform to a well-defined notion of normal behaviour’ [8]. It is quite difficult to define what is normal behaviour here. Normal behaviour is usually determined from the dataset. In some cases, anomaly situations are determined on a rule-based and low-likelihood detection [9]. It is stated that it is a suitable approach for situations where there are strict rules. For example, the journey times of the ships and the route they have to go are usually determined. Yet, it is necessary to leave a margin of flexibility, which also adds further complexity as it is difficult to define such a margin. Human factors also play a key role in anomalies. In cases where people have a direct effect, such as ship journeys, the effect on the analysis is examined [10]. The main parameters are, I) distribution of ship track distances, ii) distribution of ships' journeys and iii) distribution of ships' speed. For example, there are studies that observed that human behaviour changes over long distances. Based on the quantity of cigarettes smoked per day by rank, captains and officers smoke more cigarettes than crew members, which may affect their behaviour [11].

Journey time is an important component of safety of ships. Although the ship is sailing on its normal route, extreme situations may occur. Such as, terrorist attack, health problems, ship collision. The earlier detection, higher the chance of intervention. For this reason, anomalies on the ship's journey must be determined. When detecting the anomalies, weather effects should be determined and taken into account.

Weather conditions are not constant and have many variables. For example, wind speed, cloud ratio, temperature and wind direction. Examining the effect of these parameters may require a long-term experimentation process. Therefore, it is important that these experiments are computational reproducibility for researchers doing similar studies. With an open-source software, it is aimed to be an experiment open to development. Software will consist of functions. Parameters allow researchers to research on the desired ship.

1.1. Research Aim and Objectives

This thesis aims to develop a computationally reproducible method for anomaly detection on passenger ship travel times using AIS data. The objectives of the thesis are:

- To store the openly available Danish Maritime data [12] in PostgreSQL (Postgres), which is an open-source relational database management system with outstanding capabilities to process spatial data [13].
- To determine the journey times of a passenger ship. This objective requires a data cleaning step and data on territorial units.

- To compare possible anomalies with weather data to understand whether they could be explained by weather conditions.
- To satisfy computational reproducibility of the whole analysis.

The rest of this thesis consist of six chapters. Second chapter reviews the literature on trajectory data, and anomaly detection methods. Third chapter examines the methodology of the thesis. Fourth chapter provides the experimental results. It relies on the openly available scripts available on GitHub [14]. Additionally, results of anomaly detection in different dates are also discussed. The anomaly results are associated with weather data. Data that could not be explained by weather conditions are referred to as true anomaly (TA). Finally, concluding remarks and future studies are described in chapter six.

2. LITERATURE REVIEW

This chapter reviews the literature on trajectory data. Trajectory is a time-stamped point set generated by moving objects such as people, vehicles or animals. An exemplar trajectory data of a ship is illustrated in Figure 2. Trajectory data represented by multipoint [15].

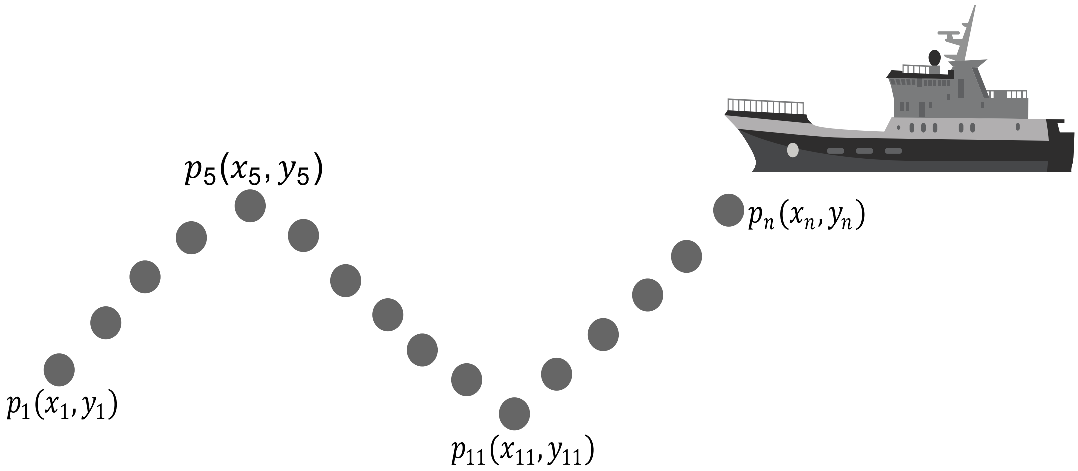

Figure 2.1. A sample visualized of trajectory data

2.1. Research on AIS Data

Research on AIS data has increased in recent years. Especially after 2016, the number of serious studies has increased exponentially. The research results obtained with the keywords AIS, trajectory and ship in Scopus are visualized in Figure 2.22.

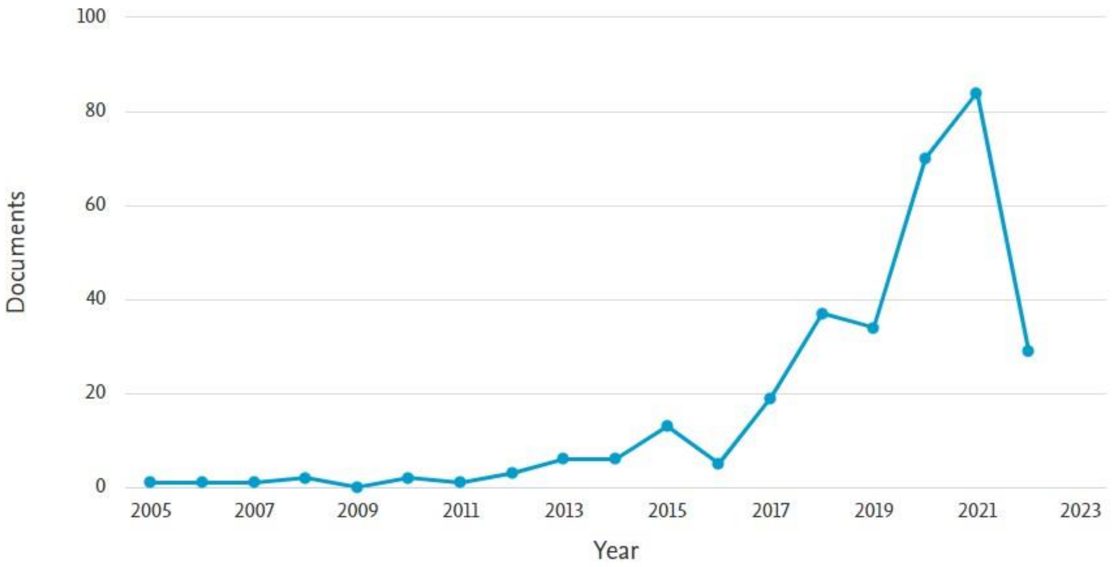

Figure 2.2. SCOPUS – ais, trajectory, ship keywords number of publications per year

Data are collected by Global Navigation Satellite Systems (GNSS) such as Global Positioning System (GPS), Global Navigation Satellite System (GLONASS) etc. One of the important uses of the collected data is to ensure the safety of vessels [16].

Safety is an important issue in maritime navigation. There are many people at different levels in marine management. 85% of safety breaches and shipwrecks are directly related to human error. Therefore, it is important to have people who have completed their education in ship traffic [17]. Human factor in ship journey is an issue that needs to be examined. In a study on this topic decision mechanism of captains are examined on ship collisions over the AIS data. In addition, collision analysis and damage assessment studies were carried out in accidents [18]. AIS data is not continuous data and contains noise. By removing outliers from the data, the ships' trajectory can be estimated using reconstruction approach. It has been found that this approach increases the accuracy of the results in the analysis [19]. Additionally, AIS data is multidimensional. Time and location information are the most important parameters. In a research, 3D graphics were obtained by using location and time [20].

Anomaly detection is not only important for the safety of ships but also for various research purposes ranging from trajectory prediction to ship exhaust emission estimation [21]. AIS data also used in route extraction studies. In such analyses, more sensitive results should be obtained by removing anomalies. The error rate is calculated because of the statistical analysis. The higher the number of anomalies in the data, the higher the error rate. Therefore, it is important to detect anomalies [22].

Different anomaly cases can occur in AIS data. For example, ships waiting in the middle of the sea or in the port should be detected separately. The stopping of the vessels in the port may not be an anomaly, but it can create an anomaly for analysis and should be detected. On the other hand, the location information may come wrong from the source. The latitude value of the AIS message must be between -90 and 90 degrees, while the longitude value must be between -180 and 180 degrees. In another case, AIS messages do not come regularly. Situations where there is no message for a long time may cause erroneous results. This situation should not be ignored throughout the analysis.

2.2. Anomaly Detection

Anomaly detection aims to detect data that does not fit the pattern in a data set. A variety of anomaly detection algorithm have been developed for various uses. There are different techniques for anomaly detection [8]. One of the commonly used techniques rely on classification-based algorithms. This technique is used if the number of classes that will occur as a result of the output is known. In these algorithms, the classes must be defined clearly because, it is forced to assign a class to each data. Another popular technique is to rely on clustering-based algorithms. These are used as unsupervised mode. These algorithms work quite well on complex data, but the computational cost can be $O(n^2)$. On the other hand, supervised algorithms computational cost is lower than unsupervised methods $O(n)$ [23]. An example anomaly is illustrated in Figure 2.3.

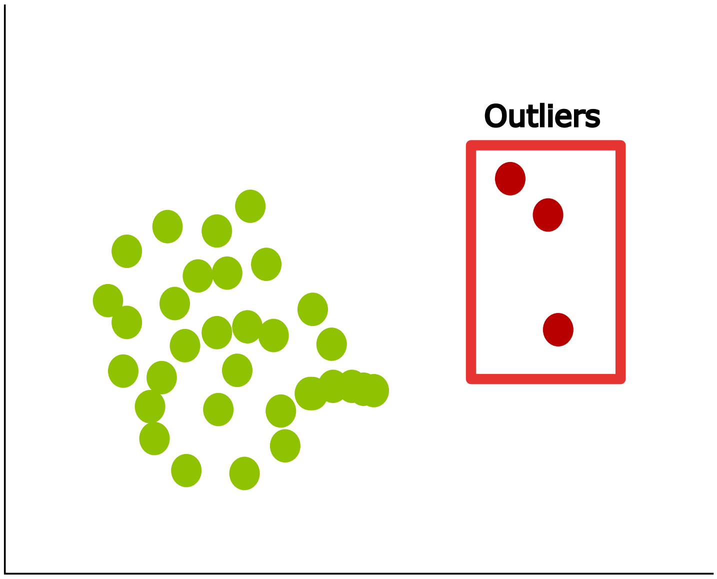

Figure 2.3. Sample anomaly detection

There are three types of anomalies [24]. First one is point anomalies. Anomaly depends on just a single property in data such as, the unusual stopping and waiting of the ship in the middle of the sea. In this case, the anomaly depends only on the speed of the vessel. The second type of anomaly is contextual anomalies. For example, it is not an anomaly for vessels to go slowly in glacial regions, while it is expected to go faster near the equator (in warm regions). The last type of anomaly is collective anomalies. With the melting of glaciers in the world, the number of glacial regions is decreasing. Therefore, it can be expected that vessels will go faster in the same areas.

In order to detect anomaly in the data, it is necessary to identify the normal behaviour. Researchers are working on different methods on AIS data [25]. Methods are divided into two as stochastic-based and machine learning-based. Stochastic-based models have four different categories. Gaussian process is the first method of stochastic models. Every finite linear combination of those random variables has a multivariate normal distribution, such as they are all normally distributed. The joint distribution of all those random variables is the Gaussian process distribution, and as such, it is a distribution over functions having a continuous domain, such as time or space. The second method and most popular method in this category is Gaussian Mixture Model (GMM). It has been observed that this model does not give good results in large areas [26]. Other methods such as kernel density estimation and bayesian network are based on stochastic models [27].

Neural network is one of the mostly used methods in machine learning recently. Another category to extract normal is clustering methods. K-means and density-based spatial clustering of applications with noise (DBSCAN) algorithms are the most popular methods in this category. DBSCAN is used in the Traffic Route Extraction and Anomaly Detection (TREAD) algorithm, which is one of the frequently used anomaly detection algorithms operating on AIS data [9]. K-means algorithm also use for anomaly detection in AIS data [28].

Tread algorithm consists of different managers, in which each manager has a different purpose. Vessel objects manager use for searching and define status of vessels. Whether or not the vessels are on the sailing is determined by their speed. On the other hand, a time threshold is defined for ships with a lack of information. These time threshold values are left to the user’s definition. It is necessary to determine the optimal values. In addition, it is calculated as the probability that the ships will go to which port in each movement of the vessels. There is no need to calculate this probability. It is already clear where the vessel will go. Entry/exit point manager cluster entry points of the vessel. Route object manager detect the anomaly points in the clusters. The final manager is that anomaly detection & route prediction. It is classifying the extracted routes.

A research comparing clustering algorithms identified the best method, which relied on the DBSCAN algorithm [29]. In addition, this research compared based on artificial intelligent method which name is long short-term memory (LSTM) and clustering methods. It has been observed that the DBSCAN algorithm gets better scores in the experimental results using 10000 training and 1000 test data.

Although many studies detect anomalies, the cause of the anomaly is not investigated. Each detected anomaly may not correspond to a real anomalous event. It should be investigated whether the anomalies can be explained. It should not be overlooked that the situations that can be explained will not be a contravention, and that the situation occurred due to an external factor. For example, when investigating ship anomalies, weather is a condition that is often overlooked. If the speed of the ships can be explained by the weather, it can be considered that the situation has occurred due to an external factor.

If the cause is explainable, this is not an anomaly. This situation divides the results into two as possible and true anomaly. The results of anomaly detection should initially be evaluated as a possible anomaly. With subsequent analyses, it should be investigated whether there is a true anomaly. For example, the fact that vessels go slow or fast in different weather conditions or follow a different route does not mean that it is an anomaly. It should be noted, weather conditions are quite variable at sea and ocean. On the other hand, the number and size of other ships on the sea may also affect the speed or route of the vessels. Therefore, possible anomalies should further be investigated.

A study researches that potential anomaly are associated with contextual verification [30]. First, the ship’s normal travel path is found by clustering via DBSCAN. After that the normal path is compared with the instantaneous path for anomaly detection. Thereon, the anomalies are used as input in the contextual verification mechanism. Several weather-related parameters including wind direction, wind speed and wave heights can be used to investigate whether a detected possible anomaly is indeed a true anomaly. If that point can be explained by these external situations, the point is not considered a true anomaly. If the speed of the ship has dropped below the average speed which attribute name is speed over ground (SOG) value and the wind direction is blowing from the opposite direction, it is said that the reason for this is the wind. With this method, it was observed that the number of true anomalies decreased between 40 and 60 percent.

Experiments performed in another overlooked case are not designed to be computational reproducibility. Being able to repeat the same experiments with the reader’s own data increases the continuity of those experiments.

There are several techniques to ensure reproducibility. Basically, the code needs to be easily run by another user. Although there is code sharing for this process, using frameworks such as jupyter-notebook makes the job easier. Another option is to develop a library and share it with users. With more developing technology, Docker is recommended for this process. Docker is a bunch of stage as an assistance item that utilization operating system level virtualization to convey programming in bundles called containers. Especially in web-based systems, it will increase the level of repeatability considerably [31]. Research carried out in the detection of anomalies are listed in Table 2.1. Working with docker containers outside of web systems is costly. The data is first transferred to the container. Then, the code is run with the command line. The outputs to be created must be downloaded back to the local system.

Table 2.1. Anomaly Detection Research

	Method	Study Area	Software	Computational Reproducibility
[32]	Tread Algorithm	N/A	No	✗
[33]	K-NN	Malaysia, Indonesia	N/A	✗
[34]	Minimum Spanning Tree (MST)	Northern Pacific	PostgreSQL	✗
[35]	Recurrent Neural Network (RNN)	Zhoushan Islands	N/A	✗
[36]	MFELCM	World	Python	✗
[37]	Neural Network	World	N/A	✓
[38]	TODDT	N/A	Python	✗
[39]	Artificial Neural Network (ANN)	N/A	N/A	✗
[40]	K-means	Indian Ocean	N/A	✗

3. METHODOLOGY

In this thesis, it is aimed to reduce the number of anomaly points by associating the anomaly results with the weather conditions. It is assumed that the speed of the ships changes according to the weather conditions. In this case, some situations that are perceived as possible anomalies will be explained by weather conditions. The methodology steps are illustrated in Figure 3.1.

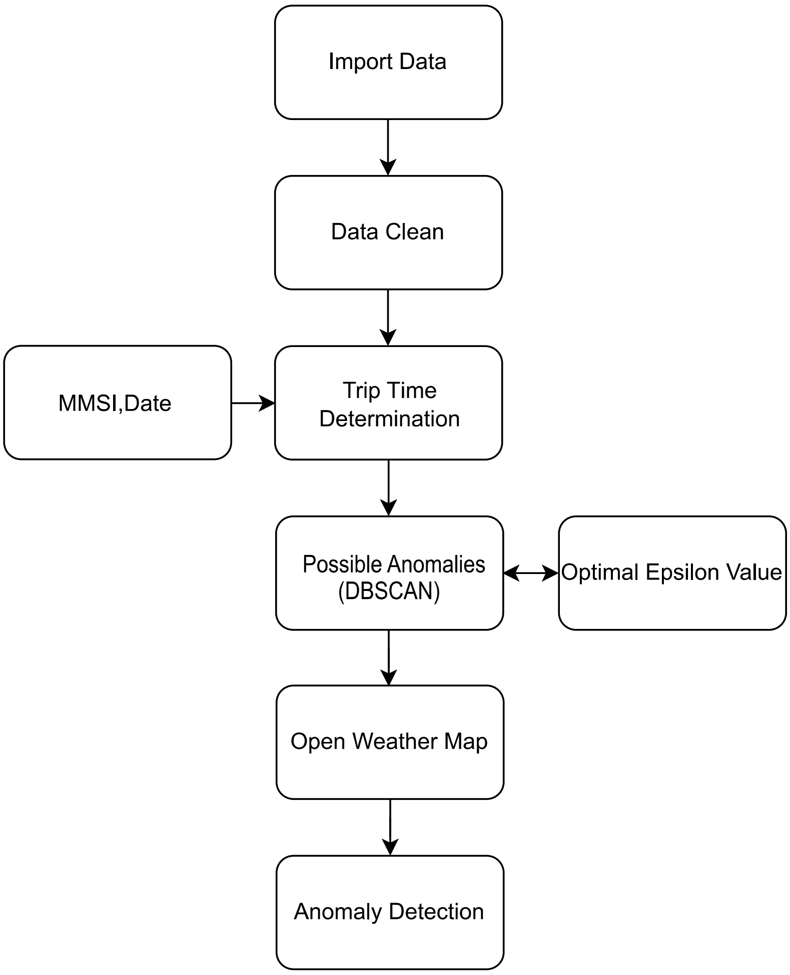

Figure 3.1. Methodology

The methodology consists of seven parts. First, AIS data is imported to a database to store and analyse trajectory data. Second, incorrectly registered data are deleted. The third one is the journey time determination, which estimates the journey time in the selected month. The fourth part of the methodology is the identification of possible anomalies. Possible anomalies are identified by the DBSCAN algorithm. It has two parameters. The first parameter is related to number of journeys. It is expected to see as many points as half of the number of journeys in a cluster. In order to achieve the best result in the DBSCAN algorithm, second parameter, the optimal epsilon value must be found. After that, the effect of traffic on the sea will be observed. Weather data will be analysed to determine the true anomalies, which are possible anomalies that are not related with weather data. Finally, the obtained results will be combined, and it will be investigated whether possible anomalies are real anomalies.

3.1. Import Data

The first step of the methodology is to import AIS data into a relational DataBase Management System (DBMS). Postgres is used as the DBMS as it provides faster results in spatial analysis [41]. Additionally, indexing method is used to speed up queries in databases [42]. While the Generalized Search Tree (GIST) index is used for spatial data in the experiments, B-tree indexing method was used for other data types such as time of the observation.

3.2. Data Cleaning

Incorrect data may occur due to wrong signal, device failure, weather, software error etc. These situations should be detected, and the data should be cleaned. There could be different situations leading to the occurrence of incorrect data.

First situation arises due to locational errors. AIS data are collected as latitude and longitude in the geographic coordinate system. In this case, the latitude value should be between -90 and 90 degrees. The longitude value should be between -180 and 180 degrees. This situation will be checked and the data that is not within these boundaries were eliminated.

The second situation is that GPS data may not be recorded regularly. Vessels with more regular data flow were taken to ensure that the analysis results gave accurate results. Therefore, ships with no more than twenty minutes of data flow between two consecutive recordings were eliminated.

Ships sailing at sea may have different purposes. Some vessels can stop in the middle of the sea while others enter a certain route. To make the best determination in relation to the weather, passenger vessels will be used in this thesis. Because, passenger ships follow a fixed route and departure, and arrival times are generally fixed. It will be more accurate to make analyses on these ships, in this situation.

3.3. Journey Time Determination

Vessels emit signals of their position continuously. The emitted position signals are recorded. At this stage, our aim is to determine the points where the vessels’ speed drop substantially, which indicates that they are close to a port.

The exact stops of the vessels are not known. Because vessels always move on the water due to waves. For these reasons, the stopping point of the vessels will be tried to be determined by speed of the vessels. When ships emit their positions, they also emit their velocity which called as speed over ground (SOG).

The challenge is to estimate the time the vessels leave and approach the ports. The passenger vessels only stop at ports on a normal voyage. Land borders (zones) are needed to determine the regions where vessels approach to a port. Therefore, land borders were imported to the database [43]. There are approximately 100 thousand zones, which are illustrated in Figure 3.2.

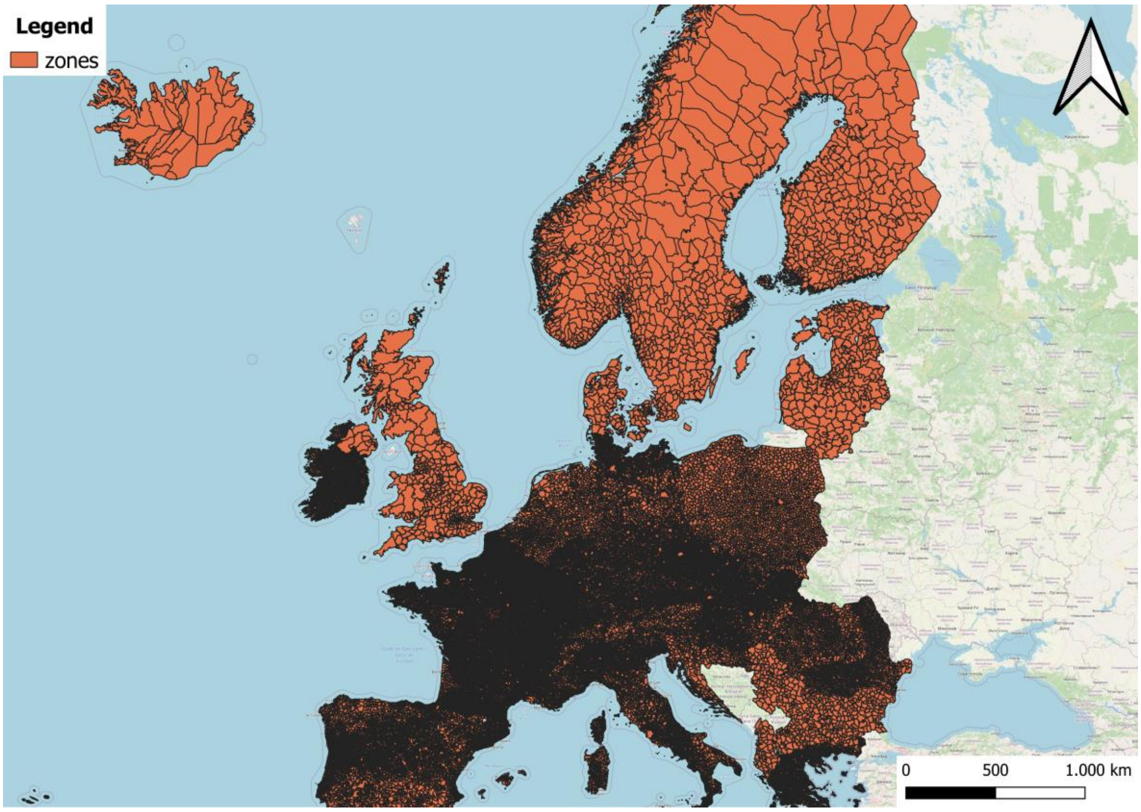

Figure 3.2. Zones of the European Union

Another case is that due to the waves and currents in the sea, SOG value is not measured as zero in general. According to a study, the appropriate SOG value for the ships at rest was taken as 0.5 [44]. As a result, there are two conditions to detect stop points of vessels.

1. The stopping point of the vessels must be within100-meters of the land at most.
2. Vessel SOG value must be lower than 0.5.

3.4. Possible Anomalies

Vessels travel between two ports. While the vessels are traveling between the same two ports, they are expected to go through similar routes. Additionally, the duration of each journey is expected to be close to each other. It is also expected that the ship will follow a similar route in different journeys. Therefore, the journeys are divided into segments (s_i). Segments are created with time information rather than distances. The starting times of the ships are taken into consideration, ignoring the point where the ships start their journey. Segments that are created based on, for example 10-minute duration, are illustrated in Figure 3.3.

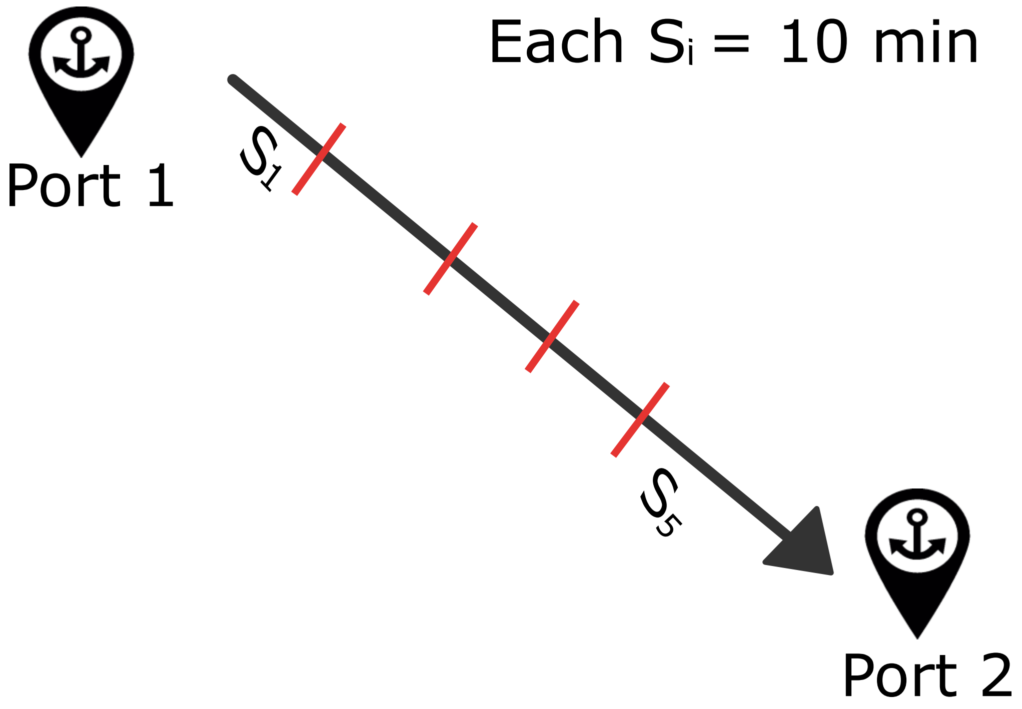

Figure 3.3. Journey Segments

3.4.1. DBSCAN Algorithm

Clustering algorithm can be applied on the same routes of a vessel. DBSCAN algorithm is used in this thesis. Silhouette score is a technique which determines goodness of a clustering algorithm [45]. Score values are change between -1 and 1. It is going from -1 to 1 means that the clustering is accurate. According to a research, the DBSCAN algorithm gave the best clustering results by taking the highest values in the experiments performed according to Silhouette score [46]. It is an advantage that the DBSCAN algorithm does not predetermine how many clusters will be formed. Because we do not have any prior knowledge about the number of clusters.

DBSCAN algorithm takes two parameters. The first parameter (min_p) specifies how many points should be in a set. This parameter will be taken as a calculated constant. It is expected that there will be half of the number of journeys ($np_{vessels}$) in each cluster. Because there are departure and return points in the data. The assumption made in this context is that the path of the trip is similar in both directions, which is usually a realistic assumption. The minimum number of points equation is shown in Equation 3.1.

$$min_p = np_{vessels} \,/\, 2 \tag{3.1}$$

The second parameter is epsilon (ε). This parameter expresses the radius from the investigated point. The most appropriate epsilon parameter should be found according to the data. How to find the most optimal epsilon value is explained in section 3.4.2.

The algorithm consists of four stages.

1. min_p and ε are determined
2. A start point selected randomly
3. Find the points in the ε radius circle inside neighbours of every point
4. Identify the points with more than min_p neighbours.
5. Set as class within inside points

The DBSCAN algorithm separates the points into three different types: core, border and noise. Core points define as at least min_p in its surrounding area with radius ε. Border points defines as if a point is reachable from a core point and there are less than min_p within its surrounding area. An example figure is illustrated in Figure 3.4. If it does not fit both cases, it is called a noise point.

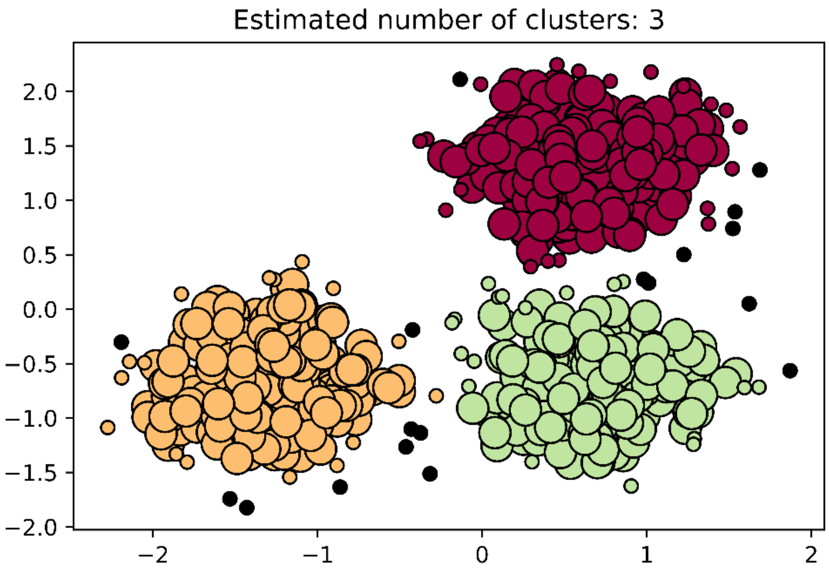

Figure 3.4. A sample figure to DBSCAN Algorithm

The results obtained as a result of clustering will be called possible anomalies. Weather and traffic conditions on a sea can vary substantially. In such cases, the ships can speed up or slow down temporarily. Such situations would validate the possible anomaly, and no true anomaly signal will be output. The optimal epsilon value should be determined in the next part.

3.4.2. Optimal Epsilon Value

DBSCAN algorithm requires an epsilon value. The optimal value varies depending on the data. While larger epsilon values are appropriate for data collected in certain regions, it may be appropriate to use smaller epsilon values for data extending in the form of lines.

Correct determination of Epsilon parameter will affect the anomaly results. The effects of the epsilon parameter were investigated in a sample data set. In a randomly generated data set, epsilon values were taken as 0.1 and 0.3, respectively. The results are shown in Figure 3.5 and Figure 3.

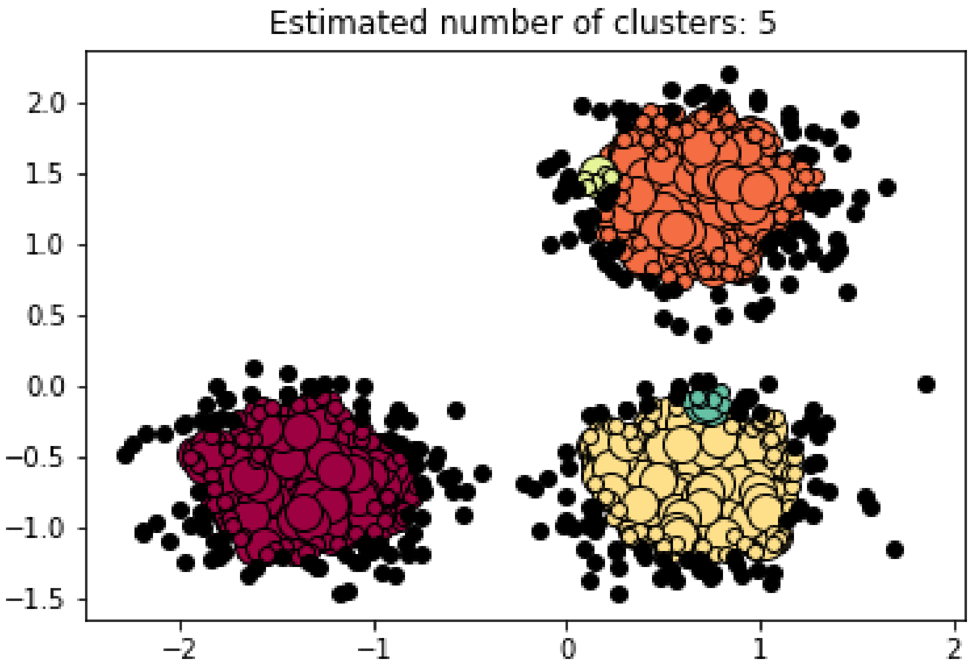

Figure 3.5. Random dataset epsilon value is 0.1

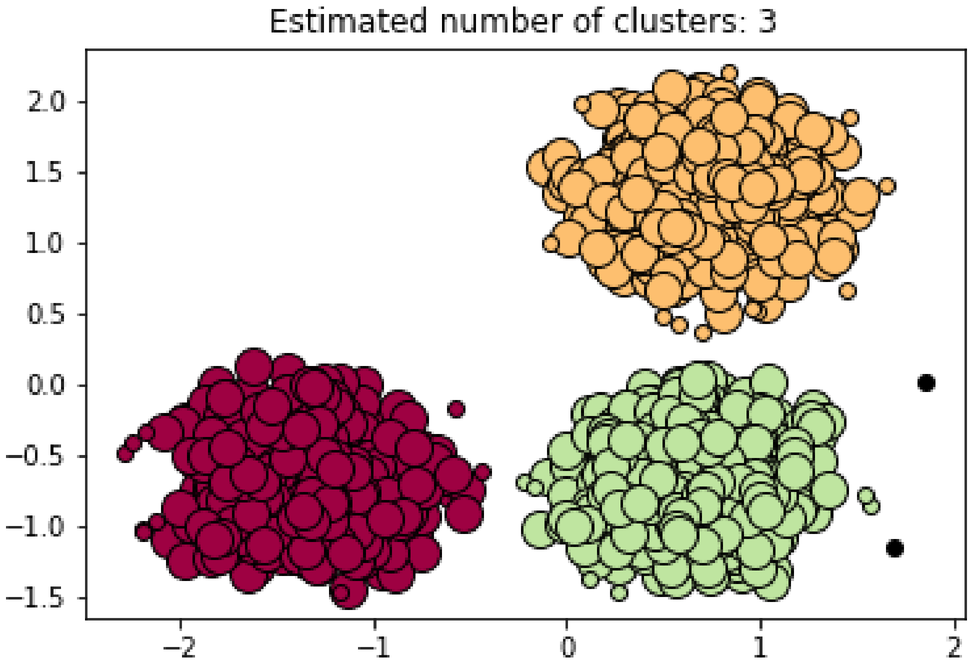

Figure 3 Random dataset epsilon value is 0.3

The change in Epsilon parameter directly affected the number of classes and the number of anomaly points. In the second case, precise results could not be obtained.

The elbow rule uses in cluster analysis to determine parameters in a data set. The rule consists of plotting the parameter as a function and defining the instantaneous change point that occurs [47]. The elbow rule will be used to determine the epsilon value.

Epsilon values will detect the number of anomaly points close to each other at certain intervals. After a value, the number of anomaly points will increase dramatically. This point will determine the most appropriate epsilon value for the data.

Optimal epsilon value was defined according to the elbow rule, which suggests that that instantaneous change point count of possible anomalies will be optimal value for our function. Epsilon value is started from one and the value is reduced by 0.02 at each step, number of anomalies are calculated up to 0.

3.5. Open Weather Map

Open Weather Map (OWM) is a web service tool that shares instant weather conditions. It is not only instant data but also historical data. Data are recorded hourly. So, there are 24 weather records for a day and location.

According to OWM website about data collection '*We collect and process weather data from different sources such as global and local weather models, satellites, radars and a vast network of weather stations. Data is available in JSON, XML, or HTML format'*. It is estimated that interpolation was made for other locations. Interpolation is widely used in weather forecasting, especially over the sea [48].

Weather data during the journey will be collected through this service. The collected data will be recorded and the relationship of the ships with the anomalies of that day will be examined. For example, the speed of ships may vary in rainy or windy weather. In such a case, if it is understood that the ship has slowed down due to the weather, and it will not be considered as a true anomaly. The features that open weather map provides are listed in Table 3.1..

Table 3.1. Open Weather Map Properties

Feature	**Unit**
Date Time	Unix, UTC
Temperature	Kelvin

Feels Like	Accounts for the human perception – Kelvin
Pressure	Atmospheric Pressure – hPa
Humidity	%
Temperature Min	Kelvin
Temperature Max	Kelvin
Sea Level	Atmospheric Pressure at sea level – hPa
Ground Level	Atmospheric Pressure at ground level – hPa
Wind Speed	Meter/sec
Wind Direction	Degree
Clouds	%
Rain	Mm
Snow	Mm

3.6. Anomaly Detection

There are four seasons in a year. Weather conditions change according to the seasons. Even in the same month, week and day, the weather can vary. The weather conditions at sea are different and variable than on land. Changes in water levels resulting from climate change affect sea conditions [49]. The journey times may change depending on weather conditions on the sea.

Possible anomalies may be explained by weather conditions, and hence not be true anomalies. On a rainy day or in foggy weather, it is quite normal for ships to go out of their ways. The anomalies obtained can be removed from being a true anomaly that can be explained by weather conditions.

In this section, the relationship between weather and possible anomaly will be examined. Cases, the reason of which can be explained, will be removed from being an inconsistent situation.

Each segment is analysed in two ways. In other words, it is not important for the analysis from which port the ship leaves. Segment one begins the moment the ship begins a travel and takes a travel on a new segment every, for example, 10 minutes. Although there is no problem in the first segments, the segments begin to overlap, especially in the middle segments. Therefore, periphery points are defined on each segment for a better understanding of the analysis. Periphery points indicate at which point the segment begins and end. Segment overlap example is illustrated in Figure 4.

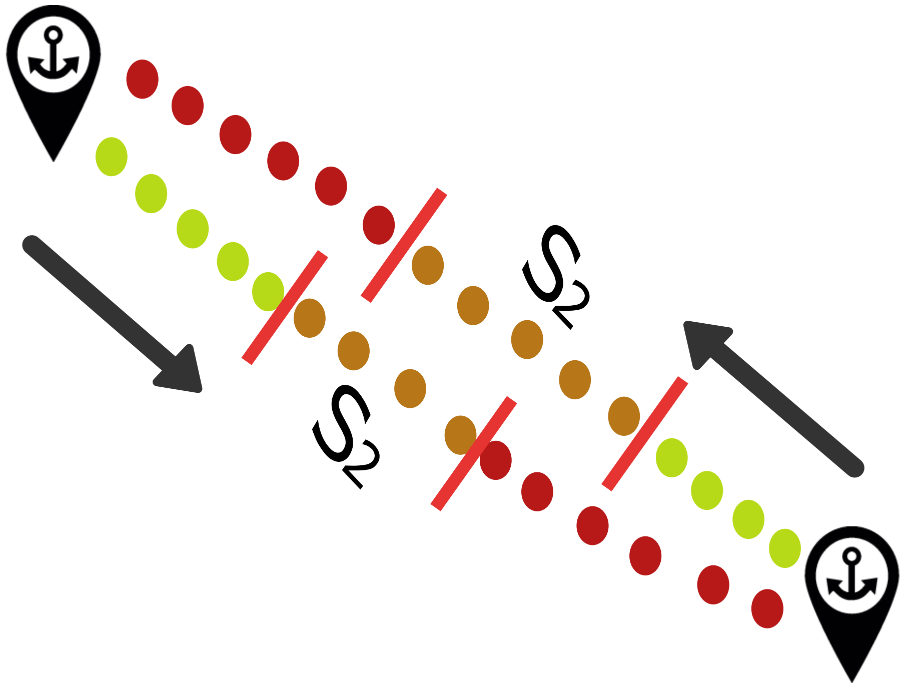

Figure 4 Segment visualization

3.7. Computational Reproducibility

All experiments carried out in this thesis are intended to be reproducible. For this purpose, a Jupyter Notebook has been created and shared on GitHub [50]. It includes all functions to detect anomalies. First, the data set to be detected is downloaded [12]. Land borders should also downloaded [43]. After the data is downloaded the whole process consists of 10 steps. Every step represents a function in jupyter notebook. The studies on the detection of anomalies are summarized in Table 2.

Table 2.2. Jupyter notebook analyse functions

Function Name	Inputs	Outputs	Purpose
connectToDatabase()	Database name, username, password, host, port	Database connection	Connection to database
createTable()	Database connection, table name	Create table result	Creates ship table
importData()	Database connection, table name, file path	Import data result	Imports AIS data to Postgres
startPostgisExtension()	Database connection	Postgis result	It creates a Postgis extension
createGeomColumn()	Database connection, table name	Create geometry column result	Adds geometry column
setGeometryColumn()	Database connection, table name	Geometry column	It creates geometry via latitude and longitude
tripTimeDetermination()	Database connection, output file name, sog value, ship id, zone	Txt File	It creates a file which contains travels start time, end time, number of travels
findAnalysePoints()	Database connection, txt file path, ship MMSI, threshold time, table name	Analyse points	It contains all points to use in the analysis.
findOptimumEpsilonValue()	Analyse points	Epsilon values	Shows a graph according to elbow rule
applyDbscan()	Txt file, analyse points	Csv file	Finds possible anomalies

The software used throughout the experiments and their versions are shown in Table 3.3.

Table 3.3. Software and versions used throughout experiments

Software	Version
PostgreSQL	14.1
Postgis	3.2
Python	3.10.01

Analyses were made in Python. Different libraries were used to make analyses easier. The first library is Numpy adds support for large, multi-dimensional arrays. Complex mathematical functions can be run on these arrays. Another library is Matplotlib plots user-interface graphs using data. Sklearn is a machine learning library for usage in Python as free. It contains classification, regression and clustering algorithms. Other used libraries are available in python by default such as CSV and datetime. The library names and versions used in the analyses are explained in Table 3.4..

Table 3.4. Python library versions

Library	Version
Numpy	1.21.5
Matplotlib	3.5.1
Sklearn	1.0.2
CSV	1.0
Datetime	4.4

4. EXPERIMENTS

The experimental results discuss in this section having four subsections. First, details about the study area provides in section 4.1. Specifically, data describes under three ways, which are I) date of analysis, ii) number of trajectory points, iii) GPS errors. Second, temporal analysis will be determined in section 4.2. After the journey time estimation section, possible anomalies will be defined via DBSCAN algorithm. Finally, real anomalies determine by using traffic data and weather data. All analyses are open source and software scripts are written in Python and shared on GitHub [51]. Fourth, the outcomes of the experiment on computational reproducibility provides.

4.1. Study Area

There are 26 different ship types in the data such as passenger ship, fishing, diving, pilot, medical etc. The ships to be used during the analysis were selected according to two rules. The first rule is that the ship must be a passenger ship. Thus, the purpose of the ship is to transport passengers from one port to another, and do not stop in the middle of the sea. The second rule is that the ship's GPS observations were recorded on a regular basis to realise a sensitive analyses. Specifically, ships with more than 10 minutes between consecutive GPS signals were eliminated. Finally, the vessel that was selected randomly in the had the MMSI ID of 219005068.

The travel takes place in Denmark, which is located on the European continent. Continental and oceanic climates are observed in Denmark [52]. Accordingly, in continental climate regions, winters are rainy and cold, while summers are dry. In the oceanic regions, the weather conditions are observed to be more stable. There is not much temperature difference between summer and winter. As the ship is a cruise ship, it travels between two ports. Name of the two ports travelled are Omo and Stigsnaes. The study area and its center point are visualized in Figure 5.

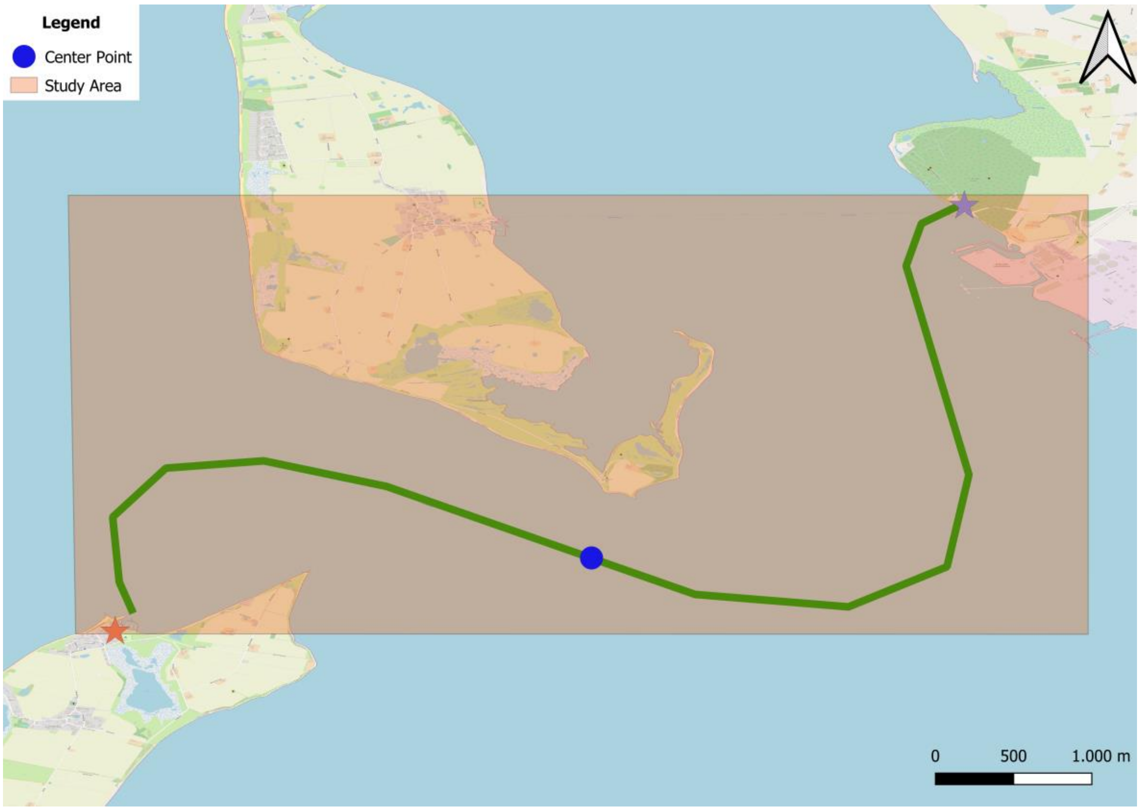

Figure 5 Expected path and study area

Weather data can be obtained hourly and on a precise coordinate. Therefore, weather data were collected from the central point of the study area. The collected data was considered to represent the weather condition of the journey. Analyses were carried out based on the weather data closest to the travel times.

4.2. Temporal Analysis

Weather conditions vary throughout the year in Denmark. While the temperature increases in summer, there is a cloudy air in winter [53]. It is important to monitor the movements of ships in different weather conditions on different dates. In this way, the relationship between weather and journey travel times is better understood. In this section, AIS data on different dates are examined. The number of journeys during the week analysed is shown in Table 4.1..

Table 4.1. Number of journeys in analysed week

Day	December 2020	September 2021
1	14	14
2	14	8

3	0	16
4	14	14
5	13	14
6	16	16
7	10	12

4.2.1. December 2020

The openly available historical weather data can be accessed to last one year. Randomly different dates were chosen from the analyses. Firstly, last week of December 2020 (i.e. 25 – 31 December) data were examined. Therefore, 2020 December data has only been reviewed for possible anomalies.

A total of 81 journeys were made in December. Journey times vary between 40 minutes and 69 minutes with a median of 45.5 minutes. Journey times are visualized in Figure 6.

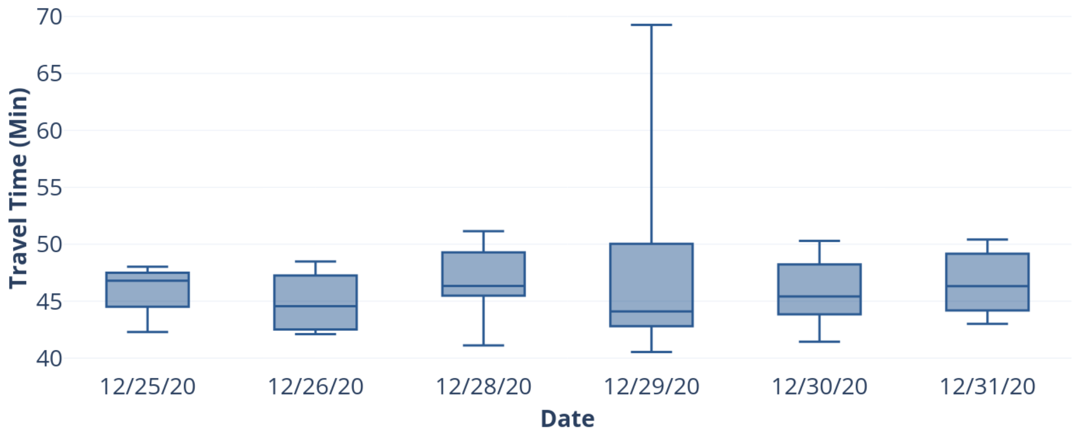

Figure 6 Travel times box plot – last week of December 2020

As can be seen in the figure, there were no journeys on 27 December due to 28 December. On the other hand, it is observed that there was an extremely long journey on 29 December taking approximately 69 minutes. On other days, it can be said that the travel times are closer to each other. It varies between 40 minutes and 52 minutes.

Anomalies will be detected with DBSCAN algorithm in this section. The results obtained with the DBSCAN algorithm will be named as possible anomaly. First, the travel was divided into segments. The chosen travel takes 45 minutes on average. The number of segments formed during the vessels' journeys varies between five and seven.

Since the anomalies will be found with the DBSCAN algorithm, the most appropriate epsilon value should be determined. For this, the elbow rule is used. The resulting graph is visualized in Figure 7.

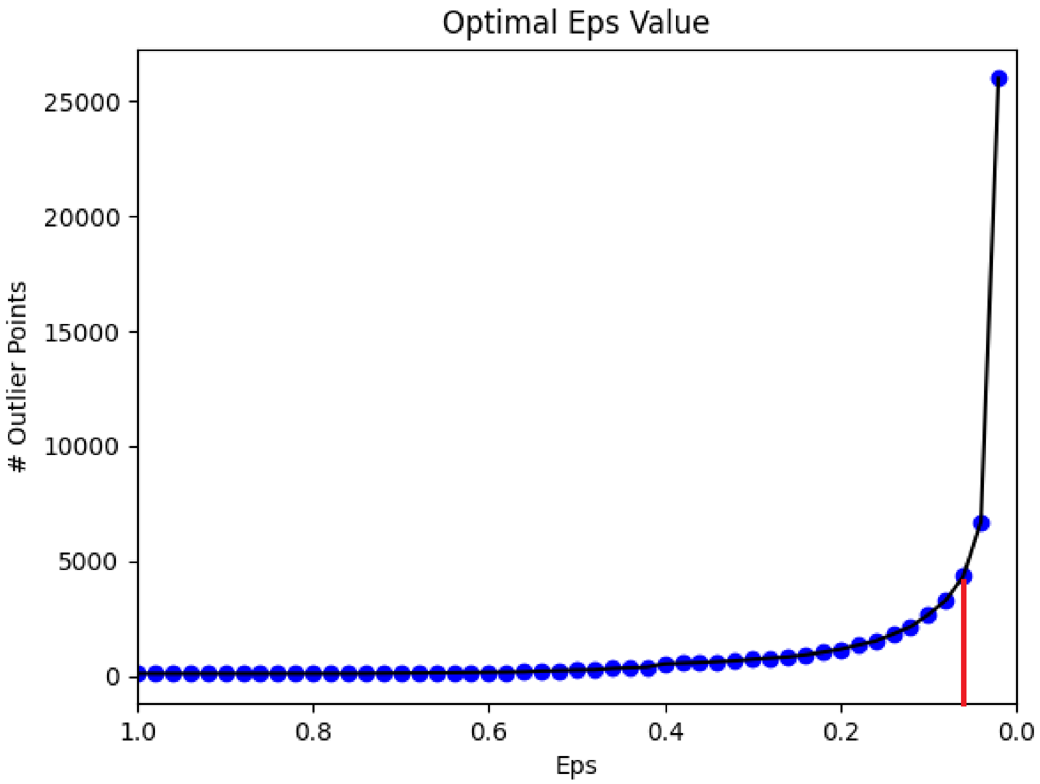

Figure 7 Optimal epsilon values according to elbow rule

According to the graph, the epsilon value was chosen as 0.06. As a result, there are totally 1738 points found as anomaly. The cases where the ship deviated from the route were observed. Possible anomaly points detected on these dates are visualized in Figure 8. Note that some of the possible anomalies are actually very close in proximity to the ship's route, but they probably were recorded because the ship was either slow or fast moving on its itinerary.

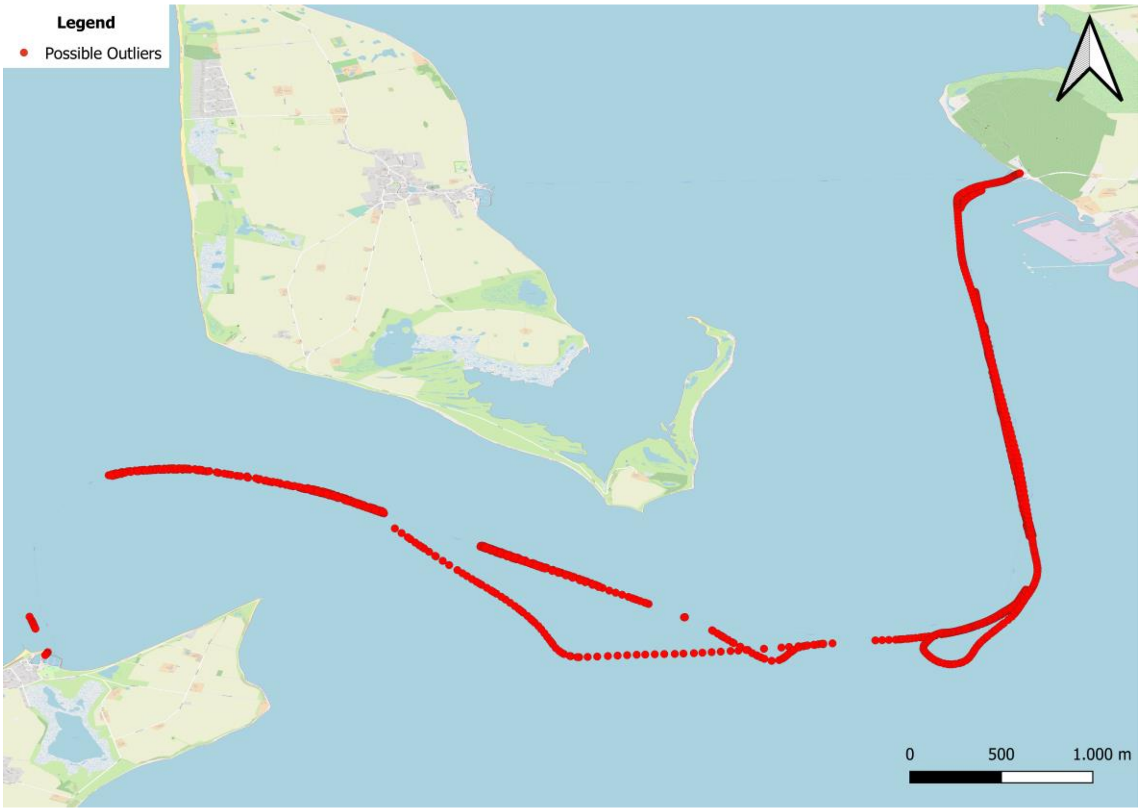

Figure 8 All possible anomalies

In addition, these possible anomalies are classified into seven segments based on the temporal information of the journey as illustrated in Figure 9. The first segment is corresponding to the first 10 minutes of the journeys, the second segment corresponds to the journey time between the 10 and 20 minutes, and so on. Since, the longest journey took 69 minutes, we have 7 segments.

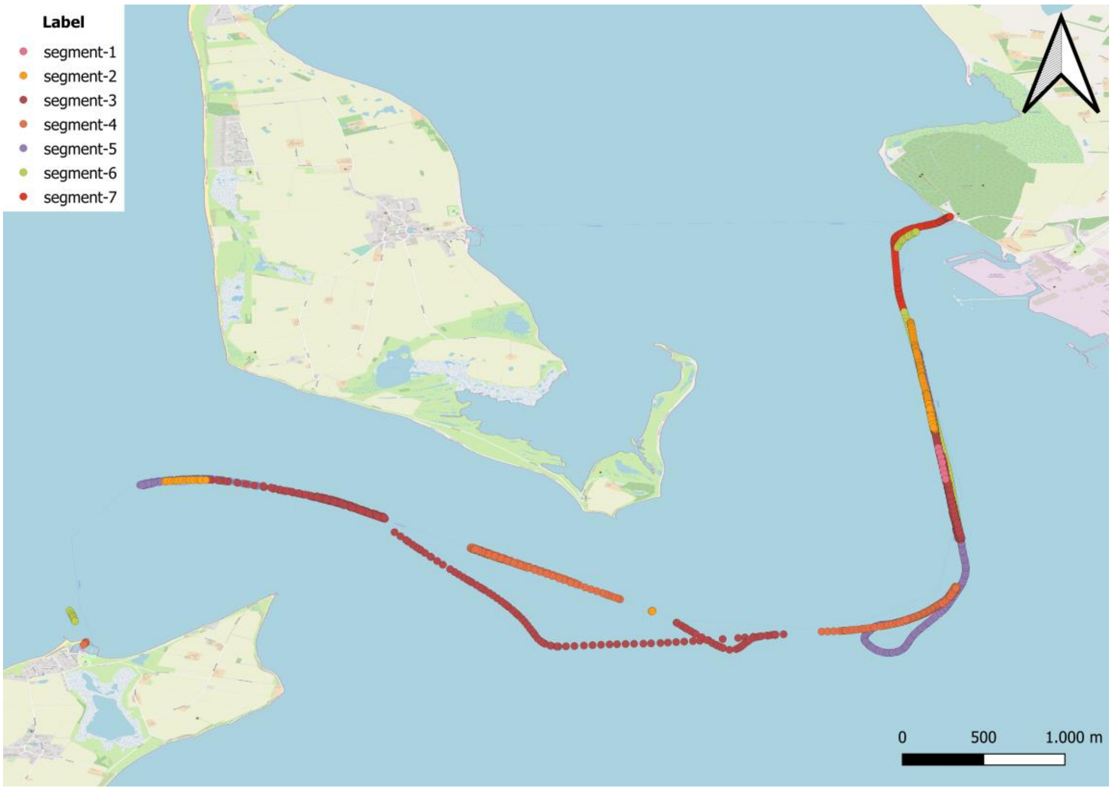

Figure 9 Classification of the possible anomalies based on their segment

An unexpected situation was detected on the journey. It was determined that the ship returned to the port after a while after starting the travel. This may be due to different circumstances. At that time, the COVID epidemic was quite common. On December 29, a quarantine decision was made in Denmark [54]. This decision may have caused the ship to return, or the passengers on the ship may have been diagnosed with illness. Regardless of these, it may have returned due to weather conditions. But the journeys continued this date. Such cases should be called anomalies as the cause is not disclosed. This case is illustrated in Figure 10.

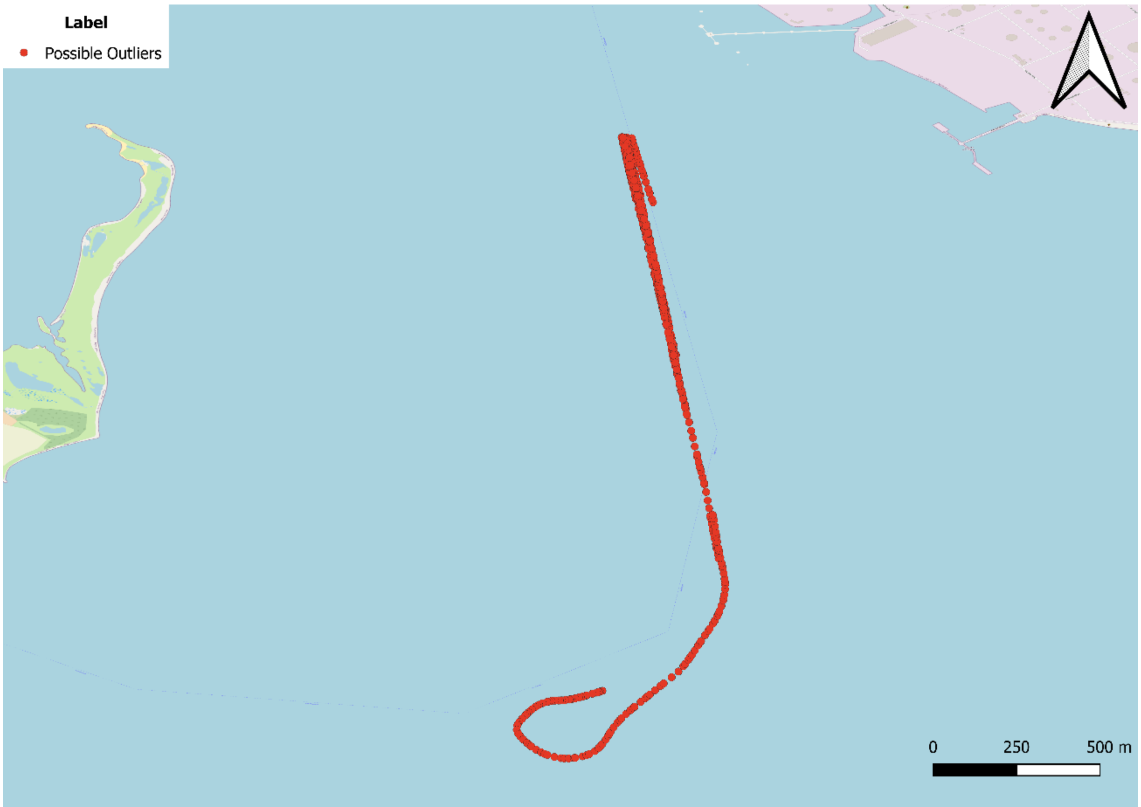

Figure 10 Possible anomaly detection example

4.2.2. September 2021

In this section, the month of September, which is in the autumn months, is examined. Firstly, it was observed how much data was available between the specified dates, which is the last week of September - 22.09.2021 to 29.09.2021. There are approximately 79.5 million records, which is approximately 11.5 million points per day. The latitude value must be between -90 and 90 degrees mathematically. Likewise, longitude values must be between -180 and 180 degrees. The number of points outside these values is searched via SQL. Approximately 1.1 million data were of bounds as illustrated in Table 4.2..

Table 4.2. Information about September 2021 data

	Count
Total Number of Points	79,652,900
Out of Bounds	1,122,039
Type of Vessels	26

The data emitted by the chosen ship is good enough to be considered regular. There are no intermittent data exceeding 7 minutes. Data is recorded regularly with an average of 10 seconds. Deviations from this observation constitute only a marginal ratio of the total collected data points. Regular GPS recordings is important to better relate the possible anomalies with weather conditions. Because the stopping and departure points of the ship should be well detectable. The boxplot illustrating the difference in consecutive GPS observations of the analysed week is shown in Figure 11.

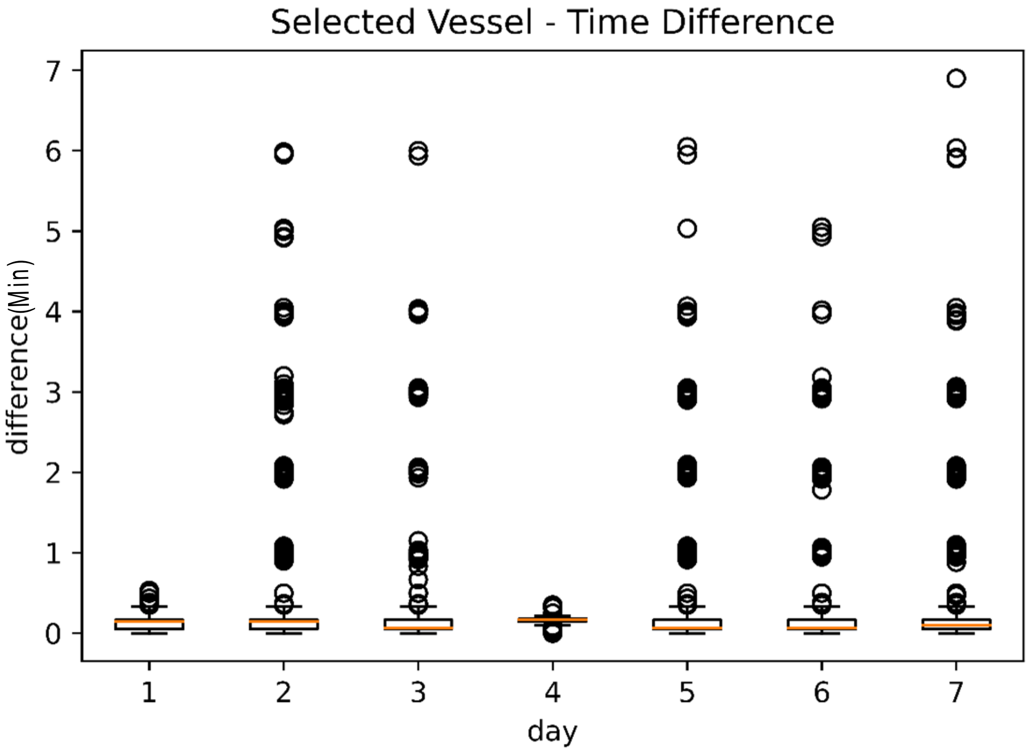

Figure 11 Boxplot illustrating the difference in consecutive GPS observations – September 2021

According to the resulting file, this vessel made 94 journeys in the analysed week. The journey time varied between 40 minutes and 55 minutes. The boxplot illustrating the journey time during the week is illustrated in Figure 12.

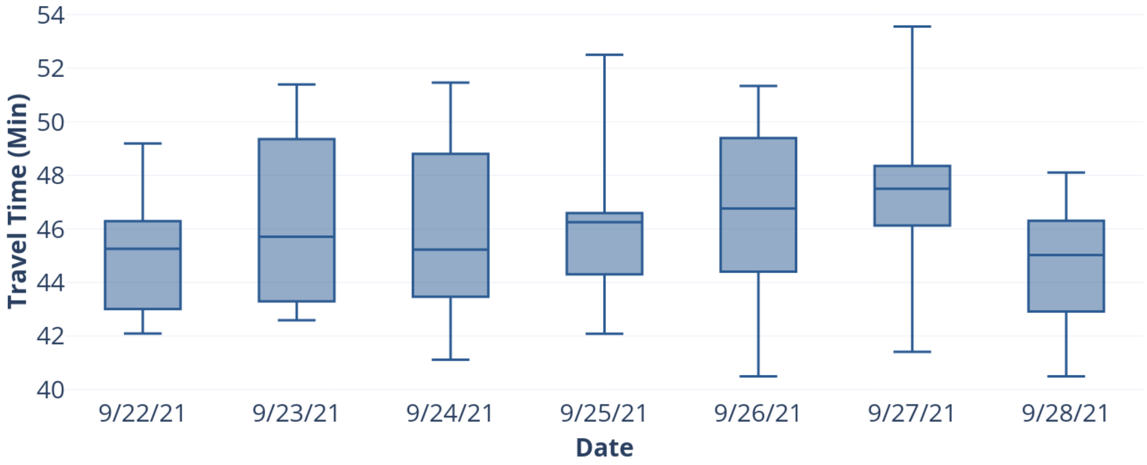

Figure 12 Travel times for September – 22 / 28 September

On September 28, the shortest journey on average was observed with 45 minutes. The longest journey time was 47 minutes on September 28.

Anomalies detect with DBSCAN algorithm in this section. The results obtained with the DBSCAN algorithm. The result points are called as possible anomaly. First, the travel was divided into segments. The chosen travel takes 45 minutes on average. The number of segments formed during the vessels' journeys varies between five and six. Segment parts are illustrated in Figure 13.

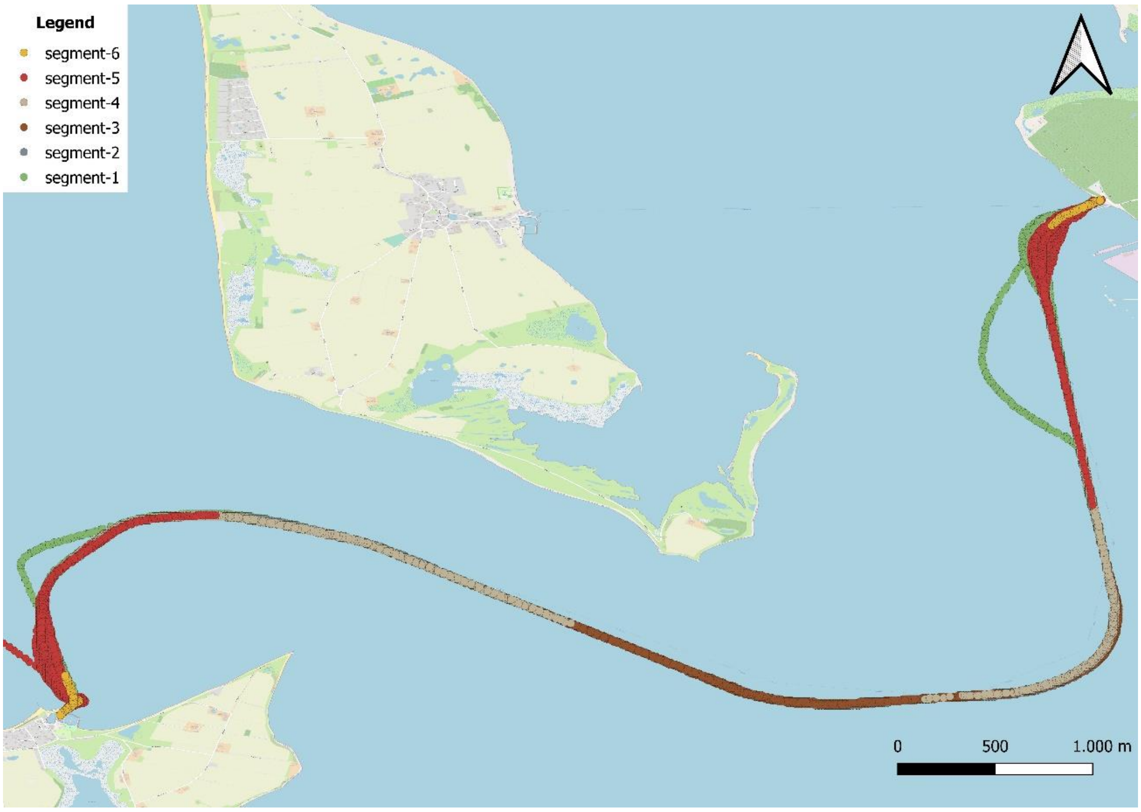

Figure 13 All segments with colors - September

While the first and last segment is the departure/arrival points from the ports, the vessel sails further in the intermediate segments.

The ship is made 94 journeys in the analysis period. Therefore, min_p value is taken 47 in every step. According to the results, the most appropriate epsilon value was found to be 0.06. The optimal epsilon value graph is illustrated in Figure 14.

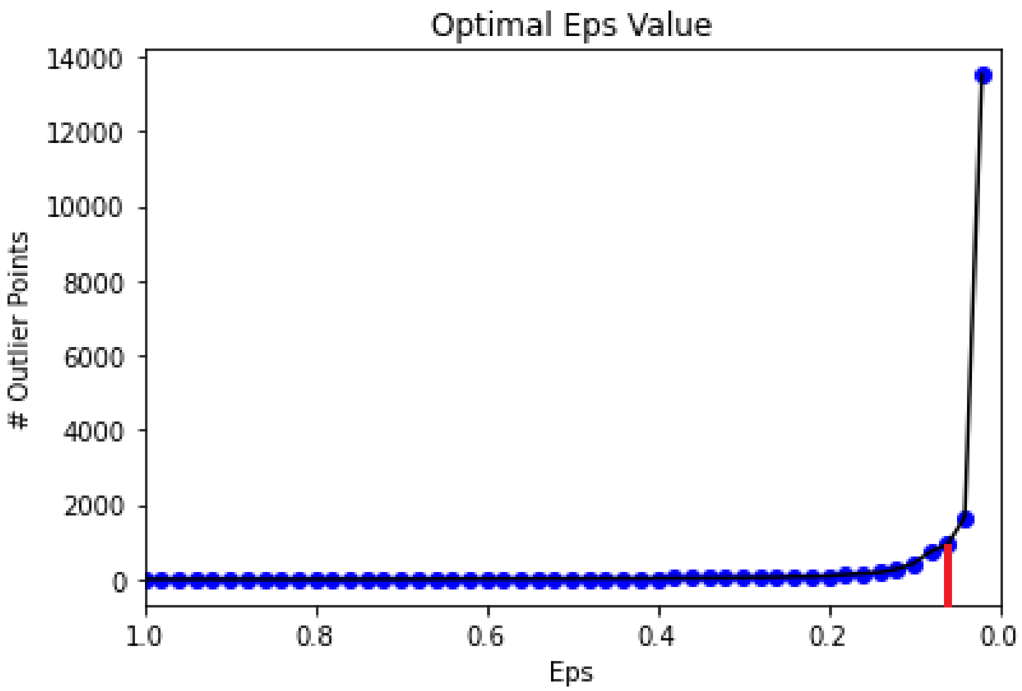

Figure 14 Optimal epsilon value according to Elbow rule

In the second analyse were made in the week of between 22.09.2021 to 29.09.2021. There are totally 1662 points found as anomaly. These anomalies may be due to the vessels path and speed. The possible anomalies are illustrated in Figure 15.

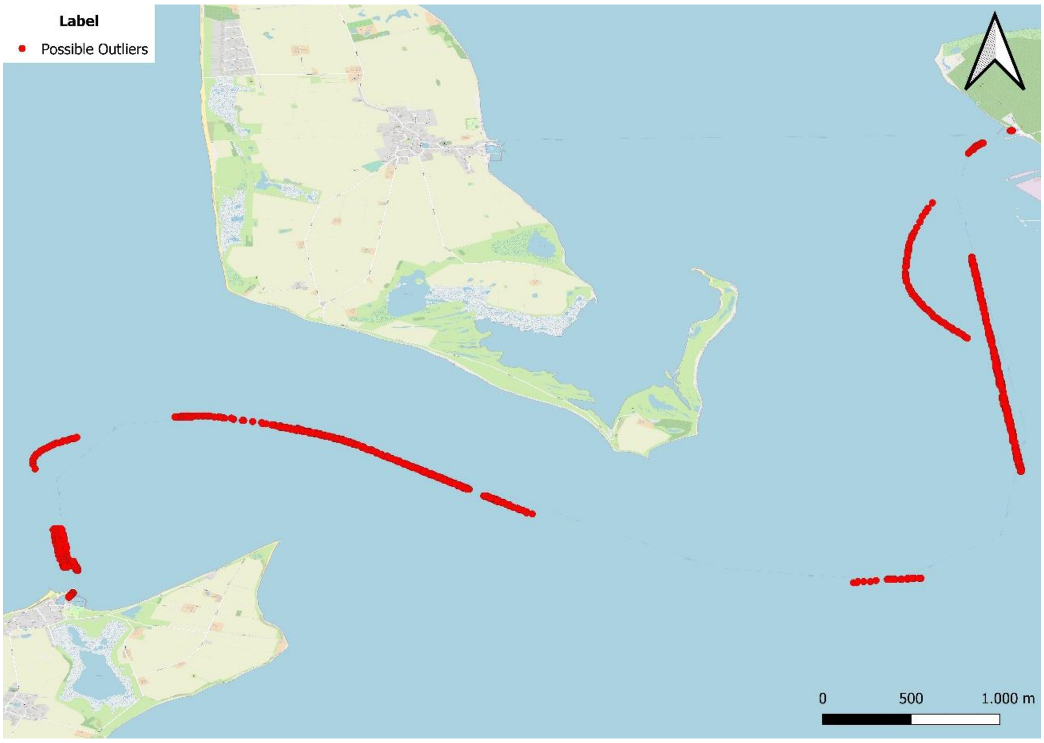

Figure 15 Possible anomalies in September

In addition, possible anomalies are coloured as stages of the cruise journey in Figure 16.

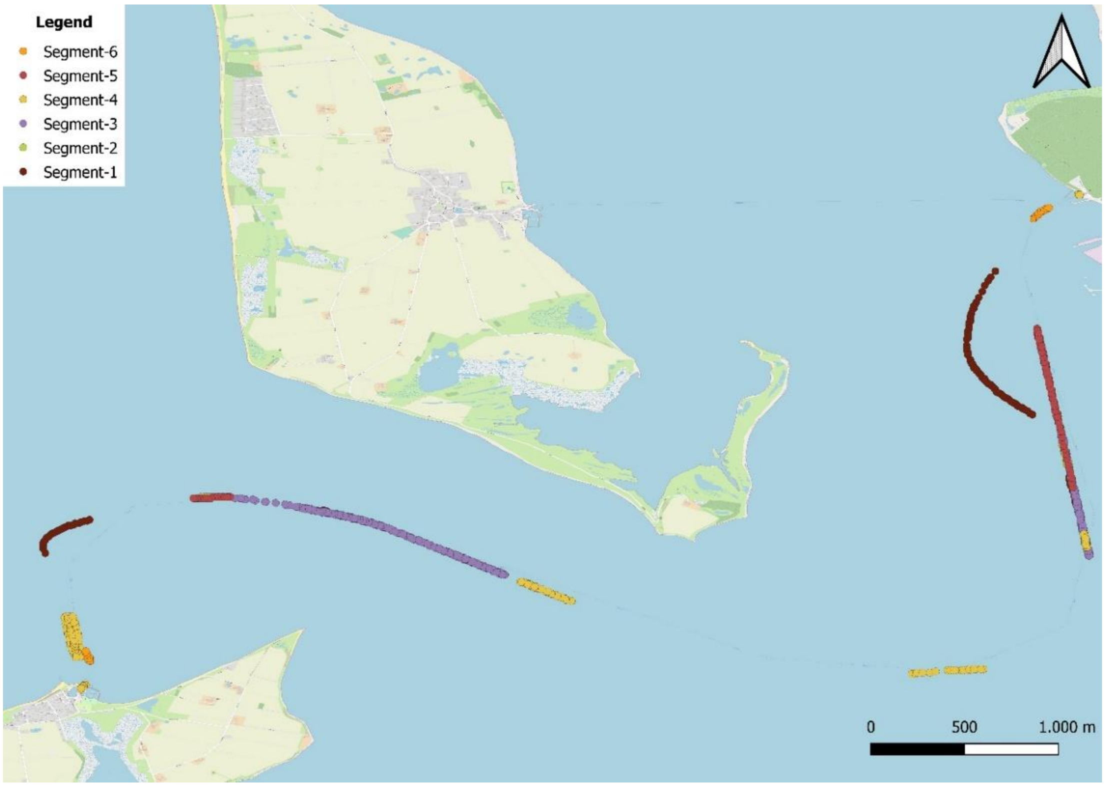

Figure 16 Possible anomalies colored with segments - September

September 2021 data is analysed to correlate with the weather. On the other hand, the windiest week of September was chosen for analysis. The windiest week 22.09.2021 to 29.09.2021 vessel data are used coming from AIS. The wind speed during the analysed week is visualized in Figure 17.

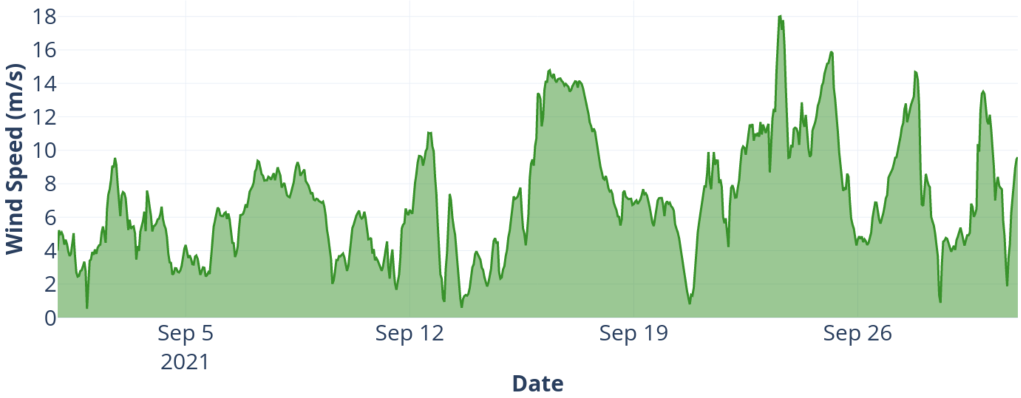

Figure 17 Wind speeds for September

4.3. Anomaly Detection

Conditions that can be explained by weather conditions shall not be called anomalies. If a ship goes off route and it is not a very stormy day, it is hard to explain with weather. In a sunny and clear weather, ships may go fast. Likewise, ships may slow down in rainy and cloudy weather. If these cases called as anomaly by the algorithm, it will be examined. In this section, this hypothesis will be examined its possible anomaly results. Weather data was collected via open weather map [55]. The collected data was recorded in a file. Since the collected data are hourly data, the closest hours to the journeys are considered in the analysis.

The wind speed/date chart for September 2021 is visualized in Figure 18.

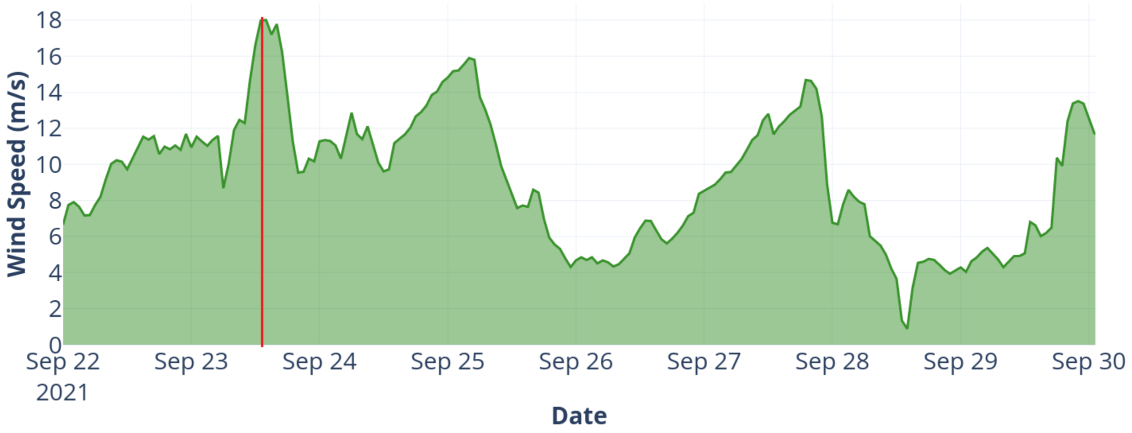

Figure 18 Wind speeds for selected week - September

Wind speed varies between 0 and 19 m/s. While the wind speed changes from day to day, the windiest day was observed on September 23. The highest wind speed is at 13.00. It is staggering to observe that no journeys were made after this time on September 23. Strong wind speed may be a factor. The windless day is September 28 which one m/s. Wind speed varied considerably throughout September. The week with the highest wind variation was between September 22 and September 29. The effect of this change on the movements of the ships has been investigated. Another factor to consider in weather conditions is the temperature data. The temperature/date chart for September 2021 is visualized in Figure 19.

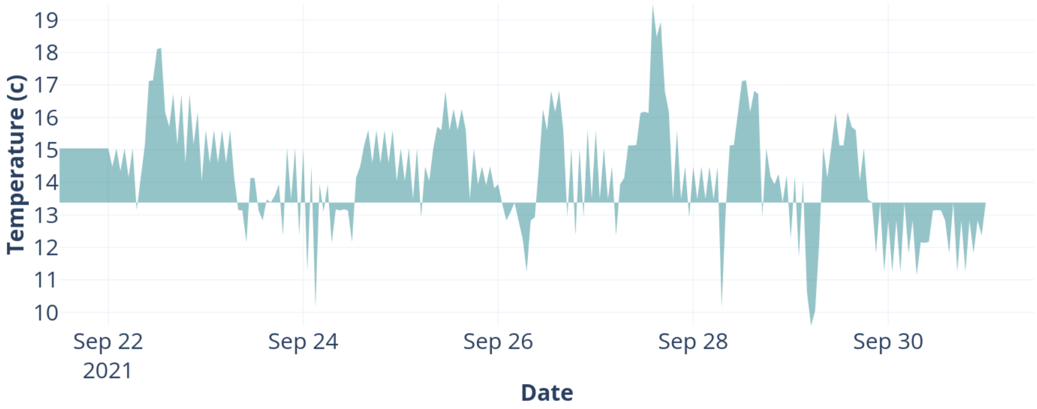

Figure 19 Temperature values for selected week - September

Temperature ranges from 9 ℃ to 20 ℃. The hottest day of the month is September 9-10. The coldest day of the month is September 19. The temperature averaged 13.4 ℃ in September. There is a positive correlation between temperature and wind directly. However, there is a decrease in temperature during the week when the wind speed reaches its maximum. Finally, the weather condition to be examined is the cloud ratio. It can affect the views of the ship captains. Cloud rate is illustrated in Figure 20.

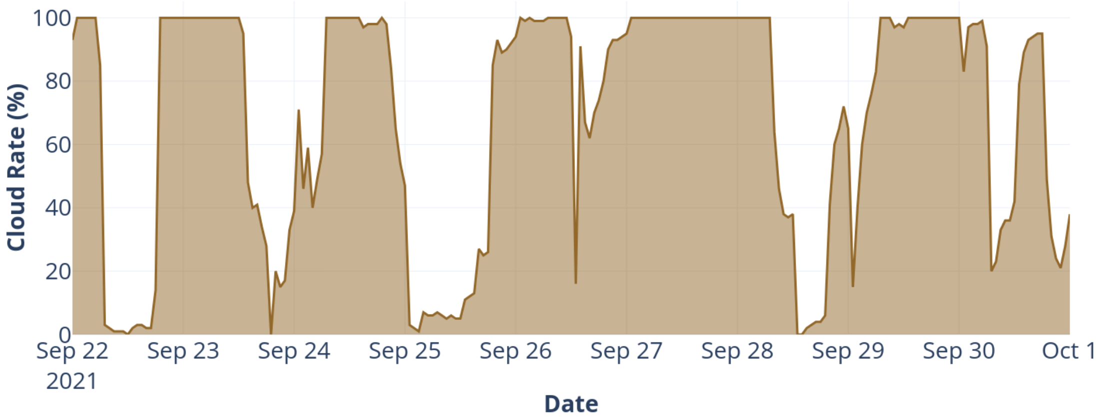

Figure 20 Cloud rates for selected week - September

Cloud coverage is above 95 percent for most of September.

At this stage, the results will be analysed on a segment-based basis. In the first part, the cases where the ship deviates from the route are observed. Segment 1 is visualized in Figure 21.

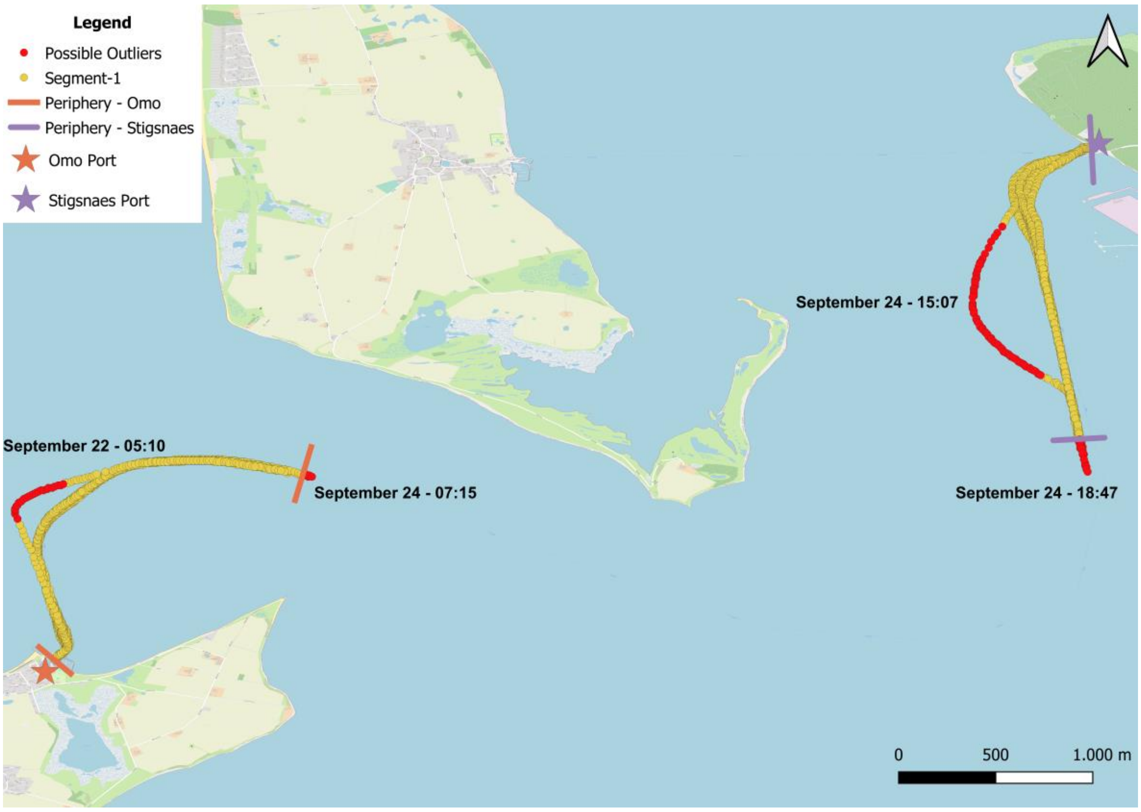

Figure 21 Segment-1 with possible anomalies

Deviations from the route occurred on September 24 at 15:07 and September 22 at 05:10 on journeys dated. Around 15:00 on September 24, the cloud rate is 97% and the wind speed is 11 m/s. Instantaneous changes were not observed. Likewise, around 05:00 on September 22, the cloud rate is 85% and the wind speed is 8 m/s. Weather remains stable. However, sudden changes in the weather are observed in possible anomalies observed on the route. At 18:45 hours on September 24, the wind speed increases from 9 m/s to 13 m/s while cloud rate is 99%. At around 7:15 am, while the cloud rate is 100%, the wind speed continues to be stable at 11 m/s.

Possible anomalies detected in segment 2 are visualized in Figure 22.

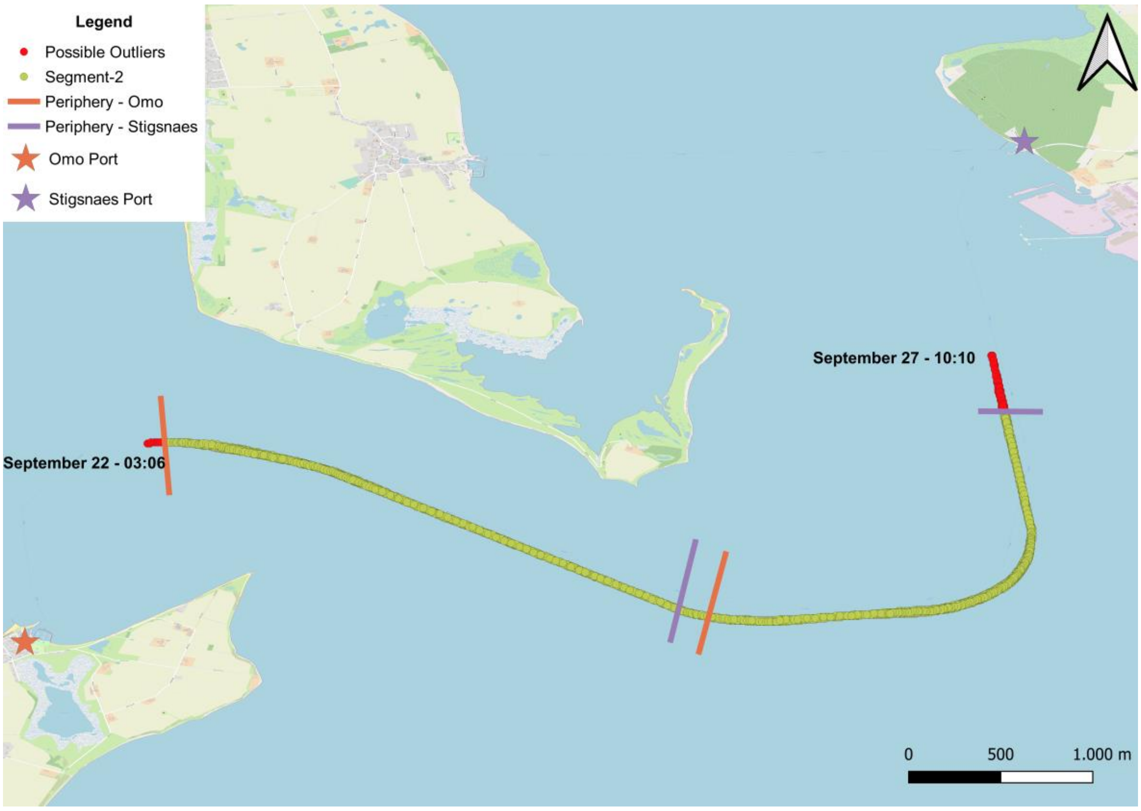

Figure 22 Segment-2 with possible anomalies

Although no deviation from the route was observed in this section, possible anomalies was found. At 03:06 on September 22, the air temperature was 14.5 ℃, the cloud rate was 100% and the wind speed was 7.5 m/s. Weather remains stable. On September 27, at 10:10, the air temperature is 14.5 ℃, the cloud rate is 100% and the wind speed is 11.5 m/s on average. Sudden wind changes are observed here on this date. Wind speed ranges from 10 m/s to 13 m/s. Possible anomalies detected in segment 3 are visualized in Figure 23.

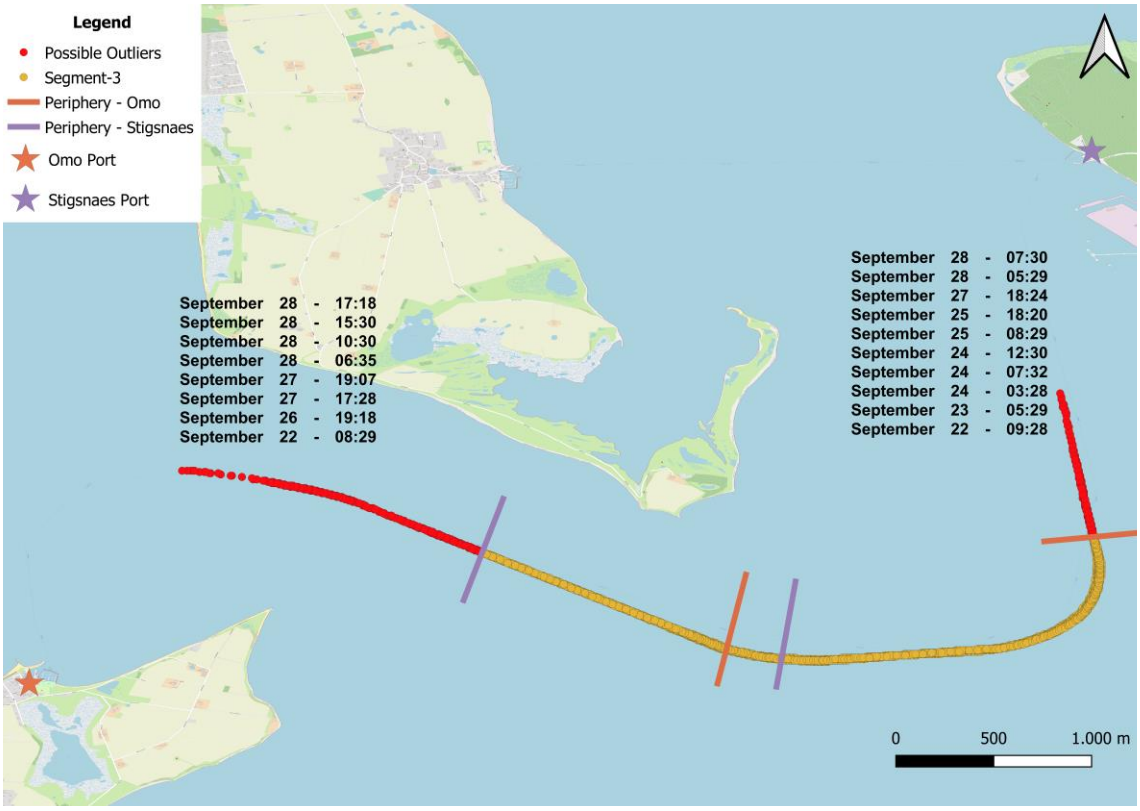

Figure 23 Segment-3 with possible anomalies

In segment 2, anomalies were detected on September 27 and 28, generally. It is observed that the ships coming out of the Stigsnaes port are going fast on September 27 and 28. The ships that started their travel from the port of Omo also went fast on 24-25 September. On September 27, the wind speed started to decrease. The cloud rate is around 99 %. September 28 is the day of the selected week when the wind speed is at the minimum and is cloudless. Wind speed decreases to 1 m/s. The cloud rate also decreases proportionally. In these two days, the weather is constantly changing. On September 24, 25, the wind speed varies between 10 m/s and 16 m/s. The cloud rate starts to rise again on September 25th. Possible anomalies detected in segment 4 are visualized in Figure 24.

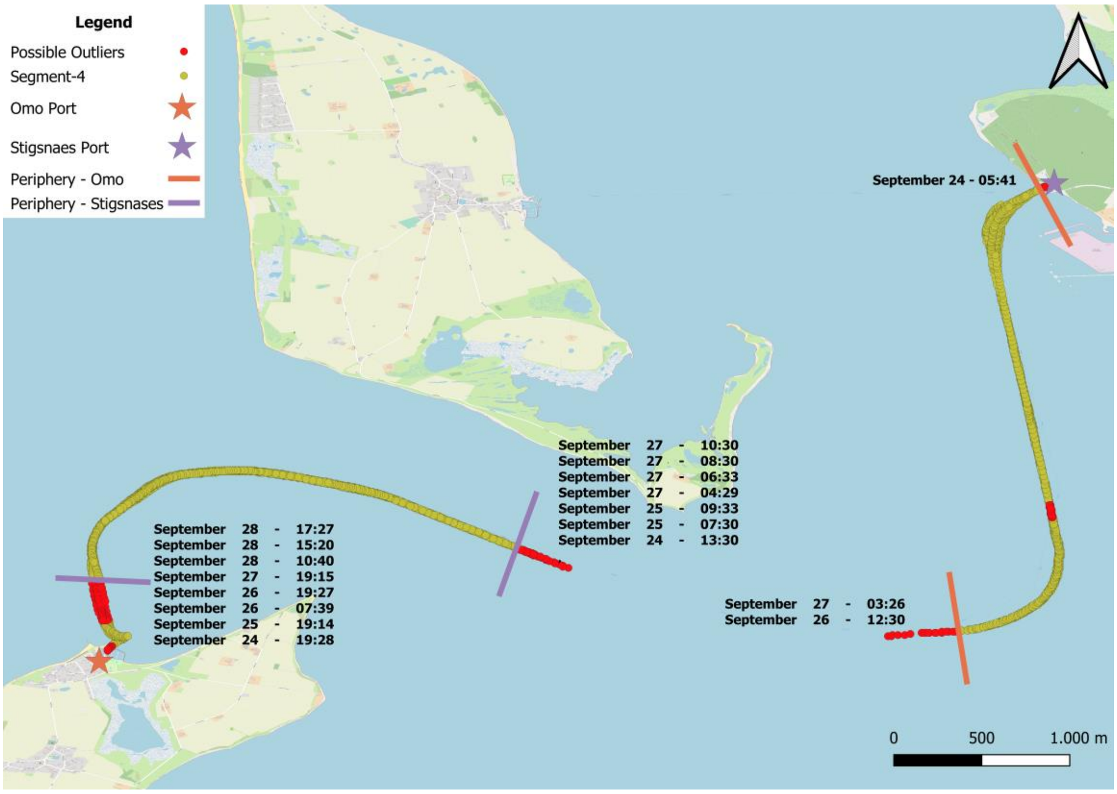

Figure 24 Segment-4 with possible anomalies

Similarly, in segment 3, it is seen that there is a possible anomaly between 27-28 September in general. Ships leaving port went faster on these dates. Likewise, on 26 September, the ship reached the other port faster in two travels. On September 26, the wind speed is on average 5 m/s and the rate of change is less than on other days. Possible anomalies detected in segment 5 are visualized in Figure 25.

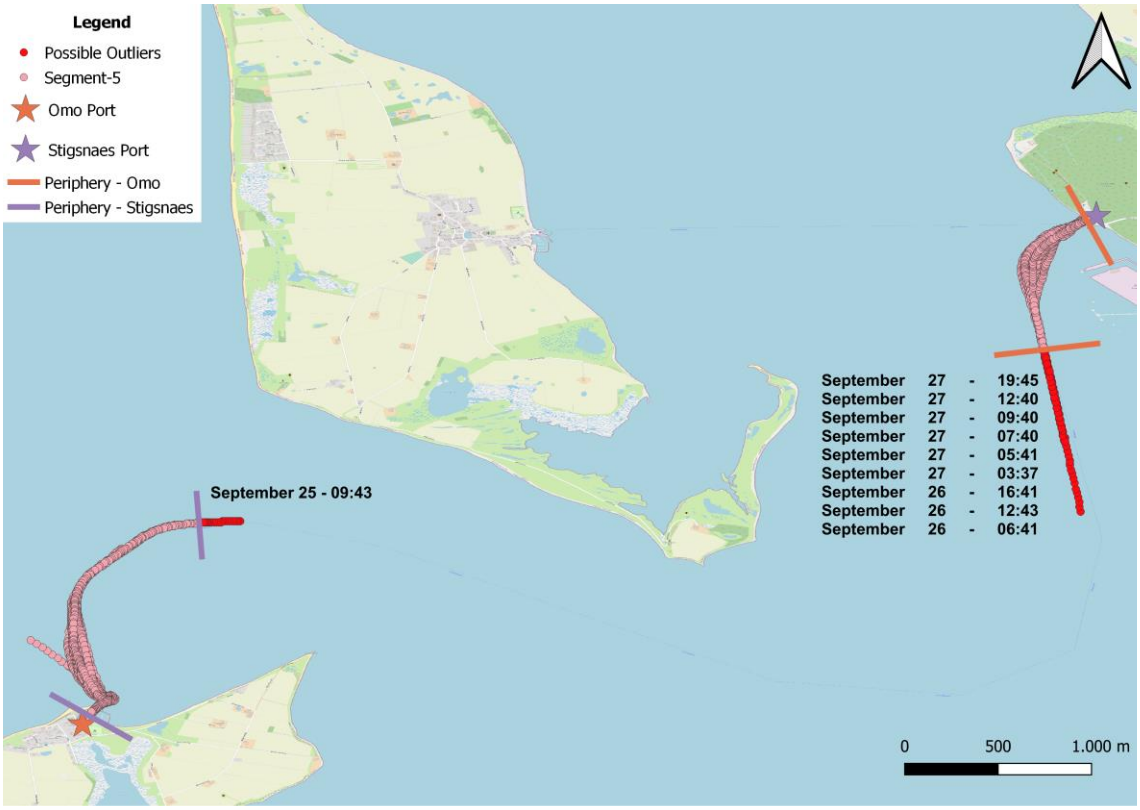

Figure 25 Segment-5 with possible anomalies

Ships leaving the port of Omo on September 26 and 27 reached the destination port more slowly. In the port of Stigsnaes, this slow down occur only on one travel on 25 September. The wind speed at the time of travel is approximately 10 m/s. Possible anomalies detected in segment 6 are visualized in Figure 26.

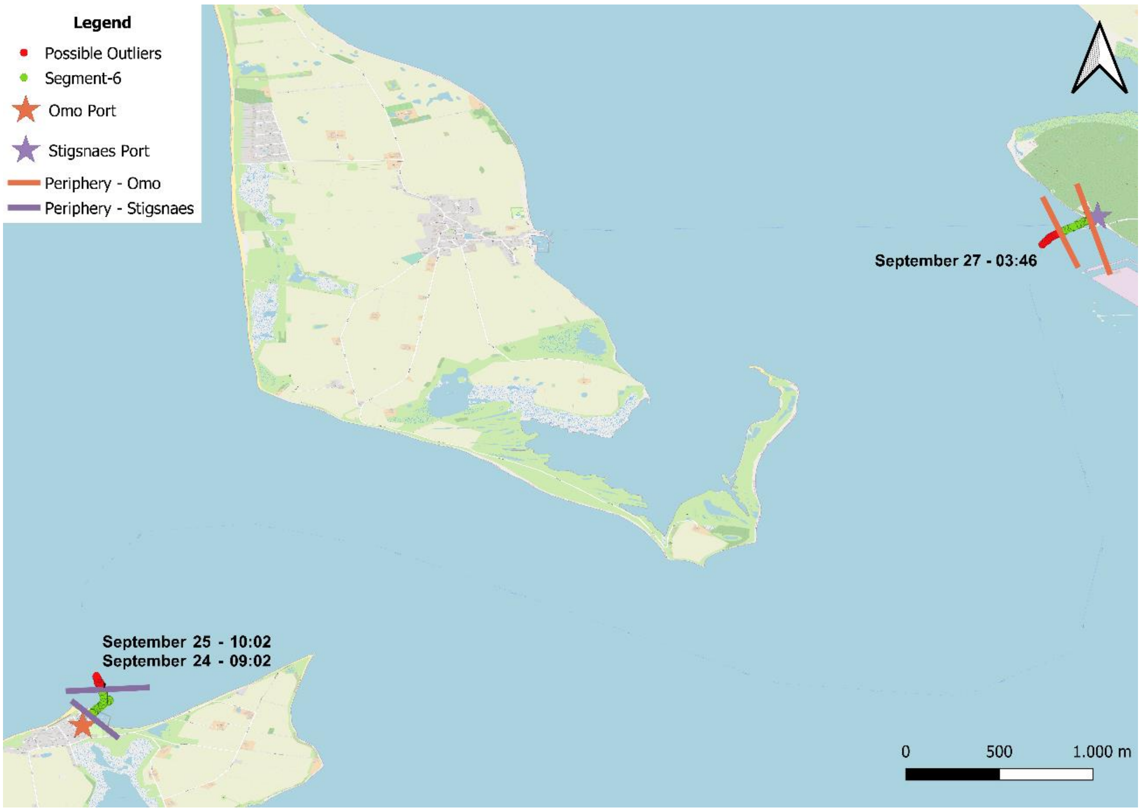

Figure 26 Segment-6 with possible anomalies

The number of journeys over 50 minutes is only nine. According to the number of journeys made, the area where it should has narrowed. In this case, possible anomalies occurred in 3 journeys. It seems that the journey on September 27 at 03:46, departing from the port of Omo, went slower than it should have been. The wind speed at around 4 am on September 27 is 9.5 m/s. While the cloud rate is close to 100%, the air temperature is 15 ℃ on average. On the other hand, possible anomaly values are observed for journeys departing from the port of Stigsnaes on September 24 at 09:02 and on September 25 at 10:02. The average wind speed on these journeys is 10 m/s. Cloud rate is below 20%. The air temperature varies between 13 and 15℃.

When all the obtained results are examined, it is seen that the most possible anomalies were detected on 27-28 September. During these dates, the highest temperature change is observed. It ranges from 10 degrees to 19 degrees. Likewise, the cloud rate drops from 100% to 0%. During the selected week, the biggest changes occur in these two days. Wind speed varies between 1 m/s and 14 m/s. All these rapid changes affect the speed and arrival times of the ships.

It has been observed that there is a correlation between wind speed and travel times throughout the experiments. Wind speeds greater than 21 knots are threat for ship journeys [56]. Threat value is assumed to be approximately 10 m/s. The wind speed to be considered may vary depending on the region. In this thesis, more than 10 m/s of possible anomaly results can be eliminated. The remaining results are called as true anomaly. Segment-based results are shown in Table 4.3..

Table 4.3. True anomaly results

Segment Id	Possible Anomaly	True Anomaly
1	143	36
2	213	108
3	746	384
4	351	224
5	179	136
6	30	16
Total	1662	904

It has been determined that 45% of the points detected as possible anomalies occur depending on the wind speed.

4.4. Computational Reproducibility Experiment

It is aimed that the experiments carried out throughout the thesis are reproducible so that those who want to do the same analysis can easily repeat it. Analysis can be extended by contributing to its open-source development. Experiments made throughout the thesis were repeated with a volunteer Master of Science student. The student performed the analysis by following the Jupyter notebook document. Additionally, two different ships and two different dates. Ships have MMSI id of 219000407. The dates are August 2021 and September 2021 of days between 22-29.

MMSI id 219000407 was investigated as the first ship. The ship is a passenger ship. The cruise is between the ports of Frederikshavn and Vestero ports. The ship's route between the two ports is visualized in Figure 27.

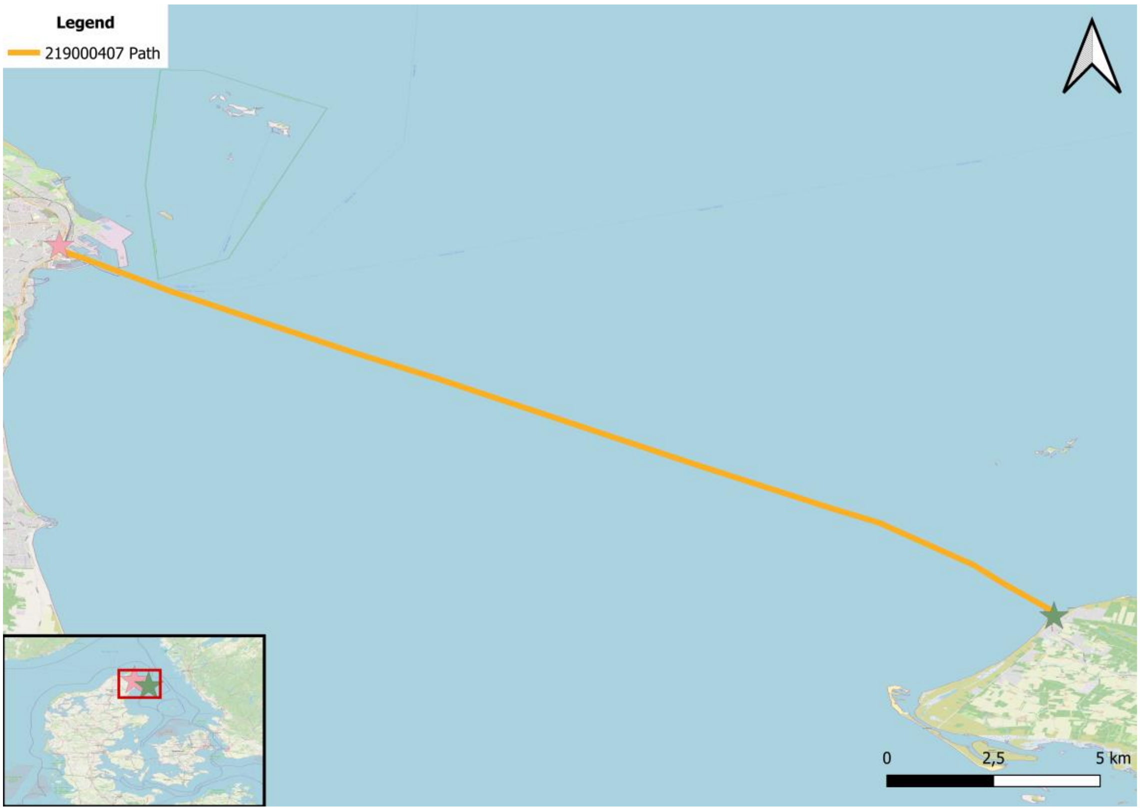

Figure 27 Study area and normal path

Travel times during the August week are visualized in Figure 28.

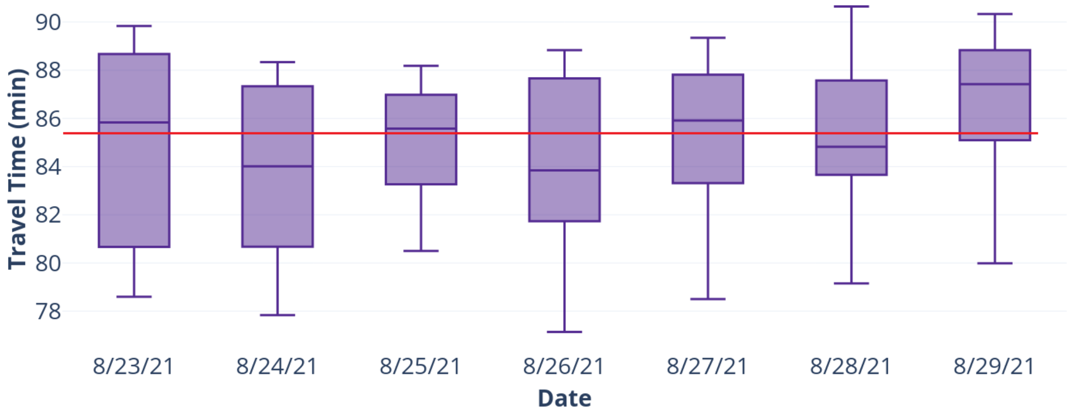

Figure 28 Travel times - August

Journey times vary between 77 and 93 minutes. However, mainly it is a journey between 81 and 89 minutes. The median mean is 85.34 minutes. Travel times in September are visualized in Figure 29.

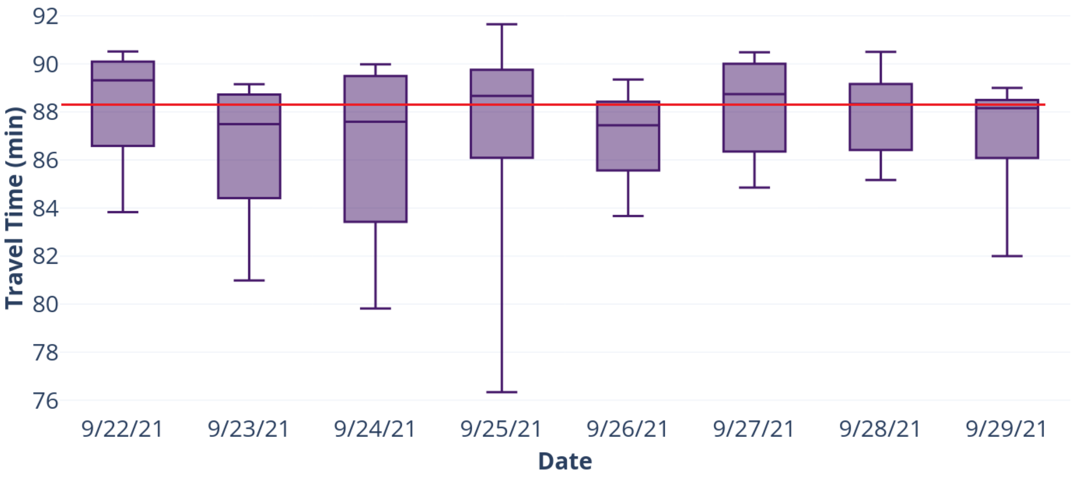

Figure 29 Travel times - September

Journey times lasted between 76 minutes and 92 minutes. It was observed that a journey on September 25 took much shorter than other journeys. The median is 88.21 minutes. The possible anomalies are visualized in Figure 30.

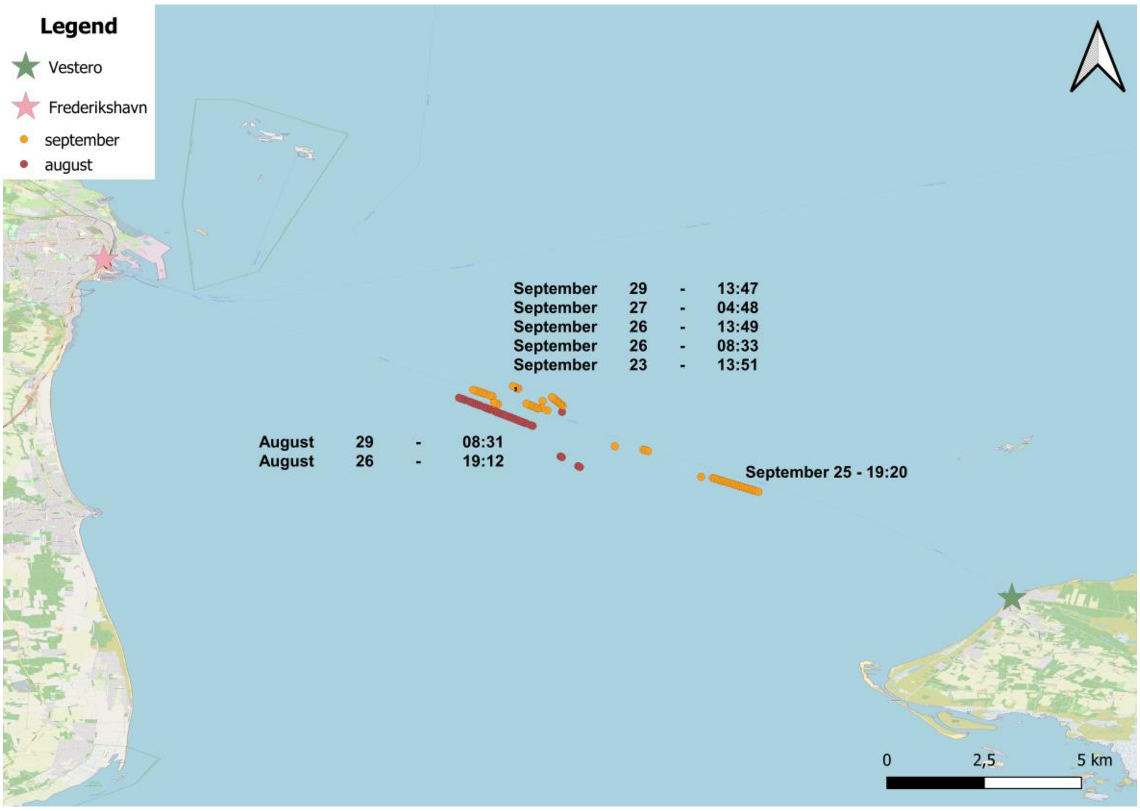

Figure 30 Possible anomalies

58 journeys were made in August and 64 journeys were made in September. Possible anomalies were detected on only two journeys in August. More possible anomalies were

identified in September. There are possible anomalies in six journeys. The wind speed in September on the analysed dates is visualized in Figure 31.

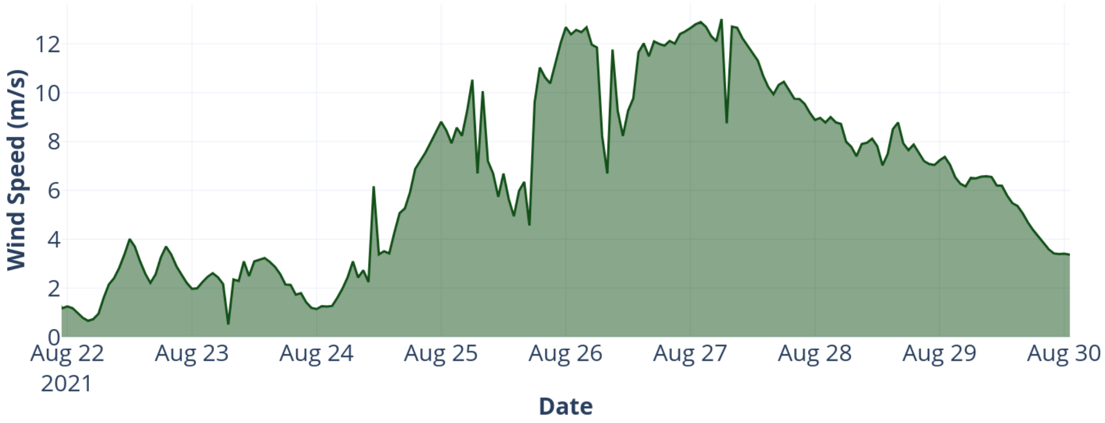

Figure 31 Wind speeds for August

On 26-27 August, the wind speed increased up to 12 m/s. Other days appear to be less windy. In September, the average travel time increased by about 3 minutes. Mean wind speed is 6.7 m/s in whole week. The wind speed in September on the analysed dates is visualized in Figure 32.

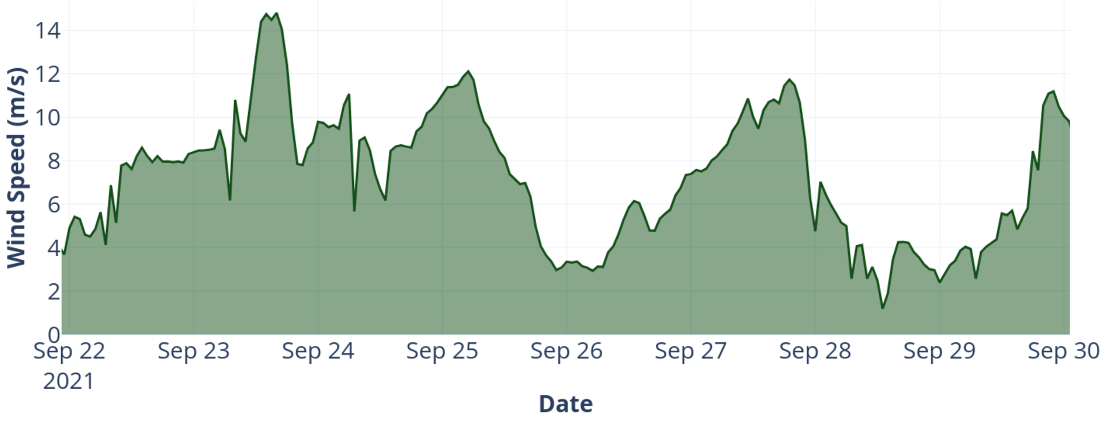

Figure 32 Wind speeds - September

While the windiest day was on September 23 with 15 m/s, the windless day was on September 28. Mean wind speed is 8.2 m/s in whole week. The 1.5 m/s wind speed difference caused 3 minutes travel time difference.

5. DISCUSSION

The anomaly points could only be detected once ship had been off course for a while. The duration of off course movement is related to the epsilon parameter in the DBSCAN algorithm. Specifically, if the most suitable epsilon value was determined by the elbow rule in the experiments. The epsilon value needs to be changed to detect earlier that the ship has deviated from the route. Here, the effect of the epsilon value on the results was investigated. There is a situation where the ship deviated from the route on September 24. For this date, the experiments were repeated with different epsilon values. In the experiments conducted throughout the thesis, the epsilon value was taken as 0.06. Possible anomalies are visualized in Figure 33.

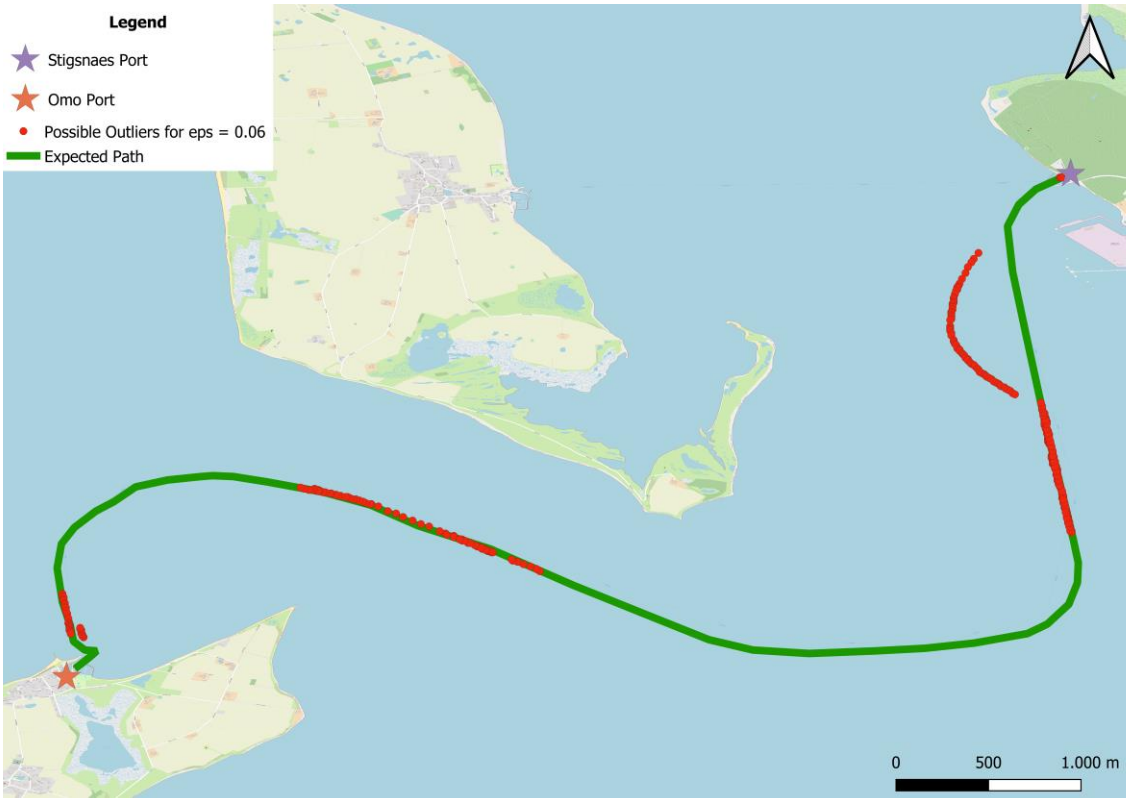

Figure 33 Epsilon value 0.06 detection possible anomalies

A total of 345 possible anomalies' point were found. Although the situation where the ship deviated from the route was detected, it was detected after a certain period. If the Epsilon value was taken as 0.04, the results would be as follows in Figure 34.

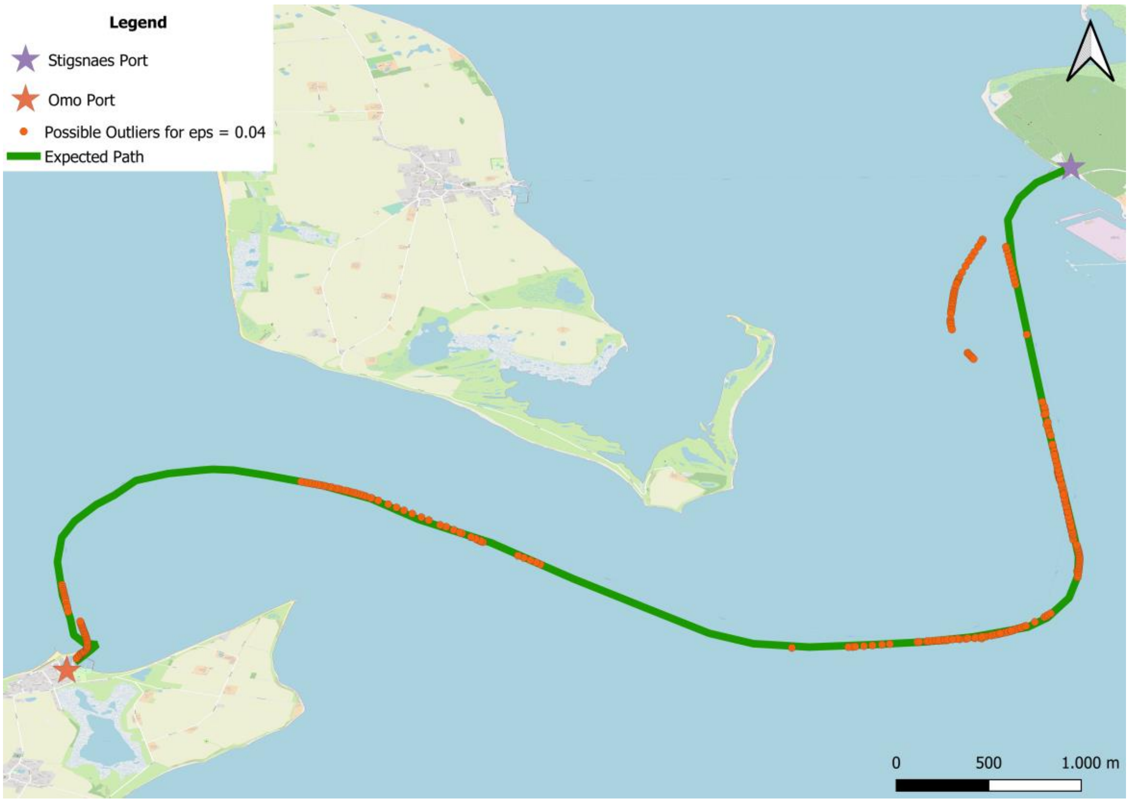

Figure 34 Epsilon value 0.04 detection possible anomalies

A total of 324 possible anomalies’ point were found. The total number of points has decreased compared to the previous situation. Half of the deviation from the route was determined as normal. The reason for this situation is that the points there provide the number that can form a class among themselves. There has been an increase in the number of points detected on the normal route. The Epsilon value was increased and taken as 0.08, the results would be as follows in Figure 35.

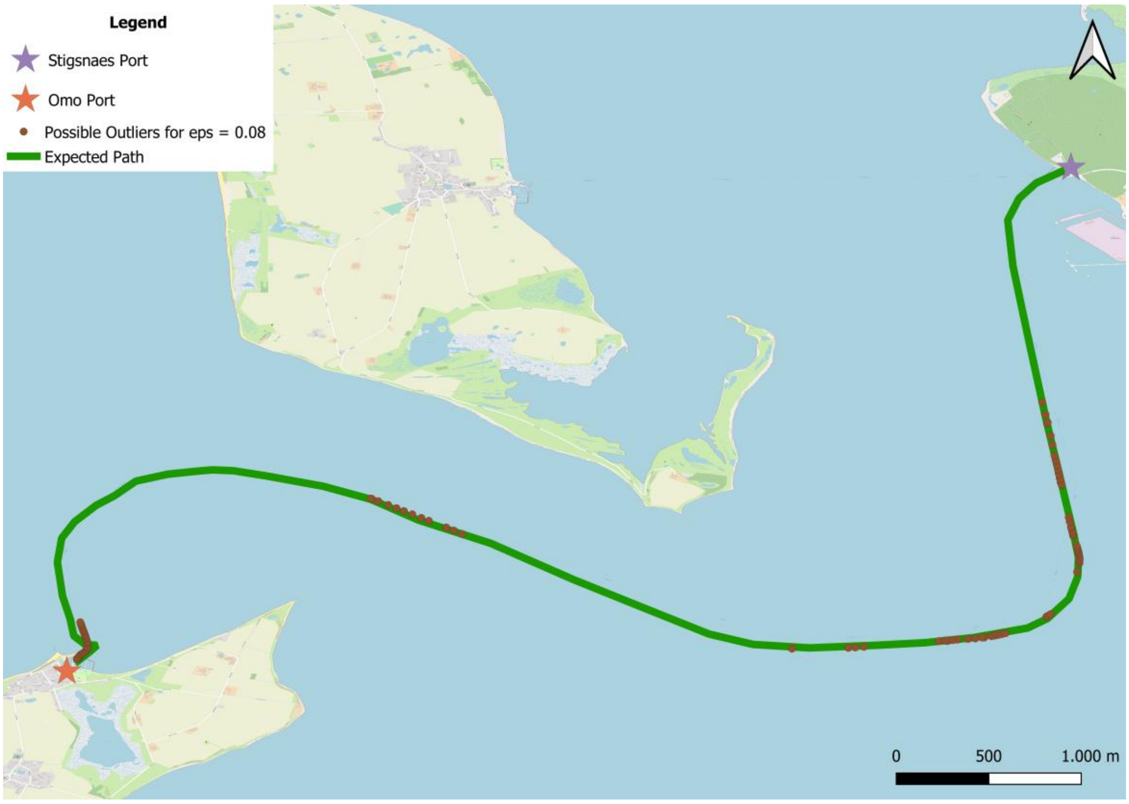

Figure 35 Epsilon value 0.08 detection possible anomalies

The number of possible anomalies decreased to 111. The cases where the ship deviated from the route could not be determined.

The focus was on wind speed throughout the thesis experiments. The direction of the wind can be effective in the sailing of the ships. With the open weather map, wind directions can be obtained angularly. Wind direction is visualized during the week of September in Figure 36.

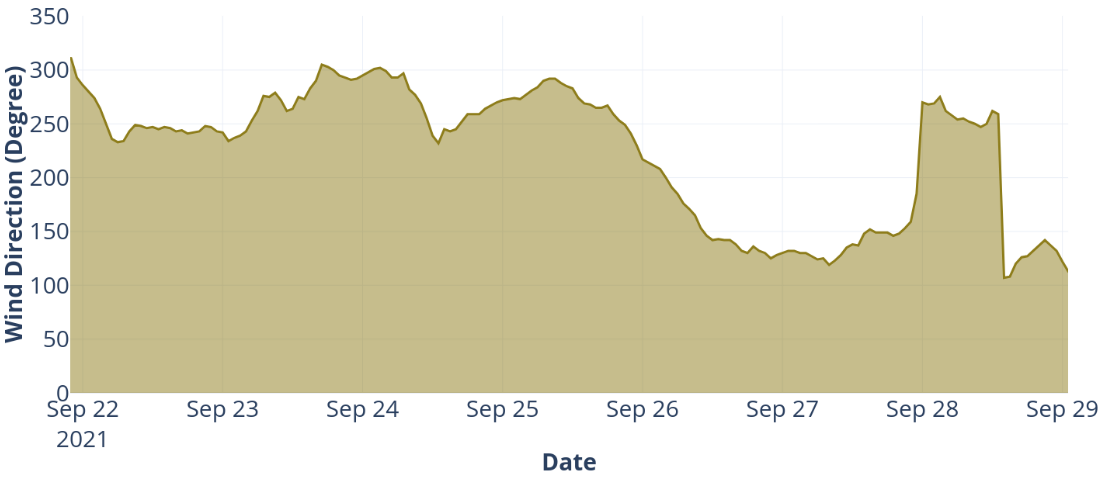

Figure 36 Wind direction values - September 2021

Wind direction varied between 107 and 311 degrees. There may be a relationship between the direction of the ships at the time of voyage and the wind speed. The values obtained during the fastest and slowest journeys throughout the analysis are shown in the Table 5.1.

Table 5.1. Fastest and slowest journey values - September 2021

Journey ID	**Journey Time (Min)**	**Wind Speed (m/s)**	**Cloud Cover (%)**	**Temperature (C)**
22 Sept - fastest	42.5	10.57	14	16
22 Sept - slowest	49.5	7.75	85	16
23 Sept - fastest	43	12.45	100	15
23 Sept - slowest	51.5	18	95	13
24 Sept - fastest	41	11.60	100	12
24 Sept - slowest	52	11	100	14
25 Sept - fastest	42	6.98	26	15
25 Sept - slowest	51.5	13	7	14
26 Sept - fastest	41	6.21	80	13
26 Sept - slowest	51.5	6.46	94	15
27 Sept - fastest	41.5	9.18	100	16
27 Sept - slowest	54	13.20	100	14
28 Sept - fastest	42.5	1.34	2	17

28 Sept - slowest	48	3.15	0	18

The experiments were carried out in two different seasons, on two different months and on two different ships. The results were examined separately. Examining more ships and dates would be useful for establishing an empirical model between weather and anomalies. With the establishment of the empirical model, weather conditions will be automatically removed from being an anomaly.

The coordinates of the ships are stored in the geographic coordinate system (GCS). A metric system is not used in GCS. In the DBSCAN algorithm, the data is used by scaling. For this reason, the results are not affected. However, the most optimal epsilon parameter is determined by the elbow rule. Epsilon parameter is related to the distribution of the data. As the distribution in the data larger, the value of the epsilon parameter increases. For example, a random sample data set was created in a metric coordinate system. The standard deviation of the data was taken as 10 and 100, respectively. The datasets images are shown in Figure 37 and Figure 38.

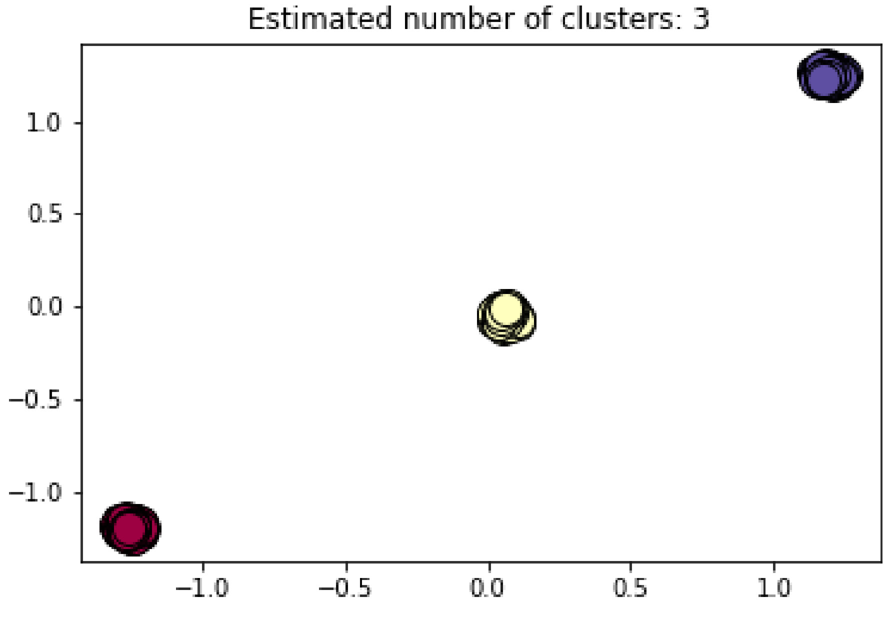

Figure 37 Random created dataset standard deviation is 10

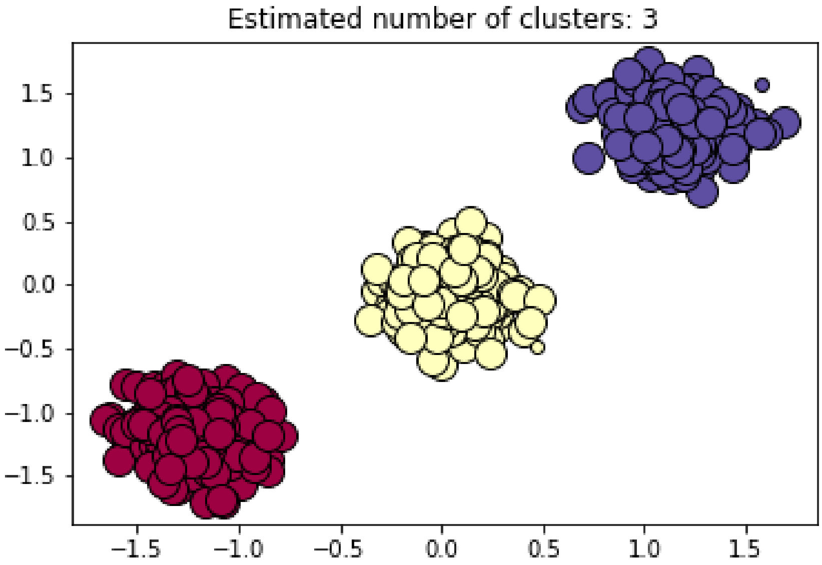

Figure 38 Random created dataset standard deviation is 100

If the optimal epsilon parameter values are found in the same data, the results are shown in Figure 39 and Figure 40.

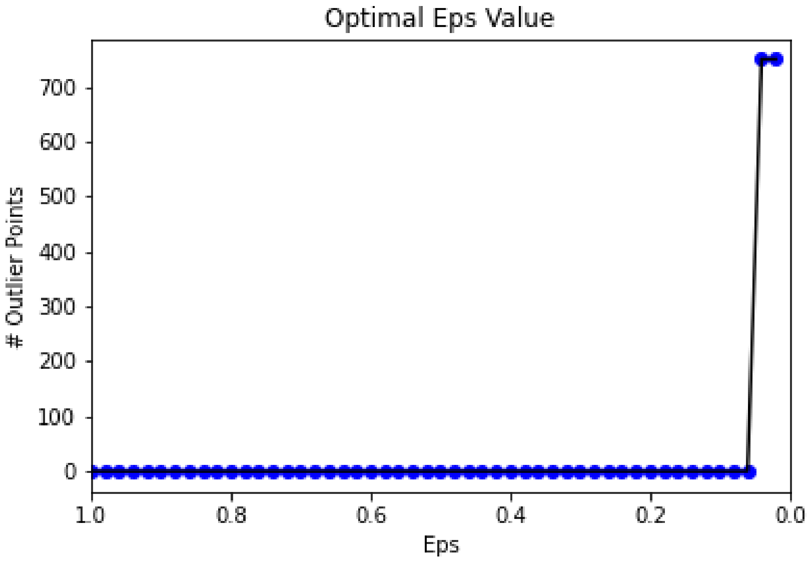

Figure 39 Optimal epsilon value for dataset standard deviation is 10

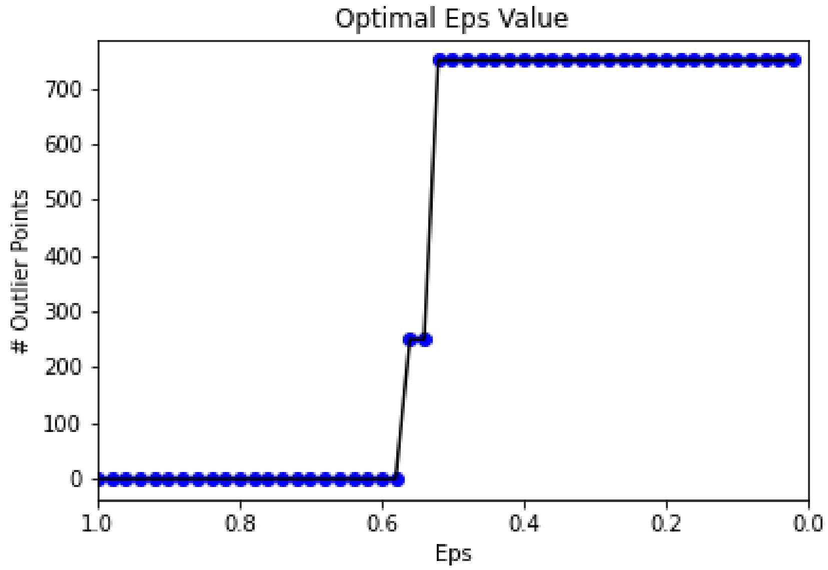

Figure 40 Optimal epsilon value for dataset standard deviation is 100

As can be seen from the graphs, an increase in epsilon values was observed. It can be said that the epsilon value is related to the distribution of the data. As the distribution larger, the parameter value increases.

Real-time application of anomaly detection is important for the safety of ships. Thus, the situations to be intervened are determined and security is ensured. Throughout the experiments in this thesis, the relationship between the weather conditions and the speed of the ships was examined.

Anomaly detection can be made in real time. For real-time anomaly detection, the clustering algorithm must be run in real-time. In a study to run the DBSCAN algorithm in real time, the RT-DBSCAN method has been proposed [57]. The study works on Apache Spark which analytics engine for large-scale data processing tool. By scaling analyses, a fast analysis result can be obtained. With this method, real-time anomaly detection can be applied in AIS data. As a result of this thesis, a relationship has been established between weather conditions and anomaly points. On the results obtained in real time, situations where the wind speed exceeds 10 m/s can be detected and weather-related anomalies can be eliminated.

6. CONCLUSION

This thesis developed a computationally reproducible approach for determining passenger ship journey times and detecting anomalous journeys. To achieve this aim, territorial units and AIS data are imported to Postgres as the initial step. Second, erroneous GPS data was removed. Third, a passenger ship was chosen randomly having an average journey time of approximately 45 minutes to have a thorough investigation. The anomalous journey times are determined using the DBSCAN algorithm that relies on the openly available Danish AIS data. In this context, Postgres was used as the spatial database. All functions were written as open source and shared as Jupyter Notebook on GitHub. In this way, other interested users can reproduce the results obtained in this thesis. The developed Jupyter Notebook was successfully replicated by a volunteer Master of Science student, and anomalous journey times of two other passenger ships were examined.

Experiments were based on passenger ships. Ships whose ship positions can be obtained regularly were filtered and a random selection was made. As a result, the study area is chosen between the Danish ports of Omo and Stigsnaes, a journey of approximately 45 minutes. Possible anomalies were obtained with the DBSCAN algorithm.

In the reproducibility experiments, a ship sailing between the Danish ports of Frederikshavn and Vestero was chosen. This journey takes about 90 minutes. Different dates were analysed in August, September and December of 2021. Journeys are divided into 10 minutes segments. Thus, where ships accelerated or slowed down could be determined. The identified possible anomalies are further compared with weather data to understand whether they could be explained accordingly.

Weather data was obtained via Open Weather Map, which is shared as open source. Historical weather data for the last year can be accessed. Weather data of a position given latitude and longitude can be obtained hourly. For this reason, weather data of the midpoint of the journey were used. In the case of weather, the closest hours are considered. In the weather data, wind speed, air temperature, cloud rate data were examined.

As a result, some of the possible anomalies could be explained by weather data. Wind speed was found to be correlated with ship journey time. In addition to wind speed, instantaneous changes in wind speed were also found to effect travel times. No direct effect of cloud rate and air temperature was observed. The cases where the weather condition is higher than 10 m/s are excluded from the possible anomaly values. The remaining values are called true anomaly. It has been determined that 45% of possible anomaly values are caused by wind speed.

Future work, it is planned to establish an empirical model between wind speed, wind direction and journeys. In this way, false alarms that may occur during journeys will be reduced.

7. REFERENCES

[1] M. Fusaro, A. Polónia, and International Maritime Economic History Association, Eds., *Maritime history as global history*. St. John’s, Nfld: International Maritime Economic History Association, **2010**.

[2] ‘Maritime transport’, *Wikipedia*. Sep. 02, 2021. Accessed: **Oct. 16, 2021**. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Maritime_transport&oldid=1041975560

[3] European Environment Agency. and European Maritime Safety Agency., *European maritime transport environmental report 2021*. LU: Publications Office, 2021. Accessed: **Sep. 19, 2021**. [Online]. Available: https://data.europa.eu/doi/10.2800/3525

[4] ‘MarineTraffic: Global Ship Tracking Intelligence | AIS Marine Traffic’. https://www.marinetraffic.com/en/ais/home/centerx:-12.0/centery:24.9/zoom:4 (accessed **Oct. 16, 2021**).

[5] ‘AIS transponders’. https://www.imo.org/en/OurWork/Safety/Pages/AIS.aspx (accessed **Sep. 12, 2021**).

[6] T. Stupak, ‘Influence of Automatic Identification System on Safety of Navigation at Sea’, *TransNav*, vol. 8, no. 3, pp. 337–341, **2014**, doi: 10.12716/1001.08.03.02.

[7] G. Cimino, ‘Sensor data management to achieve information superiority in maritime situational awareness’, p. 48, **2014**.

[8] V. Chandola, A. Banerjee, and V. Kumar, ‘Anomaly detection: A survey’, *ACM Comput. Surv.*, vol. 41, no. 3, pp. 1–58, **Jul. 2009**, doi: 10.1145/1541880.1541882.

[9] G. Pallotta, M. Vespe, and K. Bryan, ‘Vessel Pattern Knowledge Discovery from AIS Data: A Framework for Anomaly Detection and Route Prediction’, *Entropy*, vol. 15, no. 6, Art. no. 6, **Jun. 2013**, doi: 10.3390/e15062218.

[10] N. M. Quy, K. Łazuga, L. Gucma, J. K. Vrijling, and P. H. A. J. M. van Gelder, ‘Towards generalized ship’s manoeuvre models based on real time simulation results in port approach areas’, *Ocean Engineering*, vol. 209, p. 107476, **Aug. 2020**, doi: 10.1016/j.oceaneng.2020.107476.

[11] I. Grappasonni *et al.*, ‘Survey on smoking habits among seafarers’, *Acta Bio Medica Atenei Parmensis*, vol. 90, no. 4, pp. 489–497, **Dec. 2019**, doi: 10.23750/abm.v90i4.9001.

[12] ‘Index of /aisdata’. https://web.ais.dk/aisdata/ (accessed **May 01, 2022**).

[13] İ. B. Coşkun, S. Sertok, and B. Anbaroğlu, ‘K-NEAREST NEIGHBOUR QUERY PERFORMANCE ANALYSES ON A LARGE SCALE TAXI DATASET: POSTGRESQL VS. MONGODB’, *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, vol. XLII-2/W13, pp. 1531–1538, **Jun. 2019**, doi: 10.5194/isprs-archives-XLII-2-W13-1531-2019.

[14] ‘bugracoskun/trajectory’, 2021. https://github.com/bugracoskun/trajectory (accessed **Oct. 21, 2021**).

[15] X. Kong *et al.*, ‘Big Trajectory Data: A Survey of Applications and Services’, *IEEE Access*, vol. 6, pp. 58295–58306, **2018**, doi: 10.1109/ACCESS.2018.2873779.

[16] S. Chang, K. Yeh, G. Peng, S. Chang, and C. Huang, ‘From Safety to security - Pattern and anomaly detections in maritime trajectories’, in *2015 International Carnahan Conference on Security Technology (ICCST),* **Sep. 2015**, pp. 415–419. doi: 10.1109/CCST.2015.7389720.

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

[18] Y. Wang, J. Zhang, X. Chen, X. Chu, and X. Yan, ‘A spatial–temporal forensic analysis for inland–water ship collisions using AIS data’, *Safety Science*, vol. 57, pp. 187–202, **Aug. 2013**, doi: 10.1016/j.ssci.2013.02.006.

[19] L. Zhang, Q. Meng, Z. Xiao, and X. Fu, ‘A novel ship trajectory reconstruction approach using AIS data’, *Ocean Engineering*, vol. 159, pp. 165–174, **Jul. 2018**, doi: 10.1016/j.oceaneng.2018.03.085.

[20] U. Demšar and K. Virrantaus, ‘Space–time density of trajectories: exploring spatio-temporal patterns in movement data’, *International Journal of Geographical Information Science*, vol. 24, no. 10, pp. 1527–1542, **Oct. 2010**, doi: 10.1080/13658816.2010.511223.

[21] L. Huang, Y. Wen, Y. Zhang, C. Zhou, F. Zhang, and T. Yang, ‘Dynamic calculation of ship exhaust emissions based on real-time AIS data’, *Transportation Research Part D: Transport and Environment*, vol. 80, p. 102277, **Mar. 2020**, doi: 10.1016/j.trd.2020.102277.

[22] M. Goldstein and S. Uchida, ‘A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data’, *PLoS ONE*, vol. 11, no. 4, p. e0152173, **Apr. 2016**, doi: 10.1371/journal.pone.0152173.

[23] S. B. Salem, S. Naouali, and M. Sallami, ‘Clustering Categorical Data Using the K-Means Algorithm and the Attribute’s Relative Frequency’, vol. 11, no. 6, p. 7, **2017**.

[24] M. Amer, ‘Comparison of Unsupervised Anomaly Detection Techniques’, p. 44, **2011**.

[25] K. Wolsing, L. Roepert, J. Bauer, and K. Wehrle, ‘Anomaly Detection in Maritime AIS Tracks: A Review of Recent Approaches’, *JMSE*, vol. 10, no. 1, p. 112, **Jan. 2022**, doi: 10.3390/jmse10010112.

[26] R. Laxhammar, ‘Anomaly Detection in sea traffic - a comparison of the Gaussian Mixture Model and the Kernel Density Estimator’, p. 9, **Jul. 2009**.

[27] M. Riveiro, G. Pallotta, and M. Vespe, ‘Maritime anomaly detection: A review’, *WIREs Data Mining and Knowledge Discovery*, vol. 8, no. 5, Art. no. 5, **2018**, doi: https://doi.org/10.1002/widm.1266.

[28] G. Xu, C.-H. Chen, F. Li, and X. Qiu, ‘AIS data analytics for adaptive rotating shift in vessel traffic service’, *IMDS*, vol. 120, no. 4, pp. 749–767, **Mar. 2020**, doi: 10.1108/IMDS-01-2019-0056.

[29] G. Boztepe, ‘MIDDLE EAST TECHNICAL UNIVERSITY’, p. 101, **2019**.

[30] A. N. Radon, K. Wang, U. Glässer, H. Wehn, and A. Westwell-Roper, ‘Contextual verification for false alarm reduction in maritime anomaly detection’, in *2015 IEEE International Conference on Big Data (Big Data),* **Oct. 2015**, pp. 1123–1133. doi: 10.1109/BigData.2015.7363866.

[31] S. R. Piccolo and M. B. Frampton, ‘Tools and techniques for computational reproducibility’, *GigaSci*, vol. 5, no. 1, p. 30, **Dec. 2016**, doi: 10.1186/s13742-016-0135-4.

[32] G. B. Karataş, P. Karagoz, and O. Ayran, ‘Trajectory pattern extraction and anomaly detection for maritime vessels’, *Internet of Things*, vol. 16, p. 100436, **Dec. 2021**, doi: 10.1016/j.iot.2021.100436.

[33] P. Nie, Z. Chen, N. Xia, Q. Huang, and F. Li, ‘Trajectory Similarity Analysis with the Weight of Direction and k-Neighborhood for AIS Data’, *IJGI*, vol. 10, no. 11, p. 757, **Nov. 2021**, doi: 10.3390/ijgi10110757.

[34] S. A. Roberts, ‘A Shape-Based Local Spatial Association Measure (LISShA): A Case Study in Maritime Anomaly Detection’, *Geogr Anal*, vol. 51, no. 4, pp. 403–425, **Oct. 2019**, doi: 10.1111/gean.12178.

[35] L. Zhao and G. Shi, ‘Maritime Anomaly Detection using Density-based Clustering and Recurrent Neural Network’, *J. Navigation*, vol. 72, no. 04, pp. 894–916, **Jul. 2019**, doi: 10.1017/S0373463319000031.

[36] Y. Wang, L. Yang, X. Song, Q. Chen, and Z. Yan, ‘A Multi-Feature Ensemble Learning Classification Method for Ship Classification with Space-Based AIS Data’, *Applied Sciences*, vol. 11, no. 21, p. 10336, **Nov. 2021**, doi: 10.3390/app112110336.

[37] G. Spadon, M. D. Ferreira, A. Soares, and S. Matwin, ‘Unfolding collective AIS transmission behavior for vessel movement modeling on irregular timing data using noise-robust neural networks’, *arXiv:2202.13867 [cs]*, Feb. 2022, Accessed: **Apr. 30, 2022**. [Online]. Available: http://arxiv.org/abs/2202.13867

[38] S. Shuang, C. Yan, and Z. Jinsong, ‘Trajectory Outlier Detection Algorithm for ship AIS Data based on Dynamic Differential Threshold’, *J. Phys.: Conf. Ser.*, vol. 1437, p. 012013, **Jan. 2020**, doi: 10.1088/1742-6596/1437/1/012013.

[39] Y. Sun *et al.*, ‘SHIP TRAJECTORY CLEANSING AND PREDICTION WITH HISTORICAL AIS DATA USING AN ENSEMBLE ANN FRAMEWORK’, p. 17, **Feb. 2021**.

[40] M. Mieczyńska and I. Czarnowski, ‘K-means clustering for SAT-AIS data analysis’, *WMU J Marit Affairs*, vol. 20, no. 3, pp. 377–400, **Sep. 2021**, doi: 10.1007/s13437-021-00241-3.

[41] İ. B. Coşkun, S. Sertok, and B. Anbaroğlu, ‘K-NEAREST NEIGHBOUR QUERY PERFORMANCE ANALYSES ON A LARGE SCALE TAXI DATASET: POSTGRESQL VS. MONGODB’, *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, vol. XLII-2/W13, pp. 1531–1538, **Jun. 2019**, doi: 10.5194/isprs-archives-XLII-2-W13-1531-2019.

[42] J. Wang, N. Ntarmos, and P. Triantafillou, ‘Indexing Query Graphs to Speedup Graph Query Processing’. OpenProceedings.org, **2016**. doi: 10.5441/002/EDBT.2016.07.

[43] ‘LAU - GISCO - Eurostat’. https://ec.europa.eu/eurostat/web/gisco/geodata/reference-data/administrative-units-statistical-units/lau#lau19 (accessed **Apr. 03, 2022**).

[44] B. Liu, E. N. de Souza, C. Hilliard, and S. Matwin, ‘Ship movement anomaly detection using specialized distance measures’, in *2015 18th International Conference on Information Fusion (Fusion)*, **Jul. 2015**, pp. 1113–1120.

[45] A. Bhardwaj, ‘Silhouette Coefficient : Validating clustering techniques’, *Medium*, May 27, 2020. https://towardsdatascience.com/silhouette-coefficient-validating-clustering-techniques-e976bb81d10c (accessed **Mar. 20, 2022**).

[46] ‘Karataş vd. - **2021** - Trajectory pattern extraction and anomaly detectio.pdf’.

[47] T. Mullin, ‘DBSCAN Parameter Estimation Using Python’, *Medium*, **Jul. 15, 2020**. https://medium.com/@tarammullin/dbscan-parameter-estimation-ff8330e3a3bd (accessed **Mar. 20, 2022**).

[48] R. W. Reynolds, T. M. Smith, C. Liu, D. B. Chelton, K. S. Casey, and M. G. Schlax, ‘Daily High-Resolution-Blended Analyses for Sea Surface Temperature’, *Journal of Climate*, vol. 20, no. 22, pp. 5473–5496, **Nov. 2007**, doi: 10.1175/2007JCLI1824.1.

[49] N. MIMURA, ‘Sea-level rise caused by climate change and its implications for society’, *Proc Jpn Acad Ser B Phys Biol Sci*, vol. 89, no. 7, pp. 281–301, **Jul. 2013**, doi: 10.2183/pjab.89.281.

[50] B. Coşkun, *trajectory-jupyter notebook*. 2022. Accessed: **May 01, 2022**. [Online]. Available: https://github.com/bugracoskun/trajectory/blob/4f8e2b8995b51b36f4917c5287768254e96217e9/weather/outlier_file.ipynb

[51] ‘bugracoskun/trajectory’, 2021. https://github.com/bugracoskun/trajectory (accessed **Oct. 21, 2021**).

[52] J. Bergh *et al.*, ‘Modelling the short-term effects of climate change on the productivity of selected tree species in Nordic countries’, *Forest Ecology and Management*, vol. 183, no. 1–3, pp. 327–340, **Sep. 2003**, doi: 10.1016/S0378-1127(03)00117-8.

[53] E. Andersen and S. Furbo, ‘Theoretical variations of the thermal performance of different solar collectors and solar combi systems as function of the varying yearly weather conditions in Denmark’, *Solar Energy*, vol. 83, no. 4, pp. 552–565, **Apr. 2009**, doi: 10.1016/j.solener.2008.10.009.

[54] ‘Denmark extends hard lockdown until Jan. 17 amid spike in infections | Reuters’. https://www.reuters.com/business/healthcare-pharmaceuticals/denmark-extend-lockdown-measures-until-jan-17-tv2-2020-12-29/ (accessed **May 14, 2022**).

[55] ‘Current weather and forecast - OpenWeatherMap’. https://openweathermap.org/ (accessed **May 07, 2022**).

[56] V. D. Prasita, L. A. Zati, and S. Widagdo, ‘The Characteristics of West Season Wind and Wave as well as Their Impacts on Ferry Cruise in The Kalianget-Kangean Cruise Route, Madura, Indonesia’, *JST*, vol. 29, no. 3, **Jul. 2021**, doi: 10.47836/pjst.29.3.16.

[57] Y. Gong, R. O. Sinnott, and P. Rimba, ‘RT-DBSCAN: Real-Time Parallel Clustering of Spatio-Temporal Data Using Spark-Streaming’, in *Computational Science – ICCS 2018*, vol. 10860, Y. Shi, H. Fu, Y. Tian, V. V. Krzhizhanovskaya, M. H. Lees, J. Dongarra, and P. M. A. Sloot, Eds. Cham: Springer International Publishing, **2018**, pp. 524–539. doi: 10.1007/978-3-319-93698-7_40.