Journal for Mobility and Transport
ISSN 2628-4154
www.dvwg.de

Scalable In-Database Machine Learning for the Prediction of Port-to-Port Routes

Dennis Marten[1], Carsten Hilgenfeld[2] , Andreas Heuer[3]

1 Research and Development Department JAKOTA Cruise Systems GmbH | FleetMon Rostock, Germany marten@fleetmon.com

2 Research and Development Department JAKOTA Cruise Systems GmbH | FleetMon Rostock, Germany hilgenfeld@fleetmon.com

3 Chair of Database and Information Systems Institute of Computer Science, Rostock University Rostock, Germany heuer@informatik.uni-rostock.de

Abstract

The correct prediction of sub sequential port-to-port routes plays an integral part in maritime logistics and is therefore essential for many further tasks like accurate predictions of the estimated time of arrival. In this paper we present a scalable AI-based approach to predict upcoming port destinations from vessels based on historical AIS data. The presented method is mainly intended as a fill in for cases where the AIS destination entry of a vessel is not interpretable. We describe how one can build a stable and efficient in-database AI solution built on Markov models that are suited for massively parallel prediction tasks with high accuracy. The presented research is part of the PRESEA project (“Real-time based maritime traffic forecast”).

Keywords: Digitalization, Big Data, Artificial Intelligence, Markov Process, Port Destination Prediction, Traffic Management

1. Introduction and Motivation

According to the UN more than 80% of the worlds merchandise trade has been carried by sea in 2019 [1]. This shows the integral part maritime logistics plays in world economics and motivates the continuous pursuit for logistical optimization.

The project PRESEA ("Real-time based maritime traffic forecast") aims to support this cause by developing a real-time based forecast for global maritime traffic that will be integrated as a service in FleetMons infrastructure [2]. In particular, a routing network is developed that is intended to incorporate weather conditions and specific events like ships accidents. The corresponding forecast system will allow shipping companies to optimize their organization of just in time delivery, which will also optimize the fuel demand of specific vessels which can ultimately reduce the emissions of the respective ships. Furthermore, maritime security authorities can easily obtain detailed information on expected traffic volumes.

Project partner in PRESEA is the Institute for Safety Engineering / Ship Safety e.V. (ISV) located in Warnemünde (Germany). The Laeisz shipping company, Synfioo GmbH, the classification society DNV-GL and Daimler AG have pledged their active support for this project.

In this paper we present an approach that improved one basic but key aspect of this forecast system using AI technology: the accurate prediction of next port destinations or even whole subsequent port-to-port-routes. Currently, FleetMons port destination prediction is

The PRESEA project is funded by the German Federal Ministry for Economic Affairs and Energy (BMWi). The project management organisation is administrated by the Project Management Jülich (PtJ) within the framework of the call "Real-time technologies for maritime security". The project is running from June 2019 until lNovember 2021.

based on the interpretation of AIS ("Automatic Identification System") data sent from vessels and their last identified visited port.
As will be described in detail in Section II this approach harbors several challenges that are unlikely to be solved with logical approaches thus motivating a (statistical) AI based approach.
The rest of the paper is structured as follows. In the following Section III a short summarization of the state of the art is presented. In Section IV the use of Markov processes is motivated and discussed for the presented context. An evaluation of the derived models are presented in the following Section V. Finally, a short description of future projects and possible improvements of the presented work is given in Section VI.

2. AIS Data and Interpretation Challenges

Firstly, we would like to discuss the downsides of deriving the port destination from AIS data and motivate the partial use of AI in order to overcome problem cases. Therefore we give a small overview on the Automatic Identification System and challenges we have faced at inferring port destinations from its data

A. *The Automatic Identification System (AIS)*.

Currently, the main basis for the prediction of a vessels port destination is data the respective ship sent by means of the AIS. AIS is now standard equipment for all ships over 300 gross tonnes in international voyages. Via VHF, a ship transmits AIS data for its own identification and essential voyage information. The data is received from other ships and is integrated on board in an electronic navigational chart (ECDIS), which allows surrounding ships to be identified and thus the assessment of the overall navigational situation. At the same time, AIS data can also be received from satellites or shore stations and can be merged and visualised by corresponding providers. This enables a worldwide display of all ship movements. The content of the minimum data to be transmitted is internationally prescribed and comprises the following three groups:

- **Static data**: IMO number, ship name, call sign, MMSI number, type of ship, dimensions of the ship
- **Dynamic ship data**: Navigational status, ship position, me of ship position, course over ground, speed over ground, forward direction, rate of course change
- **Voyage data**: current maximum static draught, dangerous goods class of cargo, destination, estimated time, of arrival (ETA).

The destination and the expected time of arrival are manually set by the vessels navigator which is often the cause of non-matchable port destinations.
FleetMon operates one of the world's largest AIS networks consisting of thousands of globally distributed terrestrial AIS antennas as well as satellite data provided by the three largest AIS satellite data providers and several AIS research satellite constellations. While receiving, storing and processing over 480 million AIS messages a day from up to 225 thousand vessels, we need to make sure that PRESEAs routing system can accurately predict large parts of port destinations of the global fleet in order to allow precise traffic forecasts within the system. In the following we describe some challenges we have faced while interpreting the AIS destination data.

B. *Challenge in the interpretation of the AIS destination*

Currently, FleetMon uses a complex set of logical rules based on string matching that searches for identifiable ports in the destination entry. As vessels with fixed port-to-port cycles (for instances ferries) often use static entries in the form of "port 1 <-> port 2" we use information of the last known port call to identify which one of these ports is the actual destination. Anyhow, as any string sequence can be entered in the AIS destination text field, we faced a multitude of possible misspellings or misuse of the field. Trying to cope with these by mapping misspelled port names to the originally intended LOCODE turned out to be insufficient.
Besides it cannot be guaranteed that each voyage related AIS data set, which is broadcast only every 6 minutes, is received by an AIS network. Furthermore, when not updated correctly it is possible that the current destination

entry is not matching with the real port destination. All these points leads to a number of challenges that have occurred over time. Some of them will be described now by way of example:

- The ship reports "CNSHA USLAX" (for the journey from Shanghai to Los Angeles). However, the ship does not have its last port call in Shanghai, but in Hangzhou (CNHAZ). So, it is not possible to determine which was the last port and therefore which is the next port, considering the AIS destination
- Due to missing AIS coverage in the port no port call could be generated. Therefore, it is not possible to determine the last port if two LOCODEs are entered in the destination.
- Incorrectly spelled LOCODEs, for instance KRBUS for Busan (correct would be KRPUS)
- Different kinds of misspelled city names like Philadelphia
- Several ports in the world have the same city name, for example Cartagena (Spain or Colombia) or Sydney (Canada or Australia)
- The use of port name variations in different languages, for instance: Brugge (German, Dutch), Bruges (English, French, Portuguese), Brič (Macedonian, Serbian), Bruggia (Italian), Bruggy (Slovak), Brugia (Polish), Brugy (Czech), Brujas (Spanish), Brygge (Finnish). The correct name for the Port is in English Zeebrugge or Seebrugge

These examples lead to the fact that with the method used less than 80% of the destination sent in AIS can be correctly interpreted by ships over 100m. This means that for more than 20% of the ships no estimated time of arrival in the next port can be calculated.
With these restrictions in mind the rest of the paper is dedicated to present and evaluate an AI based approach to allow meaningful and efficiently computable predictions of a vessels port destination(s) without relying on AIS destination entries.

3. State of the Art

To the best of our knowledge there has been no work published for the concrete prediction of port-to-port routes via Markov models based on historical and current AIS data.
Contrary, indepth research has been done in the detection of anomalous in vessel behavior, the prediction of vessel routes or the prediction of the estimated time of arrival (ETA). Here we would like to list a few prominent representatives of this work.
In [3] an incremental statistical learning approach has been developed which detects anomalies and projects current trajectories of vessels into the future using AIS data. Different Machine Learning techniques have been evaluated for ETA predictions in [4]. Similarly, ETA predictions based on historical AIS statistics have been evaluated in [5]. Here, the next port destination has been predicted based on the last consecutive port calls, an approach we have also followed for our baseline model presented in Section IV and evaluated in Section V.
Hidden Markov models, which are extended models based on Markov processes, have also been found useful for the detection of anomalous behavior [5] or the prediction of motion patterns [6] of vessels. As these projects do not have a direct influence on our work, we would like to refer the interested reader for further state of the art analysis to the aforementioned literature.
In the following section we describe key aspects of modelling port-to-port-route predictio via Markov models. The derived method is evaluated and discussed in the subsequent Section V.

4. Modelling Port-to-Port-Routes

The most fundamental aspect for meaningful performance in machine learning is an adequate model choice. This does not only include the actual prediction precision, but also the usability of the model in an industrial context. This means that the model needs to balance the trade-off of being

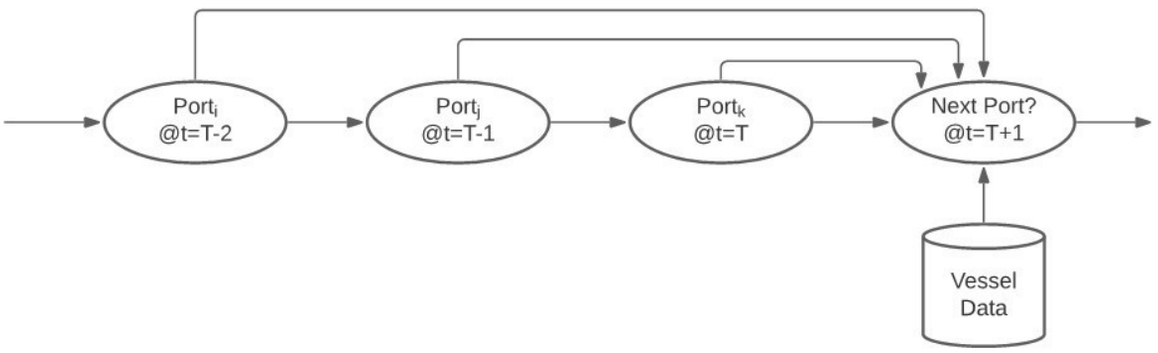

Figure 1: Dependencies of an order 3 Markov process for next port prediction.

- fairly accurate,
- massively scalable (real-time and parallel prediction of a multitude of vessels) and
- efficiently reinforce able.

The last point is especially important to ensure a long-lasting and accurate service. While the next port destination is somewhat determined, unpredictable events (like technical issues) may occur which can ultimately lead to a rerouting of the respective ships. In order to simulate this kind of possibilities it is reasonable to use statistical models, in order to cope with these kinds of uncertainties. We quickly found that discrete Markov processes fulfill all these requirements. In the upcoming subsections we give a short motivation for the usage of Markov processes followed by a short theoretical description of these models and a discussion on how and why we modelled and implemented the models that were evaluated in section V.

A. Discrete Markov Processes

Discrete Markov processes are representatives of Bayesian networks. They

- model temporary relations between states (port to port routes),
- are comparatively easy and fast to calculate as they mainly are built on basic linear algebra operations and
- are efficiently reinforce able.

Once learned, these models can not only give an estimation of the most probable port destination but can also give an estimation of the following ports and the probability of their occurrence (in the respective model).

Furthermore, these models can be extended to so called hidden Markov models which define a separate "observable" stochastic process (e.g. AIS data) from which one can infer the underlying system state (e.g. port areas). This might be of interest in future work as one could directly estimate the last port from AIS data without any event preprocessing (port calls) or even use the AIS data after departure to decide which of the most probable ports (according to the here presented underlying model) matches the real destination.

At this point, we would like to give a short introduction to the theoretical background of Markov processes in order to provide a better understanding of the implemented model. For further information the interested reader is referred to the relevant literature (for example [7])

A Markov process is a stochastic process

$$X=(X_t)_{t\in T}$$

where X_t denotes a random variable at timestamp with values of the feature space.

$$S=\{S_1,S_2,\ldots,S_n\}$$

The discrete space of time T does not hold the actual time of the respective port calls but does only determine the chronological order of the events. The main assumption of a Markov process is that the next system state X_t does only depend on the last k states:

$$P(X_t \mid X_{t-1},X_{t-2},\ldots,X_0) = P(X_t \mid X_{t-1},X_{t-2},\ldots,X_{t-k})$$

This property is often called memoryless or Markov property. In this case X is called a discrete k-th order Markov process. The transition probabilities of a first order process can be stored in a matrix $A \in \mathbb{R}^{(n \times n)}$ with entries:

$$a_{ij}=P(X_t=S_i \mid X_{t-1}=S_i)$$

With these conventions one can easily predict the next system state or rather the probabilities of all potential next states by multiplying the transition matrix by the probability distribution vector $\pi_t \in \mathbb{R}^n$:

$$\pi_{t+1}=A\,\pi_t$$

Here, the i-th element of π_t holds the probability of the system being in state S_i at timestamp $t \in T$:

$$(\pi_t)_i = P(X_t=S_i)$$

This allows fast and parallelly computable predictions as there exist many linear algebra libraries and software environments that handle these kinds of operations efficiently. This is especially important as any higher order Markov process can be transformed in a first order process by combining the sequences of the last states via cartesian products ($X_{t-1}=(X_{t-1},X_{t-2},\ldots,X_{t-k})$) [7]

B. Port-to-Port Markov Processes

After this basic introduction to Markov processes, we present and discuss our solutions for the prediction of port to port routes, in the case of untrustworthy AIS destination flags.

As we set the framework of the statistical model, the main aspect that drives the prediction accuracy is the adequate choice of the feature space. We tested three different scenarios which all share the following structure:

$$S=\wp^k \times \bigotimes_i^m \Lambda_i$$

Here S denotes the feature space, $\wp$ the discrete space of ports (LOCODEs), k the number of last port calls that should be included and the $\bigotimes_i^m \Lambda_i$ cartesian product of m distinct vessel characteristics (for instance vessel type) that are used for further clustering. As the vessel data is invariant over time it might be cleaner to describe the approach as a set of independent Markov processes for varying vessel characteristics with feature space $S=\wp^k$.
The exponent k represents in this case the order of the model that is built for the specific vessel data combination. If $\Lambda_i=\{\lambda^i_1,\ldots,\lambda^i_{|\Lambda_i|}\}$ describes the space of the i-th vessel characteristic, the overall model (neglecting the initial/ current distribution) can be described by the following set of transition matrices:

$$A_{\lambda^1_{i_1},\lambda^2_{i_2},\ldots,\lambda^m_{i_m}} = \left(P\left(X_t = S_j \middle| X_{t-1} = S_i, \lambda^1_{i_1}, \lambda^2_{i_2}, \ldots, \lambda^m_{i_m}\right)\right)_{ij}$$

At this point one needs to find a good set of parameter values for the amount of last ports considered (k) and static vessel data for clustering purposes. It is necessary to understand, that the number of transition matrices is equal to the size of the space of the static vessel data $\bigotimes_i^m \Lambda_i$ which means that if one distinguishes 10 vessel types, one needs to calculate 10 transition matrices. If one adds 5 size classes per vessel type, the number of matrices needed increases to 50. This is why it is generally not advisable to use a high number of static vessel data, although the sparsity of the matrices (that is the number of 0 probabilities [which do not need to be physically stored]) usually increases with finer categorization.
Based on this theoretical groundwork we tested three different approaches. While theoretically possible, all of these neglect the AIS destination entry as it is assumed to be invalid. The beauty of this is that there is no need to find an ever-increasing complex set of rules that map prominent mistakes like incorrect grammar or mistakenly stated subports to the respective port destination. Instead, the three scenarios can be described as follows:

1. Using the last k port calls only
2. Using the last k port calls and vessel type information
3. Using the last k port calls and the MMSI number for vessel

The evaluation of port calls without further categorization is meant to establish a baseline on how good port-to-port routes describe real vessel journeys in general. For the second scenario we used the level 2 vessel type categorization of the widely known IHS Fairplay Database which distinguishes 10 types (+1 for unknown types). Using vessel types should be suited for separating service vessels like tugs that are mainly operating at the same port and ocean-going vessels like container ships or oil tankers which follow more complex routes. In contrast to this general approach we tried to deliberately overfit the model by using the MMSI number as a categorization parameter. This means that every ship gets its own Markov process and therefore its own transition matrix. The latter are in this scenario very sparse as only the port-to-port-sequences that exist in the historical track of the respective vessels hold non-zero probabilities.

In general, it might be more advisable to use the IMO number of vessels as this number serves as an unique identifier. Anyhow, due to better coverage we choose to use MMSI numbers in this experimental setup. Due to the structure of the problem, the sparsity of the matrices and the comparatively high ratio of data selection to floating point operations the whole scenario can be conveniently implemented and processed in relational database systems.

C. In-Database Machine-Learning

At this point we would like to describe how and why we suggest implementing the described model in a (distributed) relational database system. As discussed in [8] and [9] these systems are ideally suited to provide efficient long term implementation of Markov models. The standardized and widely supported query language SQL ensures implementation independence of the concrete system used and its longevity. Furthermore, in-database solutions enable the processing of (preprocessed) data as close to the original data as possible, which in general ensures data security (provided by the database management system) and low network traffic. For a more detailed discussion on advantages of pure database solutions the interested reader is referred to [8] and [9]. Rather than storing a multitude of sparse matrices, we grouped the whole model in one big sparse tensor with the following relational schema:

```
A(
        [
        mmsi            BIGINT,|
        vesseltype      VARCHAR,
        ]
        lastport        CHAR(5),

        k_th_lastport   CHAR(5),
        nextport        CHAR(5),
        p               DOUBLE PRECISION
)
```

where p denotes the probability of "nextport" is the subsequent port after the port sequence specified by the attributes "lastport" to "k_th_lastport". Due to its compactness and its role as an identifier we only used the LOCODES for port identification. As the prediction process consists of a simple sparse matrix multiplication for a possibly selected mmsi number or a vessel type which is internally processed in the database system via a grouped aggregation of a joined table it is necessary to provide the model with a reasonable index structure. Therefore we used a nested b-tree index structure on ([mmsi, vesseltype], lastport, …, k_th_lastport). This allows the prediction of the nextport (or a sequence of next ports) in milliseconds, allowing for a high number of simultaneous real time prediction queries. In production scenarios it is feasible to use distributed database systems to ensure low latency and fault tolerance in the context of big data ("velocity").

5. Model Evaluation

In the following subsections we present and discuss the actual experimental evaluation and the implications we have derived from the results. For this we start with a brief description of the training process and the underlying training data.

A. Model Training

We used company internal port call events that were

derived from the global AIS network of FleetMon. These events were triggered (once) when a vessel entered a port zone and came to hold for a given period of time. Possible port call duplicates due to GPS jittering or brief departures from the port zone were accounted for in post processing. The port call events include amongst other AIS-data at arrival and departure (for instance the destination entry or the MMSI number) as well as processed zone data like the current port (LOCODE) and the last k ports.

For the training process we used all port call events from the years 2018 and 2019, which overall make a total of over 50 million port calls from over 600 thousand vessels and more than 4000 ports.

The calculation of the transition matrices (or the tensor) is done using the classical maximum likelihood approach (see for instance [10]):

$$a_{ij} = \frac{\#\, S_i \rightarrow S_j}{\#\, S_i}$$

Basically, the procedure counts all the transitions from one state (port call sequence + MMSI/Vessel Type) to another and the times the base state occurred in the training set. This simple structure is not only convenient to implement in SQL, but also allows a fairly easy reinforcement process. For this one only needs to store the nominator and denominator separately, so that the counts can be regularly or even continuously updated. This is especially important to account for new vessels, ports or even newly established routes.

B. Evaluation data

We evaluated the trained models against all of the aforementioned port calls that used destination entries that could at some point in time not be mapped to their real port destination. Therefore, the evaluation data with a total of over just above 20 million port calls is a real superset of the actual set of not matchable destinations. The most prominent problem class here are vessels that specify in some form Yangshan Port (e.g. “YANGSHAN”, “YANG SHAN”, “YANGSAN”, …) as their main destination, which is technically incorrect as Yangshan Port is a sub port of the Port of Shanghai. This alone accounts for well above two million port calls with non-matchable next port destinations. The other main problem classes consist of grammatical errors and deliberately unidentifiable entries like “00000000” or similar.

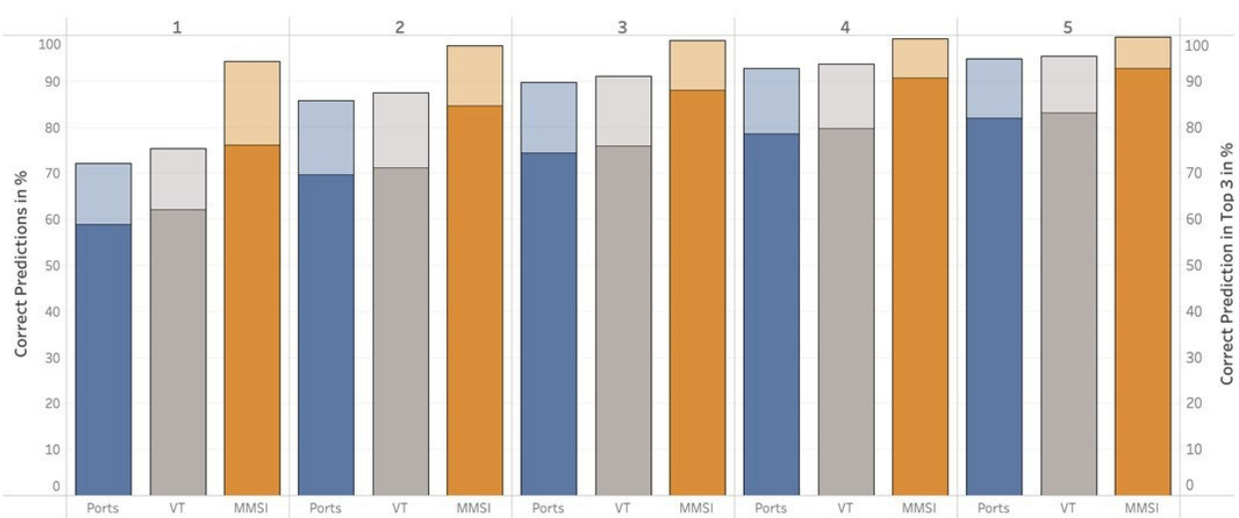

Figure 2: Results of classification differentiated between correct prediction (solid bar) and correct result being in top 3 of predicted next ports (transparent bar). The Order (top caption) represents the number of last ports that were considered for the next port prediction. The bar labels describe whether no additional static vessel data ("Ports"), the vessel type ("VT") or the MMSI number (“MMSI”) has been used as an additional classification source.

C. Evaluation Results and Discussion

The results of the three different approaches with varying number of most recent port calls are depicted in Figure 2. The solid bars represent the correct prediction by using the most probable port, while the transparent bar indicates whether the correct port destination was included in the top 3 most probable destinations. The latter might be of interest for post processing purposes when the most probable port can be dismissed as the real destination due to the observed vessel route.

In general, it can be seen that all models increase their accuracy when using more information of the last port-to-port route. The MMSI model significantly outperforms the other two models and reaches a fairly high accuracy of 93% and 99% for the top 3 case when including the most recent 5 port calls. While the use of vessel types is beneficial for the accuracy in comparison to no additional data, the increasement is close to negligible.

This might be caused by a suboptimal choice of vessel type categorization, as the level 2 categorization of the IHS statcode model uses only 10 different vessel types which leads to for instance to a category that combines passenger ships and dry cargo vessels. Two types of vessels which surely show very different behavior in their journeys (ports and frequency of port calls).

Anyhow, with correct predictions from 60% to 80% and top 3 predictions between 70% and 95% both of the non MMSI models still show potential for further use, especially after some recalibration.

This might come to effect as the vessel specific approach (MMSI) might suffer from long settling periods. That means that new port-to-port-routes of a specific vessel might need to be seen for quite some time until the model can accurately predict the route in production. While this is not that big of a problem for newly built vessels as the overall count of port calls is low, it is especially problematic for older vessels that have seen several route cycles before. Especially for this circumstance one might need to introduce some sort of possibility to forget old routes into the model. Alternatively, one can use a hybrid model where unknown routes use the vessel type approach instead of the vessel specific one. Overall the results as well as the practical feasibility of the models are very promising and will be further developed for practical use.

6. Conclusion and Future Work

This paper presented and discussed AI-based approaches for the prediction of port-to-port routes via markov processes. It has been shown that these kind of models are ideally suited for scalable services and accurate predictions for vessels destinations in the case of unidentifiable AIS destination entries.

We are currently planing on optimizing the presented solution in three main steps. Firstly we will train the model on more data. Secondly we will analyze how we can implement a vessel specific way to forget unused old port-to-port-routes that interefere with the current prediction model. Thirdly we will try to built a better vessel type categorization for a more accurate general model. This might be done by using more vessel subtypes or by adding additional information like classes of gross tonnages or vessel size.

References

[1] United Nations, "Review of Maritime Transport 2019," in United Nations Conference on Trade and Development, Geneva, 2019.

[2] FleetMon, “FleetMon - Tracking the Seven Seas,” 2020. [Online]. Available: https://www.fleetmon.com/.

[3] 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. 1, pp. 2218-2245, 2013.

[4] I. Parolas, ETA prediction for containerships at the Port of Rotterdam using Machine Learning Techniques, Delft: Delft University of Technology, 2016.

[5] K. F. Toloue and M. V. Jahan, "Anomalous Behavior Detection of Marine VesselsBased on Hidden Markov Model," in 6th Iranian Joint Congress on Fuzzy and Intelligent Systems (CFIS), Kerman, 2018.

[6] J. d. Toit and J. v. Vurren, "Semi-automated maritime vessel activity detection using hidden

Markov models," in Proceedings of the 2014 ORSSA Annual Conference, Parys, 2014.

[7] S. Russel and P. Norvig, Artificial Intelligence: A Modern Approach, Pearson, 2020.

[8] D. Marten and A. Heuer, "Machine Learning on Large Databases: Transforming Hidden Markov Models to SQL Statements," Open Journal of Databases, pp. 22-42, 2017.

[9] D. Marten, H. Meyer, D. Dietrich and A. Heuer, "Sparse and Dense Linear Algebra for Machine Learning on Parallel-RDBMS using SQL," Open Journal of Big Data, pp. 1-34, 2019.

[10] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," Proceedings of the IEEE, vol. 77, no. 2, pp. 257- 286, 1989.