Norwegian School of Economics
Bergen, Spring 2021

Short term analysis of lead time movements in dry bulk shipping

A linear approach to explore the relationship between lead time and other order-specific determinants of historical voyage orders.

Frida Austarå Botnen & Helena Ha-Anh Lam

Supervisor: Roar Os Ådland

Master thesis, Economics and Business Administration

Major: Business Analytics

NORWEGIAN SCHOOL OF ECONOMICS

Acknowledgements

We would like to express our sincere gratitude to our supervisor, Professor Roar Os Ådland, for providing invaluable guidance and constructive feedback. His profound knowledge within the maritime sector has been a great support for us throughout the process. We would also like to thank head of alternative data in Shipfix, Philippe Pagnotta, for familiarizing us with the data and offering valuable input. Lastly, we would like to express our appreciation to Professor Jonas Andersson at NHH for sharing his competence in the field of applied statistics.

Norwegian School of Economics

Bergen, June 2021

Helena Ha-Anh Lam

Frida Austarå Botnen

Abstract

This thesis investigates explanatory factors on a microeconomic level that affect lead time movements in the short term. We utilize historical voyage orders from January 2015 to April 2021 to analyze the relationship between contract-specific determinants and the lead time for dry bulk cargo. The data sample mainly includes smaller vessels carrying cargo across a wide range of routes. To determine the key drivers, we apply a linear regression model with an incorporated lasso penalty term. We divide the model into nine different commodity groups, where each model inludes the variables of cargo size, seasonality, infrastructure score, and voyage routes between loading and discharge regions. To find the most prominent factors and evaluate the accuracy of the corresponding coefficients, the model coefficients are validated through a bootstrap resampling process.

By analyzing the relationship between lead time and order-specific explanatory factors, we seek to establish whether lead time patterns are connected to the market behaviour of other attributes in the dry bulk sector. Our thesis is the first contribution to lead time analysis in the literature of maritime economics. With the application of pre-fixture order data, we are able to investigate the potential of early market insight through lead time. Results suggest that lead time is affected by the size of cargo for all commodities, except steel products. Among the route dynamics, we find strong evidence of a higher lead time for transpacific routes. On the other hand, intra-Europe voyages have a consistently lower lead time than average. Finally, no seasonal patterns seem to affect the lead time in any direction.

Keywords – Shipping, Dry bulk, Lead Time, Voyages, Linear Regression, Lasso Regression, Bootstrap

Contents

1 Introduction **1**

2 Literature Review **4**

3 Data **6**
- 3.1 Commodity Disaggregation . . . 7
- 3.2 Independent Variable: Lead Time . . . 9
- 3.3 Dependent Variables . . . 10
  - 3.3.1 Cargo Size . . . 10
  - 3.3.2 Season . . . 11
  - 3.3.3 Voyage Routes . . . 13
  - 3.3.4 Infrastructure Score . . . 14

4 Methodology **17**
- 4.1 Multiple Linear Regression . . . 17
- 4.2 Lasso . . . 18
- 4.3 Bootstrap . . . 19

5 Results and Discussion **21**
- 5.1 Linear Model Statistics . . . 21
- 5.2 Model Results . . . 23
  - 5.2.1 Cargo Size . . . 23
  - 5.2.2 Season . . . 24
  - 5.2.3 Infrastructure Score . . . 24
  - 5.2.4 Travel Routes . . . 25
- 5.3 Limitations and Further Research . . . 29

6 Conclusion **31**

References **33**

Appendix **35**
- A1 Full Overview of Region Divison . . . 35
- A2 Seaborne Coal Imports . . . 36

List of Figures

1.1 Timeline of lead time window: First enquiry to loading date. 2
3.1 Density plots of cargo sizes for all commodities 11
A2.1 Annual seaborne coal imports: Asia, EU and US (2000-2020). Data obtained from Clarkson Shipping Intelligence Network. 36

List of Tables

3.1 Data variable names, descriptions and class 7
3.2 Descriptive statistics of vessel types 8
3.3 Commodity groups, frequencies and description 9
3.4 Descriptive statistics table for cargo size, separated by cargo type 12
3.5 Region division, variable names and associated countries 15
5.1 Linear model explanatory variables (including only top 10 largest coefficients in absolute values for routes) from Lasso estimated coefficients and Bootstrap confidence intervals 22
5.5 F-statistics and Adjusted R^2 for commodity groups 22
A1.1 Overview of all initial loading/discharge locations categorized into new regions. 35

1 Introduction

The market for maritime transportation has steadily grown over the last 20 years, making the industry subject to an extensive number of studies. Dry cargo is the undoubtedly largest market segment in maritime transportation, comprising about 67% of the world fleet, whereby 44% of the total fleet represents dry bulk shipping, according to 2019 numbers (Clarksons Research, 2021). With seaborne trade accounting for such a large share of the total world trade, its organization and flow are essential in multiple aspects of the world economy.

The shipping industry comes with high risk and is subject to volatile freight rates and ship prices (Kavussanos et al., 2010). Such risk factors dominate market players' decisions. Thus, any information that can provide insight into the early stages of market changes can improve investment decisions from an economic viewpoint. Such information is highly coveted in the shipping industry.

The time perspective of market information is essential, making lead time valuable if containing information about market behaviour. We refer to the time elapsed from an order first entering the market to the loading date as the lead time. Voyage orders are known before the fixture date, enabling ahead-of-time insight into market behaviour through available order attributes (See Figure 1.1). Shipfix[1] provides a lead time index on their platform, describing it as a tool for 1) tracking oversupply or over-demand on particular routes and 2) detect seasonality for specific commodities. Overall, Shipfix implies that a lead time index provides insight into maritime market volatility. In addition to offering such information as raw data, the company utilizes lead time data to construct forward booking curves based on orders that are currently circulating in the market. The ability to perform such reliable forecasts on short-term demand is rarely possible and is a helpful tool in commodity owners' scheduling of orders to optimize freight conditions.

To our knowledge, there is no existing research on lead time in the shipping sector.

[1]See Chapter 3 for further description of Shipfix

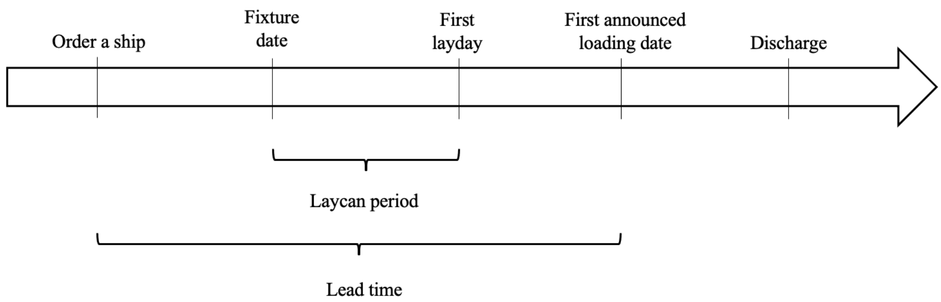

Figure 1.1: Timeline of lead time window: First enquiry to loading date.

Contrary to fixed contractual agreements, analyses of cargo orders that are present in the market before the fixture date requires access to a database that gathers such information from multiple market participants. With the steady development of AI and complex machine learning models, we expect a growth in the magnitude and precision of such platforms. As the first of its kind, this paper will assess voyage order lead times for dry bulk cargo to reveal the positive and negative influences. We perform an in-depth study of micro-level, order-specific attributes, similar to the approach of Alizadeh and Talley (2011) on trip charter fixtures. However, instead of using trip charter fixtures, data is obtained from content extraction of voyage order emails. The aim is to determine attributes of cargo orders that lead time is affected by, and how it is affected. This thesis focuses mainly on minor bulk, a segment that is rarely discussed in existing literature compared to the major bulk sector.

We suggest a two-fold objective for this thesis; (1) to define and investigate order-specific lead time determinants and (2) to compare differences in patterns for the corresponding determinants across dry bulk commodities. The thesis proposes a supervised learning method rooted in a general multiple linear regression to determine the relationship between lead time and suggested explanatory variables. Commodity segmentation is applied to improve bias and results by capturing diversity. For further improvement, we attach a tuning parameter to the linear model in the form of a Lasso penalty. Finally, an implementation of a bootstrap resampling algorithm will quantify coefficient accuracy. The remainder of this paper will first examine existing literature relevant to the thesis. Secondly, we do a presentation of data along with relevant pre-processing steps before model implementation. The theoretical framework is then introduced in the methodology section, dividing the implementation of the model into a three-step process. Finally, the results and discussion section will reveal the effects of relevant model coefficients and their validity, concluded with an evaluation of model limitations and further research suggestions.

2 Literature Review

The first part of the literature section will cover existing research on contract-specific attributes in the dry bulk market. By investigating market factors that affect other segments of the shipping industry, we apply these attributes to our model to test whether they influence the lead time of voyage orders. Next, we will move the focus to geographical research conducted in the shipping sector. The importance of scheduling and logistics in shipping has made this field a recurrent research topic. Previous findings create the basis for testing which geographical characteristics that are associated with fluctuations in lead time.

To our knowledge, there is no existing research on the lead time of dry bulk cargo using order-specific data. However, the study of Alizadeh and Talley (2011) succeed in capturing a simultaneous relationship between freight rates and laycan periods, in addition to a relationship between several microeconomic factors and the laycan period of fixtures for Capesize and Panamax vessels. They describe the laycan period as the time between the fixture date and the first layday of a chartered ship, while the lead time starts when an order enters the market to the cargo's first announced loading date. As seen in Figure 1.1, the laycan period overlaps the lead time window. Consequently, we can expect that the two attributes will respond similarly to market changes. Alizadeh and Talley (2011) finds evidence of variations in the laycan period across multiple trip charter routes.

Similarly, Köhn and Thanopoulou (2011) investigate microeconomic factors affecting the dry bulk sector for Panamax vessels. By quantifying quality-induced differences in dry bulk time charter rates, their results reveal two positive relationships; the *number of days forward*, referred to as the time between a fixture date and the delivery of a vessel, correlate positively with both the charter duration and the place of delivery[2]. Their findings suggest that longer commitments cause commodity producers to plan further in advance. Our thesis uses this claim to orders of voyage charters to answer if the results for time charter fixtures apply to the voyage market in a similar manner.

Analyses of microeconomic shipping determinants often utilize the data contained in

[2]The variable "place of delivery" is split into three levels for charters delivered in Atlantic, Pacific and worldwide.

fixtures, as seen in Alizadeh and Talley (2011); Köhn and Thanopoulou (2011). The usage of fixtures provides precise data on vessel-specific attributes for specific orders, such as age and quality, which can be implemented to analyze freight rate differences across quality-based market segments (Tamvakis and Thanopoulou, 2000; Tamvakis, 1995). We are analyzing lead time of order information available in the market before the fixture date, meaning that vessel specifications are unknown information. Orders are yet to be matched with a vessel at this point in time. On the other hand, the geographical location in which commodity owners want to load and discharge is familiar.

Commodity owners strategically put orders to the market to minimize costs without risking a supply shortage. Considering its importance, studies on geographical data is a well-established research field in shipping literature. For instance, Prochazka et al. (2019) use geographic data to construct an optimization model that captures spatial efficiency and evaluate the advantage of foresight in the dry bulk freight market. Another research focusing on geographical differences is the study by Laulajainen (2007). The study uses different shipping routes from individual freight fixtures and ship movements to verify the geographical efficiency or inefficiency in the dry bulk shipping market. Furthermore, it reveals how the ratio of demand to available ship tonnage, weighted by sailing distance to discharge/loading region, is essential in explaining dry bulk freight rates for individual routes. Brancaccio et al. (2020) propose a different perspective of spatial analysis, suggesting other factors influencing the behaviour of market participants. This study investigates the geographical effects of trade imbalance in the market, revealing how 42% of shipments between the largest importers and exporters travel in ballast.

Analyzing dynamics of geographical distributions relative to lead time incorporates its importance to our analysis. The mentioned literature provides the basis for a discussion of results obtained from the spatial data applied in this thesis. By comparing our findings with existing research, we aim to provide knowledge about the position of lead time data relative to similar market indicators.

3 Data

The Shipfix database provides the data sample applied in this paper. Shipfix possesses an algorithm for extracting order information from emails for dry bulk chartering. The data extraction technology runs an in-depth email content assessment and returns line-by-line and in-line sub-content prediction using deep learning. Around 70 maritime players are currently clients of Shipfix, providing insight on orders circulating in the market at any given point in time. We consider this data as a sample of the total dry bulk sector. The large share of private transactions present for Capesize and larger vessels are less covered in Shipfix data. Thus, the provided data sample is highly concentrated around smaller vessels.

When emails are forwarded to the Shipfix database from its clients, they are examined through textual analysis and information extraction. The algorithm structures this information in a manner that provides easy data access for data analysis. For every order entering the market within Shipfix coverage, the algorithm returns several attributes from the information contained in the specific email. The attributes include information such as the time and date of the first and last email, first and last announced loading and discharge dates, cargo sizes, vessel types (matched with stated cargo size), cargo types of different granularity levels, country of loading and discharge, and number of emails with the same offer. Our thesis explores the historical voyage order sample data. Shipfix also provides similar data for time-charter orders and tonnage daily contribution[3]. The raw voyage order data contains 383,784 observations of individual orders with 30 attributes, spanning from January 2015 to May 2021. Only historical orders observed from January 2015 to April 2021 are applied in this analysis.

Table 3.1 presents the variables used in the model. Some variables are taken directly from the original sample, while other variables reflect initial information obtained in the raw data with some adjustments. We provide elaborate variable descriptions in the continuing sections. For computation purposes, missing values are removed. The remaining data sample consists of 281,946 observations in total.

[3]Daily tonnage are ship openings sent to the market by either ship operators or shipowners daily.

Table 3.1: Data variable names, descriptions and class

Variable	Description	Class
`log_ttm`	Number of days from first order is placed to the first advertised loading date. Log transformed.	Numeric
`cargo_t0`	Type of commodity ordered for transport	Factor w/ 9 levels
`log_sz`	Nominated in metric tons, size of cargo ordered for shipping, log transformed	Numeric
`quarter`	Q1-Q4 for each quarter of the year. Q1 is January to March, Q4 is October to December.	Factor w/ 4 levels
`score`	Infrastructure score for a loading country. Increasing range of 1 - 5	Numeric
`travel`	Route for loading and discharge locations. Both are stated in regions	Factor w/ 256 levels

This chapter is structured as follows. First, we propose a commodity segmentation for the analysis, supported by existing literature. Next, there will be an assessment of the variable selection process, including relevant pre-processing steps. The independent variable lead time is the first variable considered, followed by the independent variables consisting of cargo size, seasonality, infrastructure score, and travel route.

3.1 Commodity Disaggregation

This paper proposes a model separation based on commodity type. Tsioumas and Papadimitriou (2018) explain that differences in trading patterns across commodity types are results of varying commodity prices. The price variation affects the respective import and export quantities, and naturally, not all commodities will follow identical cycles. Essentially, dissimilar market behavior suggests that aggregating over commodities can omit relevant information in a diverse market such as the maritime sector.

A contrasting approach to segmentation is made by Alizadeh and Talley (2011), separating vessel types in their analyses of microeconomic market factors. However, due to the Shipfix data sample structure, this method is considered unsuitable for this microeconomic analysis for multiple reasons. First, the provided data sample is highly concentrated around minor bulk vessels as shown in Table 3.2. Dividing by vessels will group over 70% of the sample into the same model, reducing the chances of capturing diversity. Additionally, the study of Alizadeh and Talley (2011) are based on a data sample of trip-charter fixtures, contrary to our case with a sample of pre-fixture voyage orders. The data in this analysis is obtained prior to a contract agreement, meaning that orders have yet to be assigned a vessel. Accurate information on the vessel type is therefore missing from the sample and only estimated in the Shipfix sample. Estimation of vessel types assigned to each order is based on the weight of cargo specified in the order, `cargo_sz`. We exclude this attribute from the model because estimated values cause further uncertainties in the model results. Lastly, dividing by commodity groups will indirectly consider variations in vessel type to some extent. Larger vessels usually carry a group of commodities such as iron ore and coal, whereas smaller vessels traditionally carry other commodities like grains and steel products. The division is not mutually exclusive[4], and hence, the model includes an independent variable for cargo size. This variable is discussed further in the section for cargo size.

Table 3.2 provides insight into the distribution of cargo sizes across different vessel types in the Shipfix data sample.

Table 3.2: Descriptive statistics of vessel types

`ship_design`	**Min (cargo size)**	**Max (cargo size)**	**% of orders**	**% of cargo size**
Handysize	101	36997	73.3%	43.21%
Handymax	37000	49999	9.24%	14.97%
Supramax	50000	59999	11.14%	22.22%
Panamax	60000	84000	5.39%	14.28%
Capesize	85000	218315	0.81%	3.76%
VLOC	220000	490000	0.13%	1.57%

The variable `ship_design` is estimated by the information extraction algorithm by Shipfix. It is not obtained directly from order emails.

Shipfix data includes attributes for five levels of granularity for the cargo type where the levels differ in the details of the commodity description. For example, observations categorized as *Non-ferrous metals* at level 0 and level 1, are grouped as *Aluminum* at

[4]The minimum and maximum values for the different commodities cover the same ranges of cargo size.

level 2 and *Bauxite* at level 3. This paper suggests that limiting commodity granularity to level 0 is sufficient. This categorization appears typical for the industry and improving model interpretations. The groups of commodities are presented on the left-hand side in Table 3.3, along with respective counts and descriptions. The latter represents examples from commodity descriptions in higher granularity levels of the Shipfix data.

Table 3.3: Commodity groups, frequencies and description

`cargo_t0`	**Count**	**Description**
Grains	61447	Agricultural products
Steel products	44336	Steel wire, steel spool, etc.
Ferts	36938	Phosphate, nitrogen, etc.
Other minerals	35472	Salt, sodium, clay, etc.
Coal	34834	Coking coal, steam coal, etc.
Other bulk	19754	Scrap metals, coke, etc.
Non ferrous metals	18404	Aluminium, manganese, etc.
Cement	15505	Bagged cement, cement clinker
Ferrous metals	15256	Iron, ferrous alloy

3.2 Independent Variable: Lead Time

The time between the first order email and the first advertised loading date is considered the `time_to_market` in the data set, specified in days. We refer to this variable as the lead time. We remove all orders with a time to market beyond 56 days (8 weeks). Shipfix considers any order with a lead time exceeding 56 days to belong in the speculative forward market. Lead time observations assumed to follow short-term market dynamics are the observations of interest for this analysis.

When order emails contain insufficient specifications about loading dates, the Shipfix extraction algorithm can return negative values for lead time. The algorithm is constructed to assign a value for the year, month, and day of the loading date. It will therefore predict these values if not specified in the email. Negative lead time observations from the original data should be excluded from the model because an inclusion will return model estimates based on incorrect values. On the other hand, removing observations raises concerns about biased results[5]. Shipfix considers negative lead times to generally indicate a low lead time for an order[6]. Nevertheless, the values are unknown, and because we are hesitant to tamper with original data, negative lead time observations are removed from the sample. For interpretation purposes, the regression model states the lead time in natural logarithms.

3.3 Dependent Variables

3.3.1 Cargo Size

Alizadeh and Talley (2011) reveals a direct relationship between vessel size and laycan period means and standard deviations[7]. The study emphasizes how Capesize and Panamax vessels might have a higher laycan due to their lack of flexibility surrounding the port decision-making process. Smaller vessels can easily connect with ports, while larger ships mainly transport between fixed or regular routes. In maritime trade, some cargo sizes are also ordered for shipment more frequently than others, as shown in Figure 3.1. Consequently, these orders might be easier matched with a counterparty as they are more common in the market. As a result, we can assume that commodity producers will be eager to place an early order for more unusual cargo sizes to account for possible supply shortages.

High peaks in the density plots indicate a concentration around specific cargo sizes for orders of a specific commodity. Some commodities have less visible peaks, such as *Other bulk*, *Cement* and *Ferrous metals*. Figure 3.1 proves the aforementioned statement of a high concentration around smaller cargo sizes. Orders for cargo with a weight of 80,000 metric ton or above is rare. Regardless of the deficiency of data for heavier cargo, the plots show some patterns. For example, *Steel products* have a larger majority carried in smaller amounts compared to *Coal*.

[5]Missing values should be MCAR (missing completely at random), distributing lost information uniformly among all variables and levels.

[6]Direct communication with Shipfix provides the following example: For orders placed to the market some days into a month, specifying a loading date "somewhere at the beginning of this month," will return an estimated loading date as the first day of that month. If the order is placed on the third day of a month, the algorithm will return a lead time of -2.

[7]The mean and standard deviation of laycan periods is 7.5 and 6.7 days for Capesize vessels, respectively, and 4.4 and 4.6 days for Panamax vessels, respectively.

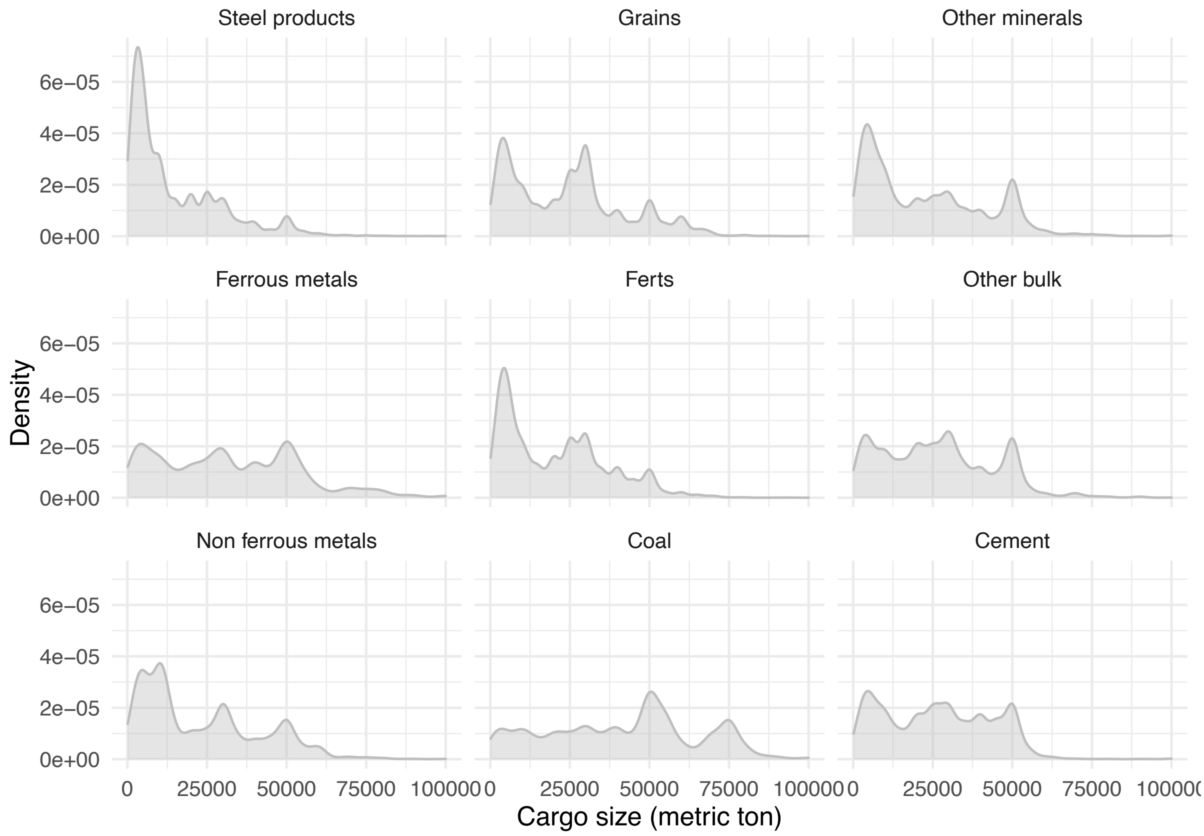

Figure 3.1: Density plots of cargo sizes for all commodities

The cargo size variable is specified in metric tons in the raw data, presented with descriptive statistics in Table 3.4. Naturally, larger vessels will carry orders of larger cargo sizes. Table 3.2 shows that Shipfix suggests a grouping of cargo size to vessel type that is perfectly correlated. The minimum and maximum values never overlap, a case which does not apply to the real world. To avoid multicollinearity, we include the variable `log_sz` only. Contrary to the estimated `ship_design`, the cargo size is extracted directly from the emails. Therefore, we consider this variable to be more precise. As a numeric variable, log transformation is applied to the cargo size to stabilize the mean and improve interpretation abilities.

3.3.2 Season

To capture potential short-term effects, we include a seasonality variable. The month of the first order email from the original data determines the assigned seasonal level for each order. Each month is then aggregated to quarterly levels and classified as a factor variable. The inclusion of a seasonal variable aims to capture potential short-term

Table 3.4: Descriptive statistics table for cargo size, separated by cargo type

Cargo type	**Count**	**% of total**	**Mean**	**Median**	**Min**	**Max**	**SD**
Grains	61447	21.79%	25046	25000	110	480000	21011
Steel products	44336	15.72%	15204	9800	105	480000	17232
Ferts	36938	13.1%	21130	20000	101	480000	20014
Other minerals	35472	12.58%	23837	20000	101	480000	23233
Coal	34834	12.35%	43084	45000	101	490000	27241
Other bulk	19754	7.01%	27489	25000	115	450000	22420
Non ferrous metals	18404	6.53%	23822	18000	105	480000	22937
Cement	15505	5.5%	27556	26000	110	450000	22766
Ferrous metals	15256	5.41%	34873	30000	105	450000	29680

seasonal effects, as suggested by Shipfix.

Seasonal cycles are a fact, depicted in for example the decline in order volumes for grain in the months of July and August (Stopford, 2009)[8]. The distinct seasonality patterns in certain commodity markets create incentives for investigating the potential of transmission effects on the lead time. A transmission effect, in this context, is present when commodity producers increase or decrease the average timing of their order placements. The distinction from an increase or decrease in volume as a result of temporary shifts in demand is irrelevant and should be clarified. Seasonality causes temporary shifts in market supply and demand, making this valuable information for shipowners and charterers in the short- and longer-term decision-making process. No previous research has been done to determine if these shifts are reflected in the lead time. This paper aims to answer if weak periods and temporary, seasonal shifts in demand influence commodity owners in their timing of the initial order enquiry for voyage charters.

Kavussanos and Alizadeh (2001) elaborate on seasonality effects on freight rates in the dry bulk sector. The study finds evidence of asymmetries in deterministic seasonality across different market conditions. The reason is rooted in changes in demand elasticities over the respective market conditions. With evidence of seasonality being present in multiple aspects of the industry, we suggest that including a seasonal variable in this

[8]The data sample also shows a significant reduction in order volume in the late summer months for a specific year.

analysis is reasonable.

Before applying a seasonal variable to the full model, we test seasonality dynamics in lead time for aggregated quarters and months individually. The results reveal that both models have low F-statistics and goodness of fits. Consequently, the seasonal dummy variables applied in the models are stated in quarterly frequencies to avoid high standard errors for the coefficient estimates. Choosing the aggregated quarterly level also reduces model complexity.

3.3.3 Voyage Routes

As mentioned, this thesis deals with a data sample that predominantly consists of orders for smaller cargo sizes. Smaller vessels have less concentrated routes as they can load and discharge at almost any port. On the other hand, larger vessels like Capesize and Panamax often travel fixed routes between a small number of ports, as discussed in Alizadeh and Talley (2011). Their paper implements four routes that cover 86.8% of the observed tip-charter fixtures[9]. However, selecting four representative routes for the Shipfix data sample would provide much lower data coverage due to the highly scattered routes for smaller vessels. The sample's emphasis on smaller shipments is reflected in the geographical distribution of transportation routes. This data sample's four main routes in decreasing order are intra-travel in South-Eastern Asia; South-East Asia to Northeast Asia; Eastern Europe to Western Asia, and intra-travel in South America. In total, the routes only account for about 12.4% of the full sample. When such a small share of the total voyage travels from the sample is represented, it raises concerns about the magnitude of information explained if the equivalent route selection process is applied to our analysis. Hence, we suggest an alternative approach that incorporates every route in the sample. A representative variable for travel distance of 256 routes in total is included in the model, derived from the original data's initial loading and discharge values.

The structure of loading and discharge in the original data is an essential factor for the construction of the `travel` variable. Loading and discharge are both stated in countries

[9]The four routes are: Trans-Atlantic Round Voyage; Continent to Far East; Trans-Pacific Round Voyage; Far East to Continent.

initially, meaning that data for specific ports are unavailable. The region aggregation process suffers from this limitation, as it is typical for maritime spatial data to be grouped by coasts or maritime regions. Grouping data in this manner is unfeasible because the information is unavailable.

The Shipfix extraction algorithm faces some challenges when order emails contain information about loading and discharge that deviate from the regular input it is programmed to detect[10]. Such challenges emerge because the data for cargo requirements applied in this thesis are known prior to the agreement of a charterparty, reducing the degree of structure for this type of data compared to fixtures. The challenge is passed on to the re-grouping process of aggregating to new regions of 16 levels. Orders that for different reasons cannot be categorized into a region will return missing values. These are essentially removed from the analysis.

The original sample contains 249 loading levels and 265 discharge levels in total. Consequently, there are 65,985 shipping routes present in the data on a country level. To include this number of routes in the model would result in high computational costs and variance in estimated coefficients. Therefore, we propose an aggregation of the loading and discharge values (countries and unspecified) based on existing region divisions present in the industry[11]. The new levels for regions are presented in Table 3.5 and are identical for the loading and discharge variables.

3.3.4 Infrastructure Score

A country's infrastructure score is determined through a survey conducted by The World Bank and the University of Turku. Logistic professionals[12] are asked to rate eight pre-selected countries based on trading experiences relative to several logistic dimensions. The survey aims to capture numerical evidence on how easy or difficult it is to transport

[10] When an order email is less specific than what the algorithm requires, the result is a variable with a combination of country and larger region levels (some levels are for example Far East and Continent in addition to country names).

[11] Clarksons Shipping Intelligence Network: Dry Fixtures (Load and Discharge).

[12] Respondents are involved in different logistic services; warehousing and distribution, customer-tailored logistics solutions, bulk or break-bulk cargo transport, and container shipping.

Table 3.5: Region division, variable names and associated countries

Region code	**Regions**	**Countries**
R1_A	Northeast Asia	China, Japan, South Korea, Taiwan, ...
R2_A		India, Pakistan
R3_A	Western Asia	Turkey, Qatar, Oman, United Arab Emirates, Saudi Arabia, Israel, Iraq, Cyprus, ...
R4_A	South-East Asia	Vietnam, Indonesia, Thailand, Malaysia, Singapore, Philippines, Cambodia, ...
R5_A	Other Asia	Bangladesh, Iran, Sri Lanka, Kazakhstan, ...
R6_ANZ		Australia, New Zealand
R7_E	Western Europe	France, Belgium, Germany, Netherlands, ...
R8_E	Eastern Europe	Ukraine, Russia, Poland, Bulgaria, Romania, Moldova, ...
R9_E	Southern Europe	Italy, Spain, Greece, Portugal, Albania, Slovenia, Croatia, Montenegro, Serbia, ...
R10_E	Baltic & Northern Europe	Lithuania, Latvia, Estonia, United Kingdom, Scandinavia, Finland, Ireland, ...
R11_AM	Northern America	United States, Canada
R12_AM	South America	Brazil, Colombia, Peru, Venezuela, Uruguay, Argentina, Chile, Ecuador, ...
R13_AM	Central America & Caribbean	Mexico, Guatemala, Panama, El Salvador, Cuba, Dominican Republic, ...
R14_AF	Other Africa	South Africa, Madagascar, Kenya, Somalia, Tanzania, Angola, Liberia, ...
R15_AF	Northern Africa	Tunisia, Morocco, Algeria, Egypt, Western Sahara, ...
R16	Other	Non-specified countries (Arabian Gulf, Black Sea, Mediterranean, ...)

See A1.1 for full overview of loading and discharge locations within each region.

goods in a country, resulting in the Logistic Performance Index scale from 1 to 5. A country with a score of 1 is considered poor, while 5 indicates a high infrastructure quality.

In concurrence with the survey, the World Bank and the University of Turku wrote a report to explain the survey and its main findings. The report describes how the Logistic Performance Index is analyzed through six different indicators (Arvis et al., 2018): (1) the efficiency of customs and border management clearance, (2) the quality of trade- and transport-related infrastructure, (3) the ease of arranging competitively priced international shipments, (4) the competence and quality of logistics services, (5) the ability to track and trace consignments and (6) the frequency in which shipments reach consignees within the scheduled or expected delivery time. In addition, the report describes the risks with such surveys. For instance, they are subject to sampling error in diverging opinions and variations in the number of received evaluations per country. Furthermore, Arvis et al. (2018) also states that the survey applies for traded products labeled as *general merchandise*. Thus, results should be interpreted carefully for goods that require special handling for transportation, such as food or pharmaceuticals.

We include the `score` variable with expectations that infrastructure qualities will vary across regions on a scale that affects the lead time in the short term. Hence, implementing this variable will test whether commodity owners consider the loading regions' infrastructure quality when placing orders in the market.

4 Methodology

This section proposes a three-step process for determining the relationship between the chosen predictors and lead time. First, a regular multiple linear regression model is applied to determine the overall relationship between predictors and the response by assessing the F-statistic and goodness of fit. Commodity types divide the individual models. Then, we introduce a lasso algorithm to improve model interpretability and account for the high variance problem of regular linear models. Finally, we run a bootstrap resampling algorithm on the lasso coefficients. Estimating standard errors and confidence intervals through resampling is included to create a basis for interpretation.

4.1 Multiple Linear Regression

The purpose of a standard multiple linear model is to reveal any existing relationship between lead time and the predictors. The null hypothesis assumes no significant nonzero coefficients. For every disaggregated model, the null hypothesis is tested by computing an F-statistic. The F-statistic expects to return a value close to 1 when there is no relationship between the response and predictors. The size of how large the F-statistic should be to reject the null hypothesis depends on the size of the sample and the number of predictors. The F-statistic, presented in Equation 4.1, will adjust for the number of predictors p in the model.

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)}, \tag{4.1}$$

where the total sum of squares $TSS = \sum(y_i - \overline{y})^2$ measure the total variance in the lead time y_i, and $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ measure the residual sum of squares. The latter quantifies the variability that is left unexplained after performing the regression. In addition to computing the F-statistic in Equation 4.1, the TSS and RSS are also utilized in the computation of R^2.

With the order-specific explanatory variables presented in Chapter 3, we formulate the following multiple linear regression model with lead time as the response:

$$ln\_LT_i = \alpha_0 + \alpha_1 ln\_sz_i + \alpha_2 IF_i + \sum_{h=1}^{S} \gamma_h SEAS_{i,h} + \sum_{k=1}^{T} \delta_k TRA_{i,k} + \epsilon_i \tag{4.2}$$

The natural logarithm of lead time ln_LT_i is a linear function of a constant α_0, the natural logarithm of cargo size in metric tons ln_SZ_i, the infrastructure score of the loading country IF_i, the quarter of the year in which an order is placed $SEAS_{i,h}$ and the travel between two regions $TRA_{i,k}$.

4.2 Lasso

A potential disadvantage of linear models with a high share of categorical data is when there is a significant increase in the number of predictors (James et al., 2013). The number of predictors interfere with the model's validity when it is considered large compared to the number of observations. Consequently, the model results may induce additional interpretation challenges.

Implementing a lasso algorithm identifies a smaller subset of predictors which exhibit the most powerful effect on the lead time. The variable selection introduces a penalty term to the initial linear model, presented in Equation 4.3. The equation shrinks the estimated coefficients of unimportant predictors to zero. We denote the lasso coefficients as $\hat{\beta}_\lambda^L$, where one coefficient represent the estimated effect on the lead time of a given predictor, for a sequence of values for λ.

$$\sum_{i}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{4.3}$$

The lasso is a shrinkage method that provides model coefficients with lower variance than ordinary least squares (OLS) without sacrificing the low bias. Hence, the model interpretation process is identical to a regular linear model interpretation. The shrinkage parameter λ adjusts for the flexibility of the model. The higher value of lambda, the lower the flexibility of the lasso regression fit, resulting in lower variance but increased bias. The OLS is the basis of the lower bound for bias, with $\lambda = 0$.

Contrary to ridge regression, where none of the coefficients are shrunk to precisely zero, the lasso yields sparse models. Common for both shrinkage methods is the importance of choosing an appropriate lambda value. We find the optimal regularization constant by repeatedly predicting linear models for a sequence of lambda values. The prediction uses 10-fold cross-validation (CV). Based on specifications of a preferred length, the lambda sequence is determined. For our model, the sequence length is set to 100 as it proves adequate. A reduction of the length is unnecessary as the computational time is minimal. The CV uses squared error for computing the loss. Hastie et al. (2013) propose a "one standard error"-rule in a model selection process rather than selecting the model with the lowest estimated MSE. The argument that one should consider the most parsimonious model whose error is no more than one standard error above the minimum obtainable MSE justifies this rule.

A new linear model is fitted to the full sample using the predicted optimal regularization constant λ from the 10-fold CV. The result of the lasso regression will return some nonzero coefficients, while other coefficients will be shrunken to zero and thereby considered non-explanatory for the lead time. In practice, using a lasso algorithm for variable selection translates to classifying the excluded predictors as unimportant and the included predictors as important.

4.3 Bootstrap

Studies have shown that the lasso often proves to have low precision in variable selection (Ayers and Cordell, 2010; Bunea et al., 2011). Specifically, Ayers and Cordell (2010) discuss the use of cross-validation to find the optimal penalization parameter as a method that potentially can return a high number of false positives. Implementing a bootstrap resampling will estimate standard errors and confidence intervals (CI) for the lasso coefficients to account for this inaccuracy.

By drawing new samples from the original data and refitting the lasso model a significant number of times, we can validate the initial lasso coefficients. When considering the risk of low precisions in lasso estimates, the bootstrap ensures more accurate result interpretation by producing safety margins of confidence intervals. In this thesis, we perform a nonparametric bootstrap where all new samples will have an equal number of observations by enabling replacement[13].

An advantage of the bootstrap algorithm is the opportunity to evaluate model uncertainties from the lasso without generating additional samples. However, the method is computationally expensive, resulting in a limitation of the bootstrap resampling in this thesis to 1000 replications with replacement for each cargo type. Standard errors and CI's are computed using a nested cross-validation vector bootstrap approach, as proposed by Efron and Tibshirani (1997). The process of determining the optimal lambda using a 10-fold CV and the "one standard error"-rule is thereby repeated for every resampling, whereby every resample have a unique lambda sequence[14]. The confidence intervals reveal that several nonzero coefficients from the initial lasso regression will not be sufficiently stable to assume an explicitly positive or negative effect on the lead time. Stability in this context refers to the consistency of the estimated coefficients for all repeated samples, assuming that "false positives" will vary between a negative lower bound and positive upper bound (or opposite) of a 95% confidence interval.

[13] Replacement meaning that one observation can be included more than once in the same model.

[14] The minimum and maximum lambda value in the sequence is determined by characteristics of the observations in the sample. From the minimum and maximum values, the rest of the sequence is computed to achieve a length of 100. Thus, every lambda sequence differ from each resampling.

5 Results and Discussion

This section presents the relevant results obtained from the lasso and bootstrap models. The first section will reveal statistics obtained from the initial linear regression model to assess the goodness of fit. Following is a discussion of relevant model coefficients for each explanatory variable. Each section provides a brief comparison of model coefficients for the selected cargo types, providing insight into the explanatory power of variables across models. Finally, we review the thesis' limitations and present suggestions for further research.

The initial linear model consists of 262 predictors. This includes all dummies for the categorical variables, meaning that each predictor reflects either a numeric predictor or one level in a categorical variable. By implementing a lasso shrinkage algorithm with the optimal lambda as suggested in Section 4.2, we see a significant reduction in nonzero coefficients for all models. Table 5.1 displays estimated coefficients from the regression model with the optimal lambda and the estimated 95% confidence intervals. The table will be discussed further in Section 5.2.

5.1 Linear Model Statistics

Models for all cargo types have a significant F-value at a 1% significance level, as seen in Table 5.5. This value reveals the presence of explanatory powers in the chosen predictors for all cargo types. Adjusted R^2 suggests that the predictive power is the strongest in the regression model for grains, steel products, and ferrous metals, in decreasing order. A further assessment of the drivers behind the value of Adjusted R^2 is the basis for extending the standard regression to a lasso regression.

Table 5.1: Linear model explanatory variables (including only top 10 largest coefficients in absolute values for routes) from Lasso estimated coefficients and Bootstrap confidence intervals

Steel products				Grains				Other minerals			
coefs	t0	lower95	upper95	coefs	t0	lower95	upper95	coefs	t0	lower95	upper95
(Intercept)	2.464	2.408	2.539	(Intercept)	1.475	1.383	1.574	(Intercept)	1.986	1.851	2.060
log_sz	0.002	-0.005	0.007	log_sz	0.110	0.102	0.119	log_sz	0.044	0.037	0.056
quarterQ2	-0.002	-0.012	0.015	quarterQ2	0.000	-0.005	0.007	quarterQ2	0.001	-0.016	0.010
quarterQ3	0.000	-0.005	0.007	quarterQ3	-0.004	-0.014	0.010	quarterQ3	0.000	-0.007	0.013
quarterQ4	0.000	-0.004	0.005	quarterQ4	0.000	-0.004	0.003	quarterQ4	0.000	-0.007	0.012
score	0.000	-0.007	0.010	score	0.005	-0.010	0.017	score	0.000	-0.009	0.016
R2-R2	-0.863	-0.996	-0.730	R10-R6	-1.438	-1.938	-0.963	R13-R9	0.697	0.541	0.803
R12-R14	0.681	0.490	0.826	R15-R6	-1.334	-3.230	-0.185	R13-R13	0.476	0.304	0.592
R1-R15	-0.661	-0.902	-0.404	R7-R10	-0.571	-0.643	-0.497	R9-R13	0.462	0.196	0.648
R7-R9	-0.660	-0.709	-0.624	R15-R13	-0.565	-1.031	-0.131	R1-R1	0.446	0.207	0.620
R15-R1	0.656	0.444	0.836	R1-R5	0.552	0.503	0.602	R1-R13	0.440	0.225	0.570
R3-R8	-0.615	-0.708	-0.530	R7-R8	-0.524	-0.753	-0.289	R1-R13	0.434	0.185	0.604
R9-R9	-0.614	-0.664	-0.579	R3-R2	-0.522	-0.760	-0.280	R6-R10	0.432	0.123	0.618
R3-R12	0.584	0.484	0.659	R1-R1	0.488	0.433	0.537	R1-R1	0.424	0.251	0.542
R7-R10	-0.570	-0.650	-0.497	R3-R10	-0.483	-0.710	-0.260	R15-R9	-0.422	-0.486	-0.335
R15-R13	0.559	0.196	0.883	R10-R8	-0.476	-0.743	-0.206	R10-R7	-0.406	-0.484	-0.305

Ferrous metals				Fertilizers				Other bulk			
coefs	t0	lower95	upper95	coefs	t0	lower95	upper95	coefs	t0	lower95	upper95
(Intercept)	2.221	2.081	2.397	(Intercept)	1.894	1.749	1.967	(Intercept)	1.677	1.505	1.800
log_sz	0.033	0.022	0.047	log_sz	0.042	0.035	0.054	log_sz	0.100	0.088	0.116
quarterQ2	-0.021	-0.048	0.008	quarterQ2	-0.009	-0.021	0.016	quarterQ2	0.000	-0.006	0.005
quarterQ3	0.000	-0.005	0.005	quarterQ3	0.000	-0.003	0.003	quarterQ3	0.000	-0.005	0.007
quarterQ4	0.000	-0.023	0.012	quarterQ4	-0.002	-0.012	0.021	quarterQ4	0.000	-0.014	0.009
score	0.007	-0.026	0.028	score	0.010	-0.002	0.028	score	0.000	-0.006	0.008
R2-R2	-0.914	-0.984	-0.848	R1-R6	0.599	0.369	0.735	R10-R10	-0.594	-0.689	-0.472
R2-R16	-0.602	-0.879	-0.277	R1-R1	0.536	0.423	0.594	R7-R8	-0.575	-0.892	-0.171
R6-R1	0.584	0.269	0.804	R1-R5	0.515	0.356	0.586	R3-R9	-0.569	-0.774	-0.324
R12-R4	0.565	0.379	0.665	R7-R7	-0.479	-0.595	-0.317	R15-R3	-0.521	-0.680	-0.323
R12-R5	0.538	0.322	0.676	R15-R3	-0.461	-0.548	-0.334	R10-R16	-0.494	-0.637	-0.319
R15-R9	-0.511	-0.692	-0.303	R15-R9	-0.450	-0.496	-0.370	R1-R6	0.482	0.352	0.582
R5-R1	-0.491	-0.788	-0.149	R7-R10	-0.444	-0.510	-0.342	R8-R3	-0.422	-0.544	-0.263
R1-R10	0.477	0.229	0.641	R1-R2	0.416	0.224	0.526	R10-R15	-0.384	-0.517	-0.219
R2-R1	-0.470	-0.531	-0.418	R14-R4	0.366	0.113	0.530	R1-R1	0.341	0.278	0.390
R12-R3	0.458	0.282	0.571	R15-R7	-0.348	-0.403	-0.257	R7-R9	-0.325	-0.398	-0.233

Non ferrous metals				Coal				Cement			
coefs	t0	lower95	upper95	coefs	t0	lower95	upper95	coefs	t0	lower95	upper95
(Intercept)	2.286	2.128	2.421	(Intercept)	2.049	1.798	2.088	(Intercept)	1.814	1.679	1.986
log_sz	0.028	0.017	0.043	log_sz	0.034	0.028	0.042	log_sz	0.064	0.047	0.075
quarterQ2	0.000	-0.005	0.004	quarterQ2	-0.007	-0.018	0.018	quarterQ2	0.000	-0.012	0.008
quarterQ3	0.000	-0.003	0.003	quarterQ3	0.000	-0.003	0.003	quarterQ3	0.000	-0.004	0.005
quarterQ4	0.000	-0.013	0.008	quarterQ4	0.000	-0.006	0.009	quarterQ4	0.000	-0.008	0.050
score	0.018	-0.007	0.040	score	0.049	0.041	0.115	score	0.000	-0.011	0.020
R10-R4	-1.293	-1.739	-0.773	R3-R9	-0.566	-0.751	-0.253	R3-R13	0.218	0.005	0.237
R7-R14	-0.724	-1.215	-0.191	R7-R10	-0.429	-0.638	-0.164	R4-R4	0.083	0.008	0.088
R3-R10	-0.577	-0.762	-0.352	R14-R14	0.395	0.305	0.457				
R7-R10	-0.508	-0.679	-0.275	R6-R4	0.376	0.291	0.404				
R2-R1	0.437	0.215	0.614	R12-R5	0.363	0.197	0.472				
R6-R1	0.400	0.303	0.478	R14-R2	0.355	0.307	0.395				
R9-R9	-0.392	-0.434	-0.329	R1-R3	0.351	0.172	0.433				
R7-R7	-0.380	-0.582	-0.132	R6-R2	0.347	0.273	0.373				
R9-R7	-0.371	-0.436	-0.280	R1-R4	-0.344	-0.408	-0.264				
R15-R9	-0.371	-0.620	-0.024	R1-R2	0.329	0.231	0.367				

Note: R1-R5 is Asia; R6 is Australia/New Zealand; R7-R10 is Europe; R11-R13 is America; R14-R15 is Africa; R16 is Other. See Table 3.5 for full description.

Table 5.5: F-statistics and Adjusted R^2 for commodity groups

Commodity	**F-statistic**	**Adjusted R^2**
Steel products	40.93^{**}	0.1978
Grains	52.25^{**}	0.2001
Other minerals	20.16^{**}	0.1301
Ferrous metals	15.72^{**}	0.2063
Fertilizers	15.92^{**}	0.1086
Other bulk	15.82^{**}	0.1592
Non ferrous metals	11.31^{**}	0.1248
Coal	22.08^{**}	0.1470
Cement	6.143^{**}	0.0654

**Significant at $<$ 1% level.

5.2 Model Results

5.2.1 Cargo Size

Estimated lasso coefficients for cargo size are significant for all commodities except for steel orders. The significant relationships are exclusively positive, indicating that larger cargo sizes result in higher lead time. Cargo size has the most prominent effect on grain orders, where a 1% increase in size results in an 11.04% increase in estimated lead time. The large impact on grain lead times compared to other commodities can be explained by the fact that agricultural commodities require specific storage facilities at the origin and destination locations (Stopford, 2009). For special cargo handling, limited storage capacities in ports of loading and discharge can cause concerns about potential supply shortages. Additionally, such transportation processes can require more comprehensive planning because both locations must be compatible in terms of storage capacity for special cargo at given point in time. Therefore, larger sizes of special cargo can expect to amplify the risk of shortage and require a longer planning horizon. This creates incentives for commodity owners to put orders to the market earlier than usual to account for potential disruptions.

Cargo size has significant coefficient estimates varying from 2.84% to 9.96% for the seven remaining commodities. These results are consistent with our expectations discussed in Section 3.3.1, where we suggested that unusual or larger cargo orders can affect the commodity owners' behavior, resulting in earlier order placements to avoid potential supply shortages. In addition, increasing cargo sizes require larger vessels. With heavier cargo and a resulting deeper draught of the vessel, it will become less flexible in route decisions (Kavussanos and Alizadeh, 2001). The costs of travelling in ballast increase with the size of the vessel, as we assume that the estimated travel distance increase with the vessel size. Charterers will preferably have a proportionally increasing planning horizon with the increasing cost of travelling in ballast. Additional time to plan longer charters can reduce the chances of travelling backhauls in ballast. Charterers will be hesitant to accept such orders if the potential loss is sufficiently large, in which case commodity owners desire to be early in the market to meet this requirement.

5.2.2 Season

As suggested by Kavussanos and Alizadeh (2001), grain exporting countries suffer from a storage facility shortage during harvest seasons. This leads to an increase in the demand for freight services, which subsequently affects the spot freight rates positively. In accordance with these findings, we expect that a storage facility shortage during seasonal peaks can lead commodity owners to place orders in advance. Beside the risk of increasing freight rates, producers of agricultural commodities fear shortage of available freight and capacity in storage facilities that can result in delayed shipments. Such delays can be expensive for raw materials and earlier order placements can reduce this risk. However, our model proves that seasonal cycles are not reflected in the lead time for any commodity. Commodity producers' timing of order placement is not affected by seasonal fluctuations in demand in the short run.

Voyage orders are usually placed for excess or unexpected cargo to supplement an existing fleet of owned or time-chartered ships. During seasonal peaks, producers can resort to voyage charters due to unexpected demand and thereby increase the volume of orders in the market. This is evident for some commodities in the sample. Because the model shows no indication of lead time variations as a result of the seasonal changes, lead time is only coincidental in seasonal context rather than a response to the market shifts.

5.2.3 Infrastructure Score

In accordance with findings of increasing transportation costs for countries with poor infrastructure (Limão and Venables, 2001), we expect poor infrastructure to result in higher lead times. Sufficient infrastructure quality is important to avoid the risk of freight delays. Prior to a planned sea transportation, cargo will normally undergo road or rail transportation. Cargo located in regions with poor infrastructure might require a longer planning horizon than regions with adequate infrastructure to avoid potential delays due to unexpectedly long transport duration. Delays can result in additional costs for commodity owners.

The `score` variable is only significant for coal, which is estimated to 0.0495. This indicates that any one-unit increase in the score results in a 5.0706%[15] increase in the lead time. This does not comply with our expectations of lower lead times for regions with better infrastructure scores. Regional division can cause vague infrastructure scores because it is aggregated over the belonging countries. When less diversity is included, minor differences are not sufficiently strong to show significant impacts.

5.2.4 Travel Routes

For the remainder of this section, references to model coefficients are limited to the coefficients shown in Table 5.1. All other coefficients for `travel` will be disregarded. These coefficients are, for every commodity, the estimated coefficients for routes that 1) are significantly positive or negative within 95% confidence intervals, and that 2) have one of the ten largest effects on the lead time for a given commodity. With a total of 256 routes included in each model, we consider an interpretation of every significant coefficient to be excessive.

Of the travel routes presented in Table 5.1, coal shows seven of the ten routes discharging in the Asian regions. Three routes loads in America (R11 and R12) and two routes loads in Australia and New Zealand (R6). These routes all have positive coefficients, indicating a higher estimated lead time than the average order for coal. Similarly, the model for ferrous metals shows six of the ten routes discharging in Asia. Half of these routes load in R12 (South America) and discharge in several regions within Asia, all of which with positive estimated coefficients. The remaining three routes discharging in Asia are intra-Asian routes. While our model proves the lead time to be longer for transpacific voyages going from America to Asia, the research of Alizadeh and Talley (2011) implies shorter laycan periods for transpacific routes of Capesize and Panamax vessels. These findings contradict the assumption of positive correlation between lead time and laycan periods.

Additionally, coal has one route travelling intra-Asia (R1 to R4) and one travelling from R3 (Western Asia) to R9 (Southern Europe) displayed in Table 5.1. These routes have

[15] 5.0706 = (exp(0.0495) -1)*100

lower estimated lead times compared to the average. Generally, cargo of coal loading in foreign regions before shipping to Asia have higher estimated lead times, while coal orders loading in Asia has lower lead times. Thus, commodity owners planning to order a charter to carry coal to Asia will put the order to the market earlier if they must load overseas (specifically America or Australia), compared to if the cargo is loaded somewhere else. All intra-Asian routes displayed in the ferrous metals section of Table 5.1 also have consistently lower estimated lead times than average, contrary to the lead times for those loading in R12, which are solely positive.

Brancaccio et al. (2020) presents a matching problem; the challenge of distributing supply and demand to continuously match at every geographical location. The matching problem can explain the mentioned differences in route lead times for coal and ferrous metals. South America is the second largest exporter of ferrous metals and Asia is the largest importing region in the sample, supporting this theory. Due to the large volume of orders discharging in Asia, many vessels will be available for loading in these regions at any point in time. The high availability of tonnage in Asia reduces the commodity producers' fear of supply shortage for ferrous metal voyages. Similarly, decreasing coal demand in America and Europe and continuously increasing demand for coal in Asia aligns with the theory of a matching problem (Clarksons Research (2021), See Appendix A2.1). In terms of order volume, Eastern Europe, Australia and New Zealand, and Northern America are the second to fourth largest coal exporters in the sample. Subsequently, the four largest importers of coal in the sample are exclusively Asian regions. Such differences in import and export regions emphasizes the imbalance in supply and demand. The results are equivalent to those of ferrous metals; commodity owners appear to be less concerned about limited supply when placing orders for loading in Asia compared to Australia or America.

According to Köhn and Thanopoulou (2011), the number of days forward to delivery correlate positively with longer time charter periods.[16] The article reveals how longer commitments require more planning. We cannot directly compare the findings with our research as the time windows are not identical. However, similarities in the time window

[16]Number of days forward is the time between a fixture date and time of delivery.

for lead time and the days forward to delivery suggest that a comparison is reasonable. For instance, voyages travelling between R11 (Northern America) and Asian ports always have a higher lead time, regardless of commodity. Transpacific voyages are expected to have longer durations than the average. However, distance effects are only speculative in this analysis, as the real distance dynamics lack precision when ports are aggregated on a regional level. Nevertheless, clear patterns indicate that an assessment of distance effects should be included.

On the other hand, routes travelling to R15 (North Africa) after loading in R11 have a significantly lower lead time than average for steel products. The route R11 to R15 has a coefficient of -0.6606, indicating an estimated lead time below average. The route going the opposite direction, R15 to R11, has a positive coefficient of 0.6565 for the same commodity. These routes have the same average distance, meaning that voyages between these regions will have the same average *commitment periods* for a voyage charter. The results contradict the aforementioned comparison between the *days to delivery* in Köhn and Thanopoulou (2011) and the lead time. Two important elements should be mentioned, however. Firstly, there is no theoretical evidence of a positive correlation between lead time and laycan periods or *time to delivery*. Second, a longer commitment in the context of time charter fixtures is distinctly different from a longer commitment in voyage charters. The latter is equivalent to travelling distance. These limitations support the theory of other explanatory factors apart from distance affecting the lead time fluctuations.

Table 5.1 shows a clear pattern in lead times for intra-European travels. Voyages within Europe always have a lower estimated lead time than average for any commodity. The top ten trade routes for non-ferrous metals contain the most with four intra-European routes, indicating that four of the ten largest coefficient estimates represent this type of voyage. This is evident despite R1 (Northeastern Asia) being the undoubtedly largest importer of non-ferrous metal in frequency, with 5746 observations in the sample. R9 (Southern Europe) is the second largest discharge region with 1668 observations. The order frequency of Europe is not particularly prominent on the loading side of non-ferrous metal trade either. Southern Europe is the fifth largest exporter of non-ferrous metals with 1501 observations, compared to South America with the highest number of orders with 4218 observations in the sample. In addition to non-ferrous metals, the segments of steel products, grain and other bulk are the remaining models in Table 5.1 that also display at least three intra-European routes. The large negative impact on lead time by the intra-Europe routes can be explained by the high share of sea transportation in Europe. Coastal and short-distance shipping account for 40% of intra-European trade today (Suárez-alemán, 2016). The concept of short sea shipping has been on the European Commission's agenda for two decades, with the aim of reducing the environmental damages caused by road transportation. Persistent work on short sea shipping within Europe improves the internal trade flow and predictability of the trade exchanges in Europe. This can explain the lower lead time for such a large number of intra-Europe routes for voyage orders.

R6 (Australia and New Zealand) will, on average, be the region located the furthest away from the other regions, meaning voyages ordered to load or discharge in R6 will generally travel a longer distance. With longer voyages, we expect higher lead times due to the risk of travelling long distances in ballast. This type of risk accounts for other underlying factors than the distance itself and results in higher expected transportation costs for the charterer if they fail to find new cargo to load for the backhaul route. Commodity owners should want to put orders to the market earlier to reduce the risk for the charterer. In return, a longer planning horizon for the charterer reduces the possibilities for loss and should be reflected in better terms for the commodity owners. Another argument supporting the expectations for a positive correlation between lead time and distance is the risk of delay. When a voyage increases in duration, the time window of which disruptions can appear will increase proportionally. The risk of delay increases the concern for extra costs for the charterer. Therefore, we assume that charterers prefer to receive additional time to evaluate this risk.

With expectations of longer distances having higher estimated lead times, routes loading or discharging in R6 should solely have positive coefficient estimates. However, Table 5.1 shows inconsistent signs of the coefficients. For instance, grain orders travelling from R15 (northern Africa) and R10 (Baltic/northern Europe) to R6 (Australia/New Zealand) has estimated coefficients of -1.3338 and -1.4379, respectively. On the other hand, orders placed for the opposite direction of R6-R10 in the model for other minerals have a positive estimated coefficient of 0.4318. Several factors can cause the inconsistency in correlations between distance and lead time. As mentioned, voyage charters are a last-minute resort for commodity owners, meaning that placements of a voyage order will be urgent in many cases. The sudden and unexpected need for additional charter causes commodity owners to search in the voyage charter market. In such situations, lead time is likely to be arbitrary. By looking at the routes R15-R6 and R10-R6, the confidence intervals are large, supporting the argument of randomness and high variations in the lead time.

5.3 Limitations and Further Research

The limitation of this thesis concerns properties of the data. Firstly, limited access to private market transactions in the sample limits the coverage of larger vessels. As orders for larger cargo often trade within private agreements, the orders omitted from the data sample are not random. Consequently, this analysis only covers a small segment of the dynamics within larger vessels. Concerns about the sample distribution compared to the total market segment should be considered. Alternatively, the interpretation of this thesis can be restricted to markets for smaller vessels. A sample that deviate from the population can potentially cause misleading model results. Iron ore and coal are examples of commodities that usually are carried by larger vessels. The models for these commodities should therefore be interpreted in the light of this limitation if the data deficiency is not random.

Another limitation of this thesis is the large difference in observation frequencies between commodities. This difference is also reflected in the number of nonzero coefficients in our models. For instance, grain has 61447 observations in our data sample, with a corresponding 70 significant nonzero coefficients in the model. On the other hand, cement has only 15505 observations with only 4 significant nonzero coefficients. Similarly, the shipping market in general will not have an even distribution of voyage orders. Some commodities are transported more frequently than others. This is visible when comparing commodities carried by Capesize vessels to those carried by smaller vessels, where voyages for smaller shipments are more recurrent. Uneven distribution in the sample is only a concern if deviating from the population.

Further work with lead time analysis can explore the new technology of AIS (Automatic identification system) to collect market data in real time. Extracting additional information on trade flows can be accessed through this technology, thereby improving the problem of low market coverage of private transactions. A wider range of market data provides the basis for increased accuracy of market indices and forecasts by reducing the risks of an unrepresentative sample. Further research can also investigate the relationship between freight rates and lead time of cargo orders in the form of forward booking curves. Using lead time and other order data to forecast freight rates can improve investment decisions and strategies of market participants.

6 Conclusion

This thesis investigates the effect of order-specific determinants on the lead time based on a data sample for voyage orders of dry bulk cargo. The goal of this paper is to provide a new perspective to existing research on voyage and time charter fixtures by analysing orders immediately when they arrive to the market. The results indicate that voyage order lead times can be explained by micro-level factors such as cargo size and travel routes, thereby enabling insight on market behaviour ahead of time.

We initiate a linear regression model to find relationships that explain the behaviour of lead time fluctuations, including a lasso penalty term to extract the main predictors of an initial model of 262 predictors. The adopted shrinkage parameter of the lasso is derived from a 10-fold cross-validation process to minimize mean squared error. Finally, confidence intervals for all coefficient estimates are obtained from a bootstrap resampling method to validate the lasso results. Lasso coefficients that are exclusively positive or negative within a 95% window are considered significant.

The analysis considers the four explanatory variables of cargo size, quarter of the year, travel route and infrastructure score. Commodities are separated into individual models to capture cargo-related diversity. Final results show that cargo size is positively related to the lead time for all commodities except for steel products, where it has no explanatory power. The seasonal component in dry bulk shipping has no significant transmission effect on the lead time for any commodity.

The variable for travel routes returns a number of significant coefficients, suggesting that some travel routes do affect voyage order lead times. Firstly, the lead time is influenced by different routes for different commodities. This confirms the presumption that commodities should be considered separately to capture diversity. While some routes are significantly affecting the lead time of one commodity, it might not have the same effect on other groups of cargo. However, there are also similarities detected between commodities. For instance, all transpacific routes travelling between America and either

Asia or Australia/New Zealand that have a significant effect on lead time is always positive, indicating a higher lead time for orders travelling transpacific routes than the average regardless of commodity. On the other hand, significant coefficients for voyage routes travelling intra-Europe always return a lower estimated lead time. Further, the model results show no evidence of a consistent pattern in lead time fluctuations relative to travel distance. This indicates that the significance in travel route coefficients contain other underlying factors aside from the distance affecting lead time.

With the relevant order attributes, we are able to investigate sparse route dynamics of minor bulk lead times in high-frequency voyage order data. The process of coefficient shrinkage and resampling through lasso and bootstrap algorithms enables our model to consider and interpret highly diverse market data.

References

Alizadeh, A. H. and Talley, W. K. (2011). Microeconomic determinants of dry bulk shipping freight rates and contract times. *Transportation*, 38(3):561–579.

Arvis, J.-F., Ojala, L., Wiederer, C., Shepherd, B., Raj, A., Dairabayeva, K., and Kiiski, T. (2018). Connecting to compete 2018 : Trade logistics in the global economy.

Ayers, K. L. and Cordell, H. J. (2010). Snp selection in genome-wide and candidate gene studies via penalized logistic regression. *Genetic Epidemiology*, 34(8):879–891.

Brancaccio, G., Kalouptsidi, M., and Papageorgiou, T. (2020). Geography, transportation, and endogenous trade costs. *Econometrica*, 88(2):657–691.

Bunea, F., She, Y., Ombao, H., Gongvatana, A., Devlin, K., and Cohen, R. (2011). Penalized least squares regression methods and applications to neuroimaging. *NeuroImage*, 55(4):1519–1527.

Clarksons Research (2021). *Seaborne Trade Monitor.*

Efron, B. and Tibshirani, R. (1997). Improvements on cross-validation: The .632+ bootstrap method. *Journal of the American Statistical Association*, 92(438):548–560.

Hastie, T., Tibshirani, R., and Friedman, J. (2013). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer New York.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer.

Kavussanos, M., Visvikis, I., and Dimitrakopoulos, D. (2010). Information linkages between panamax freight derivatives and commodity derivatives markets. *Maritime Economics Logistics*, 12(1):91–110.

Kavussanos, M. G. and Alizadeh, A. H. (2001). Seasonality patterns in dry bulk shipping spot and time charter freight rates. *Transportation Research Part E: Logistics and Transportation Review*, 37(6):443–467.

Köhn, S. and Thanopoulou, H. (2011). A gam assessment of quality premia in the dry bulk time–charter market. *Transportation Research Part E: Logistics and Transportation Review*, 47(5):709–721.

Laulajainen, R. (2007). Dry bulk shipping market inefficiency, the wide perspective. *Journal of Transport Geography*, 15(3):217–224.

Limão, N. and Venables, A. J. (2001). Infrastructure, geographical disadvantage, transport costs, and trade. *The World Bank Economic Review*, 15(3):451–479.

Prochazka, V., Adland, R., and Wallace, S. W. (2019). The value of foresight in the drybulk freight market. *Transportation Research Part A: Policy and Practice*, 129:232–245.

Stopford, M. (2009). *Maritime Economics*, volume 3. Routledge.

Suárez-alemán, A. (2016). Short sea shipping in today's europe: A critical review of maritime transport policy. *Maritime Economics Logistics*, 18(3):331–351. Copyright - Maritime Economics Logistics is a copyright of Springer, 2016; Last updated - 2020-11-17; SubjectsTermNotLitGenreText - Europe.

Tamvakis, M. N. (1995). An investigation into the existence of a two-tier spot freight market for crude oil carriers. *Maritime Policy & Management*, 22(1):81–90.

Tamvakis, M. N. and Thanopoulou, H. A. (2000). Does quality pay? the case of the dry bulk market. *Transportation Research Part E: Logistics and Transportation Review*, 36(4):297–307.

Tsioumas, V. and Papadimitriou, S. (2018). The dynamic relationship between freight markets and commodity prices revealed. *Maritime Economics Logistics*, 20(2):267–279.

Appendix

A1 Full Overview of Region Divison

Table A1.1: Overview of all initial loading/discharge locations categorized into new regions.

Regions	**Initial loading/discharge location**
Northeast Asia	China, South Korea, Taiwan, Japan, North Korea, Singapore-Japan.
India/Pakistan	India, Pakistan.
Western Asia	Kuwait, Turkey, Bahrain, Qatar, Oman, United Arab Emirates, Saudi Arabia, Israel, Jordan, Syria, Yemen, Georgia, Iraq, Cyprus, Lebanon, Azerbaijan.
South-East Asia	Vietnam, Indonesia, Thailand, NA, Malaysia, Singapore, Philippines, Brunei, South East Asia, Myanmar, Cambodia, Timor-Leste.
Other Asia	Turkmenistan, Kazakhstan, Bangladesh, Iran, Pakistan, Sri Lanka, Maldives.
Australia/New Zealand	Australia, New Zealand.
Western Europe	France, Belgium, Continent, Germany, Netherlands, Amsterdam-Rotterdam-Antwerp-Ghent, North Continent.
Eastern Europe	Ukraine, Russia, Poland, Bulgaria, Romania, Moldova.
Southern Europe	Italy, Spain, Greece, Portugal, Albania, Slovenia, Croatia, Montenegro, Gibraltar, Malta, Serbia.
Baltic & Northern Europe	Lithuania, Latvia, Estonia, Baltic, Lower Baltic, United Kingdom, Norway, Denmark, Sweden, Finland, Faroe Islands, Ireland, Iceland.
Northern America	United States, Canada, Greenland
South America	Brazil, Colombia, Peru, Venezuela, Guyana, Uruguay, Argentina, Chile, Ecuador, Suriname, Paraguay, French Guiana.
Central America & Caribbean	Mexico, Guatemala, Panama, Costa Rica, Honduras, Nicaragua, El Salvador, Belize, Puerto Rico, Bahamas, Jamaica, Cuba, Dominican Republic, Guadeloupe, Cayman Islands, Haiti, Curaçao, Barbados, Dominica, Martinique.
Other Africa	Mozambique, Djibouti, Madagascar, Kenya, Somalia, Eritrea, Tanzania, East Africa, Mayotte, Réunion, Mauritius, Angola, Cameroon, Gabon, Equatorial Guinea, Senegal, Nigeria, Guinea, Liberia, Sierra Leone, Togo, West Africa, Ivory Coast, Mauritania, Guinea-Bissau, Gambia, Ghana, Benin, South Africa, Namibia.
Northern Africa	Tunisia, Morocco, Algeria, Egypt, Western Sahara, Libya, Sudan.
Other	Non-specified countries (Red Sea, Black Sea, East Mediterranean, Arabian Gulf, North Pacific, Melanesia, West Mediterranean, Mediterranean, Micronesia, Polynesia)

A2 Seaborne Coal Imports

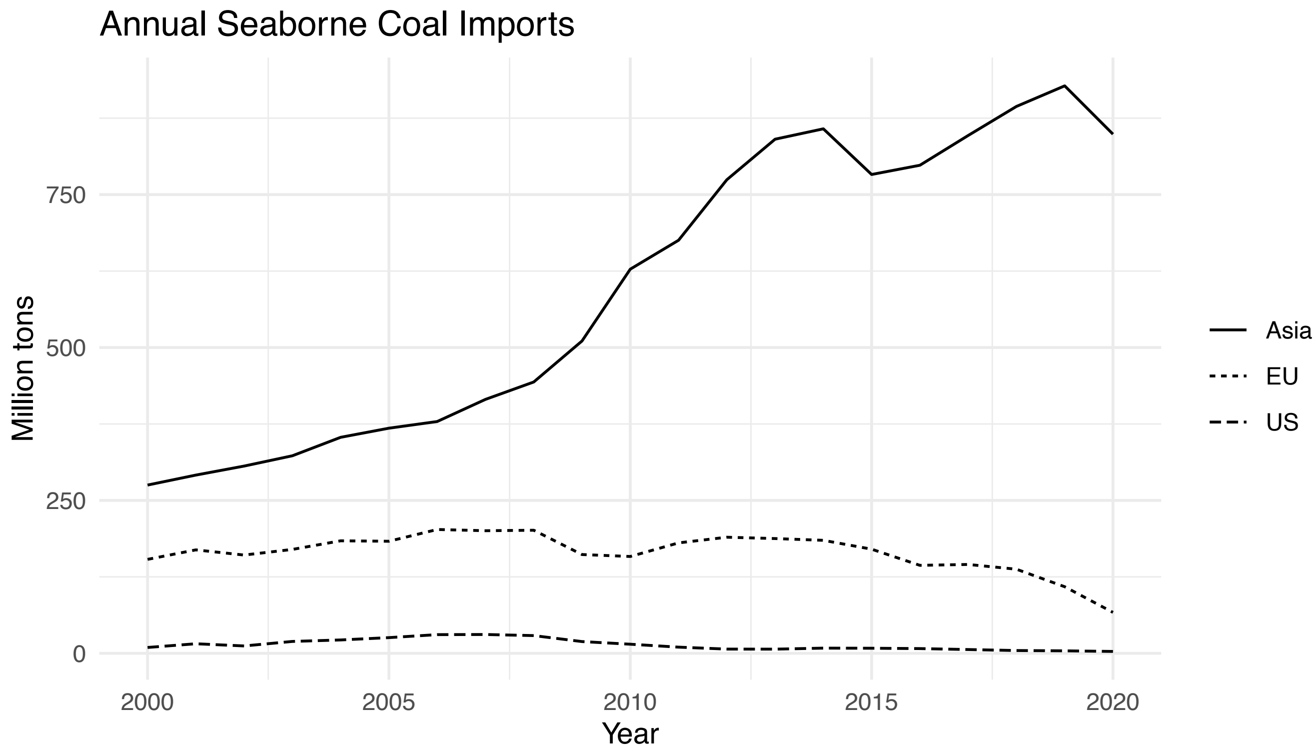

Figure A2.1: Annual seaborne coal imports: Asia, EU and US (2000-2020). Data obtained from Clarkson Shipping Intelligence Network.