Universiteit Leiden
The Netherlands

Opleiding Informatica

A Network-based Approach for Ship-type Modelling

Bram Honig

Supervisors:
Frank Takes & António Pereira Barat & Jasper van Vliet & Cor J. Veenman

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

03/11/2020

Abstract

The maritime transport sector plays a pivotal role in globalisation and economic growth. Within the European Union (EU), member-states collaborate to promote the compliance with adequate ship conduct by inspectorate entities. Depending on ship type, more or less stringent regulations apply. However, due to the number of vessels operating within European waters, these manual inspections cannot be performed on every ship.

Our goal is to help bridge the gap between the inspectorate domain and currently available data science techniques. Individual ship compliance can be assessed in terms of the expected behaviour of its type. This allows for the automatic detection of non-standard behaviour, which could alert inspectors to an increased risk. Such alerts can help the inspectors prioritize inspections of ships with an increased perceived risk to maximize the pragmatism of the limited amount of manual inspection.

Concretely, we aim to capture the underlying structure of maritime transit in a network. This, in turn, generates new insights into ship behaviour and applications within the inspectorate domain. We do so by modelling ship behaviour in a network-based approach.

For this purpose we: (1) generate a network representing ship trajectories within the EU; (2) construct a set of network-based ship features; and (3) model each ship type in a supervised-learning manner based on the constructed features. By achieving adequate AUC performance for ship type models, we validate our approach. Through model analysis, insights such as port relevance towards ship type are attained.

Contents

1 Background

Transit of goods across international waters is of paramount importance. Within the European Union, the maritime shipping sector alone accounts for roughly 75% of all international trades conducted [1]. These transactions are heavily regulated to minimise the health and environmental risks associated with ship transportation. The large number of companies within the shipping sector makes for an extremely competitive market. Furthermore, compliance with ship regulations usually involves additional costs to these companies. Therefore, there is economic incentive for companies to disregard the regulation guidelines set.

The Netherlands is the largest maritime transport hub in Europe [2], serving as both entry and an exit point to intercontinental exchange of goods. This is exacerbated by Rotterdam's status as the second largest bunker fuel port in the world [3]. Here, in the Netherlands, the body responsible for enforcing such regulations is the "Inspectie Leefomgeving en Transport" (ILT) [4]. One of the purposes of the inspectors of the ILT is to perform regular manual inspections of ships within ports. The current inspection candidate selection method by the inspectors is based solely on their domain knowledge, following a risk assessment system described in the Paris Memorandum of Understanding on Port State Control (Paris MoU) [5].

Manually assessing the compliance of all ships pertaining to every company at each port is infeasible. The methods used by inspectors are not suited to adequately address the current volume of ships. Additionally, there is a disparity between the availability of data-driven approaches and their application within the inspectorate domain. An example of this disparity is the limited amount of data used by the Paris MoU.

We propose using Network Analysis (NA) to extract graph measures which describe characteristics of ports visited. Similar NA techniques have previously been used for identifying celebrities in social media platforms [6][7]. A common approach for such a task is looking for nodes (persons) within the network, which had a high number of edges (interactions) with other nodes. We propose using similar measures for encoding ports within ship trajectories, such as centrality. There is little consensus over when to use the term "network" vs the term "graph" and since both are almost interchangeable, we will use the network terminology in this thesis since this has been a common approach when talking about data related to transportation.

Centrality allows us to estimate the importance of a node within the network. Using such centrality measures for encoding the ports visited by a ship provides context to the ports visited by this ship. An example of this is the difference between visiting an international harbour or a small fishing village. More precisely, the ports mentioned above will have greatly differing graph measures.

The data available in this research contains information regarding which ports are visited per ship, from which port they set out, and the duration of this trajectory. This allows us to create a network of ports (nodes) and trajectories (edges) between them. The network can then be used to extract graph measures for the nodes. In addition to such node based graph measures, trajectory duration is also encoded based on a comparison to trajectories in the network. As a consequence this helps to distinguish between ship behaviour based on operating speed.

The data, collected by the ILT, spans more than half a decade, so a subset of trajectories within the first year are chosen to be used to create a network. We choose to call this "normal" behaviour since it leaves out extreme values for trajectory duration and covers an entire year to account for seasonal changes. In this thesis we will refer to the network describing this "normal" behaviour as

the Maritime Movement Network (MMN).

Supervised machine learning is used in a OneVsRest approach to train and test a model to distinguish between ship types (e.g. container, bulk carrier, etc) based on a feature set partially extracted from the MMN containing node centrality, trajectory duration, and other features. A disadvantage of creating the MMN is that data used for its creation can not be be used for training nor testing purposes without leaking data to the classifier. We verify the performance of the model is adequate using stratified cross-validation. If the models performance is deemed adequate, any misclassification would mean its behaviour differs significantly from the observed general behaviour of that ship type. Any such misclassifications could automatically alert inspectors to an increased risk.

In this thesis we will present the following contributions:

1. Introduce a graph representation for ship trajectories
2. Define features for ship trajectories based on graph measures
3. Apply a supervised machine learning approach to model ship types based on the constructed features

The structure of this thesis is as follows: Section 2 will focus on the previous uses of NA within the domain, and other research concerning ship type classification. In Section 3 the graph measures are defined. Section 4 describes the data used. Section 5 describes the experimental setup in detail. Section 6 refers to the results of the experiments. Lastly Section 8 concludes our work and gives direction for future work.

2 Related Work

In this section will discuss the use of NA and other techniques used to model ship types.

2.1 Network Analysis

NA has been used to characterize a cargo ship network [8]. It was noted that there are three main subdivisions to be made between the types of ships namely: oil tankers, bulk carriers, and container ships. Each of these ship types has different characteristics such as speed or time spend unloading.

The main conclusion that could be drawn was that globally the cargo ship network tended to have large average clustering and that the strength, degree of nodes, as well as the weight of edges had broad distributions. Degree in this work is defined as the number of edges connected to a node and strength is defined as the combined weighted in- out, degree. the clustering coefficient can be interpreted as a measure of the completeness of the graph [8]. The latter characteristics regarding weight, degree and strength are very useful.

Recall that the goal of this thesis is to use Machine learning to model ship types based on this kind of data. More possible splits based on feature values allow for a split that minimizes the combined entropy [9]. Since broad distributions allow for more possible splits based on the feature values, these broad distributions are a very useful characteristic for features used in the machine learning.

It should be noted this study was on data from around the world instead of the fine-grained data in our study. This thesis covers a data set with $\tilde{2}000$ distinct ports in Europe compared to $\tilde{1}000$ distinct ports worldwide in the study [8] . The study has made useful general analyses, characterizing the network as a distribution network. This is suggested by the asymmetrical weight of the directed edges between the ports, compared to a transportation network where ships make round trajectories. Therefor a directed network is needed to fully represent ship behaviour.

2.2 Ship type modeling

Sheng et al. [10] and Kraus et al. [11] have focused on the geographical trajectories of ships to model ship type instead of the network-based approach as proposed by this paper. These trajectories are based on detailed AIS (Automatic Identification system) data which includes GPS position, orientation, speed, etc. One can easily see the potential of this data. However this is also its valency, the sheer size, since every boat will send messages every two to ten seconds, and other messages every six minutes or so. Thus besides the modeling of ship types, the main problem solved by these papers is the compression of AIS data while preserving and extracting distinguishing features.

The first of such papers focused on modeling fishing ships and cargo ships based on AIS ship trajectories [10]. Ship trajectories are subdivided into three movement patterns: anchored, straight sailing and turning. This was done by a proposed basic movement algorithm. From these movement patterns five features describe general movements such as stops and sailing distance. Another six features concerning turning behaviour, were introduced and five features describing sailing speed. Lastly main travel direction discretized by the eight cardinal directions. The logistic regression algorithm used in this paper had a AUC performance of 0.963. Within this classification speed is of huge importance and turning behaviour was important to a lesser extend. Most misclassifications by this model could be contributed to specific ship behaviour such as ocean-going fishing vessels in transit or cargo ships loading or unloading.

Kraus et al. [11] implemented similar features for their classification model. A major difference between both studies is that the turning features could not be implemented due to lack of AIS coverage in the German Bight.

To compensate a new set of features was introduced based on geographical location. The implementation of these features was done by dividing their research area into grids which were optimised for entropy based on class distribution.

This model was able to differentiate between five different ship types. An important conclusion is that ship dimensions are of major importance when differentiating between tankers and cargo ships. This is a result of both ship types having common routes as well as similar speeds.

The need for another, more global, ship type classifier is shown in the study [11]. The lack of AIS data, due to coverage issues, made it impossible to implement some of the features proposed in the study [10].

3 Definitions

This section is intended to provide a brief overview of the terminology used for graphs and additional terminology of the graph measures.

3.1 Graph

A graph G is defined as $G = (V, E)$, where V is the set of n vertices, E is a set of m edges, between the vertices in V with two attributes: weight and estimation of geographical distance.

The collection of edges within the graph consists of two vertices u, v where there are weight and distance associated with that edge of the form (u, v, w, g). A node in our graph will be a port and an edge will consist of instances of trips between two ports. The associated geographical distance $g(u, v)$ is estimated using the travel times between vertices u, v and weight $w(u, v)$ are the number of ships travelling between u, v.

3.2 Graph measures

Given a graph G, node centrality can be calculated in multiple ways. The node centrality measures used to assess node importance.

The expression for betweenness centrality of a node u is:

$$B(u) = \sum_{s \neq u \neq t} \frac{|\sigma_{st}(u)|}{|\sigma_{st}|} \tag{1}$$

Here σ_{st} is the set of shortest paths from vertex s to vertex t and $\sigma_{st}(u)$ the subset of those shortest paths that pass vertex u. These shortest paths are calculated using Dijkstra's weighted algorithm with an estimation of geographic distance using travel time as the weight [12].

The weighted inward closeness centrality for a node u is

$$C(u) = \frac{n-1}{\sum_{v=1}^{n-1} g(v, u)} \tag{2}$$

Lastly we use the strength which is the weighted combined, in and out- degree of vertices. Which is defined as:

$$weighted\_degree\_in(u) = \sum_{v=1}^{n} w(v, u) \tag{3}$$

$$weighted\_degree\_out(u) = \sum_{v=1}^{n} w(u, v) \tag{4}$$

The resulting strength of vertex u is thus:

$$S(u) = weighted\_degree\_in(u) + weighted\_degree\_out(u) \tag{5}$$

4 Data sets

Different data sets have been collected by the ILT in the past six years. Two different, related, data sets are used in this paper. First the portcall data set, which contains all trajectories made by ships within Europe. A subset of these will be used for creating the MMN. The trajectories not included in the MMN wil be used for training and testing of the machine learning model discussed in Section 5.4. Secondly a port data set which contains metadata about ports which will be used to denote extra information about ports visited by a ship.

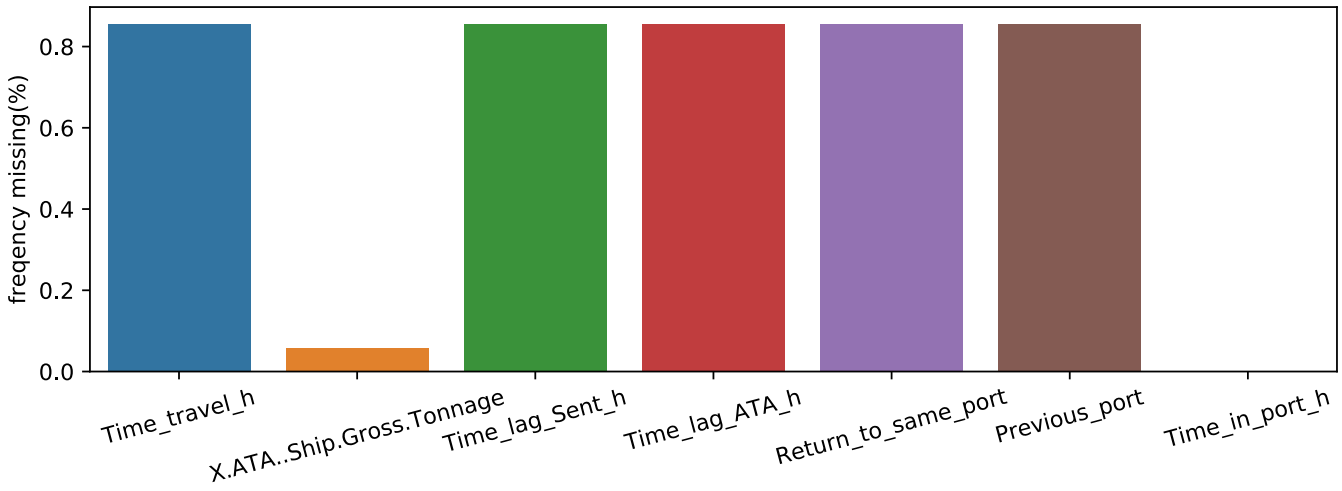

Figure 1: Relative frequency (%) of missingness per relevant column in trajectories

4.1 Trajectory

A row in the data set consists of 48 columns, which can be observed in appendix A Table 5. For this research, we define a ship trajectory by its unique ship identifier, the travel time of the voyage, and previous and destination port.

4.1.1 Missing data

Figure 1 is meant as an indication of missing values, for variables that could be used as alternatives for edge and node attributes.

The column annotating the length of a ship, misses around 40% of values. This is quite troublesome given the conclusion of related work that these dimensions are of great importance for distinguishing tankers and cargo ships.

Another interesting column is the risk profile missing in around 8% of trajectories. There are a total of 32 columns missing values, however all other columns miss less than 1% of their values One of these is especially troublesome for the network approach:

- An unknown port in the previous port: This happens around in 0.85% of the total trajectories.

This limitation is a consequence of the method used to synthesize the data set. Since all ships will have an unknown port of origin for their first trajectory in the data set, these trajectories will be left out.

4.1.2 Non representative data

Although some columns do not have any missing values, they contain extreme values. In this thesis we will use the term “non representative data” for these values, which we define as data that does not accurately describe the behaviour of the ship, within the context of the network. One of these is a non representative travel time and can be divided into three categories:

- Administrative errors

- Unknown ship behaviour such as anchorage outside a port
- Ships that leave the scope of the data set and return at a later point

Lastly we have instances of ships that change ship type within the scope of the data set.

4.2 Ports

A port in the ports data set consists of 84 columns, which can be observed in appendix A Table 6. These columns indicate the presence of facilities such as pubs, cranes and some more information about the accessibility of the port. Furthermore, it also contains the coordinates of the port itself.

4.2.1 Missing data

Although the ports do not have missing values, this does not mean it all ports are included in the data set. This is clear when comparing the number of unique ports in the portcall data set and the number of ports in the port data set, respectfully 2000 ports vs a subset of 500 ports. This means we do not have additional information for 75% of all ports. Consequently the detailed features will only be used when visiting a port within the port data set.

4.2.2 Enrichment

Most data in the ports data set is industry-specific and additional data is not easily accessible, such specific facilities or characteristics of ports. Therefor the enrichment of this data set will be limited to adding the geographical coordinates to ports not included in the ports data set through API calls to BING's geo-locate services.

5 Methodology

In this section we will describe (A) the creation of the MMN (B) the ship features (C) the machine learning model. Within Subsection 5.1 we will discuss the data used for the creation of the MMN and the calculation of its attributes used for feature extraction. The Subsection 5.3 will discuss the features used, their normalisation and present an example graph with corresponding features. In the last Subsection 5.4 the ML model used is discussed. This will include the performance measure used and creation of train and test sets. Figure 2 provides an oversight data processing, network creation, graph measure calculation and feature extraction.

5.1 Maritime Movement Network

For the MMN creation a subset of the first year of trajectory data is chosen. Trajectories in this subset are filtered out if they contain missing values in the previous port, in the travel time denoted or if the ship changes registration within the data set. The later filtering regarding change in registration introduces a bias since this happens mostly in a specific ship type namely the pleasure yachts, however these are of minimal interest in this thesis since we focus mostly on commercial shipping.

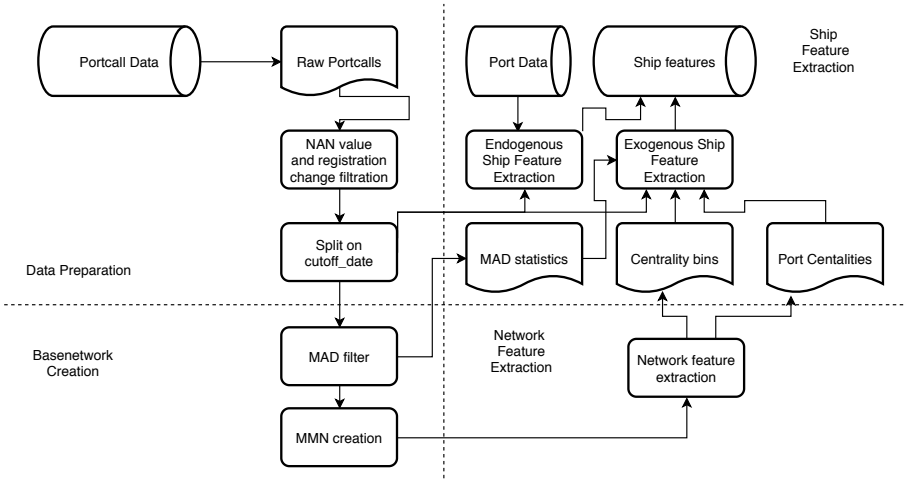

Figure 2: Diagram of ship feature creation

The frequency of missingness can be regarded in Figure 1. However, this still leaves room for variables such as travel time duration, that differ greatly from the median travel times. Therefor to ensure accurate, representative, data, we use the subset of data to derive the MMN from the data in Subsection 4.1.

This subset is then used to create a MMN between ports where trajectories are only included when their travel time falls within the double Median Absolute Deviation (MAD) cutoff, see Subsection 5.1.1. As a result 82.3% of the trajectories used for MMN creation are included in the MMN.

These trajectories are used to calculate two edge attributes, the weight and distance. The weight is the number of trajectories between two ports and the distance is the average of the travel times that fall within the MAD cutoff. The resulting MMN includes all ship type trajectories within the subset, with a travel time that falls within the MAD.

The use of MAD introduces a certain bias for estimated geographical distance due to class imbalance between ship types. This bias results in a tendency to include more trajectories from higher frequency ship types compared to lower frequency ship types.

An example of the resulting MMN can be seen in Figure 3.

5.1.1 MAD

The double MAD is able to handle the skewed distributions within trajectory duration. The method is similar to other quantile measures, which are used to clean noisy data. However by using the median instead of average it is robust against extreme values. A cutoff value of 1.5 was chosen since it was observed, the results did not change significantly for different values. In the case where the MAD is zero all trajectories are included.

5.2 Graph measures

Using the MMN, several previously described graph measures are calculated. The MMN, will have a MAD range per trajectory, which will be used to determine the relative speed of ships in the

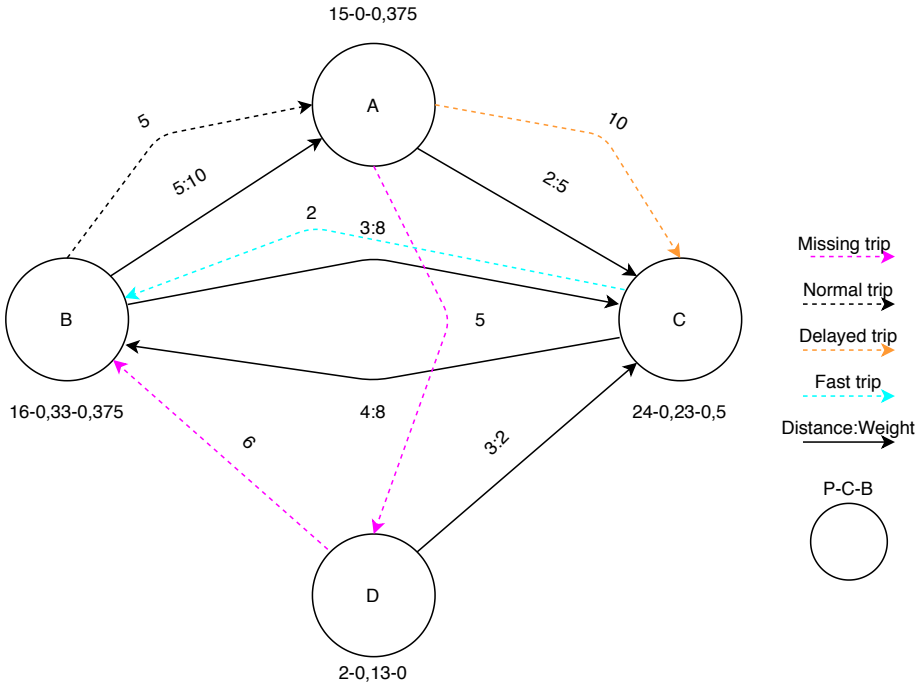

Figure 3: A example of a Maritime Movement Network, with node centrality and an example of a ship trajectory in dotted lines

training and testing data sets. Secondly, the node importance is calculated in three different ways namely: betweenness, closeness and strength.

These node features are then discretized by means of equal frequency binning. The binning is used to reduce the feature dimensions and allow for aggregation of port visiting behaviour. The number of bins is chosen to be ten due to research showing that equal frequency binning with an optimized amount of bins does not heavily influence the performance of supervised learning approaches compared to using the standard of ten bins [13].

5.3 Ship features

After the MMN is created and all necessary ranges and bins are calculated, the following feature types can be extracted for each ship in the training and test sets: exogenous features, where ship behaviour is compared to the MMN and endogenous features, which can be extracted from the data without comparison to the MMN.

5.3.1 Endogenous

- Number of visited ports with Oil
- Number of visited ports with Diesel
- Number of ports visited
- Number of unique ports visited

It should be noted that the presence of oil and diesel is only known for a subset of around 25% of all ports.

Other endogenous features are normalized by using the fraction of the value divided by the number of total trajectories per ship.

- Number of travel_times between 125-250 hours
- Number of travel_times between 250-500 hours
- Number of travel_times bigger than 500 hours

These features are intended to indicate whether a ship leaves the scope of the data set and relative speed.

5.3.2 Exogenous

The following exogenous network features are created based on the MAD range calculated in the network creation step described in Subsection 5.1. Lastly, a feature "missing trajectory" is present, which encodes trajectories that are not present in the MMN. Afterwards, these counts will be normalized by using the fraction divided by the number of total trajectories per ship:

- Number of normal trajectories
- Number of fast trajectories
- Number of slow trajectories
- Number of missing trajectories

Other exogenous features are extracted from the previously calculated graph measures and binned using the bins calculated in the MMN creation step described in Subsection 5.1. These are normalized by using the fraction of the value divided by the total number of ports visited per ship.

- Strength of ports visited
- Inbetweeness Centrality of ports visited
- Closeness Centrality of ports visited

Lastly we have a more general feature describing the total strength of all ports visited per ship:

- Total strength of ports visited

Figure 3 contains multiple trajectories of a ship starting from node A ending in node B, the exogenous features for this ship trajectory are denoted in Table 2

5.4 Machine learning

We will train a model to classify ships based on ship types using features extracted from the MMN. The performances will be compared using ten folds of stratified cross-validation. We will consider a classifier with decent performance if its performance is larger than or equal to 0.8 AUC, which we will discuss further in Section 5.4.3.

Feature description	# features	Origin
Port centrality	10	Exogenous
Port betweenness	10	Exogenous
Port strength	10	Exogenous
Unique port count	1	Endogenous
Total port count	1	Endogenous
Total port strength	1	Exogenous
Presence oil and diesel	2	Endogenous
trajectory duration	3	Endogenous
Relative trajectory duration	3	Exogenous
Missing trajectories	1	Exogenous

Table 1: Overview of features

feature	value	feature	value
Pop1	0.16	Betw1	0.16
Pop2	0.33	Betw2	0.66
Pop3	0.33	Betw3	0.16
Pop4	0.16	Fast trajectory	0.2
Cent1	0.33	Normal trajectory	0.2
Cent2	0.16	Delayed trajectory	0.2
Cent3	0.16	Missing trajectory	0.4
Cent4	0.33		
Total strength	80		

Table 2: Exogenous feature set of the example ship trajectories with a reduced # bins

5.4.1 Training and test sets

All data excluding the subset used for MMN creation and data that is filtered out, is used as the train and test set. This subset is then sorted by individual ships based on their unique identifier. The sorted data is then split into training and test sets using stratified-K-fold using a K of ten. Stratification is used to ensure the relative frequency of the ship type we try to model is the same in each split [14]. This is done because of class imbalance between ship types.

5.4.2 Classifier

For classification a non-linear gradient boosted tree framework (XGB) [15] was chosen. XGB was configured to use 100 decision trees with a max depth of three. This classifier was chosen due to its robustness to unrelated features, and its scalability [16]. This is verified by its standing as a state-of-art supervised machine learning approach [17].

5.4.3 Performance

Classification performance is measured as the Area Under the receiver operating characteristic Curve (AUC) [18]. The performance measure ranges between 0.5 and 1, where the lowest value 0.5 relates to an uninformed classifier and 1 is a perfect classifier.

6 Results

In this section we present the results of the experimental setups using the features extracted from the MMN. Table 3 below shows the performance of the different models measured with AUC.

6.1 AUC

The AUC performance yielded by our experimental setup can be seen in Table 3. Mean and standard deviation values for each ship type across folds can be regarded. Ship type names were aliased With letters due to their length, referenced in appendix A Table 8.

Ship Type	AUC	N	Ship Type	AUC	N
A	0.93 ± 0.01	5452	N	0.88 ± 0.02	2081
B	0.97 ± 0.02	236	O	0.98 ± 0.00	2291
C	0.80 ± 0.06	111	P	0.97 ± 0.01	864
D	0.86 ± 0.02	833	Q	0.89 ± 0.01	4117
E	0.84 ± 0.04	84	R	0.87 ± 0.06	70
F	0.98 ± 0.01	124	S	0.98 ± 0.02	62
G	0.93 ± 0.04	72	T	0.75 ± 0.14	57
H	0.93 ± 0.01	844	U	0.84 ± 0.01	2274
I	0.86 ± 0.02	838	V	0.97 ± 0.01	402
J	0.95 ± 0.02	452	W	0.92 ± 0.03	259
K	0.94 ± 0.01	936	X	0.97 ± 0.01	529
L	0.83 ± 0.08	170	Y	0.87 ± 0.01	617
M	0.93 ± 0.03	63			

Table 3: AUC per ship type vs Rest, N being the number of instances in training and testing

6.2 Feature importance

The individual feature are shown per ship type classifier in Figure 4 as a heat map, where a cell represents the feature importance per model. A lighter color indicates a larger reliance on that feature in the model. Appendix A, table 4 contains the ship types corresponding to the indices used in Figure 4

A notable trend within these features is the focus on endogenous node features. This indicates the importance of endogenous features for ship types to differentiate between ship types. Observing the general tendency to focus on one out of the four endogenous node features. A probable explanation for this trend is the ability to differentiate between larger transit hubs and thus larger ships and associated ship types, based on these endogenous node features. This explanation is based on the bias for bigger ports concerning the oil and diesel presence, due to the lack of port in the ports data set which can be observed in Subsection 4.2. A example of this is the importance of the oil count for commercial and non-commercial pleasure yachts (B,I), which can be interpreted as a feature excluding membership.

Regarding the exogenous features, each ship type has an apparent preference for a specific centrality bin. On closer inspections these contain ports related to that ship type. Another interesting feature importance can be observed in the ship type Warship and naval auxiliary (M). This model has a relatively high feature importance for trajectories not in the MMN. We interpret this as distinctive behaviour of warships, which do not fall within the characterisation of a distribution network. But share the characteristic of uncommon trajectories.

Lastly we can observe the use of the relative trajectory duration. These features contain the feature delayed which has the highest importance overall for the Container ship type. The importance of this feature, to that class, can be explained to due container ships having the second highest frequency within the data. Therefore an interpretation of this feature is exclusion of membership to container ship type based on speed.

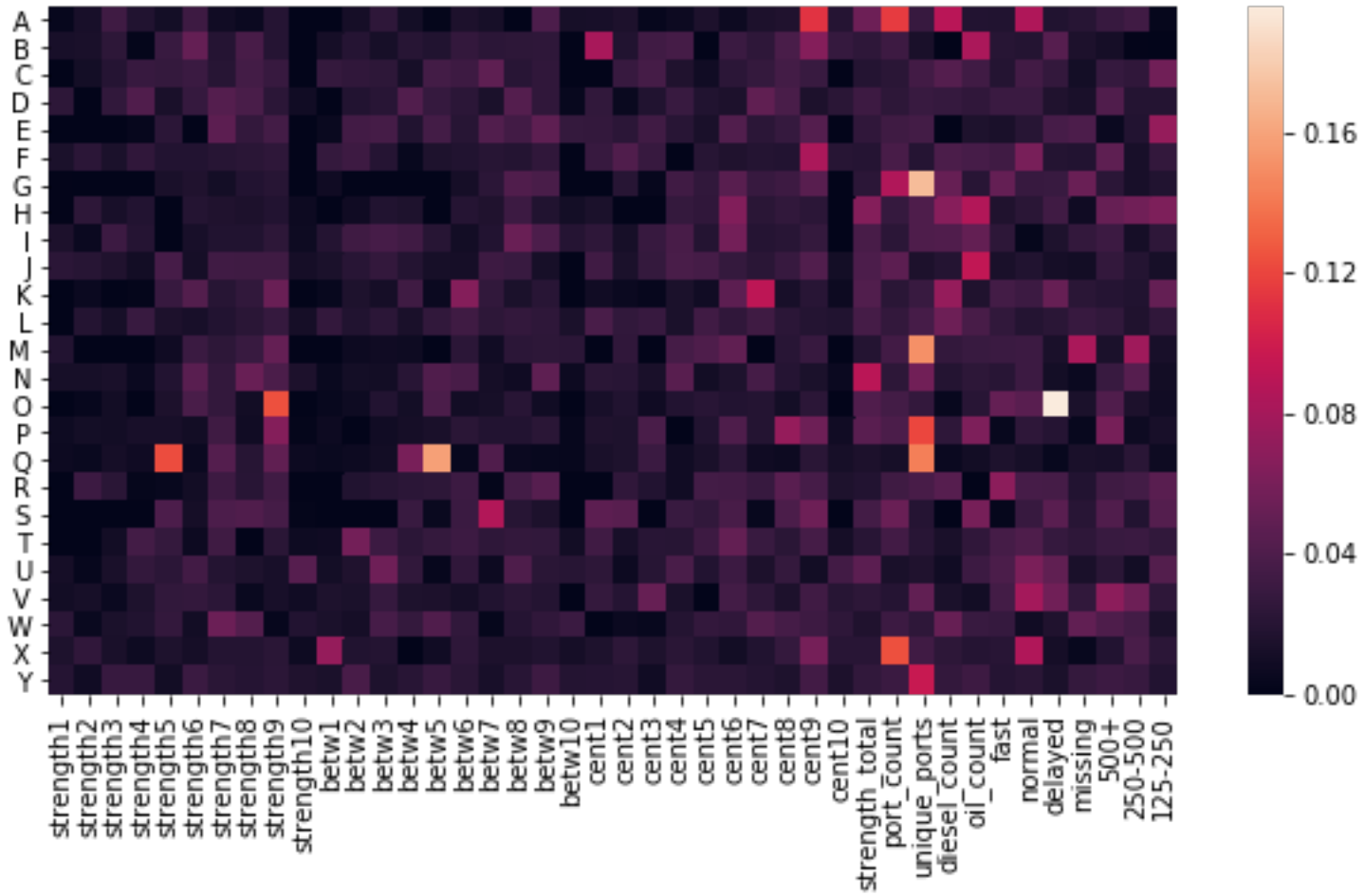

Figure 4: Feature importance per ship type

7 Discussion

When comparing the results of the MMN based model with the previous model described by Sheng et al. [10], a similar performance was achieved for 33% of ship types. However the number of ship

types that are modeled is much larger than in previous model. With a minimum performance of the NLS Tanker classifier of 0.75, the method proposed is widely applicable. This is further supported by remaining performances ranging between 0.80 and 0.98. These results lead us to conclude that the MMN model is comparable with state of the art models and could a supplemental role in the prioritizing ship inspections by giving priority to ships which are misclassified by the model.

A problem encountered within the scope of this research was figuring out which travel duration in the trajectories contained in the portcall data set were reliable. This was made difficult due to administrative errors and unknown ship behaviour such as anchorage outside of ports. The method used for extracting "reliable" travel duration was applying the robust MAD to the data. The MAD method had some disadvantages such as the need for more than three instances of travel duration to differentiate between them. Such limitations could be countered by using average speed instead of travel duration, although this would require the distances between ports to be known.

The applicability of this approach is limited by some characteristics of the ship types themselves. The main reason being, that the network approach mostly focuses on the trajectories, and thus movement of a ship between ports. This leads to more specialised ship types having a lower performance since they do not have common routes or ports. However they might share other common characteristics such as time spend in ports which was not included as a feature.

Other difficulties encountered included the lack of metadata for 75% of ports. We were able to synthesize extra information regarding the location of ports which were missing. But due to the unavailability of other data regarding the ports such as accessibility and the tonnage of cranes present, the decision was made not to include features based on these variables except for the oil and diesel counts.

Lastly, the disadvantages of the MMN creation should be mentioned. First of all, the data used to create the MMN cannot be used for training or testing without leaking information about the target variable. Secondly the choice of what subset to use for MMN creation heavily influences the performance of the classifier. This is due to the ship instances changing for example no longer including certain trajectories. But also due to graph measures changing due to changes in trajectories included in the MMN.

8 Conclusions and Future work

In this thesis we introduced (1) a graph based representation for ship trajectories, (2) a method for extracting features for ship trajectories based on graph measures, and (3) showed the effectiveness of these features for modelling ship types.

The method proposed in this thesis serves as an indication of its applicability to data describing movement, and can hopefully be expanded upon for classifying different target variables. It has been shown to be applicable to differentiate quite well between many different ship types. However, as pointed out in the discussion it is still far from a gold standard.

The proposed model resulted in multiple decent models but was not enough to classify all ship types. However, it does indicate the effectiveness of basic network analysis to classify certain behaviour.

The advantages of this approach are: simplicity, widespread applicability, and extra information concerning relations within the data. The use of these features is not limited to ship type models but could form a base for risk identification based solely on ship trajectory and other historic

behaviour. Thus one of the main advantages of the network-based approach is its flex ability to different classification goals and interpret-ability of relations. These relations could otherwise not be extrapolated by any machine learning approach, while being instinctively identified as relevant by a human inspector.

Within network analysis the idea of a network describing behaviour is not a novel concept neither is the idea of multi-network analysis to encode different interactions. However, it does not seem to have been applied to transport networks for classifying actors within those networks.

Further research could focus on applicability of network based approaches to better account for dynamic ship behaviour. Currently, risk assessment is done in a static way such that only current values are taken into account for the risk profile of a ship. When relations between these static variables can be defined, it could be used to analyse ship risk in a dynamic historical way. Thus, using the same graph measures could improve upon the understanding of such as dynamic behaviour. Such networks could include the changes of Flags of Convenience (FoC) or even ownership of a ship by different shipping companies.

Alternatively, the MMN could include cargo capacity on edges as well as encoding time in port distributions in node attributes. These additions would better capture differences between ship types as ship types usually have differing cargo capacity as well as different time in port distributions. Observing the ability of the model to identify ports indicative of different ship types, the usage of the MMN to predict the flow of goods between is feasible. Such applications would increase insight in the flows of goods within the EU and could help identify wrongly labelled cargo. Furthermore, this could provide valuable insights for future infrastructure development based on the needs of individual ports in relation to the maritime transport infrastructure as a whole.

In combination with new approaches to NA such as pattern mining, it is feasible to improve upon the performance of the current model by identifying indicative movement patterns between ports.

Lastly, interest lies in real-time calculation and classification of ships to detect possible cases of fraud as well as a method to find optimal subsets of data for MMN creation. However, for real time calculation of these features, the network structure will have to be adjusted and more efficient algorithms should be used to minimize the computation of any updates to the network and features.

References

[1] "Maritime transport." https://ec.europa.eu/transport/modes/maritime/maritime-transport_en. Accessed: 2020-07-15.

[2] "Container ports." http://www.worldshipping.org/about-the-industry/global-trade/top-50-world-container-ports. Accessed: 2020-07-15.

[3] "Bunker ports." https://maritimefairtrade.org/top-ten-bunkering-ports. Accessed: 2020-07-15.

[4] "Port-state-control." https://www.ilent.nl/onderwerpen/port-state-control. Accessed: 2020-07-15.

[5] "Paris memorandum on portstate control." https://www.parismou.org. Accessed: 2020-07-15.

[6] E. Otte and R. Rousseau, "Social network analysis: a powerful strategy, also for the information sciences," *Journal of Information Science*, vol. 28, no. 6, pp. 441–453, 2002.

[7] S. Wasserman and K. Faust, "Social network analysis in the social and behavioral sciences," *Social Network Analysis: Methods and applications*, vol. 1994, pp. 1–27, 1994.

[8] M. Barthélemy, "Spatial networks," *Physics Reports*, vol. 499, no. 1-3, pp. 1–101, 2011.

[9] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[10] K. Sheng, Z. Liu, D. Zhou, A. He, and C. Feng, "Research on ship classification based on trajectory features," *The Journal of Navigation*, vol. 71, no. 1, pp. 100–116, 2018.

[11] P. Kraus, C. Mohrdieck, and F. Schwenker, "Ship classification based on trajectory data with machine-learning methods," pp. 1–10, 2018.

[12] E. W. Dijkstra, "A note on two problems in connexion with graphs," *Numerische Mathematik*, vol. 1, no. 1, pp. 269–271, 1959.

[13] M. Boulle, "Optimal bin number for equal frequency discretizations in supervized learning," *Intelligent Data Analysis*, vol. 9, no. 2, pp. 175–188, 2005.

[14] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'95, (San Francisco, CA, USA), p. 1137–1143, Morgan Kaufmann Publishers Inc., 1995.

[15] J. Friedman, "Greedy function approximation: A gradient boosting machine," *The Annals of Statistics*, vol. 29, 11 2000.

[16] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," pp. 785–794, 08 2016.

[17] "Bench ml." https://github.com/szilard/benchm-ml. Accessed: 2020-07-18.

[18] T. Fawcett, "Introduction to roc analysis," *Pattern Recognition Letters*, vol. 27, pp. 861–874, 06 2006.

Appendix A

Ship type		Ship type	
Bulk carrier	A	Chemical tanker	N
Commercial yacht	B	Container	O
Dredger	C	Fishing vessel	P
Gas carrier	D	General cargo/multipurpose	Q
Government ship used for non-commercial purpose	E	Heavy load	R
High speed passenger craft	F	Livestock carrier	S
MODU / FPSO	G	NLS tanker	T
Offshore supply	H	Oil tanker	U
Other special activities	I	Passenger ship	V
Pleasure yacht not engaged in trade	J	Refrigerated cargo	W
Ro-Ro cargo	K	Ro-Ro passenger ship	X
Special purpose ship	L	Tug	Y
Warship and naval auxiliary	M		

Table 4: Ship type annex

Column Name	Column Name	Column Name
Port.Call.ID	External.System.ID	Sent.At
ETD_LT	ATD_LT	Is.Anchorage
Country.Code	IMO.Number	X.ATA..Ship.Name
X.ATA..Ship.Flag.Code	X.ATA..Ship.Flag.Description	ATA_LT
X.ATA..Ship.Flag.Is.PMOU	X.ATA..Ship.Gross.Tonnage	Port.Location.Name
X.ATA..Ship.Keel.Laying.Date	X.ATA..Ship.Length	Port.Name
X.ATA..Ship.Type.Code	X.ATA..Ship.Type.Description	Time_lag_ATA_h
X.ATA..Ship.Type.Is.High.Risk	X.ATA..Ship.Status	Previous_country
X.ATA..Ship.Risk.Profile	X.ATA..Ship.Priority	Postponement.Status
X.ATA..Non.elegible.for.PSC	ETA_LT	Time_travel_h
X.ATA..Non.elegible.for.PSC.Reason.Code	Time_lag_Sent_h	Time_lag_Sent_d
X.ATA..Non.elegible.for.PSC.Reason.Descriptions	Is.Detained	Previous_port
Call.Role.In.Postponement	Time_travel_d	Time_lag_ATA_d
Miss.Justification	Miss.Reason	Miss.Reason.Is.Expired
Time_travel_h_rel	Time_in_port_h	Time_in_port_d
Miss.Status	Return_to_same_port	

Table 5: Portcall column name annex

Column Name	Column Name	Column Name	Column Name	Column Name
Join_Count	TARGET_FID	JOIN_FID	OBJECTID_1	INDEX_NO
REGION_NO	PORT_NAME	COUNTRY	LATITUDE	LONGITUDE
LAT_MIN	LAT_HEMI	LONG_DEG	LONG_MIN	LONG_HEMI
CHART	HARBORSIZE	HARBORTYPE	SHELTER	ENTRY_TIDE
ENTRYSWELL	ENTRY_ICE	ENTRYOTHER	OVERHD_LIM	CHAN_DEPTH
ANCH_DEPTH	CARGODEPTH	OIL_DEPTH	TIDE_RANGE	MAX_VESSEL
HOLDGROUND	TURN_BASIN	PORTOFENTR	US_REP	ETAMESSAGE
PILOT_REQD	PILOTAVAIL	LOC_ASSIST	PILOTADVSD	TUGSALVAGE
TUG_ASSIST	PRATIQUE	SSCC_CERT	QUAR_OTHER	COMM_PHONE
COMM_FAX	COMM_RADIO	COMM_VHF	COMM_AIR	COMM_RAIL
CARGOWHARF	CARGO_ANCH	CARGMDMOOR	CARBCHMOOR	CARICEMOOR
MED_FACIL	GARBAGE	DEGAUSS	DRTYBALLST	CRANEFIXED
CRANEMOBIL	CRANEFLOAT	LIFT_100_	LIFT50_100	LIFT_25_49
LIFT_0_24	LONGSHORE	ELECTRICAL	SERV_STEAM	NAV_EQUIP
ELECREPAIR	PROVISIONS	WATER	FUEL_OIL	DIESEL
ENG_SUPPLY	REPAIRCODE	DRYDOCK	RAILWAY	Shape_Length
Shape_Area	LAT_DEG	PUB	DECKSUPPLY	

Table 6: Port column name annex