Understanding and Predicting Port Congestion with Machine Learning

Marcel Marais

4th Year Project Report
Computer Science and Artificial Intelligence
School of Informatics
University of Edinburgh

2023

Abstract

Congestion in maritime ports is a high impact problem that has yet to receive major focus from machine learning researchers. With over three quarters of the world's trade by volume flowing through ports, understanding their operational capacity is critical [8]. This knowledge is not only essential for shipping companies but also for the broader economy.

This project leveraged Automatic Identification System (AIS) data to detect, in an unsupervised manner, the regions that encompass the ports of Los Angeles, New York and Savannah. We built upon work from Abu Alhoal et al.[1] and Peng et al. [40] and quantified congestion into three metrics: spatial density, mean inter-vessel distance, and moored-to-anchored ratio. Moored-to-anchored ratio, an entirely novel metric proposed by us, captures the relationships between the number of vessels using a port and the number of vessels waiting in the vicinity of the port.

We then forecast these metrics using only historical values, framing the problem as a timeseries prediction task. We showed that a simple statistical model (exponential smoothing) is insufficient to predict our metrics but that by using a long short-term memory (LSTM) neural network, results could be improved.

In our main contribution we made a case that using only historical values to predict congestion is problematic as the magnitude of congestion can be rapidly affected by macro-economic factors of which the model has no prior knowledge. To compensate for this problem we constructed a complimentary feature extraction model that predicted the arrival time of vessels near the target port. This model served as an additional feature to a timeseries prediction LSTM with the result that mean squared error (MSE) was decreased, on average, by around 20%.

Research Ethics Approval

This project was planned in accordance with the Informatics Research Ethics policy. It did not involve any aspects that required approval from the Informatics Research Ethics committee.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Marcel Marais*)

Acknowledgements

Thank you to my parents for their unwavering support, George for having a never-ending supply of hard problems and Sid for providing structure and guidance.

Table of Contents

Chapter 1

Introduction

1.1 Motivation

Machine learning and data-driven technologies have seen widespread research effort and subsequent industry adoption over the past decade [26]. By moving away from anecdotally developed heuristics and utilising these new methods to inform business decisions, organisations across many sectors have obtained a competitive advantage [5]. The maritime and freight shipping industry - which forms around 80% of the world's trade by volume [8] - is one of the largest industries that has historically struggled to integrate this new technology into their process optimisation strategies [33]. This is largely due to the complexity and extent of maritime planning problems [8], but also due to technological hurdles such as the poor quality of cellular networks in coastal areas [49].

This paper focuses on one of the greatest challenges facing the maritime industry: congestion in seaports. Seaport congestion refers to a situation in which numerous vessels are waiting outside a port, unable to load or unload freight, because the terminal is full [43]. Congestion has clear implications for shipping companies. By increasing idle times in ports profitability is decreased. A single container ship has an operating cost of approximately $9 million USD [13]. Given that the average time spent waiting at ports is 6.5% per year [47], even a slight reduction in port wait time could lead to significant cost savings via increased utilisation and decreased crew usage. Moreover, the delays caused by port congestion have serious ramifications for global supply chains and, thus, the economy as a whole. Fundamentally, congestion is the result of a mismatch between the demand for seaborne transport as well as freight shipping, and the number of vessels a port can accommodate. The ideal solution would be to reconsider the adequacy of infrastructure, but the long construction times, high cost of expansion, and a fundamental lack of understanding of the exact demand for container throughput in ports has lead this solution to be burdened with a myriad challenges of its own [50]. In turn, this has lead to a need for vessel operators to proactively avoid congested ports. As such, there has been a significant increase in the number of publications focused on finding big data and machine learning based solutions to port congestion and related issues [33].

1.2 Aims and contributions

This work aims to improve the understanding of port congestion by leveraging Automatic Identification System (AIS) data to develop, and subsequently forecast metrics tied to congestion. The effectiveness of these methods is then analysed by applying them to three of the largest ports in North America namely the ports of Los Angeles, New York and Savannah [16]. To our knowledge, this is one of the first studies that tackles both tasks using publicly available data. We believe that this is significant as it paves the way for more transparent and reproducible work in the field. We broadly categorised the work done into congestion quantification and congestion prediction:

Congestion quantification

- *Unsupervised port region detection*: The goal was to detect bounding regions for the berth and anchorage areas of a port using only a single reference coordinate and AIS data. We built upon work done by Peng et al. [40] and Abu Alhaol et al. [1] to show that the technique of using smaller bounding polygons as well as separately identifying berth and anchorage areas, results in regions that are applicable to different types of ports, and are better suited to metric calculation.
- *Congestion metrics calculation*: We proposed three spatio-temporal metrics: spatial density, mean inter-vessel distance (MIVD) and moored-to-anchored ratio, that track congestion. The first two metrics were extensions of the work done by Abu Alhaol et al. [1], and adapted to our multi-polygon port regions. The moored-to-anchored ratio is a novel metric that we propose. By considering the ratio of moored vessels (tied to the port) to anchored vessels (attached to the seabed with a large heavy object), we can better understand the distribution of vessel activity in the port, providing valuable insights into congestion levels.

Congestion prediction

- *Traditional time series forecasting*: In order to predict future values of our congestion metrics based solely on past data, we employed a traditional time-series forecasting model, simple exponential smoothing (SES), as well a long short-term memory (LSTM) neural network. These models are widely used in the field of timeseries analysis and have demonstrated their effectiveness in predicting future. The use of a low complexity model (SES), provided a reasonable lower bound of performance that was not clear in the recent work by Peng et al [40]. We argue, however, that framing this task as a timeseries problem is inherently challenging as congestion depends on numerous latent variables, the behavior of which cannot be inferred from historical congestion data alone.
- *Congestion forecasting contextualised with proximate vessels*: Port congestion is fundamentally described by the relationship between the capacity of the port and the number of vessels expected to utilise that port within a given time period. We created a model that, on a per-vessel basis, predicts how long a vessel will take to get to a port of interest. Within a constrained area this is a much simpler task as vessel's often follow similar routes to reach ports. This model acts as a feature extractor, and supplies a timeseries forecasting model with additional context (a

window into the future) to enhance the accuracy of predictions.

Chapter 2

Background

2.1 Data maturity in the maritime Industry

Contrary to being a cornerstone of global trade, the maritime industry lags behind other industries in integrating data-centric technological innovations, most notably Artificial Intelligence (AI) and big data [48]. Several key challenges faced by the industry, such as a lack of trust and understanding of AI systems and a reluctance to adopt new software solutions for fear of job replacement[51], can be the cause for this. The lack of data-centric technological innovation also negatively impacts maritime research, with poor data accessibility being one of the biggest issues and particularly relevant for our research. Shipping companies view their data as a competitive advantage and strict silos are therefore kept between various operators. Moreover, third party data vendors charge their, mostly corporate, clientele high prices making it costly for researchers to access their services. Despite all of the above factors, the pervasiveness of data-driven decision-making has prevailed and forced the maritime industry to embrace new technologies to optimize their operations. Adding further impetus to this are new regulatory and reporting requirements such as the EU MRV (Monitoring, Reporting and Verification) [30] introduced in 2021. These regulations require large vessels to report their CO_2 emissions which almost necessitates more thorough data analysis [66]. It is becoming increasingly clear to stakeholders that understanding and communicating their uncertainties via data can provide immense benefits for all players [49].

Advances in applying AI and big data to the maritime industry is further held back by the research climate. A high level of consolidation, with just three specialised maritime universities in China accounting for 11.5% of the total research output in 2019 [33], is problematic as while results are reported, code and datasets are not, meaning researchers within these institutions have a large advantage. Moreover, maritime research is often treated in the context of defense and security rather than optimisation, meaning governments and private contractors are incentivised to keep their work confidential.

To fully realise the potential of AI and big data in the maritime industry, there needs to be a shift towards a more open research culture, with greater transparency and collaboration among all stakeholders. This will require a concerted effort from academia, industry, and government to work together and share knowledge and resources for the benefit of the entire industry.

2.2 Anatomy of a Port

A port, or seaport, is a facility that primarily serves the purpose of allowing ships to dock and load and unload cargo, materials and passengers. The layout of a typical seaport includes a variety of areas such as container yards, warehouses, transit sheds, passenger terminals, and navigation channels, among others [60]. However, for the purpose of our study on congestion, we will focus on two key areas that are most relevant: berths and anchorages.

Berths are specific locations along the wharf or dock where ships can tie up to load and unload cargo. They are usually marked by numbers or letters and are assigned ahead of time to specific ships or shipping lines [22]. Berths are often equipped with cranes, hoists, and other equipment to move containers and other cargo on and off the ships. They are also connected to a network of access roads and rail lines that provide transportation links to the rest of the port. As seen in figure 2.1 (courtesy of Shuai Jia et. al [22]), the berth is adjacent to a cargo terminal.

Anchorages, on the other hand, are more general areas in the water, designated for ships to drop anchor and wait for their turn to dock at the berths. These areas are usually located in the open water outside the port area and are marked by buoys or beacons to indicate a safe distance from other ships and underwater obstructions [60]. Anchorages may be assigned by port authorities and are monitored by personnel to ensure safe navigation.

For typical cargo vessels it is necessary to first wait at an ”outer anchorage”, before being allowed to proceed to their designated berth. However, if once they have passed through the navigation channel, their berth is still not available they may be required to further wait at an ”inner anchorage” [22]. It has been found that vessels spend between 4% and 9% of their time at anchorage [47], which is hugely significant given that this time is entirely unproductive.

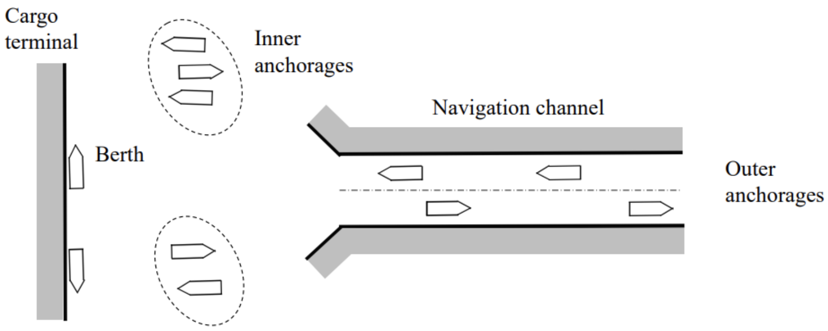


Figure 2.1: Illustrative port layout Shuai Jia et. al [22]

2.3 Port congestion and other challenges

Port congestion refers to a situation where a ship arrives at a port but cannot unload immediately due it being full. This occurs specifically when there are no appropriate berths available and ships have to wait at anchorage instead [43]. This typically results in queues and delays forming which increases overtime. This poses an external operational risk that is difficult for shipping companies to mitigate individually as congestion is often caused by inadequate infrastructure, labour shortages and poor communication between ports and vessels [6]. This directly affects the profitability of shipping companies, but congestion also has severe implications for supply chains and can affect consumers by leaving store shelves empty. It also affects businesses by delaying access to materials needed in manufacturing processes [23].

The issue of port congestion has become particularly topical recently due to several major world events over the past few years, most notably the COVID-19 pandemic's disruption of many industries heavily dependent on shipping, such as manufacturing and retail. The pandemic led to a severely reduced demand for transport goods which was reflected in a 7% year on year decrease in U.S maritime container import when compared with the first half of 2019 and 2020 [46]. The reduced number of ships in operation had a short term positive impact on congestion, but it did not last as the lockdown-induced ”e-commerce boom” in the latter half of 2020 catapulted the demand for shipping forward [57] . This is shown by the fact that U.S imports from Asia in December 2020 was 30% higher than the previous year [46]. This caused widespread delays as ports reached capacity.

Another recent global economic event, the 2021 commodities boom, which was largely fueled by Chinese demand for iron ore, also saw port infrastructure tested as delays reached historical levels and even caused dry bulk shipping rates to hit a decade high [11].

With global events being difficult to predict and port infrastructure difficult to improve, there are many factors to consider and trade-offs to be made when running a shipping company. One of the biggest decisions is choosing the type and number of ships for operations [10], often referred to as fleet size and mix. This involves assessing how large vessels should be and how many of each size is needed [38]. It is obvious that a business making a decision now about whether to expand or downsize a fleet that serves a specific route, will have an impact on congestion along that route for years to come. In light of this, one could imagine an ideal scenario in which all vessel operating companies communicated and balanced their intentions to achieve optimal utilisation of their current fleet without a change in port infrastructure being necessary. Moreover, they would know whether it made logistical-, as well as financial sense to change their current operations. However, due to the competitive nature of freight spot pricing and highly secretive practices in the industry, this is unlikely to ever be the case. This paper will, therefore, be addressing port congestion from the perspective of the port. In other words we will be investigating how the port's operations are affected by varying degrees of utilisation, and will not be directly considering the strategies or decisions of shipping companies.

2.4 Automatic Identification System (AIS)

AIS is a digital positional tracking system that operates on the Very High Frequency (VHF) maritime band. The original purpose of AIS was to increase safety and reduce collisions by continuously transmitting information such as a vessel's speed, position and identity [2]. This is done by utilising a vessel's internal GPS system along with shore based stations in order to create a map of traffic in nearby waters [62]. AIS data is transmitted at different intervals depending on the vessel's status. While the vessel is on the move, information is transmitted every 2 to 10 seconds, and while a vessel is anchored, every 3 minutes [37]. This data is typically fairly noisy as a result of inaccurate coordinates reported by ships, corrupted messages and occasionally spurious messages originating from illegitimate sources [37].

AIS messages are classified into two primary categories: static and dynamic messages. Static messages comprise information that rarely changes, while dynamic messages carry information that is updated more frequently, such as the vessel's position. This section explores the data provided by these two categories and the challenges that arise when attempting to utilise them for modeling purposes.

2.4.1 Static messages

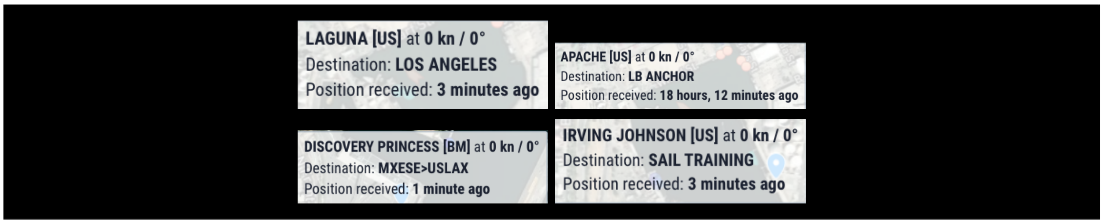
Static AIS messages are an essential source of information for identifying ships and understanding the characteristics that remain constant throughout their voyage. Typically, these messages are entered into the ship's AIS transponder by the crew or support staff through an electronic interface. These messages contain fields that relate information such as the ship's identity, ownership, type, and dimensions. Table 2.1 provides a comprehensive list of all the static messages transponded.

Data Field	**Description**
Ship Name	Name of the ship as registered with the flag state
Call Sign	International radio call sign of the ship
IMO Number	Unique seven-digit number assigned to the ship under the IMO system
Type of Ship	General category of the ship (cargo, tanker, passenger, etc.)
Dimensions	Length, breadth, and depth of the ship
Location of Position-Fixing Antenna	Vertical position of the ship's GPS antenna
Type of Electronic Position-Fixing Device	Type of GPS or other electronic device used for position fixing
Draught	Depth of the ship below the waterline
Destination	Port or area where the ship is heading
ETA	Estimated time of arrival at the destination
Ship's Information	Other relevant information about the ship (maximum speed, gross tonnage, owner/operator name, etc.)

Table 2.1: Static AIS Message Information [2] [25]

Unfortunately, data entry errors are a common cause of inconsistencies and inaccuracies in static AIS messages, and the destination field is particularly vulnerable to such errors.

Since the destination field is manually entered, it is susceptible to mistakes caused by ambiguity, misspellings, and intentional falsification. A major cause of this results from a lack of adoption and understanding of the standardised ”United Nations Code for Trade and Transport Locations” (UN/LOCODE) format, which associates each port to an abbreviation [55]. As a result, carriers tend to use their own spellings and abbreviations. Research conducted by Steidel et. al, [52] found that in a study of 4988 vessels, at least 52.2% of the destination fields were entered erroneously, and that only 1.3% were compliant with the UN/LOCODE format. Figure 2.2 shows our verification of this problem using the website marinetraffic.com[1]. Random ships were chosen from the Port of Los Angeles where the correct UN/LOCODE is USLAX - only one of the vessels matched this (and also included the port of origin).



LAGUNA [US] at 0 kn / 0°
Destination: LOS ANGELES
Position received: 3 minutes ago

APACHE [US] at 0 kn / 0°
Destination: LB ANCHOR
Position received: 18 hours, 12 minutes ago

DISCOVERY PRINCESS [BM] at 0 kn / 0°
Destination: MXESE>USLAX
Position received: 1 minute ago

IRVING JOHNSON [US] at 0 kn / 0°
Destination: SAIL TRAINING
Position received: 3 minutes ago

Figure 2.2: Variance in destination field for the same port (USLAX)

Additionally, changes to the destination, such as a modification to the port of call or voyage route, may not be updated in the AIS data, resulting in outdated or inaccurate information. Technological limitations also affect the reliability of the destination field. In particular, there are two types of transponder: class A (typically large container ships) and class B (typically domestic commercial vessels and pleasure craft), with class B containing much less detail [3]. The destination field is also susceptible to deliberate falsification by ship operators for various reasons. For instance, an operator may conceal the ship’s true destination for competitive or strategic reasons or to avoid regulatory scrutiny [52].

2.4.2 Dynamic messages

Dynamic AIS messages, in contrast to static messages, provide real-time information about a ship’s current position, speed, course, and other navigational data. These messages are automatically generated by the ship’s AIS transponder and transmitted at regular intervals, typically every two to ten seconds [3] Dynamic messages are crucial for monitoring ship movements, assessing potential collision risks, and identifying abnormal behavior. The most important dynamic messages are the position report and the voyage-related message. The position report provides the ship’s current position, speed, heading, and other navigational information, while the voyage-related message includes information about the ship’s route, intended destination, and estimated time of arrival. While dynamic messages are generally more reliable than static messages, they can still be subject to errors and intentional falsification, particularly in the case of

[1] MarineTraffic.com is a website that provides real-time information on the movements and locations of ships and vessels. The website collects data from various sources, including AIS and uses it to provide a variety of services, including vessel tracking, vessel details, port activity, and maritime news.

illicit activities such as smuggling or piracy. Table 2.2, shows the dynamic messages transmitted by a Class A AIS transponder.

Data Field	**Description**
Latitude	Geographic location (decimal degrees North)
Longitude	Geographic location (decimal degrees East
COG	Course over ground in degrees true
SOG	Speed over ground in knots
Heading	True heading in degrees (0-359)
Navigation Status	Ships current activity (e.g., underway, anchored)
ROT	Rate of turn in degrees per minute
Timestamp	UTC time stamp of the message transmission
MMSI (Maritime Mobile Service Identity)	Unique identifier for the ship's AIS transponder
Navigational Aids and Devices	Navigational devices on board (e.g., compass, GPS)

Table 2.2: Dynamic AIS Message Information [2] [25]

2.5 AIS data providers

The overwhelming majority of AIS data providers in the industry charge for access to their data. In an attempt to understand the pricing structure we reached out to one of the largest providers in the industry: Spire Global, asking how much one year of global AIS data would cost. Their representatives quoted us $36K USD for one year of vessel information (around 200K vessels). Needless to say this is outside of the $0K budget of this project.

The only publicly available sources we could find (besides the one we used) were live streaming services such as aisstream.io that provide real-time AIS data via an API but no historical data. To use something like this we would have needed to create infrastructure to store the data as it is streamed, which is outside the scope of our work.

2.5.1 Marine Cadastre

MarineCadastre.gov is an online portal that provides access to a wide range of geospatial data and tools related to ocean and coastal management in the United States [29]. The portal is a collaborative effort between the National Oceanic and Atmospheric Administration (NOAA) and the Bureau of Ocean Energy Management (BOEM), the agencies being within the U.S. Department of the Interior and Department of Commerce respectively [18].

Within their large repository of information, their AIS dataset is a key focus for this paper. The dataset spans the years from 2009 to 2022 and is frequently updated. The messages are gathered from vessels by receivers located on buoys and along the US coastline. While the dataset is fairly comprehensive, it does have several limitations that should be considered. Most importantly, the dataset only covers US waters, including the Great Lakes, coastal areas (roughly 50 miles from the coastline), and major inland waterways [36]. The dataset also does not provide destination messages at all. Their statement on the matter is: ”we collect the destination values; however, most of the entries are incomplete, inconsistent, or null”. This corroborates our findings in section 2.4.1.

Despite these shortcomings we believe that Marine Cadastre provides the most complete and up-to date AIS dataset and thus, we have chosen it for our work.

2.6 AIS Modelling Approaches

The availability of AIS datasets has opened up several opportunities for analysis and modeling. With a plethora of potential tasks and downstream applications to explore, we aim to investigate two broad predictive tasks: vessel movement prediction and port congestion forecasting.

2.6.1 Predicting vessel movement: destination, arrival time and trajectories

A simple way to think of a vessel's movement is relative to it's destination. Vessel destination and arrival time prediction involves predicting a vessel's final port of call and the time taken to reach it. This problem arises primarily from the poor quality and reliability of the destination and estimated time of arrival (ETA) fields in static AIS messages, as explained in section 2.4.1. To overcome this challenge, predictive models typically rely on the vessel's current location, past behavior, and other contextual factors such as cargo type. Specifically, most research in the field focuses on using historical vessel trajectories to predict the final destination, framing the problem as a classification task with the target variable typically being ports or cities.

Arrival time prediction, on the other hand, can be thought of as a regression task, with the goal being predicting the amount of time lapsed between the current observation of a vessel and an observation of the vessel at its destination. The objective can, however be simplified by framing it as a classification task, for example by classifying a time of day (morning, afternoon, evening) the ship will arrive. We will be discussing the noteworthy results that researchers have achieved while undertaking this task, using various techniques such as neural networks, decision trees, and statistical methods. It should be noted that, while this task is important for our method of understanding congestion, it is conventionally thought of in relation to other problems such as optimising shipping routes that often have delays [65] and surveillance [39].

Tree-based models have become a particularly popular choice due to their ability to capture the highly non-linear nature of vessel trajectories. A paper that arose as a solution to a competition centered on ”the application of machine learning to spatio-temporal maritime streaming data” [7] considered a problem in which the destination and arrival time of a given vessel, given AIS data provided by marinetraffic.com had to be predicted. These two tasks were learned separately, framing destination prediction as a multi-class classification problem in which each label is a port, and arrival time prediction as a regression task. For destination prediction, historical values of the ship's speed, longitude, latitude, departure port as well as some other vessel characteristics were used as input,. The classification problem was addressed by using an ensemble of tree-based models namely Random Forest, Gradient Boosting Decision Trees (GBDT), XGBoost Trees and Extremely Randomized Trees (ERT). A voting classifier, which predicts the label predicted by the majority of models, to obtain the final result was then used. To predict arrival time, a simple one layer feed-forward neural network with 200 neurons, with mean squared error as the chosen loss function was used. These approaches achieved an accuracy of around 97% for port destination classification, as well as good results on the arrival time task (results were not reported clearly). While this is impressive, the dataset used was highly curated and included features that may be difficult to derive at scale, such as the last port of call. The dataset also only spanned 28 days and was limited to vessels from the Mediterranean sea.

A more comprehensive paper by Zhang et. al analysed over 141 million AIS records and proposed a generalised method for destination prediction by using a Random Forest model to calculate similarity between a vessel's current trajectory and historical trajectories of vessels departing from the same port. This was done by sampling along the trajectories and calculating the perpendicular distance between themTthis is represented in figure 2.3. The idea being that similar trajectories will typically share a destination. As expected it was found that in trajectories where the vessel had been travelling for a longer time, the model achieved a higher accuracy. This was simply because, when the vessel was closer to its destination the similar trajectories are constrained to a smaller set and thus more accurate.

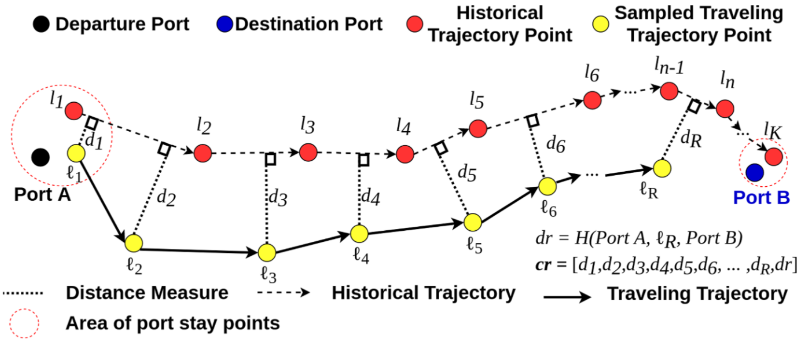


Figure 2.3: Trajectory similarity measurement in Zhang et. al

Predicting a vessel's trajectory directly is another approach to consider in vessel movement prediction. It involves anticipating a ship's future course based on its present location, previous behavior, and other contextual factors. This task is especially useful in predicting and preventing collisions [65], as well as monitoring vessel movements in and out of sensitive areas [35]. As this is a sequence modelling task, the models that are thought to be effective are somewhat limited when compared to the wide array of options typically available when dealing with classic supervised or unsupervised learning problems.

Trajectory prediction has historically been accomplished by using statistical models such as Autoregressive Integrated Moving Average (ARIMA) [63] and, while there has been some success using these techniques for tasks such as anomaly detection [39], the difficulty involved in setting hyperparameters for these models, as well as their limited ability to capture complex seasonality, [56] has caused them to largely fall out of favour. Therefore, modern research into trajectory prediction using time series AIS data has generally focused on applying deep learning methods.

Deep learning constitutes a wide array of techniques and algorithms but, due to the sequential nature of AIS data, the focus in the research community has been on architectures that have native support for sequences [64], [65]. For timeseries problems this means that the model was not trained on any data past the time horizon one is trying to predict for. The most popular group of models that have the ability to utilise this temporal dimension is Recurrent Neural Networks **Cite rnn** [27]. These neural networks have an internal memory state which acts as a summary of past information. On a practical level, this results in RNNs having the ability to "look back" in the data to inform inference. RNNs typically have issues utilising long-term temporal dependencies due to the vanishing gradient problem [54]. In order to overcome these issues a specialised type of RNN called Long Short-Term Memory (LSTM) networks was introduce which utilised gates within hidden units in order to improve the flow of data [20]. A study by Chen-Hong Yang et al. [64] addressed the trajectory prediction problem by first cleaning the raw AIS data by removing anomalies based on abnormal COG and SOG values. The setup of their study is not paritcullatry important but it was found that using a bi-directional long short term memory (BI-LSTM) model achieved the best prediction accuracy, outperforming statistical models such as ARIMA as well as other deep learning models such as an RNN.

A central issue for these sequential models is the high degree of noise present in AIS data. Thus, a significant amount of effort should be spent on feature engineering and data prepossessing. In order to combat this noise Nguyen et al. [35] proposed a four-hot encoded feature representation which essentially concatenates longitude, latitude, COG and SOG. They claim that this enhances the geometric features of AIS as well as enforces ”route-related characteristics” of trajectory data. This encoding was then used to predict anomalies using an RNN.

A similar study, also by Yang et al. [65], found most success in trajectory prediction by first clustering similar trajectories using density-based spatial clustering of applications with noise (DBSCAN) and then using the clustered data to train separate LSTM networks. Interestingly, trajectories were clustered using the COG field transmitted by AIS which essentially corresponds to the left and right turns a vessel makes over time. Their proposed method outperformed a LSTM model trained on all the data as well as an RNN.

Most recently, state of the art performance on sequence modelling tasks within fields such as natural language processing and computer vision has been achieved by Transformers - a neural network architecture that utilises a mechanism called attention [59]. Their effectiveness when applied specifically to time series forecasting tasks is somewhat unclear, however, as attention mechanism is permutation-invariant and, thus, requires positional encoding to preserve ordering [61].

2.6.2 Port congestion prediction

Port congestion prediction is a challenging task, made more challenging by the the lack of research on modelling it using machine learning. This task comprises two components: defining port congestion and forecasting it. Comparing studies is particularly challenging since there is no established standard for defining congestion. Nonetheless, as previously mentioned, this problem is significant not only for reducing waiting times but also for guaranteeing the uninterrupted movement of goods and people through ports. The two papers discussed in this section serve as the basis for our subsequent work.

AbuAlhaol et. al's paper on "Mining Port Congestion Indicators from Big AIS Data" [1] is a highly cited paper (by maritime research standards) approaches the task of defining port congestion. They filter AIS messages to create a dataset that contains "all static and dynamic messages located within a certain predefined geographical area". They then define a single large bounding region around these messages by applying Convex Hull [14] to the longitude, latitude pairs of each message. Within this region they calculate three maritime Port Congestion Indicators (PCIs): spatial complexity, spatial density and service criticality. We do not explain these here as they are defined in section 4.2. They argued that if each of these indicators is high a port is congested. Notably, they did not attempt to forecast or analyse the trends of their indicators.

Peng et al. [40] expanded upon the research conducted by AbuAlhaol et al.but instead showed that using DBSCAN to cluster AIS messages and then applying convex hull to those clusters resulted in more to refined areas where ships dock at ports. This method also allowed for distinguishing between berth/moorage and anchorage areas. The authors argue that calculating metrics on an hourly basis rather than monthly, as in AbuAlhaol et al., allows for greater granularity and improved monitoring of "slight changes in port performance". Using these smaller areas, new metrics such as "waiting time" and "load-discharge" were defined. However, unlike AbuAlhaol et al., Peng et al. did not use the convex hull area to scale their measures, making comparison between ports challenging. In addition, the authors employed an LSTM model to predict their measures, with multiple setups (uni-variate and multivariate) used for both multi-step and single-step prediction. It is worth noting that Peng et al. found predicting more than twelve hours in advance to be difficult, resulting in high errors. Their models achieved somewhat limited performance, highlighting the difficulty of this task.

In our work, we integrated effective methodologies from both of these papers to enhance the accuracy of both the quantification and prediction aspects of the task.

Chapter 3

Data pipeline and ingestion

3.1 Overview

Working with AIS data presents unique challenges due to the massive amount of data generated by the system as a result of the high frequency of transmissions. As highlighted by AbuAlhaol et. al, in their work on mining port congestion indicators [1], extracting even the most basic insights from AIS data tends to require complex and cumbersome data engineering. This is because we often want to calculate statistics based on the movement of only a few vessels over many weeks or even months but this, due to the streaming nature of AIS, requires looking through all the messages for the date range you are interested in. Thus, careful consideration needs to be taken to optimise performance and memory usage - especially with the limitations imposed by the University teaching cluster[1]. This chapter will provide an explanation of how we filtered and sampled the raw AIS dataset from Marine Cadastre, to allow for tractable metric computation and model training.

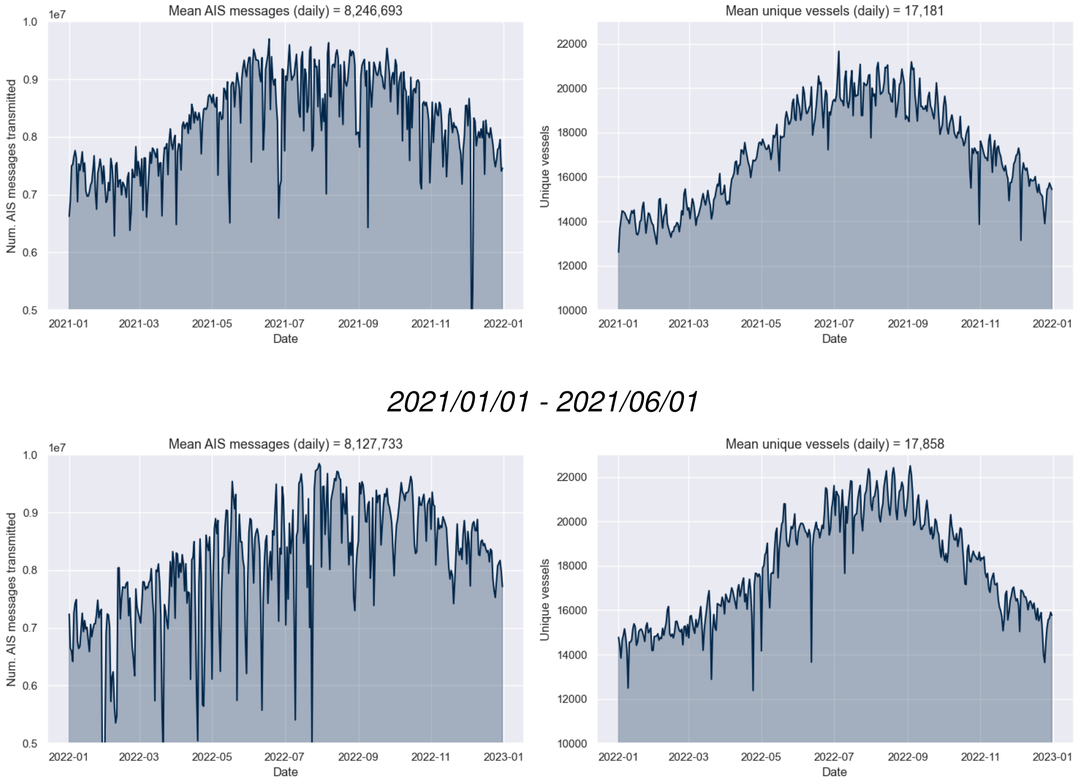
3.2 Marine Cadastre dataset preliminary investigation

Some of the data quality issues raised by Marine Cadastre have serious implications for the quality of our analysis. Most notably, some days in the dataset have considerably less AIS messages than the average. In the following quote they claim this is out of their control and rather in the hands of the U.S Coast Guard, ”at this time, the U.S. Coast Guard does not provide an explanation for the variability of data coverage and data volume for their AIS network” [36]. This, coupled with the COVID-19 pandemic having an outsize impact on the shipping and travel industries [23], meant that even selecting a date range to focus our efforts on was non-trivial and required prior analysis. To ensure the data is recent and minimally impacted by the pandemic - we chose to investigate the AIS messages in 2021 and 2022.

For the years in question (2021 and 2022) the AIS transmissions are segmented

[1] NVIDIA GTX1060 6GB graphics cards are provided (https://computing.help.inf.ed.ac.uk/teaching-cluster)

into days. Specifically, each file has the format: `AIS_YEAR_MONTH_DAY.csv`, so `AIS_2021_01_01.csv` will contain all the transmissions for the 1st of January 2021. Both years have over 250 gigabytes of data when uncompressed corresponding to around 3 billion AIS messages each. This means that for this initial look, around half a terabyte of data was loaded in and out of memory. Our aim was to find a date range that had a reasonably constant (or at least constantly increasing) number of AIS messages transmitted per day, or in other words, as few anomalous dips in the number of transmissions as possible. In figure 3.1 we show the number of AIS messages transmitted as well as the number of unique vessels transmitting those messages in 2021 and 2022.



2021/01/01 - 2021/06/01

2022/01/01 - 2022/06/01

Figure 3.1: Marine Cadastre AIS data statistics 2021 & 2022

Overall, figure 3.1 reveals the seasonal nature of the maritime industry. With both the number of AIS messages transmitted and the number of unique vessels peaking in around August, likely due to retailers scaling up their inventories in preparation for the holiday season [4]. Moreover, we can see far fewer significant dips in the number of messages transmitted, as well as the number of unique vessels in 2021 when compared to 2022. This leads us to believe that for long periods of time in 2022 the U.S coast guard either deliberately blocked coverage, or suffered from transceiver outages in several large areas. Either way it is clear that this is less of an issue in the latter half of 2021, specifically between and June and December, and therefore that will be the date range we use for the proceeding work.

3.3 Dataset Creation

Even within the reduced time frame we identified there is around 1.5 billion AIS records to be processed. However our goals of port detection, congestion metric calculation and congestion prediction each have different computational limitations and thus have differing data demands. Consequently, we have created three key port-specific subsets for each task that tailor their features and sampling methods to the method at hand.

These datasets each required many distance calculations between coordinates, which should be done using a distance measure that considers the spherical surface of the Earth. We evaluated two options: Haversine distance, assuming a perfect sphere, and Vincenty distance, accounting for the flattened ellipsoid shape of the Earth. Haversine distance is faster computationally, but less accurate for longer distances or high latitudes, while Vincenty distance is more precise but slower [28]. We chose Haversine distance for our short distance and high volume calculations, resulting in over twice the speed compared to Vincenty distance. The formula for the haversine distance between two coordinate pairs X and Y is given by formula 3.1.

$$d = 2r\sin^{-1}\left(\sqrt{\sin^2\left(\frac{X_{lat}-Y_{lat}}{2}\right)+\cos(X_{lat})\cos(Y_{lat})\sin^2\left(\frac{X_{lon}-Y_{lon}}{2}\right)}\right) \tag{3.1}$$

3.3.1 Port bounding region dataset

In this dataset we require the coordinates of static vessels in port. This enables the detection of the berth and anchorage areas of that port with DBSCAN & Convex Hull in section X. To achieve this we performed several filtering steps to reduce noise (ships not at berth or anchorage). Firstly, using the ”navigation status” AIS field - which provides an indication of a vessels activity - we filter for values 1 (anchored) and 5 (moored / tied to another object to limit free movement). Next, we calculated the distance from every coordinate transmitted by every ship and select only those that are within a specified threshold d, which is decided on a port specific basis. Algorithm 1 shows the filtering in further depth.

Algorithm 1 Port bounding region data-set construction

```
Input: AIS Messages $A_0 \ldots A_n$, Distance threshold $d$, Port coordinates $(p_{\text{lat}}, p_{lon})$
filteredMessages = []
for i = 0 to n do
    distToPort = distance(coordinates($A_i$), $(p_{\text{lat}}, p_{\text{lon}})$)
    if (Status($A_i$) == 1 or Status($A_i$) == 5) AND distToPort $\leq d$ then
        filteredMessages.append($A_i$)
    end if
end for
```

Algorithm 1 was applied using the coordinates of the three ports on interest. This reduced the amount of data to be considered for port-region identification by over 90%.

3.3.2 Congestion metric set

The purpose of this dataset was to facilitate the computation of congestion metrics, which were defined in Section 4.2. Therefore, the data should only include vessel movements within port moorage and anchorage areas that we encapsulate by the polygons created in section 4.1. Ideally, this should be a subset of the port bounding region dataset, however while the imperfect and sporadic use of of the *"status"* field is fine for approximating berth and anchorage areas it will not be suitable for calculating the congestion metrics as we desire a complete set of all vessels that visited the port. To this end we reran algorithm 1, without the status filtering step. This took over **40 hours** per port on the University teaching cluster and resulted in a three datasets of size of around 3.5gb, 4.5gb and 1.5gb for the Port of Los Angeles, New York and Savannah respectively. Next, to further subset these datasets and only include only vessel movements within port berth and anchorage areas, we framed the task as a "point-in-polygon" problem which asks whether a point lies within a polygon's boundary. We leveraged the computational geometry python library, Shapely, and its built-in `contains` method, which employs the Ray Casting algorithm [42] to perform this check. The time complexity of the algorithm is linear in the number of vertices in the polygon, or $O(n)$, where n is the number of vertices [21]. Given the large number of coordinates and polygons (each with multiple vertices) in our dataset, it is evident that the computational demands are significant.

With this approach every coordinate in the dataset was associated with the polygon (moorage anchorage) that it belongs to or was filtered out. Calculating the three metrics over the 6 month period took over 10 hours per port.

3.3.3 Arrival time dataset

In this final dataset our goal is to determine, for each vessel in a 25km radius of the port, if that vessel is going to the port we are interested in and if so calculate it's arrival time in order to train a model to predict it. By utilising the fact that we know the identifier of every vessel that entered the port during the period, from the congestion metric dataset we just explained, the number of vessels we need to calculate arrival time for is restricted to only vessels that are guaranteed to enter the port. This may seem obvious but is not possible without the prior effort.

Given our modelling task is predicting arrival time on a per vessel basis, with a 12 hour look back, our training samples are formed by shifting a 12 hour rolling window through our 6 month subset. Then for each hour in the window we calculate the mean haversine distance from the port and the mean speed over ground value (SOG) for each vessel within our radius that entered the port. The target for each observation is the difference in time between the latest observation of the vessel in the window and the first observation of the vessel in a berth/moorage in the target port. Vessels that did enter port but are within the radius get a target of -1. It was crucial that we add every vessel in the radius to this dataset because at inference time the model does not know which vessels are going to the port. Due to the high amount of vessels near busy areas, this data took over 50 hours to construct for each port.

Chapter 4

Modelling

4.1 Port area recognition

In order to understand congestion it is first necessary to identify the anchorage and moorage areas of the ports we wish to analyse. The methodology we present follows the core ideas of AbuAlhaol et al. [1], in using DBSCAN and convex hull but also draws from Peng et al. [40] in differentiating between berth and anchorages. Identifying these regions will allow us to calculate the necessary metrics and subsequently derive metrics to enable prediction.

> ***Task formalisation:*** for a port P find two sets of polygons A and M. For $M = \{m_1 \ldots m_n\}$, the majority of the area enclosed by each m_i overlaps with a moorage area, and contains with as little open ocean as possible. The total area $\sum_{i=0}^{n} m_i$ should also cover as much of the total moorage area of P as possible. For $A = \{a_1 \ldots a_n\}$, each polygon a_i capture a possible anchorage region.

In section 3.3.1 we detail the construction of a data set, with respect to a port, containing ships that are moored or anchored. In order to overcome the noise, introduced by incorrect AIS messages (ships reporting to be anchored or moored when they are not), in this data set, as well as segment berths anchorages, we will detect high density areas using the unsupervised clustering algorithm Density-based spatial clustering of applications with noise (DBSCAN).

DBSCAN works by defining a neighbourhood around each point in the dataset, coordinates in our case, and grouping together points that have a minimum number of neighbours within a specified radius. DBSCAN has two main parameters: Epsilon (ε) and `minPoints` [17]. Epsilon determines the radius around each point that will be considered in the neighborhood. `MinPts` sets the minimum number of points required within the epsilon radius to form a cluster [17]. Points that are not within any cluster are considered noise. By adjusting these parameters, we can control the sensitivity of the algorithm to the density of the data, allowing for effective clustering in ports with varying levels of noise as well as anchorage and moorage sizes.

DBSCAN was chosen over other unsupervised clustering algorithms such as k-means and hierarchical clustering, because of its ability to handle noise and clusters with irregular shapes [40]. It also does not require specifying the number of clusters a priori - which is unlikely to be known for every port.

DBSCAN clustering was applied to *anchored* and *moored* vessels individually, the goal being to assign each point in the dataset to either an anchorage cluster $c_1^a \dots c_n^a$ or a moorage cluster $c_1^m \dots c_n^m$ or, otherwise, is flagged as noise. The goal is to convert each of these clusters into a polygon to allow for simplified computation of port statistics as shown in section 4.2. To achieve this goal, the convex hull algorithm was chosen.

Convex Hull is used to find the smallest convex polygon that encloses a given set of points, meaning that it identifies the outermost boundary of a set of points such that no point lies outside the polygon [14]. It works by first sorting the points in increasing order of their polar angle with respect to a fixed point (often the leftmost point). Then, it constructs a convex hull by iteratively adding points to the boundary of the polygon. Specifically, it adds a point to the boundary if it makes a left turn from the previous two points on the boundary. This process continues until the starting point is reached again. It is possible to use more accurate algorithms such as concave hull, however due to the large number of points in each cluster it becomes very inefficient to calculate. The convex hull was applied to each cluster in both cluster sets to create the corresponding polygon set $\{\text{ConvexHull}(c_1^a) \dots \text{ConvexHull}(c_n^a)\} \rightarrow \{a_1 \dots a_n\}$.

4.2 Port congestion quantification

With the berthage and anchorage areas identified for a given port, it is now possible to calculate statistics with respect to it's polygon sets. In work done by AbuAlhaol et. al, the metrics spatial density, spatial complexity and time criticality with respect to a single convex hull region [1] are presented. In this section we formalise the extension of two of these metrics to the multi-polygon case, first introduced by Peng et. al [40]. We argue that our choice of metrics are robust to different types of ports, as well as more precise, due to the fact that we reduce the amount of open ocean considered. Moreover, using the the convex hull regions we created, we distinguish between moored (berthed), and anchored vessels and present a novel metric that directly tracks the relationship between them.

Another important distinction between our work and AbuAlhaol et. al, is that, when normalising using the maximum historical value, we are careful to use the maximum historical value of only our train set, as using the maximum value across the entire dataset would be a form a target leakage. Target leakage is when information that would not be available at the time of prediction in a real world scenario is used in the feature engineering process [53]. This can lead to optimistic performance metrics during training, and poor performance in practice.

4.2.1 Notation

We will use the following notation and high-level functions to describe statistics and attributes of the moorage and anchorage polygons

a_i / m_i	Anchorage / moorage bounding polygon i for a Port.
$a_i^{(t)}$ / $m_i^{(t)}$	All vessel movement in polygon i during time period t
$\text{ConvArea}(x_i)$	Convex Hull area of polygon i, where $x \in \{a, m\}$
$\text{VesselCount}(x_i^{(t)})$	# Unique vessels in polygon i during time period t
$\text{dist}(u, v)$	Haversine distance between vessel u and vessel v
T_{train}	The set of time periods in the training set

For readability, we define the weighting factor α in equation 4.1. This will be used to scale the metrics from individual polygons based on their relative convex hull size.

$$\alpha_i = \frac{\text{ConvArea}(m_i)}{\max\limits_{i \in I}\{\text{ConvArea}(m_i)\}} \tag{4.1}$$

4.2.2 Spatial density

The spatial density of a sea port is a normalised density measure that considers the number of vessels in the port's moorage areas as well as the size of those moorage areas. We are only interested in moorage areas for this metric, as anchorage areas tend to be much larger and less well defined, leading to noisy observations. We define a weighted sum of the densities of each polygon, for a time period t as $\theta^{(t)}$ in Equation 4.2.

$$\theta^{(t)} = \sum_{i=0}^{n} \alpha_i \text{VesselCount}(m_i^{(t)}) \tag{4.2}$$

The weighting with α (Equation 4.1) allows us to extend the metric to multiple polygons, the intuition being that we want to discount the density of small moorage areas, as they can become highly dense without contributing much to the overall congestion. We now define the normalised spatial density for a time period t in terms of θ^t in Equation 4.3.

$$\textbf{SpatialDensity}^{(t)} = \frac{\theta^{(t)}}{\max\limits_{t \in T_{\text{train}}}\{\theta^{(t)}\}} \tag{4.3}$$

4.2.3 Mean inter-vessel distance (MIVD)

This metric is an reframing of ”spatial complexity” by AbuAlhaol et. al. Firstly, in Equation 4.4 we define the *average vessel proximity* within a single moorage polygon over a time period as the average distance between every pair of vessels, $(u, v)_j$, in the polygon. Since each vessel in the polygon is assumed to be stationary the position taken to calculate the distances is the longitude latitude pair that is transmitted the most by that vessel over the time period. To calculate the average inter-vessel distance across all polygons we perform a weighted sum with the α and the inverse of δ in Equation 4.5. This means when vessels are close to each other (distances are low), the Δ value will be high.

$$\delta_i^{(t)} = \sum_{j=0}^{\text{VesselCount}(m_i^{(t)})} \text{dist}(u,v)_j \quad \forall (u,v)_j \in m_i^{(t)} \tag{4.4}$$

$$\Delta^{(t)} = \sum_{i=0}^{n} \frac{\alpha_i}{\delta_i^{(t)}} \tag{4.5}$$

We then simply define the normalised mean inter-vessel distance (MIVD), by dividing the max historical value:

$$\mathbf{MIVD}^{(t)} = \frac{\Delta^{(t)}}{\max\limits_{t \in T_{\text{train}}} \{\Delta^{(t)}\}} \tag{4.6}$$

This metric also only considers moorage areas because vessels at anchorage can be very far from each other or the port, leading to the metric losing some meaning.

4.2.4 Moored-to-Anchored Ratio

In this novel metric we are interested in understanding how many vessels are waiting at anchorage compared to how many vessels are moored and engaging in port activities. To do this we calculated a ratio between the total count of moored vessels and the total count of anchored vessels, shown in equation4.8. This should give an indication of the overall efficiency and capacity of the port. If there are a large number of vessels waiting at anchorage, it could indicate that the port is experiencing congestion and delays in processing incoming vessels. On the other hand, if there are a smaller number of vessels waiting at anchorage and a larger number of vessels engaged in port activities, it could suggest that the port is operating efficiently and can handle a higher volume of traffic.

$$\beta^{(t)} = \frac{\sum_{i=0}^{n} \text{VesselCount}(a_i^{(t)})}{\sum_{j=0}^{l} \text{VesselCount}(m_j^{(t)})} \tag{4.7}$$

$$\mathbf{Moor2Anch}^{(t)} = \frac{\beta^{(t)}}{\max\limits_{t \in T_{\text{train}}} \{\beta^{(t)}\}} \tag{4.8}$$

4.3 Data Preprocessing

Data preprocessing is a fundamental step in the modeling workflow and plays a crucial role in improving the quality and accuracy of models. Although by calculating our metrics we have in some sense already preprocessed our raw data, additional steps still need to be taken to ensure the data is fit for modeling.

Resampling: The inconsistent rate of transmission of AIS messages is a big challenge when calculating the congestion metrics and training models. A vessel can transmit any number of messages in a time period (or none at all). This is not an issue for the static messages (since we do not have access to the destination and ETA fields in the marine cadastre dataset), but dynamic messages need to be resampled in order to obtain a consistent time series that can be used for analysis. Resampling involves converting the data from one time interval to another, for our use case we convert the irregular continuous AIS messages into hourly observation by taking the median of the fields we are interested in. Using the mean in this instance would be a problematic choice due to outliers caused by transponder faults, as well as that, taking the mean of geographic coordinates may result in vessels being placed on land (this is most relevant to the mean inter-vessel distance metric).

Scaling drastically impacts the accuracy and performance of models. Proper scaling ensures that features are on a comparable scale, allowing models to effectively learn from each feature without being biased towards those with larger values. Scaling also helps speed up the training process by reducing the number of iterations required for models to converge [34].

Although the congestion metrics are already normalised (all being between 0 and 1), they have drastically different ranges. Specifically, MIVD was found to have much lower variance than the other two metrics. To combat this we opted to use the Z-score normalisation (also known as standard scaling), which scales numerical features to have zero mean and unit variance. This is done by subtracting the mean of the feature from each value and dividing by the standard deviation, as seen in equation 4.9.

$$x_{\textbf{(standard scaled)}} = \frac{x - \bar{X}}{\sigma_X} \tag{4.9}$$

Null values: There are two cases of missing values to consider. In some rare cases cases a vessel will transmit a message with a critical field, specifically MMSI, position (longitude / latitude) or SOG, missing. This is particularly anomalous as they should be included automatically. We found that this happens less than 2% of the time in the constructed datasets for all ports and, thus, opted to simply filter them out. Another case is if there are no transmitted values near the port for a given hour, making it impossible to calculate the metrics. This happened for only a few hours, but to overcome it we filled the values using the linear interpolation method in Pandas[1]. This is preferred over another method such as replacing with the mean as the trends are not interrupted.

[1] https://pandas.pydata.org/docs/reference/api/pandas. DataFrame.interpolate.html

4.4 Congestion prediction

In this section, we investigate the application of machine learning and statistical models to predict the metrics associated with seaport congestion, as defined in section 4.2. Our focus is on understanding models which utilise historical observations to make future predictions, as well as to explain our novel arrival-time informed setup. The models we used include exponential smoothing an LSTM neural network. However we also experimented with an ARIMA model.

4.4.1 Time Series Forecasting Models

Our time series forecasting task involves predicting future congestion values for a given port based on historical congestion observations. In our case, a single observation of congestion at time period (hourly) t is described by the three normalised metrics in Equation 4.10.

$$x_t = \{\mathbf{SpatialComplexity}^{(t)}, \mathbf{MIVD}^{(t)}, \mathbf{Moor2Anch}^{(t)}\} \tag{4.10}$$

Our task is to predict congestion values, $x_{t+1} \ldots x_{t+\beta}$, given congestion observations $x_{t-\alpha} \ldots x_t$. To achieve this, we have set β to a fixed value of 12 hours, meaning that each model aims to forecast the next 12 hours of congestion. However, we will vary the value of α in order to determine which context is most effective for each model.

4.4.1.1 Simple Exponential Smoothing (SES)

SES is a basic method for time series forecasting that works by taking a weighted average of past observations, with the weights decreasing at an exponential rate as one goes back in time [31]. This is shown in the equation

$$\hat{y}_{t+1} = \alpha y_t + (1-\alpha)\hat{y}_t \tag{4.11}$$

This technique is designed for uni-variate data and, thus, does not consider the interactions between variables. Therefore, we will simply fit three SES models to our dataset - one for each metric. It also does not have native support for multi-step prediction and has to be applied recursively, resulting in predictions being a flat line [31]. But this does not inherently make it a bad model given the large amount of variance in our data.

4.4.1.2 LSTM

Long Short-Term Memory (LSTM) is a specialised type of recurrent neural network (RNN) architecture that excels at capturing long-term dependencies in time series data [20]. Unlike traditional RNNs that suffer from the vanishing gradient problem, LSTMs overcome this issue by using a specialized memory cell that allows the network to selectively remember or forget information over time. The memory cell is controlled by three gates: the input gate, the forget gate, and the output gate, which enable the network to selectively process and update information [19].

The **input gate** is responsible for regulating the amount of new information that is allowed to enter the memory cell. It determines which information is relevant and important enough to be stored in the cell. The input gate is controlled by a sigmoid activation function, which outputs a value between 0 and 1. If the input gate output is close to 0, it means that the input is not relevant, and if it is close to 1, it means that the input is highly relevant [20].

The **forget gate** determines which information from the previous time step is no longer important and should be removed from the memory cell. It is also controlled by a sigmoid activation function. If the forget gate output is close to 0, it means that the information in the memory cell should be retained, while if it is close to 1, it means that the information should be forgotten.

The **output gate** regulates the amount of information that should be passed to the output at the current time step. It determines how much of the information stored in the memory cell should be used to make a prediction. If the output gate is close to 0, it means that the information in the memory cell should not be used for prediction, while if it is close to 1, it means that all the information should be used.

Figure 4.1 from Graves et al. [19] shows the flow of information through the various gates in an LSTM memory block.

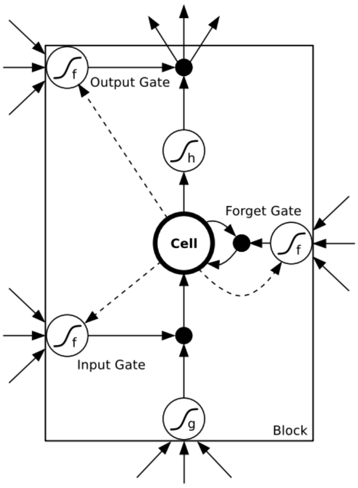


Figure 4.1: LSTM memory block with one cell from Graves et al. [19]. Each gate applies an activation function, then perform summation. Black dots indicate multiplication. f is the sigmoid activation function, while g and h are either the sigmoid or tanh activation function.

LSTMs are particularly effective when dealing with non-linear and non-stationary time series data because their many parameters allow for more complex relationships between the input and output than traditional statistical models. The ability of LSTMs to capture long-term dependencies in the data makes them ideal for modeling time series data with seasonality, trend, or cyclical patterns.

The performance of a LSTM model is highly dependent on the hyperparameters used during training. Below, we describe some of the most important hyperparameters and their impact on the LSTM model:

- ***Hidden units*** and ***layers*** significantly impact the model's ability to capture complex patterns. By increasing the number of hidden units and layers we can enhance the model's learning capacity but can also increase the risk of overfitting.
- ***Sequence length*** is effectively the amount of temporal context the model can utilize during training. Longer sequence lengths may provide benefit, but also require additional computational resources and could introduce noise.
- ***Learning rate*** controls how quickly the model's weights are updated during training, with a higher learning rate promoting faster convergence but increasing the likelihood of overshooting the optimal weights.
- ***Batch size*** determines the number of samples used in each training batch, with larger batch sizes often leading to more stable training but requiring additional memory resources.

In addition to hyperparameter tuning, choosing an appropriate optimiser is also crucial in LSTM training. The Adam optimizer, for instance, has been shown to provide superior performance in many applications due to its ability to adaptively adjust learning rates based on gradient statistics [9]. Other popular optimisers include RMSprop and Adagrad.

4.4.1.3 Other models

Autoregressive integrated moving average (ARIMA): is a time series forecasting method that combines autoregressive (AR), differencing (I), and moving average (MA) components [56]. Unlike SES, this model has support for complex trends and seasonality in the data, and incorporates forecast errors from past predictions to improve future predictions. Due to the highly subjective nature of choosing parameters, this model was very difficult to optimise manually [56]. And while there exists packages to do this automatically, such as AutoARIMA [41], the high computational cost due to the multivariate nature of our problem, and given that the `statsmodels` package is CPU bound, made achieving satisfactory results very difficult. This was especially problematic as we found that high numbers of moving average terms (the parameter q) was needed to achieve reasonable results, which added exponentially to the computational complexity, as more maximum likelihood estimations were required [58]. So, while this model was attempted in several configurations, it was, overall, not found to be particularly suitable for the task and did not provide a more rigorous baseline than SES.

4.4.2 Arrival Time Informed Time Series Model

We introduce a novel approach that seeks to address the difficulties inherent in time series forecasting using only historical values. To provide the model with additional context, we propose constructing a supplementary model that predicts whether vessels within a designated radius of a port are headed to that port and, if so, their expected arrival times. This model serves as a feature extractor for a time series congestion prediction model, allowing us to factor in the expected arrival times of ships and improve the accuracy of congestion metric calculations. Although predicting the precise coordinates and arrival times of vessels would obviate the need for the second model, this task is challenging because ships are not consistently assigned the same berth or anchorage each time they visit a port, and it may not be feasible to disambiguate this using AIS data.

This approach requires two models: the feature extractor/arrival time prediction model and a time series forecasting model that utilizes the predictions from the former, in addition to historical congestion metric values, to forecast future congestion metric values. In this section we will explain the architecture and design of each model, as well as the additional data processing needed to optimise their performance.

4.4.2.1 Arrival time prediction model

The primary goal of this model is to forecast a vessel's estimated time of arrival, based on a fixed lookback window of information. For each vessel we used 12 hours of historical positions (relative to the port) and speed over ground values to predict the time difference between the last observation of the vessel (typically the twelfth hours of our context window) and its initial observation in a berth at the target port.

For this task we use once again use an LSTM, taking advantage of its ability to model sequential data. Figure 4.2 shows an illustrative example of how the model processes an observation for a single vessel.

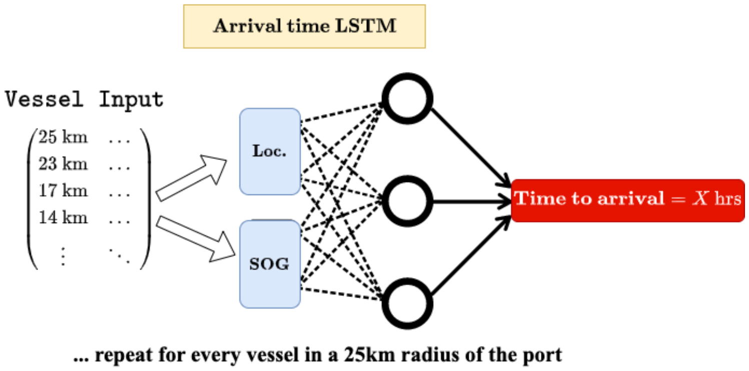


Figure 4.2: Arrival time LSTM inference on a single vessel

To obtain a complete set of features for our congestion model, the model needed to perform inference on every vessel within a specific radius - denoted as r - from the port. This resulted in an output vector - denoted as v_{AT} - that contained the estimated time of arrival for each vessel within the radius. The size of this vector was determined by the number of unique vessels within the radius.

However, for the congestion forecasting model to utilise the sequential nature of this data, we needed to reshape the v_{AT} vector to match the context length provided to the congestion forecasting model. Specifically, if the congestion prediction model used α hours of congestion metrics as a lookback, the arrival time vector, v_{AT}, needed to be reshaped to size α. This was achieved by grouping the estimated arrival times by the hour of arrival and counting the number of observations within each hour.

For example, suppose the model predicted that two vessels would arrive in four hours. In this case, the fifth element of v_{AT} would be set to 2, indicating that two vessels were expected to arrive at the port in the fifth hour. It is worth noting that we set the fifth index instead of the fourth index because the congestion model could predict vessels that were ”zero hours away”, meaning vessels that were less than one hour away from the port or already in a berthage area. This restructuring of the data is illustrated, with the lookback $\alpha = 12$hrs in figure 4.3.

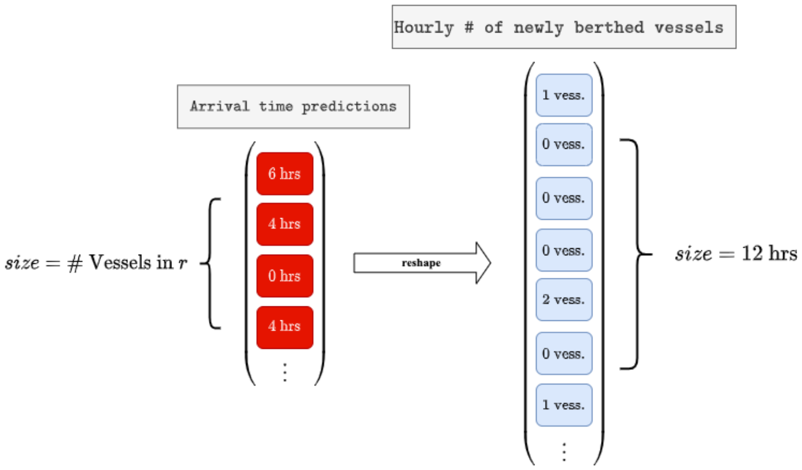


Figure 4.3: Arrival time data restructuring

4.4.2.2 Proximate vessel LSTM

This model is simply the original timeseries LSTM model, but with an extra input feature the V_{AT} vector as shown in 4.12. We kept this consistent because we wanted to measure the impact of the new input feature in isolation.

$$x_t = \{\mathbf{SpatialComplexity}^{(t)}, \mathbf{MIVD}^{(t)}, \mathbf{Moor2Anch}^{(t)}, \mathbf{NumArrivingVessels}^{(\alpha+t)}\} \tag{4.12}$$

4.4.2.3 Limitations

This technique is held-back by the radius we use for the feature extractor model. In other words, if the radius was increased the model would have a more accurate picture of the vessels that are going to arrive. However, considering that we only wish to validate that this technique is beneficial a smaller radius is sufficient.

4.5 Evaluation

In this section we will describe the measures we used to analyse the constructed congested metrics, as well as the performance of the models that predicted those metrics.

Our expectation for the constructed congestion metrics was that they would be somewhat correlated. For example, a more spatially dense port should result in more ships at anchorage and, thus, a higher moored-to-anchored ratio. As such, we used Pearson's correlation coefficient to measure the degree of linear correlation between each pair of congestion metrics [32]. The coefficient ranges from -1 to 1, where -1 indicates a perfectly negative correlation, 0 indicates no correlation, and 1 indicates a perfectly positive correlation. The Pearson correlation coefficient between two variables x and y, is given by equation 4.13.

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{4.13}$$

To evaluate the forecasts produced by the models we used the popular error metric Mean Squared Error (MSE) and Root Mean Squared Error (RMSE). The MSE and RMSE between a prediction $\hat{y}_i$ and a true value y_i, are given in equation 4.14.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad RMSE = \sqrt{MSE} \tag{4.14}$$

We opted to use MSE instead of other popular choices such as Mean Absolute Error (MAE) because MSE placed a higher penalty on large errors, which was desirable in our case as we were most interested in cases of extreme congestion.

The training set consisted of the first 70% of the congestion metrics and the test contained the final 30%, the split is shown in figure 4.4. Because we worked with a relatively small number of observations, we did not include a separate validation set that would be used for hyperparameter tuning.

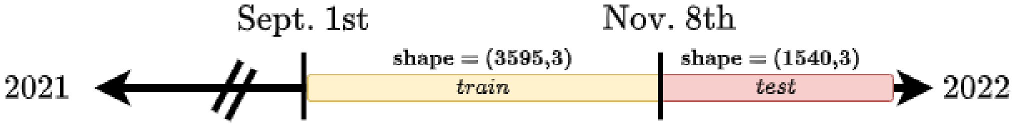


Figure 4.4: Train, test split (not to scale)

Chapter 5

Results

In this chapter we present the results of the three tasks outlined in the modelling section, namely port area identification, port congestion quantification and port congestion prediction. We critically evaluated the performance of the models developed and techniques used for each task using appropriate metrics and used visualisations to aid in their interpretation.

5.1 Port area identification

The port area recognition method described in section 4.1 was applied to three of the busiest ports in North America: the ports of *Los Angeles, New York and Savannah* [16]. In order to tune the DBSCAN parameters `minPts` and ε, and in turn verify the correctness of the polygons created, we had to rely on manual visual comparison with satellite imagery, as in Peng et al, [40]. Specifically, we checked that the polygons created from applying convex hull to the DBScan clusters corresponded with berth and anchorage areas present in satellite imagery, and did not contain too much open ocean. We then adjusted the parameters accordingly. An example of two extremes of the tuning process is illustrated in Figure A.1 in appendix A. We found that setting the ε parameter too high resulted in very large, overlapping clusters containing mostly ocean, whereas setting ε too low resulted in many small clusters - in some cases containing only a single vessel.

After completing the tuning process for the three ports, we calculated various statistics about the resulting polygons, specifically we were interested in the moorage polygons. In table 5.1 we report the number of moorage polygons for each port, the total area spanned by those polygons, as well the DBSCAN parameters used to create them. We also visualise the results for each port on a map in figures 5.1, 5.2 and 5.3. In general the identification was successful, however several issues did arise. Notably, moorage and anchorage polygons occasionally overlap indicating that many vessel operators have either said they are at anchor when they are not, or have switched the message to anchored only just before mooring.

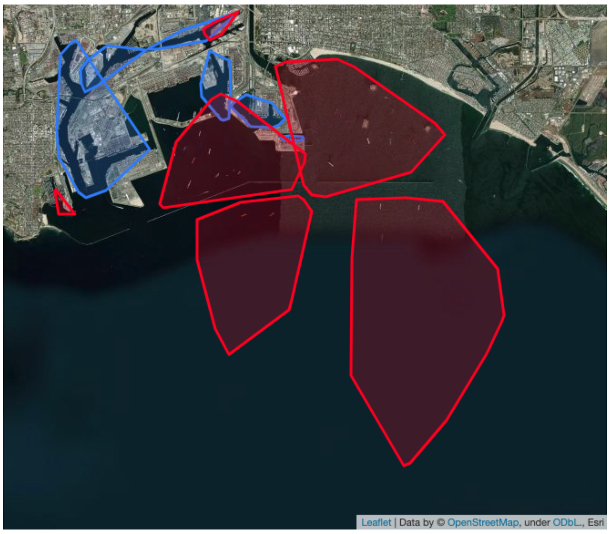


Figure 5.1: Port of Los Angeles region identification (berths in blue, anchorages in red)

Remarks: The operational components of the port of *Los Angeles* encompasses around 30km^2 of land and water [44]. During 2021 it was one of the busiest ports in America and historically received around 20% of the cargo entering the United States [12]. Our convex hull polygons capture, and clearly distinguish, between the anchorage areas in nearby coastal waters, and the moorage areas nearer to land. There is however some overlap in these areas, likely due to misuse of the "*status*" AIS field.

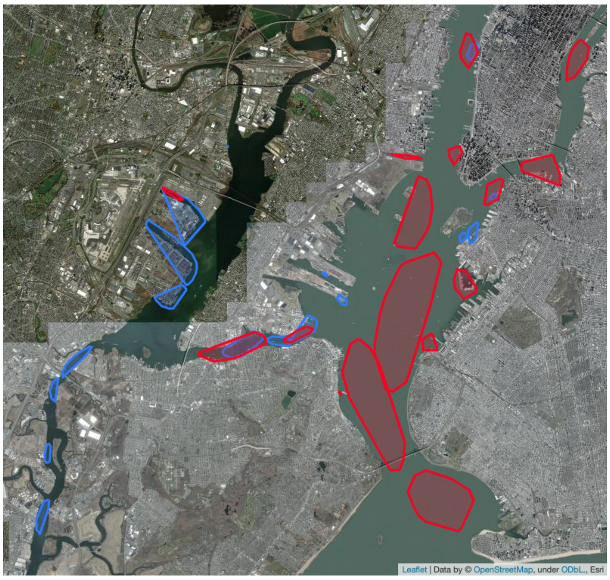


Figure 5.2: Port of New York region identification (berths in blue, anchorages in red)

Remarks: The *Port of New York* (and New Jersey) is another of the largest and busiest ports in the United States. It consists of a complex system of waterways, dock and terminals, and facilities spanning along the shores of New York City and northern New Jersey [45]. Our method is shown to work well over this very large area identifying many of the major moorage areas. However, some smaller berths have indeed been omitted Our hypothesis is that this is due to the limitations of DBSCAN when dealing with clusters of different densities [15].

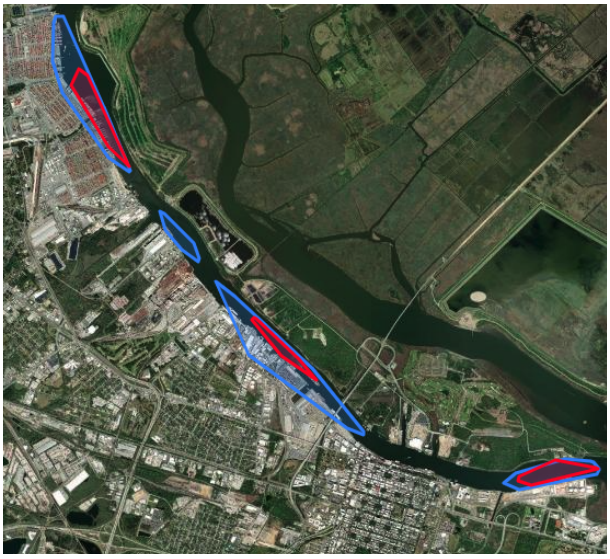


Figure 5.3: Port of Savannah region identification (berths in blue, anchorages in red)

Remarks: The Port of Savannah in Georgia is the fourth busiest container port in the United States, handling over 4.5 million twenty-foot equivalent units (TEUs) in 2021 [44]. Notably, the berthage areas are about 29*km* from the Atlantic Ocean, meaning that the majority of ships likely have to anchor far from the actual port. This showcases one of the limitation of using a 15*km* radius to detect berth and anchorage areas, as the anchorage areas shown are most likely a result of vessel operators forgetting to change the *status* AIS field from *anchored* to *berthed*.

Port Name	minPts	ε	# Anchorage Polygons	Tot. berth area (km^2)
Los Angeles	50	0.20×10^{-2}	6	22.35
New York	100	0.10×10^{-2}	14	11.90
Savannah	50	0.15×10^{-2}	4	3.09

Table 5.1: Final DBSCAN parameters and polygon statistics for all ports

5.2 Port congestion quantification analysis

The congestion metrics presented in section 4.2 were mined from the congestion dataset, which spans the dates 01/06/2021 - 31/12/2021 for each port. The metrics were calculated on an hourly time interval, resulting in 5136 observations of the form described in equation 4.10. In this section we provide an analysis of the metrics individually and investigate the correlation between them.

5.2.1 Spatial density analysis

Spatial density is arguably the congestion metric that is most closely tied to port utilisation. This is because a high spatial density indicates a relatively large number of vessels docked in the berthage areas of the port, while a low spatial density implies a lower level of activity. Our objective, therefore, is to validate this metric by analyzing trends over time and contextualizing them with external information.

In figure 5.4 we show the mean congestion across the hours of the day for our 6 month period. Given that the ports are some of the busiest in North America, and are, thus, highly utilised (used at all times of the day), we do not expect to see major fluctuations when averaging across many hours. However, there is a clear trend - particularly in the Port of Savannah and Port of New York - towards the ports having a higher spatial density during the traditional business hours of 9am to 5pm. The port of Los Angeles slightly subverts this trend, becoming slightly more spatially dense in the evening and late morning (7pm - 2am). We believe this is due to an announcement by the Biden administration in October stating that the port of Los Angeles will start operating 24 hours a day to try clear the pent up demand caused by Covid-19, thus resulting in a historic number of vessels utilising night-time berthage slots [24].

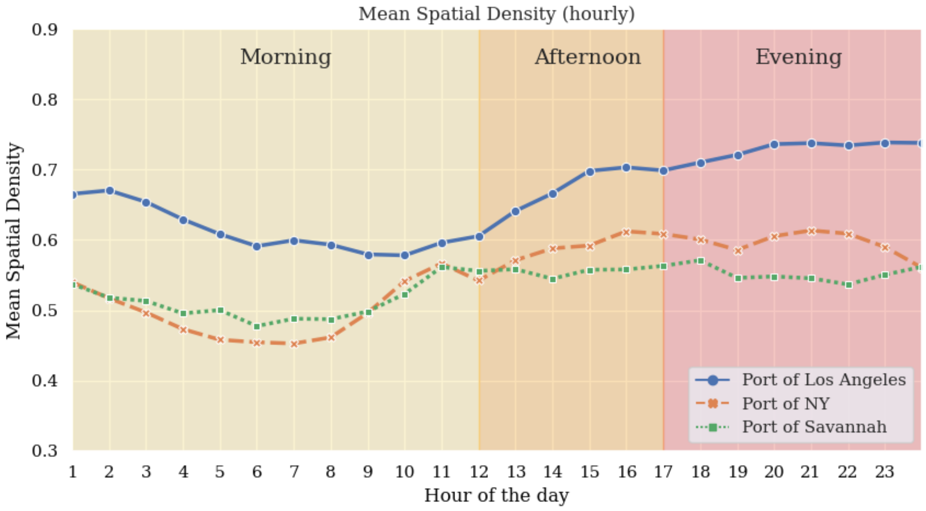


Figure 5.4: Mean spatial density (hourly)

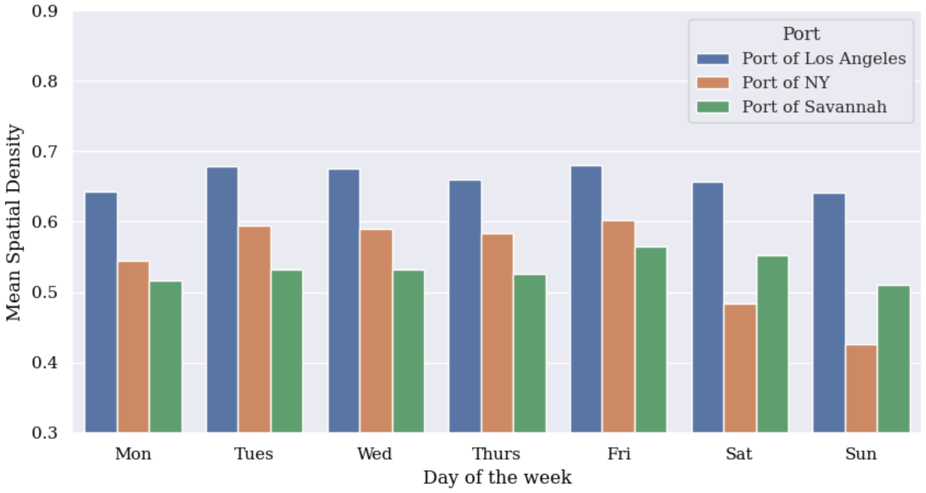


Figure 5.5: Mean spatial density (day of the week)

In figure 5.5 we show the mean spatial density across the days of the week. Once again, the fact that, during 2021, the Port of Los Angeles was the busiest container port in North America [16] is evident by its consistent and high spatial density across the days of the week. However, we noticed that the Port of New York sees drastically reduced spatial density on the weekend compared to weekdays, indicating a lower level of activity during those days. This could be attributed to a variety of factors, but most likely is simply a result of opening hours for container yards being reduced during the weekend, meaning fewer vessels can unload their cargo during those days.

5.2.2 Mean inter-vessel distance (MIVD) analysis

Mean inter-vessel distance calculates the average distance between every pair of vessels within a given port. Compared to Spatial Density, it provides an alternative perspective on the utilisation of a port and gives insight into how freely vessels can move within the port. A high mean inter-vessel distance does not, however, mean that a port is congested. In our analysis, we investigated the values of mean inter-vessel distance and aimed to validate their accuracy by taking into account the layout and characteristics of the respective ports.

Figure 5.6 displays the distribution of mean inter-vessel distance (MIVD) observations for each port over a 6-month period. Overall, these distributions appear to be approximately normal and exhibit three distinct means. We hypothesised that these means were related to the size and layout of the ports, as illustrated in Figures 5.1, 5.2, and 5.3. Specifically, larger and more dispersed ports, such as the Port of New York, tended to have higher MIVDs than more compact ports like the Port of Los Angeles. Interestingly, the Port of Savannah, while, not as physically extensive as the other two in terms of acreage, had a higher MIVD mean due to its elongated shape. Furthermore, the more evenly spread observations around the means of Savannah and New York suggest that the variance in the distance between the berths in these ports is higher than those in Los Angeles. This is intuitive when considering the ”horseshoe” like shape of the port of Los Angeles.

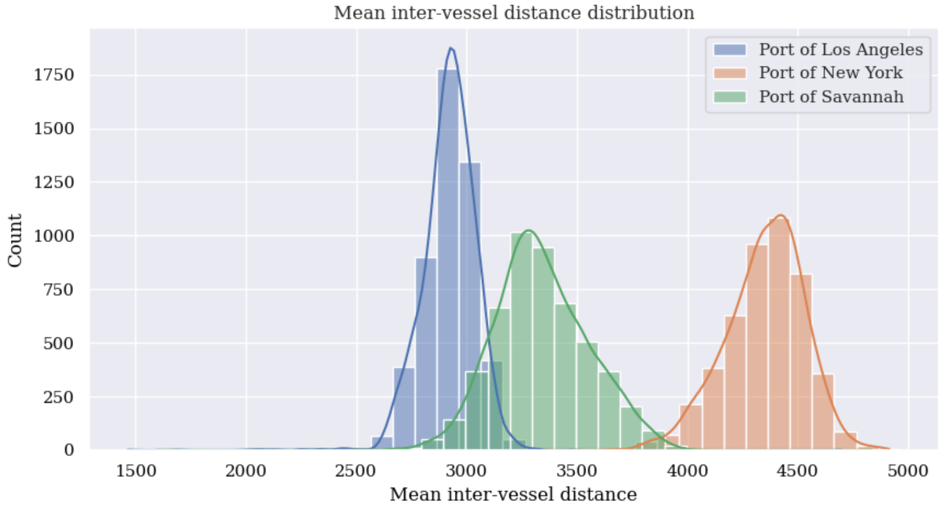


Figure 5.6: Mean spatial density (day of the week)

5.2.3 Moored-to-anchored ratio analysis

The moored-to-anchored ratio provides an indication of how many ships are waiting at anchorage, compared to the number of ships that are currently docked and unloading/loading cargo at the port. A high value, therefore, means that many vessels are waiting to be served.

Figure 5.7 shows the mean moored-to-anchored ratios across the weeks of the second half of 2021. It reveals that, for the port of Savannah, the poor anchorage detection (figure 5.3), resulted in less variance and a significantly lower magnitude than the other two ports. However, we believe the metric has been very effective at capturing congestion for the port of New York and Los Angeles. For the port of New York the spike between week 35 and 40 (Sept 2 - Oct 4) coincides with the Daily Mail reporting a historic number of ships at anchorage [67]. And for the port of Los Angeles the spike in mid October corresponds with retailers upping their supplies in preparation for the holiday season [4].

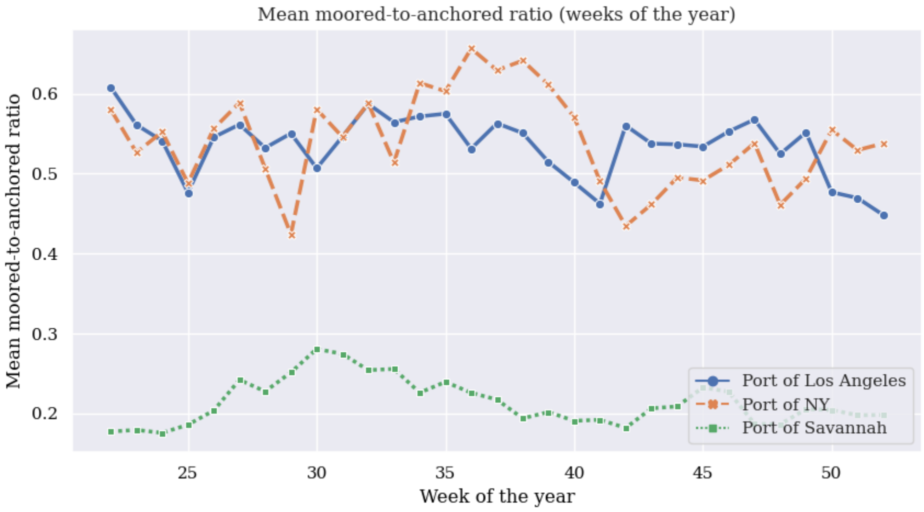
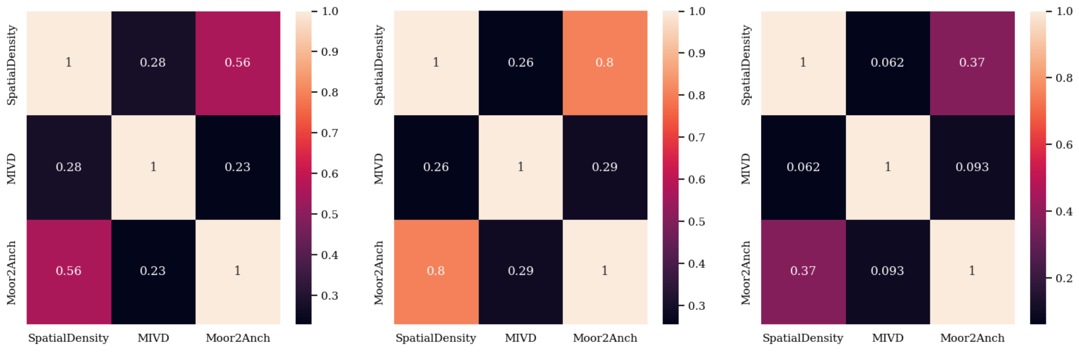


Figure 5.7: Mean moored-to-anchored ration (across weeks of the year)

5.2.4 Correlation

To visualise the Pearson's correlation coefficient between the metrics for each port we used a heatmap. As expected, there was a strong positive correlation between spatial density and moored-to-anchored ratio, especially in the Ports of Los Angeles and New York. This indicated that, as the port became busier, more vessels had to wait at anchorage. This correlation was weaker (0.37) in the Port of Savannah which was most likely a result of the poor anchorage detection as shown in figure 5.3.

We also observed a weak positive correlation between MIVD and the two other measures, most notably in the Port of Los Angeles and New York. Since, in our definition of MIVD, we took the inverse of the distances, this meant that, as the average proximity between vessels decreased, the port became more spatially dense and had more vessels waiting at anchorage. We did not expect this to be a strong correlation because, as explained in section 4.2.3, a port can have a very high MIVD but be almost empty.



Los Angeles New York Savannah

Figure 5.8: Pearson's correlation coefficent between each metric pair

5.3 Congestion prediction evaluation

This section outlines our findings with regard to the effectiveness of three time-series forecasting models: Simple Exponential Smoothing (SES), Long Short-Term Memory (LSTM) and the Proximate Vessel LSTM in predicting the three congestion metrics. The parameters of models were tuned to predict the three port congestion metrics with a focus on optimising their generalisation. The results and analysis are presented with respect to the train and test sets of each port.

5.3.1 Simple Exponential Smoothing (SES)

During tuning of the Simple Exponential Smoothing (SES) model only one parameter can be changed, the smoothing parameter, α. We experimented with varying values of α as well as let the `statsmodels` package find an optimised value for the test set automatically. In table 5.2 we report the results on the train and test set.

Smooth. (α)	**Port**					
	Los Angeles		*New York*		*Savannah*	
	Train MSE†	**Test MSE**†	**Train MSE**†	**Test MSE**†	**Train MSE**†	**Test MSE**†
0.2	0.38	**0.89**	0.80	2.12	0.47	**0.52**
0.8	0.30	1.17	0.54	1.87	0.47	0.67
Optimised*	**0.28**	1.11	**0.54**	**1.87**	**0.44**	0.58

Table 5.2: Evaluation of SES model at different smoothing values, best results shown in **bold**.*Optimised values = 0.85, 0.81, 0.53.†All loss values $\times 10^{-2}$

The results firstly showed that the SES model had varying levels of success across the ports. The Port of New York had particularly high MSE, likely due to the high variance in the metrics, relative to Los Angeles and Savannah. These two ports also performed best when setting α to a lower value (0.2). A low value of alpha indicated that the model placed more weight on past observations, and less weight on the most recent observation, meaning that the underlying trend was more stable, when compared to the Port of New York which benefits ted from a high α, indicating more volatility. Figure 5.9, shows that this volatility mainly came from the MIVD and Moored-to-anchored ratio congestion metrics, where the straight line prediction struggled to keep up with rapid changes in the metrics. The results also showed that the test set optimised smoothing value did not always produce the lowest test MSE.

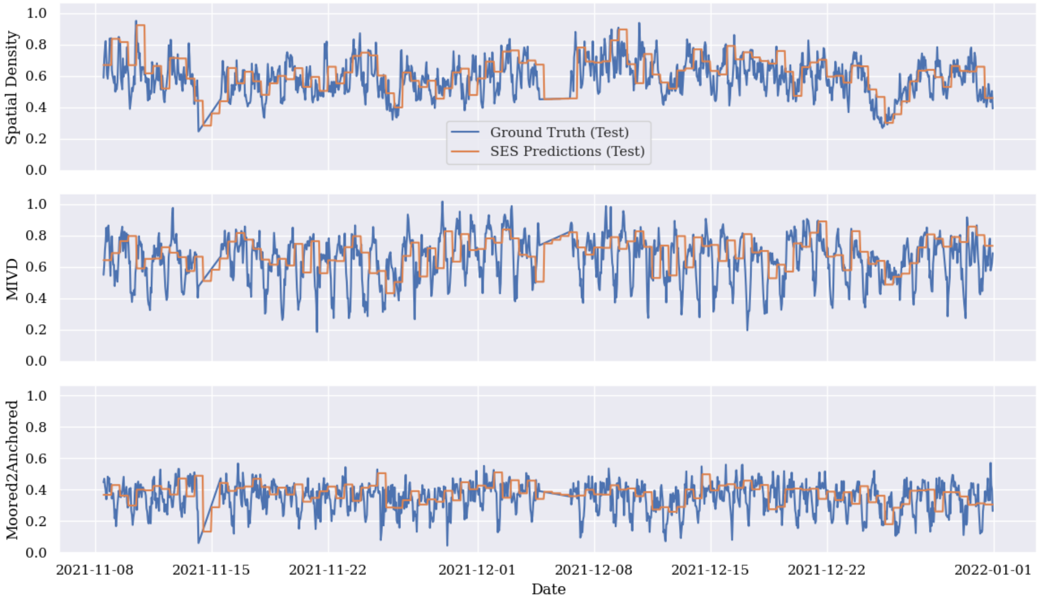


Figure 5.9: **Port of New York** SES test predictions ($\alpha = 0.81$)

5.3.2 Timeseries LSTM

Tuning an LSTM model presents a challenge due to the vast search space of possible hyperparameter combinations. Furthermore, establishing a clear relationship between hyperparameters and model performance can be elusive, further complicating the search for optimal settings. To ease this process and improve the clarity of the results, we kept certain hyperparameters, including the optimiser (Adam), learning rate (0.01), full batching, and the number of hidden units per layer (64), consistent across all models for each port.

We were left with two remaining parameters to tune: the number of hidden layers and the sequence length. These adjustments alter the complexity of the model and the extent to which it can leverage past information to inform future decisions. Below, we present the results of our tuning process for the Port of New York in Table 5.3. In Appendix A, we also include the results for the Port of Los Angeles (Table A.1) and the Port of Savannah (Table A.2). We trained each model for 200 epochs and report the lowest MSE achieved on the test set, as well as the corresponding MSE for the training set at that epoch.

Seq. length (hours)	# Hidden Layers	Train MSE $(\times 10^{-2})$	Test MSE $(\times 10^{-2})$
12	1	0.98	1.48
	2	0.95	1.53
	3	1.02	1.51
24	1	0.94	***1.03***
	2	0.94	1.04
	3	***0.93***	1.04
48	1	0.99	1.16
	2	1.00	1.15
	3	0.99	1.13

Table 5.3:
Port of New York, LSTM Tuning Sequence length and # Hidden Layers (best results show in ***bold and italics***)

In general the results showed that models with 1 or 2 layers and a sequence length of 24 hours performed the best. This indicated that the LSTM started to overfit with 3 layers. This observation was supported when evaluating the loss at the final epoch for the Port of New York. Specifically, the 3 layer models achieved the lowest training set loss at epoch 200 when compared to all other models for that port. This suggested that, although the increased capacity of the 3 layer models benefited the training set, it did not generalise well to the test set. Interestingly, doubling the model's historical context from 24 to 48 hours, did not prove to be beneficial. This could be because the added historical context resulted in more noise and less relevant information for the model to learn from.

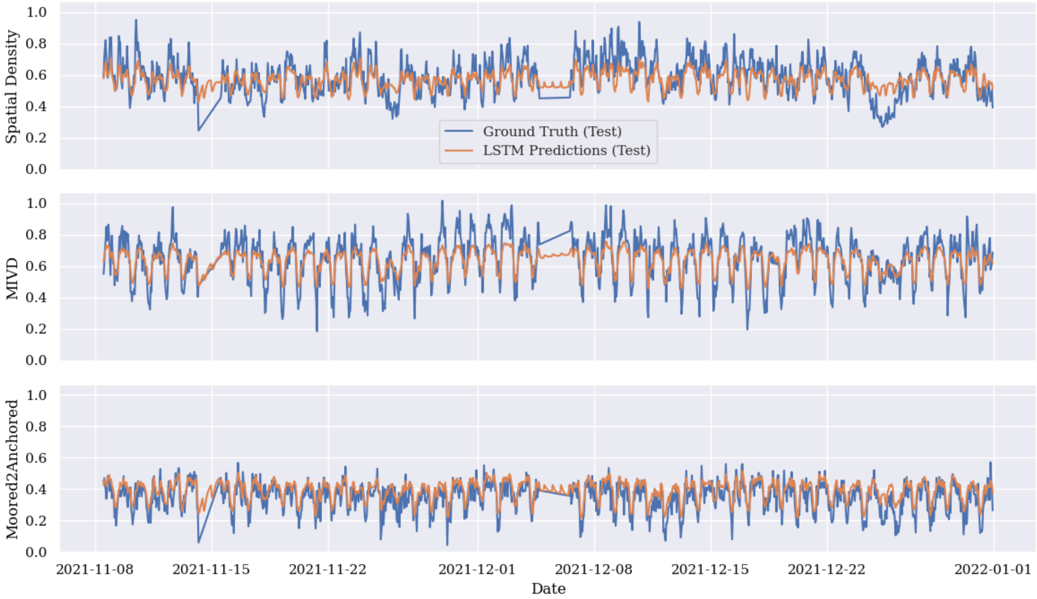


Figure 5.10: **Port of New York** LSTM test predictions (# Hidden layers = 1, sequence length = 24hrs)

Figure 5.10, shows that predictions on the test set of the best performing time series LSTM model. We noticed that it did a particularly good job of capturing the erratic movement of the moored-to-anchored ratio metric, but struggled to capture sudden dips in the spatial density metric.

Table 5.4 is a summary of the hyperparameter tuning process and shows the results for each each port

Port	**Seq. Length**	**# Hidden Layers**	**Train MSE** $(\times 10^{-2})$	**Test MSE** $(\times 10^{-2})$
Los Angeles	24	2	0.35	0.67
New York	24	1	0.94	1.03
Savannah	12	1	0.33	0.38

Table 5.4: Summary of **timeseries LSTM** hyperparameter tunning

5.3.3 Proximate Vessel LSTM

Arrival time prediction model The aim of this model is to predict the arrival time of vessels within a fixed radius of the port. We found that extensive hyperparameter tuning was not needed for this model and that a fairly simple model worked well. We believe this to be because our very limited radius (25km) reduced the training and test set to only a small subset of paths that vessels could take to reach the ports. The small radius also meant that the overwhelming majority of the vessels were very near the port. In fact, over 60% of the vessels (for every port) were found to be less than 1 hour away.

In table 5.5 we show the RMSE of an LSTM, with 64 hidden units and two hidden layers that was trained for 100 epochs. Overall, the model for each port achieved a RMSE of less than 1 hour. It was particularly interesting to note that the port of Savannah achieved the lowest error. We suspect that this was because of it being more inland than the other two and, thus, most vessels in the radius would be following a restricted water way, rather than navigating open ocean making the arrival time easier to predict. Our focus was, however, on the downstream task of congestion prediction.

Port	**Train RMSE (hours)**	**Test RMSE (hours)**
Los Angeles	0.41	0.62
New York	0.33	0.47
Savannah	0.29	0.36

Table 5.5: Arrival time LSTM results (# hidden layers = 2, # hidden units = 64)

To ensure a valid comparison and isolate the impact of the vessel arrival information, we maintained consistency in the architecture of the timeseries LSTM and proximate vessel LSTM models, and used the same random seed. The sole variation between the two models was an increase in input size from three to four. We selected the model that delivered the most promising outcomes on the test set, as indicated in table 5.5.

Table 5.6 shows the result of adding the arrival time feature to the model. We observed an improvement in test MSE for every port, suggesting that adding information about future arrival times was beneficial to congestion prediction. The improvement, we believe, is not necessarily limited by the accuracy of the feature extraction model but rather by the area considered. The port of New York sees the least benefit from this technique likely because it is so extensive and the radius of vessels considered does not capture as many vessels that are further away.

Port	Seq. Length	# Hidden Layers	Train MSE†	Test MSE†	Improvement*
Los Angeles	24	2	0.33	0.51	24%
New York	24	1	0.72	0.87	16%
Savannah	12	1	0.26	0.31	18 %

Table 5.6: **Proximate vessel LSTM** results. *Improvement is the decrease in test loss between this table and table 5.5.†All loss values $\times 10^{-2}$

Additionally, we compared the predictions on the test of the timeseries LSTM (figure 5.10) and proximate vessel LSTM for the port of New York (figure 5.11). We observed a significant improvement in predicting the magnitude of the metric. This is likely because even if the feature extractor model predicted an arrival time incorrectly by a few hours, the timeseries model could still leverage the fact that that vessel was going to arrive within the inference period and thus would have a better idea of the overall ”fullness” of the port.

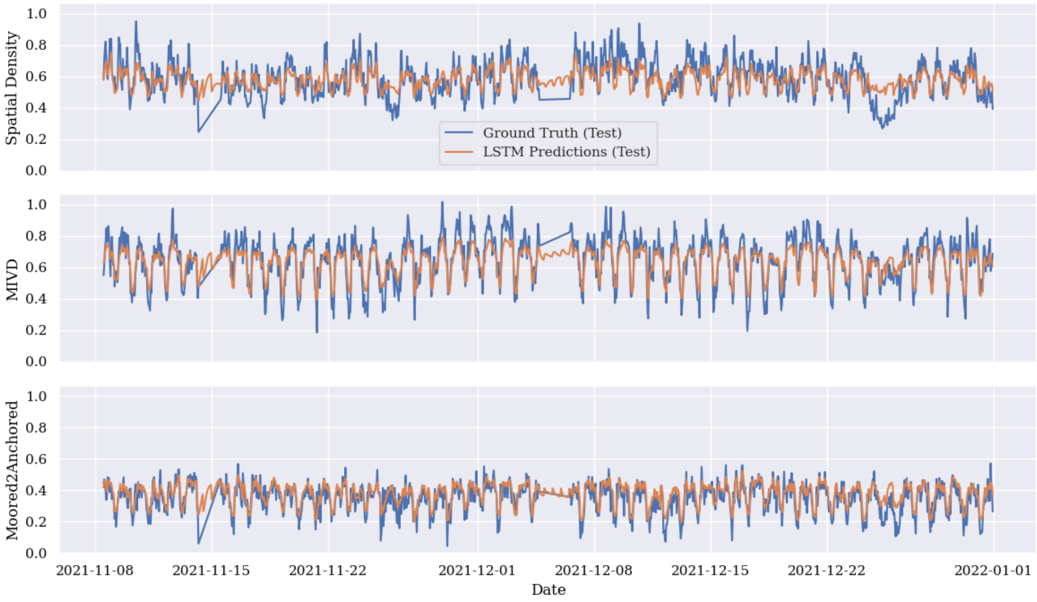


Figure 5.11: **Port of New York** Proximate LSTM test predictions (# Hidden layers = 1, sequence length = 24hrs)

Chapter 6

Conclusion

In our work we explored the challenges and complexities of applying machine learning and data analysis techniques to the maritime use case of port congestion.

It is clear that leveraging Automatic Identification System (AIS) data to detect and quantify congestion in maritime ports has the potential to provide valuable insights for the shipping industry and the broader economy. Through our work, we have shown that it is possible to detect the regions that encompass the ports of Los Angeles, New York, and Savannah in an unsupervised manner and quantify congestion using three spatio-temporal metrics: spatial density, mean inter-vessel distance (MIVD), and moored-to-anchored ratio. We also showed that by linking the capacity of moorage and anchorage areas with our novel metric, moored-to-anchored ratio, that it is possible to capture a poignant aspect of congestion

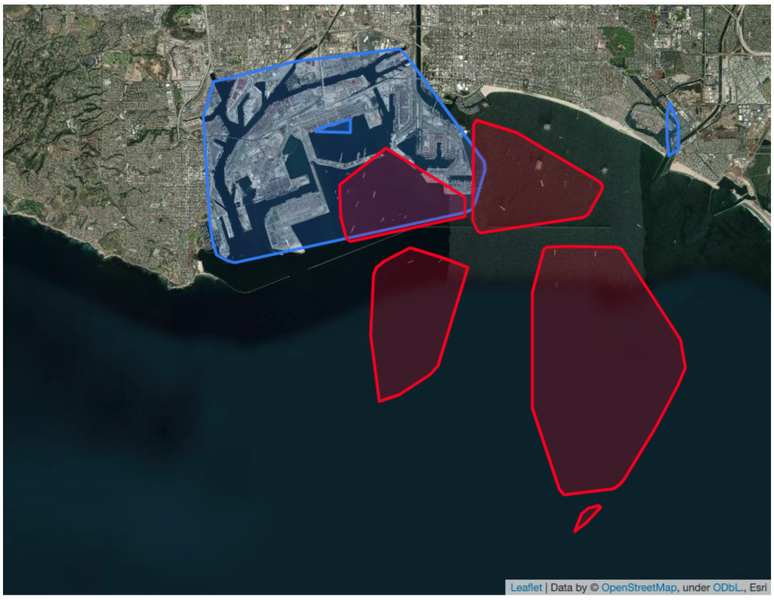
Furthermore, we demonstrated that challenges of forecasting these metrics using only historical values, even when using a suitably complex model (LSTM). Subsequently, we showed that by using a complimentary feature extraction model that predicted the arrival time of vessels near the target port, that it was possible to provide additional context to a timeseries model and decrease error by on average around 20%.

Our work represents one of the first studies to tackle both congestion quantification and congestion prediction tasks using publicly available data. By broadening our understanding of port congestion and providing an improved means of predicting it, we believe that our work can help vessel operators avoid congested ports, reduce waiting times in ports, and increase overall efficiency. We hope that our study will open the door for more transparent and reproducible work in the field and inspire further research on the use of machine learning techniques for addressing complex maritime planning problems.

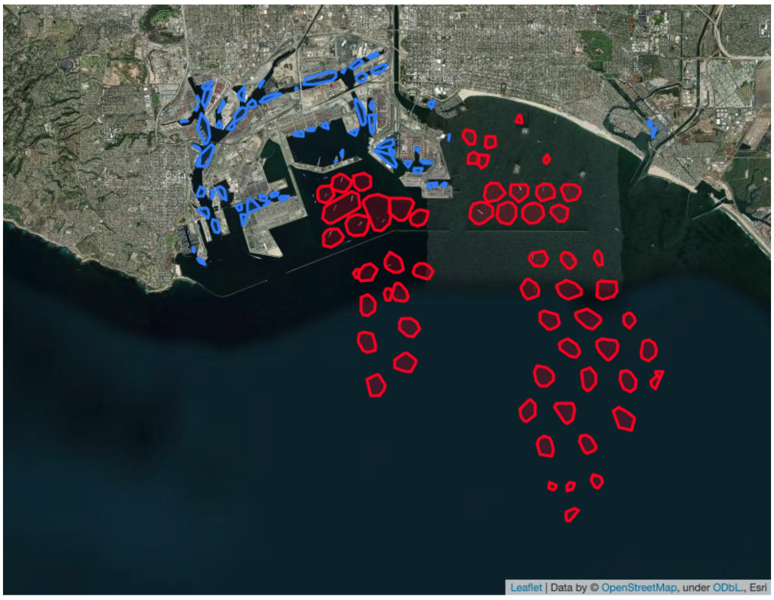
Appendix A

Results

A.1 DBSCAN parameter tuning visualisation



Course grain cluster (high ε)



Fine grained clusters (low ε)

Figure A.1: DBSCAN hyperparameter tuning example on the Port of Los Angeles

A.2 Timeseries LSTM hyperparameter tuning

Sequence length (hours)	# Hidden Layers	Train MSE $(\times 10^{-2})$	Test MSE $(\times 10^{-2})$
12	1	0.64	0.78
	2	0.65	0.80
	3	0.67	0.80
24	1	0.41	0.70
	2	0.35	***0.67***
	3	***0.33***	0.73
48	1	0.40	0.83
	2	0.39	0.70
	3	0.34	0.86

Table A.1: **Port of Los Angeles**, LSTM Tuning Sequence length and # Hidden Layers (best results show in ***bold and italics***)

Sequence length (hours)	# Hidden Layers	Train MSE $(\times 10^{-2})$	Test MSE $(\times 10^{-2})$
12	1	0.33	***0.38***
	2	0.31	0.42
	3	0.28	0.45
24	1	0.28	0.41
	2	***0.24***	0.43
	3	0.28	0.46
48	1	0.33	0.54
	2	0.31	0.58
	3	0.30	0.58

Table A.2: **Port of Savannah**, LSTM Tuning Sequence length and # Hidden Layers (best results show in ***bold and italics***)

Bibliography

[1] Ibrahim AbuAlhaol et al. “Mining Port Congestion Indicators from Big AIS Data”. In: *2018 International Joint Conference on Neural Networks (IJCNN)*. ISSN: 2161-4407. July 2018, pp. 1–8. DOI: 10.1109/IJCNN.2018.8489187.

[2] *AIS (Automatic Identification System) Overview*. en. URL: https://shipping.nato.int/nsc/operations/news/2021/ais-automatic-identification-system-overview.aspx (visited on 10/17/2022).

[3] *AIS Fundamentals*. en. URL: https://documentation.spire.com/ais-fundamentals/ (visited on 03/19/2023).

[4] Dani Anguiano. “‘Like a freeway in traffic’: America’s busiest ports choked by a pandemic holiday”. en-GB. In: *The Guardian* (Dec. 2021). ISSN: 0261-3077. URL: https://www.theguardian.com/business/2021/dec/21/inside-americas-busiest-port-during-holidays (visited on 04/12/2023).

[5] Mohsen Attaran and Promita Deb. “Machine Learning: The New ’Big Thing’ for Competitive Advantage”. In: 5 (Jan. 2018), pp. 277–305. DOI: 10.1504/IJKEDM.2018.10015621.

[6] Xiwen Bai, Haiying Jia, and Mingqi Xu. “Port congestion and the economics of LPG seaborne transportation”. In: *Maritime Policy & Management* 0.0 (June 2021). Publisher: Routledge _eprint: https://doi.org/10.1080/03088839.2021.1940334, pp. 1–17. ISSN: 0308-8839. DOI: 10.1080/03088839.2021.1940334. URL: https://doi.org/10.1080/03088839.2021.1940334 (visited on 10/18/2022).

[7] Oleh Bodunov et al. “Real-time Destination and ETA Prediction for Maritime Traffic”. In: *Proceedings of the 12th ACM International Conference on Distributed and Event-based Systems*. DEBS ’18. New York, NY, USA: Association for Computing Machinery, June 2018, pp. 198–201. ISBN: 978-1-4503-5782-1. DOI: 10.1145/3210284.3220502. URL: https://doi.org/10.1145/3210284.3220502 (visited on 02/21/2023).

[8] Berit Brouer, Christian Vad Karsten, and David Pisinger. “Big Data Optimization in Maritime Logistics”. In: May 2016, pp. 319–344. ISBN: 978-3-319-30263-8. DOI: 10.1007/978-3-319-30265-2_14.

[9] Zihan Chang, Yang Zhang, and Wenbo Chen. “Effective Adam-Optimized LSTM Neural Network for Electricity Price Forecasting”. In: *2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS)*. ISSN: 2327-0594. Nov. 2018, pp. 245–248. DOI: 10.1109/ICSESS.2018.8663710.

[10] Marielle Christiansen et al. “Chapter 4 Maritime Transportation”. en. In: *Handbooks in Operations Research and Management Science*. Ed. by Cynthia Barnhart and Gilbert Laporte. Vol. 14. Transportation. Elsevier, Jan. 2007, pp. 189–284.

DOI: 10.1016/S0927-0507(06)14004-9. URL: https://www.sciencedirect.com/science/article/pii/S0927050706140049 (visited on 10/21/2022).

[11] "Commodities boom sends bulk shipping costs to decade highs". In: *Financial Times* (May 2021).

[12] *Competitors are eating into L.A. ports' dominance*. en-US. Section: Business. Apr. 2016. URL: https://www.latimes.com/business/la-fi-la-ports-competition-20160427-story.html (visited on 03/30/2023).

[13] *Container Facts —Costamare IR*. URL: https://www.costamare.com/industry_containerisation (visited on 04/07/2023).

[14] *Convex hull*. en. Page Version ID: 1138722281. Feb. 2023. URL: https://en.wikipedia.org/w/index.php?title=Convex_hull&oldid=1138722281 (visited on 04/12/2023).

[15] Madhuri Debnath, Praveen Kumar Tripathi, and Ramez Elmasri. "K-DBSCAN: Identifying Spatial Clusters with Differing Density Levels". en. In: *2015 International Workshop on Data Mining with Industrial Applications (DMIA)*. San Lorenzo, Central, Paraguay: IEEE, Sept. 2015, pp. 51–60. ISBN: 978-1-4673-8111-6. DOI: 10.1109/DMIA.2015.14. URL: http://ieeexplore.ieee.org/document/7544972/ (visited on 03/30/2023).

[16] Ria Dutta. *Top 10 busiest ports in the US: [+get the best leasing rates]*. en-US. Feb. 2023. URL: https://www.container-xchange.com/blog/busiest-ports-in-the-us/ (visited on 03/29/2023).

[17] Martin Ester et al. "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise". en. In: ().

[18] *Frequently Asked Questions — Bureau of Ocean Energy Management*. URL: https://www.boem.gov/frequently-asked-questions (visited on 03/19/2023).

[19] Alex Graves. *Supervised Sequence Labelling with Recurrent Neural Networks*. en. Vol. 385. Studies in Computational Intelligence. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012. ISBN: 978-3-642-24796-5 978-3-642-24797-2. DOI: 10.1007/978-3-642-24797-2. URL: https://link.springer.com/10.1007/978-3-642-24797-2 (visited on 04/04/2023).

[20] Sepp Hochreiter and Jürgen Schmidhuber. "Long Short-term Memory". In: *Neural computation* 9 (Dec. 1997), pp. 1735–80. DOI: 10.1162/neco.1997.9.8.1735.

[21] Chong-Wei Huang and Tian-Yuan Shih. "On the complexity of point-in-polygon algorithms". en. In: *Computers & Geosciences* 23.1 (Feb. 1997), pp. 109–118. ISSN: 00983004. DOI: 10.1016/S0098-3004(96)00071-4. URL: https://linkinghub.elsevier.com/retrieve/pii/S0098300496000714 (visited on 03/21/2023).

[22] Shuai Jia, Qiang Meng, and Haibo Kuang. "Equitable Vessel Traffic Scheduling in a Seaport". In: *Transportation Science* 56 (Jan. 2022). DOI: 10.1287/trsc.2021.1076.

[23] Andras Komaromi, Diego Cerdeiro, and Yang Liu. *Supply Chains and Port Congestion Around the World*. en. SSRN Scholarly Paper. Rochester, NY, Mar. 2022. URL: https://papers.ssrn.com/abstract=4076355 (visited on 10/18/2022).

[24] "LA port to open round the clock to tackle shipping queues". en-GB. In: *BBC News* (Oct. 2021). URL: https://www.bbc.com/news/business-58901777 (visited on 03/29/2023).

[25] Philipp Last et al. "Comprehensive Analysis of Automatic Identification System (AIS) Data in Regard to Vessel Movement Prediction". en. In: *The Journal of Navigation* 67.5 (Sept. 2014). Publisher: Cambridge University Press, pp. 791–809. ISSN: 0373-4633, 1469-7785. DOI: 10.1017/S0373463314000253. URL: https://www.cambridge.org/core/journals/journal-of-navigation/article/comprehensive-analysis-of-automatic-identification-system-ais-data-in-regard-to-vessel-movement-prediction/95E9218CA2796FEF05A216E4F2376B9D (visited on 04/12/2023).

[26] Ting-Peng Liang and Yu-Hsi Liu. "Research Landscape of Business Intelligence and Big Data analytics: A bibliometrics study". en. In: *Expert Systems with Applications*. Big Data Analytics for Business Intelligence 111 (Nov. 2018), pp. 2–10. ISSN: 0957-4174. DOI: 10.1016/j.eswa.2018.05.018. URL: https://www.sciencedirect.com/science/article/pii/S0957417418303099 (visited on 10/17/2022).

[27] Bryan Lim and Stefan Zohren. "Time-series forecasting with deep learning: a survey". EN. In: *Philosophical Transactions of the Royal Society A* (Apr. 2021). Publisher: The Royal Society Publishing. DOI: 10.1098/rsta.2020.0209. URL: https://royalsocietypublishing.org/doi/10.1098/rsta.2020.0209 (visited on 10/20/2022).

[28] Hagar Mahmoud and Nadine Akkari. "Shortest Path Calculation: A Comparative Study for Location-Based Recommender System". In: Mar. 2016, pp. 1–5. DOI: 10.1109/WSCAR.2016.16.

[29] *MarineCadastre.gov*. URL: https://marinecadastre.gov/about/ (visited on 03/19/2023).

[30] *Monitoring, reporting and verification of EU ETS emissions*. en. URL: https://climate.ec.europa.eu/eu-action/eu-emissions-trading-system-eu-ets/monitoring-reporting-and-verification-eu-ets-emissions_en (visited on 04/10/2023).

[31] *Moving average and exponential smoothing models*. URL: https://people.duke.edu/~rnau/411avg.htm (visited on 04/11/2023).

[32] MM Mukaka. "A guide to appropriate use of Correlation coefficient in medical research". In: *Malawi Medical Journal : The Journal of Medical Association of Malawi* 24.3 (Sept. 2012), pp. 69–71. ISSN: 1995-7262. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3576830/ (visited on 04/12/2023).

[33] Ziaul Haque Munim et al. "Big data and artificial intelligence in the maritime industry: a bibliometric review and future research directions". In: *Maritime Policy & Management* 47.5 (July 2020). Publisher: Routledge _eprint: https://doi.org/10.1080/03088839.2020.1788731, pp. 577–597. ISSN: 0308-8839. DOI: 10.1080/03088839.2020.1788731. URL: https://doi.org/10.1080/03088839.2020.1788731 (visited on 10/17/2022).

[34] Nazri Mohd Nawi, Walid Hasen Atomi, and M. Z. Rehman. "The Effect of Data Pre-processing on Optimized Training of Artificial Neural Networks". en. In: *Procedia Technology*. 4th International Conference on Electrical Engineering

and Informatics, ICEEI 2013 11 (Jan. 2013), pp. 32–39. ISSN: 2212-0173. DOI: 10.1016/j.protcy.2013.12.159. URL: https://www.sciencedirect.com/science/article/pii/S2212017313003137 (visited on 03/25/2023).

[35] Duong Nguyen et al. “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)*. arXiv:1806.03972 [cs, stat]. Oct. 2018, pp. 331–340. DOI: 10.1109/DSAA.2018.00044. URL: http://arxiv.org/abs/1806.03972 (visited on 04/10/2023).

[36] Noaa Nos Ocm and Jesse Brass. “Frequently Asked Questions: AIS Data and Tools”. en. In: ().

[37] *Overview of AIS dataset - AIS Handbook - UN Statistics Wiki*. URL: https://unstats.un.org/wiki/display/AIS/Overview+of+AIS+dataset (visited on 03/26/2023).

[38] Giovanni Pantuso, Kjetil Fagerholt, and Lars Magnus Hvattum. “A survey on maritime fleet size and mix problems”. en. In: *European Journal of Operational Research*. Maritime Logistics 235.2 (June 2014), pp. 341–349. ISSN: 0377-2217. DOI: 10.1016/j.ejor.2013.04.058. URL: https://www.sciencedirect.com/science/article/pii/S0377221713003846 (visited on 10/17/2022).

[39] Spyridon Patmanidis et al. “Maritime surveillance, vessel route estimation and alerts using AIS data”. In: *2016 24th Mediterranean Conference on Control and Automation (MED)*. June 2016, pp. 809–813. DOI: 10.1109/MED.2016.7535966.

[40] Wenhao Peng et al. “A deep learning approach for port congestion estimation and prediction”. In: *Maritime Policy & Management* 0.0 (Apr. 2022). Publisher: Routledge _eprint: https://doi.org/10.1080/03088839.2022.2057608, pp. 1–26. ISSN: 0308-8839. DOI: 10.1080/03088839.2022.2057608. URL: https://doi.org/10.1080/03088839.2022.2057608 (visited on 10/21/2022).

[41] *pmdarima: ARIMA estimators for Python — pmdarima 2.0.3 documentation*. URL: http://alkaline-ml.com/pmdarima/ (visited on 04/11/2023).

[42] *Point in polygon - GIS Wiki — The GIS Encyclopedia*. URL: http://wiki.gis.com/wiki/index.php/Point_in_polygon (visited on 03/21/2023).

[43] *Port Congestion: Meaning, Causes, Solutions & More*. en-US. Section: Ship Tracking. Jan. 2020. URL: https://www.marinetraffic.com/blog/port-congestion-explained/ (visited on 03/13/2023).

[44] *Port of Los Angeles*. en. Page Version ID: 1144878189. Mar. 2023. URL: https://en.wikipedia.org/w/index.php?title=Port_of_Los_Angeles&oldid=1144878189 (visited on 03/30/2023).

[45] *Port of New York and New Jersey*. en. Page Version ID: 1149451742. Apr. 2023. URL: https://en.wikipedia.org/w/index.php?title=Port_of_New_York_and_New_Jersey&oldid=1149451742 (visited on 04/12/2023).

[46] “Review of Maritime Transport 2020”. en. In: *REVIEW OF MARITIME TRANSPORT* (2020).

[47] *Routine losses: ships spend up to 9 percent time at anchorage*. en. URL: https://www.wartsila.com/insights/article/routine-losses-ships-spend-up-to-9-percent-time-at-anchorage (visited on 04/07/2023).

[48] Pedro-Luis Sanchez-Gonzalez et al. "Toward Digitalization of Maritime Transport?" en. In: *Sensors* 19.4 (Jan. 2019). Number: 4 Publisher: Multidisciplinary Digital Publishing Institute, p. 926. ISSN: 1424-8220. DOI: `10.3390/s19040926`. URL: `https://www.mdpi.com/1424-8220/19/4/926` (visited on 03/13/2023).

[49] Ola Sekkesaeter. "Shipping in the digital age: how feasible is the application of big data to the maritime shipping industry, and under what conditions can it be developed to become an integral part of its future?" eng. PhD thesis. University of Geneva, 2017. URL: `https://archive-ouverte.unige.ch/unige:95049` (visited on 10/17/2022).

[50] Vitali Serafimov, Oleg Stets, and Andriy Shkolyk. "Seaports in the BRI: Challenges, Solutions and Emerging Regulations". eng. In: *Lex Portus* 7.5 (2021), pp. 14–41. URL: `https://heinonline.org/HOL/P?h=hein.journals/lxportus27&i=292` (visited on 04/12/2023).

[51] Priya Singh. *Dawn of AI in Shipping Industry — BOXXPORT*. en-GB. Jan. 2022. URL: `https://blog.boxxport.com/2022/01/21/dawn-of-ai-in-shipping-industry/` (visited on 03/13/2023).

[52] Matthias Steidel et al. *Correcting the Destination Information in Automatic Identification System Messages*. June 2019.

[53] *Target Leakage*. en-US. URL: `https://www.datarobot.com/wiki/target-leakage/` (visited on 04/12/2023).

[54] Ashutosh Tripathi. *What is the main difference between RNN and LSTM — NLP — RNN vs LSTM*. en. July 2021. URL: `https://ashutoshtripathi.com/2021/07/02/what-is-the-main-difference-between-rnn-and-lstm-nlp-rnn-vs-lstm/` (visited on 10/20/2022).

[55] *UN/LOCODE — UNECE*. URL: `https://unece.org/trade/uncefact/unlocode` (visited on 02/21/2023).

[56] *Understanding ARIMA Models for Machine Learning*. en. URL: `https://www.capitalone.com/tech/machine-learning/understanding-arima-models/` (visited on 10/21/2022).

[57] *Understanding the COVID-19 Effect on Ecommerce + Trends*. en-us. Nov. 2021. URL: `https://www.bigcommerce.com/blog/covid-19-ecommerce/` (visited on 03/20/2023).

[58] Vaib. *Why can't ARIMA model large lags and/or long range dependence?* Forum post. May 2018. URL: `https://stats.stackexchange.com/q/294881` (visited on 04/11/2023).

[59] Ashish Vaswani et al. *Attention Is All You Need*. arXiv:1706.03762 [cs]. Dec. 2017. DOI: `10.48550/arXiv.1706.03762`. URL: `http://arxiv.org/abs/1706.03762` (visited on 04/10/2023).

[60] Dmitry Shafran I. worked as an officer in the deck department on various types of vessels et al. *What Is Berthing And Unberthing Of A Ship? - Maritime Page*. en-us. Section: Nautical Science. Aug. 2022. URL: `https://maritimepage.com/what-is-berthing-and-unberthing-of-a-ship/` (visited on 04/07/2023).

[61] Qingsong Wen et al. *Transformers in Time Series: A Survey*. arXiv:2202.07125 [cs, eess, stat]. Mar. 2022. DOI: `10.48550/arXiv.2202.07125`. URL: `http://arxiv.org/abs/2202.07125` (visited on 10/20/2022).

[62] *What is AIS & How Does It Work? - Marine Radio Articles - Icom UK*. URL: https://icomuk.co.uk/What-is-AIS-and-How-Does-It-Work/3995/165/ (visited on 10/18/2022).

[63] Zhixian Yan. "Traj-ARIMA: A Spatial-Time Series Model for Network-Constrained Trajectory". In: (Jan. 2010). DOI: 10.1145/1899441.1899446.

[64] Cheng-Hong Yang et al. "AIS-Based Intelligent Vessel Trajectory Prediction Using Bi-LSTM". In: *IEEE Access* 10 (2022). Conference Name: IEEE Access, pp. 24302–24315. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2022.3154812.

[65] Cheng-Hong Yang et al. "Deep Learning for Vessel Trajectory Prediction Using Clustered AIS Data". en. In: *Mathematics* 10.16 (Jan. 2022). Number: 16 Publisher: Multidisciplinary Digital Publishing Institute, p. 2936. ISSN: 2227-7390. DOI: 10.3390/math10162936. URL: https://www.mdpi.com/2227-7390/10/16/2936 (visited on 10/20/2022).

[66] Ibna Zaman et al. "Challenges and Opportunities of Big Data Analytics for Upcoming Regulations and Future Transformation of the Shipping Industry". en. In: *Procedia Engineering*. 10th International Conference on Marine Technology, MARTEC 2016 194 (Jan. 2017), pp. 537–544. ISSN: 1877-7058. DOI: 10.1016/j.proeng.2017.08.182. URL: https://www.sciencedirect.com/science/article/pii/S1877705817333386 (visited on 04/10/2023).

[67] Ariel Zilber. *Dozens of ships off coast of New York wait to dock in port*. Section: News. Sept. 2021. URL: https://www.dailymail.co.uk/news/article-10029043/Dozens-ships-coast-New-York-wait-dock-countrys-second-largest-port.html (visited on 04/12/2023).