

# **Deep Learning & Graph Clustering for Maritime Logistics: Predicting Destination and Expected Time of Arrival for Vessels Across Europe**

**Álvaro Orgaz Expósito**

## **School of Science**

Thesis submitted for examination for the degree of Master of  
Science in Technology.

Helsinki 31.7.2020

## **Supervisor**

Prof. Arno Solin

## **Advisor**

Dr. Jussi Poikonen

Copyright © 2020 Álvaro Orgaz Expósito

---

**Author** Álvaro Orgaz Expósito**Title** Deep Learning & Graph Clustering for Maritime Logistics: Predicting Destination and Expected Time of Arrival for Vessels Across Europe**Degree programme** ICT Innovation (EIT Digital Master School)**Major** Data Science **Code of major** SCI3095**Supervisor** Prof. Arno Solin**Advisor** Dr. Jussi Poikonen**Date** 31.7.2020 **Number of pages** 57+3 **Language** English**Abstract**

In recent years, the need for improving operational processes internationally has drastically increased in the maritime logistics field. The lack of streamlined systems that provide reliable information about real-time maritime traffic for the main agents across countries, such as ports operators and ships authorities, has prompted several research questions. In this work, we propose Deep learning and Machine Learning based methods for (i) clustering ports across Europe using their maritime traffic connectivity, (ii) predicting the next destination of vessels, and (iii) forecasting their expected voyage duration. Several experiments based on public AIS data are developed to analyse and verify these methods, and the results of these experiments indicate that the proposed models achieve the state-of-the-art predictive performance considering the wide geographical scope of the problem across all over Europe. Furthermore, a big advantage of the proposed methods respect to other solutions is that the input data configuration and the intrinsic nature of the models enable the users to predict the aforementioned targets about the next destination of vessels right after they arrive at any European port, instead of waiting for the information given by the first submitted AIS messages once their corresponding next voyage has started. When deployed into production, the resulting system will help maritime industry agents to enhance their real-time situational awareness and operational planning.

---

**Keywords** Maritime Logistics, AI, Deep Learning, Graph Clustering

## Preface

I want to thank the startup Awake.AI for giving me the opportunity of developing this project and financing this research, as well as my instructor Dr. Jussi Poikonen and Professor Arno Solin for their guidance.

Álvaro Orgaz Expósito

# Contents

<b>Abstract</b>	<b>3</b>
<b>Preface</b>	<b>4</b>
<b>Contents</b>	<b>5</b>
<b>Symbols and abbreviations</b>	<b>6</b>
<b>1 Introduction</b>	<b>7</b>
1.1 Aims . . . . .	9
1.2 Motivation . . . . .	10
1.3 Research questions and hypothesis . . . . .	11
<b>2 Background</b>	<b>13</b>
2.1 Clustering ports based on maritime traffic . . . . .	14
2.2 Predicting next destination of vessels . . . . .	19
2.3 Forecasting ETA of voyages until next destination . . . . .	19
<b>3 Research material and methods</b>	<b>21</b>
3.1 Data and resources available . . . . .	21
3.2 Hardware and software used . . . . .	28
3.3 Proposed methods . . . . .	29
<b>4 Results</b>	<b>36</b>
4.1 Clustering ports based on maritime traffic . . . . .	36
4.2 Predicting next destination of vessels . . . . .	38
4.3 Forecasting ETA of voyages until next destination . . . . .	43
<b>5 Discussion</b>	<b>47</b>
5.1 Industrial and commercial benefits . . . . .	50
5.2 Future work . . . . .	52
<b>6 Conclusion</b>	<b>54</b>
<b>References</b>	<b>55</b>
<b>A Appendix</b>	<b>58</b>

## Symbols and abbreviations

### Symbols

$T$	Transpose
$\emptyset$	Empty set
$\log$	Logarithm
$\Sigma$	Summatory
$\subseteq$	Subset of set
$\cup$	Union of sets
$\cap$	Intersection of sets
$\in$	Membership in set
$\bar{S}$	Complementary of set S

### Abbreviations

AI	Artificial Intelligence
AIS	Automatic Identification System
ANN	Artificial Neural Network
API	Application Programming Interface
CNN	Convolutional Neural Networks
CNM	Clauset, Newman and Moore
CPU	Central Processing Unit
DL	Deep Learning
ETA	Expected Time of Arrival
ETD	Expected Time of Departure
GB	Gigabyte
GPS	Global Positioning System
GPU	Graphical Processing Unit
h	Hour
IMO	International Maritime Organization
LSTM	Long Short Term Memory
ML	Machine Learning
MMSI	Maritime Mobile Service Identity
NOAA	National Oceanic and Atmospheric Administration
nmi/h	Nautical Miles per Hour
OP	Operating System
RAM	Random Access Memory
RCNN	Recurrent Convolutional Neural Network
R&D	Research & Development
RMSE	Root Mean Square Error
RNN	Recurrent Neural Network
ReLU	Rectified Linear Unit
SSD	Solid State Drive

# 1 Introduction

In this research, the problem in question arises from the issues and challenges in the maritime logistics field. Nowadays, globalization has scaled across the planet and the sea is one of the most used means of transport for passengers, goods, products, etc. However, the transportation process using vessels between ports is not well optimized internationally and several severe problems and inefficiencies appear in the whole process, such as delays due to unexpected ports full capacity reached. The consequences of these difficulties in international maritime logistics have a big negative impact in the industry worldwide, and that is why some organizations and companies are providing technological solutions for improving the universal information exchange for maritime logistics to reduce emissions, save time and costs from port calls, and improve operational planning.

Today, many ports are still struggling with inaccurate predictions of expected time of arrival (ETA) and expected time of departure (ETD), since the communication between maritime logistics actors can be slow due to siloed port operations and the inefficiency of tools for resource and task planning. Moreover, overseeing vessels and cargo movement can also be difficult with poor mapping tools. All this results in lower operational efficiency so more efficient port call operations are needed.

Furthermore, traditionally ships are served in the order they arrive at ports, which leads to queues and waiting, effectively causing loss of revenue and unnecessary emissions. Environmental sustainability has become a serious concern for policymakers [1] and thus a significant transitional force for the entire maritime transport business. Then, when ships berth, it is essential that all processes, such as unloading and loading, as well as documentation, start instantly. So digitalization and automation are rapidly changing the maritime outlook to reduce fuel and energy consumption as well as to improve the overall efficiency. Thus, profits can be increased when digital technologies are integrated into the entire operational maritime system. That is why the companies and organizations that are establishing greater digital competence are gaining a significant advantage in the ever-changing landscape of regulations and performance requirements.

In summary, there are problems and issues in the maritime logistics field that create inefficiencies in the transportation process, and the reasons why those happen can be diverse across countries. The main point to tackle is the lack of good and streamlined systems for optimization and fluent communication between international ports. In order to get a visual overview of the maritime traffic across Europe and the geographical dimension of the problem, in Figure 1 we can find a representation of the European maritime traffic density during 2017.

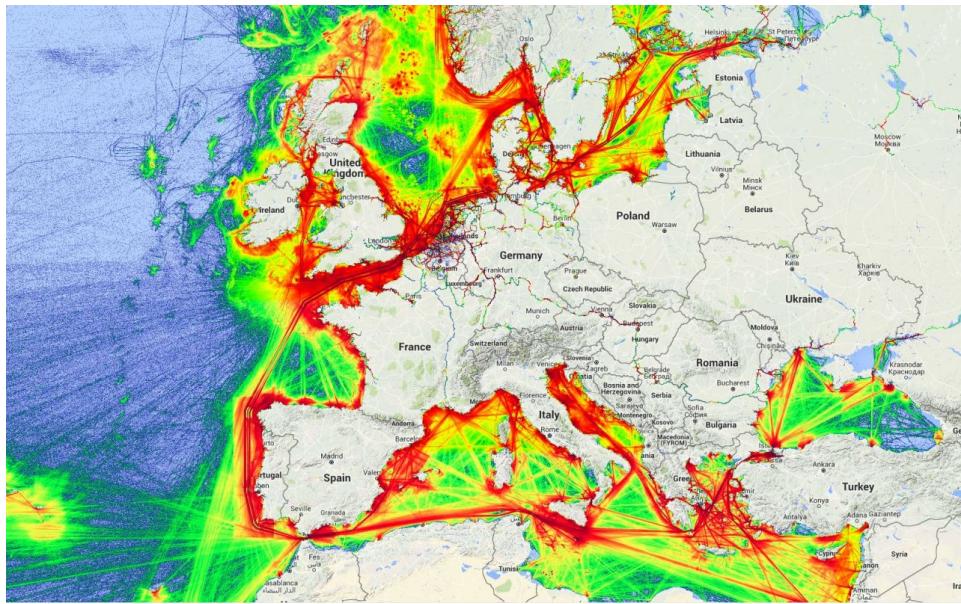
This thesis is a research and development project financed by Awake.AI [2] which is a company based in Finland that provides a top all-in-one solution for real-time collaboration and decision making between all port operations across Europe. Their portfolio of customers is diverse and they are working with multiple port operators, authorities, and logistics companies internationally. Moreover, the main benefits of their product and service are well aligned with the aims of the present research:

- Maximize the use of existing port capacity by
  - Establishing a true real-time situational awareness for the port community
  - Reducing emissions in the port area
  - Optimizing every port call
  - More efficient use of the port assets
- Operational efficiency and transparent communication by
  - Port call optimization with AI insights
  - Optimizing resource needs and plan operations more efficiently (personnel, equipment, warehouse, etc.)
  - Closer collaboration with other port call actors enabling just-in-time operations
  - Efficient and real-time operational status sharing for other port call actors
- Spend less time waiting and arrive just in time by
  - Avoiding rush to wait, but plan the best possible cargo flow instead
  - Ensuring increased capacity utilization
  - Smarter voyage planning and optimization
  - Reducing emissions and fuel costs
  - Powerful communication channel with all port actors

The plausible methods to approach these aforementioned questions can be wide, but this research focuses on computer science based techniques and more specifically, the Artificial Intelligence areas well known as Deep Learning and Machine Learning. The impact of AI has exploded over the last years in different fields and many successful applications of Deep Learning and Machine Learning techniques can be found across sectors. In the end, this research is an innovative work on the maritime logistics field based on the application of Deep Learning and Machine Learning. A summary of the impact of Artificial Intelligence on innovation is introduced by Iain M. Cockburn et al. [3] and citing from this article:

Artificial intelligence may greatly increase the efficiency of the existing economy. But it may have an even larger impact by serving as a new general-purpose “method of invention” that can reshape the nature of the innovation process and the organization of R&D.

It is important to point out that maritime logistics is not one of the most explored fields in the area of AI so this research is an innovation project, financed by a company specialized in the topic, that could be deployed into the industry if the research aims are achieved providing a successful solution for the aforementioned problems.



**Figure 1:** European maritime traffic density in 2017. Source: MarineTraffic.

## 1.1 Aims

Given the problem, a set of aims for this research have been defined to tackle the aforementioned issues in the maritime logistics field from an Artificial Intelligence based perspective. These aims are mainly:

- **clustering ports** based on maritime traffic
- **predicting the next destination** of vessels when berthed in a port
- **forecasting the expected voyage duration** until the next destination

The geographical scope of these objectives is all European waters and ports, so in all stages of the research development process, this target is very important since it affects and determines many decisions such as the selection of the methods and techniques to use, as well as the data gathering and preprocessing.

In later sections, the data and resources available are explained more extensively, but for introducing the work and aims, the technical details and concepts needed are that this research will explore how to apply Deep Learning and Machine Learning techniques to historical data of voyages across European waters and ports in order to solve the stated research aims.

In conclusion, this thesis will provide the necessary information for developing a predictive analytics system, from raw data to deployment of the optimal models into production, as a tool to improve the aforementioned current problems in the maritime logistics field. The output is intended to be a fully integrated system based on the optimal researched algorithms for each task, that complement each other to provide real-time forecasts of the overall maritime transportation scenario in Europe.

## 1.2 Motivation

This section covers the motivation and reasons why achieving the aforementioned aims can contribute as an approach for solving or enhancing the described problems.

Ideally, a system that gives a real-time and precise overview of the sea, ports and land information would be needed by decision-makers of the maritime logistics, so they can get the most accurate predictions for managing logistic events optimization. Concretely Awake.AI, the company that finances this research, has created a platform aiming for better collaboration and planning between all maritime actors. Sharing of data is an integral component of the platform which is bringing together sea, port, and land operators in one ecosystem, enabling users to allocate maritime resources better, save time and materials, as well as to schedule operations in a smarter way. This is a completely new way of thinking in terms of openness and information exchange between the maritime collaborators compared to the traditional way of operating in silos, and the outcomes of this research can drastically boost the predictive performance of this tool enabling better results and more robust and reliable information.

Regarding the first aim of the present research, describing mathematically and analytically the maritime connectivity of ports across Europe can provide very useful information to the techniques applied for other aims. For example, it would be possible to train more specialized models for each cluster of ports separately in order to boost the performance of the AI algorithms. Thus, finding a way to cluster the harbours based on their maritime connectivity can be crucial for addressing the problem more specifically at port level.

Regarding the second and third aims, these are directly connected since it is trivial to think that, for example, when informing that a vessel V berthed now at port A is heading next to port B, the expected time T until arrival to the destination harbour is needed for having complete information. Achieving a good predictive model at the European level for the next destination and ETA of vessels berthed at any port can be significant for optimizing maritime logistics internationally and not only between ports in the same geographical area. All this situational information and real-time predictive tools pave the way for efficient port call planning, maritime operations ports, determining and optimizing the fleet utilization, reducing the shipping emissions, and solving other aforementioned challenges. Furthermore, it is crucial to analyze the main stakeholders or users that would directly benefit out of it:

- **Port authorities.** Those can maximize the throughput of the port without expensive investments on physical infrastructure. The port authorities and local port communities can have highly improved situational awareness and thus can effectively optimize every port call. The research outcomes can also be used by port authorities to provide informative digital services to their community. All this leads to more efficient use of port assets and fewer emissions.
- **Ship operators.** Those can reduce turnaround times, slow steam to save cost on fuel, reduce emissions, and increase the utilization of their fleet. They could have the possibility of a better communication channel with all European

maritime actors and then make sure their fleet spends less time waiting and arrives just in time.

- **Terminal operators.** Those would be able to increase the loading and unloading efficiency (tonne per hour) by planning the service needs earlier and optimizing future operations with needed detail. They could collaborate with other port call actors in real-time, allowing them to have better preparation for port operations by getting a more accurate and real-time operational overview and status sharing with other port call actors.
- **Cargo owners.** Those would get more transparency and planning of the cargo flow at sea, port, and on land, enabling better tracking, tracing, and optimization of their cargo flow going through the port gates. This provides them with an excellent overview to better prepare for deliveries and pick-ups at the right time and ensures efficient and reliable operations for days. Besides, all of this has a positive impact on emissions in their logistics chain as well.

In conclusion, the motivation under the research aims lies in enabling smarter operational planning and improved situational awareness for all maritime logistics chain actors in real-time. This is achieved through enhancing the visibility of future operational status and international voyages across European waters several days ahead, as well as enabling a better flow of information between all maritime actors.

### 1.3 Research questions and hypothesis

The research questions of this thesis rely on the main aims of the project. Firstly, understanding better the European ports structure based on their maritime connectivity can end up being important information for solving and have better interpretability of the research problem. There is not a clear or straightforward way to cluster the ports but understanding their connectivity as a mathematical graph, where nodes or vertices are the ports and the edges are the number of voyages between ports, can lead to successful approaches and experiments for clustering the harbours. Thus, the hypothesis is that different Machine Learning techniques and methods that have been applied for graph clustering could be applied to this problem too.

Secondly, predicting the next destination of vessels when berthed in a port may not have been extensively approached using Deep Learning models. Then, the hypothesis is that there is open room for innovation and discover suitable applications of Deep Learning techniques for this aim. However, beforehand it seems challenging to come up with a suitable approach based on the wide international scope of the research.

Thirdly, forecasting the expected voyage duration is a problem that has been well studied in different areas, concretely by using Deep Learning and Machine Learning methods as a regression problem. However, the hypothesis for it applied to the maritime logistics field, considering the available research data, is that it is feasible but not trivial to find or reproduce solutions that reach the necessary accuracy when predicting voyages duration until the next destination. This is also because from the

data engineering point of view, the model or system created should be robust respect to the data gathering and ingestion difficulties of this real-time big data problem.

Finally, in terms of deployment of the research outcomes into the industry through Awake.AI, the company that finances this thesis, some hypotheses are also considered. Leading ETA prediction accuracy and near future port call forecasts can be critical for the success of the digitalization of the maritime logistics, however many challenges and difficulties can arise when deploying the optimal models into production since for example these will be based on the data sources shared across Europe in real-time from thousands of vessels. So managing this real-time big data problem can present unexpected impediments.

## 2 Background

The application of AI into the maritime logistics field has not been extensively studied in the literature since the international maritime transportation has scaled up in the last decades, so the need to optimize and enhance maritime logistics processes from an academical point of view is recent. In this section, an overview of major previous and current researches on this topic is firstly provided, and secondly, a more mathematical literature review is developed separately for each research aim by focusing on Deep Learning and Machine Learning based methods.

Early work in this area was performed in 2006 by Panayides et al. [4] whose work studied and explained how the derived demand for maritime transport has evolved from a demand for the possession of goods to an integrated demand for the possession of goods that has added value, timely, reliably and cost-efficiently, and has given rise to the concept of maritime logistics. This paper discussed the evolution of this concept by reviewing the contributions in the field made by the best papers on the topic presented at the International Association of Maritime Economists (IAME) 2005 conference.

After some years in 2013, the same author Panayides et al. [5] formulated the evolution of maritime logistics as an emerging discipline that has resulted, to a large extent, from the increasing and varied demands of shippers and customers, and the rapidly changing role of ports in the context of supply and logistics chains. This paper explained that scholars are becoming increasingly aware of the need to integrate logistics and supply chain management concepts in the maritime transportation chain and operations, providing a review and foundation for understanding the domain of this field and to assess its potential as an emerging discipline that adopts economic and management perspective.

Focusing now into AI based research, more recent researches in 2018 include work by Leclerc et al. [6] which leveraged Deep Convolutional Neural Networks in order to do ship classification for maritime target tracking. In the last years, the state-of-the-art in computer vision has improved greatly thanks to increased use of this Deep Learning technique, advances in graphical processing units (GPU) acceleration and the availability of large labelled datasets such as ImageNet. This paper explained that obtaining datasets as comprehensively labelled as ImageNet for ship classification remains a challenge. As a result, in this paper, the authors experimented with pre-trained CNNs based on the Inception and ResNet architectures to perform ship classification. Instead of training a CNN using random parameter initialization, transfer learning is used. Then, fine-tuning was performed in the pre-trained CNNs to perform maritime vessel image classification on a limited ship image dataset, achieving a significant improvement in classification accuracy compared to the previous state-of-the-art results for the Maritime Vessel (Marvel) dataset.

An interesting variant of the ship classification problem was developed by Dao et al. [7]. This paper considered the ability to identify maritime vessels and their type as an important component of modern maritime safety and security, and this work presented the application of Deep Convolutional Neural Networks to the classification of maritime vessel images.

Furthermore, Marie et al. [8] leveraged real-time maritime situation awareness based on Deep Learning with dynamic anchors in 2018. This paper conceived that situation awareness in the maritime environment entails early detection and classification of maritime targets of varying sizes, depths, shapes, textures, and contrasts. Thus, this work described a novel Deep Learning based maritime situation awareness approach using high-definition video in which object detection is achieved in three main steps. At first, a key region based tracking algorithm allows to, dynamically and parsimoniously, extract high-quality region proposals mainly focalized around rigid video locations. The latter are, further, fed into a Fast-RCNN for carrying out objectness detection and box regression. Finally, a mere box post-regression operation enables the extraction of maritime objects. Furthermore, the found object detections are fed into a second classification RCNN, specifically, trained to recognize up to 40 vessel classes. The experiments showed that the proposed approach achieves state-of-the-art speed and accuracy.

Shortly thereafter in 2019 Ellefsen et al. [9] formulated a solution for a completely different topic inside maritime logistics, which is a fault detection algorithm for maritime components. This paper explained that in recent years, the reliability and safety requirements of ship systems have increased drastically. This has prompted a paradigm shift toward the development of prognostics and health management approaches for these systems' critical maritime components. Any solution should include independent and intelligent fault detection algorithms that can report faults automatically, and this work proposed an unsupervised reconstruction based fault detection algorithm for maritime components. The results suggested that the algorithm is highly suitable to be included as part of a pure data-driven diagnostics approach in future end-to-end system solutions.

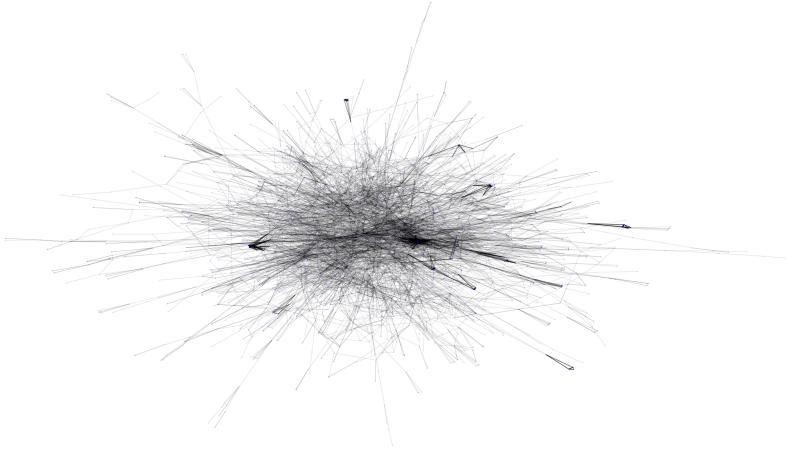
## 2.1 Clustering ports based on maritime traffic

This section describes literature about methods that have been used or are suitable for achieving the first aim of the present research, which is clustering ports based on maritime connectivity and understanding it as a mathematical graph where nodes or vertices are the ports and the edges the number of voyages between them.

The task of partitioning a graph has become quite popular since it is possible to find ubiquitously in real-world network several sets of densely connected nodes, joined by a small number of edges. Moreover, as it can be seen in Figure 2, this task is not trivial at all as normally no clear partitions are found in the graph, and in fact, graph partitioning is an NP-hard problem. Therefore, there are a lot of different approaches to tackle the problem, so some of them are introduced to get an overview of the problem and its complexity before implementing the proposed method.

Before going into further details about the literature for graph clustering, a mathematical formulation of this research problem is provided.

Given a directed graph  $G = (V, E)$  and an integer  $k > 1$  we want to split the set of vertices  $V$  into  $k$  communities  $V_1, \dots, V_k$  so that  $\bigcup_{i=1}^k V_i = V$  and  $V_i \cap V_j = \emptyset$  for all  $i \neq j$ . We want the communities  $V_1, \dots, V_k$  to be



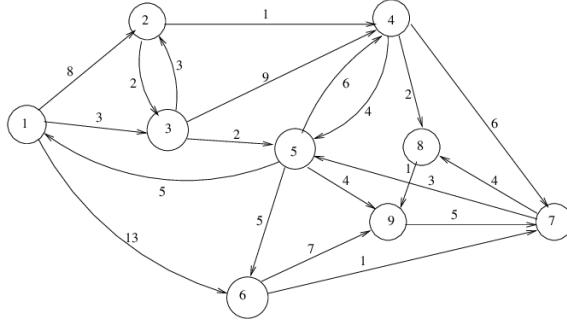
**Figure 2:** Example representation of the popular graph ca-GrQc.

as much separated from each other as possible. We also want that the communities have roughly equal size.

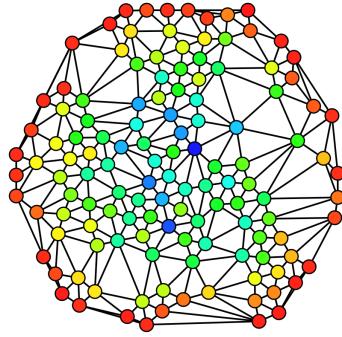
Then, the goodness of a partition  $V_1, \dots, V_k$  is based on reducing the values of  $E(V_i, \bar{V}_i)$  where  $V_i, \bar{V}_i \subseteq V$  with  $V_i \cap \bar{V}_i = \emptyset$  and  $E(V_i, \bar{V}_i)$  is defined to be the set of edges of  $G$  with one endpoint in  $V_i$  and the other endpoint in  $\bar{V}_i$ , i.e.,  $E(V_i, \bar{V}_i) = \{(u, v) \in E \mid u \in V_i \text{ and } v \in \bar{V}_i\}$ . Also defining  $\bar{V}_i = V \setminus V_i$ .

### 2.1.1 Dealing with directed graphs

Even though undirected graph clustering has been well studied in the literature, an interesting variant of the problem is directed graphs in which the order of the origin and destination vertices matter and an example representation of it is provided in Figure 3. A relevant and extensive paper was developed in 2013 by Fragkiskos et al. [10]. This work described how networks or graphs appear as dominant structures in diverse domains, including sociology, biology, neuroscience, and computer science. In most of the aforementioned cases, graphs are directed, in the sense that the edges are directional, making the semantics of the edges non-symmetric. Then, revealing the underlying community structure of directed complex networks has become a crucial and interdisciplinary topic with a plethora of applications. This paper contained a review of the methods for clustering directed networks along with the relevant necessary methodological background and also related applications. Finally, it presented the relevant work along with two orthogonal classifications: the first one was mostly concerned with the methodological principles of the clustering algorithms, while the second one approached the methods from the viewpoint regarding the properties of a good cluster in a directed network.



**Figure 3:** Representation of a directed mathematical graph.



**Figure 4:** A graph colored based on the betweenness centrality [12] of each vertex from least (red) to greatest (blue). Source: Wikipedia.

### 2.1.2 The Girvan-Newman method

The method introduced in 2002 by Girvan M. and Newman M. [11] is based on the fact that edges connecting communities should have high **betweenness centrality** which means that for every pair of vertices in a connected graph, there exists at least one shortest path. The betweenness centrality for each vertex is the number of shortest paths that pass through the vertex. Therefore, it removes links in order to decrease betweenness and returns the remaining components of the network as the communities obtained. Its main downside is that it makes heavy demands on computational resources, running in  $\mathcal{O}(m^2n)$  with  $m$  edges and  $n$  vertices or  $\mathcal{O}(n^3)$  on a sparse graph.

---

**Algorithm 1:** The Girvan-Newman method.

---

1. Calculate the betweenness for all edges in the network.
  2. Remove the edge with the highest betweenness.
  3. Recalculate betweennesses for all edges affected by the removal.
  4. Repeat from step 2 until no edges remain.
-

### 2.1.3 CNM method

The Clauset, Newman and Moore (CNM) method [13] is based on increasing **modularity** of a network. The modularity is an evaluation metric for community detection and it tests a given division of a network against the random division. The algorithm measures when a division is a good one, in the sense that there are many edges within communities and only a few between them. It starts with  $N$  communities and at every step of the algorithm two communities that contribute maximum positive value to global modularity are merged.

### 2.1.4 Spectral methods

Spectral clustering methods use the spectrum of a graph to perform dimensionality reduction before clustering in fewer dimensions. It is better to use this approach than directly using KMeans [14] because it performs poorly since it can only find spherical clusters. The idea is based on computing the eigenvectors corresponding to the smallest eigenvalue of the normalized Laplacian or some eigenvector of some other matrix representing the graph structure [15], [16], [17]. The resulting eigenvector is used as a vertex embedding of the graph to determine the clustering. Its main downside is that computing eigenvalues and eigenvectors for graphs are slow and hence such methods might face scalability issues when applied to massive graphs. Normally, one common way to tackle this problem is to use the more efficient method *Implicitly Restarted Lanczos* where only the  $k$  largest or smallest eigenvalues and its eigenvectors are needed.

For example, if a graph  $G$  is a collection of  $k$  disjoint cliques, the normalized Laplacian is a block-diagonal matrix that has eigenvalue zero with multiplicity  $k$  and the corresponding eigenvectors serve as an indicator of the membership of each clique: the eigenvector  $v_i$  has a different value or a larger magnitude for the vertices that are inside the clique  $i$  than the other vertices. Thus, there is an underlying structure that can be seen using the eigenvectors of the Laplacian. Moreover, if we introduce edges between the cliques we will find that  $k - 1$  of the  $k$  eigenvalues that were zero will become slightly larger than zero so it is a robust method.

## Recursive Spectral Bi-partitioning

This method uses one of the typical ideas of spectral clustering: computing the eigenvector corresponding to the second-smallest eigenvalue of the normalized Laplacian. The resulting eigenvector works like an embedding of the vertices to split the graph into two clusters: a positive value in the  $i$  position of the eigenvector indicates that vertex  $i$  belongs to a cluster  $C_1$  and a negative value that it belongs to a cluster  $C_2$ .

In order to do more than two partitions, we can perform a two-classification iteratively, using the spectra of the resulting induced subgraphs previously created. This will yield a divisive hierarchical clustering algorithm which is called Recursive Spectral Bi-partitioning [18].

---

**Algorithm 2:** Recursive Spectral Bi-partitioning algorithm.

---

**Input:** A graph  $G$  and the desired number of clusters  $k$ .

**Preprocessing:** Build Laplacian matrix  $L$  of graph  $G$ .

**Decomposition:**

1. Find eigenvectors  $X$  and eigenvalues  $\lambda$  of the matrix  $L$ .
2. Sort the eigenvalues:  $\lambda_1 < \lambda_2 < \lambda_3 \dots < \lambda_n$ .
3. Map vertices to the components of the eigenvector corresponding to  $\lambda_2$ .

**Grouping:**

4. Sort components of reduced 1-dimensional vector.
5. Identify clusters by splitting the sorted vector into negative and positive.
6. Induce two subgraphs using the clusters identified.

**Recursion:**

7. Recurse with the induced subgraphs until  $k$  partitions are reached.
- 

## K-Way Spectral Clustering

Another popular method to tackle the multicluster problem that is actually more commonly used is the K-Way Spectral Clustering introduced back in 2000 by Jianbo Shi and Malik [19]. This method instead of using only the eigenvector associated with  $\lambda_2$  from the Laplacian matrix it uses multiple eigenvectors. Afterwards, any multidimensional clustering algorithm can be performed such as KMeans.

---

**Algorithm 3:** K-Way Spectral Clustering algorithm.

---

**Input:** A graph  $G$  with  $n$  nodes, the desired number of clusters  $k$ , and the desired number of components to keep  $c$ .

**Preprocessing:** Build Laplacian matrix  $L$  of graph  $G$ .

**Decomposition:**

1. Find eigenvectors  $X$  and eigenvalues  $\lambda$  of the matrix  $L$ .
2. Embed the space from eigenvectors associated to  $c$  smallest eigenvalues.

**Clustering:**

3. Apply KMeans to the reduced  $n \times c$  space to produce  $k$  clusters.
-

## 2.2 Predicting next destination of vessels

This section describes literature about methods that have been used or are suitable for achieving the second aim of this research, which is to predict the next European destination of vessels berthed in a port using their historical voyages data.

Recent approaches include work in 2018 by Nguyen et al. [20] which leveraged maritime surveillance with a multi-task Deep Learning architecture for using AIS data streams. Although this paper does not relate directly to destination prediction, some techniques included have been used as an inspiration for the proposed method later on for this task. This paper described how in a world of global trading, maritime safety, security, and efficiency are crucial issues. Then, it proposed a multi-task Deep Learning framework for vessel monitoring using Automatic Identification System (AIS) data streams, which combines Recurrent Neural Networks with latent variable modelling and an embedding of AIS messages to a new representation space to jointly address key issues to be dealt with when considering AIS data streams. This work demonstrated the relevance of AI for maritime surveillance which is another important factor in the maritime logistics field.

However, for the sake of this research, it is very important to mention some earlier work in a completely different field that was performed by Graves et al. [21] back in 2013, and that has been an inspiration for the innovative solution proposed in this research. This paper is about generating sequences with Recurrent Neural Networks and shows how Long Short Term Memory RNNs can be used to generate complex sequences with long-range structure. The approach was demonstrated for text (where the data is discrete) and online handwriting (where the data is real-valued). It is then extended to handwriting synthesis by allowing the network to condition its predictions on a text sequence. The resulting system could generate highly realistic cursive handwriting in a wide variety of styles.

## 2.3 Forecasting ETA of voyages until next destination

This section describes literature about methods that have been used or are suitable for achieving the third aim of this research, which is forecasting the voyage duration or expected voyage duration until the next destination.

Recent Machine Learning approaches for forecasting time of arrival include work by Yin et al. [22] in 2017 which leveraged a prediction model of bus arrival time at stops with multi-routes. As in the maritime logistics field, accurate bus arrival time is fundamental for efficient bus operation and dispatching decisions. This paper proposed a new prediction model based on Support Vector Machine (SVM) and Artificial Neural Networks (ANN) to predict arrival time at an objective destination with multi-routes. The preceding arrival time of the objective route and all other routes passing by the same destination as well as the travel speed of the target were inputs of the model.

Furthermore, more specifically related to this research topic, shortly thereafter Bodunov et al. [23] proposed a solution for real-time destination and ETA prediction for maritime traffic. This paper presents an approach to provide a prediction for

a destination and the arrival time of ships in a streaming-fashion using geographic spatial data in the maritime context. Novel aspects of this approach include the use of Ensemble Learning based on several Machine Learning models such as Random Forest, Gradient Boosting Decision Trees, XGBoost and Extremely Randomized Trees in order to provide a prediction for a destination, while for the arrival time they propose the use of Feedforward Neural Networks. Although the aims of this paper were very similar to the aims of the present research, the geographical scope in there was only the Mediterranean sea so a way lower number of ports is considered.

## 3 Research material and methods

In this section, firstly the main data and resources used in the research will be described in section 3.1 as well as the preprocessing applied to it for experimenting with the proposed methods. Secondly, in section 3.2 a list of the hardware components and software versions used will be provided. Thirdly, a conceptual and mathematical explanation of the proposed methods by this research will be given in section 3.3 for each research aim separately.

### 3.1 Data and resources available

The data available for this research is mainly provided by the company Awake.AI which finances this project. They are specialized in maritime logistics with several real-time AI based products, so they have solid data pipelines that fetch several data sources every day to update their databases. Concretely, the content of the database used for this research contains historical maritime traffic data across Europe, including positional information about the vessels and their voyage route. It is more extensively described in the following subsections but basically, the dataset has been built from two main data sources:

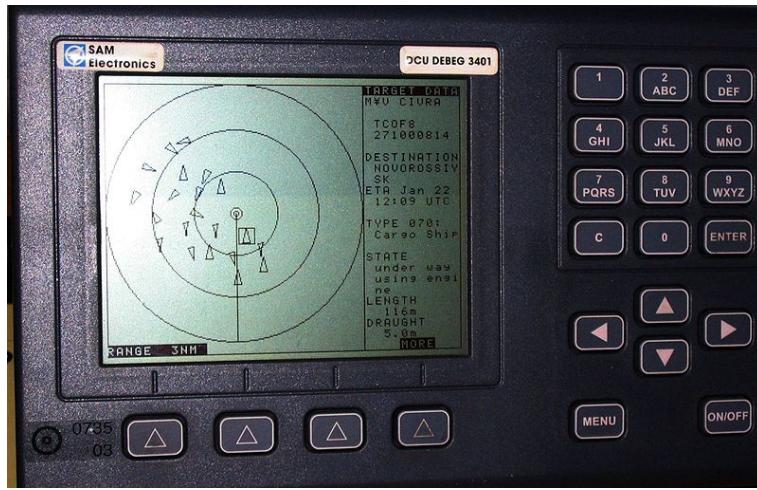
- Public governments information and databases about their maritime infrastructure and ports, in order to obtain trustful information such as geographical coordinates across Europe and other relevant metrics related to harbours capacity, dimensions, main maritime routes, etc.
- Automatic Identification System (AIS) [24] messages data which is extensively described in the following subsection.

It is important to remark the essence of this research from a data engineering perspective. The deployment into the industry of the outcomes of this project is a real-time big data problem that requires solid data infrastructure and pipelines to support robustly without failures the huge amount of information sent across Europe every second. In the following subsections, more details about the scope and data dimensions are described to have a more comprehensive overview of the problem dimensions.

#### 3.1.1 Automatic Identification System (AIS)

The Automatic Identification System (AIS) [24] is an automatic tracking system that uses transponders on ships and is used by vessel traffic services. AIS is a maritime navigation safety communications system, that supplements marine radar, standardized by the International Telecommunication Union (ITU) and adopted by the International Maritime Organization (IMO).

It provides vessel information, including the vessel's identity, type, position, course, speed, navigational status, and other safety-related information automatically to appropriately equipped shore stations, other ships, and aircraft. Moreover, it



**Figure 5:** AIS equipped system on board a ship presents the bearing and distance of nearby vessels in a radar-like display format. Source: Wikipedia.

receives automatically such information from similarly fitted ships, monitors and tracks ships, and exchanges data with shore based facilities. Imagine a shipboard radar or an electronic chart display that includes a symbol for every significant ship within radio range, each with a velocity vector and heading and with a symbol that reflects the actual size of the ship, with position to GPS. By clicking on a ship symbol, you can learn the ship name, course and speed, classification, call sign, registration number, MMSI, and other information. Thus, display information that was previously available only to modern vessel traffic service operations centres can be available to every AIS user as shown in Figure 5.

Moreover, it is intended to assist a vessel's watchstanding officers and allow maritime authorities to track and monitor vessel movements. AIS integrates a transceiver with a positioning system such as a GPS receiver, with other electronic navigation sensors, such as a gyrocompass or rate of turn indicator. Vessels fitted with AIS transceivers can be tracked by AIS base stations located along coastlines or, when out of range of terrestrial networks, through a growing number of satellites that are fitted with special AIS receivers which are capable of deconflicting a large number of signatures.

Finally, about the content of the AIS messages fetch from APIs, there are several types based on their priorities and access schemes used to transmit data. The two main types of messages that have been used to create the dataset of the present research are:

- **Position report.** This type of messages are broadcasted every 2 to 10 seconds during voyages, and every 3 minutes while at anchor. The metadata list of parameters or variables included in them is in Appendix Table A1.
- **Static and voyages data.** This type of messages is broadcasted every 6 minutes during voyages. The metadata list of parameters or variables included in them is in Appendix Table A2.

However, in the case of *static and voyages data*, some variables of the message information such as the expected time of arrival, destination, and ship type are provided by the responsible on board of the ship instead of automatically by a transceiver. This leads to low quality and non-reliable data which opens the room to the research problem for enhancing the maritime logistics across European waters.

### 3.1.2 Metadata

This section provides a thorough description and details of the data used in this research. The main dataset provided by Awake.AI for this project is based on historical maritime traffic data, containing preprocessed information of all available maritime voyages in European waters in 2020 from approximately 557 million public AIS messages sent by vessels during their voyages.

Concretely, the timeframe is all maritime voyages between January 2020 and May 2020 included. Before any data preprocessing in this thesis, the provided dataset contains 3,528,644 voyages of 37,650 different vessels (considering their distinct MMSI) and from/to 6,377 different departures/destinations harbours all over Europe. However, the data ingestion pipelines are continuously running in real-time and in case of deployment of the research outcomes, a more wide and updated timeframe could be considered for retraining the optimal models and system.

Since the dataset is provided by the company Awake.AI, there is intrinsic preprocessing in the data ingestion pipelines used for fetching the data from the original APIs of AIS. As aforementioned, vessels across Europe send real-time AIS messages to stations every a specific amount of time with a unique Maritime Mobile Service Identity (MMSI) during the voyage, berth and anchor. However, the shape of the provided dataset is one unique observation row per voyage, which means that all the messages across the vessel trajectory from the departure to the destination port have been summarised.

In Table 1 you can find the list of variables in the dataset used for the experiments of the proposed methods in this research. It excludes all the parameters and variables from, for example, AIS message that are not robust or reliable for the modelling.

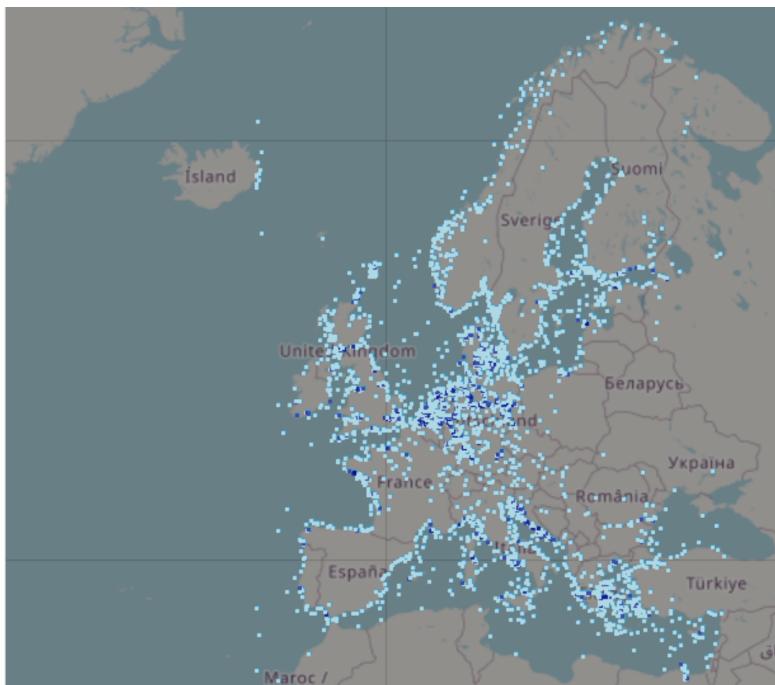
### 3.1.3 Data preprocessing

This section describes the preprocessing done to the aforementioned data before applying the proposed methods in all the experiments of this research.

Firstly, the departure and destination codes have been cleaned by deleting the extra port code information such as: ‘ANCHR near’, ‘anchorage\_area\_’, ‘anchorage\_’, ‘berth\_area\_’, ‘berth\_’, ‘pilot\_boarding\_’. Once the port codes are cleaned, the total number of different ports is 3,374 and in Figure 6 you can see the European map of all the ports. Note that this map is showing the coordinates of the departure and destination codes included in the dataset of this research, however, as you can see there are some ports inside land which correspond to harbours in rivers or canals, and also few coordinate points in the middle of the sea which correspond to anchoring areas used, for example, when ports berth capacity is reached.

**Table 1:** List of available variables used in this research.

Variable	Description
MMSI	MMSI number of the vessel.
Ship type	Type of vessel.
Departure code	Code of the departure port.
Destination code	Code of the destination port.
Departure latitude	Coordinate latitude of the departure port.
Destination latitude	Coordinate latitude of the destination port.
Departure longitude	Coordinate longitude of the departure port.
Destination longitude	Coordinate longitude of the destination port.
Departure date	Timestamp of the departure date.
Destination date	Timestamp of the arrival date.
Maximum velocity	Maximum velocity during the voyage in nmi/h.
Median velocity	Median velocity during the voyage in nmi/h.
Voyage length	Total length of the voyage in nmi/h.

**Figure 6:** Map of European ports included in the research dataset. The color scale is based on ports concentration by coordinate from least (light blue) to greatest (dark blue).

Secondly, the voyages with unknown departures and destinations have been filtered out. This can happen due to multiple issues with messages such as lost data or wrong information. So in this step, 11% of the voyages are dropped.

Next, due to the aim of the research and its international scope across Europe, voyages with the same departure and destination are filtered out. This happens in 53% of the voyages since in this data voyages of, for example, rescue boats or pilot vessels are included, as well as fishing boats that end voyages in the same port than the departure. Overall, after filtering out these cases 1,130,148 out of 3,528,644 initial voyages are remaining. Moreover, in Table 2 you can find the distribution of the length of the voyages.

**Table 2:** Distribution of the voyages length.

Quantile	Voyage length (nmi)
10%	0.92
20%	2.13
30%	3.45
40%	5.14
50%	7.84
60%	12.34
70%	20.03
80%	33.15
90%	75.72
95%	159.16
97.5%	290.41

Thereafter, since there are ports quite close to each other, as you can see in Figure 6, an aggregation of ports by closeness is developed. The algorithm pseudocode description of the aggregation strategy used is shown down below. In the experiments section, the threshold distance to group ports close to each other is a parameter to experiment with, since logically the lower number of different ports the easier is to achieve higher predictive performance in the models. However, the bigger the threshold the lower geographical resolution of the model or system. Down below you can also see the number of resulting ports after aggregation by closeness thresholds.

**Table 3:** Resulting number of ports after aggregation by closeness thresholds.

Closeness threshold (km)	Ports after grouping
$\leq 1$	3,189
$\leq 15$	1,616
$\leq 25$	1,012

---

**Algorithm 4:** Pseudocode of the strategy used to aggregate ports.
 

---

**Input:** List of ports with their coordinates and closeness threshold  $T$  in km.

**Preprocessing:**

1. Build matrix  $D$  with distances between all pairs of ports coordinates using  
**function**  $distance(lat1, lon1, lat2, lon2)$ :  
 $lat1, lon1, lat2, lon2 \leftarrow rad(lat1), rad(lon1), rad(lat2), rad(lon2)$   
 $dlon, dlat \leftarrow lon2 - lon1, lat2 - lat1$   
 $a \leftarrow \sin^2(dlat/2) + \cos(lat1) \cdot \cos(lat2) \cdot \sin^2(dlon/2)$   
**return**  $6373 \cdot 2 \cdot \arctan2(\sqrt{a}, \sqrt{1-a})$

**Grouping:**

2. Filter pairs of ports close to each other creating a mask  $M = D \leq T$ .
  3. Iterate pairs of ports close to each other and group them using  
 $ports\_rows, ports\_cols \leftarrow M.where()$   
 $group\_id \leftarrow 0$   
 $groups\_of\_near\_ports \leftarrow dictionary()$   
**for**  $port\_row, port\_col$  **in**  $zip(ports\_rows, ports\_cols)$  **do**  
 $row\_group, col\_group \leftarrow None, None$   
**if**  $length(groups\_of\_near\_ports) = 0$  **do**  
 $groups\_of\_near\_ports[group\_id] \leftarrow set(port\_row, port\_col)$   
 $group\_id \leftarrow group\_id + 1$   
**continue**  
**for**  $group$  **in**  $groups\_of\_near\_ports$  **do**  
**if**  $port\_row$  **in**  $groups\_of\_near\_ports[group]$  **do**  
 $row\_group \leftarrow group$   
**if**  $port\_col$  **in**  $groups\_of\_near\_ports[group]$  **do**  
 $col\_group \leftarrow group$   
**if not**  $row\_group$  **and not**  $col\_group$  **do**  
 $groups\_of\_near\_ports[group\_id] \leftarrow set(port\_row, port\_col)$   
 $group\_id \leftarrow group\_id + 1$   
**elif not**  $row\_group$  **do**  
 $groups\_of\_near\_ports[col\_group].add(port\_row)$   
**elif not**  $col\_group$  **do**  
 $groups\_of\_near\_ports[row\_group].add(port\_col)$   
**elif**  $row\_group \neq col\_group$  **do**  
 $group\_row \leftarrow groups\_of\_near\_ports[row\_group]$   
 $group\_col \leftarrow groups\_of\_near\_ports[col\_group]$   
 $groups\_of\_near\_ports[group\_id] \leftarrow group\_row \cup group\_col$   
 $group\_id \leftarrow group\_id + 1$   
 $groups\_of\_near\_ports.pop(row\_group)$   
 $groups\_of\_near\_ports.pop(col\_group)$
-

Then, the specific feature engineering for the models is applied. One the one hand, for predicting next destination, chronological sequences of ports by vessels are generated. So given a sequence length, all the possible sequences of visited ports with this length are created for each vessel. Down below you can see the pseudocode used for creating them and the distribution of the number of voyages by vessels.

---

**Algorithm 5:** Pseudocode of the strategy used to create voyages sequences.

---

**Input:** Desired sequence length  $L$  and voyages data including MMSI vessels identifier as well as departure and destination ports and dates.

**Sequences generation:**

```

sequences ← list()
for mmsi in data.mmsi.unique() do
    data_mmsi ← data.subset(data.mmsi = mmsi)
    if length(data_mmsi) ≥ L do
        data_mmsi.sort(by=departure_date)
        departures ← data_mmsi.departure
        for i in range(length(df_mmsi) - L + 1) do
            sequences.append(departures[i : (i + L)])

```

---

**Table 4:** Distribution of the number of voyages by vessels.

Quantile	Number of voyages
0%	1
25%	2
50%	8
75%	29

Another specific feature engineering for forecasting ETA is applied. On the one hand, the ship type variable has been summarized in 5 levels *Missing*, *Cargo*, *Passenger*, *Tanker* and *Others*, and also one hot encoding [25] has been applied to it. In Table 5 you can find the resulting distribution of the ship types. On the other hand, min-max normalization [26] has been applied to numerical features such as the *maximum velocity*, *median velocity*, and *voyage length*.

Finally, two different target variables for forecasting ETA have been created in order to perform experiments.

- Firstly, the voyage duration in hours computed from the difference between the departure and destination dates, and in Table 6 you can find the distribution of the duration of the voyages in the dataset.

**Table 5:** Distribution of the ship types.

<b>Ship type</b>	<b>Distribution</b>
<b>Cargo</b>	33.45%
<b>Missing</b>	10.98%
<b>Other</b>	9.89%
<b>Passenger</b>	30.42%
<b>Tanker</b>	15.26%

- Secondly, the voyage duration in hours divided by the median duration between the corresponding departure and destination ports.

**Table 6:** Distribution of the voyage duration in hours.

<b>Quantile</b>	<b>Voyages duration (h)</b>
25%	0.43
50%	1.17
60%	1.83
70%	2.84
80%	4.72
90%	9.31
95%	16.62
97.5%	28.25

Moreover, it is trivial to think that when predicting new cases in future, some information about the voyages such as the voyage length or velocity, will be unknown beforehand. However, the voyages history of each vessel and also all the voyages between the corresponding departure and destination ports will enable us to find good estimates of, for example, the velocity as well as the voyage length. Furthermore, to predict fairly with the trained models, it will be necessary to do the same preprocessing than in the training data including, for example, the same minimum and maximum values in the min-max normalization for the numerical features.

### 3.2 Hardware and software used

This section covers the specifications of the hardware components and software versions used to run the research experiments.

#### 3.2.1 Hardware

- CPU: Intel(R) Core(TM) i7-7700HQ 2.8GHz 4 Core(s) 8 Logical Processor(s)

- RAM: DDR4-2400 16GB
- GPU: 1x NVIDIA GEFORCE GTX 1050 Ti

### 3.2.2 Software

The Deep Learning framework used is Keras running on GPU and CUDA, which is the NVIDIA's parallel computing architecture that enables dramatic increases in computing performance by harnessing the power of the GPU.

- CUDA version: v9.0
- NVIDIA drivers: v22.21.13.8554
- Programming language: Python 3.6.8
- Python modules: keras v2.2.4, numpy v1.16.0, pandas v0.25.3, sklearn v0.0, scipy v1.2.0, networkx v2.4

## 3.3 Proposed methods

This section describes the concept and mathematical formulation of the proposed methods for each research aim, including their input and output, model, and training protocol. As aforementioned the research aims are:

- clustering ports based on maritime traffic
- predicting the next destination of vessels when berthed in a port
- forecasting the expected voyage duration until the next destination

### 3.3.1 Spectral graph clustering (to cluster ports)

This section describes the proposed model for the first research aim, which is to cluster ports based on maritime traffic, as well as the training procedure implemented.

#### Input & Output

The input for this task is the directed graph containing the maritime traffic or voyages between ports, composed by different nodes which are the ports available and a list of edges between nodes which are all the voyages in the research dataset. Moreover, the number of desired clusters has to be provided too as a model parameter.

As aforementioned in the previous data cleaning section 3.1.3, the list of ports available in the research dataset has been preprocessed, including ports codes cleaning, filtering out unknown departures and destinations, as well as voyages with the same departure and destination, and aggregation of ports by closeness due to the international scope of the aims.

The output for this task is a cluster identifier for each port or node in the input graph. So basically, the list of ports is grouped in as many clusters as specified based on their maritime traffic connectivity.

## Model

The proposed approach is based on the algorithm 3 mentioned in the previous background section 2, but including modifications to deal with directed graphs. The proposed method is described in detail down below.

---

**Algorithm 6:** Proposed method used to cluster ports.

---

**Input:** A directed graph  $G$  with  $n$  nodes or ports containing the maritime traffic or voyages between ports as graph edges, the number of desired clusters  $k$ , and the desired number of components to keep  $c$ .

**Preprocessing:**

1. Compute the adjacency matrix  $A$  of the directed graph.
2. Transform the directed graph into undirected by  $A = A + A^\top$ , as explained in Fragkiskos et al. [10] paper.
3. Build the Laplacian matrix of  $G$  as  $L = D - A$  where  $D$  is the diagonal matrix containing the sum of edges/connections to other nodes/ports for each port.

**Decomposition:**

4. Find eigenvectors  $X$  and eigenvalues  $\lambda$  of the matrix  $L$ .
5. Embed the space from eigenvectors associated to  $c$  smallest eigenvalues.

**Clustering:**

6. Apply KMeans to the reduced  $n \times c$  space to produce  $k$  clusters, using the distance metric *Euclidean*, and initializing centroids based on quantiles of the embeded space dimensions.
- 

**Training**

In the experiments of this research task, the algorithm has been implemented on Python3 using the modules *numpy* for basic mathematical operations, *networkx* to deal with the graphs sparsely and create their adjacency matrix as well as the Laplacian matrix, *sklearn* for the KMeans algorithm, and *scipy* for the eigendecomposition. In fact, the first implementation did not use specific functions for sparse matrices, but later it was noticed that the graph used for the research is very sparse.

### 3.3.2 RNN with LSTM & Embedding layers (to predict destination)

This section describes the proposed method for the second research aim, which is to predict the next destination of vessels when berthed in a port, as well as the training procedure implemented.

#### Input & Output

The input for this task has been previously described in the section of data preprocessing 3.1.3. Basically, it consists of all the generated chronological sequences of ports by vessels, using MMSI as vessel identifier. So given a sequence length, which in the proposed model is 10, all the possible chronological sequences of departure ports with this length are created for each vessel by rolling the full sequence. As aforementioned in the previous section 3.3.1 about the proposed method for clustering ports, the list of considered ports has been preprocessed according to data cleaning section 3.1.3.

The output for this task is the probability for each available port in the dataset to be the next destination for a vessel. So the port with the highest probability can be chosen as the next destination, but also other top candidates can be considered.

#### Model

The proposed approach is based on Recurrent Neural Networks (RNN) which are a rich class of dynamic models that have been used to generate sequences in domains as diverse as music, text, and motion capture data. RNNs can be trained for sequence generation by processing real data sequences one step at a time and predicting what comes next. Based on novel solutions, although several approaches and variations have been tried and implemented, the proposed architecture is based on embedding [27] and Long Short Term Memory (LSTM) [28] layers.

On the one side, in principle a large enough RNN should be sufficient to generate sequences of arbitrary complexity. In practice, however, standard RNNs are unable to store information about past inputs for very long and model long-range structure, this ‘amnesia’ makes them prone to instability when generating sequences. Then, having a long memory has a stabilising effect, because even if the network cannot make sense of its recent history, it can look further back in the past to formulate its predictions.

Long Short Term Memory (LSTM) is an RNN architecture designed to be better at storing and accessing information than standard RNNs since it uses purpose-built memory cells to store information. LSTM has recently given state-of-the-art results in a variety of sequence processing tasks, including speech and handwriting recognition. Due to the complexity of the research problem and its wide geographical scope across Europe, the main goal of including them in the proposed approach is to use its memory to learn complex and realistic sequences of ports considering the long-range structure of the European maritime traffic.

On the other side, RNNs are fuzzy in the sense that they do not use exact templates from the training data to make predictions but rather, like other neural networks,

use their internal representation to perform a high-dimensional interpolation between training examples. However, when input sequences are sparse for example one-hot encoded vectors with length in the order of thousands such as in the case of words in a dictionary or ports in this research problem, it can be very beneficial for the model to find a more suitable mapping representation of them.

Embedding is the concept of mapping or projecting an element of a space (a port, a word, a phrase, a sentence, an entity, a relation, an image, etc.) into an element of a vectorial space, frequently low dimensional, with fixed dimension. The main goal of including an embedding layer in the architecture of the proposed approach is to learn a map which will embed each port into a continuous vector space representation of all the set of ports, and aim for embedding similar ports to similar embedded space regions from the maritime traffic point of view. The embedding vectors get updated while training the neural network.

The following Table 7 shows the exact architecture of the proposed neural network. As aforementioned it contains an embedding layer right after the input, and later two consecutive LSTM layers with the first one returning the input sequence shape too. Finally, a fully connected layer connects to the output layer with as many nodes as available ports including Softmax activation function.

**Table 7:** Proposed neural network architecture for predicting the next destination. S=Input sequence length, O=Output layer shape, A=Activation function, RS=Return sequence, P=Number of ports, ReLU=Rectified linear unit, Tanh=Hyperbolic tangent.

Layer	O	A	RS
Input	S	-	-
Embedding	Sx50	-	✓
LSTM	Sx100	Tanh	✓
LSTM	100	-	-
Dense	100	ReLU	-
Output	P	Softmax	-

Finally, since predicting the destination can be understood as a discrete classification problem, the output can be formulated as a multinomial distribution which can be naturally parameterised by a Softmax function [29] at the output layer. Then the loss function proposed is the Sparse Cross Entropy [30], which is a variant from the Cross Entropy that covers a subset of use cases and the implementation is different to speed up the calculation. However, their formula (eq. 1) is the same

$$\text{Cross Entropy}(Y, \hat{Y}) = - \sum_{i=1}^N \sum_{p=1}^P Y_{ip} \cdot \log(\hat{Y}_{ip}) \quad (1)$$

where  $P$  is the number of output nodes or ports,  $Y$  the one hot encoded true port for each observation, and  $\hat{Y}$  the predicted probability distribution of ports for each observation.

## Training

In the experiments of this research task, the algorithm has been implemented on Python3 using the Deep Learning framework *keras*, as well as the modules *numpy* and *pandas* for basic mathematical operations and data processing. It is standard to use gradient descent [31] approaches to train neural networks. However, there are different variants and the chosen one for the research experiments is Adam [32]. The batch size used is 128. And the epochs have been limited by early stopping [33] monitoring the loss function and the accuracy in a validation set with a minimum improvement of 0.01 and patience of 3 epochs.

### 3.3.3 Feedforward neural networks (to forecast ETA)

This section describes the proposed model for the third research aim, which is to forecast the expected voyage duration of the voyage until the next destination, as well as the training procedure implemented.

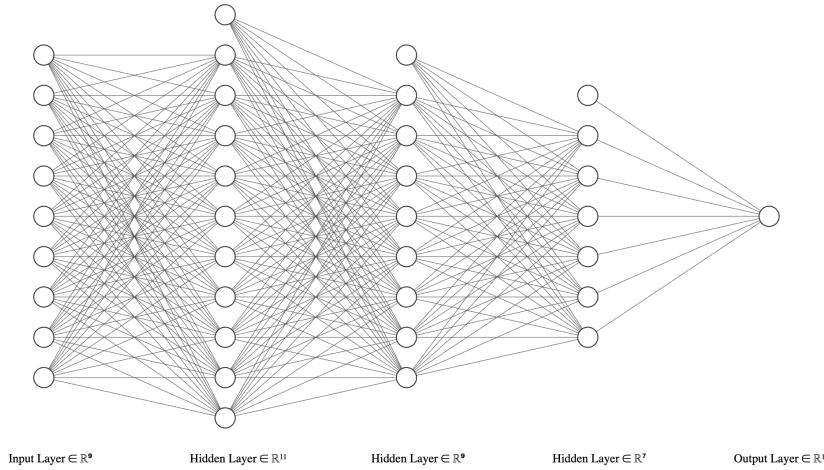
#### Input & Output

The input for this task has been previously described in the section of data preprocessing 3.1.3. Basically, it consists of all the historical voyages and their corresponding variables *voyage length*, *median velocity*, *maximum velocity*, and *ship type* in one hot encoded shape, which all of them correspond to 9 different input nodes in the proposed neural network.

The outputs for this task are two different target variables in order to experiment which option gives the best outcomes:

- On the one hand, the voyage duration in hours computed from the difference between the departure and destination dates, and in Table 6 you can find the distribution of the duration of the voyages in the dataset.
- On the other hand, the voyage duration in hours divided by the median duration between the corresponding departure and destination ports. In this case, before computing the RMSE, the output of the model has to be multiplied by the same median used when creating the target value of the corresponding voyage to undo the transformation and have voyages duration in hours scale.

Moreover, it is trivial to think that when predicting new cases in future, the destination as well as the information about the voyages, such as the voyage length or velocity, will be unknown beforehand. However, by using the proposed method to predict the next destination and the history of the voyages of each vessel as well as all the voyages between ports will enable to find good estimates of, for example, the velocity as well as the voyage length between the corresponding departure and predicted destination.



**Figure 7:** Proposed neural network architecture for forecasting ETA.

## Model

The proposed approach is based on Feedforward Neural Networks which are the most common class of Deep Learning models that have been used for a wide range of tasks. These models are trained without any sequential input and each observation input vector is mapped to its corresponding target vector. Although several approaches and variations have been tried and implemented, the proposed architecture is shown in Figure 7 and Table 8.

**Table 8:** Proposed neural network layers for forecasting ETA. I=Number of input variables, O=Output layer shape, A=Activation function, BN=Batch normalization layer, D=Dropout layer, ReLU=Rectified linear unit.

Layer	O	A	BN	D
Input	I	-	-	-
Dense	10	Sigmoid	-	-
Dense	8	Sigmoid	-	-
Dense	6	Sigmoid	-	-
Output	1	ReLU	-	-

Since forecasting the voyages ETA can be understood as a non-negative regression problem, the output can be formulated as a single node which can be naturally parameterised by a ReLU function [34] at the output layer. Then the loss function proposed is the popular Mean Square Error [35] which can have the same units than the target variable when measuring the error by applying square root (eq. 2) to it

$$\text{RMSE}(Y, \hat{Y}) = \frac{\sum_{i=1}^N (Y_i - \hat{Y}_i)^2}{N} \quad (2)$$

where  $N$  is the number of observations or voyages,  $Y$  the true duration and  $\hat{Y}$  the predicted duration. Moreover, in the experiments a variant of this loss function is considered by adding a regularization term

$$\text{RMSE}(Y, \hat{Y}) + \lambda \sum_{m=1}^M \beta_m^2 \quad (3)$$

where  $M$  is the number of parameters in the neural network architecture,  $\beta$  the fitted parameters, and  $\lambda$  the regularization term.

## Training

In the experiments of this research task, the same training protocol used in the previous task [3.3.2](#) has been used but with batch size 64.

## 4 Results

In this section, the main experiments done will be described as well as their results and performance analysis for all the proposed methods of each research aim. Firstly, it will be necessary to understand the list of variations and parameters involved that have been used to experiment with each method, and then their respective results will be shown and compared.

During this research, a wide range of experiments and trials have been developed and implemented by trying different configurations and parameters for the proposed approaches. In this section, we are going to define and explain concrete experiments which are useful to understand behaviours and make comparisons.

Across the results of the experiments, you will find 3 groups which represent 3 different ports aggregations by closeness in the initial dataset: <1km, <15km, and <25km. The details of how this aggregation is done are extensively explained in the previous section about data preprocessing 3.1.3, but the concept is simply that, for example, in <15km group all the ports at less than 15km of geographical distance are aggregated into the same group. It is trivial to think that the lower the distance threshold the more ports there will be after the aggregation, but more difficult will be for the models to predict, for example, the next destination. The effect of this trade-off is well studied and reasoned in this section.

The training methods and configurations have been explained in the previous section 3.3 separately for each aim. However, across the following results tables, you will find performance metrics values and outcomes for both training and validation sets in the format of *training/validation* sets values. This split is based on 75/25% of input data respectively, and concretely for the second aim of predicting the next destination of vessels, the split is not done randomly but applying it chronologically to the sequences of departure ports for each vessel.

### 4.1 Clustering ports based on maritime traffic

This section covers the experiments and results of the first research aim. However, this task is also related to other research aims so more results about clustering ports will be shown in the following sections 4.2 and 4.3.

First of all, according to the proposed method for clustering ports based on maritime traffic explained in section 3.3.1, a concrete number of desired clusters has to be chosen and the selected value for all the research experiments has been 3. Then, after computing the single value decomposition of the Laplacian matrix, also a concrete number of eigenvectors has to be chosen to apply the clustering KMeans:

- For the <1km ports aggregation, 1000 eigenvectors are selected out of 3,189 possible ones equivalent to the number of ports after this aggregation.
- For the <15km ports aggregation, 400 eigenvectors are selected out of 1,616 possible ones equivalent to the number of ports after this aggregation.
- For the <25km ports aggregation, 400 eigenvectors are selected out of 1,012 possible ones equivalent to the number of ports after this aggregation.

Secondly, after applying the proposed method to each different ports closeness aggregation case, as we can see in Table 9 with the maritime voyages data grouped by the resulting departure ports clusters, there is a cluster called B containing more unique departures and destinations, and then the rest of clusters A and C have way less. This first analysis gives us an orientation about the maritime traffic from/to of the resulting clusters. It is important to note that logically, the lower the threshold distance the more ports in the resulting data.

**Table 9:** Number of unique departures and destinations in maritime voyages data grouped by departure ports clusters and by ports closeness aggregation levels.

Cluster of departure	# of unique departures			# of unique destinations		
	<1km	<15km	<25km	<1km	<15km	<25km
A	522	167	119	1154	373	349
B	2112	1232	718	2855	1554	980
C	466	190	160	1049	424	422

Thirdly, the Table 10 shows the comparison of the maritime connectivity between the resulting clusters. More specifically, in both rows and columns, we can find the clusters of the departure ports in the voyages data, and the table content corresponds to the percentage of shared destination ports between clusters and ports closeness aggregation levels. As we can see, the proposed spectral clustering method has worked because it has identified 3 separate groups of ports which have different destination ports across Europe. For example, for the aggregation case <15km, only 22% of voyages destinations from cluster B ports are shared with A, only 26% of voyages destinations from cluster B ports are shared with C, and only 34% of voyages destinations from cluster C ports are shared with A.

**Table 10:** Percentages of shared destination ports by resulting clusters of departure ports.

Cluster of departure	<1km			<15km			<25km		
	A	B	C	A	B	C	A	B	C
A	100%	84%	44%	100%	95%	39%	100%	97%	56%
B	34%	100%	32%	22%	100%	26%	34%	100%	42%
C	49%	88%	100%	34%	96%	100%	46%	98%	100%

Then, as we can see in Table 11 it is quite interesting to see that, in the ports aggregation case <15km, a higher rate of voyages departing from cluster A ports correspond to ships of type *Missing*, a higher rate of voyages departing from cluster B ports correspond to ships of type *Passenger*, and a higher rate of voyages departing from cluster C ports correspond to ships of type *Cargo*. Although some ship types have similar rates across clusters.

**Table 11:** Distribution of ship types by departure ports clusters in the ports aggregation case <15km.

Cluster of departure	Ship type				
	Cargo	Tanker	Passenger	Other	Missing
A	34.6%	9.7%	11.9%	9.2%	34.5%
B	33.3%	15.3%	30.5%	9.9%	10.9%
C	54.7%	10.7%	12.7%	9.1%	12.9%

Finally, as we can see in Table 12, there is a clear proof about the efficacy of the proposed method to cluster ports based on maritime connectivity. In the ports aggregation case <15km, the resulting cluster A contains most of the ports with a low amount of voyages departing from them, cluster B contains most of the ports with a high amount of voyages departing from them, and cluster C has a bit more homogeneous distribution containing a relatively high percentage of ports with a high amount of voyages departing from them but also many other ports with less maritime traffic.

**Table 12:** Distribution of departure ports groups created by maritime traffic levels for each departure ports cluster in the ports aggregation case <15km. The maritime traffic levels are created based on the sum of voyages from each port, so 0-25% is the 25% of ports with lower traffic departing from them, and 75-100% is the 25% of ports with higher traffic.

Cluster of departure	Groups of ports by traffic levels			
	0-25%	25-50%	50-75%	75-100%
A	58.0%	32.4%	9.5%	0.0%
B	0.0%	0.7%	3.5%	95.8%
C	10.2%	7.8%	16.9%	64.9%

## 4.2 Predicting next destination of vessels

This section covers the experiments and results of the proposed method in section 3.3.2 for the second research aim. However, this task is also related to the third research aim of forecasting ETA in the sense that when predicting new cases in future, the next vessel destination will be unknown as well as the necessary input voyages variables for forecasting the ETA. But based on the destination predicted by the proposed method here, it will be possible to obtain good estimates of these variables from the voyages history of each vessel and all the voyages between the corresponding ports.

Firstly, since an extensive search for optimal optimization hyperparameters would have been prohibitively expensive and time consuming given the available GPU

resources, a limited list of experiments have been tried for the proposed RNN architecture which can be found in the Table 13. As you can see, the experiments configurations are done by playing with the resulting clusters of voyages departure ports, the ports aggregation by closeness, the length of the input sequence, the number of minimum different ports in the input sequences, and the embedding layer output size.

**Table 13:** List of experiments for the proposed RNN to predict the next destination port.

ID	Cluster of departure	Agg. of ports	Unique ports	Sequence length	Sequence # of ports	Embed layer size
1	All	<25km	1012	10	1	50
2	All	<25km	1012	10	1	100
3	All	<25km	1012	5	1	50
4	All	<25km	1012	10	3	50
5	B	<25km	1012	10	1	50
6	All	<15km	1616	10	1	50
7	All	<1km	3189	10	1	50

Secondly, in Table 14 we can see the performance results after implementing each proposed experiment. These are the respective conclusions:

- **Experiment 1.** After completing the first experiment, we conclude that the proposed method is clearly working well for solving the research aim, and it achieves very high accuracy in both training and validation sets. By considering as destination only the top 1 predicted port, which means the port with the highest predicted probability of being the next destination, the model has accuracies around 90%. Furthermore, by using the top 2 and top 3 predicted ports as candidate next destinations, the model is able to achieve accuracies up to 95%.
- **Experiment 2.** In this experiment, the embedding layer output size has been increased in order to check if the embedding of the input sequences could be done with a different architecture that improves the model performance, but the model results have not improved.
- **Experiment 3.** In this experiment, the input sequence length has been reduced from 10 to 5 in order to test if the performance can be kept with shorter sequences. Remember that short sequences are important for different factors, on the one side, less historical information about the voyages of each vessel is needed, and on the other side, more training sequences can be generated with the same data. The accuracy results of this experiment are very positive because the performance has not dropped so much but the sequence length has been reduced by half.

- **Experiment 4.** In this experiment, all the input sequences with 1 or 2 unique ports have been filtered out in order to test if the next destination for sequences with very different ports can be learnt better. However, the results are worst with an accuracy around 55%, which could be reasonable if the accuracy for the sequences with very different ports were also at this low level in the first experiment, but it is not the case as it will be described when commenting the Table 15.
- **Experiment 5.** In this experiment only the voyages sequences with the last departure port in cluster B are considered, using the clustering results of previous experiments section 4.1. However, the results are not improved which make sense since not too much sequences are filtered out.
- **Experiment 6.** In this experiment, the distance threshold for the ports aggregation by closeness has been lowered to 15km. This is a very relevant experiment because, keeping in mind that the scope of this research is long voyages across all European waters, if the model still works well with this configuration it means that the geographical accuracy of the system is drastically increased. According to its results, the performance keeps high achieving almost 80% and 75% top 1 accuracy in the training and validation sets respectively, but for the top 2 and top 3, the performance results are even better achieving accuracies around 90%. In fact, for the rest of the analysis in this section, this is the model configuration selected.
- **Experiment 7.** In this experiment, again, the distance threshold for the ports aggregation by closeness has been lowered to 1km. In this case, the number of ports drastically increase and the model performance decreases. However, considering the top 3 accuracy, its performance is quite high achieving accuracies above 70% for both training and validation sets.

**Table 14:** Results of experiments for the proposed RNN to predict the next destination port. Top X means that the X ports with highest predicted probability have been used as next destination candidates. The format of the content is *training/validation* sets values.

ID	# of input data sequences	Accuracy incl. top 1	Accuracy incl. top 2	Accuracy incl. top 3
1	947,018	90.2/87.2%	94.6/92.1%	95.9/93.7%
2	947,018	89.9/87.6%	94.6/92.2%	96.0/93.7%
3	1,031,246	87.8/86.1%	92.6/90.3%	94.2/93.9%
4	146,287	57.4/55.1%	72.4/70.2%	79.9/76.8%
5	942,584	89.7/87.1%	94.4/92.5%	95.8/93.5%
6	947,018	78.2/73.8%	87.1/83.2%	90.6/86.9%
7	947,018	55.3/53.1%	68.6/66.3%	74.6/72.3%

Thirdly, in Table 15 we can find an analysis of the accuracy of the model in experiment 6 by sequences grouped by the number of unique ports in them. Logically, the more variance in the input sequence of departure ports of a vessel, the harder is to predict the next destination. This behaviour can be found in the results table since the next destination for sequences with only 1 unique port have perfect accuracy (which sounds trivial and easy but it is indeed a great validation check for a model with such a geographical scope across all Europe), even sequences with only 2 unique ports have very high accuracies, but then lower 50% top 1 accuracy is reached in sequences of 5 unique ports, although at least 60% top 3 accuracy is kept whatever is the number of unique ports in the sequences.

**Table 15:** Results of experiment 6 by number of unique input sequence ports. The format of the content is *training/validation* sets values.

# unique ports in sequences	Distribution	Accuracy incl. top 1	Accuracy incl. top 2	Accuracy incl. top 3
1	36.8/36.9%	98.1/97.9%	98.7/98.5%	98.9/98.7%
2	24.1/23.1%	82.9/77.3%	94.7/92.5%	96.6/95.0%
3	14.0/14.2%	67.6/62.5%	83.7/78.6%	90.1/86.1%
4	9.1/9.1%	59.3/52.4%	75.1/67.6%	82.6/76.0%
5	6.3/6.4%	51.2/44.1%	66.4/58.1%	75.2/66.1%
6	4.2/4.4%	44.1/36.7%	58.4/49.1%	66.8/56.9%
7	2.8/3.1%	38.8/28.6%	52.4/40.2%	60.4/47.3%
8	1.6/1.9%	33.1/24.5%	46.1/34.5%	54.2/40.8%
9	1.1/1.0%	47.2/36.3%	57.2/44.8%	63.8/50.9%

Then, in Table 16 we can find another interesting analysis related to the experiments for the proposed method to cluster ports based on maritime traffic in section 3.3.1. Here the accuracies of the model in experiment 6 are shown by voyages departure clusters. Surprisingly, the model behaves very reasonably since the clusters A and C which corresponds to the clusters of ports with less maritime traffic connectivity across Europe are the ones for which it is harder to predict the next destination.

**Table 16:** Results of experiment 6 by clusters of departure ports. The format of the content is *training/validation* sets values.

Cluster of departure	Accuracy incl. top 1	Accuracy incl. top 2	Accuracy incl. top 3
A	54.4/51.3%	71.4/68.1%	79.1/75.8%
B	78.7/74.3%	87.5/83.5%	90.7/87.1%
C	61.4/57.9%	75.6/72.1%	80.5/77.3%

Also, as we can see in Table 17, it is quite interesting to see that, in the ports aggregation case <15km, almost perfect accuracies are achieved by the model of experiment 6 for those voyages that correspond to ships of type *Passenger*, lower accuracies are achieved for those voyages that correspond to ships of type *Cargo* and *Tanker*, and very bad accuracy performance is found for those voyages that correspond to ships of type *Missing*.

**Table 17:** Results of experiment 6 by ship type. The format of the content is *training/validation* sets values.

Ship type	Distribution	Accuracy incl. top 1	Accuracy incl. top 2	Accuracy incl. top 3
<b>Passenger</b>	35.5/34.5%	94.8/91.4%	99.1/97.4%	99.5/98.3%
<b>Cargo</b>	30.8/32.4%	62.9/58.1%	76.1/70.3%	82.2/76.4%
<b>Tanker</b>	14.8/15.2%	69.3/65.5%	78.9/75.1%	83.7/80.0%
<b>Other</b>	10.0/10.5%	77.0/10.5%	87.1/84.1%	91.1/88.1%
<b>Missing</b>	8.9/7.4%	8.0/7.6%	9.2/8.8%	9.5/9.2%

Finally, a crucial analysis for the experiment 3 results, especially for the third research aim of forecasting the ETA, is done in Table 18. The average distance between the top 3 predicted ports as next destination candidates and the true target destination harbour has been computed for each sequence, and then the quantiles of it are shown in this table. It is important to understand that even if the destination has been well predicted as top 1 option, this analysis includes also the top 2 and top 3 candidates, so that is why for example the minimum value is 7.6km instead of 0km. However, that is exactly the target to study, the closeness between the top 3 destination candidates to the true destination. As we can see, considering that the threshold for the ports aggregation by closeness is <15km in this experiment, the model is performing very well in this analysis since in 50% of the voyages, the average closeness to the true destination is only 60km.

**Table 18:** For experiment 6, the average closeness in km of the top 3 predicted destination ports to the true destination harbour.

Quantile	Closeness to true port incl. top 3 (km)
0%	7.6
25%	37.8
50%	60.7
75%	161.8

### 4.3 Forecasting ETA of voyages until next destination

This section covers the experiments and results of the proposed method in section 3.3.3 for the third research aim of forecasting the voyages ETA. Concretely, all the experiments tried for this proposed model are done with the distance threshold <15km for the ports aggregation by closeness.

As aforementioned in previous sections, this task is also related to the second research aim of predicting next destination port in the sense that when predicting new cases in future, the next vessel destination will be unknown as well as the necessary input voyages variables for forecasting the ETA. But based on the destination predicted by the proposed method of this research, it will be possible to obtain good estimates of these variables from the voyages history of each vessel and all the voyages between the corresponding ports.

In the experiments, several configurations and combinations of parameters have been tried, and some of them are not included in the proposed method section 3.3.3. All these experimented model variants have been done by playing with the following well known Deep Learning concepts or techniques: hidden layers and number of nodes for the neural network architecture, activation function of the nodes in the hidden layers, regularization in the loss function, batch normalization layers, autoencoder or glass shape architecture, pre-training of the network weights, and dropout layers. Moreover, as aforementioned, the output target variables considered in these experiments are two different ones:

- Firstly, the voyage duration in hours computed from the difference between the departure and destination dates, and in Table 6 you can find the distribution of the duration of the voyages in the dataset.
- Secondly, the voyage duration in hours divided by the median duration between the corresponding departure and destination ports.

**Table 19:** List of experiments for forecasting ETA and their results. The format of the content column *RMSE* is *training/validation* sets values.

ID	Target variable	Changes respect to proposed method	RMSE
1	Duration	None	1.6/1.6
2	Duration	Hidden layers of network architecture [8,4,2]	1.7/1.7
3	Duration	Hidden layers of network architecture [8,4]	9.8/9.9
4	Duration	Activation function of hidden layers ReLU	9.1/9.4
5	Duration	Regularization 0.01 & Batch normalization	8.8/9.1
6	Duration	Autoencoder & Pre-training & Dropout 0.1	1.8/1.8
7	Duration	Only voyages $9h \leq \text{Duration}$	5.2/5.3
8	Duration/Median	None	6.5/6.7
9	Duration/Median	Only voyages $0.5 \leq \text{Duration}/\text{Median} \leq 2$	1.8/1.8

Then, in the Table 19 we can find the list of experiments tried for the proposed method and their performance results. These are the respective conclusions:

- **Experiment 1.** After completing the first experiment, we conclude that the proposed method performs well for solving the research aim and has a very reasonable RMSE which in practice can be understood in the same units as the target. So in both training and validation sets, the model has a mean error of 1.6 hours overall. By looking at Table 6, this mean value looks very reasonable and will be more aggregated by voyages duration in the following analysis.
- **Experiment 2.** In this experiment, the network architecture has been changed by reducing the number of hidden nodes at each hidden layer to analyze how the model can perform with a simpler architecture. As we can see from the table, in this experiment the performance has not been improved but it has not decreased much so it is also a positive outcome.
- **Experiment 3.** In this experiment, again the network architecture has been changed by reducing the number of hidden layers to only 2 hidden layers with 8 and 4 hidden nodes respectively. As we can see in the table, in this experiment the performance has drastically decreased which means that actually, the network architecture of experiment 1 is not a too complex choice at all for the research problem and data.
- **Experiment 4.** In this experiment, the activation function of the hidden layers has been changed from Sigmoid to ReLU. As we can see in the table, in this experiment the performance has decreased too with respect to experiment 1.
- **Experiment 5.** In this experiment, the network architecture has been changed by adding batch normalization layers into the hidden layers of experiment 1. As we can see in the table, in this experiment the performance has drastically decreased too with respect to experiment 1.
- **Experiment 6.** In this experiment, the network architecture has been completely reformulated. By using the Deep Learning technique or concept of autoencoder, the network architecture has been switched to the network shown in the Table 20 and Figure 8. As you can see, a hidden layer with 8 nodes is used, followed by another hidden layer with the same amount of nodes as the input layer, and then it is fully connected to the output layer node. Then, the 2 hidden layers are pre-trained trying to fit as output the same inputs which is the basic idea of autoencoders. Moreover, dropout layers have been added on top of the aforementioned hidden layers. Finally, the performance has been very surprising since it is very similar to the best model so far in experiment 1 but with a way simpler network architecture.
- **Experiment 7.** In this experiment, the voyages with less than 9 hours of duration have been filtered out to test if the duration of long voyages can be learnt better. However, the results are worst with an RMSE around 5, which

could be reasonable if the RMSE for the long voyages were also at this level in the first experiment, but it is not the case as it will be described when commenting the Table 21.

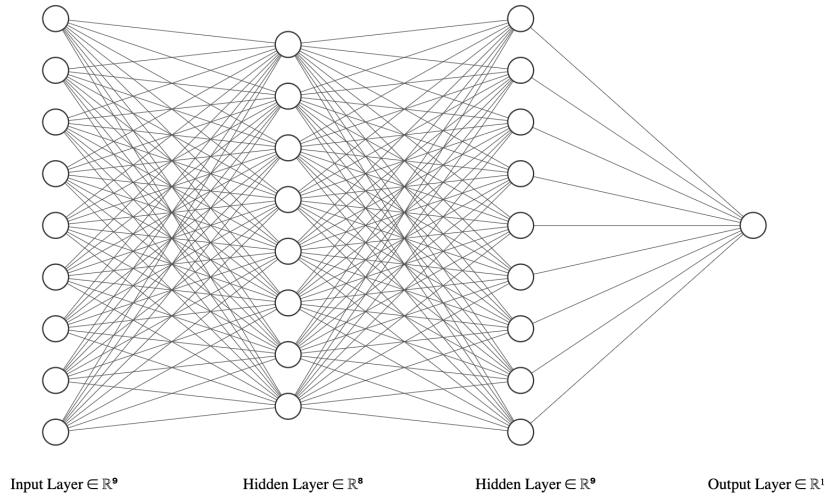
- **Experiment 8.** In this experiment, the target variable during the training has been changed to the second option which is the voyage duration in hours divided by the median duration between the corresponding departure and destination ports. The hypothesis is that it could be easier to forecast the duration of a voyage compared to the respective median, however, the RMSE has drastically increased.
- **Experiment 9.** In this experiment, again the target variable during the training has been changed to the second option but also the voyages with less duration than half of the median or more than twice have been filtered out to test if the duration of more common or not outlier voyages can be learnt better using this target variable. As we can see in the table, unlike the previous experiment 8, a very similar performance compared to experiment 1 is achieved.

**Table 20:** Proposed network architecture of experiment 6 for forecasting ETA. I=Number of input variables, O=Output layer shape, A=Activation function, BN=Batch normalization layer, D=Dropout layer, ReLU=Rectified linear unit.

Layer	O	A	BN	D
Input	I	-	-	-
Autoencoder	8	Sigmoid	-	✓
Dense	I	ReLU	-	-
Output	1	ReLU	-	-

Moreover, as we can see in Table 21, an analysis of the RMSE by duration intervals has been done for the proposed method. This is in fact what we are most interested in evaluating since the RMSE over all the data is not informative enough due to the wide different voyages in the dataset. As we can see in the results, voyages of less than 3 hours have an RMSE below half an hour, voyages with a duration between 3 and 9 hours have an RMSE around 1 hour, voyages with a duration between 9 and approximately 1 day have an RMSE around between 2 and 3 hours, and finally, voyages of more than 1 day of duration have an RMSE around 5 hours. These performance outcomes are significant considering the wide geographical scope of the data all over Europe, and the potential implications of this high precision are discussed in the following section 5.

Finally, as we can see in Table 22, it is quite interesting to see that again the best performance is achieved by the model for those voyages that correspond to ships of type *Passenger*, achieving a low RMSE of half an hour for the whole data, and the worst results are found for those voyages that correspond to ships of type *Cargo* and *Tanker*.



**Figure 8:** Network architecture of experiment 6 for forecasting ETA.

**Table 21:** Results of experiment 1 by voyages duration intervals. The format of the content column *RMSE* is *training/validation* sets values.

Quantile group	Duration interval (h)	RMSE
0-25%	0.1-0.4	0.2/0.3
25-50%	0.4-1.2	0.2/0.2
50-60%	1.2-1.8	0.3/0.3
60-70%	1.8-2.9	0.4/0.5
70-80%	2.9-4.7	0.8/0.9
80-90%	4.7-9.3	1.3/1.4
90-95%	9.3-16.5	2.0/1.9
95-97.5%	16.5-27.9	2.9/2.8
97.5-99%	27.9-47.0	5.2/4.6

**Table 22:** Results of experiment 1 by ship type. The format of the content is *training/validation* sets values.

Ship type	Distribution	RMSE
<b>Passenger</b>	30.2/31.1%	0.5/0.6
<b>Tanker</b>	15.1/15.6%	2.1/2.2
<b>Cargo</b>	33.1/34.5%	2.2/2.1
<b>Other</b>	9.7/10.4%	1.3/1.4
<b>Missing</b>	11.9/8.3%	1.6/1.3

## 5 Discussion

This section aims to reflect and discuss the results of the present research in light of the findings reported in previous sections. Firstly it will cover the success level of the proposed methods and their research experiments in order to answer the formulated research questions and hypothesis. Secondly, the potential impact of the research outcomes will be commented as well as a restatement of the main findings and experimental conclusions. Finally, a reflection is given about the hardest parts of the project development or the most difficult steps in the whole process, followed by what would be done differently by the author if the project started again.

In this research, as aforementioned in the introduction section 1, the problem in question arises from the issues and challenges in the maritime logistics field, since nowadays globalization has scaled across the planet and the sea is one of the most used means of transport for passengers, goods, products, etc. but the maritime transportation is not optimized internationally and several maritime agents are struggling with inaccurate ETA and ETD predictions which create inefficiencies in the whole process, such as delays due to, for example, unexpected ports full capacity reached. During the research process, it has been possible to realize the lack of reliable and robust data systems for the real-time awareness situation because, for example, most of the fields in the AIS messages have poor quality and are not sufficient for modelling or being used in AI systems.

The principal aim of this thesis was to propose and implement a predictive AI based system that predicts the next destination and expected voyage duration for vessels across Europe as a tool that can be used in production to improve the aforementioned maritime logistics inefficiencies, enhancing the maritime situational awareness and also the communication across ports for better task planning and ports calls. Once the proposed methods have been introduced and their respective results and predictive performance have been exposed, we conclude that this research successfully provides significant advances in the field with a pure AI based solution. Concretely, one of the most important accomplished concepts in the definition of the research problem and aims is the wide geographical scope across European waters and ports. This has affected all project stages, especially the selection of the methods and techniques to use as well as the data preprocessing applied to the research data available such as the ports aggregation by closeness described in section 3.1.3.

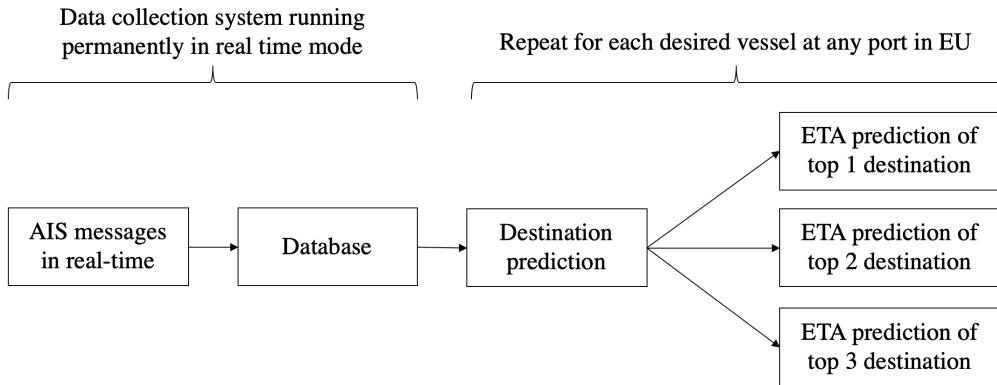
Then, according to the stated research questions, hypothesis and aims, we can argue now that the aforementioned issues and problems in the maritime logistics field can be tackled from a perspective based on Artificial Intelligence, and more specifically for each research aim:

- For clustering ports based on maritime traffic, the proposed method and its experimental results establish a way to understand better the European ports structure based on maritime connectivity which can lead to important information for solving and have better interpretability of the research problem. A mathematical method to cluster the ports has been successfully developed understanding the maritime connectivity across Europe as a mathematical

graph, where the nodes or vertices are the ports and the edges the number of voyages between ports. The hypothesis for this first research aim is corroborated because it is true that several Machine Learning techniques and methods have been applied for graph clustering and specifically Spectral Graph Clustering, which is the basis of the proposed method, has been a successful selected approach with very informative and useful results as described in section 4. Furthermore, the resulting clustering has been used in the experiments of the rest of aims giving very interesting description and performance by clusters, since the method has detected a couple of small clusters of ports from which is harder to predict the next destination and ETA.

- For predicting the next destination of vessels when berthed in a port, it is corroborated after doing the research literature review that this topic has not been widely studied especially with a full European scope so there has been open room for innovation and discovering a suitable application of Deep Learning techniques for this aim. As proved in the results section 4.2, the proposed method in section 3.3.2 based on RNN with LSTM and embedding layers, using as input the sequence of previous departure ports to predict the following destination of vessels, is an innovative application of popular neural network techniques, which have been applied in other areas such as text generation, into the maritime logistics field. After the research experiments, the performance reached has been significant providing results with very high performance in terms of accuracy and robustness.
- For forecasting the expected voyage duration of voyages until the next destination, a well-integrated method with the destination prediction model has been proposed and implemented successfully. In this case, the proposed method is not an innovative model since its structure is based on well documented and studied Feedforward Neural Networks, but its application and performance reached for the research aim in question has been very positive, not only for the predictive results considering the wide geographical scope across European waters but also because these results are achieved without using too much input information from the data engineering point of view. So the model proposed should be robust respect to the data gathering and ingestion difficulties of this real-time big data problem. One of the most interesting and insightful analysis can be found in Table 21, where an estimated of the mean error in the predicted duration is given by different voyages duration intervals. These results present very low errors and are a big step for solving the aforementioned maritime issues, especially for long voyages.

Overall, the proposed methods for all the research aims can work as modules of a whole integrated AI based system, represented in Figure 9, that complement each other to improve the situational awareness and provide real-time forecasts of the whole maritime picture in European waters. This is because predicting the next destination port and ETA are directly connected since it is trivial to think that, for example, when informing that a vessel V berthed now at port A is heading next to



**Figure 9:** General representation of the proposed methods in the present research as a full system for improving the maritime logistics.

port B, the expected time T until arrival to the destination harbour is needed for having complete information of the voyage. And based on the proposed methods, the same ETA model can be used to predict the duration of several top candidate destinations. In conclusion, achieving a good predictive model at the European level for the next destination and ETA of vessels berthed at any port has been significant for optimizing maritime logistics internationally and not only between ports in the same geographical area. All this situational information can be now achieved real-time using the research outcomes which pave the way for efficient port call planning, maritime operations ports, determining and optimizing the fleet utilization, reducing the shipping emissions, and solving other challenges that will be more extensively described.

As aforementioned, efficiency at ports is everything and it strongly correlates with the level of competitiveness which is one of the key drivers for the entire maritime transport business. That is why achieving the research aims and corroborating the hypothesis of this thesis, a better driving or step forward is achieved for the digital transformation in maritime logistics, since the situational awareness can be drastically improved across Europe with help of these AI insights, which would enable better planning, transparent and real-time status sharing of European ports operations, improved efficiency, sustainability, competitiveness, and lower costs for all maritime actors, including port authorities, terminal operators, cargo owners and many more.

Furthermore, the proposed methods in this project have a relevant advantage concerning the time frame of the models' application into the industry, since with these models the prediction of the next vessels' destination and its expected ETA can be done before departure, which means that there is no need to wait until the next voyage has started and its first AIS messages from the vessel are sent. What is more, it enables to run the prediction right after the vessel arrives at a port in Europe and this is a really important aspect since especially industrial vessels tend to stay several hours or even days in the ports, so all this time is gained in advanced achieving a significant improvement.

So it is expected that this research contributes to the application of AI into the

maritime logistics field, which is quite recent in history and has not been extensively studied in the literature since the international maritime transportation has scaled up in the last decades. Especially, it is important to point out that this field is not one of the most explored areas in Artificial Intelligence neither, and that is why this research has a very innovative component and could be directly tested in the industry as a solution tool for the aforementioned issues.

Finally, it is interesting to discuss too what has been the bottlenecks or main difficulties during the whole research process, as well as what would be implemented differently or more optimally if the project started again. On the one side, it was hard during the literature review to find articles and papers working on similar aims especially with such a wide geographical scope as the entire European waters. Many research of AI applied to maritime logistics are about object recognition in ports to detect, for example, type of vessels. However, other papers and articles about applications of AI into areas such as social networks graphs clustering or text generation using RNN served as inspiration for the proposed methods in the present research. On the other side, several issues with the data were tried until the right techniques or preprocessing were found for creating the inputs and output targets of all the proposed models. Then, about what would be done differently if the project started again, probably it would have been interesting from the beginning to be aware that some extra data such as weather information would be difficult to obtain for the whole historical voyages across Europe. In fact, it was planned to study how weather data could improve the ETA forecasting model but since this data collection was started too late, several problems arose to get good quality historical weather information for all European voyages during the full research dataset period.

## 5.1 Industrial and commercial benefits

Digitalization and automation are rapidly changing the maritime outlook to reduce fuel and energy consumption as well as to improve the overall efficiency. Then, profits can be increased when digital technologies are integrated into the entire operational maritime system. The companies and organizations that are establishing greater digital competence are gaining a significant advantage in the ever-changing landscape of regulations and performance requirements. As we can see in results section 4, many real industrial cases can be successful after applying the proposed methods in this thesis due to the high predictive performance and characteristics of the designed AI based system. Mainly, that is why once we can predict the next destination of vessels and their expected voyage duration across Europe with the high levels of predictive performance provided by the methods in this research, then several improvements can be done in the overall European maritime transportation process by offering the right service and real-time information to the involved maritime agents. Concretely, a crucial characteristic of the proposed system is that it enables to predict hours or even days in advance before the next departure, so directly once a vessel arrives at a port, then the models can be run already so the time frame horizon expands to plan operations ahead.

So the predictive information that the outputs of the proposed methods provide

can have several benefits for different agents involved in the whole process of maritime logistics and operational planning. Moreover, since we have seen that nowadays vessels can come from everywhere at any port, the successful wide international scope of this research covers this difficulty or complexity pretty well. Now, examples of concrete actions that the main maritime agents can take are provided:

- **Ship operators.** Traditionally ships are served in the order they arrive at ports, which leads to queues and waiting, effectively causing loss of revenue and unnecessary emissions. Then, by using the researched system the ship operators can, for example, order to accelerate or reduce the velocity during the voyage in case that they are communicated that there is or not space in the port. This can reduce turnaround times, save cost on fuel by slow steam, reduce emissions, and increase the utilization of their fleet by making sure their fleet spends less time waiting and arrives just in time. Of course, a better communication channel with all European maritime actors should be needed to stream align the operational planning and this point will be described later.
- **Ports.** Currently, most of the ports do not have accurate and updated arrivals and ETA estimates, so with more accurate information such as the outcomes of the present research they can take actions to improve their logistics processes. On the one side, they can spend less time waiting and ensure that vessels arrive just in time by planning the best possible cargo flow, ensuring capacity limits utilization, preparing smarter voyage planning and optimization, reducing emissions and fuel costs, and having a better overview and connection with other port actors. On the other side, they can maximize the use of existing port capacity by establishing a true real-time situational awareness for the port community, optimizing every port call, having a more efficient use of the port assets, and optimizing resource needs and plan operations more efficiently in terms of personnel, equipment, warehouse, etc. in order to avoid wasting resources like personnel waiting a full day in the port just because of a vessel delay. So overall, ports could maximize the throughput of the port physical infrastructure. And other port authorities and communities can have informative digital services which improve the situational awareness and thus the efficiency of every port call.

Moreover, as aforementioned one of the main points to tackle in the maritime logistics field is the lack of good and streamlined systems for optimization and fluent communication between international ports, so in the last decades, several companies have started to provide software and services for it. This thesis is a research and development project financed by Awake.AI, which is a company based in Finland that provides a top all-in-one solution for real-time collaboration and decision making between all port operations across Europe. So far, they have created a platform aiming for better collaboration and planning between all maritime actors which has different product modules. However, the modules related to the same aims of this research are based on developed approaches that use real-time AIS messages during the vessels voyage, whose essential idea is to consider the current coordinates to

check if the vessel is arriving at one of the predefined and known voyage trajectories. This is a key point for the company related to the outcomes of the present research because now by using the proposed methods they can provide the service of these product modules way earlier in time, even before the voyages departure. In fact, right after a vessel has arrived at any port in Europe, its next destination and its ETA can be predicted even if the vessels have to wait in the port for hours or days. This is going to drastically change the time frame horizons of their AI based insights and how their tool is used by their customers.

## 5.2 Future work

This section covers the potential future tasks that can be developed to advance in the present research problems and the proposed methods. Several directions for future work suggest themselves and here is a list of some of them.

- Study and add new additional input information for the models such as the proposed method for forecasting ETA. New features could include weather data at, for example, the moment of departure or even during the voyages, the mean historical speed of the vessel and other input metrics instead of only during the corresponding voyage, etc.
- Train models for specific data groups such as ship types or ports clusters.
- Try out other neural network architectures for both the destination prediction and its ETA forecast. Since in AI and Machine Learning algorithms finding the optimal parameters for a model is the key and the most time-consuming point during the training process, then an important thesis extension would be to play with several configurations and parametrizations that improve the predictive performance.
- Integrate more the resulting ports clustering into the models for predicting the next destination and forecasting ETA.
- It has not been studied in the experiments but due to the nature of the proposed RNN for predicting next destinations, it is possible to predict and generate sequences of more than 1 destination. These type of models are very used in the field of text generation, so the output predicted for an input sequence can be used as the next input for predicting the following sequence component and so on. As you can imagine, if this worked it would open a completely new room in terms of predictive capabilities for the maritime logistics since the time frame horizons for optimization would be drastically improved.
- In terms of application and deployment of research outcomes into the industry, it would be a natural next step for Awake.AI, the company that finances the project, to deploy the proposed method into production and test out its performance. However, many challenges and difficulties can arise when deploying the models since these are based on data sources shared across

Europe in real-time from thousands of vessels, so managing this real-time big data problem can present unexpected difficulties and impediments. Actually, this data engineering part is already well covered by them so the deployment could be straightforward.

- Finally, it would be very interesting to analyse the business impact of the system. In other words, researching how applying these proposed methods can impact the profit of different maritime agents and society in economic terms. After all, the new era of AI is beginning and it will create a strong impact in the maritime logistics field too.

## 6 Conclusion

This research project has investigated the possibilities for automatic AI based insights that can help to solve issues and inefficiencies in the maritime logistics field. Due to the lack of voyages arrival information and good estimates of the expected arrival times across ports, several maritime agents are facing severe problems for their ports calls planning, resources operations, reducing emissions and save cost, etc. Therefore, a method has been proposed for the research aims (i) clustering ports across Europe based on their maritime traffic connectivity, (ii) predicting the next destination of vessels, and (iii) forecasting the expected voyage duration for voyages. These AI based models can be used as a whole or full system that enables different maritime agents such as ports operators or ships responsible to take actions that improve all the processes based on this information.

Several experiments and analysis have been performed using maritime traffic data across Europe during 2020 provided by Awake.AI, the Finnish company that finances this research. This data is based on public AIS messages which are informative messages that vessels send to stations during their voyages about their position, status and other aforementioned variables about the voyage.

Overall, we can conclude that the proposed methods and their implementation have been a success in terms of predictive performance at the wide scale of the entire European waters. Moreover, we have measured and experienced empirically how useful can be to use the proposed method into a validation set, achieving (i) the detection of 3 ports clusters with very different profiles based on maritime traffic, (ii) an accuracy over 90% for predicting the next destination in some experiments and configurations, and (iii) a low mean error for forecasting ETA in relation to the duration of the voyages. So the experiments of this research thesis demonstrate quantitatively the ability of the proposed system to solve the research aims. However, it could be enhanced by, for example, the future work steps proposed. Furthermore, we also have learnt the computational difficulties of working with a large dataset in real-time big data problems with millions of observations.

By the proposed approach, the system is considered generic and applicable across European waters, firstly because the necessary inputs for achieving these high predictive performances are very light compared to other approaches, and secondly, because the intrinsic nature of the models enable the users to predict the aforementioned targets about the next destination of a vessel right after it arrives at a port, instead of waiting for the first AIS message data once the next voyage starts.

Our work validates that AI based insights can be suitable and create positive impact into the maritime logistics field, and contributes to the literature of applied AI which, in particular for this field of maritime logistics, had not developed previously extensive research mainly because the need to explore more this area has been growing in the last decade due to globalization.

## References

- [1] J. S. Henry Schwartz, Magnus Gustafsson, “Emission abatement in shipping – is it possible to reduce carbon dioxide emissions profitably?,” vol. 254, Elsevier, 2020.
- [2] Awake.ai, “Universal information exchange of maritime logistics.” [www.awake.ai](http://www.awake.ai).
- [3] I. M. Cockburn, R. Henderson, and S. Stern, *The Impact of Artificial Intelligence on Innovation: An Exploratory Analysis*, pp. 115–146. University of Chicago Press, January 2018.
- [4] P. Panayides, “Maritime logistics and global supply chains: Towards a research agenda,” *Maritime Economics and Logistics*, vol. 8, pp. 3–18, 03 2006.
- [5] P. M. Panayides and D.-W. Song, “Maritime logistics as an emerging discipline,” *Maritime Policy & Management*, vol. 40, no. 3, pp. 295–308, 2013.
- [6] M. Leclerc, R. Tharmarasa, M. C. Florea, A. Boury-Brisset, T. Kirubarajan, and N. Duclos-Hindrié, “Ship classification using deep learning techniques for maritime target tracking,” in *2018 21st International Conference on Information Fusion*, pp. 737–744, 2018.
- [7] C. Dao-Duc, H. Xiaohui, and O. Morère, “Maritime vessel images classification using deep convolutional neural networks,” in *SoICT*, ACM, 2015.
- [8] V. Marié, I. Béchar, and F. Bouchara, “Real-time maritime situation awareness based on deep learning with dynamic anchors,” in *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance*, pp. 1–6, 2018.
- [9] A. L. Ellefsen, E. Bjørlykhaug, V. Åsøy, and H. Zhang, “An unsupervised reconstruction-based fault detection algorithm for maritime components,” *IEEE Access*, vol. 7, pp. 16101–16109, 2019.
- [10] F. D. Malliaros and M. Vazirgiannis, “Clustering and community detection in directed networks: A survey,” *CoRR*, vol. abs/1308.0971, 2013.
- [11] M. Girvan and M. E. J. Newman, “Community structure in social and biological networks,” *Proceedings of the National Academy of Sciences*, vol. 99, pp. 7821–7826, June 2002.
- [12] Wikipedia contributors, “Centrality — Wikipedia, the free encyclopedia.” <https://en.wikipedia.org/w/index.php?title=Centrality&oldid=963626406>, 2020. [Online; accessed 28-Mar-2020].
- [13] A. Clauset, M. E. J. Newman, and C. Moore, “Finding community structure in very large networks,” *Physical Review E*, vol. 70, Dec 2004.
- [14] Wikipedia contributors, “K-means clustering — Wikipedia, the free encyclopedia,” 2020. [Online; accessed 30-July-2020].

- [15] M. Belkin and P. Niyogi, “Laplacian eigenmaps for dimensionality reduction and data representation,” *Neural computation*, vol. 15, no. 6, pp. 1373–1396, 2003.
- [16] U. von Luxburg, “A tutorial on spectral clustering,” *CoRR*, vol. abs/0711.0189, 2007.
- [17] A. Y. Ng, M. I. Jordan, and Y. Weiss, “On spectral clustering: Analysis and an algorithm,” in *Advances in neural information processing systems*, pp. 849–856, 2002.
- [18] A. Dasgupta, J. Hopcroft, R. Kannan, and P. Mitra, “Spectral clustering by recursive partitioning,” in *Algorithms – ESA 2006* (Y. Azar and T. Erlebach, eds.), (Berlin, Heidelberg), pp. 256–267, Springer Berlin Heidelberg, 2006.
- [19] J. Shi and J. Malik, “Normalized cuts and image segmentation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, 2000.
- [20] D. Nguyen, R. Vadaine, G. Hajduch, R. Garello, and R. Fablet, “A multi-task deep learning architecture for maritime surveillance using ais data streams,” in *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 331–340, 2018.
- [21] A. Graves, “Generating sequences with recurrent neural networks,” *CoRR*, vol. abs/1308.0850, 2013.
- [22] T. Yin, G. Zhong, J. Zhang, S. He, and B. Ran, “A prediction model of bus arrival time at stops with multi-routes,” *Transportation Research Procedia*, vol. 25, pp. 4627–4640, 12 2017.
- [23] O. Bodunov, F. Schmidt, A. Martin, A. Brito, and C. Fetzer, “Real-time destination and eta prediction for maritime traffic,” in *Proceedings of the 12th ACM International Conference on Distributed and Event-Based Systems*, (New York, NY, USA), p. 198–201, Association for Computing Machinery, 2018.
- [24] US Department of Homeland Security, “AIS system of the Navigation Center.” [www.navcen.uscg.gov/?pageName=AISmain](http://www.navcen.uscg.gov/?pageName=AISmain).
- [25] Wikipedia contributors, “One-hot — Wikipedia, the free encyclopedia.” <https://en.wikipedia.org/w/index.php?title=One-hot&oldid=966055631>, 2020. [Online; accessed 7-April-2020].
- [26] Wikipedia contributors, “Feature scaling — Wikipedia, the free encyclopedia.” [https://en.wikipedia.org/w/index.php?title=Feature\\_scaling&oldid=938494090](https://en.wikipedia.org/w/index.php?title=Feature_scaling&oldid=938494090), 2020. [Online; accessed 15-April-2020].
- [27] Keras, “Embedding.” [www.keras.io/api/layers/core\\_layers/embedding](http://www.keras.io/api/layers/core_layers/embedding).

- [28] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [29] Wikipedia contributors, “Softmax function — Wikipedia, the free encyclopedia.” [https://en.wikipedia.org/w/index.php?title=Softmax\\_function&oldid=965608678](https://en.wikipedia.org/w/index.php?title=Softmax_function&oldid=965608678), 2020. [Online; accessed 2-May-2020].
- [30] Wikipedia contributors, “Cross entropy — Wikipedia, the free encyclopedia.” [https://en.wikipedia.org/w/index.php?title=Cross\\_entropy&oldid=963499515](https://en.wikipedia.org/w/index.php?title=Cross_entropy&oldid=963499515), 2020. [Online; accessed 2-May-2020].
- [31] Wikipedia contributors, “Gradient descent — Wikipedia, the free encyclopedia.” [https://en.wikipedia.org/w/index.php?title=Gradient\\_descent&oldid=965468445](https://en.wikipedia.org/w/index.php?title=Gradient_descent&oldid=965468445), 2020. [Online; accessed 5-May-2020].
- [32] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA* (Y. Bengio and Y. LeCun, eds.), 2015.
- [33] Wikipedia contributors, “Early stopping — Wikipedia, the free encyclopedia.” [https://en.wikipedia.org/w/index.php?title=Early\\_stopping&oldid=959991526](https://en.wikipedia.org/w/index.php?title=Early_stopping&oldid=959991526), 2020. [Online; accessed 15-May-2020].
- [34] V. Nair and G. E. Hinton, “Rectified linear units improve restricted boltzmann machines,” in *ICML International Conference on Machine Learning*, 2010.
- [35] Wikipedia contributors, “Mean squared error — Wikipedia, the free encyclopedia.” [https://en.wikipedia.org/w/index.php?title=Mean\\_squared\\_error&oldid=963224523](https://en.wikipedia.org/w/index.php?title=Mean_squared_error&oldid=963224523), 2020. [Online; accessed 1-May-2020].

## A Appendix

**Table A1:** List of variables contained in AIS messages of type *positioning report*.

Variable	Description
Message ID	Identifier for this message.
Repeat indicator	Used by the repeater to indicate how many times a message has been repeated.
User ID	MMSI number of the vessel.
ROT	Rate of turn: 0 to +126 = turning right at up to 708 deg per min or higher, and 0 to -126 = turning left at up to 708 deg per min or higher. Values between 0 and 708 deg per min coded by $4.733 \cdot \sqrt{ROT_{sensor}}$ degrees per min.
SOG	Speed over ground in 1/10 knot steps.
Position accuracy	Flag determined in accordance with 1 = high which means $\leq 10$ meters, and 0 = low which means $> 10$ meters.
Longitude	Coordinate longitude.
Latitude	Coordinate latitude.
COG	Course over ground in 1/10 meters.
True heading	Heading direction in degrees 0-359.
Time stamp	UTC second when the report was generated.
Special manoeuvre	Described by 0 = not available = default, 1 = not engaged in special maneuver, 2 = engaged in special maneuver.
Navigational status	Described by 0 = under way using engine, 1 = at anchor, 2 = not under command, 3 = restricted maneuverability, 4 = constrained by her draught, 5 = moored, 6 = aground, 7 = engaged in fishing, 8 = under way sailing, 9 = reserved for future amendment of navigational status for ships carrying pollutant content, 10 = reserved for future ships carrying dangerous goods and harmful substances, 11 = power-driven vessel towing astern (regional use), 12 = power-driven vessel pushing ahead or towing alongside (regional use), 13 = reserved for future use, 15 = undefined.

**Table A2:** List of variables contained in AIS messages of type *static and voyages data*.

<b>Variable</b>	<b>Description</b>
Message ID	Identifier for this message.
Repeat indicator	Used by the repeater to indicate how many times a message has been repeated.
User ID	MMSI number of the vessel.
Call sign	Craft associated with a parent vessel, should use 'A' followed by the last 6 digits of the MMSI of the parent vessel. Some examples of these craft include towed vessels, rescue boats, tenders, lifeboats and liferafts.
Name	The name of the vessel should be as shown on the station radio license.
Type of ship and cargo type	Number starting by digit 0 = missing, 6 = passenger, 7 = cargo, 8 = tanker, 9 = other.
Overall dimension	Indicates the dimension of the ship in meters.
Type of electronic position fixing device	Described by 0 = missing, 1 = GPS, 2 = GLONASS, 3 = GPS/GLONASS, 4 = Loran-C, 5 = Chayka, 6 = integrated navigation system, 7 = surveyed, 8 = Galileo, 9-14 = not used, 15 = internal GNSS.
ETA	Estimated time of arrival provided by responsible in vessel.
Maximum present static draught	Draught in 1/10 meters.
Destination	The use of this field may be decided by the responsible administration.