Contents

[Background 2](#_Toc98273963)

[Problem Statement and Research objectives 3](#_Toc98273964)

[Literature Review 4](#_Toc98273965)

[Data preprocessing 7](#_Toc98273966)

[Present the summary statistics of raw data. 7](#_Toc98273967)

[Data pre-processing solutions and the justification of your final choice of solution 10](#_Toc98273968)

[References 12](#_Toc98273969)

# Background

Since over 5000 years ago, sea transportation has always been a major supporter of global trade (NAMEPA, 2018). Today, seaborne trade remains as the backbone of international trade and global economy. According to the Organisation for Economic Co-operation and Development (OECD), marine transport accounts for over 90% of worldwide trade volume (OECD, 2021). Maritime traffic is predicted to rise substantially over the future decades, propelled by increasing demand for primary resources and container shipping, despite a reduction in shipping activity in 2020 due the consequences of the COVID-19 pandemic. Therefore, maritime transport can have a significant impact on transport costs and trading prices.

The maritime trading environment can be complex as there are wide range of cargoes which can be transported via different types of vessels. Vessels can be classified into five major groups based on the types of cargoes they are designed to carry, namely oil tankers, gas carriers, bulk carriers, general cargo ships, and container ships. They can then be further segmented based on their sizes. Statistics from the United Nations Conference on Trade and Development (UNCTAD) show that bulk carriers represent nearly 43% of the total Deadweight Tonnage (DWT) of the total vessel capacity in 2021. This is equivalent to around 17% of total trade values (UNCTAD, 2021). In contrast to the familiar container ships and container-based cargo, which we are familiar with, container vessels only make up of around 13% percent to total DWT and 9% of total trade value.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **World fleet by principal vessel type,**  **2020–2021** | **2020** | | **2021** | | **% Change** |
| **Principal types** | **million DWT** | **%** | **million DWT** | **%** |  |
| Bulk carriers | 880 | 42% | 913 | 43% | 3.8% |
| Oil tankers | 601 | 29% | 619 | 29% | 3.0% |
| Container ships | 275 | 13% | 282 | 13% | 2.5% |
| Other types of ships: | 239 | 12% | 244 | 11% | 2.2% |
| *Offshore supply* | 84 | 4% | 84 | 4% | 0.1% |
| *Gas carriers* | 74 | 4% | 77 | 4% | 5.1% |
| *Chemical tankers* | 47 | 2% | 49 | 2% | 2.9% |
| *Other/not available* | 26 | 1% | 25 | 1% | -0.4% |
| *Ferries and passenger ships* | 8 | 0% | 8 | 0% | 1.5% |
| General cargo ships | 77 | 4% | 77 | 4% | -0.12% |
| **World total** | **2072** |  | **2135** |  | **3.0%** |

Table 1: World fleet by principal vessel type, 2020-2021

Source: UNCTAD calculations, based on data from Clarksons Research

As the largest part of the world’s merchant fleet in terms of DWT capacity, bulk carriers usually carry raw materials that fuel the modern economy. Such raw materials include ores, coal, and grains etc. They can also carry other products used and seen across the various stages of manufacturing processes. These include intermediate products such as steel coils, and final products like industrial machines or train carriages. The wide coverage makes them truly the backbone and workhorse of the maritime transport industry. As such, to narrow down the scope while still capture the gist of the maritime market, our research will mainly be focusing on dry bulk sectors.

Port congestion has always been one of the major problems faced by the industry players. Over the past 2 years, the impact of Covid-19 exacerbated the problem as lockdowns imposed around the world led to manpower shortages for port operations. With lesser workers handling the same number of ships, productivity dropped drastically. This led to longer waiting time for vessels wanting to enter ports for loading or unloading works.

The impact of long waiting time was further manifested by the blocking of the Suez Canal in March 2021 where strong winds caused the ship Ever Given to wedge along the waterway, causing traffic jam at the canal for around 6 days. The blockage caused approximately US$54 billion trade loss, affected not just businesses, but also the end consumers.

A long waiting time at the port can be fatal to all the market players, as the costs of the supply chain will eventually be driven up due to the shortage of dry bulk supply capacity:

* 1. For shipowners, while supply shortage could mean a higher freight rate, their vessels will be locked up in the port, thus affecting the supply of their next voyage.
  2. For cargo owners, the short shelf lives of these products could mean potential value loss of the products, which can be detrimental to their businesses.
  3. For retailers and end consumers, increasing freight rates due to supply shortage will eventually increase import and consumer prices.

# Problem Statement and Research objectives

Our research aims to predict the expected waiting time of a new arrived ship given the current observed port congestion levels of dry bulk carriers and their distribution of carrier sizes. By investigating whether the vessel size and entry date play a part in affecting waiting time at the port, ship operators can better plan their voyage to minimise possible delays, thus prevent unnecessary costs from incurring. The port can also better schedule their port activities and improve on their logistics planning, which can eventually help to alleviate the port congestion problems.

The Port of Santos is selected as our piloting port for our prediction model as it is one of the busiest dry bulk ports in Latin America, with a diverse range of cargoes handling terminals for solid and liquid bulk, containers, and general cargoes. The substantial daily cargo transaction volumes at the port can ensure that there are enough feeding data for building a robust model.

As mentioned previously, we will only focus on dry bulk sector to narrow our scope of analysis. Dry bulk cargo plays a crucial role at the Port of Santos with its significant trading volume. Besides, the Port of Santos plays a crucial role in dry bulk delays due to port congestion. According to Danish Ship Finance, congestions in Brazilian ports during 1Q 2021 is one of the main reasons drove up the number of vessels caught up in port congestion to 5% of the dry bulk fleet, increased from 4% in 4Q 2020 (Danish Ship Finance, 2021).

The data from Refinitiv shows that in 2021, dry bulk carriers had to wait at the Santos anchorage for around one month to enter the Port of Santos to perform loading or unloading work on average.

|  |  |
| --- | --- |
| **Anchorage** | **Average waiting time (days)** |
| Outer Yangtze Estuary Anchorage North | 51 |
| **Santos Anchorage** | **35** |
| Paranagua Outer Anchorage | 28 |
| Tubarao (and Vitoria) Anchorage | 25 |
| Tianjin Northern Anchorage | 24 |
| Rizhao Anchorage | 23 |
| Ponta da Madeira (Itaqui) inner Anchorage | 23 |
| Port Hedland Anchorage | 23 |
| Outer Zhoushan Anchorages | 23 |
| Huanghua Anchorage | 22 |

Table 2: Top 10 average waiting days of dry bulk carriers in 2021

Source: Refinitiv

# Literature Review

Port congestion is a key problem that impacts the performance, productivity, and efficiency of seaports, and is one of the most essential elements for gauging the port performance. Port congestion is caused by a wide range of factors, many of which are complex. Several studies on port congestion have been conducted, both for the purpose of optimising port efficiency and for the examination of policies relating to the expansion of psychical infrastructure, capacity, and modernisation.

Bolat et al., for example, attempted to apply Analytic Hierarchy Process (AHP) method to determine the most crucial elements affecting port congestion based on the information provided by numerous key industry stakeholders, primarily from surveyors’ perspectives (Bolat, Gizem, Emine, Furkan, & Soysal, 2020). According to their research, the most important factors contributing to port congestion are documentation procedures, port operation and administration, ship traffic inputs, port layout and strategy, and government interactions.

Jonquais et al., on the other hand, used supervised Machine Learning and predictive analysis to improve the estimated time of arrival (ETA) for a shipment. Linear Regression, Random Forest and Fuzzy Rule-Based Bayesian Network techniques were deployed to address the fuzzy and incomplete data, as a result of disruptions and congestions at the port. The models aim to reduce delays farther down the supply chain by estimating the transit time and ETA of vessels (Jonquais & Krempel, 2019). The selected features are shown in the table below:

|  |  |
| --- | --- |
| **Features** | **Definitions given by the authors** |
| Average time spent at Port of origin (Port of origin mean) | Average time shipments spend in the port of origin from the moment when they are received by the shipper until the vessel leaves the port, based on the different shipper and port combinations. |
| Standard deviation of time spent at Port of origin (Port of origin SD) | Standard deviation of the average time that shipments spend in the port of origin from the moment when they are received by the shipper until the vessel leaves the port, based on the different shipper and port combinations. |
| Average time spent at Port of destination (Port of destination mean) | Average time shipments spend at the port of destination from the moment when the containership arrives at the port of destination until the containers are discharged at the port, based on the different customer and port combinations. |
| Standard deviation of time spent at Port of destination (Port of destination SD) | Standard deviation of the time that shipments spend at the port of destination from the moment when the containership arrives at the port of destination until the containers are discharged at the port, based on the different customer and port combinations. |
| Average time spent per Route (Route mean) | Average transit time a containership needs to travel from one port to another, based on the different carrier and route combinations. |
| Schedule | A refined route mean: this average also depends on the day of the week the containership left the port of origin. |
| Origin Service | A binary column indicating if the shipment has to be consolidated at the port of origin after it is received by the shipper. |
| Holiday | A binary column indicating if the shipment falls within a two-week period before Chinese New Year. |
| Quarter | Z-score of the transit time for every shipment in the dataset normalized by the carrier route combination. |
| Expected Time to port | The time period between the booking date and the expected time the shipper will receive the shipment at the port of origin, given by the customer. |
| Port of origin late | A binary column indicating if the shipment is received after the expected received date. |
| Port of origin latest | A binary column indicating if the shipment is received after the latest possible received date given by the shipper. |
| Capacity of port of destination | The percentage of total capacity historically occupied at the port of destination in a two-day window around the expected arrival date of the containership. |
| Late departure | A binary column indicating if the containership left the port of origin later than the expected date of departure. |

Table 3: Selected features and their definitions

Source: Predicting Shipping Time with Machine Learning by Antoine Charles Jean Jonquais and Florian Krempl

The study used the same set of performance metrics to compare the results of all three models. These include:

* R-squared
* Mean absolute error
* Mean absolute percentage error
* Root mean squared error

Although the research attempted to apply several methods and models for predictive analysis, the importance scores for the selected features are not specified. Therefore, it is unclear how these features are selected. Furthermore, our study has a different goal: instead of estimating a shipment’s ETA, we are more interested in predicting the waiting time at a port by examining the port congestion situations.

A recent paper by Stepec et al. provides a more comprehensive perspective that is more aligned with our research objective (Stepec, Martincic, Klein, Vladusic, & Costa, 2021). In their paper, Dejan investigated the possibility to predict turnaround time for ships making a call at the Port of Bordeaux. Departing from the classical literature on berth scheduling optimization which are restrictive in their assumptions (Golias, Boile, & Theofanis, 2009), their approach to the issue was to train a gradient boosting machine learning method, CatBoost, using a collection of historical data ranging from data relating the vassals making the port call, activities and duration of the port calls, weather and tidal information in the region, seasonality of the week, congestion within the port, and the local holiday schedule as their input features.

Stepec et al. opted to segment their prediction via cargo type, and their achieved model performance on the CatBoost outperformed slightly as compared to the baseline linear regression on the MAE, with most categories achieving sub 30%. Overall, their model can achieve a MAPE score between 8% to 45%, with model scores congregating at the 20% and 30% region for most cargo categories, which is significant in predictive capabilities. They have also shown that their model vastly outperforms the estimation provided by the port as a service to their customers by a large margin. While impressive, one must note the inherent motivations between the researchers aim to model an accurate prediction of the turnaround time while the port administration may be overbiased to provide a greater amount of slack to the estimation of time (Xu, Chen, & Quan, 2012) to manage expectations of customers. Below is the summary of performance matrices used against the models explored:

|  |  |
| --- | --- |
| **Models** | **Performance Metrics** |
| CatBoost | MAE, RMSE, MAPE |
| Linear Regression (Baseline model) | MAE |
| Port proprietary model | MAE |

Table 4: Performance metrics

Source: Machine Learning based System for Vessel Turnaround Time Prediction by Stepec et al.

As part of their findings, Stepec et al. reported that the most significant features are the type of cargo, its tonnage, the day of the week of entry and the berth used. Lesser significant features revolve around the holiday schedule of the region and no observed impact from tides and congestion within ports. Weather data has some weak significance to certain cargo categories such as dry bulk cargo. Taking these as references, we will be very interested to investigate if similar significance of these features is observed in the prediction of waiting time in anchorage. We should be able to acquire all features that are deemed to be strong predictors in this paper, except for berth being unique to the ports. While the authors did not find any significance of port congestion to prediction of turnaround, it will be our aim to investigate the significance of anchorage congestion in waiting time at the anchorage.



Following are the features selected and used in the final CatBoost model:



|  |  |
| --- | --- |
| **Feature** | **Importance** |
| cargo type (U) | 17.4 |
| cargo tonnage (U) | 16.15 |
| day of entry | 12.71 |
| berth (U) | 11.13 |
| cargo type (L) | 8.54 |
| hour of entry (round 4) | 8.21 |
| berth (L) | 8.03 |
| fiscal cargo type (L) | 6.7 |
| fiscal cargo type (U) | 5.56 |
| cargo tonnage (L) | 4.72 |
| holiday on entry | 0.31 |
| holiday in 2 days | 0.2 |
| holiday 1 day ago | 0.18 |
| holiday in 1 day | 0.16 |







# Data preprocessing

## Present the summary statistics of raw data.

The raw data include bulkers congestion data and waiting time for each anchorage from 2015 to 2021. There are two data sets in bulkers congestion data for each year. For Bulk Waiting Time files, there are 9 features and the description is in Table 3-1. For Bulkers Congestion Daily History, there are 30 features and the description is in Table 3-2.

|  |  |  |
| --- | --- | --- |
| Feature(s) | Description | Type |
| IMO | unique ID for each vessel when they register | int64 |
| Anchorage | an area off the coast which is suitable for a ship to anchor. The locations usually have conditions for safe anchorage in protection from weather conditions, and other hazards. | object |
| Anchorage Entry | date and time when vessel enter the anchorage | object |
| Port | the port that the vessel is waiting to enter | object |
| Berth | a ship's allotted place at a wharf or dock. | object |
| Berth or Port Entry | date and time when vessel enter the berth or port | object |
| Waiting Time (Days) | number of days a vessel waited at anchorage before it enters a port or berth (berth or port entry - anchorage entry) | float64 |
| Vessel Type | types of vessels, indicates different vessel sizes as well, refer to the [Notice] tab for the corresponding range of sizes for each type of vessel. | object |
| Date | Date of entry, I think this is the same as the day when vessel reaches the anchorage | object |

Table 3-1 feature description for Bulk Waiting Time files

|  |  |  |
| --- | --- | --- |
| Feature(s) | Description | Type |
| Date | date of the data | object |
| Anchorage | an area off the coast which is suitable for a ship to anchor. The locations usually have conditions for safe anchorage in protection from weather conditions, and other hazards. | object |
| Zone | Anchorage zone | object |
| Country | country where the anchorage locates | object |
| Total Vessels (Number) | number of vessels waiting at the anchorage | int64 |
| Total Vessels (DWT) | deadweight tonnage, a measure of how much weight a ship can carry. It is the sum of the weights of cargo, fuel, fresh water, ballast water, provisions, passengers, and crew. The total dwt of all the vessels waiting at the anchorage | object |
| Capesize | vessel type, refer to notice tab for dwt range | int64 |
| Capesize Laden | the laden weight is the weight of a vehicle when it is carrying cargoes and other passengers in the vessel. | int64 |
| Capesize Unladen | The unladen weight of any vehicle is the weight of the vehicle when it's not carrying any passengers, goods or other items. It includes the body and all parts normally used with the vehicle or trailer when it's used on a road. | int64 |
| Panamax | vessel type, refer to notice tab for dwt range | int64 |
| Panamax Laden | the laden weight is the weight of a vehicle when it is carrying cargoes and other passengers in the vessel. | int64 |
| Panamax Unladen | The unladen weight of any vehicle is the weight of the vehicle when it's not carrying any passengers, goods or other items. It includes the body and all parts normally used with the vehicle or trailer when it's used on a road. | int64 |
| Handymax | vessel type, refer to notice tab for dwt range | int64 |
| Handymax Laden | the laden weight is the weight of a vehicle when it is carrying cargoes and other passengers in the vessel. | int64 |
| Handymax Unladen | The unladen weight of any vehicle is the weight of the vehicle when it's not carrying any passengers, goods or other items. It includes the body and all parts normally used with the vehicle or trailer when it's used on a road. | int64 |
| Handysize | vessel type, refer to notice tab for dwt range | int64 |
| Handysize Laden | the laden weight is the weight of a vehicle when it is carrying cargoes and other passengers in the vessel. | int64 |
| Handysize Unladen | The unladen weight of any vehicle is the weight of the vehicle when it's not carrying any passengers, goods or other items. It includes the body and all parts normally used with the vehicle or trailer when it's used on a road. | int64 |
| Capesize (DWT) | total DWT of all the capesize vessels waiting at the anchorage. | object |
| Capesize Laden (DWT) | total laden weight of all capesize vessels waiting at the anchorage | object |
| Capesize Unladen (DWT) | total unladen weight of all capesize vessels waiting at the anchorage | object |
| Panamax (DWT) | total DWT of all the panamax vessels waiting at the anchorage. | object |
| Panamax Laden (DWT) | total laden weight of all panamax vessels waiting at the anchorage | object |
| Panamax Unladen (DWT) | total unladen weight of all panamax vessels waiting at the anchorage | object |
| Handymax (DWT) | total DWT of all the handymax vessels waiting at the anchorage. | object |
| Handymax Laden (DWT) | total laden weight of all handymax vessels waiting at the anchorage | object |
| Handymax Unladen (DWT) | total unladen weight of all handymax vessels waiting at the anchorage | object |
| Handysize (DWT) | total DWT of all the handysize vessels waiting at the anchorage. | object |
| Handysize Laden (DWT) | total laden weight of all handysize vessels waiting at the anchorage | object |
| Handysize Unladen (DWT) | total unladen weight of all handysize vessels waiting at the anchorage | object |

Table 3-2 feature description for Bulk Congestion Daily History files

Besides, I use describe function in python to summarize these two data sets for each year, the results for each year are in the .ipynb files.

## Data pre-processing solutions and the justification of your final choice of solution

I checked these data and found the following problems. For each problem, I used suitable methods to cope with.

1) Missing value:

There is no missing value in Bulkers Congestion Daily History. The two features, “Berth” and “Vessel Type”, have missing Bulk values in Waiting Time files. For handling missing values, I delete all rows with missing values because we just delete about 1.5% of rows for each data set. Listwise deletion is reasonable in this case.

2) Outliers:

The feature “Waiting Time (Days)” in Waiting Time files has outliers. I set the 95 percentile of waiting time as a threshold and delete all rows with “Waiting Time (Days)” larger than the threshold. Besides, I delete the rows with “Waiting Time (Days)” equal to zero. We keep 90% of the original data after implementing this strategy.

3) Join waiting time files and congestion files

For the convenience of subsequent work, we join these two tables for each year. The primary key is (Anchorage, Date) because all names in Anchorage in the waiting time file are the same as congestion data.

4) Transform the waiting time into four bins

We used the 25 percentile, 50 percentile, 75 percentile, and 100 percentile as four thresholds and created four bins to represent four levels of waiting time. Level 1 includes the data from the first 25 percentile. Level 2 includes the data from 25 percentile to 50 percentile, and so on. We added a new column to store the levels of waiting time.

5) Encoding of categorical variables

We used one-hot encoder to transform the categorical features because machine learning models and regression models have difficulty in inputting features with the object data type.

6) Select the data with respect to Santos

7) The evidence that we have solved all data preprocessing problems and the prediction performance could be better.

We checked the missing value, outliers, and the type of categorical variables again. There are no missing values and too many outliers. The types of categorical variables we need to use are all int. The results are shown in xxx.ipynb.

Moreover, we established a random forest model to show the improvement of data. The AUC score is xxx before we do data preprocessing. The AUC score is xxx after we do data preprocessing. This result shows the improvement of data.

# References

Bolat, P., Gizem, K., Emine, G., Furkan, K. E., & Soysal, O. (2020). Weighting key factors for port congestion by AHP method. *Journal of ETA Maritime Science*.

Chambers, S. (2021, August 2021). *splash247.com*. Retrieved from splash247.com: https://splash247.com/record-number-of-dry-bulk-ships-hit-by-port-congestion/

CODESP. (2014, 09 08). *Porto de Santos*. Retrieved from http://www.portodesantos.com.br

CTS. (2020). *Container Trades Statistics*. Retrieved from https://www.containerstatistics.com

Danish Ship Finance. (2021, May). *Shipping Market Review - May 2021.*

Fao. (2014, 09 09). *Organização das Nações Unidas para Alimentação e Agricultura.* Retrieved from FAO: fao.org

Golias, M. M., Boile, M., & Theofanis, S. (2009). Berth scheduling by customer service differentiation: A mutil-objective approach. *Transportation Research Part E: Logistics and Transportation Review*, 878-892.

Gui, D., Wang, H., & Yu, M. (2022). Risk Assessment of Port Congestion Risk during the COVID-19 Pandemic. *School of Transportation and Logistic Engineering, Wuhan University of Technology, Wuhan 430063, China; <email>249002@whut.edu.cn</email> (D.G.); <email>hywang777@whut.edu.cn</email> (H.W.) , School of Transportation and Logistic Engineering, Wuhan Unive*.

Hilsdorf, W. d., & Neto, M. (2016). *Port of Santos: prospection on the causes of access difficulties.*

International chamber of shipping . (n.d.). *bulk Carriers*. Retrieved from International chamber of shipping website: https://www.ics-shipping.org/explaining/ships-ops/bulk-carriers/

Jonquais, A. C., & Krempel, F. (2019). *Predicting Shippimg Time with Machine Learning.*

Maersk. (n.d.). *Special Cargo*. Retrieved from Maersk Website: https://www.maersk.com/industry-sectors/special-cargo?gclid=f461af77ab2a16c841e2cd5cdcf98518&gclsrc=3p.ds&

NAMEPA. (2018). *The history of the maritime industry.* Retrieved from https://namepa.net/wp-content/uploads/2018/08/Lesson-3-The-History-of-the-Maritime-Industry.pdf

OECD. (2021). *Ocean Shipping and Shipbuilding*. Retrieved from https://www.oecd.org/ocean/topics/ocean-shipping/

Pedro, B. (2010). *Muito a Navegar: Uma análise logística dos portos brasileiros. Rio de Janeiro: Topbooks.*

Samora, R. (2022, 03 11). *South American nations push to exclude fertilizer from Russia sanctions*. Retrieved from Reuters: https://www.reuters.com/world/americas/south-american-nations-push-exclude-fertilizer-russia-sanctions-2022-03-10/

Soliani, R. D., Mauro, V., Pedro, A. D., & Argoud, A. T. (2015). *Analysis of Load Scheduling System Implementation At Brazilian Santos Port.*

Stepec, D., Martincic, T., Klein, F., Vladusic, D., & Costa, J. P. (2021). *Machine Learning based System for Vessel Turnaround Time Prediction.* XLAB Research.

UNCTAD. (2021). *Review of Maritime Transport.* The United Nations Conference on Trade and Development (UNCTAD).

Xu, Y., Chen, Q., & Quan, X. (2012). *Robust berth scheduling with uncertain vessel delay and handling time.* Annuals of Operations Research.