Predicting the Time of Arrival at Port for Maritime Surface Ships in the Baltic Sea Using Recurrent Neural Networks

Jonatan Lahtivuori

Abstract

Here is where I would say what is in this document.

Contents

1 Introduction

Describe the need for accurate ETA predictions for maritime traffic and what data is available what has been done already and what the thesis will contribute.

Structure of the thesis.

2 Maritime vessels and traffic

The Baltic Sea with an approximate surface area of 420,000 km^2 is rather small compared to other seas and navigating in the Baltic Sea is often confined to fairways, especially when getting closer to land and travelling in the archipelago. This creates highways of vessel traffic coming from and going to the many ports in the Baltic Sea merging into one another at some point during the voyage. This is even more evident in the winter months when there are restrictions on where vessels are allowed to navigate due to the ice, which is governed by the Finnish and Swedish authorities. During the winter months, vessels might also be required to get help from an icebreaker to enter the port, and with only around 30 operational icebreakers in the Baltic Sea could even create queues to get assistance where vessels have to anchor until assistance arrives. Other characteristics are the areas of international waters in the Baltic Sea; vessels travelling from and to the east part of the Baltic Sea often use these areas as an example. To enter the Baltic Sea from the Atlantic, all vessels have to travel through the Danish straits and Kattegat.

The past years, approximately 90 % of all import and export of Finland was done by sea [1]. There is no question about the importance of sea transports and maintaining efficient operations at sea and at port is of interest for all parties, especially in Finland where most of all imports and exports passes through ports at some point during the transportation.

One important port for the Finnish industry and economy is the Port of Naantali. Located close the city of Turku and having a central location in the Baltic Sea it was deemed to be a suitable candidate for the thesis.

2.1 Automatic identification system

Automatic identification system (*AIS*) was introduced by the International Maritime Organization (*IMO*) in the early 2000's in accordance with the Safety of Life at Sea (*SOLAS*) treaty as an open communication tool for all maritime traffic. The main objectives of the treaty is to improve the safety of life at sea, protection of the marine environment and safety and efficiency of navigation [2]. In its general operating form, AIS operates on two VHF channels and all vessels or base stations within the vessels transmission range receive the messages transmitted and vice versa the vessel receives all messages broadcasted in its receiving range. The broadcasting is based on Self-organized Time Division Multiple Access (*(S)TDMA*) with a minimum of 2000 time slots per minute broadcasting capacity rate and the ship-to-ship communication for vessels closer to each other, takes precedence over vessels farther away from one another, which allows for sharing time slots, and thus overloading the available time slots [2].

The IMO requires in accordance with regulation V/19 of SOLAS, that all vessels with a gross tonnage of 300 or more on international voyages, cargo vessels of 500 gross tonnage or more not on international voyages and all passenger vessel disregarding the gross tonnage to be equipped with an AIS. Further, the EU requires new-built fishing vessels longer than 15 meters to be fitted with an AIS from November 2010 and existing vessels to install a AIS by May 2014 at the latest [3]. There are two types of AIS, Class A and Class B. Class A transceiver are the more common and are compliant with the IMO regulations whereas Class B transceivers are not subject to the full IMO AIS requirements. Class B transceivers are typically simpler, of lower cost and installed on smaller vessels of one's own choosing for improving the situational awareness. Class A transceivers take precedence over Class B transceivers when broadcasting.

The vessels onboard sensors and positioning systems interface with the AIS to allow for communicating the status of the vessel, both static and dynamic information. AIS also allows for voyage related information and safety-related information to be transmitted via the system, see Table 1. The AIS guidelines requires that there is a minimum display and keyboard for input and receiving data, for updating information manually and retrieving data received via AIS. Some data for the AIS is entered only once, other has to be entered for every voyage and other data will be automatically gathered from the vessels onboard sensors.

The information received by the AIS does not provide the full situational picture of the vessels surroundings and should only be used as a navigational and situational awareness aid.

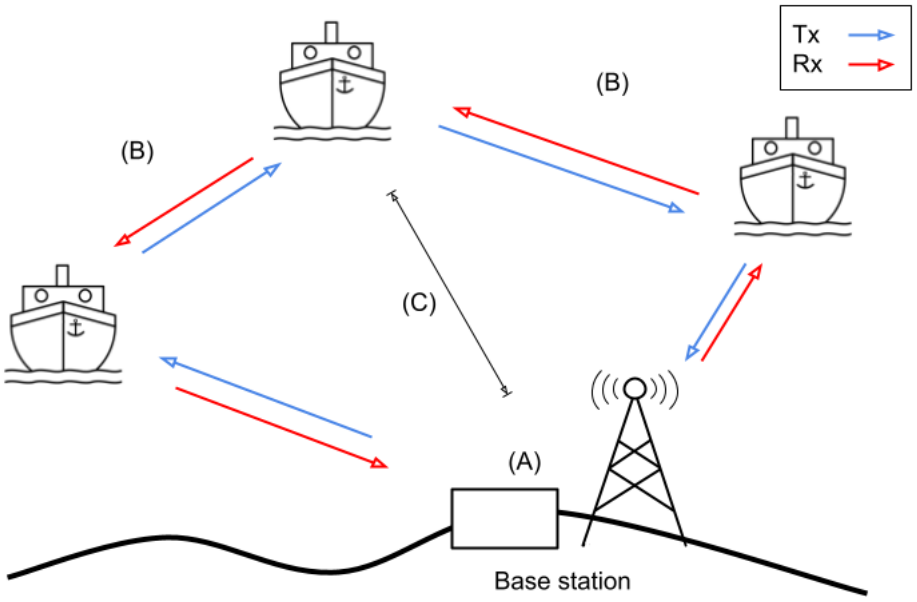

Figure 1: (A) Base station transmits information about other vessels, port data and possible hazards. Base station receives information from all vessels in range. (B) Ship-to-ship communication. Vessel broadcasts its current state (identifier, speed over ground, course over ground etc.) and receives data from all vessels in range, only limited by the number of possible time slots available. (C) Vessel outside the range of broadcasting AIS data to the base station, still broadcasts information to other vessels.

The AIS integrated checks for data integrity, built-in integrity test (*BIIT*), is not capable of validating data, for example a non-functional sensor will report as not available to the AIS. This can introduce faulty data being broadcasted to surrounding vessels and could create dangerous situations.

The nature of AIS being unencrypted and unauthenticated with the fact that there is no data validation before transmitting does entail that the veracity of the data is poor. The information available through AIS should not be trusted by itself and it is up too the OOW to safely operate the vessel and ensure safe navigation for the vessel and vessels in its surroundings.

The information available to transmit over AIS is divided into three groups; static, dynamic and voyage-related.

Data	Description
Static	
Maritime Mobile Service Identity (MMSI)	Set on installation, can change during vessels operational lifespan
Call sign and name	Set on installation, can change during vessels operational lifespan
IMO number	Set on installation
Length and beam	Set on installation or if changed
Type of ship	Selected from list of predefined values
Location of electronic positioning system	Set on installation, can be changed
Dynamic	
Ships position with accuracy indication and integrity status	Automatically updated from the position sensor
Position time stamp in UTC	Automatically updated from the main position sensor
Course over ground (COG)	Automatically updated from the main position sensor if available
Speed over ground (SOG)	Automatically updated from the main position sensor
Heading	Automatically updated from the vessels heading sensor
Navigational status	Manually entered by the OOW as needed
Rate of turn (ROT)	Automatically updated from the vessels ROT sensor or from the vessels gyro
Voyage-related	
Draught	Manually entered at the start of the voyage
Hazardous cargo (type)	Manually entered at the start of the voyage confirming the presence of such cargo
Destination and ETA	Manually entered at the start of the voyage and updates as needed
Route plan (waypoints)	Manually entered at the start of the voyage
Safety-related	
Short safety-related messages	Free format manually entered broadcasted to specific receiver or all vessels

Table 1: AIS data content and description, source [2].

Depending on the type of message sent by the AIS, the rate at which the AIS transmits messages autonomously changes. For static and voyage-related data the interval is six minutes between messages or at the request of another user. Dynamic information on the other hand takes the vessels current navigational status into consideration at which rate to transmit messages.

The different states of a vessel and at which rate it will transmit autonomously AIS messages with dynamic information can be seen in Table 2. Faster moving

vessels will broadcast its location and trajectory more frequently as a faster moving vessel will cover longer distances and can alter its course more over a shorter time. It is still important for the OOW to understand that not all vessels will be transmitting AIS data and therefore AIS is not guaranteed to provide a complete overview of the surroundings. It is also not guaranteed that other vessels are receiving AIS data and could therefore be a risk factor.

Vessel state	**General reporting interval**
At anchor or moored and not moving faster than 3 knots	3 min
At anchor or moored and moving faster than 3 knots	10 s
0-14 knots	10 s
Changing course at 0-14 knots	3 1/3 s
14-23 knots	6 s
Changing course at 14-23 knots	2 s
Faster than 23 knots	2 s
Changing course at more than 23 knots	2 s

Table 2: AIS message broadcast interval for dynamic information for class A vessels [2].

2.1.1 Estimated Time of Arrival

Estimated time of arrival is the estimated time when a vessel will reach its destination, typically transmitted to the correct authorities 24 to 72 hours before arrival [4, 5]. The means of communicating this ETA varies and the accuracy of the ETA is not guaranteed to be within any margin of error. Ports, being very complex infrastructures with many moving parts, has to operate with certain uncertainties of vessels arrival times and required demands and it is essential to plan port operations for a 24-hour period at a minimum to ensure smooth operations [6]. All uncertainties that can delay or alter any operations at one port are therefore a risk that can delay and disrupt all operations for not only the port in question but also all other ports that have any relations to the port.

There are many different practices for relaying the ETA to the correct authorities, for example via the AIS, phones and email. However, a study by Harati Mokhtari *et al.* [7] discovered that almost a majority of the transmitted AIS messages had faulty or inaccurate data for the destination and ETA. Another risk discovered with the destination and ETA, when the data was available, was that it was not updated and thus could lead to more problems if any party thought the information to be true. An incorrect or inaccurate ETA could mean port authorities failing to plan its operations to handle incoming and outgoing traffic within the expected time.

2.2 The Baltic Marine Environment Protection Commission

The Baltic Marine Environment Protection Commission or Helsinki Commission (*HELCOM*) as it is also known, was formed in 1974 parallel with the *Convention on the Protection of the Marine Environment of the Baltic Sea*. The participating parties are Denmark, Estonia, the European Union, Finland, Germany, Latvia, Lithuania, Poland, Russia and Sweden with the goal to protect the marine environment from increasing exploitation and ensuring protection of the ecosystem in the Baltic Sea [8].

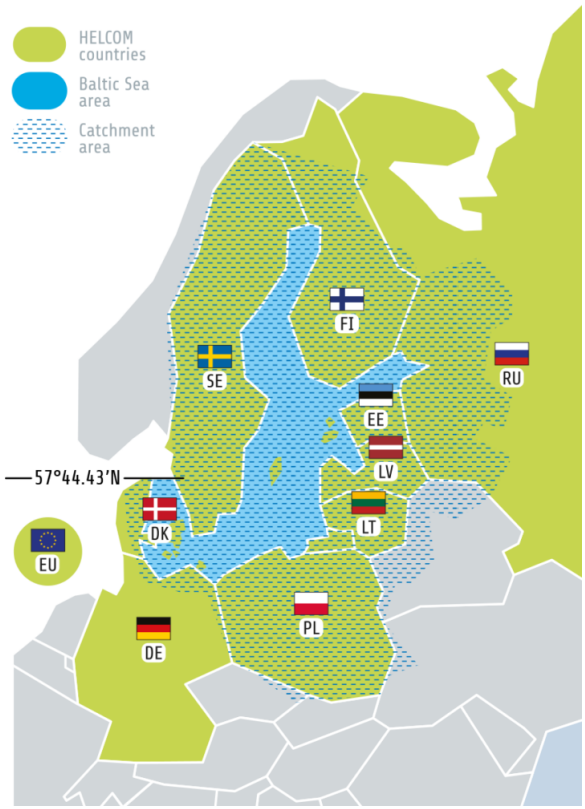

Figure 2: Area of the Baltic Sea of the Helsinki Convention, source `https://helcom.fi/about-us/`

2.2.1 HELCOM dataset

The HELCOM dataset consisting of AIS data, covers the period from January 2009 to December 2019. Each month for each year is stored as its own comma separated value *.csv* file.

The AIS data in the HELCOM dataset has already been processed from what is originally transmitted from the AIS. All rows in the dataset has been made uniform to include the same data, every entry in the dataset contains exactly the same features. This has been possible by merging many different databases into one and

this has also made it possible to cover the whole operating area of HELCOM, which includes all of the Baltic Sea seen in Figure 2.

2.2.2 Description of HELCOM dataset features

The HELCOM dataset have twelve unique features for each entry described in Table 3. The features are static, dynamic and voyage-related AIS data merged from multiple sources.

The combination of these features gives information about the vessels identification, position, time stamp, trajectory, draught and physical size.

Name	Description	Value
timestamp	Unix epoch time in milliseconds when AIS message was created	Min 1230768000000
mmsi	Maritime Mobile Service Identities, unique for each vessel can change	9 digit identifier
lat	Latitude position when AIS message was generated	Coordinate in WGS 84
long	Longitude position when AIS message was generated	Coordinate in WGS 84
sog	Speed over ground in knots	0.1 knot resolution
cog	Course over ground in degrees relative to true north	0.1 degrees
draught	Vertical distance from waterline to the keel	0.1 meters
dimBow	Reference point for position of positioning system on the vessel	Meters from bow
dimPort	Reference point for position of positioning system on the vessel	Meters from port side
dimStarboard	Reference point for position of positioning system on the vessel	Meters from starboard side
dimStern	Reference point for position of positioning system on the vessel	Meters from stern
imo	Unique identifier for each vessel, does not change	7 digit identifier

Table 3: Variables in HELCOM data set and description.

2.2.3 Statistical information

The HELCOM dataset covers a large window of time and gives an insight into historical vessel movements on the Baltic Sea. The dataset has been created by merging many databases and processing the data to create a uniform dataset for the whole period. In this dataset it is possible to find vessels movements on the Baltic Sea, where they are going and where they have been, by grouping the data

by unique vessels and then order the data for each vessel in chronological order. It is not given that a vessel will have a contiguous timeline without gaps, this due to the possibility when a vessel has travelled outside the area covered by the AIS or have turned off their AIS for any reason. These gaps can be many months even years long and does not impact the quality of the routes. Smaller gaps in the timeline of a vessel does however impact the quality of the routes. There are many reasons for why vessels timelines can have smaller gaps, less than a couple of hours, and all of them will negatively impact the possibility of finding good routes. Some of the most likely reasons are malfunctioning AIS, the AIS has been turned off or the data received has been faulty and not stored.

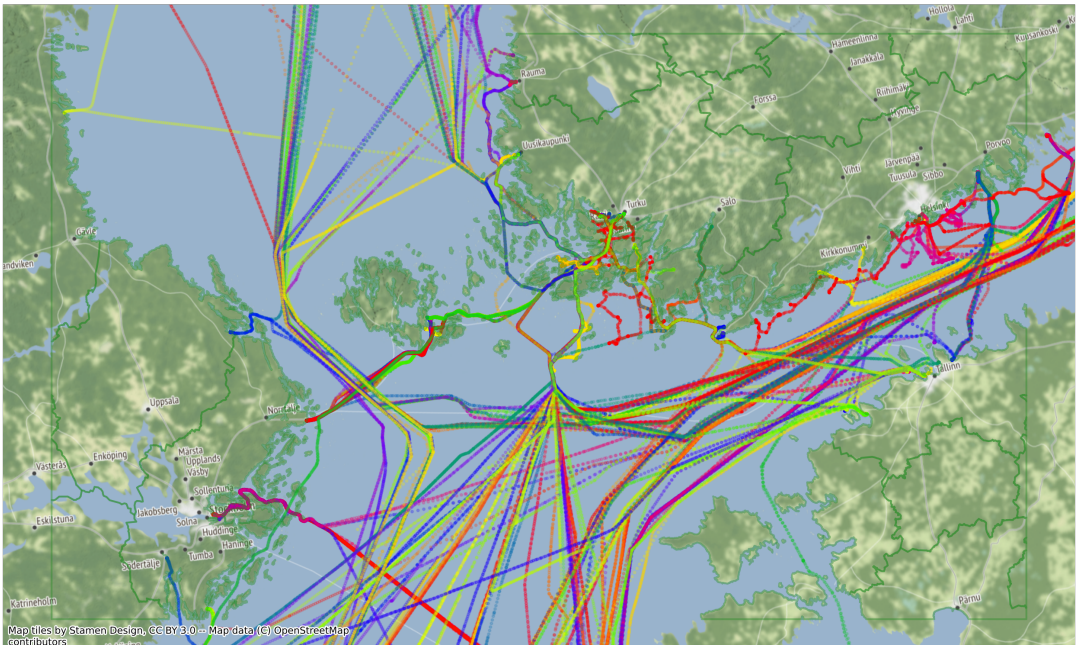

Figure 3: Small window of time of raw AIS data from the HELCOM dataset. Coloured by unique vessels.

Figure 3 shows a small window of time from the HELCOM dataset of all AIS data. Only a fraction of this data is viable for further processing, but a first pass over the whole dataset is required to capture all possible vessels for each month.

The average speed over ground for routes going to the Port of Naantali seen in Figure 4, indicates a even distribution of vessels travelling about 14 knots. This is expected from the types of vessels that are equipped with AIS and have been recorded in the dataset.

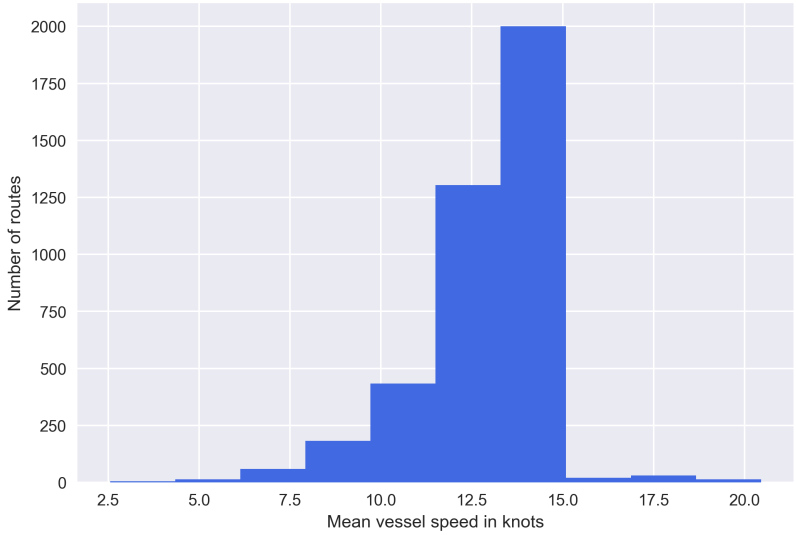

Figure 4: Average vessel speed for a route for all routes used, from the period of 2009 to 2018.

The time difference between AIS messages does not correlate to what raw AIS transmission rates should be, see Table 2. A concentration of time difference between one and four minutes is another indication of pre-processing done by HELCOM, as the timesteps compared are from windows of time where vessels have been under way during a route.

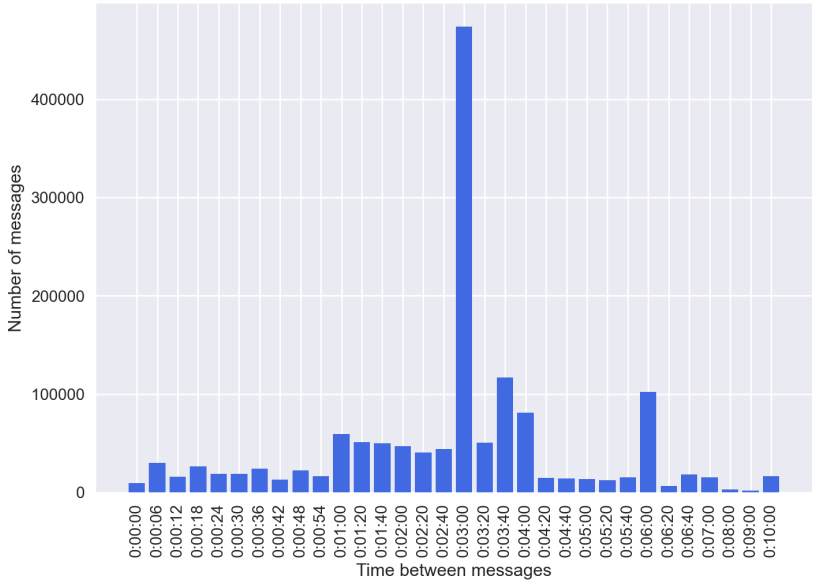

Figure 5: Time difference between AIS messages for any routes found during a five year period. All time differences greater than ten minutes have been collected in the same bar.

Vessels physical features and the grouping of different vessels by size. In Figure 7 an example of one year of unique vessels which have visited the port of Naantali can

be seen. Classifying vessels according to the physical size of the vessel has shown to be a efficient way of discerning different vessels capabilities of manoeuvring [9]. It is expected that vessels of similar dimensions manoeuvre in a similar fashion. The classification was done by estimating different groups of vessels from their length and width, after which they were classified into classes according to the boxes seen in Figure 7.

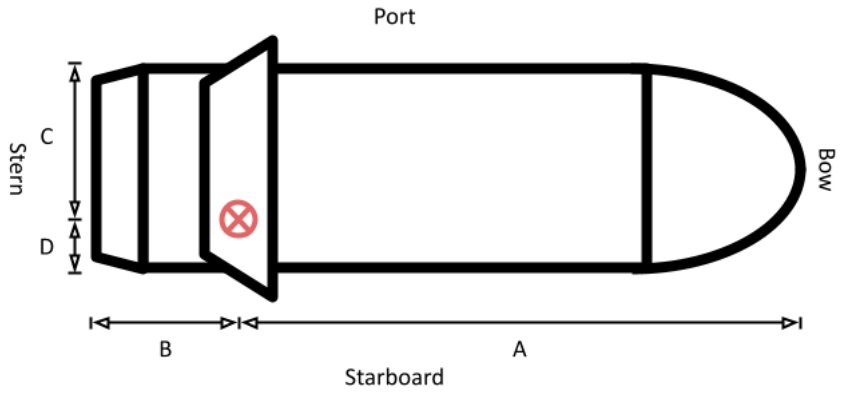

Figure 6: Vessel class definition from HELCOM data, the positioning system marked with red. A is *dimStern*, B is *dimBow*, C is *dimPort* and D is *dimStarboard*.

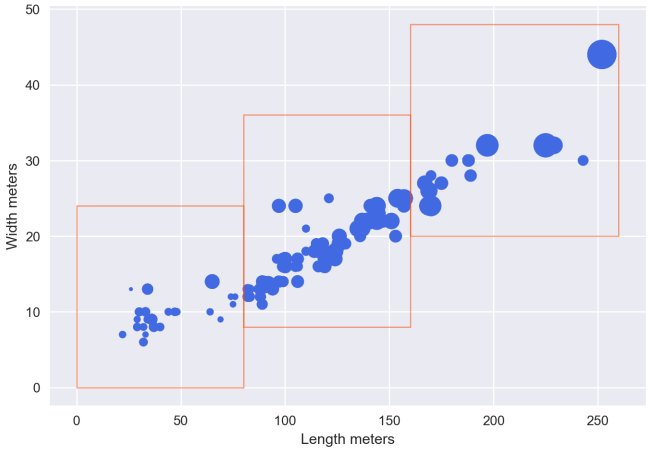

Figure 7: Defined vessel classes and distribution of all unique vessels for the year 2015. Size of the circle is related to the draught of the vessel.

2.3 Port of Naantali

The Port of Naantali is a rather important and busy port in the Baltic Sea and it is considered one of Finlands most busy ports [1]. With a reported total of over 8 million tonnes of cargo passing through the port and over 1,000 port calls during

2020 its logistical importance is of significance [10]. The port did at average for the years 2018 through 2020 handle at average 5,8 % of all of Finlands transports by sea and combined with the port at Kilpilahti handled all of Finlands crude oil transport [11].

This is also recognised in the HELCOM data by the amount of the data that has any connections to the Port of Naantali.

Year	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**
% of data	8.21	7.16	7.30	13.19	12.30	12.30	12.30	10.17	10.87	10.61

Table 4: Total amount of vessel data with connection to the Port of Naantali for each year of the HELCOM data.

The percentage of the data in Table 4 is the complete timeline for each vessel that has at any point during the years visited the Port of Naantali, including other voyages that are not going and coming from the Port of Naantali. Only a fraction of this data is actual viable data for the routes, described in 4.2.1. Still, the fact that 10 % on average of all data has any connection to the Port of Naantali further indicates its importance for the Finnish shipping industry and corroborate it being chosen of all possible ports.

The port is defined, for the scope of this work, by manually defining a area that covers the whole operational area of the port, to include all berths and the possible routes to approach the port. Port of Naantali is for all vessels except small pleasure boats approachable from one direction, see Figure 8. The bounding box in red in Figure 8 is what defines the area of when a vessel has entered the port *i.e.* what defines a port. For the scope of this thesis, it is the port authorities responsibility to ensure vessels are allowed to berth as close to the ETA as possible or if not possible, communicate another ETA for the vessel so it can adjust its speed and manage just in time arrival (JIT).

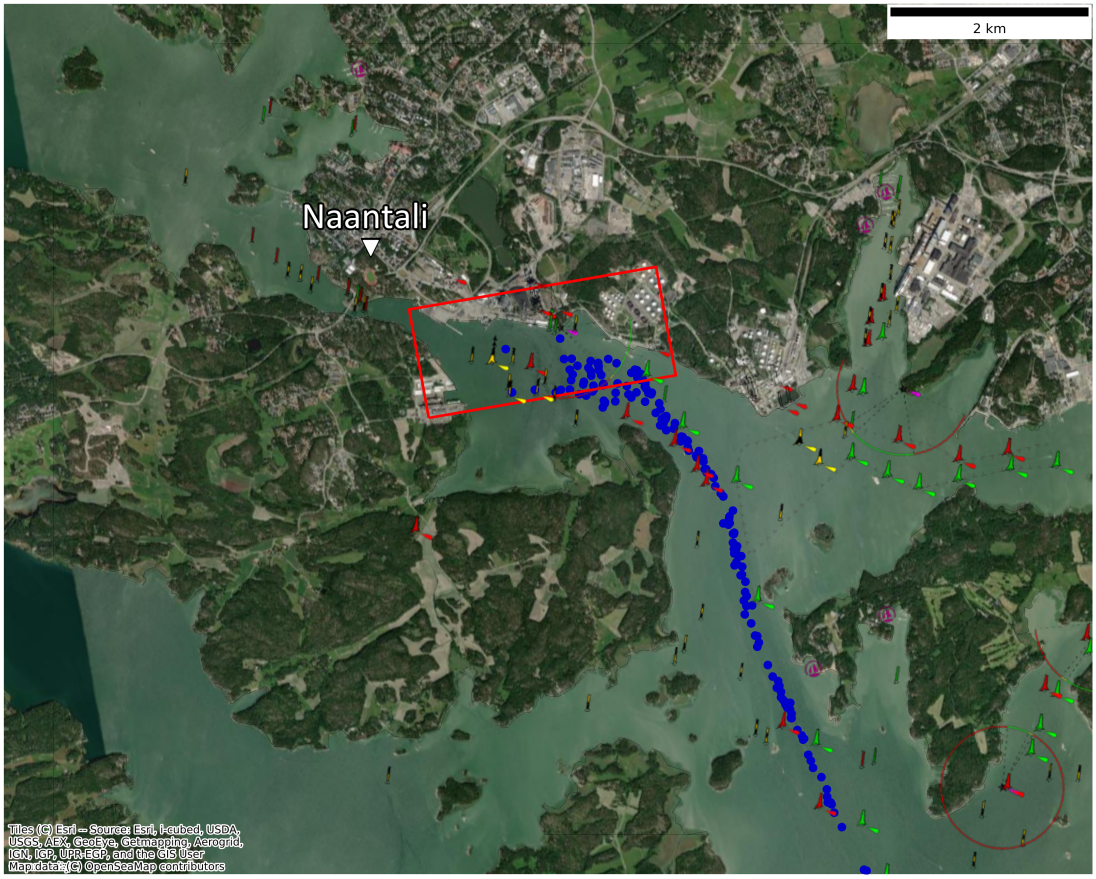

Figure 8: Defined bounding area for the Port of Naantali outlined in red. A small sample of the incoming routes shown in blue.

3 Recurrent Neural Networks

Introduction to what neural networks and more precisely recurrent neural networks and why they are ideal for solving this problem (remembering the order of inputs and time is linear which means all previous timesteps have an impact on the next)

3.1 Long Short Term Memory

Description of what exactly LSTM networks are

4 Implementation

Python

4.1 Libraries used

Keras

Tensorflow (GPU)

Pandas

Numpy

Geopandas

Packages and programming language used.

4.2 Data pre processing

The years 2009 through 2018 is chosen to train and evaluate the performance of the model. The year 2019 was omitted due to unusual file sizes, which lead to the belief that the files for that year are not complete. All these files total approximately 130 gigabytes, which will introduce some limitation on how much of the data is processed at once. The process of finding possible routes in this raw data for further processing and validation, involves finding all vessels that have visited the port of interest at any point. Due to the file sizes and amount of data for each file this process has to be split into smaller chunks of time windows, *i.e.* the number of consecutive months read in at once and processed. With the possibility of vessels starting a route at the end of a month and arriving at the beginning of the next month, larger time windows are preferred, but memory limitations impeded the possibility of very long time windows. Splitting the year into quarters proved to be a good compromise computationally wise.

First step involves finding which vessels have visited the port of interest, the port of Naantali, for each month. This step is quite computationally time consuming, since every row in each file has to be checked, except rows of a previously known vessel. Only once a vessel has been identified to have visited the port of interest it can be skipped, else the row has to be checked. It is not possible to known at what point in time a vessel has visited the port of interest, so every row has to be checked and since a vessel could have only made one visit to the port for the whole period covered, it is unnecessary to get the whole timeline of said vessel. Rather, only getting the timeline of when the vessel has visited the port can be fetched, drastically decreasing the required amount of data. The process involves checking each rows latitude and longitude and whether these coordinates are within the bounding box seen in Figure 8. This process has to be done once for each file,

after which the unique identifiers, *imo*, for every vessel can be stored for future use. Every month can have more than 30 unique vessels, a quarter can have around 80 unique vessels, and the whole timeline is extracted for all vessels for the quarter.

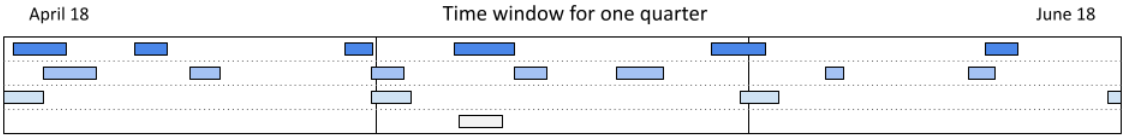

Figure 9: Example of one quarter of the the year 2018, a *time window*, and from this time window each unique vessels complete timeline is extracted for further processing. The coloured boxes represents raw AIS data for four unique vessels. Does not represent actual data, rather visualizing the process of finding relevant data in the raw HELCOM dataset.

From the complete time window seen in Figure 9 only a fraction of the data is actual viable routes going to the port of interest. Merging the complete time window with AIS data for every vessel generates a timeline for each of the unique vessels discovered in the first step. Within that timeline, at some point a or many routes going the port of interest exists and the next section describes how the routes are extracted and validated.

4.2.1 Algorithm for extracting routes from dataset

A route is defined as a series of consecutive AIS messages where the first message in the series is the starting point for a vessel travelling to a port of interest. The last message in the series is the first messages when the vessel has entered the port of interest bounding area, see Figure 8. The routes does not need a specific starting point as this point bound to any position, rather it will get chosen from one of two possibilities. The end of the route will be the same for all vessels as the goal is to find all routes coming from somewhere travelling to the port of interest.

To find routes coming from anywhere and going to a area of interest or port a iterative approach is used. With a complete historical timeline of one vessel, the routes can be discovered by traversing the timeline in reversed chronological order. The end of a route, *i.e.* the start of the search, is the first point where the vessel is leaving the bounding box. From this position, traversing the timeline, in reversed chronological order, until a predefined condition is met will give the section of a vessels timeline which equals a route coming from somewhere travelling to a area of interest.

The condition for when the start of the route, end of the search, is met, is whenever the duration of the route has exceeded 48 hours or the vessels speed over ground

is zero, standing still, for more than 5 consecutive transmitted messages. The 48 hour time limit was defined for the reason that the maximum travel duration in the Baltic Sea is below 48 hours for a vessel moving at the mean speed discovered and vessels travelling from outside the Baltic Sea has reached Kattegat and the edge of the Baltic Sea Area seen in Figure 2 at this point in time. The reason behind checking that multiple transmitted messages have recorded the speed to be zero is to guarantee that the vessel has actually stopped and not slowed down due to other reasons. This condition can still introduce routes that have not started from another port but will minimize these routes.

Algorithm 1: Find all routes going to a area of interest

Input: DataFrame for one vessel sorted in descending time, a *timeline*
Result: List R of all routes found
The algorithm searches in reverse order of time from reaching the destination until the start of the route;

```
foreach row in DataFrame do
    index ← Keep track of current row;
    if route found then                    // Only true after finding first route
        Start searching from first unknown point;
        Skip rows until index = start;
    if current point is in PORT then
        foreach row in DataFrame starting from index do
            if point is outside PORT then      // Vessel is entering the port
                end ← First point reaching the destination;
                foreach row in DataFrame starting from end do
                    if start of route reached then
                        start ← Current row;
                        R ← Save route from end to start;
    /* When a route has been found and saved start search
       again from the last not visited point.                    */
```

Route validation rules were defined to give an efficient way of discarding any routes that would impact the models performance and routes that were off no interest for the scope of the thesis.

All routes that did not travel for more than eight hours from start to finish were discarded. The reason for this rule comes from the fact that all vessels will be required to travel through the same area for the final part of the voyage, and the model should perform well in this area even though these routes are discarded.

A minimum distance travelled rule was used to discard any routes that travelled less than 200 kilometres, that is the total route distance and not how far away from the port the vessel travelled. The navigate out from the port of Naantali and the largest bodies of islands outside of Naantali takes approximately 70 kilometres following the fairway. Moving at 14 knots the time to cover the distance takes just under three hours, however, the speed in this area will most likely be lower for most vessels due to the limited area to manoeuvre.

Any routes that have a time difference between two consecutive AIS messages greater than 12 minutes is also discarded. The reason for this is explained more in details in Section 4.2.4.

Vessels that have returned to the port of Naantali or are returning to the area around Naantali are also discarded. This combined with the minimum distance travelled required will ensure that, for example pilot vessels and other leisure vessels, will not be in the final training data. Vessels travelling back to the area are not of interest and could impact the model negatively, as these vessel will probably have an abnormal voyage compared to cargo vessels.

Routes that have faulty data, *i.e.* NaN values or other bad values, and which can not be inferred from the data in any other way is discarded. Examples of this was found where the draught was not recorded only for a small section of the route but could be inferred from the parts where it was recorded. If the values could not be inferred at all the route was discarded completely.

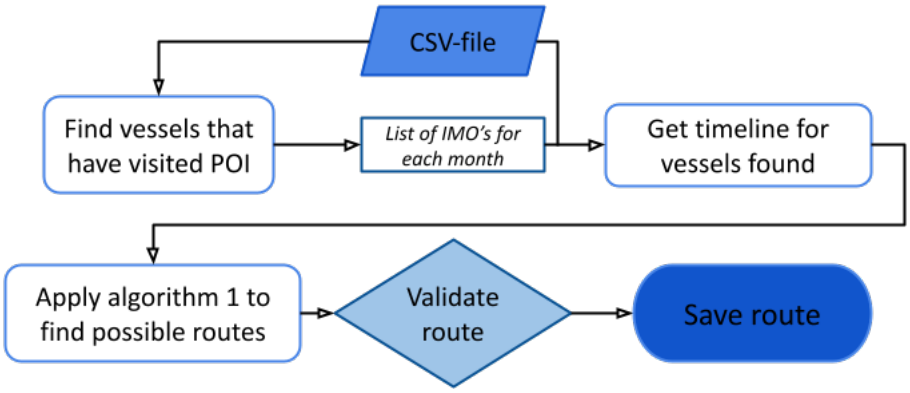

Figure 10: Getting routes.

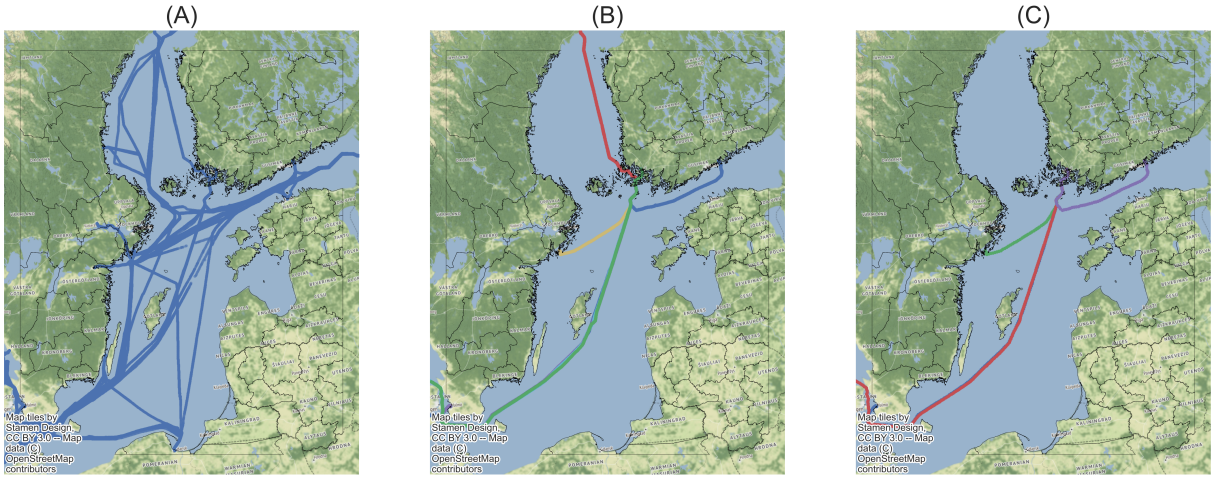

Figure 11: (A) Complete timeline for one vessel year 2013. (B) All possible routes for the same vessel. (C) All validated routes which are used for training.

The number of routes found from each complete year was at average 1395 routes.

The total number of routes for the period chosen was 13,955. However, only a fraction of these routes are valid routes. The final route validation and which rules are defined to find viable routes are explained in the next section. What range of data was used for the final model and reason for so. The problems with AIS data and getting clean routes from start to finish and primarily enough routes for training and testing.

POI (port of interest could also be thought of as Area of Interest). Timeline is the vessels historical data which is fed into algorithm 1. Route validation according to rules and then save the complete route.

4.2.2 Coordinate accuracy

The latitude and longitudes in the HELCOM dataset has an accuracy of six decimal places, 0.111 meters. This accuracy is not required as the difference in time between a lower accuracy will be within a possible prediction error.

The level of scaling the coordinate accuracy with, was chosen to a decimal degree of three places. That will say 0.001 which is 111 meters. The accuracy of the vessels location will be within approximately 100 meters.

Degrees	Distance
1.0	111 km
0.1	11.1 km
0.01	1.11 km
0.001	111 m
0.0001	11.1 m
0.00001	1.11 m
0.000001	0.111 m

Table 5: Coordinate accuracy by decimal places in decimal degrees for latitude and longitude, from [12]

AIS messages transmitted within approximately 100 meters of each other will be thought of as sent from the same location using this scaling level. With a mean speed for the routes found and used, see Figure 4, of approximately 14 knots (25 km/h) and a time difference between messages of 10 minutes a vessel will have moved 2.2 nmi (4.1 km) during this time. With a positional accuracy of approximately 100 meters this distance can vary but on the open sea a difference in $\pm$ 100 meters does not impact the overall time it will take to reach the destination.

4.2.3 Feature selection

The features chosen for training the model on was decided by what is available in the HELCOM dataset, what is possible to infer from available data and what has been proven to be efficient features by previous studies [13, 9].

Most of the features available in the HELCOM dataset was used for the model; latitude and longitude for the position of the vessel at a given point in time, sog to indicate the vessels moving rate, cog for where the vessels trajectory is heading, draught to indicate the vessels physical capabilities in combination with the vessel class feature, which is a combination of features, see Figure 6 and finally the distance to the destination calculated from the historical route the vessel has taken.

Using the distance left to the destination as a feature does mean that the problem to be solved is in its simplicity learning the time it will take to traverse the distance left at the current speed. However, there are many factors that will impact the time it will take to reach the destination which can not be learned from only the distance to travel and the speed. It is these factors that could be learned by training on historical data and improve the prediction accuracy compared to other simpler solutions.

The target to predict is time to destination (TTD) which is in minutes. The TTD prediction is added to the current time and thus give an ETA. The TTD is calculated as the difference between the current time and what time the vessel has arrived at port.

4.2.4 Time series data windowing and time distribution

With the routes validated a final processing step is required to generate the time series data that is fed to the network. Since regular RNN suffers from poor performance when the data is unevenly sampled in terms of time difference between samples, it is vital to normalize the time difference between time steps. The nature of AIS and the HELCOM dataset means that it is only viable to normalize the sampling interval to a certain degree.

A compromise with the time difference between messages was achieved by as evenly as possible sample data with longer time intervals, discarding messages in that interval, and by this achieving a improved time series with preferable time difference normalization. In this improved time series the difference between messages is on average at most three minutes. A lower limit of nine minutes was chosen to conform with what was discovered in Section 2.2.3 and Figure 5 and the upper bound of twelve minutes has been set when validating the routes. On average, two messages will get discarded so the time difference is nine minutes. Shorter time intervals should be possible with raw AIS data, but for the dataset used the number of routes that were viable without generating data between messages were too few. Decreasing the longest allowable gap between AIS messages in a route discarded too many routes, *i.e.* a route with at any point a time difference between messages smaller than twelve minutes would have been rejected.

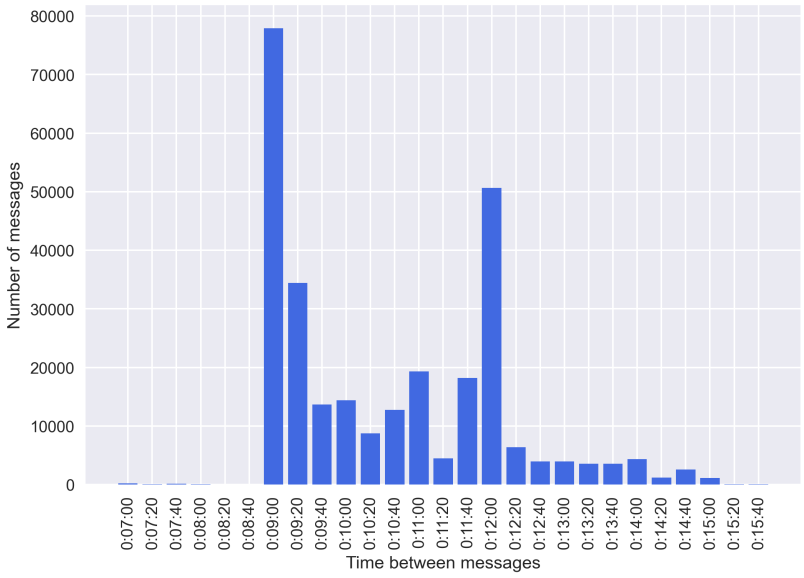

Figure 12: The distribution of time differences between messages for the normalized routes.

In Figure 12 the normalized time differences are in the range from nine to twelve minutes with only some exceptions that have a time difference shorter than or longer than this range. The exceptions are for the most part due to the inconsistencies in

the recorded time intervals, *e.g.* given one message at time zero t_0 with the following messages at t_{+4}, t_{+4} and t_{+7} (minutes from the previous message), the first two messages will not be more than nine minutes from the first and the third message will therefore be chosen as the next point in time, with a 15 minute difference.

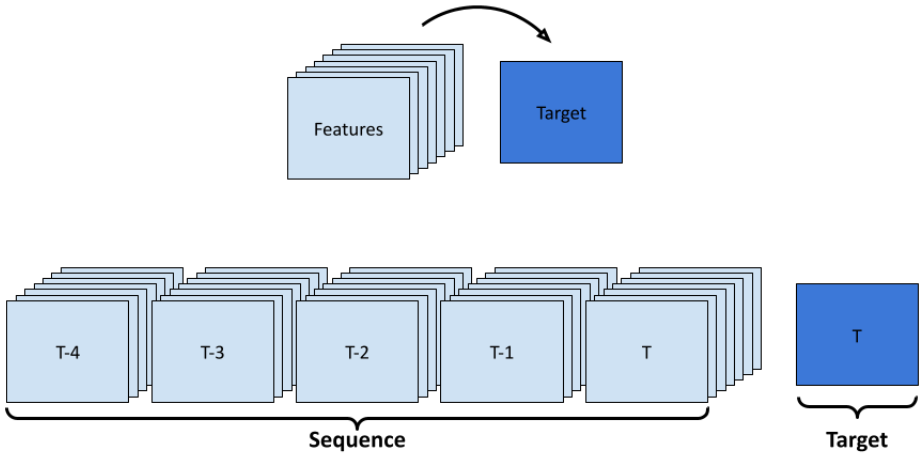

Figure 13: Timeseries sequenced data, with a time window width of five. Each timestep in the sequence has a time difference in the range defined.

With a route consisting of AIS messages with a time difference distribution in the range mentioned above, a time window of n timesteps is generated for the whole route. A time window is defined as n number of consecutive timesteps that create a window of time for the route. Each timestep in this time window consists of all chosen features and the target TTD. For all but the last timestep, the target is discarded and only the final timesteps target is set as the target to predict. Sliding the time window over the whole route one timestep at a time generates a set of windows that each covers n timesteps.

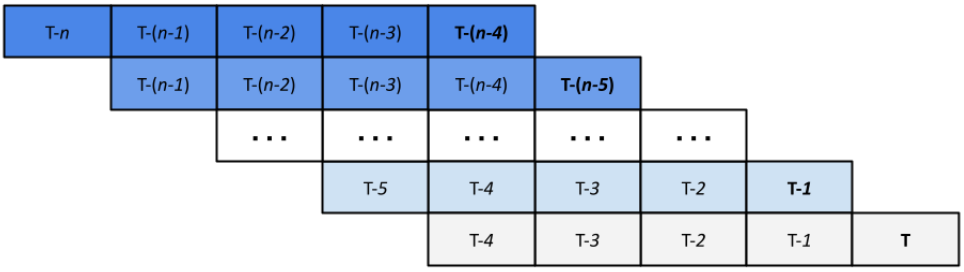

Figure 14: Sliding time window for a complete route starting from the end at $T - n$ and ending at the last timestep T. A prediction will be made for every last timestep in the time window, marked with bold.

A prediction will be made for the last timestep in the time window, which in

terms of real time prediction will require that for the first four timesteps, AIS messages received, a prediction can not yet be made. Only after the fifth message can a prediction be made. This also means that any five historical AIS messages could be used, if they fulfil the requirements for the time window.

4.3 ML model

Description of the neural network model used and tested to find the optimal performer

4.4 Comparison model

!!! If the travel distance left to destination is used in nmi for example, test the accuracy against simply calculating time left by the current speed and distance left

5 Results

5.1 ETA prediction

5.1.1 Biased training data by direction

5.2 Navigation in the archipelago

Models difficulties to predict ETA within the archipelago

6 Discussion

Discussion about the results and validity future work

7 Conclusion

The possibility to use historical AIS data, example HELCOM dataset, to train a model on predicting the ETA for ports that have good coverage of vessels inbound or some amount of routes coming to the port

References

[1] Suomen Varustamot Ry, *Merenkulun avainluvut*, 2022. [Online]. Available: https://shipowners.fi/kilpailukyky/merenkulun-avainluvut/ (visited on 04/01/2022).

[2] International Maritime Organization, *Revised guidelines for the onboard operational use of shipborne automatic identification system*, 2015. [Online]. Available: https://www.imo.org/en/OurWork/Safety/Pages/AIS.aspx (visited on 12/14/2021).

[3] "Commission Directive 2011/15/EU of 23 February 2011 amending Directive 2002/59/EC of the European Parliament and of the Council establishing a Community vessel traffic monitoring and information system," 2011. [Online]. Available: https://eur-lex.europa.eu/eli/dir/2011/15/oj.

[4] A. Veenstra and R. Harmelink, "On the quality of ship arrival predictions," *Maritime Economics & Logistics*, vol. 23, no. 4, 655–673, 2021. DOI: 10.1057/s41278-021-00187-6.

[5] "Directive 2009/16/EC of the European Parliament and of the Council of 23 April 2009 on port State control," 2009. [Online]. Available: http://data.europa.eu/eli/dir/2009/16/oj.

[6] G. Fancello, P. Claudia, M. Pisano, P. Serra, P. Zuddas, and P. Fadda, "Prediction of arrival times and human resources allocation for container terminal," *Maritime Economics & Logistics*, vol. 13, pp. 142–173, 2011. DOI: 10.1057/mel.2011.3.

[7] A. Harati Mokhtari, A. Wall, P. Brooks, and J. Wang, "Automatic Identification System (AIS): A Human Factors Approach," 2008.

[8] HELCOM, *Convention on the Protection of the Marine Environment of the Baltic Sea*, 2014. [Online]. Available: https://helcom.fi/about\-us/convention/ (visited on 03/29/2022).

[9] C. Jahn and T. Scheidweiler, "Port call optimization by estimating ships' time of arrival," in *Dynamics in Logistics*, M. Freitag, H. Kotzab, and J. Pannek, Eds., Springer International Publishing, 2018, pp. 172–177. DOI: 10.1007/978-3-319-74225-0.

[10] *Total transports in the Port of Naantali over 8 million tonnes in 2020*, 2021. [Online]. Available: https://portofnaantali.fi/en/press-release/total-transports-in-the-port-of-naantali-over-8-million-tonnes-in-2020/ (visited on 03/23/2022).

[11] Traficom, *Vesikuljetusten Kuljetusmäärät*, 2021. [Online]. Available: https://tieto.traficom.fi/fi/tilastot/vesikuljetusten-kuljetusmaarat (visited on 04/01/2022).

[12] Wiki.GIS.com, *Decimal degrees*, 2011. [Online]. Available: http://wiki.gis.com/wiki/index.php/Decimal_degrees (visited on 04/04/2022).

[13] S. El Mekkaoui, L. Benabbou, and A. Berrado, "Predicting ships estimated time of arrival based on ais data," in *Proceedings of the 13th International Conference on Intelligent Systems: Theories and Applications*, Association for Computing Machinery, 2020, ISBN: 9781450377331. DOI: 10.1145/3419604.3419768.