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

Vessels navigating the Baltic Sea

Importance of maritime traffic globally and timely operation

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 [1]. 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 closer vessels takes precedence over vessels farther away which allows for sharing time slots, and thus overloading the available time slots [1].

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 an AIS May 2014 at the latest [2]. 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 receive of data, for updating information manually and retrieving data received via AIS.

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.

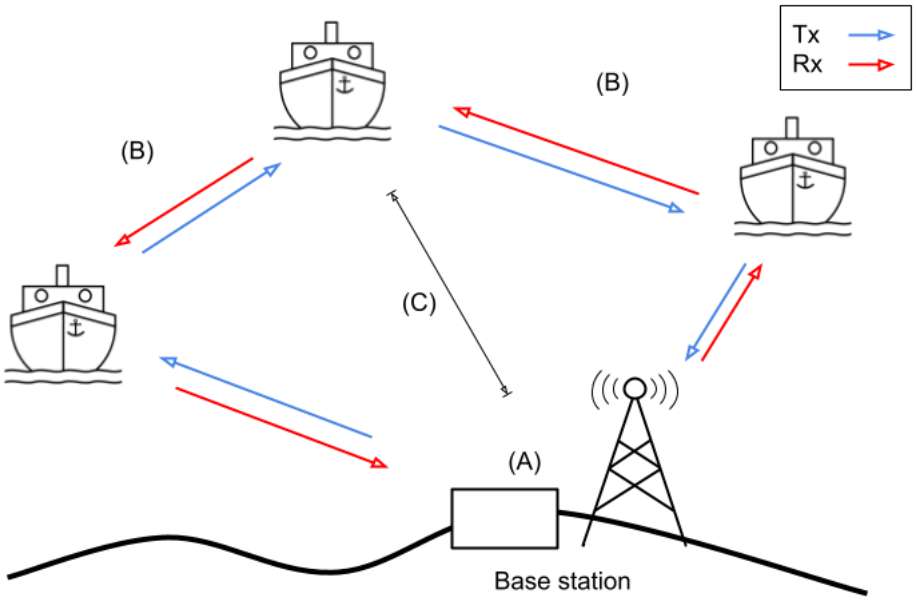

Figure 1: (A) Base station transmits information about other vessels, port data and possible hazards. Base station receives information from all vessels in range. (B) Ship-to-ship communication. Vessel broadcasts its current state (identifier, speed over ground, course over ground etc.) and receives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

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, from [1].

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 [3, 4]. 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 [5]. All uncertainties that can delay or alter any operations at port are therefore a risk.

2.2 HELCOM dataset

The HELCOM dataset 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 is already processed to some extent compared too raw AIS data, see example of raw AIS data in Section **ref to raw ais**.

Describe the features and quirks with the HELCOM dataset as it is a processed dataset that contains AIS data already processed from the raw format. Merging of A and B AIS message types

HELCOM dataset has no ETA entries meaning to use this dataset for ETA prediction the ATA is calculated for all points on the route until the vessel has arrived, the true arrival time. Can't compare vessels ETA with the estimation but knows the estimated time left

Showing examples of the data as plots of the Baltic, also shows the area covered

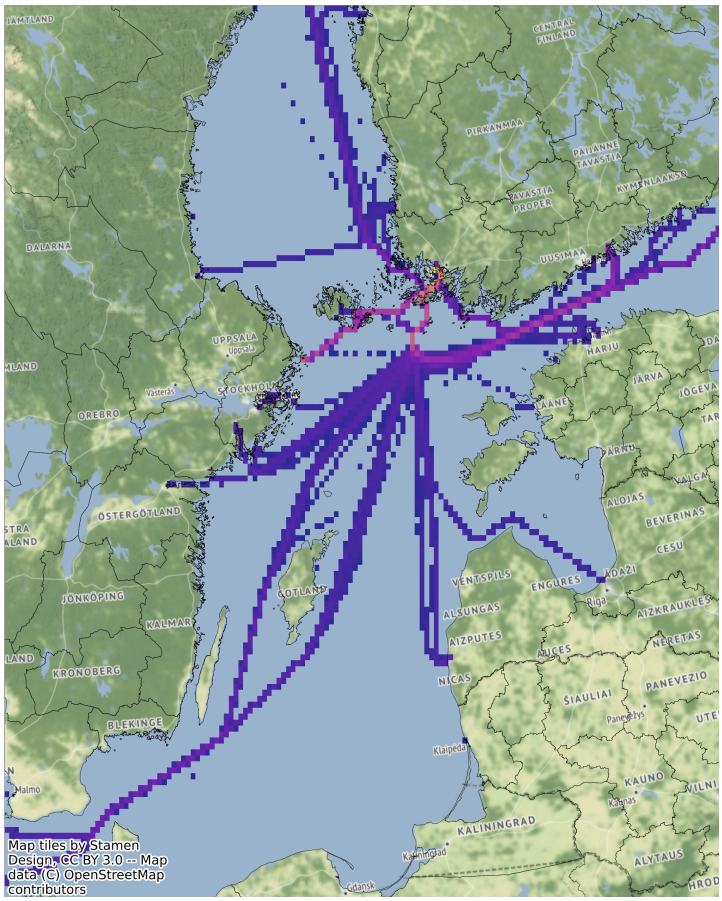

Figure 2: Heatmap the Baltic Sea with a limit of more than 10 unique messages per grid cell.

2.2.1 Description of HELCOM dataset features

The HELCOM dataset have twelve unique features for each entry described in Table 2. The features are static, dynamic and voyage-related AIS data merged from multiple sources.

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 digits long
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 2: Variables in HELCOM data set and description.

2.2.2 Statistical information

Unevenness of the recorded messages (time intervals distribution even) but large gaps in routes. Reasons can be merging of many databases, preprocessing of the data to unify the records. The number of "good" valid routes for training is much lower than the total number of routes. Some of that could be handled by generating routes (data) where there are gaps. In a realistic dataset with AIS data there are going to be faulty data and missing, so HELCOM data is quite accurate to real data without generating missing data.

Vessels physical features and the grouping of different vessels by size. In Figure 4 an example of one year of unique vessels which have visited the port of Naantali can be seen.

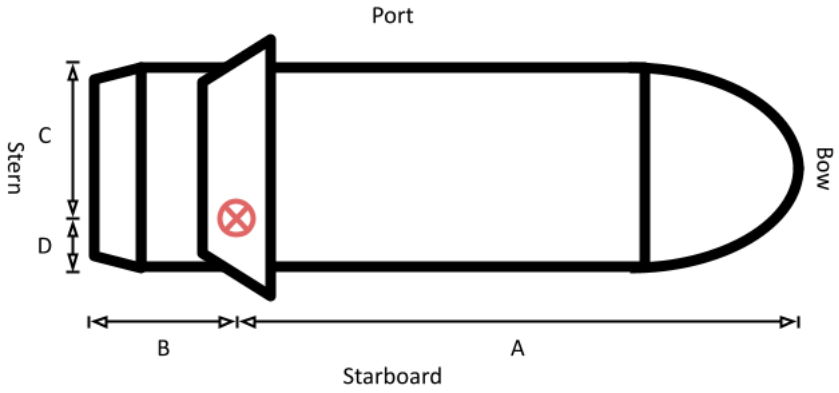

Figure 3: Vessel class definition from HELCOM data, the positioning system marked with red. A is *dimStern*, B is *dimBow*, C is *dimPort* and D is *dimStarboard*.

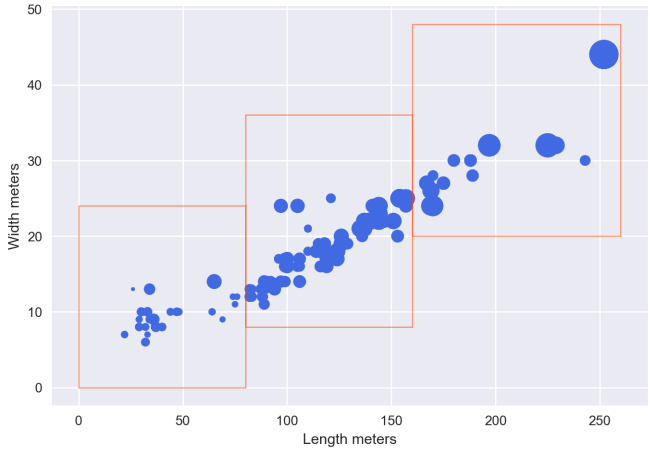

Figure 4: Defined vessel classes and distribution of all unique vessels for the year 2015. Size of the circle is related to the draught of the vessel.

2.3 Port of Naantali

The Port of Naantali is a rather important and busy port in the Baltic Sea. 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 [6].

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

This data is the complete timeline for each vessel that has at any point during the years visited the Port of Naantali. Only a fraction of this data is actual viable data for the routes, described in **SECTION REFERENCE**.

The port is defined 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 5. This bounding box in red in Figure 5 is what defines the area of when a vessel has entered the port.

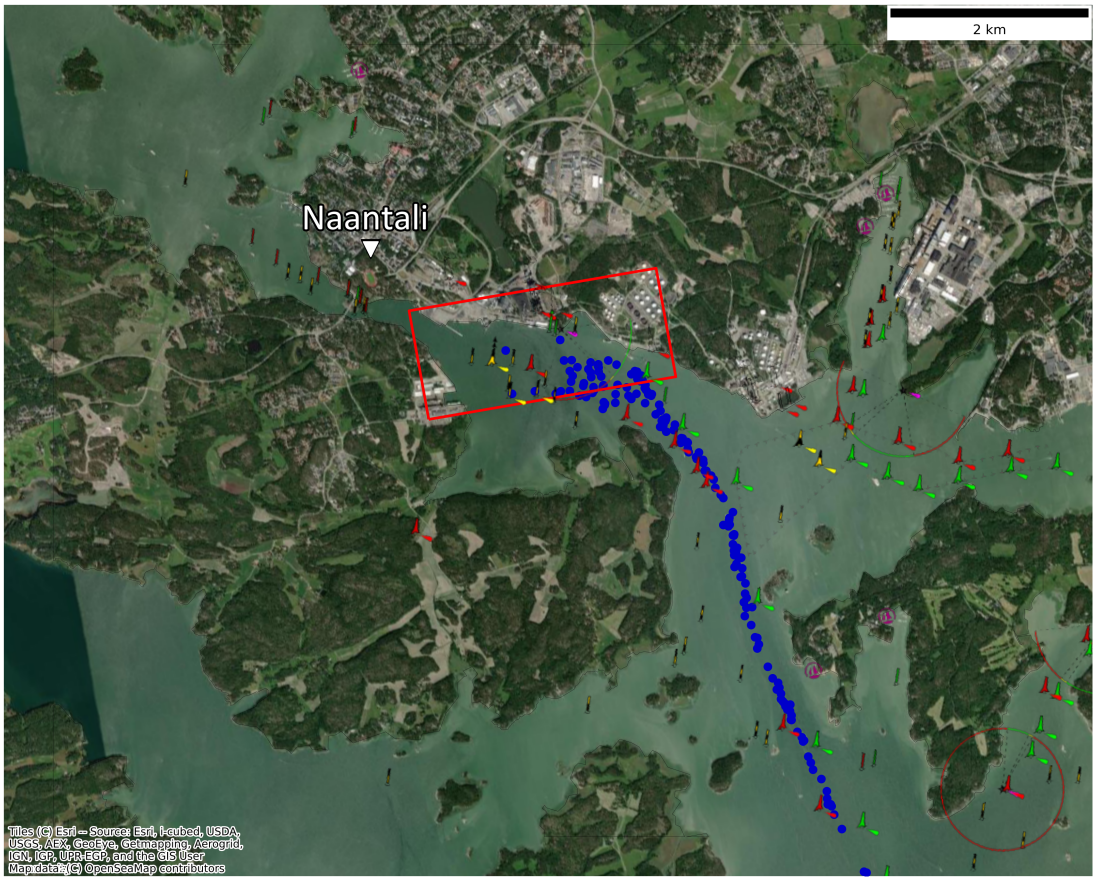

Figure 5: Defined bounding area for the Port of Naantali outlined in red. A small sample of the incoming routes shown in blue.

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 model. 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. 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. Due too the possibility of vessels starting a route at the end of a month but arriving the first day of the next month, larger time windows are preferred but limitations with memory limit this. Splitting the year into quarters proved to be a good compromise.

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. The process involves checking each rows latitude and longitude and whether the point is within the bounding box seen in Figure 5. This process has to be done once for each file, after which the unique identifiers for every vessel can be stored for future use. The raw files containing millions of row of data at average and begin around one gigabyte on file size limits the number of files that can be processed at once. Every month can have more than 30 unique vessels and the whole timeline has to extracted for all vessels.

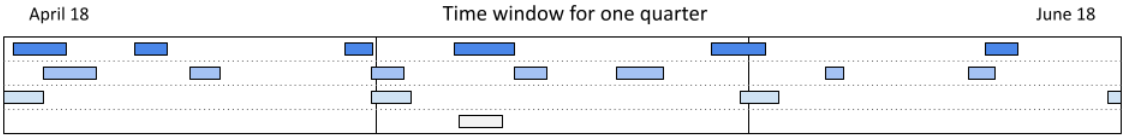

Figure 6: Example of one quarter of the the year 2018, a *time window*, and from this time window each unique vessels complete timeline is extracted for further processing. The coloured boxes represents raw AIS data for four unique vessels. Does not represent actual data, rather visualizing the process of finding relevant data in the raw HELCOM dataset.

From the complete time window seen in Figure 6 only a fraction is used for the final route finding algorithm, see Table 3. Joining this raw AIS data for every vessel found generates a timeline for each vessel.

4.2.1 Algorithm for extracting routes from dataset

Explain the algorithm behind getting the routes from the raw data set and that the idea can be applied to any port or area of interest really by defining the area of interest.

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. */

The end of a route is either vessels sog less than 0.1 for more than 5 consecutive messages or has travelled longer than 48 hours

Duration of route travel time

Minimum distance travelled

The largest gaps allowed 12 minutes

Travelling back to FINLI

Faulty data that can not be inferred from other data example draught

The port chosen for the evaluation has its caveats and should perhaps be identified.

Validation of the routes extracted i.e gaps, no shorter than n number of messages where the vessel is going, not returning to port, longer than n hours, no faulty data etc.

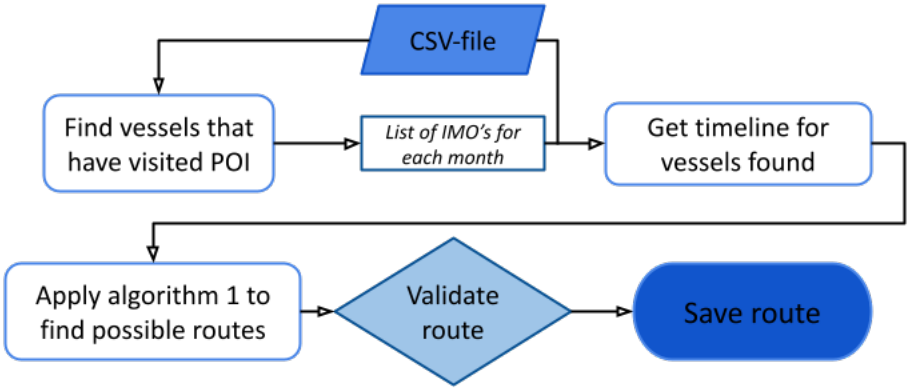

Figure 7: Getting routes.

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

Scaling the accuracy from raw data to a suitable accuracy. GIS decimal degrees

Vessel travelling at 14 knots (25 km/h) and a message interval of 10 minutes approx. will move 2.2 nmi (4.1 km) per message

4.2.3 Feature selection

The features chosen for the final training

Latitude, longitude, sog, cog, vessel class, draught, ?distance?

The target is Time To Destination, how many minutes are left to the destination

4.2.4 Time series data preparation and cleanliness

Further details on how the processed routes are handled to generate the training data to utilize multiple timesteps per prediction which improves performance. Also

why the number of timesteps per prediction was chosen, with the time normalized data and how many steps then per time window.

The largest allowed difference between messages in the time normalized data

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

5.2 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] Internation 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).

[2] "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.

[3] 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.

[4] "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.

[5] 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.

[6] *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).