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 **1**

2 Maritime vessels and traffic **2**
2.1 Automatic identification system 2
2.1.1 Estimated Time of Arrival 2
2.2 HELCOM dataset 2
2.2.1 Description of HELCOM dataset features 3
2.2.2 Statistical information 3
2.3 Port of Naantali 4

3 Recurrent Neural Networks **5**
3.1 Long Short Term Memory 5

4 Implementation **6**
4.1 Libraries used 6
4.2 Data pre processing 6
4.2.1 Algorithm for extracting routes from dataset 6
4.2.2 Coordinate accuracy 7
4.2.3 Feature selection 7
4.2.4 Time series data preparation and cleanliness 7
4.3 ML model 7
4.4 Comparison model 8

5 Results **9**
5.1 ETA prediction 9
5.2 Navigation in the archipelago 9

6 Discussion **10**

7 Conclusion **11**

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

What is AIS, what is is features, problems. Example of raw AIS data.

Type A and B message types

2.1.1 Estimated Time of Arrival

Importance of ETA and what current systems there is.

2.2 HELCOM dataset

Introduction to HELCOM

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

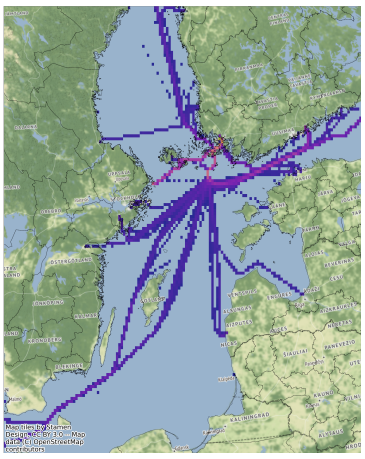

Figure 1: Heatmap the Baltic Sea with a limit of more than 10 unique messages per grid cell.

2.2.1 Description of HELCOM dataset features

The HELCOM dataset have twelve unique features for each entry described in Table 1. There are bad rows in the dataset, i.e. rows that have a or many bad values.

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 1: 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.

2.3 Port of Naantali

The amount of data collected travelling to FINLI from the HELCOM dataset, rather busy port approx. 2 % of the total data per month

The projects definition of the port, the bounding box with image.

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

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.

Processing the HELCOM data to extract routes that can be used to train a model with the goal in mind.

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.

The largest gaps allowed

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.

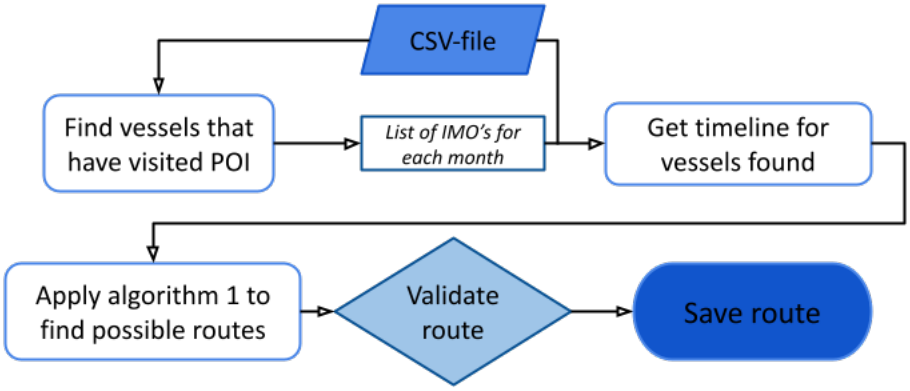

Figure 2: Getting routes.

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