An Applicability Study of Data Mining to Improve the Secondhand Market Model of the Maritime Business Game

K. Anggelia

Master of Science Thesis

TU Delft

An Applicability Study of Data Mining to Improve the Secondhand Market Model of the Maritime Business Game

Master Thesis Report
by

K. Anggelia

Study program
Master of Science in Marine Technology
track Ship Design, Production and Organization
with specialization in Ship Production

Participating Faculty
Mechanical, Maritime & Materials Engineering

Thesis number: SDPO.19.031.m.

To be defended publicly on Thursday, 14 November 2019 at 10.00.

Supervisor:	Dr. J.F.J. Pruyn MSc.	TU Delft
Thesis Committee:	Ir. J.W. Frouws	TU Delft
	Dr. W.W.A. Beelaerts van Blokland MSc.	TU Delft

TU Delft

Whatever you do, work at it with all your heart, as working for the Lord, not for human masters.
(Apostle Paul to the People of Colossae)

Preface

I would like to extend my thanks to everyone that have helped make this graduation project possible. I would especially thank my daily supervisor, Jeroen Pruyn for helping me to find a fascinating research topic. And along the way, thank you for all the precious feedback, advice and guidance that you have provided throughout the project. I really appreciate your expertise, availability, flexibility and directness; I could not hope for a better supervisor. My classmates can testify of how happy I am to have you as my supervisor.

Furthermore, I want to express my gratitude to Mr. Frouws and Mr. Beelaerts van Blokland for allocating your busy time to be a part of my graduation committee. Especially to Mr. Frouws, thank you for reminding the added values which a marine engineer should have. I would like to extend my thanks to Drs. Lourdes Gallastegui Pujana who has been mentoring me from the beginning of my study. The first time we met was when I just had started with bachelor Mechanical Engineering (and then I switched my study to MT one year later). Thank you for walking with me through my darkest academic days.

Also thanks to my family. Especially for Dad, who initially disagreed with my decision to be an engineer. After the first year of my MT bachelor, he finally gave in and became my biggest supporter. I am glad that we have learned to disagree agreeably. I want to thank my Mom who has been allowing me to dream (sometimes unrealistically) and has been supporting me (even when my Dad wouldn't). I also want to thank my siblings, Meli and Markus, whom I truly love. They are my reality check. Thank you for still having faith in me, even after I had changed my study 3 times (and have spent 1 year for each "trial", until I ended up in bachelor MT in Sept. 2013). Most of Asian (minded) parents won't tolerate such a thing, so thank you for tolerating me.

Furthermore, I want to thank Cao Yan who has provided some valuable insight regarding GAM modelling. I want to thank my life-group family and friends from Redeemer Delft who have provided constant encouragement and prayers. I especially thank Marian, Sinjin, Yuchen, Marjorie, Helen, Alister and Hsoc for encouragement and very practical supports. What my church family has done (both in Delft and Trondheim), has meant so much.

Also, I would like thank all my thesis buddies, especially Adam, Angin, Rani, Ges, Agung, Inggrid, Fareza, Kevin & others. Fighting together is much better than fighting alone. In addition, I want to thank Shen, Giovani, Enri, Teddy, Jessica, Albert & Amira for the weekend hangouts. I also want to thank Fany, Ava, tante Hana, Sarang, Anggrit & Ficky who have been constantly supporting me from a far. Lastly, I want to thank my jogging partner, Ajeng, thank you for running with me and hearing all my stories, both good & bad.

Last but not least, without trying to be religious, I want to give the utmost credit to Jesus. Studying at TU Delft has been a very challenging period for me and I have been through some very difficult things (such as extreme culture shock, language barrier, being very discouraged, burn-out and even depression). I believe that it is mostly my faith in Jesus and the genuine support from my blood-family, church-family & friends who have been giving me the hope & strength to keep on going. I have given so much grace to finish this (difficult) study within 6 years, I am still very amazed of it.

Abstract

This study aims to improve of secondhand market model used in the **Maritime Business Game (MBG)** & to judge the relevance of three data mining algorithms. Pruyn developed MBG to simulate the long term consequences of various shipping scenarios. MBG consists of three multilevel models to represent the maritime economy. Starting from the Multi Country model, it goes down to the shipping markets model & ends at the fleet scheduling model. Secondhand market is a sub-part of shipping market module, where the shipowners trade their old ships. In real life, the valuation is done by shipbrokers under the assumption that vessels are in good working condition. Thus's, physical characteristics have a little influence on the sales price. A structural model based on the physical characteristics can be a good complement to the broker's judgment.

The MBG secondhand market considers five main important variables namely vessels' ***age***, ***size***, ***income***, ***LIBOR*** & ***orderbook size***. General Additive Methods (GAM) is used as the pricing method because it requires less data points to predict more dependent variables. The first aim of the project is to increase model sensitivity towards three additional features of the vessels. These are variety in $\frac{Volume}{DWT}$ and smaller vessels features like ***ice class*** & ***crane***. Furthermore, other factors are considered by reviewing the literature. These are the ***efficiency index***, ***builder reputation*** & ***various sustainability indicators***. By matching the literature finding with data availability the parameters to represent those factors are selected.

Furthermore, six relevant data mining algorithms are considered to be used as potential basis for the pricing model. These are ***Generalized Additive Models (GAM)***, ***Multivariate Adaptive Regression Splines (MARS)***, ***Random Forests***, ***Gradient Boosting Machines (GBM)***, ***Artificial Neural Networks (ANN)*** & ***Support Vector Machines (SVM)***. To select the model, the *Analytic Hierarchy Process* is applied to select the most potential candidates. Six selection criteria are applied, namely ***usability***, ***result quality & fitting tendency***, ***preprocessing ease***, ***postprocessing ease***, ***input compatibility*** & ***computational time***. Among these techniques, GAM, Random Forest & GBM are voted as the most relevant algorithms.

Before starting the modelling process, two statistical analysis, *Correlation Test* & *Principal Component Analysis*, are conducted. Several variables are eliminated to prevent collinearity & fulfill the statistical assumptions. After performing these tests, 19 variables remains & they are summarized in table 6.6. The sales data is obtained from the Clarksons website, with a total of 1227 individual sales contract from July 2016 to July 2019. It consists of 356 handysize vessels, 441 handymax vessels, 329 panamax & 101 capesize.

The results obtained from this research generally validate the results of previous research. Namely ***Age, DWT, TC-rate & LIBOR*** are among the most important variables to consider regarding the vessels price. The results have been consistent across five different vessel types & three different algorithms. In addition, the importance of new variables have been discovered. These are $\frac{Volume}{DWT}$ ***Ratio & Normalized Admiralty Constant (Ca)***. According to Random Forests outcomes, $\frac{Volume}{DWT}$ is important for every vessel types except for handymax. Furthermore, GBM results suggest that normalized Ca is influential for all vessel types except for Capesize. Lastly, among all variables, ***age & DWT*** are proven to has the most influence on price.

To conclude, considering the strengths & weaknesses of algorithms along with the improvement goals; all algorithms are useful, but GAM is the most applicable one for this project. This statement is no longer valid when the number of variables significantly increases. In such situation, machine learning approach will be more suitable. By this, the main research question, '***What is the most practical way to improve the current secondhand bulk carriers pricing model used in Maritime Business Game based on the available data & what is the implementation result?***', that is proposed in the beginning of the project is answered.

Contents

List of Figures

List of Tables and Equations

List of Tables

List of Equations

List of Abbreviations

AHP	Analytic Hierarchy Process
ANN	Artificial Neural Networks
ARCH	Autoregressive Conditional Heteroscedasticity
AR(I)MA	Autoregressive (Integrated) Moving Average
BF	Basis Functions
CI	Consistency Index
CO_2	Carbon Dioxide
CR	Consistency Ratio
DCF	Discounted Cash Flow
DUT	Delft University of Technology
DWT	Dead Weight Tonnage
ECA	Emission Control Area
EMF	Efficient Market Hypothesis
FS Model	Fleet Scheduling Model
GLM	Generalized Linear Models
GAM	Generalized Additive Models
GBM	Gradient Boosting Machines
GCV	General Cross Validation
GLM	Generalized Linear Model
GT	Gross Tonnage
ICE	Individual Conditional Expectation
IMO	International Maritime Organization
LBT	Length-Breadth-Draft
LIBOR	London Inter Bank Offered Rate
LNG	Liquefied Natural Gas
LTAV	Long Term Asset Value
LWT	Lightweight Tonnage
MARS	Multivariate Adaptive Regression Splines
MBG	Maritime Business Game
MCM	Multi Country Model
MDE	Multivariate Density Estimation
MGO	Marine Gas Oil
MLE	Maximum Likelihood Estimation
NIESR	National Institute of Economic and Social Research
NO_x	Nitrogen Oxides
NPV	Net Present Value
OECD	Organization for Economic Co-operation and Development
OLS	Ordinary Least Squares
PhD	Philosophiae Doctor
RBF	Radial Basis Function
RF	Random Forests
RI	Random Index
RMSE	Root Mean Square Error
ROE	Return On Investment
RSS	Residual Sum of Squared

SAE	Sum of Absolute Errors
SM Model	Shipping Market Model
SO_x	Sulfur Oxide
SVM	Support Vector Machines
TC	Time Charter
TPRS	Thin Plate Regression Splines
UA	University of Antwerp
VAR	Vector Autoregressive
VOC	Vereenigde Oostindische Compagnie
VECM	Vector Error Correction Model
WS100	World Scale Flat Rate

1 Introduction

Und jedem Anfang wohnt ein Zauber inne, Der uns beschützt und der uns hilft, zu leben.

Hermann Hesse

1.1. Background

Seagoing vessels have always been playing an important role throughout the human civilizations. Dated from 1571 to 1862 was the age of sail. It was the period when the international trade and the naval warfare were dominated by sailing ships. European kingdoms were expanding their territory to the rest of the world during that time when the most expansive human migrations were recorded in history[2]. Beginning with exchanging the essential resources to meet basic needs, international trading then developed as a mean to gain income. As an example, the Dutch had sailed all the way through to South Africa, India and even to Indonesia to gather natural resources. Gradually, the Dutch trading company, VOC (Vereenigde Oostindische Compagnie) had grown into the most profitable trading companies ever existed in human history[1].

And as for this time, although other means of transportation such as air and road transports are available, ships are still largely used. The International Chamber of Shipping states that as in 2018, 90% of world trade is transported by seagoing vessels. Moreover, among all the goods that are shipped, major dry bulk dominates as it makes up to one third of the shipped commodities. All these facts have proven that shipping industries have been playing an essential role in human civilization, and it will continue to be a major driving force of the global economy for a long time to come[3].

Since the shipping industry holds a very important role, having a good insight about it is thus indispensable. This is especially true for people who are directly involved in it such as shipowners, shipping operators and shipyards; as well as for the traders who make use of its service and for the financial institutions who makes investments in it. In this way, a wide variety of people involve in it and ever more are, directly and indirectly, affected by it. Moreover, making heavy investment is an essential requirement for the main players. This fact in combination with the very dynamic nature of the industry make shipping business becomes not only challenging and profitable, but also risky and speculative at the same time. Accordingly, understanding this business well will help them to know what the right decision is and when to take it[4].

However, understanding the dynamic of maritime market is not that easy, let alone predicting the future. This is especially true for the new players and small scale shipping enterprises, including the yards, investors and shipping companies. Since new players do not enjoy the benefit of years of experiences, and small enterprises do not have the privilege of big model and database which are usually possessed by the large companies[5]. While the industry remains immensely competitive by nature and shipping cycles efficiently squeeze the weak players out of the markets, the main objective remains the same for everyone, namely to minimize loss and to maximize profit[4]. Thus all of these statements lead to this one question: *What are the possibilities for new and small players to get a better grip on the maritime business future?*

1.2. Maritime Business Game

In order to answer the question in the previous subsection, Jeroen Pruyn, a researcher from TU Delft, provides an alternative. He has developed an integrated model consists of all relevant markets within the maritime business. This model is further developed into a game, so that it can be conveniently used by a wide range of users. This is called the **Maritime Business Game (MBG)** which simulates various consistent shipping scenarios. Hereby, its players might learn what the medium or long term consequences are of making a certain decision in a particular scenario. As users experiencing different shipping scenarios, they will get a deeper practical insight and enhanced understanding over the shipping business dynamic. This eventually might help them to make wiser operational and investment decisions.

As an example of real-world application of this game, MBG has been played by the students who take at the Maritime Finance, Business and Law course at TU Delft. A group of three students constitutes a team who will act as a shipping company with a limited initial capital. In this context, the players will operate their vessels in a simulated freight market in order to obtain as much profit as possible within a specific time period. In the beginning of the game, they can buy a new-building or secondhand vessel(s). Buying a new-building will require extra building time that player could not enter the freight market immediately; thus this is not a preferable option for most. By playing MBG, players obtain an understanding of shipping market.

Technically, MBG is built upon the integration of three models within the maritime business context. The first highest level model is the Multi Country macroeconomic model which represents world's leading maritime capitals. The second level model represents the shipping market consists of the new-building market, the secondhand market and the scrapping market. The third level model represent the fleet scheduling which allocates the available fleet and the available cargo based on market's supply and demand. The advantage of having such an integrated model is that one can create various scenarios which are consistent in all three level[5]. The variation between scenarios is measured relative to a base scenario, to see the medium or long term effect instead. To sum up, the conceptual overview of the integrated model is shown in the figure 1.1.

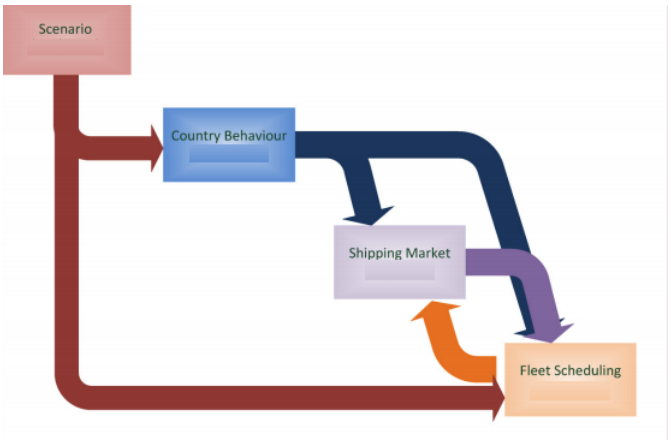


Figure 1.1: An overview of the integrated MBG model. Source: Jeroen Pruyn[5]

1.3. Research Focus

By and large, this research specifically concentrate on the shipping market module. The shipping market consists of four sub-markets which are closely related to one another. These are the new-building market, the secondhand market the scrapping market and the freight market. Among the four markets, the shipping market model in MBG represents only the first three markets mentioned above; whereas the freight rate market is modelled separately in the fleet scheduling module[5]. Henceforth, this study will focus exclusively on the secondhand market part.

In the initial secondhand market pricing model, Pruyn has used the Generalized Additive Model (GAM) as the basis for the MBG pricing model. GAM is used because it requires less data points to predict much more dependent variables. Currently, five main factors determine the (secondhand) vessels prices; namely age, size, vessel income, newbuilding price level and order size. However, there are more variables which might influence the secondhand vessel price which have not been considered in Pruyn's model. Thus, the first main focus of this study is ***to assess the relevance of other deterministic factors based on literature study and model assessment***. The second main focus of this research is ***to compare the application results of various data mining algorithms in building the pricing model***, to find the most suitable algorithm.

1.4. Research Objective

To conduct this research, the database from Clarksons is used as the main source of information where a vast array of data is available[7]. Succinctly, the main objective of this research is to improve the current secondhand vessels pricing model used in MBG by considering more influencing factors and statistical methods. These objectives lead to the following principal research question:

What is the most practical way to improve the current secondhand bulk carriers pricing model used in Maritime Business Game based on the available data and what is the implementation result?

The main research question will be answered systematically. First by clearly defining the secondhand market and its relation with other sub-markets. Secondly, the current model will be evaluated to see which potential improvements can be made. Afterwards, the relevant factors will be investigated based on the existing studies. Furthermore, the expected performance criteria for the improved model will be explained. Moreover, the potential statistical methods will be analyzed and the most fitting method is going to be selected. Lastly, all of these findings shall be implemented in the initial model and the augmented model is proposed. All these steps can be translated into following research sub-questions:

1. *What are the characteristics of the secondhand vessels market and how does it relate to other sub-markets in shipping market context?*
2. *What are the shortcomings of the current pricing model & which improvements are expected from an enhanced model?*
3. *"What are the suitable algorithms for building the structural pricing model?"*
4. *What are the relevant influencing factors to be considered in the secondhand vessel pricing model according to the literature?*
5. *"How can the points of improvement be translated into model set-up?"*
6. *What is the final result of applying the suggested improvement points in the current pricing model?*

1.5. Report Outline

This report is structured in such a way that it orderly answers the research sub-questions and then the principal research question. Chapter 2 reviews the (secondhand) shipping market theory. Afterwards, chapter 3 evaluates the current pricing model to set-up the benchmarks for a better performance. Correspondingly, the potential statistic models are introduced in chapter 4, along with the selection process of the potentially suitable models. Moreover, chapter 5 examines the potentially influential factors as long as ship sales price is concerned. What comes next is the data preparation and model set-up which is outlined in chapter 6. Afterwards, the modelling processes and corresponding results of selected algorithms are given in chapter 7 and 8. Finally, chapter 9 outlines the conclusion of this project and the recommendation for further research.

2
Secondhand Vessel Market

> Just as the weather dominates the lives of seafarers, so the waves of shipping cycles ripple through the financial lives of shipowners.
>
> *Martin Stopford*

In this chapter, the shipping business theory is given as a prelude to the secondhand market. The concept of shipping economics is introduced in section 2.1. Afterwards, the essential components that build up the shipping markets are elaborated in section 2.2. Subsequently, the secondhand market which is the focus of this project is highlighted in section 2.3. And lastly, the summary of this chapter is presented in section 2.4.

2.1. Shipping Economics

Shipping industry is a very volatile market as it is heavily influenced by the shipping cycles. The world seaborne trade volume has been growing considerably over the last centuries. In 2017 it is recorded that the total of 10.7 billions tons goods were shipped worldwide. Meanwhile, this number has doubled what it was 20 years ago, when there were only around 5 billions tons of seaborne trade worldwide[8]. Reciprocally, the total amount of global shipping capacity has increased vastly to catch up with the global shipping demand. In the same manner, over the last century there has been an major increase in shipping efficiency that leads to the reduction in shipping costs[4].

Despite the increasing shipping efficiency, volatility and risk still dominate the industry. Due to the heavy risk and volatile nature of the shipping industry, Stopford compares the nature of shipping market with the Poker game. Therefore, he suggests that it is essential to understand the dynamic of shipping cycle to play the business strategically[4]. In fact, economic cycles occur in almost all industries. As an example in the past, an early economist has observed a general 7-year cycle in corn prices. While shipping cycles have its own characteristics, they are very dependant on cycles of other industries. For examples, the cycle of oil, mining product and other commodities which are transported by sea. This implies that the freight rate movement is indirectly dependant on the supply and demand of goods which ships transport[4].

Following the economic theory by Adam Smith, the commodity price rises when the demand is higher than supply. Simultaneously, it is followed by more transport demand. Demand leads to the increase of the goods production until the supply capacity reaches the point of equilibrium. As the supply exceeds demand, the commodities price decreases. Shortly afterwards, it is followed by the downturn of the transport demand. Correspondingly, the commodities production will keep decreasing until it reach the trough where supply is much less than demand. Afterwards, the commodities prices will start raising again and the cycle will repeat itself[9]. In the next subsection, the key influences of the supply and demand of sea transport are explained.

2.1.1. Supply and Demand of Sea Transport

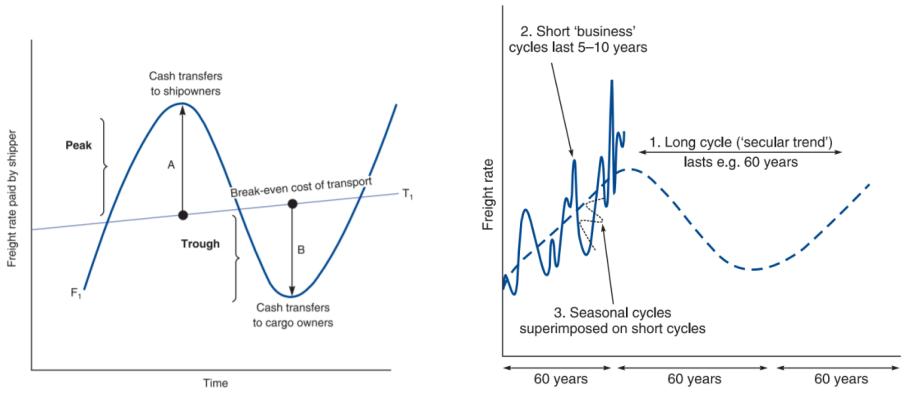
In the context of shipping market, each shipping cycle is unique and unrepeatable. It is mainly caused by the movement of supply and demand of sea transport. In order to get a better insight of shipping cycle, it is important to first understand which factors generate these cycles. From the many factors which influence the supply and demand movement, few most important key influences are selected and summarized in table 2.1. Each of these factors are elaborately explained and the discussion can be found in appendix A section A.1.

Demand	**Supply**
1. The World Economy	1. World Fleet
2. Seaborne Commodities Trade	2. Fleet Productivity
3. Random Shocks	3. Shipbuilding Production
4. Transport Cost	4. Scrapping and Losses

Table 2.1: Influencing Factors of Supply and Demand for Sea Transport. Source: Stopford[4]

2.1.2. Shipping Cycles

The movement in ship supply and demand leads to the shipping cycles. As any other cycles, shipping cycles consist of periodic peaks and troughs. During the peak, the freight rate is high and the cash flows into the shipping business. On the contrary, during the bad market, the freight rate is low and the cash flow out from shipping business into the cargo owners. Moreover, using advanced statistical technique, modern-day economists classify economic cycles in three different categories. These are long-term cycles, short-term cycles and seasonal cycles. The same concept is applicable in the shipping market context. Stopford has investigated the freight rate over the last 180 year and he is able to identify these components within the shipping market context[4]. His observation is illustrated in figure 2.1.



((a)) One Period of Shipping Cycle ((b)) Various Cycles Illustration

Figure 2.1: Shipping Cycles Illustration. Source: Stopford[4]

From figure 2.1, one can observe three different components which compose the global shipping cycles. The first most fundamental component is the **long cycles**. This cycle lasts in about 60 years. Long-term cycle generally follows the trend of world's economic cycle. The second component is the **short cycles**, or also known as business cycles. They last in about 5 year to 10 year. The short cycles are an important characteristic of shipping markets as they also mark the raises and falls of shipping companies. They efficiently eliminate bad players out of the shipping market[4]. The last type is the shortest and most periodic cycles called **seasonal cycles**. They occur between the short cycles with higher frequency and are caused by seasonal patterns of sea transport demand. More elaborate information is discussed in appendix A section A.2.

Shipping Economics Conclusion
Shipping industry is characterized by the shipping cycles. They are dependent from the economic cycles of transported commodities. Shipping cycles are observed based on the freight rate fluctuation. The biggest influencing factors that dictate the ship demand are the world economy, seaborne commodity trade, random shocks and transport cost. Key influences that control the fleet supply are world fleet, fleet productivity, shipbuilding production rate and removal of old vessels by scrapping or losses. Each shipping cycle is unique and not repeatable. However, based on the historical statistic, the estimated long-cycle is 60 year and short-cycle is 8 year[4]. The average cycle length from different eras can be calculated and summarized in the table 2.2. An interesting trend is that the gradual reduction in the cycles length for each era.

Era	Year	Total cycles	Average Peak	Average Trough	Average Total
Sail Era	1741 - 1871	7	6.1 Year	8.7 Year	14.9 Year
Tramp Era	1871 - 1937	7	2.6 Year	6.7 Year	9.2 Year
Bulk Era	1947 - 2007	8	3.0 Year	5.0 Year	8.0 Year

Table 2.2: Short Shipping Cycles Summary. Data is compiled by Stopford[4]

2.2. Shipping Markets

The most general definition of ***market*** is a place where two or more parties can meet with the purpose of exchanging goods[11]. In the past, a market is an actual place where people meet to trade goods. At this point, the definition of market has developed into any place where an economic transaction take place. For example the virtual market such as amazon is now considered as marketplace. The varieties of things which is traded has also evolved beyond literal goods into service, information, currency, financial derivatives and many others. The amount of sellers, buyers and money involved determine the size of the market[11].

Shipping market is a place where ships and shipping services are traded. Ships are generally considered as fixed assets. However, they are easy to liquidate as shipowners actively buy and resale of their assets. It is one of the markets where the capital assets are being actively traded[4]. It has various sub-segments, for example dry bulk, tanker, container and ferries. Each market segment has the different business characteristics. *The focus of this project lies on the dry-bulk markets* which is briefly outlined in section 3.2.1. Maritime economists distinguish shipping market into four categories, namely ***the freight market***, ***the newbuilding market***, ***the secondhand market*** and ***the scrap market***. Each sub-market has peculiar characteristics.

By nature, the shipping business is heterogeneous, dynamic and very competitive. Ships are heterogeneous product because of its vast array of varieties of which serve different purposes. Also, in most cases it is not possible to substitute one type of ship for another. For example, one cannot use the tanker to substitute the dredger's job to excavate the river bottom[12]. The competition within the business is so intense that the market is quite close to what classical economist describe as ***perfect competition model***[19]. Most prevalent similarity between shipping and perfectly competitive market are the availability of accurate information about market status and the accessibility of market for the new player[13]. To sum up, the important characteristics of each sub-market are outlined as following:

1. ***Freight Market***
   The freight market is the place where sea transport service is the main commodity. The main stakeholders are charterers and shipowners. Freight market has three sub-sectors, namely ***the voyage market***, ***the time-charter market*** and ***freight derivatives market***. The voyage market provides transport service for a single voyage. The time-charter market is characterized by periodic hiring of ships. Thus during this period, ship will repeatedly transport the cargo along a certain route. In the derivatives market, the stakeholders trade market derivatives using freight indices. Two main products are the Freight Futures Trading and Forward Freight Agreement. Freight rates revenue acts as the prime mover which govern the investors activities within the shipping business[4].

2. ***Newbuilding Market***
   The new building market provides newly constructed vessels. The major participants of this sub-market are shipyards, banks and shipowners. The majority of shipping industry cash flows out through this sub-market as the shipyard uses it to pay for materials and labour for ship construction. There are around 300 major shipyards in the world and many smaller yards[4].

   Few yards specialize in specific vessel type and majority of them is flexible. It takes somewhere between 1 and 4 year to build a new ship. Thus, the future expectation is one of the most important thing that drives shipowners' decision. Newbuilding costs are determined by the orderbook size which depends on the market situation[4]. Since 1960s, the newbuilding market of bulk carriers has stable growth. This means that bulk transport supply has steadily increased over the last centuries[4].

3. ***Secondhand Market***
   The secondhand market is the place to trade older vessels with shipowners, financial institutions and ship-brokers as the key actors. Old vessels investment involves agreement between shipowners and investors[20]. Regarding the cash flow, the secondhand market plays a more subtle role since the cash changes hand between the shipowners, yet it stays in the shipping business. In comparison to other sub-markets, sale and purchase market is the most volatile and speculative market. The current freight rate and investors expectation are the two biggest drivers of market fluctuation[4]. In extreme case, old vessels had higher value than new-build because of their fast availability. Moreover, this research will focus on secondhand dry-bulk market which will be further discussed in section 2.3.

4. ***Scrapping Market***
   The scrap markets are the places to dismantle the old vessels and to sell the deconstructed steel into steel market. This is the secondary income source for the shipping business. By referring back to the shipping cycle, the ships scrapping rate increases during the market trough or recession period. The scrapping price is determined by the steel price. Most of scrapping yards are located in China, Bangladesh, Pakistan and India. Sustainable ship recycling currently gains a global attention. This issue is also coupled with the economic circularity[4].

2.2.1. Markets Integration

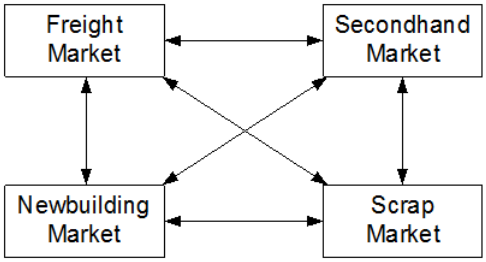


Figure 2.2: Shipping Markets Model. Source: Engelen, et al.[20]

Within the shipping business, each sub-market actively interacts with each other. Shipowners are the most important stakeholder since they involve in all the markets. The freight market is the major income source, thus the changing in freight rates influences the investors decision. This will create a domino effect on all the other markets. Secondary source of income is the scrapping market. The cash flows into the shipyards as the shipowners invest in new vessels. Otherwise, the money changes hand between the shipowners as they trade in the secondhand market. During the strong market, there are more activities in the newbuilding and secondhand markets. On contrary, the scrapping market becomes more active during the weak market. The interaction between these four markets with emphasize on balance sheet is depicted in figure 2.2[4].

In conclusion, this cash flow drives the shipping market cycles which was previously discussed in section 2.1. When transport demand is high, the freight rates rises and cash flows into the shipping business. It allows shipowners to invest more in secondhand and newbuilding. Several years after newbuilding vessels arrive in market, shipping markets are oversupplied, consequently the freight rate will go down. During the market trough, shipowners will sell or scrap their old vessels. Inefficient shipping companies will be force out of market. As the scrapping rate increases, the ship supply decreases. When the transport supply goes below transport demand, the freight rates start increasing again and the cycle restarts[4]. In the next section, in depth discussion about the focus of this project, the secondhand vessel market is evaluated.

2.3. Secondhand Vessels Market

Among other commodities markets, secondhand vessel market has extremely unique characteristics. By studying the diagram in figure 2.2, one can conclude that this market segment plays an active role in shipping market. As what being previously discussed, shipping markets are very cyclical. Thus, the timing of investment holds a very important place here, especially in the secondhand market. Extreme profit can be obtained here by playing the buy low and sell high technique. Secondhand ship values follow closely the freight rates. Thus when the freight rates are low, ships price falls. As a near perfectly competitive market, new investors can easily enter this market by participating in selling and purchasing the readily available vessels[4].

As what was mentioned, this is one of the few places where the players so easily trade the capital assets worth tens of millions of dollars. For this reason, many econometric researchers have been devoting their time to study the price fluctuations trend of secondhand vessels[23]. Many different models based on various economic theories have been built[21][24][26]. Different advanced statistical technique have been used and developed to build reasonable models[36]. This research project will focus on improving a pricing model for secondhand bulk carriers. But before going further into the modelling part, this section will discuss deeper about the normal practice in the sale and purchase market.

The key players of secondhand market are the charterers, shipping companies (including shipowners) and investors who speculate to obtain profits. Based on the freight rates fluctuation, shipping companies on behalf of shipowners will buy or sale secondhand vessels in order to raise or to spend cash. Besides the future expectations, there are also several other reasons why shipowners put their vessels on sale. For example, several shipping companies have a regulation to replace the vessels when they reach a certain age to avoid high maintenance cost, with little regard to the shipping market status. In another case, during the depression time (trough of cycle), shipowners are 'forced' to sell their ships in order to fund their daily operation[4].

Status of ship during sales might varies, for example a ship might be sold with or without ongoing charter. In the same manner, purchasers of secondhand markets have a variety of reasons behind their transaction. It can be a future expectations of raising freight rates, so speculators take a chance to play their buy low and sell high strategy. Based on the same expectation, shipowners will buy readily available vessels to anticipate the raising freight demands in the near future. In other case, a shipowner will buy a readily available vessel with specific criterion in order to bid for a new shipping contract[4].

In secondhand market, shipbrokers acts as the intermediaries to facilitate the economic transaction between shipowners. Shipowners will make use of broker service to valuate and to find potential buyers for their vessels. Valuation is mainly done based on the shipping market economic status. This is because most of the shipbrokers have no qualification to do the technical assessment. To avoid the price bias, it is common for shipowners ask multiple brokers to conduct the valuation and using the average price[4]. The brokers will actively seek the potential buyers which look for vessels with particular specifications.

In the depressive market, buyers are more selective and careful since there is low competition. They will conduct the technical inspection and carefully bid for the vessel. In the high market, the competition is high, thus ship inspection is not likely to be done[4]. When the buyer is found, shipbrokers will help the seller

and purchaser to finalize the economic and legal transaction. While different authors might have different opinion, Stopford suggests the following steps[4]:

1. ***Advertisement Stage***
   When a shipowner wants to sell his vessel, he normally will contact a broker to valuate and to put his ship to sale. In rarer case, shipowner will decide to take care of the sale himself. The advertisement about this ship will then be circulated between interested parties within the market.

2. ***Negotiation Stage***
   Once potential buyers are found, they will discuss the price with broker or directly with shipowner. The negotiation process is heavily influenced by the state of market. During the peak of the cycle, a buyer has to compete with several other buyers. Thus he has to make a quick decision based on limited information as they might not have a chance to inspect the ships. In the depressive market, there will be less buyers, thus buyers will have the privilege to conduct an inspection and to negotiate the price if ship is in less than perfect condition. Once an agreement is reached, the broker will prepare a contract.

3. ***Contract Stage***
   When the buyer and seller agree on a certain price, the broker will prepare a purchase agreement. This contract is called the **Memorandum of Agreement** which specifies the details terms and conditions of the ship purchase. There are several possible framework that is used for Memorandum of Agreement. The most commonly used version is the Norwegian Sales Form (1993). This document consist of information such as the transaction administrative details (such as the place, date, price, terms and method of payment), technical conditions (including documentation by classification society) and contractual rights based on agreement between the parties.

4. ***Inspections Stage***
   Depending on the market status, both parties might agree on inspection. Then shipowners will ask surveyor to assess the ship. Detail inspection based on dry-docking or underwater inspection by diver might be done. The buyer might also check the status of the ship with respect to the classification society. When significant defects are found, the sale may fail. Otherwise, the the potential buyer might use this to further re-negotiate the price of ship until a new agreement is reached.

5. ***Closing Stage***
   To conclude the ship sales, the seller will deliver the ship on agreed conditions (time and place). Afterwards, the buyer will transfer the money to the appointed bank. Lastly, shipbroker will help seller and buyer to finalize the legal process and to complete all the necessary documents for both parties.

2.4. Chapter 2 Conclusion

In this chapter, the theory behind shipping economics and shipping markets are discussed. Shipping economics is driven by supply and demand of sea transport, therefore it has a cyclical characteristic. Shipping markets are generally divided into 4 sub-markets namely freight rate markets, newbuilding market, secondhand market and scrapping market. They interact with each other and the change in one sub-market affect the other market segments[4]. This research focuses on the secondhand market. To conclude, this chapter has answered the first research sub-question, namely *what are the characteristics of the secondhand vessels market and how does it relate to other sub-markets in shipping market context ?*. In the next chapter, the current pricing model of Maritime Business Game will be reviewed.

3
Maritime Business Game

> Learning through a game motivates the students to learn better and retain the knowledge longer.
>
> *Jeroen Pruyn*

In this chapter, the existing pricing model for secondhand bulkers market is reviewed. As a prelude, the Maritime Business Game (MBG) is introduced in section 3.1. It starts from the explanation of development process of the game, followed by the highlight about the shipping market model used in MBG. Furthermore, the literature review about the practical and scientific valuation process for the secondhand vessels is presented in section 3.2. Moreover, the potential points of improvements for the current model are suggested in section 3.3. By this, the second research sub-question, *"What are the shortcomings of the current pricing model & which improvements are expected from an enhanced model?"* is answered. Lastly, the conclusion of this chapter is outlined in section 3.4.

3.1. A Game for Maritime Education

As what was previously explained in chapter 2, it is important to get a solid understanding about how the shipping market works in order to manage the shipping business appropriately. The best way to learn about something is by doing it. This is also the case for the shipping business. However, during the "real-life" learning process, one bad decision may cause millions of dollars loss. This is due to the capital and risk intensive nature of shipping business. For this reason, a joint initiative is taken by Delft University of Technology (DUT) and University of Antwerpen (UA) to develop the Maritime Business Game. It is a "safe" simulated environment where the players can experience the dynamic of shipping market.

They will also gain an insight of probable consequences of certain decision relative to their own "shipping company". It is "safe" because the consequence of a bad decision will only cost "simulated" millions of dollars. Since the time is simulated, players would be able to understand the long-time consequences within a short playing time. For years, the game has been played yearly by Maritime students of DUT, UA and University of South Eastern Norway. During the game, the students act as various shipping companies within the shipping market. The main objective of shipping companies is to optimize their performance which is measured by comparing the companies' Return On Investment (ROE) at the end of the game. Altogether, MBG has proven itself to be a successful educational tool which students enjoy to play [18].

While MBG is not the first business game created, it has several distinctive features in comparison to the already existing games. Its main purpose is to simulate the medium or long term consequences of decision taken in various shipping scenarios, focusing on the dry bulk market. The most prominent feature is the integration of all relevant economics levels which participate in the shipping business. The students act as the shipping company which manage the fleet operation. On contrary, the shipyards role is currently played by the computer to manufacture standardized vessels. In the current version, MBG only has the shipping

market as the playable entity, however in the future, it will incorporate the ship yards[20]. Moreover, the structure of MBG integrated model will be reviewed in the next subsection.

3.1.1. Integrated Model Review

As what was previously mentioned in chapter 1 in section 1.2, MBG is built upon an integrated model which consist of three modular entities. Combining these entities will allow this integrated model to capture all the relevant sectors within the maritime economy. Starting from the **Multi-Country Model** which represents the macro-economy status of chosen countries. These countries are chosen based on their reputation as the world's leading maritime capitals. It is followed by the **Shipping Markets Model** that represents the interaction between the shipping markets. Finally it ends at the **Fleet Scheduling Model** which manages the match between fleet and cargo availability of the dry-bulk market. Each module actively interacts with each other and the overview is given in figure B.1 in appendix B.

Looking at previous research, many individual models on each level are already available. Large research institutions (like OECD and NIESR) and financial institutions are the most 2 common parties who developed the Multi-Country models[5]. In the last 3 decades, they use Multi-Country model to evaluate the possible after-effects of implementing a certain monetary policy. On the other hand, researchers have been evaluating the Shipping Market model since the 1930s. Two major research purposes are to assess the vessels' potential future earning and to optimize the vessel management that leads to the operational cost reduction. For this reason, most of the models are focused on a specific shipping company and have rather limited applicability. However, most of models are stand-alone model and MBG is the first integrated model.

At this point, MBG integrated model is proven to be robust and stable[5]. Besides for educational purpose, this integrated model could be used for another essential purpose, namely as a simulation tool. An appropriate integration of these three modules enables MBG to generate simulations which are consistent in all levels. Another advantage is because MBG has a more generic applicability in the context of bulk shipping. Thus it would be possible to use MBG to simulate a certain global scenario change to identify the possible aftermaths in shipping market level. Lastly, it could be used by the industry to to test the possible company-level outcomes of applying certain business or operational strategy in their business.

As a side note, generating the possible medium or long-term consequences is not essentially the same as predicting the near future of an interested parameter such as the freight rate. However this implies that one can see the general aftereffects which would be experienced by shipping companies when a certain event takes place. For example, one may ask if the shipping business would be positively affected upon the opening of North Sea route. If this event would ever take place, the MBG simulation predicts as following:

> *"Opening up the Northern Sea Route for the dry bulk trade between Europe and Asia will only be beneficial if either the rates in shipping far exceed the cost price or the fee for passage drops at least below 15 USD per Gross Tonnage.[17]"*

From this simulation, one can see the minimum required condition to make the Northern Sea Route passage to be economically viable. In the same way, it is possible to simulate other scenarios to see the implication of that events from the shipping business perspective. In the next subsection, the development process of MBG and the interaction between the modules are explained.

3.1.2. Model Framework Development

An initial framework is designed by Pruyn to fit together all the relevant entities. Modularity is a central concept in the MBG development. Modular means that each sub-module is built separately, mainly by borrowing and tailoring the already existed (individual) model. The developer argued that this approach has accelerated the game development process, however considerable adaptations had to be made. Another challenge is in the fact that different input data is needed by different sub-models. Lastly, the integration process requires lots of effort since developer has to make all sub-models works collectively, while retaining the modular struc-

ture. The main advantage of modular structure is that one can make an update in a single module, without having to make extra adaptation in the complete model.

The MBG framework is presented in figure B.1 in appendix B. The three main modules are highlighted with red box and the flow of information is represented by arrows between modules. Economic scenario is the starting point before the simulation which has to be first adapted. However, when being used as an educational gaming tool, no scenario adaptation is needed since the main purpose is to educate students. The only playable entities are the shipping market module and the fleet scheduling module[18]. With a given capital, students determine which type of vessel they are going to acquire and which freight assignment they are going to take. Students have much influence on the Shipping Market and Fleet scheduling module, while they can only respond to the situation created in the country level.

Shipping Market is the main focus of this project, so observing the flow of information around this module might be useful. By observing the flow of information in figure B.1 in appendix B, it can be seen that the main input for Shipping Market module comes from Multi-Country module, which generates relevant economic information. This country-level information is especially relevant to the newbuilding market since the price of a new vessel is heavily dependant on the country of yard. In short, variables such as national wage, exchange and inflation rates determine the final cost of the new vessel. Conversely, the secondhand model has a global characteristic, thus they are not influenced by country level information[5].

Furthermore, Shipping Market model generates the information about the total number and the price of all vessels in the system which represents the ship supply. This information is then needed as input to Fleet Scheduling module, which in turn will generate the transport demand. This includes the Freight market related information such us shipping assignments. After processing the input from Shipping Market model, this freight information acts as a feedback input which is routed back to Shipping Market model. Nevertheless, the further explanation about 2 other modules is outside the scope of this study. Thus in the next subsection 3.1.3, the details of shipping market model is discussed.

3.1.3. Shipping Market Model Review

Shipping Market module is one of the three main modules within the MBG integrated framework. In a nutshell, every sub-market, except the freight market, are modelled as part of Shipping Market module. The freight market, where the freight contract is arranged, is modelled separately as a part of Fleet Scheduling model. The two most important consideration in the MBG shipping market module are the **price/ cost** and total **number of vessels** in the system which represents the *transport supply*. After generated by the Shipping Market module, these variables will flow further as the input to Fleet Scheduling module[5].

The ***vessel's price*** is meaningful in all sub-markets model. In newbuilding and secondhand trade, price/cost indicates the cost of capital. As for the scrapping market, the price does not have a significant role except as a means to compare the scrapping price and secondhand price. Based on this comparison, the economically advantageous decision is automatically generated by the system, whether to scrap or to sell their old vessels. Newbuilding and secondhand price is modelled separately[5]. However, modelling the scrapping price is less complicated, thus it is modelled as a part of secondhand price. On the other hand, **number of vessels** in the system is only determined by the activities in the newbuilding and scrapping market.

New order in newbuilding market increases the ship supply while order in scrapping market reduces ship supply. However, trading in secondhand market does not change the total amount of vessel in market, thus order quantity is not a relevant variable for secondhand market. At last, Ship supply information from Shipping Market model, coupled with transport demand information from Multi-Country model, will determine the freight rates and available assignments. The order quantity is modelled separately for newbuilding and scrapping market[5]. This can be seen in figure B.1 in appendix B, where the relationship between relevant variables within the building blocks of shipping market model are highlighted.

Furthermore, from figure 3.1 one can see what are the main variables in different market segments and how are they related to one another. For example, *secondhand price* variable is related to the *newbuilding order* quantity by the demand and price arrows. It means that when newbuilding demand is high, the secondhand price will positively be influenced. This is because due to a longer waiting time for delivering the new vessel, some shipowners will prefer to buy secondhand vessel as an alternative.

Figure 3.1: Building blocks of the Shipping Market Model. Source: Pruyn [5]

Similarly, higher newbuilding demand also means higher newbuilding price, following the supply and demand theory. For this reason, both the newbuilding order and newbuilding price are also related by the *price* and *demand* arrows. Furthermore, in high market, comparing the newbuilding and secondhand price is a common practice by shipowner to establish their purchase decision. Thus these two variables are connected by bidirectional price arrows.

Similarly, in low market, shipowner will compare the secondhand and scrapping price to determine their sale decision. For this reason, there is only one-directional arrow goes from secondhand price to scrapping order quantity. It signifies that when scrapping price is better than secondhand price, the shipowner will automatically sell their old vessel and this influences the scrapping order quantity. Among these important variables, this study focuses exclusively on the secondhand market. Thus, despite of the importance of other variables from other market segments, the newbuilding and scrapping markets will not be further explained. To finish, this section has discussed the development of MBG framework and its shipping market module. To continue, the review about the valuation practices is discussed in the next section.

3.2. Review of Secondhand Market Models for Dry-bulk Carriers

Before discussing what can be improved from the MBG pricing model for dry-bulk carriers, it might be relevant to ask two questions. These are ***'What is the essence of the dry-bulk market?'*** & ***'How secondhand vessel pricing model is usually modelled?'***. A brief answer to the first question is outlined in subsection 3.2.1 while the answer to the second question might lead to a prolong discussion. The most common valuation methods which is practiced by the certified brokers without technical knowledge[4]. Two most important factors are the market status (based on supply and demand principal) and expected future income. For the sake of clarity, the broker's valuation procedures are explained in subsection 3.2.2.

Although brokers' experiences are undeniably valuable, brokers do not thoroughly consider the vessel's technical characteristics[23]. Historically speaking, overlooking the technical parameters might lead to inappropriate estimation which cause millions of dollar loss in this capital intensive business. There are several reported cases where the price suggested by one broker company is completely different than another broker [30]. Thus, utilizing the scientific valuation model might be a good way to complement the normal broker valuation process. The review of scientific valuation models are given in subsection 3.2.3.

3.2.1. Dry-bulk Market Overview

Bulk carriers are traded in MBG shipping market. These vessels have special design and physical characteristics compared to other types, as their main purpose is to carry bulk cargoes. The major dry-bulk commodities are *iron ore, coal & grains* which make up almost two-thirds of global dry bulk trade[91]. Furthermore, minor bulks accounts for the remaining one-third of dry bulk trade. These consist of products such as agricultural products, fertilizer, steel products and minor ores (alumina, bauxite, nickel ore)[91]. The geographical locations of producers and consumers of three major bulks mainly determine their operational area. Besides oil, coal is used as the main energy source due to its abundance, affordability and wide distribution across the world. It is predicted that coal will soon replace oil's position as the largest energy source[93].

As for 2018, world's top-3 coal exporters are Australia, Indonesia and Russia and coal is distributed worldwide since it still accounts for almost 40% of the world's electricity production[92][93]. Besides coal, iron ore is another very important commodity. As the fourth most abundant elements, it plays a critical role in civilization since thousands of years ago. Due to its abundance and price advantages, iron can be found in almost every infrastructure, vehicles and modern tools[94]. As for 2018, world's top-3 iron-ore exporters are Australia, Brazil and South Africa and 65% of them are shipped to China[95][96]. Additionally, distinctions are also made based on the ships size. Based on their size, bulk-carriers are divided into four categories, namely ***Handysize***, ***Handymax/ Supramax***, ***Panamax*** and ***Capesize***, as summarized in figure 3.2.

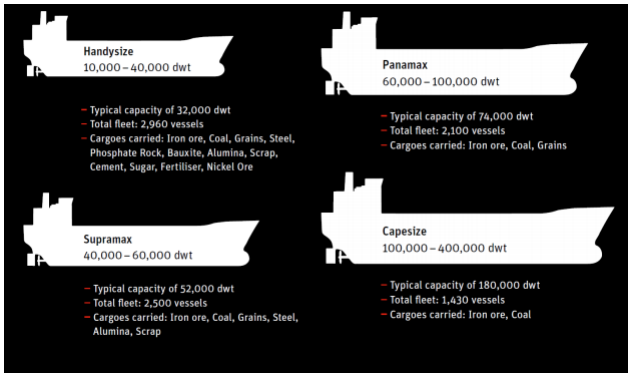


Figure 3.2: Bulk-Carriers Classification based on size. Source: Schmitz [98].

Smaller vessels are the workhorses of the dry bulk market. This sub-market has been undergoing a transformation from a "traditional" market with limited cargo types to a niche market with diversified cargo types. The market is now dominated by conventional handysize and handymax vessels with growing numbers of specialized vessels[98]. Contracts are usually short-term, thus it provides "flexibility" for the shipowners. The most important players are Norwegian and Greeks operators[108]. Norwegian operators generally focuses on advanced technology and provides new technical solutions.

Panamax is the largest vessel that can pass the Panamax canal. The demand for this market mostly comes from the industries buying coking coal, power stations buying steam coal and national government buying grain[98]. Since the last decade, the types of cargo have diversified. The Panamax market provides more transparency in comparison to other types, thus there is a fierce competition in this segment[108]. There are a substantial amount of large panamax operators. They consist of private companies and shipping pools. Bigger operators are engaged in Contract of Affreightment and other long-term contracts, while the smaller operators handle the voyage market[98]. The main players come from Far East, Norway and UK[108].

Capesize has size advantage and due to applicability of the economies of scale[105]. However it cannot pass the Panamax canal and thus has a limited the operation area[103]. Also, only few ports can accommodate a fully loaded capesize due to its large draft[106]. Market demand comes mostly from iron ore and coal business with few chances of minor bulk and grain[108]. The cargo has mostly low value (per DWT) and

requires homogeneous (not-specialized) service. Generally, this market has the purest feature of free competition due less diversified transported commodities. Due to the homogeneous service, Capesize operators experience severe competition. This market is dominated by small, independent owners[108].

3.2.2. Practical Valuation Procedure

In the real life, the valuation is done early in purchase procedure to determine the market price in order for the seller to make a public offering. Valuations is also necessary for aspirant buyer and bank to determine how much loan the buyer will take. Without a definite collateral value, financial institution will not give mortgage to the buyer[4]. This collateral value will be used to make purchase documentation and used by the company for the bookkeeping purpose. In some cases, this is done to issue company's bonds or prospectus based on their purchase. Thus, ship valuation is mainly done for accounting and financial purpose. Just like other assets, ship valuation is done based on professional standards following specific commercial guidelines[4].

Despite of the available guidelines, ship's value mainly reflects the broker assessment of ship's market value at certain date. *Valuation is mainly done based on a common assumption that vessels are in good condition[4].* Thus, there is a tendency to overlook ship's physical condition. Generally, high market means more expensive vessels and weak market mean cheaper vessels. Since the valuation is done based on broker's judgement, it is sometimes reported that 2 brokers give entirely different vessel value based on their expectation[23]. If there are no willing buyers, ships price will be considerably discounted. After valuation is done, when both parties agree on inspection, it will be done by the aspirant buyer. There are four well-established valuation procedures which are used by the shipbrokers, namely:

1. ***Market Approach***
   This method suggests that a vessel is valued based on market comparison with similar vessel, mainly the last done transaction[14]. Price adjustment is made based on age, cargo capacity and other technical specifications. The problem comes during the inactive market period, or for a vessel type which is not commonly sold. In this case, making comparison will be difficult. The vessel's value mainly reflects the market's condition[15].

2. ***Replacement Cost***
   This method specifies that vessel is sold as much as the price needed to replace the vessel at the present time according to its current worth. In this case, ship is valued based on its bookkeeping value. It is less common to use this method for ordinary merchant vessels. This method is usually applied for heavily customized or refitted vessels because comparison is difficult to impossible to be made. For example, this method is used to value a bulk carrier which is refitted to have a hotel for a total of 120 people on board[14]. Usually, discounted replacement cost is used[15].

3. ***Income Approach***
   This mostly used method indicates that ship value should reflect its future freight revenue prospect. The ship is valued using Net Present Value of all expected future earnings plus the expected scrapping value of vessels. Normal operating age used in this method is between 25 and 30 years. Assumptions are made regarding vessel's operating expenses. This approach provides a rational value, however might be inaccurate since many assumptions are made[14].

4. ***Hamburg Ship Valuation Standard***
   Also known as Long Term Asset Value (LTAV), this method is based on a Discounted Cash Flow model. This method completely excludes the market cycle effect and it focuses on long-term earning potential of the ship. LTAV Association claimed that while this method should not be used to substitute the traditional income method. However, this can be used as a supplement that particularly useful during the disturbed market[16].

To illustrate the difference between these methods, a calculation of ship value is done using different approaches. As an example, these approaches are used to determine the price of secondhand tankers to be delivered in 2009. The computation result is presented in table 3.1. Based on the result, one can conclude

that there is a slight variation of result between the first three methods. However, Hamburg method delivers a much higher price since the market influence is excluded. To conclude, this sub-section has elaborately discussed the practical valuation procedures which are used by shipbrokers. Beside these procedures, the other scientific alternative is outlined in the next sub-section.

Valuation Method	Tanker Size		
	52k DWT	105k DWT	300k DWT
Market Approach	$34,000	$53,000	$96,000
Replacement Cost	$37,000	$52,000	$98,000
Income Approach	$34,000	$46,000	$91,000
Hamburg Approach	$59,000	$80,000	$150,000

Table 3.1: Tanker Prices based on different Pricing Approaches. Calculation Source: Tanker Operator Magazine [14]

3.2.3. Scientific Valuation Approach

Within the secondhand context, there are two main valuation methods that have been developed so far. The first one is the *theoretical structural method* which uses the cross-sectional data. This approach suggests that the vessel price can be inferred from determining factors such as vessel's age and expected future income. The second way is by the *atheoritical time series method* which suggests that the future vessel price can be determined by the historical prices of the vessel. The existing methods are discussed based on their timeline.

Beenstock Model & Efficient Market Hypothesis Discussion

The first econometrist who actively attempted to model shipping market and pricing model for secondhand vessels was Beenstock in 1985[21]. He argued that supply and demand function is not sufficient to represent the secondhand vessels price formation. Furthermore, Beenstock and Vergottis suggested a structural model as pricing instrument for secondhand drybulk vessels[22]. However, due to assumptions which are taken, Beenstock model is controversial and received serious criticisms.

The main criticism is due to the assumption that market efficiency is present in shipping market. This means that the shipping market behaves rationally. Thus, there is no correlation between one commodity price with the other. The second heavily criticized assumption is because they considered the perfect correlation between newbuilding and secondhand price. Thus, the price between these two commodities is perfectly interchangeable[23]. As the result, not long after Beenstock model is proposed, many researchers dedicated their time to evaluate the validity of Efficient Market Hypothesis (EMH) in the shipping market[6].

In 1992, Hale & Vanags evaluated the long-term correlation between secondhand vessel prices and concluded that EMH does not hold due to the existence of correlation between commodities[24]. However, this research is further advanced by Glen in 1997. He also found the correlation between prices but argues that EMH might hold in this situation given that the common stochastic trends is present. In this instance, Glen did not explicitly reject EMH[25]. In 2002, Kavussanos and Alizadeh also tried to test the EMH validity and concluded that shipping market is inefficient[26]. In short, the previous researches over the validity of EMH in shipping market prevail to reach a solid conclusion about its validity.

In addition, Tsolakis et al rejected the second assumption in Beenstock model regarding the perfect correlation between the newbuilding and secondhand vessel price. He suggests that since newbuilding market is not as volatile and as speculative as the secondhand. Thus, newbuilding ship price is different than secondhand vessels[27]. Also, the newbuilding price and secondhand price in most cases are not similar, so a conclusion can be made about the invalidity of the second assumption. After Beenstock, the other secondhand price estimation model is proposed by Strandenes in 1986 using a simple Net Present Value.

She claims that secondhand values is equal to the weighted average of short and long term profit. She suggested that secondhand price is influence by the future expectation of the development of other market sectors. Individual cash flow is substituted by the average of expected yearly income in this case[29]. However,

this method is disputed by Kavussanos and Alizadeh in 2002. Specifically due to the unrealistic assumption that the ships have infinite economic lifetime and inclusion of a depreciation factor. In further research, they have re-modelled the work of Strandenes with more pragmatic expectation; namely the usage of a finite economic lifetime and a realistic depreciation factor[28]. However, the change of assumption has resulted in a different outcome which implies that Strandenes method might be unreliable.

Time Series Model

In response to the earlier structural model proposed by Beenstock, other researchers implicitly rejected it by moving to another direction. Namely by using the atheoritical pricing model which is based on the time series values of the generic vessels. This research direction is also supported by the development of more advanced statistical techniques which improves the overall quality of time series analysis. The first prominent time-series model application is done by Kavussanos in 1996. He examined the volatility in the drybulk and tanker markets by applying the Autoregressive Conditional Heteroscedasticity (ARCH) technique. He concluded that the volatility varies across ship sizes with smaller vessels being less volatile[31].

Afterwards, Glen and Martin conducted similar research in 1998 using different set of data and methodology. Finally, they reached the same conclusion as Kavussanos[34]. Another model is made by Kavussanos in 1997 to predict the freight market, using the Autoregressive (Integrated) Moving Average (AR(I)MA)[33]. Moreover, by Vector Autoregressive (VAR) co-integration method, Veenstra and Haralambides established the secondhand ship prices which are stationary in first difference. Veenstra believes that the value of vessel is proportional to the discounted sum of its future income[32]. Following this line of research, Haralambides et al applied the *Theoretical Error Correction* model to estimate the newbuilding and secondhand price[23].

Back to the previous issue regarding EMH, by applying the Discounted Cash Flow approach, Veenstra implicitly agreed that the shipping market can be considered as efficient[6]. DCF along with Net Present Value techniques are commonly used to consider an investment decision. They have been extensively studied and adapted to include a realistic risk-factor. Several recent examples of their applications in other industries are for aircraft acquisition[41] and for real estate investment[40]. In shipping market context, Bendall & Stent used this improved version to consider investment in an express liner service[39].

The use of DCF or NPV approach agrees on EMF because such expectation about future income will only be valid when the market is efficient. Therefore, DCF approach will not be considered further in this project since the previous research could not reach a solid conclusion about the validity of EMH. Furthermore, by considering the previous comments about time-series model, an improved pricing model is proposed by Tsolakis et al in 2003. They tried to bridge the time series with structural models by using the *Error Correlation model*. While accepting the criticism about the lack of theory, Tsolakis et al supplemented their model with the maritime economic theory and solved the correlation problem in the structural models[27].

Tsolakis study is further extended by Thalassinos and Politis by using a more advanced time-series technique, *Vector Error Correction Model* (VECM). They used the cash flow analysis and co-integration approach. Using the supply and demand principal, they consider more variables in comparison to Tsolakis. These are secondhand price, US treasury bond, operating profit & industrial production index. They concluded that small vessels are the most market driven. Furthermore, the handysize and capesize are more capital intensive compared to larger vessels[30]. Lastly, Geomelos et al used linear panel data model which combines the time series and cross-sectional data analysis. They concluded that spot rates, newbuilding prices and vessel size have considerable influence on secondhand price[37]. These are the most relevant time series models.

Structural Model

Reflecting on the discussion above, time-series models seem to be more reliable in regards to the consistency of the result and the absence of unrealistic assumptions. However, an important remark is made by Adland and Koekebakker in 2007 about time series approach, namely regarding the data source. They pointed

out that the published (time-series) values are usually based on the broker best estimates for certain type of vessels[35]. Furthermore, reliance on broker judgment may introduce subjectivity and human error in measurement. *Since the subjectivity is the main thing to avoid in this research, the time-series approach will not be further considered in this project.*

Looking back to the already existing researches, the second significant structural model is proposed by Tsolakis in his PhD dissertation. In contrary to Beenstock, Tsolakis utilized the supply and demand theory as the base for his model. He considers the time-charter rate, newbuilding price and LIBOR as the influencing factors[38]. Unfortunately, he does not consider vessel age which might be the most influential variable regarding the sales price. Additionally, he used the time-series data in a yearly level which causes this approach to be flaw and will not be further considered for this project.

By the same token, Adland and Koekebakker introduced a potentially more objective approach; namely by analyzing the actual sales data instead of the constructed time series values. They propose the usage of generic market factor (such as time charter rate) as well as the vessel individual characteristics (such as age). They aimed to investigate the available cross-sectional (sales) data using non-parametric approach, namely the Multivariate Density Estimation (MDE). An advantage of using such a non-parametric approach is that it can capture the non-linearity that might present in analyzed variables[35].

While Adland and Koekebakker have proposed an ingenious approach, their model has a shortcoming. Namely the number of explanatory variables, because only three variables from a moderate size of sales data are analyzed. These variables are ship size, age and time charter rate[35]. However, these variables are not sufficient to explain the observed market price because other important factors such as type of engine and fuel consumption are not taken into account[36]. The main reason lies in the MDE technical limitation, namely its inability to take up a large numbers of selected variables. In short, the model would suffer when large number of variables are considered.

Therefore, Adland and Koekebakker suggested that in the future research, more relevant variables should be taken into account in a semi-parametric framework. An extra advantage of semi-parametric model is that it can take both non-parametric and parametric components (such as dummy variables), and therefore it has the potential to produce a more reliable result[35]. In response to Adland and Koekebakker remarks, the following research in the line of micro-economic approach is done by Koehn. He suggested to use Generalized Additive Model (GAM) instead of MDE[36].

GAM is a more powerful semi-parametric model with capacity to undertake more variables and to capture the possibly non-linear relationship between them[36]. GAM is an extension to the *Generalized Linear Model (GLM)*, combined with the *additive model*. Kohn considered much more influencing factors; such as newbuilding price level, number of tanks, pump capacity, speed, horsepower, classification society, IMO rating, hull type and country of build[36]. Up to the point when MBG was created, Pruyn argues that Kohn's research can be considered as the most complete work which provides reliable results[6]. *Thus, adopting Pruyn's point of view, Kohn's method can be seen as the best structural model created up to 2013.*

Initial Model Review

To sum up, Pruyn has "borrowed" Kohn's approach to model the secondhand market for MBG[5]. Pruyn used GAM because it requires less data size to predict more dependent variables. Pruyn obtained the ship sales data from Clarkson, dating from June 1998 to December 2010. This database consists of various type of bulkers, namely 3228 handysize vessels, 1254 panamax and 444 capesize[5]. Pruyn has considered 6 price determinants, namely age, size, vessel income, newbuilding orderbook level, cost of finance and bunker price. These variables are considered based on literature study and the assessment of the strength of relation. At the end, bunker cost was eliminated because it was insignificant. In this study, this initial model will be improved. This could be done by assessing more factors and studying the applicability of other statistical methods.

Literature Review Conclusion

To conclude, the practical and scientific valuation procedures are evaluated. In addition, the characteristic of initial model is shortly outlined. Kohn's approach is voted as the best technique and used in the initial model. GAM is undoubtedly a suitable algorithm for the structural model[5][36]. However, GAM is not the only algorithm for structural modelling. Data mining has experienced advanced growth in the past few centuries[64], thus there are many other algorithms that are still unexplored. Furthermore, the improvement goals of current model are discussed in the next section.

3.3. Improvement Goals

After analyzing the shipping market and reviewing the literature, a better understanding is obtained. Based on these efforts and an interview with Pruyn, three improvement points are finally suggested, namely:

1. **Increasing Model Sensitivity**
   Additional features such as ***high*** $\frac{Volume}{DWT}$ might influence the type of cargo which ships can carry. High $\frac{Volume}{DWT}$ is suitable for low density cargo such as grain whereas low $\frac{Volume}{DWT}$ is suitable for high density cargo like steels. Furthermore, having ***ice class*** and ***crane*** on-board is typical for smaller vessels[91]. having an ice class will enable vessels to take cargo to the Northern hemisphere during the winter. Lastly, having a crane will enable vessels to sail to crane-less ports such as Lamu and Bagamoyo in Africa[97]. These factors are not yet considered in the current model, thus analysing effects of these features in handysize and handymax submarkets, might lead to a useful result.

2. **Integrating Additional Influencing Factors**
   In addition to the used variables, Pruyn and other literature suggest that a better model would consider more influencing factors. This is because inclusion of additional factors might potentially increase the model accuracy and reliability. For example, factors like speed, country of build and classification society might have a considerable effect on sales price[36]. Therefore, more factors based on further literature study will be taken into considerations for the future model. In addition to literature study, logical analysis and significant test will be performed to determine their importance.

3. **Compacting the Final Mathematical Equation**
   The vessel price is estimated as a function of relevant variables. The resulting graphs are produced by a non-parametric approach. This means that there is no limitation on the graphs' shape as long as they are continuous. Several variables such as ship size, age and orderbook level are characterized by non-linearity. To translate them into mathematical equation, Pruyn used "discontinuous approach". The disadvantage of using this approach is there are multiple mathematical equation needed to represent one graph. For each variable mentioned above, there are 2-3 mathematical equations needed. While these equations can still be used, this situation is not particularly desirable.

To sum up, the the current MBG pricing model is reviewed with three suggested points of improvements. This particular information answers the second research sub-question, namely *"What are the shortcomings of the current pricing model & which improvements are expected from an enhanced model?"*.

3.4. Chapter 3 Conclusion

In conclusion, An elaborate explanation about MBG framework is given in section 3.1. An emphasize is given to the MBG shipping market model. Furthermore, the literature review of dry-bulk market and valuation procedures is elaborated in section 3.2. Lastly, the potential points of improvement are suggested in section 3.3. By this, the second sub-question, *"What are the shortcomings of the current pricing model & which improvements are expected from an enhanced model?"*, are clarified. Three main improvement goals are **Increasing Model Sensitivity**, **Integrating Additional Influencing Factors** and **Compacting the Final Mathematical Equation**. Considering the advanced development in data analysis, one may ask whether there are any suitable data mining algorithms to build a structural model. This shall be answered in the next chapter.

4
Data Mining Algorithms

> All models are approximations. Essentially, all models are wrong, but some are useful.
>
> *George Box*

The data mining algorithms can be used for various purposes, and one of them is to build the structural pricing model to determine ship price. Among many data mining algorithms, there are several algorithms that are suitable to build the structural model. In this chapter those potential algorithms will be assessed. The main purpose of chapter 4 is to answer the third research sub-question; *"What are the suitable algorithms for building the structural pricing model?"*. As a starting point, section 4.1 shortly clarifies *Analytic Hierarchy Process (AHP)*, the principal selection method. Here, the selection criteria and scoring system will be explained. Furthermore, section 4.2 outlines the development of the statistic to machine learning.

Furthermore, two potential advanced regressions techniques, Generalized Additive Models (GAM) and MARS, are explained in section 4.3. Moreover the potential tree-based models, Random Forests and Gradient Boosting Machines, are evaluated in section 4.4. Subsequently, two alternatives from the novel machine-learning approach, Artificial Neural Networks (ANN) and Vector Support Machine (VSM), are shortly discussed in section 4.5. After explaining all the candidates, section 4.6 shows the AHP calculation. Finally, section 4.7 outlines the essential content of chapter 4.

4.1. Model Selection Process

To make a (quantitative) selection between the potential candidates, the *Analytic Hierarchy Process (AHP)* is used. AHP is a pairwise comparison method popularized by Saaty in 1970. The main purpose is to make a single selection from a group of fixed alternatives in a rather complex decision marking process. Pairwise comparison is established within the framework of selected criteria. The score for each criteria is given based on technical data or expert opinion. The major advantage of AHP is because pairwise comparison prompts user to make a more rational decision, since they only have to compare 2 items at a time[53]. Another advantage of AHP is that it is one of the few mathematical methods which provide a well-documented decision.

The main reason of using AHP is to eliminate author's subjectivity in choosing the algorithms. However, one has to realize that normally the scoring should be done by a team of experts; whereas, in this case, the whole process is done independently by author's judgement. This might be also considered as vulnerable point of using AHP in this instance. There are four major steps to perform AHP[53], namely:

1. **Criteria evaluation - Establishing the Criteria Matrix**
   To do this, one should first establish the selection criteria and then rank their importance. The framework for the scoring system is available in table 4.1. After the pairwise comparison between each criterion is done, a criteria matrix can be set up. The criteria matrix for the pricing method selection is given in the first matrix presented in section 4.6.

2. **Alternatives Evaluation - Establishing the Comparison Matrices**
   For each criterion, a comparison matrix can be developed by evaluating each alternative based on that criterion. This iterative step is repeated for each criteria, thus at the end, the number of comparison matrices is equal to the number of criteria. In this case, to substitute the expert opinion, literature study is carried out to determine the criteria scoring for each alternative.

3. **Consistency Assessment - Evaluating the *Consistency Ratio (CR)* for each matrix**
   After each matrix is established, one should assess the cardinal consistency of those matrices to maintain the scoring objectivity. To be acceptable, each matrix should have a CR values that is less than 0.1. When this term is violated, the scores should be re-evaluated. CR can be calculated using equation 4.1. RI is a Random Index (a fix value for 6X6 matrix is ***1.24***[53]) and CI is Consistency Index which can be calculated using the dominant eigenvalue (λ)[53].

$$CR = \frac{CI}{RI} = \frac{\left(\frac{\lambda - n}{n-1}\right)}{RI} \tag{4.1}$$

4. **Establishing the *criteria weights* and *local priorities matrix* - Finding the *principal eigenvector of each matrix***
   To get to the final result, one should establish the local priorities matrix and criteria weights. From criteria matrix, one can find the principal eigenvector from the largest eigenvalue which represent the criteria weights ($v_{criteria}$). Afterwards, one can set up a local priorities matrix by calculating the principal eigenvector of each comparison matrix. The result is summarized in subsection 4.6.2. Lastly, the final result is obtained by multiplying the local priorities matrix with criteria weights (matrix 4.6.2).

AHP model will be built by implementing the procedures which are mentioned above. In the next sub-section 4.1.1, the first step of establishing the criteria is explained.

4.1.1. Selection Criteria

The first step of making AHP model is by defining the selection criteria. In this case, six selection criteria are chosen to assess the suitability of the pricing model, namely:

1. ***Usability***
   Usability indicates the method's competency for regression problem, the ease of use and learnability of the method. Some methods are more suitable than others for regression. Also, more famous methods have more accessible learning resources, so they are considered to more advantageous. In this case, ***R*** is the main program. Thus, all potential candidates have readily available package in R.

2. ***Result Quality & Fitting Tendency***
   This criteria indicates the quality of the prediction. Several algorithms have better predictive ability than others. In addition, Some of the methods are proven in pricing model context. Fitting tendency is included because candidates might have the overfitting or underfitting tendency. Most of machine learning algorithm has overfit/ overtrain tendency. There are extra steps that could be taken to compensate this tendency. Some requires more extra step than other and this is considered as disadvantageous.

3. ***Input Compatibility***
   This criteria is mentioned because there are varieties of input data used for the model, such as numerical data, categorical data and binary variable. Some models are only capable to handle limited data type. Several methods can handle this issue better than other.

4. ***Preprocessing Ease***
   This criteria is included as some algorithms can handle missing data and outliers better than others. In this instance, less preprocessing effort is required. In addition, several algorithms require more tuning than others, while others provide "automated tuning feature".

5. ***Postprocessing Ease***
   The last criteria is important regarding the second "goal" of the project which is presented in subsection 3.3, namely to have a continuous final mathematical formulation for each regressor. Several algorithms produce not-so-interpretable results, although this can be solved with tool like LIME or DALEX.

6. ***Computational Time***
   Some algorithms are computationally more expensive than others. Although in general, the amount of time needed to run the model is not particularly substantial (maximum few hours). The algorithms which work faster are generally considered to be better.

After criteria are established, the next step is to create the criteria matrix. The scoring system is further explained in figure 4.1 to compare one alternative with another. The emphasize of AHP lies on pairwise comparison, both for comparison between criteria and between candidates. Starting from these pairwise comparisons, several pairwise comparison matrices will be set up. The pairwise matrix will be made for all criteria previously mentioned, and from each matrix, the principal eigenvector is calculated. The pairwise matrices and correspondent eigenvectors are presented in section 4.6.

Score	***Description***
1	Equal Comparison
2	Moderate Comparison
3	Strong Comparison
4	Very Strong Comparison
5	Extreme Comparison

Table 4.1: Weighting Factor Description. Source: Kana[53]

4.2. From Statistical Modelling to Machine Learning

Historically speaking, ancient human beings have been analyzing the data pattern as far as 3000 years back. It was recorded that ancient scholars from 1000 BC in China had used statistics to help Chinese emperors and their ministers in their decision making process[55]. In the more recent time, statistical analysis has developed into 2 distinct branches, namely the Bayesian and the classical statistic. Both are considered as the first generation of modern statistic. At the first place, the Bayesian concept is developed by Reverent Thomas Bayes in early 1700 by relying on the ***conditional probability***[54]. Some times later at the late 1800, mathematician Ronald Fisher disagreed with Bayesian view and developed an inference testing system based on his concept of standard deviation.

On contrary to Bayesian view, Fisher's statistic relies on the ***joint probability***. Fisher's point of view has been widely adapted as the classical statistic[54]. Fisher's initial aim was to equip medical investigators with an inference system to observe the effect of different medical treatments on patient. He wanted to develop a framework that works objectively when used by a large number of different investigators. Bayesian probability was based on individual belief and therefore was not applicable here since it would create a large variance. At the end, Fisher invented what is now known as the ***parametric model*** with following assumptions[54]:

1. **Collected data fits a known distribution.**
   Examples of known distributions are normal, Binomial and Poisson distribution. This is assumed so that one can find the important features such as mean and standard deviation easily.

2. **Factor independency.**
   Each predictor (Xs) are assumed to have independent effects on Y. When one predictor is strongly related to another, it causes a problem called ***collinearity***. When between more than 2 predictors, it is known as ***multicollinearity***. Multicollinearity can be compensated with ***interaction term***.

3. **Linear additivity.**
   Other very important assumption besides the independence between X-variables is that Y is the linear

combinations of all X-variables. Thus, the initial model cannot deal with non-linearity.

4. **Constant variance (homoscedasticity).**
   This assumption infers that the variance throughout the range of each variable should be constant. Variation may occurs in insignificant amount. When the variance of variables differ significantly, it is known as ***heteroscedastic***. Combined heteroscedasticity might cause significant prediction error.

5. **Variables must be numerical and continuous.**

Besides this linear model, other major contribution that Fisher has been made the statistic field is the introduction to the term *"Likelihood"* and the concept of *"Confidence Interval"*. The term likelihood refers to the intrinsic probability of event occurrences which is solely based on the current experiment. And this concept is difference than the Bayesian concept of *"probability"* which Bayes referred as the future probability, which is equal to past occurrence divided by the probability of all competing events. Moreover, the Fisherian concept of likelihood is interchangeable with probability in classical statistic under one condition; namely when the sample size becomes very large and the effects of investigators' subjectivity approaches zero. This is statistically known as *the Law of Large Number*[54].

Furthermore, the Confidence Interval is introduced due to two types unavoidable errors. The first one is the *alpha error* or false positives; which refers to the probability of being wrong when one thinks he is right. The second type is the *beta error* or false negatives; which refers to the probability that one thinks that he commits error while he is not. In practice, reducing one type of error comes at an expense of increasing other type of error. Type 1 error rate is related to the statistical significant level which is represented by *p-value*. On contrary, type 2 error is used to compute the "power" or "robustness" of an analytic test[54]. By stating 95% confidence level, someone infers that he is willing to be right only 95% of the time. Further in this research, the concept of p-value and confidence level will be repeatedly used.

While the parametric method has been widely used and proven to be useful in wide array of discipline, one of its major drawback is its potential to force-fitting the real world data into an idealistic mathematical construct. For this reason, semi and non-parametric models are assumed to be more useful due to its fitting flexibility. Another major disadvantage of the classical linear model is its disability to capture the non-linear relationships between variables in dataset[54]. However, non-linear relationships occurs very frequently in reality, including the current secondhand market model. Thus, only the semi and non-parametric approaches are further considered to be used in this project.

To deal with non-linearity issue, the *second generation* of modern statistical analysis rises by introducing the non-linear versions of parametric methods. One prominent example is the ***logistic regression*** where the dependent variable (Y) is assumed to be an exponential natural log function of the independent variables (X-s). This further develops into the ***Generalized Linear Model (GLM)*** which make use of so-called *link functions*; by which GLM can take up a limited degree of non-linearity. The most brand-new statistical techniques normally adopts the underlying principals of GLM. Two of them will be discussed in section 4.3. Finally, the development of *third generation of statistic* is marked by the rise of so-called *"machine learning"*[64].

Within the discipline, there are two distinct pathways, namely the *Artificial Neural Networks (ANN)* and the *Decision Trees*. The main idea behind the ANN is to express the non-linear function directly by assigning weights to the input variables (X-s or the *"cause"*), sum their effects to produce an output value which forms a decision function (Y or the *"effect"*). ANN framework imitates the way how human brain works in by passing multiple neural impulses between neurons through the synapses (gaps) to generate a response[54]. Biologically, the synaptic connections is "trainable" and able "learn" to respond faster. The computer scientist strive to create an ANN that could "learn" and recognize the complex patterns in dataset.

Tree-based methods focused on expressing the rules explicitly and therefore bypass the assumptions of parametric model. This makes tree-based models to be more flexible and suitable for non-linear events. Simple decision tree in combination with such as bootstrapping or boosting method are developed into more

advanced techniques. Few important examples are *Random Forests*, *boosted trees* and *Gradient Boosting Machines (GBM)*. Random Forests and GBM are used a lot for regression problems, thus they are selected as potential algorithms and they will be further discussed in section 4.4.

The newest machine learning techniques are coming from the various tweaks (such as *banging* or *boosting*) in combinations with the basic machine learning techniques (such as ANN or decision tree). To summarize, a prominent British statistician, Brian Ripley describing the relationship between regression statistic and machine learning as following: "*To paraphrase provocatively, 'machine learning is statistics minus any checking of the models and assumptions.[64]*" This statement is to some extent true because the main focus of machine learning is to find pattern, making prediction and *generalization*.

For this reason, it might be interesting to see if machine learning technique could be applicable in this project. Among all the available data mining algorithms, six potential algorithms are selected. These are the most commonly used models for regression problem[54]. A brief explanation of these methods are given in the following three section, with main focus lies on the practicality of applying chosen algorithms in building the pricing model. Generally, these algorithms came from 3 categories of statistical and/or machine learning approaches. The statistical methods are presented in section 4.3. The decision tree based models are explained in section 4.4. The machine-learning approaches are given in section 4.5.

4.3. Alternatives 1: Advanced Regression Techniques

4.3.1. Generalized Additive Models (GAM)

As what have been mentioned earlier, the advanced regression method stems originally from the famous standard linear model which in a matrix notation can be described as:

$$\mu = X\beta = b_0 + b_1 x_1 + b_2 x_2 + ... + b_p x_p$$
$$y \sim N(\mu, \sigma^2 I)$$

This model is valid under specific conditions, such as Y is a normally distributed response with expected value μ and constant variance σ^2. Furthermore, X is a matrix of predictor variables where the first column is a vector of intercepts or *bias*. Lastly, β is the vector of corresponding weights to all intercept and predictors. One can apply the linear model in any situation,. However, when the relationships between variables are not linear it will create a large *error (ϵ)/ residuals/ deviance* which will negative influence the fitting accuracy[61].

The fitting accuracy is measured by the *coefficient of determination* $\left[R^2 = 1 - \left(\frac{\sum_i \epsilon_i}{\sum_i (y_i - \bar{y})^2}\right)\right]$, where i is the index to each point estimate. And error (ϵ) can be represented in 2 ways, namely the *Residual Sum of Squared (RSS) Errors* $\left[\sum_{i=1}^{n} \left(y_i - f(x_i)\right)^2\right]$ or *Sum of Absolute Errors (SAE)* $\left[S = \sum_{i=1}^{n} \left|y_i - f(x_i)\right|\right]$. In short, R^2 measures how well a model explains and predicts future outcomes, and in a perfect fit the value is close to 1. The main focus of classical statistic is to maximize R^2 by minimizing errors; because parametric regression is bound to a certain fitting model[64]. There are various techniques of reducing errors, and the most prominent one is the *Ordinary Least Squares (OLS)* Estimator.

The linear model developed into Generalized Linear Model (GLM) regression which can capture the more general situations by allowing the response variable (Y) to have an exponential-family distributions (such as Bernoulli, Poisson, etc). Thus, the option is not only limited to the Gaussian distribution. In addition to the flexible distribution, the response variable (Y) is also mapped using the link-function (*(g(.))* which can vary linearly with the predicted/ fitted values[61]. Thus, Y can take up value beyond the linear combination of predictors (X). Depends on what is being modelled, the most commonly used link functions are log *g(μ)= log(μ)*, log-odds *g(μ)= log(μ/(1-μ))*, identity (g(μ)=μ), negative inverse (g(μ)=-μ^{-1}) and inverse squared(g(μ)=μ^{-2}).

To summarize, GLM can mathematically be described as following:

$$Y = g(\mu) = \eta = X\beta = b_0 + b1x_1 + b_2x_2 + ... + b_px_p$$
$$E(y) = \mu = g^{-1}(\eta)$$
$$y \sim ExpFamily(\mu, \text{etc.})$$

GLM can handle the non-linearity in the dataset, given that users would manually add model terms (link function) in the beginning. However, users must already understand the specific non-linear nature of the data and this is not always possible. To deal with this issue, Hastie and Tibshirani developed Generalized Additive Models (GAM) by combining the concept GLM with *additive models*[61]. The major difference between GAM and GLM lies in the addition of the additive terms (undefined, univariate smooth functions *f(X)*) which enables GAM to find a non-restricted correlation (both linear & non-linear) between each X and Y. Generally, GAM can mathematically be represented as following:

$$Y = g(\mu) = \eta = X\beta + \mathbf{f(X)} = b_0 + f(x_1) + f(x_2) + ... + f(x_p)$$
$$E(y) = \mu = g^{-1}(\eta)$$
$$y \sim ExpFamily(\mu, etc.)$$

The implementation of GAM in *R* employs the *reduced rank smoothing approach.* By doing this, a considerably less computational power and time is needed to execute the model. At its simplest mathematical concept, the main idea is to replace the smooth function *f(X)* with *basis function (B_j)* with weights β_j. Typical basis function B_j are *B splines* for 2D model or *thin plate splines* for 3D model. Additionally, β_j will be later estimated as part of model fitting. In addition to the previous equation, an associated error (ϵ) is now considered to take over the intercept/ bias (b_0)[61]. Furthermore, the principal mathematical equation can be rewritten in equation 4.2.

$$Y = g(\mu) = \eta = f(x) + \epsilon = \sum_{j=1}^{d} B_j(x)\beta_j + \epsilon \tag{4.2}$$

The last distinction between GLM and GAM is the method used to find their estimates ($\hat{Y}$). In GLM, the most common method used is the classic *Maximum Likelihood Estimation (MLE),* whereas in GAM, the *Penalized Likelihood* is used. MLE computes estimates by *maximizing the likelihood function.* The likelihood function is a property specific to the type of Y-distribution. In addition to MLE, Penalized Likelihood takes into account the model complexity, by subtracting a *penalty term* λ to the *likelihood function.* The penalty term λ depends on the model complexity and the value generally increases with the number of variables. The penalized regression generally deliver a better result in comparison to MLE[64]. Since the focus is to judge its practicality, the detail mathematical discussion about GAM is outside the scope of this project.

Strengths and Weaknesses
Lastly, figure C.2 in appendix C section C.2 shows the head acceleration in a simulated motorcycle accident. One can see that GAM is better than standard non-linear regression technique. While GAM is not the only non-linear fitting method, it is selected in this project due to its proven usefulness when it comes to price modelling[36][5]. GAM is a powerful algorithm with some advantages. In the pricing context, GAM has successfully been used by Koehn[36] & Pruyn[5] (proven usability). Furthermore, GAM is *good in handling large dataset with many variables*[78], also good in handling various data types (good input compatibility).

On the other hand, this algorithm is not without weaknesses. GAM has to be complemented with 'backward elimination', thus it requires much effort and longer time to build the model. From this view, GAM has long computational time. Another disadvantage is *its tendency to overfit* and *limitation in forecastingability* when the values of smoothed variables are outside of the dataset range[78]. From that view, GAM has fair result quality. These advantages and disadvantages will be further translated into quantitative scores.

4.3.2. Multivariate Adaptive Regression Splines (MARS)

MARS is the extension of multiple linear regression which makes no assumptions about the relationship between the response (Y) and the predictors (X-s). This way, it can provide a non-parametric framework which automatically capture the non-linearity and interactions between variables. The MARS algorithm can be categorized in two steps[58], namely:

1. ***Create a series of relevant Basis Functions (BF)***
   Before building the BF, range of predictor values are partitioned in several groups (bin). A separate linear regression model is made for each group. The connection between each group is called "knots" and in general, MARS algorithm automatically estimates the most suitable places for these knots. Each knot has a pair of basis functions which represent the relationship between *environmental variables (X-s)* and *response (Y)*. Mathematically, MARS model can be mathematically described as equation 4.3. Thus The model is a weighted sum of basis functions $B_i(x)$ with c_i acts as constant coefficients. BF can take various forms such as *linear function, polynomial function* or *step function*[57].

$$\hat{f}(x) = \sum_{i=1}^{k} c_i B_i(x)) \tag{4.3}$$

2. ***Generate a least-squares model for each BF***
   After sufficient BF is created, the next step is to minimize the error of each BF. Each basis function is treated as independent variables. When the model is very large/ complicated, model pruning (reduction/ simplification) can be done to reduce the overfitting.

Strengths and Weaknesses

Using MARS algorithms has several advantages such as *high accuracy* (decent result quality), *relatively short computational time* (short computational time). MARS is also *suitable for various size of dataset*, thus from this view MARS has (good input compatibility). However, this algorithm *does not work as good as more advanced algorithms* such as Random Forests & GBM[64]. MARS also *incapable of handling missing values*, thus a more thorough data preprocessing step is needed (low score in preprocessing). The last major drawback of using MARS is its *high tendency to overfit* (fair result quality) and in this case, pruning is needed[58]. The selection of the number of knots and type of BF will determine the fitting quality, and this is illustrated in figure C.3 in appendix C section C.3.

4.4. Alternatives 2: Tree-based Models

Before understanding how the Random Forests and Gradient Boosting Machine work, one has to understand the general concept of ***Regression Tree***. The starting point of ***Regression Tree*** is the (multiple) Linear regression which is extensively explained in chapter 4 section 4.3. It is a global model, where there is a single predictive formula holding over the entire data-space. When the data has lots of (nonlinear) variables, assembling a single global model can be very difficult. An alternative approach to nonlinear regression is to sub-divide, or *partition*, the space into smaller regions, where the interactions are more manageable. The partition process is done repetitively (*"recursive partitioning"*), until one can fit simple models to the partitioned spaces[109]. The global model consists of the *recursive partition* and a simple model for each cell of the partition[54].

The regression tree consists of three main parts namely the ***Root Node***, ***Branch*** and ***Leaves or Terminal Nodes***. ***Root Node*** is the first partition which is the starting point of the regression trees. Afterwards, tree will grow its ***Branches*** until each branch stops at the ***Leaves or Terminal Nodes***. User can determine how deep a tree will grow by specifying the number of *minimum observations* before a branch split[109]. Figure 4.1 illustrates how a general regression problem is translated to decision tree by *partitioning the data-space*. Two numerical variables (Wheelbase and Horsepower) are used to determine the cars price. When more variables or/and observations increases, the complexity of a tree increases. Since Tree is susceptible to overfitting, the advanced algorithms such as **Random Forests** and **Gradient Boosting Machine** are developed[54].

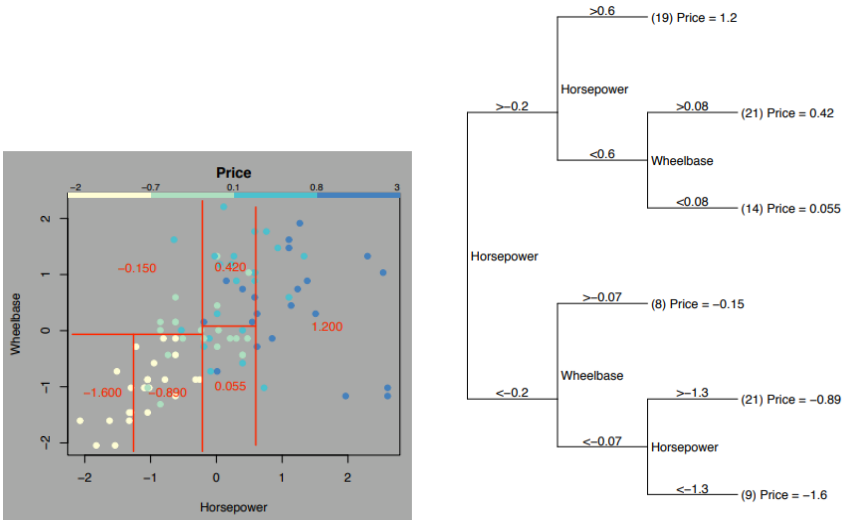


Figure 4.1: Translating Regression Problem into Regression Tree. Source: CMU Statistics[109]

4.4.1. Random Forests

Random Forests is the extension of the regression tree to solve its overfitting tendency. Regression tree tends to have a low *bias* with very high *variance*, as trees grow very deep and create an irregular pattern[60]. Bias is the difference between the average prediction (estimate) of our model and the true population value which model tries to predict. High bias means there are a big discrepancy between the predicted value with the true value, or shortly *inaccurate prediction/ underfit model*. On the other hand, variance measures the "spread" of the data. High variance model means that more attention is given to train/ fit the data and less is given to "generalize" the result. *A perfect model will have low bias and low variance*. However this is not possible due to the trade-off between minimizing bias and variance[60] which is further explained in appendix C.1.

How is a (Random) Forest created ?

It combines the principal of regression tree and *bootstrapping*. Statistical bootstrapping refers to the process of repeatedly taking the random samples from the dataset until the normal distribution is achieved. In this context, 2/3 *"in the bag"* of the dataset is used to create random subsets, while 1/3 *"out of the bag"* data will be used to test the model later. Random Samples are repeatedly taken from *"in the bag"* dataset. For each random sample, a regression tree is grown. At the end, one will end up with hundreds to thousands of (random) regression trees. The collection of these (random) trees are called ***Random Forests***[59]. At the end, the estimate is made by taking the average of all the trees within the forest. To conclude, the process can be illustrated using figure 4.2 and the estimate can be described by equation 4.4.

$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B} f_b(x) \tag{4.4}$$

Strengths and Weaknesses

Shortly, this approach turns a tree model with low predictive power into an accurate prediction function. Random Forests is a more advanced version of decision tree, and *it is one of the most accurate learning algorithms available* (excellent result quality). Another advantages are *its ability to handle many predictors* (good input compatibility), *its ability to maintain high accuracy even when a large proportion of data is missing* and it automatically provides *estimation of the importance level of different predictors* (great usability).

In addition to them, little tuning is needed when working Random Forests (great preprocessing ease)[60]. Despite its strength, this algorithm is not without shortcomings, few examples of Random Forests are *its*

tendency to overfit "noisy" data and *the result is less interpretable* (fair postprocessing ease). However, the less interpretable result can be improved using tools such as LIME or DALEX.

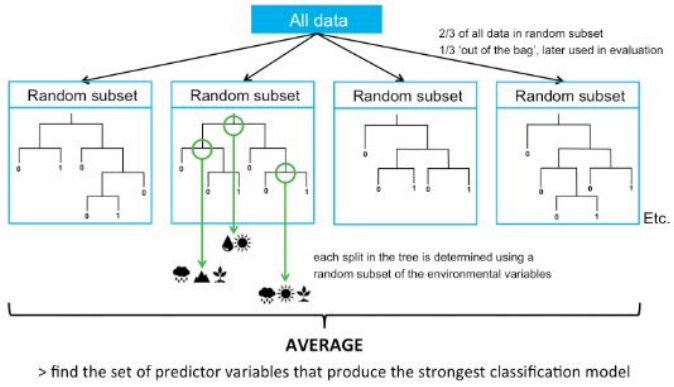


Figure 4.2: Random Forests Process Illustration. Source: BCCVL[59].

4.4.2. Gradient Boosting Machines

How does GBM work ?

Gradient Boosting Machines (GBM) is also a modified tree-based model. It is also known as *Gradient Boosting* or *Generalized Boosted Model*. GBM combines two machine learning principals, namely the regression tree and boosting methods. On contrary to Random Forests, GBM builds an ensemble of shallow and *weak successive trees*. The tree is "weak" because the numbers of maximum branches per tree is determined, so they cannot grow very deep. Because the trees are shallow, the performance of each individual tree is generally bad. However performance correction will be made with each successive trees. Also, each tree in Random Forest is completely independent from other trees since they are selected randomly; whereas each tree in GBM is connected to previous tree as each tree learns from their predecessors to reduce their errors[62].

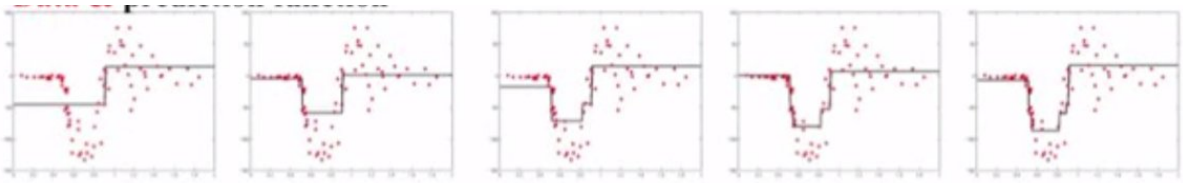


Figure 4.3: Improvement in Fitting Performance with each (successive) Tree. Source: Boehmke[63].

The selection of dataset to create the successive trees is done using the *boosting method*. Weights are assigned to new trees in such a way that trees with poor performance have higher chance to be "re-selected". When being "re-selected", trees performance will be (gradually) improved. When combined, these trees produce a powerful “committee” which make excellent prediction. Mathematically, optimization is done using the *loss function*. This boosting process can be illustrated using figure 4.3. The algorithm of GBM is also outlined in appendix C section C.4. Shortly, GBM model can be generalized to the equation 4.5[62].

$$f(x) = \sum_{b=1}^{B} f^b(x)) \tag{4.5}$$

From optimization point of view, *loss function* is an *objective function* which machine learning algorithm seeks to minimize since it represents errors (ϵ) and penalty terms (λ). There are several types of loss function which commonly used in machine learning[63], however further technical discussion about it is outside the scope of this chapter. Several common loss functions are mentioned in appendix C in section C.4.

Strengths and Weaknesses

To sum up, GBM is a powerful algorithm that is *suitable for large datasets*, or a model with *a large number of predictors* (good usability). The advantages of GBM are *its exceptional prediction accuracy* (great result quality), *high flexibility, in terms of response distribution, loss functions and tuning options*[62]. Also, *its robust ability to handle outliers and missing values*[62] (good input compatibility).

Also, *GBM works well with any type of data and no data preprocessing is required* (good preprocessing ease). Besides its advantages, this algorithm *requires a substantial computational power* since a massive number of trees is needed[63] (slow computational time). In addition, high flexibility comes at the cost of many parameters interaction, which *requires a substantial grid search during the tuning* (fair preprocessing ease). Lastly, this algorithm is *less interpretable* although this issues can be solved with extra postprocessing work[62] (fair postprocessing ease).

4.5. Alternatives 3: Machine Learning Approaches

4.5.1. Artificial Neural Networks (ANN)

Artificial neural networks (ANN) is a group of models that forms connections like brain neural networks. It is a "trainable" algorithm which obtain knowledge by extracting patterns from data[66]. Conceptually, ANN consist of a large number of nodes and connections which is organized in *"layers"*. These nodes and connections represent many interconnected computational units, called neurons. Every single neuron has a (limited) intrinsic approximation capability. However, when these are combined, they have a cumulative effect resulting in a remarkable learning performance[67].

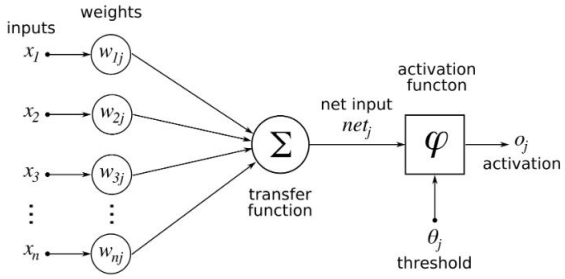


Figure 4.4: Mechanism of Single Neuron from the Input Layer. Source: Boehmke[65]

Generally, there are three categories of layers, namely *input*, *hidden* and *output*. The input layer receives the input data and the number of nodes signifies the number of input variables. The step of processing the input data can technically be represented using diagram in figure 4.4. Afterwards, the leaning process takes place inside the hidden layers. Lastly the result is presented in the output layers, the result can be a single or multiple functions. When appropriately tuned, ANNs are able to "approximate" any functions which map input of a number of explanatory variables to a number of outputs[67]. This can be mathematically represented using equation 4.6. $\hat{y}$ represents the output values of the output layer, X represents the input values, $\vec{w}$ represents the random initial weight and θ is the selected *threshold values/ function.*

$$\vec{\hat{y}} = f(X, \vec{w}, \theta) \tag{4.6}$$

Referring to the figure 4.4, for a model with n total number of regressors, index i indicates the number of input variables while j refers to the *j-th* neuron. Between the input variable x_i and *j-th* neuron node of the corresponding middle layer, there is a certain weight (w_{ij}) assigned to it. The main challenge is to determine the number of neuron units per (hidden) layer. Most ANN implementations do not permit the specification of the number of hidden layers, but it requires the specification of neuron units.

More neuron units increases the neural net capacity to recognize non-linear patterns, but also exponentially increases the computational time and the chance of higher variance[54]. Overtrain and overfit tendency increases with the variance. Overtrain model may give an excellent prediction to one dataset but unable to predict others (not-generalized)[54]. Afterwards, these weighted inputs will be summed using the *transfer function* resulting in the *net input*. For each *j-th* node, there is an associated *bias* (b_j) term. In short, it can mathematically written as following:

$$net_j = b_j + \sum_{i=1}^{n} w_i x_i$$

Subsequently, the activation process takes place and it is represented by some mathematical function $\phi_j(.)$. The activation functions enables ANN to learn non-linear relationship of the variables. Depends on the problem, the selection of appropriate $\phi_j(.)$ will majorly determine the model ability to make good prediction[65]. For example, linear $\phi_j(.)$ is suitable for linear estimation problems and the logistic $\phi_j(.)$ works best for classification problems. The graphic overview of the functions can be found in appendix C.5. For each activation function, there is an associated *threshold value* θ_j.

Threshold value is a predetermined value/ function in which the activation function generates an output signal (o_j), when threshold is reached. At the end, o_j will be forwarded to *k-th* neuron with an assigned random weight $_{jk}$ at *time = t*, together they form what is known as *propagation function*. They can be formulated using equation 4.7 and 4.8[67].

$$o_j = \phi(b_j + \sum_{i=1}^{n} w_i x_i) \tag{4.7}$$

$$p_j(t) = \sum_{j} o_j(t) w_{jk} \tag{4.8}$$

Regarding the direction of data propagation, there are few types of ANN such as *recurrent* and *feed-forward* ANN. However, the most common and relevant type for regression problem is the *feed-forward ANN*, where the information moves consistently toward one direction (from input to the output layer)[54]. In feed-forward case, the information never goes backwards. The learning process of ANN follows a popular algorithm called *back-propagation*. which is also biologically inspired by the mechanism of the neurons of the mammalian[68]. At its simplest, back-propagation mechanism can be defined as an iterative and recursive method to provide the weight updates (see figure 4.4) to train ANN model until expected performance level is achieved. This process can be visualized in figure C.5 in appendix C section C.5[65].

Strengths and Weaknesses

To conclude, ANN is an outstanding non-linear machine learning algorithm with can be used for regression problem. Examples of advantages are *its ability to handle very large size of data efficiently* (good result quality), thus giving *good prediction with relatively low computational power* (relatively short computational time). Also, it has (limited) *tolerance to correlated input variables*[64], thus ANN is able to incorporate the predictive power of *different combinations of inputs types* (good input compatibility).

In addition to them, ANN is one of the most acclaimed data mining algorithms and it has undergone a substantial number of advancements in the last few centuries[54]. This makes *learning source for ANN becoming very accessible* in comparison to others. On the other hand, the main shortcoming of ANN is its *low ability to handle outliers and missing values* (weak preprocessing ease), which might make ANN susceptible to irrelevant features[66] (limited usability).

4.5.2. Support Vector Machine (SVM)

Support Vector Machine is originally developed for classification problem, namely by using *decision planes/ hyperplanes* to define *decision boundaries*. The hyperplanes are *pre-determined requirements* given by the users at the beginning of modelling. They separates objects based on *classes* and SVM can handle any number of classes/ variables of any number of observations[54]. SVM uses the *kernel* function to *maps* objects to separate them between the *hyperplanes*. Kernel are versatile mathematical functions, thus able to map any

transformations. As the result, SVM can handle any classification problems. Figure 4.5 illustrates how the SVM classification works in not-complex environment.

To optimize the separation (especially in complex cases where observations are non-linearly separable), SVM uses the *maximal margin algorithm*. Thus, this algorithm strives to create hyperplanes with the maximum distance from the data points. When handling the regression problem, SVM uses the same principals of *kernel transformation* and *maximum separation*. The process of SVM regression is illustrated in figure 4.5. The Kernel function transform the non-linear planes (or lines) into linear spaces. Maximum Separation principal plays the rule as users define the maximum ϵ (a margin of tolerance) which is depicted as dotted lines in second figure of 4.5. That would say that it is "tolerable" to have errors within the margin. However when errors are outside the tolerable ares (ξ), the algorithm will seek to minimize them[69].

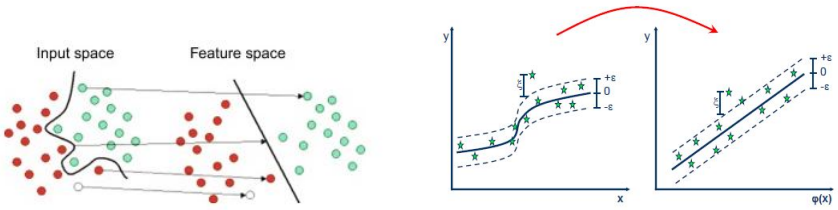


Figure 4.5: SVM for Classification (left) and Regression (right) Problem. Source: Nisbet, et al[54]

The Kernel functions are used to linearize higher dimensional data, two most commonly used kernel for non-linear SVM regression are the *polynomial kernel* and *Radial Basis Function (RBF) kernel*. Respectively, they can be mathematically written as equation 4.9 and 4.10. At the end, the general SVM equation can be written as 4.11, where b refers to the bias term. The selection of kernel function will determine the fitting result, figure C.6 in appendix C.6 illustrates the different results by different kernel function[56].

$$\langle \varphi(x_i), \varphi(x_j) \rangle = K(x_i, x_j) = \left(x_i^T x_j + 1\right)^d \tag{4.9}$$

$$\langle \varphi(x_i), \varphi(x_j) \rangle = K(x_i, x_j) = \exp\left(-\frac{1}{2\sigma^2} \left\| x_i - x_j \right\|^2\right) \tag{4.10}$$

$$y = \sum_{i=1}^{n} \left(\alpha_i - \alpha_i^*\right) \cdot \langle \varphi(x), \varphi(y) \rangle + b \tag{4.11}$$

Strengths and Weaknesses
To summarize, SVM is a suitable machine learning method for regression problem with several advantages. These are the *good ability prediction in different situations* (decent result quality) & thanks to Kernel function, it can *handle complex problem quickly* (relatively short computational time). The main disadvantage of SVM regression when compared to others is its *stronger black box nature* which makes it even harder to grasp the underlying processes[56] (fair usability). Another weakness is its *low competency in handling of mixed data types* (limited input compatibility). In addition to it, when being compared to famous algorithm like ANN, there are less learning material available for SVM regression[69]. In the next section, the advantages and disadvantages of SVM and 5 other methods are quantified to make a logical comparison.

4.6. AHP Calculation

Pairwise Matrix for Criteria by Criteria

	Usability	Result & Fitting	Input	Preprocessing	Postprocessing	Time	$v_{criteria}$
Usability	1	1	1	2	2	3	0.2287
Result & Fitting	1	1	1/2	2	2	3	0.2084
Input Compatibility	1	2	1	2	2	2	0.2487
Preprocessing Ease	1/2	1/2	1/2	1	1	2	0.1213
Postprocessing Ease	1/2	1/2	1/2	1	1	1	0.1074
Processing Time	1/3	1/3	1/2	1/2	1	1	0.0855

Pairwise Matrix for Alternatives by Usability

	GAM	MARS	Random Forests	GBM	ANN	SVM	v_1
GAM	1	2	2	2	3	3	0.2983
MARS	1/2	1	1	1/2	1	1	0.1175
Random Forests	1/2	1	1	1	2	3	0.1804
GBM	1/2	2	1	1	3	2	0.2000
ANN	1/3	1	1/2	1/3	1	2	0.1058
SVM	1/3	1	1/3	1/2	1/2	2	0.0980

Pairwise Matrix for Alternatives by Result Quality & Fitting Tendency

	GAM	MARS	Random Forests	GBM	ANN	SVM	v_2
GAM	1	2	2	1	3	2	0.2644
MARS	1/2	1	1	1/2	1	1	0.1229
Random Forests	1/2	1	1	2	1	2	0.1884
GBM	1	2	1/2	1	2	2	0.2006
ANN	1/3	1	1	1/2	1	1	0.1159
SVM	1/2	1	1/2	1/2	1	1	0.1078

Pairwise Matrix for Alternatives by Input Compatibility

	GAM	MARS	Random Forests	GBM	ANN	SVM	v_3
GAM	1	1/2	1	3	1	3	0.2186
MARS	2	1	1	1/2	1	2	0.1851
Random Forests	1	1	1	2	1	2	0.1924
GBM	1/3	2	1/2	1	2	2	0.1778
ANN	1	1	1	1/2	1	1	0.1385
SVM	1/3	1/2	1/2	1/2	1	1	0.0877

Pairwise Matrix for Alternatives by Preprocessing Ease

	GAM	MARS	Random Forests	GBM	ANN	SVM	v_4
GAM	1	1/2	1	1	2	2	0.1819
MARS	2	1	1/2	2	1	2	0.2073
Random Forests	1	2	1	2	2	2	0.2518
GBM	1	1/2	1/2	1	1	1	0.1239
ANN	1/2	1	1/2	1	1	1	0.1259
SVM	1/2	1/2	1/2	1	1	1	0.1092

Pairwise Matrix for Alternatives by Postprocessing Ease

	GAM	MARS	Random Forests	GBM	ANN	SVM	v_5
GAM	1	1/2	1	3	2	3	0.2306
MARS	2	1	1	1	1	2	0.2003
Random Forests	1	1	1	2	2	3	0.2187
GBM	1/3	1	1/2	1	3	4	0.1775
ANN	1/2	1	1/2	1/3	1	1	0.1010
SVM	1/3	1/2	1/3	1/4	1	1	0.0718

Pairwise Matrix for Alternatives by Processing Time

	GAM	MARS	Random Forests	GBM	ANN	SVM	v_6
GAM	1	1/2	1	3	2	2	0.2002
MARS	2	1	2	3	3	3	0.3223
Random Forests	1	1/2	1	2	1	3	0.1820
GBM	1/3	1/3	1/2	1	1	1	0.0916
ANN	1/2	1/3	1	1	1	1	0.1119
SVM	1/2	1/3	1/3	1	1	1	0.0921

4.6.1. Matrices Consistency Assessment

In this subsection, the calculation of consistency index for each pairwise matrix is presented. From table 4.2, one can observe that all matrices are cardinally consistent, thus one can proceed to the result calculation.

Pairwise Matrix	***Consistency Index (C.I.)***	***Cardinally Consistent C.I. <0.1***
Criteria	0.0195	Yes
Usability	0.0626	Yes
Result Quality	0.0393	Yes
Input Compatibility	0.0933	Yes
Pre-processing Ease	0.0355	Yes
Post-processing Ease	0.0811	Yes
Processing Time	0.0191	Yes

Table 4.2: Matrix Consistency Assessment.

4.6.2. Selection Result

From the previous steps, the local priorities matrix can be set up. Afterwards, this local priorities matrix can be multiplied with $v_{criteria}$. The result is calculated in the following matrix.

	Usability	Fitting	Input	Preprocessing	Postprocessing	Time		Weight		Result	
GAM	0.2983	0.2644	0.2186	0.1819	0.2306	0.2002		0.2287		***24.33%***	**GAM**
MARS	0.1175	0.1229	0.1851	0.2073	0.2003	0.3223		0.2084		16.95%	MARS
RF	0.1804	0.1884	0.1924	0.2518	0.2187	0.1820		0.2487		***19.74%***	**RF**
GBM	0.2000	0.2006	0.1778	0.1239	0.1775	0.0916	*	0.1213	=	***17.52%***	**GBM**
ANN	0.1058	0.1159	0.1385	0.1259	0.1010	0.1119		0.1074		11.79%	ANN
SVM	0.0980	0.1078	0.0877	0.1092	0.0718	0.0921		0.0855		9.57%	SVM

From the calculation result, it is concluded that GAM is the most suitable pricing method for this project. The second and third best candidates are Random Forest and GBM respectively. The three best algorithms will be used to build the pricing model, and at the end, the results will be compared.

4.7. Chapter 4 Conclusion

The discussion about the algorithms selection is elaborated in this chapter. Six potential candidates are evaluated. Two statistical approaches are considered, namely the Generalized Additive Models (GAM) and Multivariate Adaptive Regression Splines (MARS). Moreover, four machine learning algorithms are evaluated, namely the Random Forests, Gradient Boosting Machines (GBM), Artificial Neural Networks (ANN) & Support Vector Machine (SVM). Their important characteristics are clarified in sections 4.3, 4.4 and 4.5 respectively. Furthermore, these candidates are ranked using the Analytic Hierarchy Process (AHP) framework. GAM, Random Forest and GBM are voted as the three best models to be used in this project. By this, the third research question, *"What are the suitable algorithms for building the structural pricing model?"* is answered.

5
Features Selection

Bezint eer ge begint.

Dutch Proverb

In this chapter, the key influences that affect the price of a secondhand vessel are evaluated. This evaluation clarifies the fourth research sub-question, *"What are the relevant influencing factors to be considered in the secondhand vessel pricing model according to the literature ?"*. All these factors are already included in current model, and they will be discussed individually in section 5.1. In addition to them, extra parameters based on literature study will be considered. Optional factors such as ***high*** $\frac{Volume}{DWT}$ ***ratio***, ***ice class*** and ***crane capacity*** are considered by analyzing the initial system. They are discussed in 5.2. Lastly, factors suggested by the literature review are reviewed in 5.3. Finally, the chapter summary is presented in section 5.4.

5.1. Current Variables

Five factors which are already used in current model will be re-used in the improved model as their importance are repeatedly emphasized in multiple literature[27][23][6][35][36]. Their importance is confirmed by the most recent study done by Merika et al. where they explore the price heterogeneity in sale and purchase market of bulk carriers. They concluded that vessel age, 3 month LIBOR and annual charter rate have the highest influences on sales price[42].

Age

Age indicates the expectation of how much longer the ships will be productive and therefore profitable. Stopford argued that vessels age plays an the prominent role in secondhand vessel price determination[4]. He explained that older vessel have more disadvantages mainly due to performance declining and higher maintenance costs[4]. As a vessel becomes older, the secondhand price will eventually falls below the scrapping value. The average bulk carriers life-time before being scrapped is 26 years, and 35 years when protected trade is applied. As what was mentioned before, this variable is also used by Adland & Koekebakker as determinant in their structural model[35]. In other case, Koehn took the same approach to consider vessel age in his model[36]. Adopting the experts opinions, age is used in the new model.

Vessel Income

Tsolakis explained that secondhand price is the function of vessel revenue[27][38]. In a way, vessel income reflects the future expectation. Income incorporates the time-series effect (like economic cycles) in this structural model. However, one need to bear in mind that the registered selling date (when they handed in the vessel) is not exact time when both parties agree to trade their vessel. The agreement took place earlier, it could be weeks up to months in advance.

Unfortunately, this information is not disclosed, thus the best indicator that one can use is the selling date. Stopford used ***Freight rates*** as income representation[4]. However, Tsolakis claimed that ***time charter***

(TC) rate is a more subjective indicator since it is independent of the route[27]. In other research, Geomelos & Xideas used ***spot rate*** as the income indicator. However, Tsolaskis argued that TC-rate is better since it has a greater stability in comparison to the spot rate[27].

The last alternative to represent income is ***Worldscale flatrate (WS100)***, which is used by Beenstock & Veenstra in their research[22][32]. However, Tsolakis argued that WS100 is not the best indicator because it also takes into account bunker prices, port dues and currency exchange rate[38]. In addition, this index is especially made for tankers and not easily accessible. Thus, the usage is irrelevant in this project. Adopting Tsolakis view, Pruyn has used 1 year TC rate. In this research, the applicability of *6 months*, *3 years* and *5 years* TC rate is investigated. At the end, it can be concluded that these indices follow the exact same trend and do not differ significantly (maximum standard deviation is 9.1%). Thus it is decided to used the (time) average TC-value per vessel type, more elaborate discussion can be found in appendix D section D.1.

In addition to the average value, *the normalized TC-rate rate* is also considered. Normalized TC-rate is a dimensionless index, obtained by dividing the current TC rate with the TC-rate in the beginning of the entry (in this case July 2016). This is because, TC-rate per vessel type might vary reasonably, for example it is common for TC-rate for capesize to be twice as much as of handysize vessels. In this case, using the normalized TC-rate might be advantageous since the value will be relatively taken to limit the range of TC. In this way, the normalized TC-rate might reflect the fluctuation of shipping market more subjectively. The formulation for TC-average and TC-normalized in given in equation 5.1.

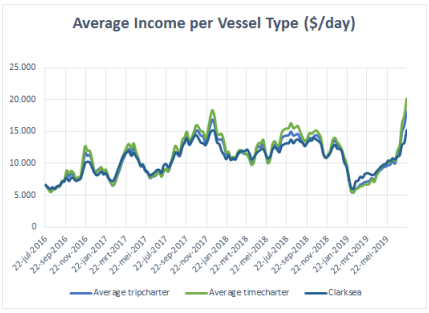


Figure 5.1: Comparing Average Income Representations. Data is compiled and processed from Clarksons.

for each vessel type

$$TC_{average,i} = \frac{TC_{6mt,i}+TC_{1yr,i}+TC_{3yr,i}+TC_{5yr,i}}{4}$$
$$TC_{normalized,i} = \frac{TC_i}{TC_{begin}} \tag{5.1}$$

where i indicates the i-th week

Beside TC-rates, ***Clarksea index*** and ***average trip charter*** are researched. Clarksea index is a weighted average of bulk carriers' earnings where the weighting is based on the number of vessels in each fleet type. By nature, Clarksea index is a generic variable since it represents the whole bulk carriers group. Also, the trip charter rate has never been researched since the rate varies based on the travelling route. In this case, an average value is used. The comparison between these indices are presented in figure 5.1.

By observing figure 5.1, one can observe an almost perfect collinearity in the three income indicators. Elaborate discussion about collinearity can be found in chapter 4 section 4.2. Since collinearity is undesirable, it is decided to only use one income representation. The average TC-rate is chosen due to its higher stability compared to trip charter and it gives information per vessel type while Clarksea gives only a general information.

Size

An observable fact is a larger vessel costs more than a smaller vessel of same type. This makes the vessel size becomes an important price determinant both for newbuilding and secondhand. Furthermore, the applicability of economic of scale in sea transport infers that bigger ships have more profit potential both for shipowners and cargo owners[105]. However, bigger ship owners are also exposed to a bigger risk due to (much) higher initial investment and chance of sailing with less loading than what is profitable[4].

Moreover, there are several possible parameters to represent size such as Length-Breadth-Draft (LBT), Gross Tonnage (GT) and DWT. Among them, Koehn argued that DWT is a more proper representation since it is directly related to the potential future income and the information is always avaialble; whereas LBT or GT are sometimes missing from Clarksons[36]. Thus, DWT is used as size representation in current model.

Cost of Finance

In most cases, buying a new or secondhand vessel will require extra financial help from financial institution. LIBOR index indicates the average global interest which also indicates the ease of vessels financing. It is the second time-series variable which is incorporated into model, thus the same concern about the possible time-gap between the "deal-date" and "sold-date" applied in this case. LIBOR is a market indicator; higher LIBOR rate means a thriving economic and lower rate means low market. According to market theory, sales price soars during the high market due to high demand and vice versa[4]. Thus, positive effect will be expected. In many econometric models, 3 months LIBOR are used to representation because this data is rather accessible. Thus, 3 months LIBOR will be used in the new model.

Orderbook Percentage

In this context, orderbook percentage is used to incorporate the newbuilding price index in the model, since newbuilding price is considered as essential in secondhand price formation by multiple literature[42][23][27][37][30]. While newbuilding price index is provided by Clarksons, they do not disclosed the calculation method. Thus, the estimate might be unreliable[5]. Thus, orderbook percentage is a preferable indicator. In this case, the orderbook size is presented relatively as percentage of total fleet. This is a better indicator because it encompasses the fluctuation of total fleet number orderbook[38]. Just like TC-rate and LIBOR, orderbook is a time dependent variable; thus similar precaution is applied.

Higher orderbook means that the shipyard has lots of order, thus the waiting time to get the newbuilding is high. Thus, buyer who wants to acquire vessel quickly would go to secondhand market as an alternative. However, higher orderbook also means that many investors have already ordered their ships. Since there are limited number of players in shipping market, there would only be a handful number of people who has not ordered their ships. Thus, one might expect a positive or/and negative effect of orderbook on sales price.

5.2. Optional Features

Optional features are suggested by analyzing the current model. These factors are analyzed to meet the first improvement goal which is explained in section 3.3 of chapter 3, namely to *increase the model sensitivity*.

$\frac{Volume}{DWT}$ Ratio

$\frac{Volume}{DWT}$ **Ratio** is used in MBG as a " quick" indicator of which type of cargo a bulk-carrier can carry. Higher $\frac{Volume}{DWT}$ in MBG means indicates ship can carry heavy cargo such steel-coils. Unlike normal cargo such as grain, carrying steel-coil is a tricky business because steel-coils will cause not-uniform load distribution on the top platting of double-bottom[101], such as illustrated in figure 5.2. Eventually, improper placements might cause deformation on top platting stiffeners. In reality, there is software available for cargo planners to do calculation and planning to ensure the most optimum weight distribution[102]. However, in MBG, players do not have this privilege. Thus, this ratio is given as a quick rule of thumb.

There are five level of $\frac{Volume}{DWT}$ **Ratio** for the vessels in the MBG. Lower $\frac{Volume}{DWT}$ indicates that the vessels are suitable for grain, coal or cargo with lower density. Higher $\frac{Volume}{DWT}$ indicates that vessels are suitable for higher density material such as ores or even steels coil. In reality, *"ore-carrier"* is specially designed for lower volume cargo. They have usually combined function also as oil-carrier since they have larger side-tanks to maintain the location of the center of gravity[104]. In MBG, a higher ratio also means cargo flexibility and it is more common for handysize and panamax vessels. Thus, exploring the co-relation between vessels price and $\frac{Volume}{DWT}$ might lead to a useful finding.

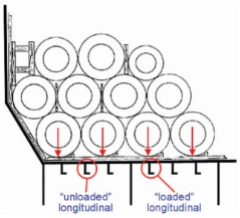


Figure 5.2: Weight Distribution Illustration for Steel Coils Cargo. Source: TheNavalArch Team[102].

Optional Ice-Class

Sailing in the ice-infested water will require ships to overcome a rougher environmental condition. Ice will cause additional challenges for ship such as reduced ship stability because reduction in metacentric height due to the additional weight of the ice which[51]. There is also extra risks associated with sailing in icing condition such as limited operability of the safety equipment such as lifeboats and fire fighting systems, thus extra training will be needed for the crews on board[51]. In addition, IMO also prescribes extra regulations for ship sailing in Polar area such as heavy oil banned[99]. At the end, operational expenses of those ships will also be higher due to higher fuel consumption for the propeller and extra heating system on board (and heating for certain types of cargo)[51][105].

There are extra requirements for ships sailing in ice-infested water namely the ice-class hull structure and stronger propulsion power[50]. Thus higher investment cost is needed[52]. Despite of the extra expenses, there is also potentially extra profit when ships are sailing from China to Europe through the Northern Sea route due to the much shorter distance compared to the Suez canal route[105]. It was reported that at least 6 ice class oil tankers were sailed through the eastern part of the Northern Sea Route in early October 2019[100]. Considering these, it might be useful to study the effect of ice class in the secondhand price. From 1227 vessels, 46 vessels (3.75%) are registered as ice-class; in which 42 of them are handysize vessels. Thus, ice-class will only be considered for handysize vessels in the further work.

Crane Capacity

Not every ports in the world are equipped with cranes, thus having a crane will enable ship to sail to crane-less ports. Such ports are usually smaller in size, thus handysize and handymax vessels will deliver cargo to those ports. Examples of crane-less ports are East London port in south Africa, Lamu port in Kenya and Bagamoyo port in Tanzania[97]. In addition, having crane(s) will accelerate the loading and unloading process. This will beneficial for shipowners because their bulk-carriers will spend less time in the port, thus they will pay less port-tariff[107]. Thus "geared" vessels are beneficial compared to " gearless" vessels as they will have broader operational area and they will potentially pay less operational expenses (due to lower port-cost).

However, crane is only common for smaller vessels because larger vessels tend sail to ports with better infrastructure[97]. Although not significant, the crane installation itself will cost an additional building cost. In the future model, crane capacity variable will be analyzed to see if there is significant relationship between them. From the database with 1227 vessels, 759 vessels (61.86%) are registered as having crane on-board. Among them, 89% handysize vessels and 100% handymax vessels are equipped with crane. In the new model, crane capacity is expressed in ton.

5.3. Additional Factors

Additional features are suggested by the literature review. This is to meet the second improvement goal as explained in section 3.3 of chapter 3, namely to *integrate additional influencing factors.*

Builder Reputation

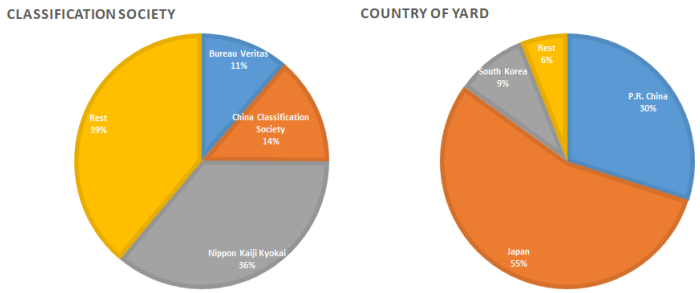


Figure 5.3: Percentage of Classification Society and Country of Yard. Data is compiled and processed from Clarksons[7].

Builder reputation can be viewed from two points, namely the ***country of yard*** and the ***classification society***. The country of yard influences the newbuilding cost since in some countries the labor costs are lower. Furthermore, the classification society, to a limited extent, might represent the vessel's quality[42][36]. For these reasons, they are considered in the new model. From the vessels' database, there are 10 countries of yards and 23 Classification Societies. They are summarized in table D.1 and D.2 in appendix D. From these tables, one can observe that some countries and classification society only make up a small percentage of population, thus their effect might be insignificant. Thus, only criterion which make up more than *5% for country of yard* and *10% for classification society* are considered individually. These are outlined in figure 5.3.

Sustainability Indicators

Sustainability has been becoming an increasingly important issue. In its simplest essence, it suggests that the resources exploitation to meet human's need is permissible, as long as one makes the effort to preserve the environment's welfare. In the shipbuilding industry, maintaining the environmental and social responsibility while trying to meet the financial goal can be a challenging task. Many efforts have been made toward sustainable shipbuilding, starting by the research proposed by academic institutions. Examples of such research are the reduction of sea emission[71] and the responsible ship recycling[72]. In this project, an effort is made to include the sustainability factor. Based on Clarksons database availability, ***anti-fouling coating***, ***scrubber*** and the ***type of fuel*** are considered.

Parameter	With	Without
Anti-fouling coating	9.78%	90.22%
Scrubber Installation	4.89%	95.11%
Using MDO-fuel	1.63%	98.37%

Table 5.1: Sustainability Indicators Summary.

Anti-fouling coating prevents corrosion and development of marine organism on the underwater hull. This way, good vessel's performance can be preserved because fouled hull creates more resistance and therefore it requires more fuel to sail[73]. Thus, the ships with anti-fouling can be considered, to some extent as more sustainable. Furthermore, with increasingly stricter regulation, shipowners are compelled to switch to more sustainable energy sources. International Maritime Organization (IMO) defines the Emission Control Area's (ECA's) where limitation on SO_x & NO_x applies[46].

To comply with them, a ship can either use the expensive Marine Gas Oil (MGO), install the scrubber or use alternatives such as LNG. Up to this point, the cheapest and most accessible solution is by installing scrubber on board[47]. Thus, the presence of scrubber is considered as sustainability indicator. However, one has to bear in mind that some of the ships are produced much earlier than the validity of ECA's regulation. Thus, the decision to install scrubber might be made independent from this regulation. The sustainability indicators is summarized in table 5.1.

Efficiency Index

Improving energy efficiency may have the potential to increase the profit margin by reducing the vessel operational expenses. From environmental viewpoint, higher energy efficiency means reducing the emission. For this reason, some shipowners are often willing to pay more for higher energy efficiency[45]. Chen et al claims that a standard newbuilding panamax vessel with a 5% reduction in fuel consumption per would cost 4.89% more. However they also mention that reduction in fuel consumption per unit have sizable impacts on costs and earnings in the long run[70].

Moreover, Adland et al discovered in their study that energy efficiency has an important influence on the vessel sales price. Using *daily fuel consumption* and *Fuel Efficiency Index (FEI)* as main indicators, they concluded that vesselswith higher energy efficiency tend to have higher price[43]. From the database with 1227 vessels, the information about daily fuel consumption (in ton) is available for 74% of the ships. Fuel consumption has a high correlation with the ship size since bigger ships consume more fuel. Thus, ***fuel consumption is normalized against its average value*** to prevent collinearity in model. *Normalized fuel consumption* is computed as a function of ship's displacement. The data spread and average value (*presented as dotted black line*) are given in figure 5.4 (right). Lastly FEI (in gram/ton.mile) is calculated using equation 5.2.

$$FEI = \text{Daily Fuel Consumption} * \frac{10^6}{(24 * V * DWT)} \tag{5.2}$$

In addition, the presence of gearbox is considered in this research. This is because modern engines tend to be slow speed engines and more efficient. These engines do not require gearbox but cost more than medium/ high speed engines. However, due to its lower efficiency and associated gearbox loss (when it is used), vessels with gearbox tend to consume more fuel *gearbox*. For vessels without gearbox, the voyage cost would be lower in the long run. Thus analyzing gearbox effect on salesprice might lead to a useful result.

Gearbox installation is only essential when higher speed diesel engines (speed between 300 RPM - 1000 RPM) are used. When the low speed engines (speed below 300 RPM) are used, gearbox is not needed[80]. Furthermore, two main reasons why higher speed (4-strokes) engines are favored above low speed (2-strokes) engines are because they are generally cheaper and smaller in size[79]. Therefore only smaller ships are equipped with gearbox; in which 43 handysize and 4 handymax vessels. For this reason, the gearbox analysis is only appropriate for handysize vessels, where 12.1% are registered to have higher speed engines. Thus, gearbox is only considered for handysize vessels.

Another parameter which indicates ship efficiency (in its operation) is the speed[49]. There is always an optimum speed depends on the vessel type, propulsion system and fuel type[48]. Operational speed has an influence on the fuel consumption, building cost and hydrodynamic ship design. It is calculated based on the economical advantage and the service requirements[70]. As a performance indicator, speed is always coupled with the engine power. And higher speed does not mean a better performance, but ships which require least power to achieve certain speed can be considered better than the other. Thus, what researchers try to optimize is usually the speed and power relationship.

To represent speed and power relationship, the *Admiralty coefficient (Ca)* can be used[70]. The general formulation for admiralty constant is expressed using equation 5.3. Generally, values range from 400 to 600, the higher the value the more economic the vessel[44]. To calculate the ship's displacement (Δ, equation 5.4), the necessary block coefficient (C_b) is approximated using two formulations suggested by Barras (C_{b1},

equation 5.5) and Katsoulis (C_{b2}, equation 5.6), where the average value is taken. Just like fuel consumption, Ca has a high correlation with the ship size since it is a factor of ship's displacement. Thus, ***Ca is normalized against its average value*** to prevent collinearity. The data spread against *normalized Ca* as a function of ship's displacement and the average value (*presented as dotted black line*) is given in figure 5.4 (left).

$$C_a = \frac{\Delta^{2/3} \cdot V^3}{P} \tag{5.3}$$

$$\Delta = L \cdot B \cdot T \cdot C_b \tag{5.4}$$

$$C_{b1} = 1.20 - 0.39 \cdot \left(\frac{V}{\sqrt{L_{pp}}} \right) \tag{5.5}$$

$$C_{b2} = 0.8217 \cdot f_{bulk} \cdot L_{pp}^{0.42} \cdot B^{-0.3072} \cdot T^{0.1721} \cdot V^{-0.6135} \qquad f_{\text{bulk}} = 1.04 \tag{5.6}$$

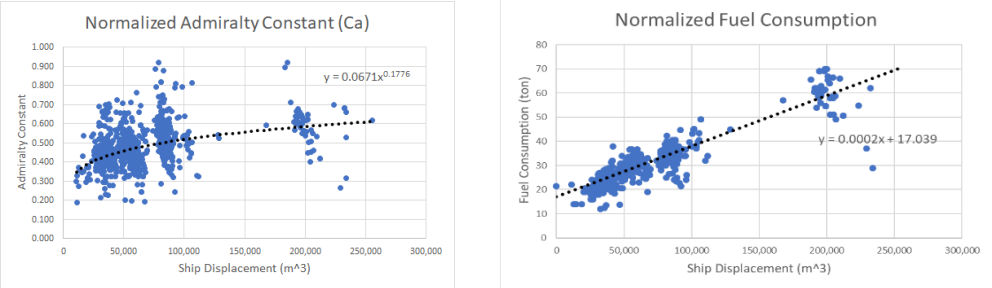


Figure 5.4: Normalization of Admiralty Constant and Fuel Consumption.

5.4. Chapter 5 Conclusion

To conclude, the variables that affect the ship sales price are discussed. These factors are divided into 3 categories. The first category consists of the five variables that are used in the initial model. The second category is deducted by analyzing the initial model to include unexplored variables such as ice class, crane capacity and $\frac{Volume}{DWT}$ ratio in the current pricing model might *increase the model sensitivity* . Lastly, the third category consists of various additional variables which are suggested by the literature study. By this, the fourth research sub-question, *"What are the relevant influencing factors to be considered in the secondhand vessel pricing model according to the literature ?"*, is answered through this chapter. There are *11 variables which are represented by 19 parameters* which are compiled in table 6.1 in chapter 6.

6
Data Preparation

Percaya itu baik, tetapi mengecek lebih baik lagi.

B. J. Habibie

In this chapter, the data preparation process is outlined. First, section 6.1 outlines the variables and statistical overview of the data prior to the statistical testings. Subsequently, the statistical testing process is explained in section 6.2. Two tests are performed, namely the correlation test in subsection 6.2.1 and principal analysis test in subsection 6.2.2. Subsequently, the benchmark for *(variables-based) model combinations* are explained in section 6.3. By this, the fifth research question; *"How can the points of improvement be translated into model set-up?"*, is answered. Finally, section 6.4 concludes this chapter.

6.1. Data Description

6.1.1. Initial Variables Overview

The selected variables are explained in chapter 5. There are initially *11 variables which are represented by 19 parameters* which might influence the secondhand ship prices. They are summarized in table 6.1. They will be statically tested before being used as the explanatory variables in the final model.

No	Variable	Parameter	Unit	Expected Effect
1	*Age*	Age sold	Year	Negative
2	*Vessel Income*	TC-Average TC- Normalized	\$/day -	Positive
3	*Size*	Deadweight	tons	Positive
4	*Cost of Finance*	LIBOR	%	Positive
5	*Orderbook*	% All Bulk Carriers % Vessel Type	% %	Negative
6	$\frac{Volume}{DWT}$ *Level*	Grain Capacity/DWT	tons/m^3	Positive
7	*Ice Class Indicator*	Ice Class Indicator	dummy	Positive
8	*Crane Indicator*	Crane Capacity	tons	Positive
9	*Builder Reputation*	Country Yard Classification Society	categorical categorical	Positive/ Negative
10	*Sustainability Indicator*	Scrubber Indicator Anti Fouling Indicator Fuel Type	dummy dummy categorical	Positive
11	*Efficiency Index*	Normalized Fuel Consumption Gearbox Indicator Fuel Efficiency Index Normalized Admiralty Constant	- dummy gram/ton.mile -	Positive/ Negative

Table 6.1: Initial Key Influences Summary.

6.1.2. Observations Overview

The sales data is obtained from the Clarksons website; initially with the total of 1570 individual sales contract from July 2016 to July 2019. However, there are 343 observations with missing price information; those observations are useless and therefore they are removed. From the remaining 1227 observations, there are numbers of observations with negative age and extreme prices. These may happen for several reasons, but the most obvious one is the human error factor.

Because negative age is physically impossible and a buyer with rational mind would not spend 350 million dollar for secondhand vessel. After filtering these faulty observations, the final database consists of 1200 sales contract and the detail is presented in the first part of table 6.2. Lastly, the information about variables with missing data is presented in second part of table 6.2. Several observations have missing data from multiple variables, thus at the end the *total missing data* is counted based on the whole data.

((a)) Total Observations.

Type	*Initial*	*After Filtering*
Handysize	356	355
Handymax	441	438
Panamax	329	328
Capesize	101	79
Total	**1227**	**1200**

((b)) Missing Observation Overview.

Variable	**Missing Data**
$\frac{Volume}{DWT}$ *Ratio*	28
(Normalized) Fuel Consumption	302
Fuel Efficiency Index	302
Speed (Admiralty Cons.)	47
Total Missing Data	***302***

Table 6.2: Observation Overview.

6.1.3. Ship Sales Overview

The graphs in figure 6.1 illustrates the sales quantity when the observations are taken. From the bar-chart below, one can observe the cyclical tendency of shipping market which has been elaborately discussed in chapter 2 section 2.1. Furthermore, the total price changes overtime is represented in the third sub-figure using stack-lines. From the figure, one can observe that all vessels type generally follows the same selling trend, with the highest sale takes place in April 2017. This marks the global economic upswing which happen as investors positively respond to Trump's pro-growth policies and higher inflation[76].

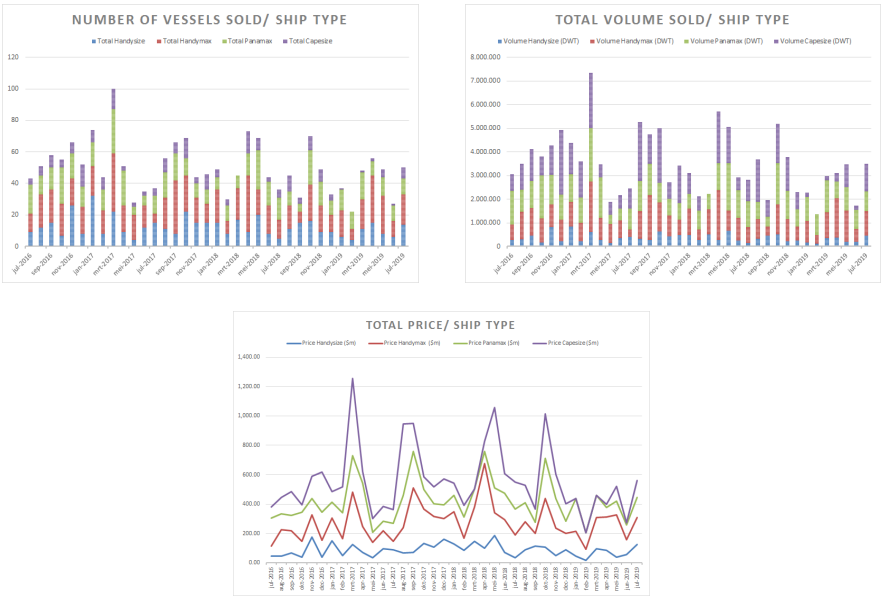


Figure 6.1: Sales Overview per Ship Type.

6.1.4. Numerical Data Overview

To understand the nature of the data, the statistical overview of numerical data is given in table 6.3. The spread of observation is presented as density plots in figure 6.2. From the plots, one can see that observations for most of variables are not normally distributed. Most variables have positively skewed distributions. One possible reason for this is due to the capesize vessels who are considerably larger than the rest. For the financial variables, irregular distribution of density plots represent the common fluctuations of world economy.

Variable	Min	Max	Range	Median	Mean	Std.dev
Age	0.27	32.88	32.61	9.93	11.15	5.70
TC-Average	5,250	20,000	14,750	9,250	9,280	2,224
TC-Normalized	0.98	2.50	1.52	1.57	1.53	0.29
DWT	10,124	229,186	219,062	43,340	51,705	35,833
LIBOR	1.0%	2.9%	1.9%	1.6%	1.9%	0.6%
Orderbook All	8.7%	14.8%	6.2%	11.9%	11.4%	1.6%
Orderbook per Type	5.0%	16.9%	11.9%	9.1%	9.8%	2.5%
$\frac{Volume}{DWT}$ *Ratio*	0.64	2.66	2.02	1.29	1.52	0.42
Crane Capacity	0	200	200	120	76	60
Fuel Efficiency Index	0.60	3.53	2.93	2.06	2.02	0.46
Normalized Fuel Consumption	0.45	1.50	1.04	1.03	1.02	0.12
Normalized Admiralty Constant	0.40	1.85	1.45	1.01	1.02	0.19

Table 6.3: Statistical Summary of Initial Numerical Variables.

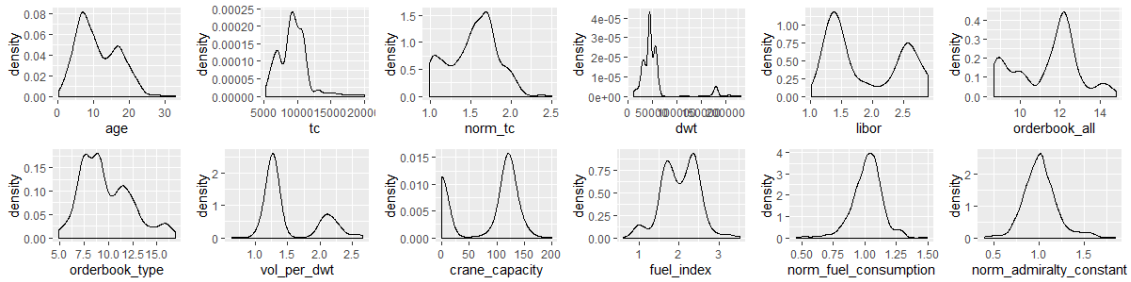


Figure 6.2: Density Plot for All Numerical Variables.

6.2. Statistical Testings

6.2.1. Correlation Test

As what is explained in chapter 4, multicollinearity should be avoided. To assess this, a correlation matrix are made and presented in figure 6.3. Hair et al suggested to remove variables with correlation value above 0.9; whereas the software (R) uses stricter threshold of 0.8 to avoid multicollinearity[75]. In this case, threshold of 0.8 is used. From figure 6.3, one can observe that *DWT* and *Fuel Efficiency Index (FEI)* are highly correlated. In addition, $\frac{Volume}{DWT}$ *Ratio* and *crane capacity* are also highly correlated. Relation between DWT and FEI happens since FEI is a factor of *absolute fuel consumption. Absolute fuel consumption* is highly correlated with size because larger ships consume more fuel. Thus, FEI is removed as DWT has no missing information. Furthermore, $\frac{Volume}{DWT}$ *ratio* is chosen over *crane capacity* since there are less missing observations.

From figure 6.3, one can deduce that the rest of variables are weakly correlated. However, there are few variables such as TC-average, TC-normalized and LIBOR. TC-average and TC-normalized are highly correlated because they represent the same thing. The highly correlation between TC-normalized and LIBOR confirms the validity of hypothesis given in section 5.1 under the **vessel income** discussion. It says that *the normalized TC might represent the fluctuation of world's economy since the value is taken relatively.* Despite of their high correlation, these three variables will be further considered in the model due to their essential role. To conclude, *crane capacity and fuel efficiency index are eliminated* to prevent multicollinearity.

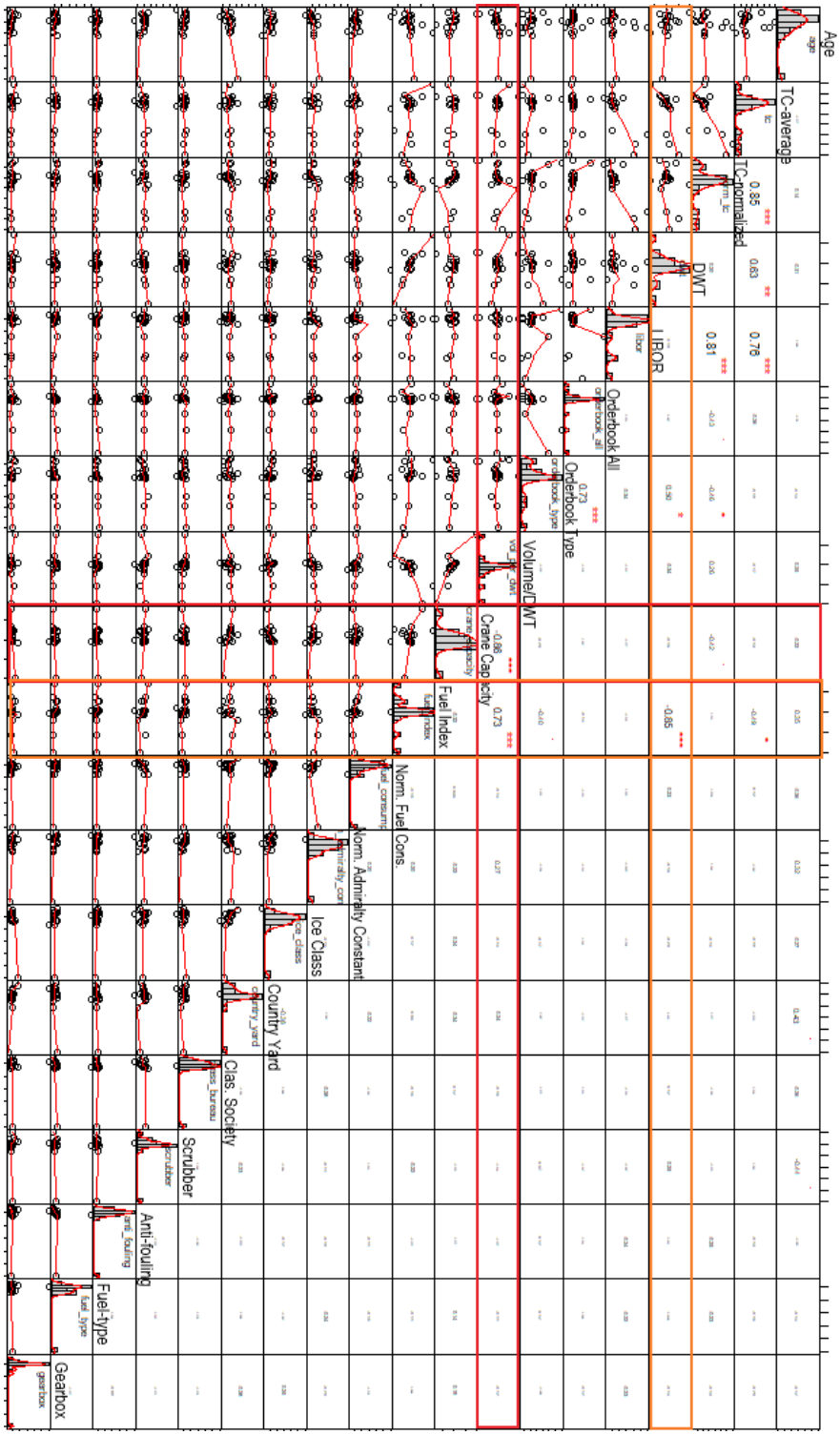


Figure 6.3: Correlation Matrix for All Variables.

6.2.2. Principal Component Analysis (PCA)

PCA is a mathematical algorithm that reduces the dimensionality of numerical variables. It starts to identify directions and the (quantified) variation in the data. Generally, combination-tests are performed, to see the variation that every variable bring in the final result. Insignificant variables will give little variance or contribution. Few problems of using PCA are: measurements are projected to the lower dimensional space and only linear relationships are considered. Lastly, PCA has an univariate computational nature thus it does not consider the multivariate interactions between variables. PCA is useful for GAM, however, decision trees and random forest algorithm, have their own way to assess the multivariate nature[77].

((a)) Combination 1

Variable	Proportion of Variance
Age	24.0%
TC-average	17.8%
DWT	15.5%
LIBOR	13.4%
Orderbook All	11.4%
$\frac{Volume}{DWT}$ *Ratio*	9.4%
Normalized Fuel Consumption	7.5%
Normalized Admiralty Constant	1.1%

((b)) Combination 2

Variable	Proportion of Variance
Age	23.9%
TC-average	19.9%
DWT	16.5%
LIBOR	12.1%
Orderbook per Type	11.4%
$\frac{Volume}{DWT}$ *Ratio*	8.6%
Normalized Fuel Consumption	6.0%
Normalized Admiralty Constant	1.5%

((c)) Combination 3

Variable	Proportion of Variance
Age	22.4%
TC-normalized	17.8%
DWT	15.8%
LIBOR	14.6%
Orderbook All	11.7%
$\frac{Volume}{DWT}$ *Ratio*	9.3%
Normalized Fuel Consumption	7.4%
Normalized Admiralty Constant	1.2%

((d)) Combination 4

Variable	Proportion of Variance
Age	24.5%
TC-normalized	18.7%
DWT	16.1%
LIBOR	12.4%
Orderbook per Type	11.7%
$\frac{Volume}{DWT}$ *Ratio*	8.5%
Normalized Fuel Consumption	5.9%
Normalized Admiralty Constant	2.1%

Table 6.4: Proportion Variance for Various Combinations of Numerical Variables.

In this case, 4 combinations are made out of the 8 numerical variables. For each combination, either TC-average or TC-normalized is used. This also applies for the orderbook% for all vessels and orderbook% per vessel type. The result is given in table 6.4 and visualized in figure E.1 in appendix E. In all cases, one can see that age consistently gives the highest contribution while *Normalized Admiralty Constant (Ca)* consistently contributes the lowest. However, *Normalized Ca* will still be considered further model because it is the only efficiency index which is applicable to the entire size-based type.

6.3. Model Set-up

Combination 1	Combination 2	Combination 3	Combination 4
TC-average	*TC-average*	*TC-normalized*	*TC-normalized*
Orderbook All	*Orderbook per Type*	*Orderbook All*	*Orderbook per Type*

Table 6.5: Model Combinations - Main Set-up.

Vessels are distinguished in five size-based types, namely *all, handysize, handymax, panamax* and *capesize*. This has been discussed in section 3.2.1. Since they represent different market, these size-based types will be tested separately. On top of it, ***four (variables-based) model combinations*** are set up. Based on the extensive discussion in chapter 5, an exception is made for the *gearbox & ice-class indicators* variable. which are only applicable to handysize vessels. Furthermore, the ***four (variables-based) model combinations*** are made based on the previous PCA test. These combinations are outlined in table 6.5. By this, the fifth research question; *"How can the points of improvement be translated into model set-up?"*, is answered.

6.4. Chapter 6 Conclusion

This chapter starts with outlining the selected variables based on the literature study, which are given in table 6.1. Afterward, the sales overview and statistical summary is presented in section 6.1. Moreover, the statistical testings are presented in section 6.2. Lastly, the *model combinations* are outlined in section 6.3. By this, the fifth research question; *"How can the points of improvement be translated into model set-up?"*, is answered. After the multicolinearity test, *fuel index and crane capacity are eliminated.* The remaining variables are listed on table 6.6. They will be used as the regressor in new models. Lastly, the statistical summary of remaining variables for each vessel group are presented in table 6.7.

No	Variable	Parameter	Unit	Type
1	*Age*	Age sold	Year	Numerical
2	*Vessel Income*	TC-Average	$/day	Numerical
		TC- Normalized	-	Numerical
3	*Size*	Deadweight	tons	Numerical
4	*Cost of Finance*	LIBOR	%	Numerical
5	*Orderbook*	% All Bulk Carriers	%	Numerical
		% Vessel Type	%	Numerical
6	$\frac{Volume}{DWT}$ *Level*	Grain Capacity/DWT	tons/m^3	Numerical
7	*Ice Class Indicator*	Ice Class Indicator	-	Categorical
8	*Builder Reputation*	Country Yard	-	Categorical
		Classification Society	-	Categorical
9	*Sustainability Indicator*	Scrubber Indicator	-	Categorical
		Anti Fouling Indicator	-	Categorical
		Fuel Type	-	Categorical
10	*Efficiency Index*	Gearbox Indicator	-	Categorical
		Normalized Fuel Consumption	-	Numerical
		Normalized Admiralty Constant	-	Numerical

Table 6.6: Final Key Influences Summary.

Variable	*All*				*Handysize*				*Handymax*			
	min	*max*	*mean*	*std.dev*	*min*	*max*	*mean*	*std.dev*	*min*	*max*	*mean*	*std.dev*
Age	0.27	32.88	11.15	5.70	0.27	32.88	11.15	5.86	0.66	27.35	11.20	5.73
TC-Average	5,250	20,000	9,280	2,224	5,250	11,500	8,466	1,648	6,000	13,250	9,348	1,753
TC-Normalized	0.98	2.50	1.53	0.29	0.98	1.78	1.48	0.25	0.98	1.75	1.44	0.24
DWT	10,124	229,186	51,705	35,833	10,124	38,907	28,856	6,756	40,064	63,679	53,633	5,031
LIBOR	1.0%	2.9%	1.9%	0.6%	1.0%	2.9%	1.9%	0.6%	1.0%	2.9%	1.9%	0.6%
Orderbook All	8.7%	14.8%	11.4%	1.6%	8.7%	14.8%	11.4%	1.6%	8.7%	14.8%	11.3%	1.6%
Orderbook Type	5.0%	16.9%	9.8%	2.5%	5.0%	13.2%	8.5%	2.2%	7.2%	16.9%	9.8%	2.4%
$\frac{Volume}{DWT}$ *Ratio*	0.641	2.656	1.517	0.424	0.682	2.403	1.304	0.089	0.690	2.120	1.260	0.069
Norm. Fuel Consump.	0.45	1.50	1.02	0.12	0.51	1.50	0.96	0.12	0.52	1.31	1.06	0.10
Norm. Admiralty Cons.	0.40	1.85	1.02	0.19	0.51	1.66	1.03	0.19	0.40	1.61	0.98	0.18
Price	1.00	50.00	10.86	7.16	1.00	41.00	7.67	4.91	2.00	50.00	10.58	5.94

Variable	*Panamax*				*Capesize*			
	min	*max*	*mean*	*std.dev*	*min*	*max*	*mean*	*std.dev*
Age	0.30	27.94	11.69	5.60	0.27	17.93	8.70	4.56
TC-Average	6,000	11,500	9,020	1,541	6,350	20,000	13,623	3,751
TC-Normalized	0.99	2.08	1.66	0.32	1.05	2.50	1.67	0.41
DWT	42,579	43,658	43,089	321	114,248	229,186	178,893	16,633
LIBOR	1.0%	2.9%	1.9%	0.6%	1.2%	2.9%	1.7%	0.5%
Orderbook All	8.7%	14.8%	11.4%	1.6%	8.7%	14.4%	11.3%	1.8%
Orderbook Type	6.5%	13.0%	10.2%	2.1%	10.3%	16.2%	13.5%	2.1%
$\frac{Volume}{DWT}$ *Ratio*	1.659	2.656	2.170	0.159	0.641	1.195	1.105	0.057
Norm. Fuel Consump.	0.61	1.27	1.01	0.10	0.45	1.24	1.03	0.16
Norm. Admiralty Cons.	0.48	1.85	1.06	0.20	0.44	1.59	1.02	0.17
Price	2.76	48.50	11.83	6.80	7.50	50.00	22.71	9.68

Table 6.7: Statistical Summary for All Vessel Types.

7
General Additive Models

Betre å vite visst enn berre å gjette.

Norwegian Proverbs

This chapter outlines GAM models and results. The *Backward Elimination* will be used to systematically remove insignificant variables. The brief explanation about *Backward Elimination* and *Significance Level Threshold* is outlined in section 7.1. Afterward, the *Single Significant Test* for each numerical variables used in model is presented in section 7.2. Subsequently, the result of model first iteration is presented as the starting point in section 7.3. After applying the *Backward Elimination*, every insignificant variables are eliminated. Moreover, the models with the most promising results are chosen and presented in section 7.4. Lastly, the result discussion in the light of initial model and the project goals are reviewed in 7.5.

7.1. Backward Elimination Procedure

To complement GAM, the *Backward Elimination* is used. The main purpose is to eliminate insignificant variables and get the final model which only consists of important variables[84]. The significance level of a variable is indicated by the ***p-value***. This concept is introduced because there are two types unavoidable errors in a statistical analysis. The first one is the *alpha error* or false positives; which refers to the probability of being wrong when one thinks he is right. The value of type 1 error is is given by ***p-value***[54]. ***P-value*** is used as main indicator here because the GAM result is given with the significance level of each variable. Various significant codes are applied for different ***p-value***. These values is summarized in table 7.1. In regression problem, retaining the variable with significance level 5% or less is the most common practice[83].

Significance Code	'***'	'**'	'*'	'.'	' '
P-value	0.1%	1%	5%	10%	100%

Table 7.1: R-Significance Code Explanation.

Furthermore, the steps of implementing Backward Elimination can be summarized as following[83]:

1. Select a significance level. Based on literature, the (maximum allowable) threshold of ***5% p-value ['*']*** is selected[54]. 5% is the most common threshold to set regarding the confidence interval[54][89].

2. Run the model with all variables to discover the *p-values* of model with *all variables*. Then gradually remove the variables with *large p-values [' '] & ['.']*, until only significant variables are left.

3. When all insignificant variables are removed, final model is ready. Based on model set-up, there will be 4 models for each vessel type (see table 6.5). Lastly, the best model is selected based the highest R-squared value[85].

7.2. Single Significance Test

First, all variables are tested to see their one-on-one effect on price. The nature of relationship between all variables and price can be observed from figure 7.1. *Age* & *Orderbook* has negative effect in (almost) all cases. However, *LIBOR*, *DWT*, *TC-rate* and *norm. Ca* generally have positive trends. At last, it will be used to judge the collinearity in the model by comparing the single test plots with the integrated models.

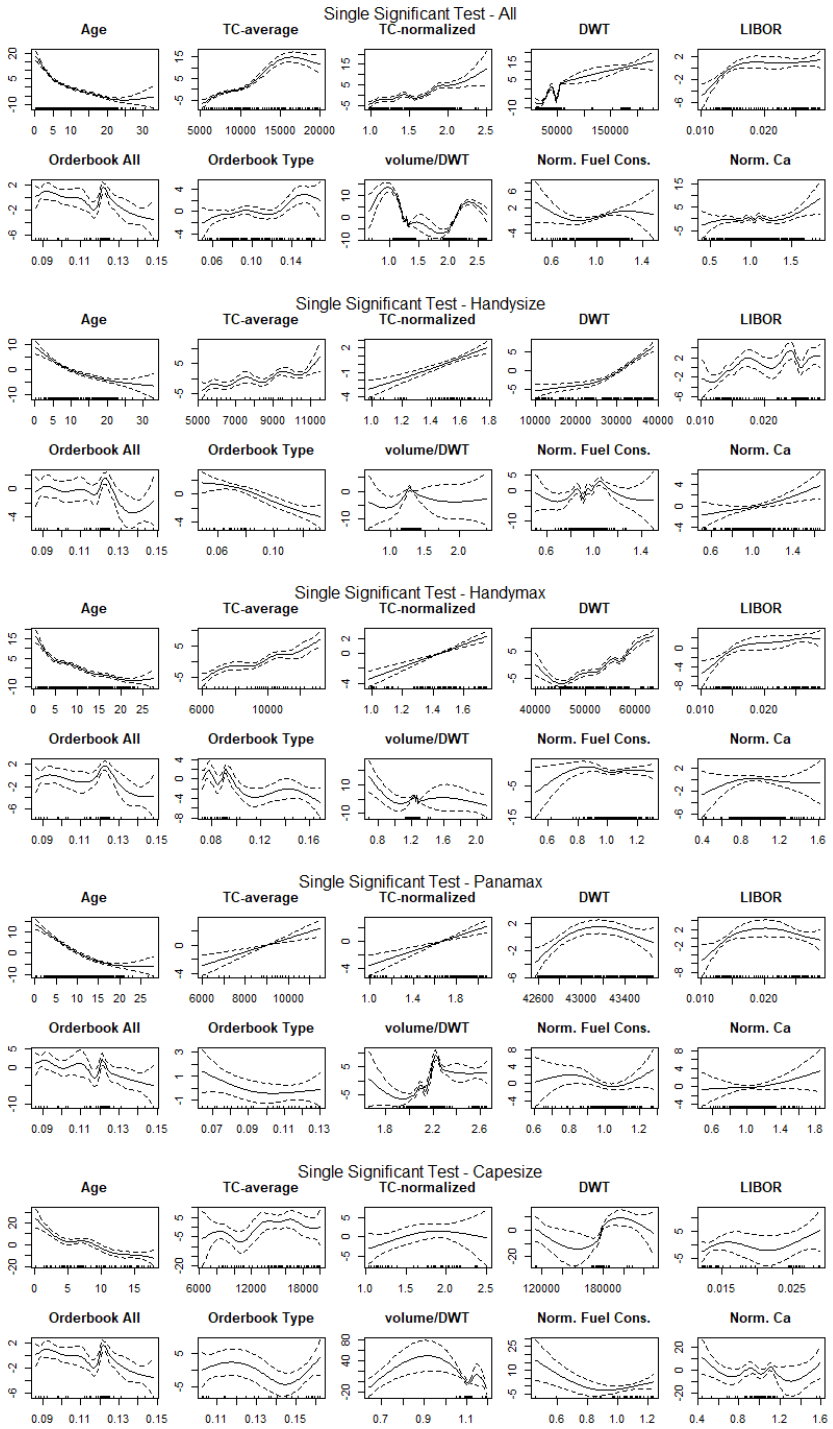


Figure 7.1: Single Significant Test per Vessel Type.

The summary about the significance level of all variables for every vessel type are given in table 7.2 from which one can judge the relationships strengths. Among all variables, *age* and *DWT* are shown to have high influence on sales prices. Moreover several remarks about unique (unexpected) trends of single test plots are:

Variable	***All***		***Handysize***		***Handymax***		***Panamax***		***Capesize***	
	R-sq.	*S.L.*	*R-sq.*	*S.L.*	*R-sq.*	*S.L.*	*R-sq.*	*S.L.*	*R-sq.*	*S.L.*
Age	0.437	***	0.360	***	0.537	***	0.608	***	0.626	***
TC-Average	0.257	***	0.134	***	0.194	***	0.043	***	0.108	.
TC-Normalized	0.121	***	0.102	***	0.088	***	0.060	***	0.027	
DWT	0.388	***	0.403	***	0.532	***	0.049	**	0.290	***
LIBOR	0.035	***	0.108	***	0.096	***	0.038	**	0.028	
Orderbook All	0.030	***	0.048	**	0.059	***	0.053	**	0.067	
Orderbook Type	0.015	**	0.094	***	0.099	***	0.005		0.065	
$\frac{Volume}{DWT}$ *Ratio*	0.290	***	0.031	*	0.092	***	0.463	***	0.223	**
Norm. Fuel Consumption	0.014	*	0.103	***	0.011		0.022		0.166	*
Norm. Admiralty Constant (Ca)	0.008	.	0.032	**	0.003		0.004		0.083	

Table 7.2: GAM - Single Significant Test Summary.

All Vessels

As expected, *DWT* has a positive trend except for a valley at around 50k DWT vessels. This valley corresponds to the handymax which has similar DWT as panamax, but lower in price. Moreover, orderbook type has oddly a positive trend although they are always negative when observed individually (per type). This happens as certain percentage range are typical for particular type. Lower range is more typical for handysize while higher range is for capesize. Since it encompasses every types, the initially small value is related to (cheaper) handysize, whereas higher price at the higher orderbook% corresponds to more expensive capesize.

Lastly, *vol/DWT* exhibits unique behaviour which also happens as certain values are specific for particular vessel type. For instance, capesize's vol/DWT is concentrated at 1.05-1.1 while handysize & handymax are at 1.2-1.5. Moreover, panamax concentrated range is at around 2-2.4 and around 2.6. Thus the initial peak around 1.0 indicates the capesize, then it rapidly decreases following the lower price of smaller vessels. Finally the second peak at around 2-2.4 indicates the panamax market.

Handysize Vessels

First, *vol/DWT* has a mountain-like shape with peak at 1.25 where most of handysize are. The beginning and end have odd shape due to low data density. Lastly, *fuel consumption* seems to fluctuate at a steady level (near 0 mil.$). This might indicate that for handysize, fuel consumption has a little effect on price.

Handymax Vessels

DWT has initially a negative trend which might indicate the specialized or younger vessels with higher price. Another odd trend is the increasing *fuel consumption* corresponds to an increasing sales price; whereas it is expected to negatively influence the price. This might happen due to specialized vessels which cost higher and consume more fuel. Lastly, *Ca* seems to level off at 0 mil.$, which indicates its little effect on price.

Panamax Vessels

The only odd tendency of panamax is the decreasing trend at the end for *DWT*. This might happen due to another submarket for panamax bigger than 43k DWT which is less popular thus cost less.

Capesize Vessels

Capesize is a unique market as vessels' cost is very high and there are bigger risk. It is more difficult to find buyers for old capesize. This might explains the different trends of many capesize variables when compared to other types. The first unique trend is decreasing trend in the beginning of *DWT* which might correspond to low data density. The same reasoning might applies to the odd shape in the beginning of *norm. fuel consumption* and *norm. Ca*; as well as *vol/DWT* when ratio is less than 1.1. The last anomaly happens as orderbook type got really high (above 14.5%). Higher orderbook means higher waiting time for the newbuilding. Rich owners who do not want to wait would go to secondhand market as an alternative, and they will pay more for the secondhand.

7.3. GAM Modelling

Before starting the model, it is worth to mention to GAM will automatically eliminate observations with missing variables. Thus, by looking at table 6.2, including variables like *normalized admiralty constant* and *normalized fuel consumption* with 302 missing observations in total will substantially reduce sample size (n).

All Vessels

Iteration	First											
Model	All1			All2			All3			All4		
Parametric	Estimate	p-value		*Estimate*	p-value		Estimate	p-value		Estimate	p-value	
(Intercept)	6.66	< 2e-16	***	6.65	< 2e-16	***	6.66	< 2e-16	***	6.64	< 2e-16	***
scrubberY	2.54	0.00	***	2.56	0.00	***	2.45	0.00	***	2.43	0.00	***
anti_foulingY	-0.32	0.33		-0.31	0.35		-0.33	0.31		-0.32	0.33	
fuel_typeMDO	-1.55	0.16		-1.70	0.13		-1.71	0.12		-1.68	0.13	
japanY	2.28	< 2e-16	***	2.29	< 2e-16	***	2.30	< 2e-16	***	2.31	< 2e-16	***
chinaY	-0.10	0.63		-0.10	0.65		-0.09	0.67		-0.07	0.76	
southkoreaY	2.85	< 2e-16	***	2.85	< 2e-16	***	2.74	< 2e-16	***	2.70	< 2e-16	***
rest_countryY	1.63	0.00	***	1.62	0.00	***	1.71	0.00	***	1.70	0.00	***
bvY	1.42	0.00	***	1.44	0.00	***	1.47	0.00	***	1.49	0.00	***
ccsY	1.82	0.00	***	1.78	0.00	***	1.78	0.00	***	1.71	0.00	***
nkkY	1.59	< 2e-16	***	1.60	< 2e-16	***	1.56	< 2e-16	***	1.56	< 2e-16	***
rest_classY	1.82	< 2e-16	***	1.83	< 2e-16	***	1.85	< 2e-16	***	1.88	< 2e-16	***
Smooth-term	edf	p-val.	sig.	edf	p-val.	sig.	edf	p-val.	sig.	edf	p-val.	sig.
s(age)	6.82	< 2e-16	***	6.88	< 2e-16	***	6.86	< 2e-16	***	6.81	< 2e-16	***
s(tc)	6.03	0.00	***	6.22	< 2e-16	***						
s(norm_tc)							3.27	0.00	***	2.81	< 2e-16	***
s(dwt)	8.53	< 2e-16	***	8.52	0.00	***	8.50	< 2e-16	***	8.51	< 2e-16	***
s(libor)	1.00	0.02	*	1.00	0.21		1.89	0.05	.	1.00	0.07	.
s(orderbook_all)	1.00	0.03	*				1.28	0.13				
s(orderbook_type)				1.01	0.61					1.35	0.24	
s(norm_fuel_consumption)	1.00	0.29		1.00	0.31		1.00	0.23		1.00	0.18	
s(vol_per_dwt)	8.39	0.00	***	8.40	0.00	***	8.30	0.00	***	8.32	0.00	***
s(norm_admiralty_constant)	1.68	0.61		1.67	0.62		1.69	0.62		1.66	0.61	
Result	R-sq.	GCV	n	R-sq.	GCV	n	R-sq.	GCV	n	R-sq.	GCV	n
	0.761	9.281	897	0.760	9.325	897	0.760	9.313	897	0.759	9.331	897

Figure 7.2: Comparison of Initial Models for All Vessels.

Following the *backward elimination procedure*, the models will include all variables during the first iteration. The initial result for all vessels are summarized in table 7.2. *The insignificant variables which will be removed in the next iteration are highlighted with grey color. There are considerable missing observations from the fuel consumption and admiralty constant; thus removing them has increased the number of observations in later iterations.The optimum models are obtained on the third iteration.*

Iteration	Third											
Model	All1			All2			All3			All4		
Parametric	Estimate	p-value		*Estimate*	p-value		Estimate	p-value		Estimate	p-value	
(Intercept)	7.56	< 2e-16	***	7.56	< 2e-16	***	7.52	< 2e-16	***	7.53	< 2e-16	***
scrubberY	2.01	0.00	***	2.01	0.00	***	2.13	0.00	***	2.09	0.00	***
anti_foulingY												
fuel_typeMDO												
japanY	2.01	0.00	***	2.00	0.00	***	2.07	0.00	***	2.04	0.00	***
chinaY												
southkoreaY	2.84	0.00	***	2.80	0.00	***	2.88	0.00	***	2.85	0.00	***
rest_countryY	1.23	0.01	**	1.18	0.01	*	1.38	0.00	**	1.31	0.00	**
bvY	1.73	0.00	***	1.74	0.00	***	1.72	0.00	***	1.72	0.00	***
ccsY	2.11	0.00	***	2.06	0.00	***	2.10	0.00	***	2.07	0.00	***
nkkY	1.77	< 2e-16	***	1.79	< 2e-16	***	1.72	< 2e-16	***	1.75	< 2e-16	***
rest_classY	1.95	< 2e-16	***	1.97	< 2e-16	***	1.98	< 2e-16	***	1.99	< 2e-16	***
Smooth-term	edf	p-val.	sig.	edf	p-val.	sig.	edf	p-val.	sig.	edf	p-val.	sig.
s(age)	7.04	< 2e-16	***	7.03	< 2e-16	***	7.05	< 2e-16	***	7.12	< 2e-16	***
s(tc)	6.92	0.00	***	7.10	0.00	***						
s(norm_tc)							4.04	0.00	***	3.93	0.00	***
s(dwt)	8.39	< 2e-16	***	8.39	< 2e-16	***	8.42	< 2e-16	***	8.39	< 2e-16	***
s(libor)	1.79	0.00	***	4.29	0.00	**	2.01	0.00	***	7.71	0.00	**
s(orderbook_all)	1.00	0.00	**				1.00	0.00	**			
s(orderbook_type)												
s(norm_fuel_consumption)												
s(vol_per_dwt)	8.55	< 2e-16	***	8.53	< 2e-16	***	8.58	0.00	***	8.58	0.00	***
s(norm_admiralty_constant)												
Result	R-sq.	GCV	n	R-sq.	GCV	n	R-sq.	GCV	n	R-sq.	GCV	n
	0.768	11.976	1172	0.767	12.021	1172	0.769	11.883	1172	0.769	11.937	1172

Figure 7.3: Final Results After Backward Elimination - All Vessels.

The final result is given in figure 7.3. The intermediate results are presented in appendix F. The compari-

son is done based on the R-squared and the General Cross Validation (GCV) values. Higher R-squared implies better fitting, while a lower GCV is related to the (lower) error[78]. Finally, ***model 3 has the highest R-squared***.

Handysize Vessels

Iteration	First											
Model	Handysize1			Handysize2			Handysize3			Handysize4		
Parametric	Est.	p-val.	sig.	Est.	p-val.	sig.	Est.	p-val.	sig.	Est.	p-val.	sig.
(Intercept)	5.66	< 2e-16	***	5.67	< 2e-16	***	5.66	< 2e-16	***	5.63	< 2e-16	***
ice_classY	-1.26	0.18		-1.25	0.18		-1.27	0.17		-1.19	0.19	
anti_foulingY	-0.20	0.79		-0.11	0.89		-0.13	0.86		-0.11	0.89	
gearboxY	-0.14	0.95		-0.19	0.93		-0.21	0.92		-0.48	0.82	
japanY	0.54	0.22		0.54	0.22		0.51	0.25		0.48	0.28	
chinaY	-0.11	0.82		-0.07	0.89		-0.03	0.95		0.03	0.94	
southkoreaY	4.90	0.00	***	4.82	0.00	***	4.83	0.00	***	4.86	0.00	***
rest_countryY	0.33	0.69		0.37	0.65		0.35	0.67		0.25	0.76	
bvY	1.73	0.00	**	1.78	0.00	**	1.77	0.00	**	1.70	0.00	**
ccsY	0.47	0.63		0.51	0.60		0.44	0.65		0.37	0.70	
nkkY	1.35	0.00	**	1.38	0.00	***	1.43	0.00	***	1.43	0.00	***
rest_classY	2.11	0.00	***	2.01	0.00	***	2.02	0.00	***	2.13	0.00	***
Smooth-term	edf	p-val.	sig.	edf	p-value		edf	p-value		edf	p-value	
s(age)	2.24	< 2e-16	***	2.26	< 2e-16	***	2.28	< 2e-16	***	2.25	< 2e-16	***
s(tc)	4.20	0.55		1.00	0.18							
s(norm_tc)	1.35	0.01	**				1.00	0.18		4.91	0.22	
s(dwt)	1.35	0.01	**	1.35	0.00	**	1.37	0.00	**	1.32	0.01	**
s(libor)	1.00	0.08	.	1.00	0.42		1.00	0.14		1.00	0.65	
s(orderbook_all)	1.00	0.27					1.00	0.47				
s(orderbook_type)				1.00	0.15					1.00	0.14	
s(norm_fuel_consumption)	1.00	0.92		1.00	0.84		1.00	0.85		1.00	0.79	
s(vol_per_dwt)	1.00	0.86		1.00	0.64		1.00	0.69		1.00	0.80	
s(norm_admiralty_constant)	1.00	0.10	.	1.00	0.11		1.00	0.11		2.30	0.26	
Result	R-sq.	GCV	n	R-sq.	GCV	n	R-sq.	GCV	n	R-sq.	GCV	n
	0.594	11.222	247	0.588	11.216	247	0.591	9.199	247	0.599	11.167	247

Figure 7.4: Comparison of Initial Models for Handysize Vessels.

The result for first iteration is summarized in table 7.4. Handysize is the only type which considers the gearbox and ice-class indicator, as discussed in chapter 5 section 5.3 under the *Sustainability Indicator*. However, there is no single handysize vessel which has scrubber installed. *The insignificant variables which will be removed in the next iteration are highlighted with grey color. There are 108 missing observations from the fuel consumption and Ca; thus removing them has increased "n" in later iterations.*

Iteration	Fifth						**Sixth (Final)**					
Model	Handysize1			Handysize2			Handysize1			Handysize2		
Parametric	Est.	p-val.	sig.	Est.	p-val.	sig.	Est.	p-val.	sig.	Est.	p-val.	sig.
(Intercept)	6.55	0.00	***	6.74	0.00	***	7.45	< 2e-16	***	7.45	< 2e-16	***
ice_classY												
anti_foulingY												
gearboxY												
japanY												
chinaY												
southkoreaY	4.26	0.00	***	4.29	0.00	***	4.33	0.00	***	4.33	0.00	***
rest_countryY												
bvY	1.10	0.21		0.85	0.33							
ccsY												
nkkY	0.94	0.23		0.79	0.30							
rest_classY	0.94	0.22		0.70	0.34							
Smooth-term	edf	p-value		edf	p-value		edf	p-value		edf	p-value	
s(age)	6.07	< 2e-16	***	6.26	<2e-16	***	6.03	< 2e-16	***	6.21	<2e-16	***
s(tc)	5.69	0.00	***				5.09	0.00	***			
s(norm_tc)				1.00	<2e-16	***				1.05	<2e-16	***
s(dwt)	1.00	< 2e-16	***	1.00	<2e-16	***	1.00	< 2e-16	***	1.00	<2e-16	***
s(libor)												
s(orderbook_all)												
s(orderbook_type)												
s(norm_fuel_consumption)												
s(vol_per_dwt)												
s(norm_admiralty_constant)												
Result	R-sq.	GCV	n	R-sq.	GCV	n	R-sq.	GCV	n	R-sq.	GCV	n
	0.648	8.968	355	0.644	8.953	355	0.648	8.866	355	0.646	8.829	355

Figure 7.5: Final Results After Backward Elimination - Handysize Vessels.

The optimum models are obtained on the 6th iteration. The final models are given in figure 7.5. *Orderbook* is removed on the 4th iteration and it reduces the number of models from four to two. The intermediate

results can be found in appendix F. Among all, **model 1 is voted as the best.**

Handymax Vessels

The first iteration for handymax is given in table 7.6. *The insignificant variables which will be removed in the next iteration are highlighted with grey color. There are considerable missing observations from the fuel consumption and admiralty constant; thus removing these (insignificant) variables has increased the number of observations in later iterations.* In this case, *the optimum models are obtained on the fourth and fifth iteration.*

Iteration	First											
Model	Handymax1			Handymax2			Handymax3			Handymax4		
Parametric	Est.	p-val.	sig.	Est.	p-val.	sig.	Est.	p-val.	sig.	Est.	p-val.	sig.
(Intercept)	6.70	< 2e-16	***	6.72	< 2e-16	***	6.72	< 2e-16	***	6.72	< 2e-16	***
scrubberY	2.62	0.00	***	2.60	0.00	***	2.51	0.00	***	2.50	0.00	***
anti_foulingY	0.13	0.73		0.14	0.71		0.10	0.78		0.13	0.73	
fuel_typeMDO	-0.72	0.43		-0.79	0.39		-0.69	0.45		-0.69	0.45	
japanY	3.03	< 2e-16	***	3.00	< 2e-16	***	2.98	< 2e-16	***	2.97	< 2e-16	***
chinaY	-0.70	0.01	*	-0.72	0.01	*	-0.69	0.02	*	-0.71	0.01	*
southkoreaY	2.46	0.00	***	2.48	0.00	***	2.41	0.00	***	2.42	0.00	***
rest_countryY	1.92	0.00	***	1.95	0.00	***	2.02	0.00	***	2.03	0.00	***
bvY	1.57	0.00	***	1.61	0.00	***	1.54	0.00	***	1.55	0.00	***
ccsY	2.07	0.00	***	2.03	0.00	***	2.08	0.00	***	2.07	0.00	***
nkkY	1.57	0.00	***	1.57	0.00	***	1.58	0.00	***	1.57	0.00	***
rest_classY	1.50	0.00	***	1.51	0.00	***	1.52	0.00	***	1.52	0.00	***
Smooth-term	edf	p-val.	sig.	edf	p-val.	sig.	edf	p-val.	sig.	edf	p-val.	sig.
s(age)	4.73	< 2e-16	***	4.83	< 2e-16	***	4.99	< 2e-16	***	4.99	< 2e-16	***
s(tc)	2.40	0.00	***	2.25	0.01	*						
s(norm_tc)							2.08	0.00	***	2.01	0.01	**
s(dwt)	3.13	0.00	***	3.18	0.00	***	3.58	0.00	***	3.82	0.00	***
s(libor)	1.00	0.09	.	1.00	0.18		1.44	0.11		1.31	0.03	*
s(orderbook_all)	1.00	0.05	.				1.00	0.60				
s(orderbook_type)				1.00	0.02	*				1.00	0.28	
s(norm_fuel_consumption)	1.00	0.63		1.00	0.62		1.00	0.47		1.00	0.48	
s(vol_per_dwt)	2.39	0.00	***	2.37	0.00	***	2.34	0.00	**	2.33	0.00	**
s(norm_admiralty_constant)	1.24	0.77		1.18	0.83		1.62	0.67		1.59	0.70	
Result	R-sq.	GCV	n	R-sq.	GCV	n	R-sq.	GCV	n	R-sq.	GCV	n
	0.82	5.298	363	0.82	5.281	363	0.823	5.223	363	0.823	5.238	363

Figure 7.6: Comparison of Initial Models for Handymax Vessels.

The final result is given in figure 7.7. The intermediate results can be found in appendix F subsection F.2.3. On the second iteration, the *orderbook and libor*, thus the number of models decreases from four to two. Among the final models, **model 3 from the 4th iteration has the best fitting quality.**

Iteration	Fourth									**Fifth (Final)**					
Model	Handymax1			Handymax2			**Handymax3 (final)**			Handymax1			Handymax2		
Parametric	Est.	p-val.	sig.	Est.	p-val.	sig.	Est.	p-val.	sig.	Est.	p-val.	sig.	Est.	p-val.	sig.
(Intercept)	7.39	< 2e-16	***	7.40	< 2e-16	***	7.41	< 2e-16	***	7.54	< 2e-16	***	7.60	< 2e-16	***
scrubberY															
anti_foulingY															
fuel_typeMDO															
japanY	2.19	0.00	***	2.12	0.00	***	2.12	0.00	***	1.97	0.00	***	1.83	0.00	***
chinaY															
southkoreaY															
rest_countryY	0.87	0.15		1.12	0.06	.	1.17	0.05	*						
bvY	1.65	0.00	***	1.54	0.00	***	1.58	0.00	***	1.69	0.00	***	1.60	0.00	***
ccsY	1.68	0.00	***	1.81	0.00	***	1.83	0.00	***	1.64	0.00	***	1.76	0.00	***
nkkY	1.87	0.00	***	1.84	0.00	***	1.82	0.00	***	1.95	0.00	***	1.96	0.00	***
rest_classY	2.20	0.00	***	2.21	0.00	***	2.19	0.00	***	2.25	0.00	***	2.28	0.00	***
Smooth-term	edf	p-val.	sig.	edf	p-val.	sig.	edf	p-val.	sig.	edf	p-val.	sig.	edf	p-val.	sig.
s(age)	1.99	< 2e-16	***	1.90	< 2e-16	***	1.87	< 2e-16	***	1.89	< 2e-16	***	1.75	< 2e-16	***
s(tc)	3.40	< 2e-16	***							3.35	< 2e-16	***			
s(norm_tc)				1.83	0.00	***	1.00	0.00	**				1.78	0.00	***
s(dwt)	2.73	0.00	***	2.85	0.00	***	2.87	0.00	***	2.71	0.00	***	2.89	0.00	***
s(libor)				1.46	0.03	*	1.28	0.01	*				1.56	0.04	*
s(orderbook_all)															
s(orderbook_type)							1.00	0.03	*						
s(norm_fuel_consumption)															
s(vol_per_dwt)															
s(norm_admiralty_constant)															
Result	R-sq.	GCV	n	R-sq.	GCV	n	R-sq.	GCV	n	R-sq.	GCV	n	R-sq.	GCV	n
	0.685	11.489	438	0.684	11.513	438	0.684	11.524	438	0.683	11.162	438	0.682	11.564	438

Figure 7.7: Final Results After Backward Elimination - Handymax Vessels.

Panamax Vessels

The result for first iteration for panamax vessels is summarized in table 7.8. When all variables are included, one can conclude that *model 1 has initially the highest R-squared value.* This vessel group initially considers all variables except gearbox. However, a change might take place as the backward elimination procedure is implemented. For panamax vessels, optimum models are obtained on the sixth iteration. The intermediate results are presented in appendix F subsection F.2.4.

Iteration	First											
Model	Panamax1			Panamax2			Panamax3			Panamax4		
Parametric	Est.	p-val.	sig.	Est.	p-val.	sig.	Est.	p-val.	sig.	Est.	p-val.	sig.
(Intercept)	6.53	< 2e-16	***	6.52	< 2e-16	***	6.50	< 2e-16	***	6.51	< 2e-16	***
scrubberY	2.29	0.04	*	2.40	0.03	*	2.48	0.03	*	2.29	0.04	*
anti_foulingY	-0.37	0.59		-0.31	0.64		-0.35	0.60		-0.36	0.59	
fuel_typeMDO	-6.69	0.07	.	-7.23	0.05	*	-6.86	0.05	.	-7.04	0.05	*
japanY	2.55	0.00	***	2.55	0.00	***	2.58	0.00	***	2.60	0.00	***
chinaY	0.61	0.24		0.64	0.21		0.71	0.16		0.63	0.22	
southkoreaY	2.22	0.00	***	2.25	0.00	***	2.24	0.00	***	2.30	0.00	***
rest_countryY	1.15	0.19		1.08	0.23		0.97	0.27		0.98	0.26	
bvY	0.10	0.86		0.08	0.89		0.10	0.85		0.13	0.82	
ccsY	2.32	0.00	***	2.34	0.00	***	2.35	0.00	***	2.34	0.00	***
nkkY	1.76	0.00	***	1.79	0.00	***	1.70	0.00	***	1.72	0.00	***
rest_classY	2.35	0.00	***	2.30	0.00	***	2.34	0.00	***	2.32	0.00	***
Smooth-term	edf	p-val.	sig.	edf	p-val.	sig.	edf	p-val.		edf	p-val.	sig.
s(age)	2.33	0.00	***	2.37	< 2e-16	***	2.31	< 2e-16	***	2.34	< 2e-16	***
s(tc)	1.69	0.52		1.00	0.94							
s(tc_norm)				2.63	0.01	*	1.00	0.00	**	2.35	0.00	**
s(dwt)	2.11	0.17		2.63	0.01	*	1.88	0.26		1.67	0.23	
s(libor)	1.00	0.46		1.00	0.27		1.00	0.21		1.00	0.34	
s(orderbook_all)	1.50	0.62					1.75	0.34				
s(orderbook_type)				3.35	0.44					1.00	0.86	
s(norm_fuel_consumption)	2.43	0.22		1.00	0.08	.	1.00	0.06	.	1.00	0.04	*
s(vol_per_dwt)	7.59	0.00	***	7.44	0.00	***	7.49	0.00	***	7.19	0.00	***
s(norm_admiralty_constant)	1.97	0.23		2.07	0.18		2.07	0.18		2.10	0.16	
Result	R-sq.	GCV	n	R-sq.	GCV	n	R-sq.	GCV	n	R-sq.	GCV	n
	0.744	10.426	243	0.745	10.395	243	0.749	10.136	243	0.749	10.138	243

Figure 7.8: Comparison of Initial Models for Panamax Vessels.

The result of 6th iteration is presented in figure 7.9. *The insignificant variables which will be removed in the next iteration are highlighted with grey color. There are considerable missing observations from the fuel consumption and admiralty constant; thus removing these (insignificant) variables has increased the number of observations in later iterations.* Removing orderbook and TC-rate has decreased the model combinations from 4 to 1. The final outcome is presented in figure 7.9.

Iteration	Fourth			Fifth			**Sixth (Final)**		
Model	Panamax			Panamax			Panamax		
Parametric	Est.	p-val.	sig.	Est.	p-val.	sig.	Est.	p-val.	sig.
(Intercept)	9.80	< 2e-16	***	10.74	< 2e-16	***	10.99	< 2e-16	***
scrubberY	2.54	0.00	**	2.31	0.01	**	2.29	0.01	**
anti_foulingY									
fuel_typeMDO									
japanY	1.37	0.00	**	1.33	0.00	**	1.26	0.00	**
chinaY									
southkoreaY									
rest_countryY									
bvY									
ccsY	1.89	0.02	*	0.92	0.12				
nkkY	0.87	0.22							
rest_classY	1.20	0.08	.						
Smooth-term	edf	p-val.	sig.	edf	p-val.	sig.	edf	p-val.	sig.
s(age)	7.15	0.00	***	7.08	0.00	***	7.27	0.00	***
s(tc)									
s(tc_norm)									
s(dwt)	2.91	0.00	***	3.09	0.00	***	3.51	0.00	***
s(libor)									
s(orderbook_all)									
s(orderbook_type)									
s(norm_fuel_consumption)									
s(vol_per_dwt)	8.18	0.00	***	8.27	0.00	***	8.22	0.00	***
s(norm_admiralty_constant)									
Result	R-sq.	GCV	n	R-sq.	GCV	n	R-sq.	GCV	n
	0.748	12.611	328	0.747	12.567	328	0.747	12.561	328

Figure 7.9: Final Results After Backward Elimination - Panamax Vessels.

Capesize Vessels

The result for first iteration for capesize is presented in table 7.10. *The insignificant variables which will be removed in the next iteration are highlighted with grey color. There are considerable missing observations from the fuel consumption and admiralty constant; thus removing them has increased the number of observations in later iterations.* Also, removing TC-rate and orderbook has decreased model combinations from 4 to 3.

Iteration	First											
Model	Capesize1			Capesize2			Capesize3			Capesize4		
Parametric	Est.	p-val.	sig.	Est.	p-val.	sig.	Est.	p-val.	sig.	Est.	p-val.	sig.
(Intercept)	14.41	0.00	***	16.86	0.00	***	17.66	0.00	***	17.61	0.00	***
scrubberY	-0.45	0.71		1.61	0.34		1.15	0.48		1.81	0.28	
anti_foulingY	-3.09	0.08	.	-2.16	0.19		-2.40	0.20		-3.05	0.07	.
japanY	6.49	0.01	*	1.77	0.56		1.09	0.73		0.51	0.87	
chinaY	2.46	0.39		-0.67	0.83		-1.05	0.73		-1.64	0.61	
southkoreaY	4.17	0.12		1.23	0.69		0.04	0.99		0.71	0.83	
bvY	4.74	0.00	***	5.58	0.00	**	5.97	0.00	***	6.06	0.00	***
ccsY	2.10	0.10	.	2.07	0.24		2.36	0.16		1.79	0.33	
nkkY	3.27	0.00	**	3.96	0.00	**	4.33	0.00	***	4.39	0.00	***
rest_classY	4.30	0.00	***	5.25	0.00	***	4.99	0.00	***	5.37	0.00	***
Smooth-term	edf	p-val.	sig.	edf	p-val.	sig.	edf	p-val.	sig.	edf	p-val.	sig.
s(age)	9.00	< 2e-16	***	1.00	< 2e-16	***	1.00	< 2e-16	***	1.00	< 2e-16	***
s(tc)	4.66	0.00	***	5.25	0.30							
s(norm_tc)							4.583	0.34		1.000	0.82	
s(dwt)	7.030	0.00	***	1.000	0.50		2.85	0.07	.	1.00	0.76	
s(libor)	1.00	0.00	***	6.72	0.00	***	1.91	0.14		5.93	0.00	**
s(orderbook_all)	8.70	0.00	***				5.34	0.01	**			
s(orderbook_type)				1.00	0.79					1.00	0.89	
s(vol_per_dwt)	7.78	0.00	***	6.289	0.02	*	8.82	0.00	**	5.871	0.02	*
s(norm_admiralty_constant)	4.382	0.00	***	1.00	0.16		1.000	0.60		1.00	0.15	
Result	R-sq.	GCV	n	R-sq.	GCV	n	R-sq.	GCV	n	R-sq.	GCV	n
	0.964	9.3267	78	0.873	18.84	78	0.889	17.673	78	0.852	19.678	78

Figure 7.10: Comparison of Initial Models for Capesize Vessels.

For capesize, the back-fitting process ends at the fourth iteration for combination 1 and 2; and at 7th iteration for combination 3. The final result is summarized in figure 7.11. The intermediate results are presented in appendix F subsection F.2.5. Among them, **combination 1 has the highest R-squared value**.

Iteration	Fourth								
Model	Capesize1 (Final)			Capesize2 (Final)			Capesize3		
Parametric	Est.	p-val.	sig.	Est.	p-val.	sig.	Est.	p-val.	sig.
(Intercept)	19.48	<2e-16	***	17.70	0.00	***	18.26	0.00	***
scrubberY							2.68	0.06	.
anti_foulingY							-4.58	0.00	**
japanY									
chinaY									
southkoreaY									
bvY	5.34	0.01	**	6.70	0.00	**	6.15	0.01	*
ccsY									
nkkY	3.19	0.05	*	5.20	0.01	**	4.40	0.03	*
rest_classY	3.42	0.02	*	5.38	0.00	**	5.00	0.01	**
Smooth-term	edf	p-val.	sig.	edf	p-val.	sig.	edf	p-val.	sig.
s(age)	1.00	< 2e-16	***	1.00	0.00	***	1.00	< 2e-16	***
s(tc)	5.80	0.00	***						
s(norm_tc)									
s(dwt)	5.031	0.00	***						
s(libor)	1.704	0.00	***	7.03	0.00	***	2.869	0.00	***
s(orderbook_all)	5.063	0.00	***				5.476	0.012	*
s(orderbook_type)									
s(vol_per_dwt)	7.86	0.00	***	5.047	0.01	**	4.988	0.01	*
s(norm_admiralty_constant)									
Result	R-sq.	GCV	n	R-sq.	GCV	n	R-sq.	GCV	n
	0.924	11.547	78	0.860	16.753	78	0.868	16.622	78

Figure 7.11: Final Results After Backward Elimination - Capesize Vessels.

7.4. GAM Results

For each type vessel type, the "best" model is selected and the result is given in figure 7.12. The two indicators are considered, namely R-squared value GCV-value. Afterwards, *the smooth-terms plot will be compared with the single test to detect the multicollinearity in model. When the multicollinearity is found, variable will be regarded as parametric instead.* Afterwards, using MATLAB curve-fitting toolbox, fitting can be obtained. Used functions are given in table F.10 in appendix F section F.3.

Final Model	All			Handysize			Handymax			Panamax			Capesize		
Parametric	Estimate	p-value		Est.	p-val.	sig.	Est.	p-val.	sig.	Est.	p-val.	sig.	Est.	p-val.	sig.
(Intercept)	7.52	< 2e-16	***	7.45	< 2e-16	***	7.41	< 2e-16	***	10.99	< 2e-16	***	19.48	<2e-16	***
ice_classY															
scrubberY	2.13	0.00	***							2.29	0.01	**			
anti_foulingY															
fuel_typeMDO															
japanY	2.07	0.00	***				2.12	0.00	***	1.26	0.00	**			
chinaY															
southkoreaY	2.88	0.00	***	4.33	0.00	***									
rest_countryY	1.38	0.00	**				1.17	0.05	*						
bvY	1.72	0.00	***				1.58	0.00	***				5.34	0.01	**
ccsY	2.10	0.00	***				1.83	0.00	***						
nkkY	1.72	< 2e-16	***				1.82	0.00	***				3.19	0.05	*
rest_classY	1.98	< 2e-16	***				2.19	0.00	***				3.42	0.02	*
Smooth-term	edf	p-val.	sig.	edf	p-value		edf	p-val.	sig.	edf	p-val.	sig.	edf	p-val.	sig.
s(age)	7.05	< 2e-16	***	6.03	< 2e-16	***	1.87	< 2e-16	***	7.27	0.00	***	1.00	< 2e-16	***
s(tc)				5.09	0.00	***							5.80	0.00	***
s(norm_tc)	4.04	0.00	***				1.000	0.004	**						
s(dwt)	8.42	< 2e-16	***	1.00	< 2e-16	***	2.867	0.00	***	3.51	0.00	***	5.031	0.00	***
s(libor)	2.01	0.00	***				1.282	0.01	*				1.704	0.00	***
s(orderbook_all)	1.00	0.00	**										5.063	0.00	***
s(orderbook_type)							1.00	0.032	*						
s(norm_fuel_consumption)															
s(vol_per_dwt)	8.58	0.00	***							8.22	0.00	***	7.86	0.00	***
s(norm_admiralty_constant)															
Result	R-sq.	GCV	n	R-sq.	GCV	n	R-sq.	GCV	n	R-sq.	GCV	n	R-sq.	GCV	n
	0.769	11.883	1172	0.648	8.8659	355	0.684	11.524	438	0.747	12.561	328	0.924	11.547	78

Figure 7.12: Models after Back-fitting

Comparing Smooth-terms of Integrated Model with Single Test

All Vessels

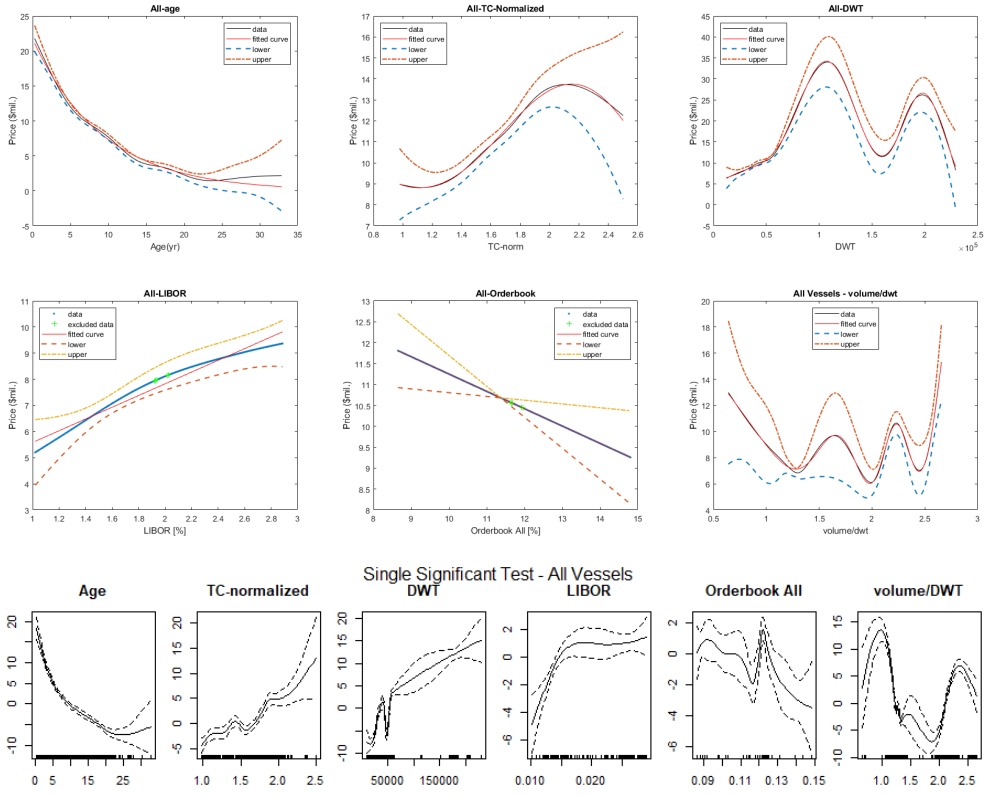


Figure 7.13: Smooth Terms for All Vessels - compared with Single Test.

One can observe from figure 7.13 that among all variables, *DWT, Orderbook and vol/DWT* have quite different plots compare to the single test. This difference indicates strong multicollinearity, thus they will be

treated as parametric. on contrary, *age, LIBOR and TC-norm* have comparable plots. Just as expected, older vessel costs less. High LIBOR indicates high market and higher secondhand demand, thus the price increases. Lastly, secondhand market is the 'substitute' to newbuilding. Thus when there are less newbuilding demand (low orderbook%), buyers turn to secondhand (higher secondhand demand, and sales price).

Handysize Vessels

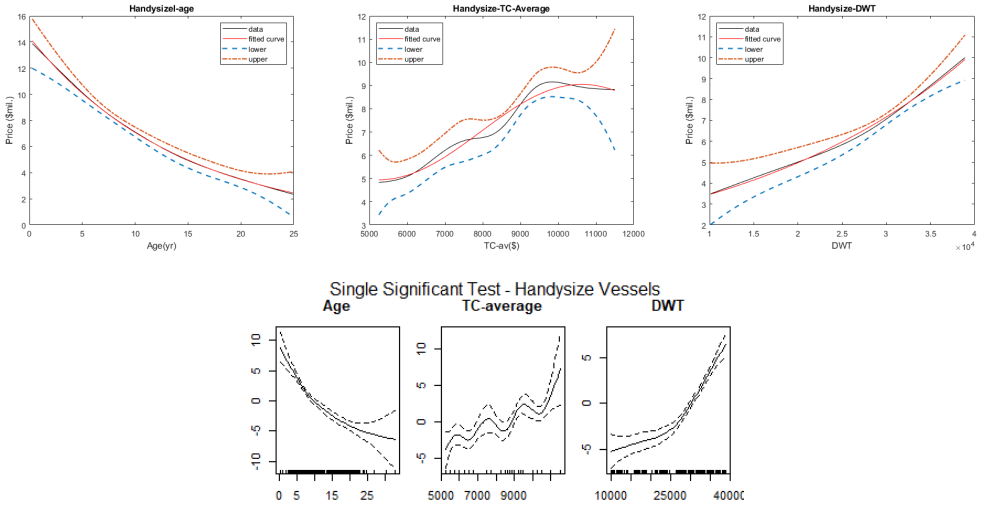


Figure 7.14: Smooth Terms for Handysize Vessels.

One can observe from figure 7.14 that no multicollinearity occurs since plots of entire variables are comparable to the single test. *Age has a declining trend while TC-rate and DWT have a positive trend.* Younger and/or bigger vessel would cost more. Also, the expected income (TC-rate) increases along with the price.

Handymax Vessels

One can observe from figure 7.15 that although both has negative trend, they orderbook smooth term is quite different than the single test. Strong collinearity appears in the first half of the plot; thus it is decided to regard orderbook as parametric at the end. *Age has a declining trend while average TC-rate and DWT have a positive trend.* The similar reasoning as those of handysize. Furthermore, LIBOR is the market status indicator. High LIBOR indicates high market and higher secondhand demand, thus the price increases.

Panamax Vessels

From figure 7.16, one can observe that all smooth term plots are following the single test. Although the $\frac{vol}{dwt}$ plot is initially not exactly the same, they follow the same trend. The difference indicates some degrees of (acceptable) multicollinearity. Just like other vessel types, *age has a negative effect on price.* The DWT has first positive at the first half and follows by negative trends. This anomaly (decreasing trend) might occur due to the sub-market between panamax and capesize vessels which are slightly less popular; thus the price is lower. Another possible factor is the low data density regarding this sub-market. Also, $\frac{vol}{dwt}$ has negative trend in the beginning where the data density is low. This peak corresponds to the majority of panamax vessels and the rest to the sub-markets in between which are less popular thus has lower price.

Capesize Vessels

Lastly, one can observe from figure 7.17 that all smooth-terms variables except the TC-average have comparable plot to the single test. TC-average has a weak increasing trend in the single test while it has a negative trend in the integrated model. This indicates unacceptable collinearity, thus TC-rate will be consider as parametric. Just as in previous *age has a declining trend.* On contrary, other variables have a quite different trend

in comparison to other vessel types. One has to bear in mind that capesize is a quite different market compared to other types because the newbuilding cost is very high and there are much bigger risk related to the income. Secondhand capesize is also more difficult to be sold and only big players can afford them. This explains the trends of TC-average, Orderbook and DWT which are different than other size-based types.

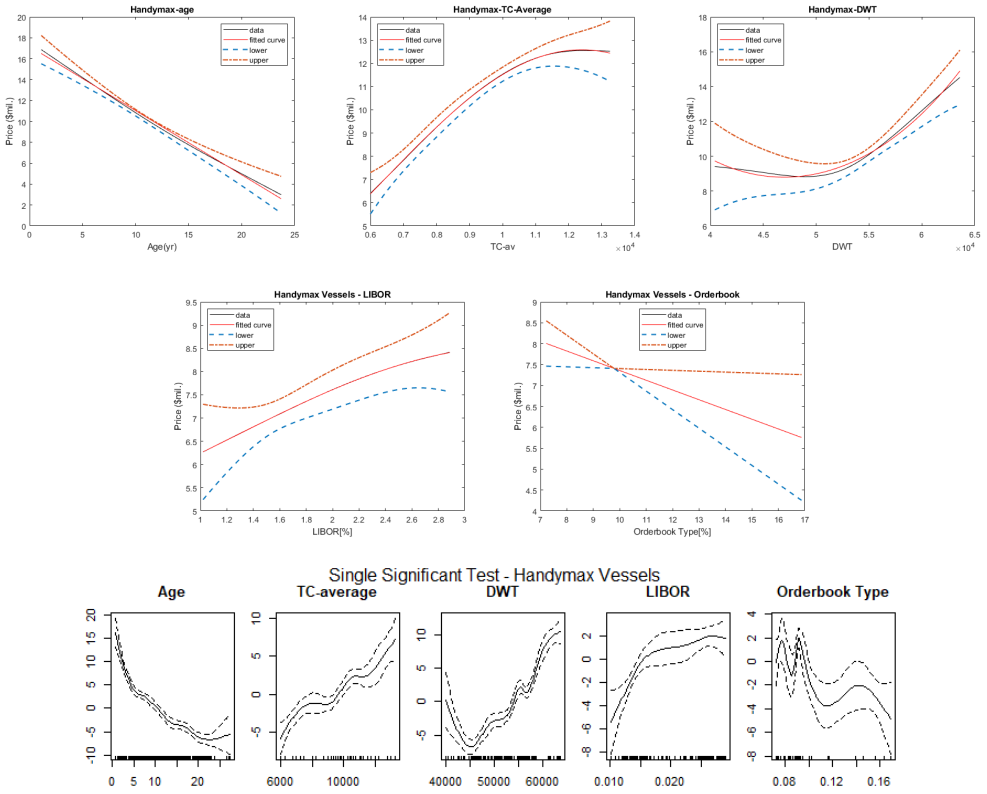


Figure 7.15: Smooth Terms for Handymax Vessels.

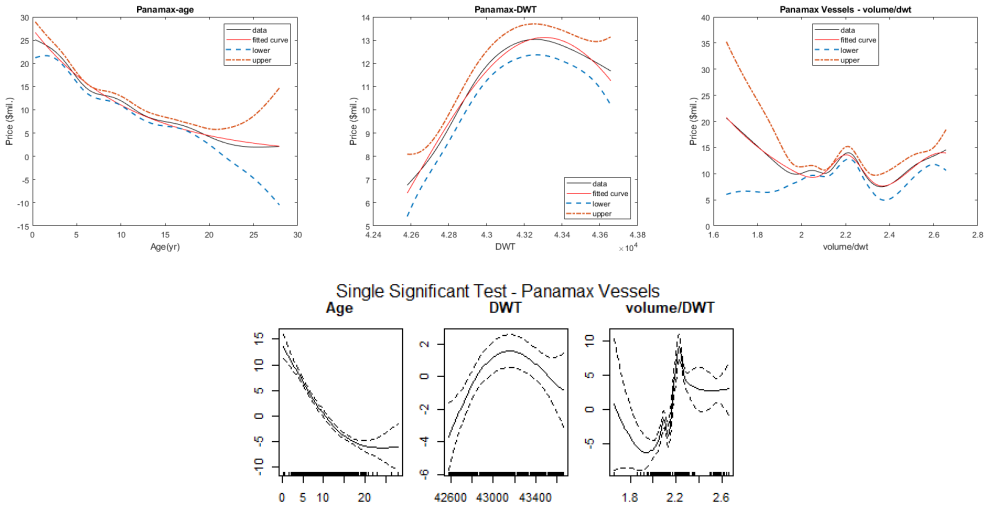


Figure 7.16: Smooth Terms for Panamax Vessels.

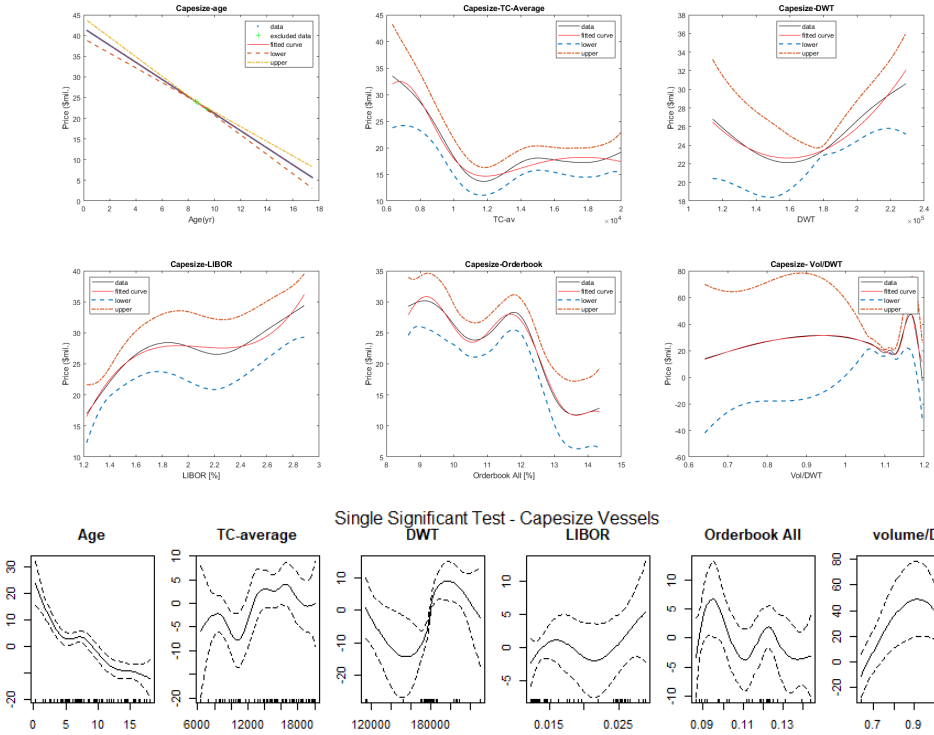


Figure 7.17: Smooth Terms for Capesize Vessels.

Final Result

Final Model	All			Handysize			Handymax			Panamax			Capesize		
Parametric	Estimate	p-value					Estimate	p-value		Estimate	p-value		Estimate	p-value	
(Intercept)	6.00	0.00	***	7.45	< 2e-16	***	9.23	< 2e-16	***	10.99	< 2e-16	***	19.48	<2e-16	***
dwt	0.00	< 2e-16	***												
orderbook	-51.81	0.00	**				-23.28	0.03	*						
vol_per_dwt	2.28	0.00	***												
tc													5.80	0.00	***
ice_classY															
scrubberY	2.86	0.00	***							2.29	0.01	**			
anti_foulingY															
fuel_typeMDO															
japanY	1.85	0.00	***				2.12	0.00	***	1.26	0.00	**			
chinaY							1.17	0.05	*						
southkoreaY	2.47	0.00	***	4.33	0.00	***	2.03	0.00	***						
rest_countryY	1.44	0.00	**				1.17	0.05	*						
bvY	1.37	0.00	**				2.03	0.00	***				5.34	0.01	**
ccsY	1.61	0.00	***				2.28	0.00	***						
nkkY	1.41	0.00	***				2.28	0.00	***				3.19	0.05	*
rest_classY	1.61	0.00	***				2.64	0.00	***				3.42	0.02	*
Smooth-term	edf	p-val.	sig.	edf	p-val.	sig.	edf	p-val.	sig.	edf	p-val.	sig.	edf	p-val.	sig.
s(age)	6.65	< 2e-16	***	6.03	< 2e-16	***	1.87	< 2e-16	***	7.27	0.00	***	1.00	< 2e-16	***
s(tc)				5.09	0.00	***									
s(norm_tc)	4.19	0.00	***				1.00	0.00	**						
s(dwt)				1.00	< 2e-16	***	2.87	0.00	***	3.51	0.00	***	5.03	0.00	***
s(libor)	2.54	0.00	***				1.28	0.01	*				1.70	0.00	***
s(orderbook_all)													5.06	0.00	***
s(orderbook_type)															
s(norm_fuel_consumption)															
s(vol_per_dwt)										8.22	0.00	***	7.86	0.00	***
s(norm_admiralty_constant)															
Result	R-sq.	GCV	n	R-sq.	GCV	n	R-sq.	GCV	n	R-sq.	GCV	n	R-sq.	GCV	n
	0.75	12.457	1172	0.648	8.8659	355	0.684	11.524	438	0.747	12.561	328	0.924	11.547	78

Figure 7.18: Final Models for All Vessel Types.

After judging the collinearity by comparing the smooth terms with single test result; the highly correlated variables are eliminated by treating them as parametric terms. These models are simulated again and the final result is presented in figure 7.18. The next step is to obtain the explicit mathematical equations for each smooth terms. Each vessel types will be discussed separately. Expected value for all vessels is represented in

equation 7.1. For handysize vessels, secondhand price is represented in equation 7.2. Furthermore, expected value for handymax vessels is outlined in equation 7.3. For panamax vessels, expected secondhand price is given in equation 7.4. Lastly, Expected value for capesize can be mathematically written as in equation 7.5.

All Vessels

$$g(E(SH))_{all} = \beta_1 + s(age)_1 + s(tc_{norm})_1 + s(libor)_1 + s(vol/dwt)_1 + \beta_{1a} \cdot japan + \beta_{1b} \cdot china + \beta_{1c} \cdot korea + \beta_{1d} \cdot country\_rest + \beta_{1e} \cdot BV + \beta_{1f} \cdot NKK + \beta_{1g} \cdot CCS + \beta_{1i} \cdot class\_rest + \beta_{1j} \cdot scrubber + \beta_{1k} \cdot dwt + \beta_{1l} \cdot orderbook_{all} \tag{7.1}$$

All	*R-sq.*	*Equation*
s(age)	0.992	$21.7 \cdot \exp(-0.11 \cdot x)$
s(tc_norm)	0.995	$7.9 + 0.4 \cdot cos(x \cdot 3.6) + 1.5 \cdot sin(x \cdot 3.6)$
s(libor)	1	$2.24 \cdot x + 3.34$

Table 7.3: Mathematical Equations for All smooth-terms.

Handysize Vessels

$$g(E(SH))_{handysize} = \beta_2 + s(age)_2 + s(tc_{av})_2 + s(dwt)_2 + \beta_{2a} \cdot korea \tag{7.2}$$

Handysize	*R-sq.*	*Equation*
s(age)	1	$14.4 \cdot \exp(-0.07 \cdot x)$
s(tc_norm)	0.95	$7.8 - 1.15 \cdot \cos(x \cdot 8.8e-4) + 1.07 \cdot \sin(x \cdot 8.8e-4)$
s(dwt)	0.997	$2.40 \cdot \exp(3.65e-5 \cdot x)$

Table 7.4: Mathematical Equations for Handysize smooth-terms.

Handymax Vessels

$$g(E(SH))_{handymax} = \beta_3 + s(age)_3 + s(tc_{norm})_3 + s(dwt)_3 + s(libor)_3 + \beta_{3a} \cdot japan + \beta_{3b} \cdot country\_rest + \beta_{3c} \cdot BV + \beta_{3d} \cdot NKK + \beta_{3e} \cdot CCS + \beta_{3f} \cdot class\_rest + \beta_{3g} \cdot orderbook_{type} \tag{7.3}$$

Handymax	*R-sq.*	*Equation*
s(age)	0.998	$-0.61 \cdot x + 17.17$
s(tc_av)	1	$7.02 - 5.51 \cdot \cos(2.64e-4 \cdot x) - 0.72 \cdot \sin(2.64e-4 \cdot x)$
s(dwt)	0.992	$50.6 \cdot exp(-4.9e-5 \cdot x) + 2.9e-3 \cdot \exp(1.39e-4 \cdot x)$
s(libor)	0.990	$1.2 \cdot x + 5.2$

Table 7.5: Mathematical Equations for Handymax smooth-terms.

Panamax Vessels

$$g(E(SH))_{panamax} = \beta_4 + s(age)_4 + s(dwt)_4 + s(vol/dwt)_4 + \beta_{4a} \cdot scrubber + \beta_{4b} \cdot japan \tag{7.4}$$

Panamax	*R-sq.*	*Equation*
s(age)	0.987	$27.4 \cdot exp(-0.09 \cdot x)$
s(dwt)	0.992	$6.9 - 5.2 \cdot cos(x \cdot 2.3e-3) - 3.5 \cdot sin(x \cdot 2.3e-3)$
s(vol/dwt)	0.98	$5.3e14 \cdot \exp - \left(\frac{x-26}{5}\right)^2 + 11.8 \cdot \exp - \left(\frac{x+2.7}{0.24}\right)^2 + 7.2 \cdot \exp - \left(\frac{x+2.2}{0.1}\right)^2$

Table 7.6: Mathematical Equations for Panamax smooth-terms.

Capesize Vessels

$$g(E(SH))_{capesize} = \beta_4 + s(age)_5 + s(tc_{av})_5 + s(dwt)_5 + s(libor)_5 + s(orderbook_{all})_5 + s(vol/dwt)_5 + \beta_{5a} \cdot BV + \beta_{5b} \cdot NKK + \beta_{5c} \cdot class\_rest \tag{7.5}$$

Capesize	*R-sq.*	*Equation*
s(age)	1	$-1.89 \cdot x + 36$
s(dwt)	0.965	$2.15 \cdot x^2 + 1.63 \cdot x + 22.92$
s(libor)	0.972	$19.4 \cdot x^3 - 122 \cdot x^2 + 253 \cdot x - 146$
s(orderbook_all)	0.996	$27.7 \cdot \sin(0.28 \cdot x + 5.08) + 3.85 \cdot \sin(2.26 \cdot x - 12.7)$
s(vol/dwt)	0.93	$34.8 \cdot \exp - \left(\frac{x+1.2}{0.02}\right)^2 + 31.7 \cdot \exp - \left(\frac{x+0.9}{0.3}\right)^2$

Table 7.7: Mathematical Equations for Capesize smooth-terms.

7.5. Chapter 7 Discussion and Conclusion

Comparison with Initial Models

Y	Result elements	DWT, age, TC-rate, bunkers, LIBOR, Orders per type	DWT, age, TC-rate, bunkers, LIBOR, Orders totals	DWT, age, earnings, bunkers, LIBOR, Orders per type	DWT, age, earnings, bunkers, LIBOR, Orders totals
Price	Minimum R^2 Region	0.4853	0.4571	0.4038	0.4576
	R^2 All	0.7634	0.7645	0.7058	0.6964
	Deleted Variables	LIBOR, Bunk	Bunk	LIBOR, Bunk	LIBOR, Bunk
	Significance	***	***	***	***
Price/ DWT	Minimum R^2 Region	0.5280	0.5317	0.5018	0.4639
	R^2 All	0.6306	0.6223	0.6372	0.6212
	Deleted Variables	Bunk	Bunk	LIBOR, Bunk	LIBOR, Bunk
	Significance	***	***	***	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Figure 7.19: Initial Models for All Vessels by Pruyn[5].

In this section, the (new) GAM results are compared with the initial results by Pruyn. The result from this chapter is the only one that can be compared directly with initial model. However, this is not applicable for the machine learning approach (in chapter 8) since the results have different forms. The main findings is summarized in table 7.19. From that table, one could see that Pruyn has only assessed all vessels type. Among the four model combinations, the second model has the highest R-squared value (0.7645), thus it is selected as the best model. Furthermore, the sales price can be formulated as equation 7.6.

$$E(SH/DWT)_{all} = c_1 + s(DWT) + s(Age) + s(Orderbook_{type}) + c_2 \cdot TC + c_3 \cdot LIBOR \tag{7.6}$$

Furthermore, an important question to ask is *'what kind of improvement has been brought to the new model in comparison to Pruyn's model?'* The standard way to assess the quality of a model which has been used throughout the whole project is by comparing the R-square value. So Pruyn's model is simulated using the current dataset and the result is presented in figure 7.20. By this, a direct comparison can be done since the same dataset is used. *For all vessels type, the final result indicates an improvement since the R-value for new model (0.75) is higher than the initial model (0.622).*

Furthermore, by considering the number of variables used in this project, one can also considered that improvement has been made. This is because 15 variables are used in new model whereas the initial model only consists of 5 variables. Thus, the result from new model might represent the reality better. The last improvement point is by compacting the smooth-term formulations. The initial formulations is given in figure F.11, whereas the formulations of new models are summarized in table 7.13. Considering all these factors, an improvement has been made from the initial model.

Comparison	All (New)			Puyn (Initial)		
Parametric	Estimate	p-value		Estimate	p-value	
(Intercept)	6.00	0.00	***	5.63E-05	3.16E-05	***
dwt	0.00	< 2e-16	***			
orderbook	-51.81	0.00	**			
vol_per_dwt	2.28	0.00	***			
libor				2.98E-03	6.72E-07	***
tc				1.25E-08	1.49E-09	***
scrubberY	2.86	0.00	***			
fuel_typeMDO						
japanY	1.85	0.00	***			
chinaY						
southkoreaY	2.47	0.00	***			
rest_countryY	1.44	0.00	**			
bvY	1.37	0.00	**			
ccsY	1.61	0.00	***			
nkkY	1.41	0.00	***			
rest_classY	1.61	0.00	***			
Smooth-term	edf	p-val.	sig.	edf	p-val.	sig.
s(age)	6.65	< 2e-16	***	5.69	< 2e-16	***
s(tc)						
s(norm_tc)	4.19	0.00	***			
s(dwt)				8.02	< 2e-16	***
s(libor)	2.54	0.00	***			
s(orderbook_all)						
s(orderbook_type)				1.00	6.25E-04	***
s(norm_fuel_consumption)						
s(vol_per_dwt)						
s(norm_admiralty_constant)						
Result	R-sq.	GCV	n	R-sq.	GCV	n
	0.75	12.457	1172	0.622	6.66E-09	1200

Figure 7.20: Comparison between the Initial and New Models.

Comparison between Smooth-terms

From figure 7.21, a comparison can be made between the smooth-term plots of Pruyn's integrated model with the single test. One can observe that all variables have comparable plots. Orderbook has quite different plot but both are with the same negative trend. This indicates collinearity in acceptable level. The smooth terms formulation for Pruyn's model is presented in table 7.8. Lastly, from figure 7.22, one can compare the smooth-term result between the initial data-set and the new data-set. One can see a comparable trend between Pruyn's and current result for *age and DWT*. Such as expected, age has negative effect on price. For the DWT-plot, the first valley might indicate the handymax and panamax concentrated area; whereas the second valley implies capesize concentrated area.

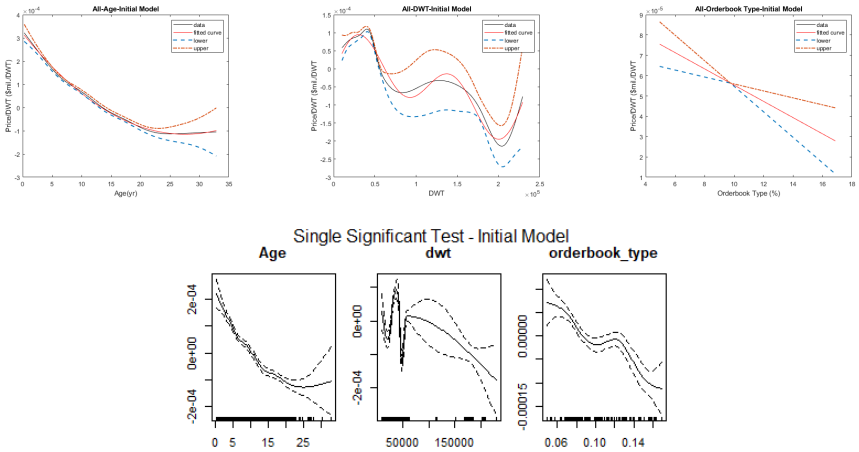


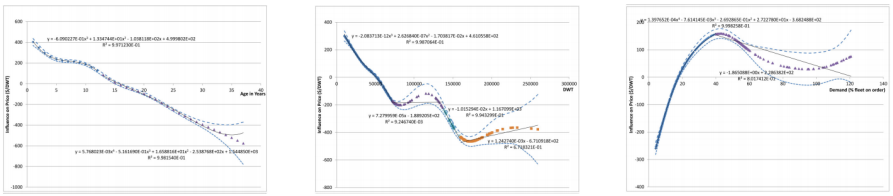
Figure 7.21: Smooth Terms of New Model. In comparison to single test.

On contrary, *the orderbook plot is completely different.* One possible explanation is because the samples were take in different time period. *Thus the different plot characteristic reflects the time-series nature of the orderbook variable.* Pruyn's model encompasses ship sales between June 1998 - December 2010[5]. However, the new database consists of sales between July 2017 - July 2019, a period when world economic is relatively

stable. Between 1998-2010, there was an extreme stagflation in 2008 which marks the beginning of global recession[90]. For this reason, Pruyn's orderbook ranges from 0% to 120%, whereas orderbook of new model ranges from 8% to 15%. In a stable condition, investor might be able to take the rational decision and it result on a linear graph. However, they might be intimidated by the circumstances during the difficult seasons.

Panamax	***R-sq.***	***Equation***
s(age)	0.987	5.7e-7 $\cdot x^2 + -3.2e-5 \cdot x + 3.2e-4$
s(dwt)	0.96	-5e-5 + 9e-6 $\cdot cos(x \cdot 3e-5) + 8e-5 \cdot sin(x \cdot 3e-5) + 2e-5 \cdot cos(2 \cdot x \cdot 3e-5) + 8e-5 \cdot sin(2 \cdot x \cdot 3e-5)$
s(orderbook_type)	0.999	4e-5 $\cdot x + 9.5e-5$

Table 7.8: Mathematical Equations for Initial Model smooth-terms.



((a)) Age ((b)) DWT ((c)) Orderbook Type

Figure 7.22: Smooth Terms from Pruyn's data. Source: Pruyn[5].

GAM Conclusion

Effects of *age, size, income indicator (TC-rate)* & *financial indicator (LIBOR)* on sales price has confirmed what was given by the previous research[36][42][5]. This evidence might suggest the robustness of *structural approach* in modelling the secondhand market, as a comparable conclusion is obtained even when different dataset is used. Based on comparison of the initial and new model; one can conclude that most of the are similar with few differences that can be logically explained. Furthermore, this research discovers the correlation between $\frac{Volume}{DWT}$ *Ratio* and sales price, although the pattern is rather irregular. Lastly, the improvements that has been made in the model has been highlighted in the last part of this chapter.

8
Machine Learning Approach

> The goal is to turn data into information, and information into insight.
>
> *Carly Fiorina*

The machine learning approach is tested in this chapter. According to the algorithm selection which is done in chapter 4, *Random Forests* and *Gradient Boosting Machines (GBM)* are voted as the second and third best candidates. The introduction to data preparation, tuning of both algorithms and model selections are presented in section 8.1. Furthermore, the most essential result for Random Forests is outlined in section 8.2 and for GBM is presented in section 8.3. Lastly, the discussion about the results in comparison with GAM results and the goals of this project are reviewed in 8.4. By this, the last research sub-question; *"What is the final result of applying the suggested improvement points in the current pricing model?"* is answered.

8.1. Machine Learning Implementation

Since the variables are already chosen and the algorithms are already determined, the next steps are *data splitting* and then *model tuning*. Generally, data has to be split into 2 parts, namely *training dataset* and *testing dataset*. The *training dataset* is where model fitting is made whereas the result will be tested in the *testing dataset* to determine model's accuracy[82][54]. The most common splitting ratios to use are 60(training)-40(testing), 70(training)-30(testing) and 80(training)-20(testing).

However, one needs to bear in mind that spending too much on training (above 80%) will make the model to excellently fit the training data, but is not generalizable in other data-sets (*conservative model*). On the other hand, spending too much data in testing (above 40%) will result in generalized model with the risk of poor parameters assessment[64][82]. Thus a balanced proportion has to be chosen. Therefore, the middle proportion value (70-30) is chosen for this project.

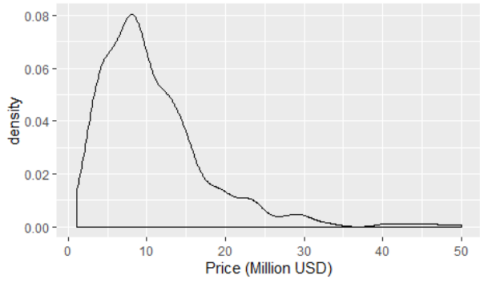


Figure 8.1: Training dataset - Density Function for Price.

There are two common data splitting method, namely *simple random sampling* and *stratified sampling*. In the first method, splitting is done by taking sample randomly. In the second method, the data is first split

into subgroups, then sample is take systematically from each subgroups. The first method is common for regression. The second method is more suitable for classification problem[64]. Thus, in this case, the *simple random sampling* is used. The distribution of ship sales price from the training dataset is presented in figure 8.1. The response variable distribution is right skewed. When the dataset is very large (hundred of thousands of observations), it is common to *normalize* the skewed response variable to speed up the simulation[82]. However it is unnecessary since the data-size is relatively small.

8.1.1. Random Forest Hyperparameters Tuning

For Random Forests analysis, there are two packages available in R, namely *"randomForest package"* and *"ranger package"*[60]. Here, ***ranger package*** is chosen because it performs the simulations in considerably lower time in comparison to *randomForest package*[86]. The number of parameters to tune in Random Forests are less than GBM[60], these are:

- ***ntree:number of trees.*** The goal is to use enough trees to stabilize the error, but not too many because it is inefficient. In this case, Out-Of-Box (OOB) error is used to measure model stability[60]. The result is presented below.

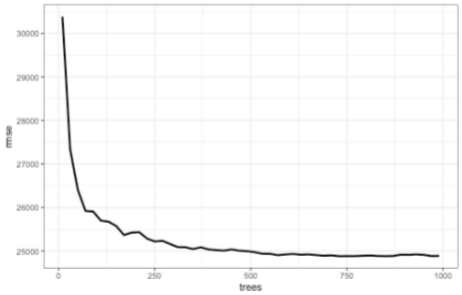


Figure 8.2: Optimum tree numbers - Random Forests.

- ***mtry: the number of variables to randomly sample as candidates at each split.*** When mtry is equal to the number of variables, the model equates to *bagging*, which is is one other machine learning algorithm[64]. When mtry is equal to 1, the split variable is completely random, thus all variables get even chance. However too much randomness might lead to an overly biased result.

- ***splitrule: splitting rule.*** For regression problem, "variance" and "extra trees" rules are usually used. From these one can get the value of accuracy, where the optimum method has the highest accuracy.

- ***min.node.size: minimum number of samples within the terminal nodes.*** This controls the complexity of the trees. Bigger node size allows for deeper, more complex trees and smaller node results in shallower trees. This is another bias-variance trade-off where deeper trees introduce more variance (risk of overfitting) and shallower trees introduce more bias[63].

8.1.2. GBM Hyperparameters Tuning

There are three available GBM package in R, namely (traditional) *"GBM"*, *"Extreme Gradient boosting (XGBoost)"* and *"h20"*[63]. In this case, ***XGBoost*** is used. This algorithm provides a few advantages over traditional boosting such as considerably faster simulation, built-in regularization (to reduce the overfitting) and built-in cross validation feature (which means there is one less parameter to tune)[87]. In comparison with *XGBoost*, *h20* also performs really well. However, *h20* is not chosen because the final function plots will be given in alphabetical order instead of the rank of importance[63]. On the other hand, *XGBoost* presents the plots according to the rank of importance which will make the result more intuitive.

The main challenge of GBM is to tune the parameters due to amount of parameters that can possibly be tuned. On the other hand, it also means that GBM is highly flexible, given that one has an adequate knowledge regarding the tuning. However, finding the optimal combinations are generally time consuming. For more information about how this algorithm works, one can refer back to chapter 4 subsection 4.4.2 of this report. When being tuned appropriately, it will provide a more accurate result in comparison to Random Forests[63]. The main hyperparameters to tune are:

- ***eta: total number of trees.*** In general, GBM requires many trees. However, many tree increases the chance to overfit. Therefore, the tuning goal is to find an optimal number of trees to minimize the loss function (associated with errors) and the number of cross validations[62].
- ***max_depth: depth of trees.*** This indicates the number of splits in each tree. The depth controls the complexity of the boosted ensemble. Increasing tree-depth will make model more complex, accurate and more likely to overfit[62]. The typical value is between 1 and 10 for regression problem. In this cases, since the data-size is relatively small, lower value can be chosen.
- ***min_child_weight: minimum number of observations required in each terminal node.*** Algorithm stops to split once the sample size in a node goes below this specified threshold. Higher values prevent model from learning relations of specific tree and therefore reduce the overfit tendency[63]. However it also means that the model become more conservative. This variable is tuned to control the over-fitting and it has relatively small impact on overall performance.
- ***subsample: percent of training data to (randomly) sample for each tree.*** It controls the fraction of available training observations. Using less than 100% of the training observations means that stochastic gradient descent is implemented and this can minimize overfitting[63]. Subsampling is performed once in every boosting iteration.
- ***colsample_bytrees: percent of columns to sample from for each tree.*** Similar role as *subsample*[63].
- ***alpha: L1 regularization term on weights.*** Increasing this value will make model more conservative[87].

Parameter	Tested Values	Used Vales
num.trees	0-1000	1000
mtry	15, 18, 20	20
splitrule	"variance", "extratrees"	"variance"
min.node.size	63%, 70%, 75%, 80%, 100%	100%

((a)) Random Forests

Parameter	Used Vales
eta	0.5
max_depth	3
min_child_weight	3
subsample	0.8
colsample_bytree	1
alpha	100

((b)) GBM

Figure 8.3: Machine Learning Hyperparameters.

8.1.3. Best Models Selection - Machine Learning

Four combinations presented in table 6.5 is used for each vessel type. Afterwards, the best model for each vessel type is selected based on the lowest RMSE and/or the highest R-squared value[85][88]. In this instance, Random Forests algorithm gives out R-squared value along with RMSE, whereas GBM only provides the RMSE value. However, one can take a general conclusion that low RMSE value corresponds to the high R-squared value[88]. The final result of all models testing is presented in the table in figure 8.4. From best models, one can rank the importance of each variable. Top 3 variables are chosen to be visualized. However, machine learning has an advantage over GAM, namely they are able to generate plots for both numerical and categorical variables whereas GAM only able to plot the numerical[64][63][60].

The most common ways to visualize machine learning results are through the *Partial Dependence Plot (PDP)* and the *Individual Conditional Expectation (ICE)*[63][60][81]. The *Partial Dependence Plot (PDP)* is the

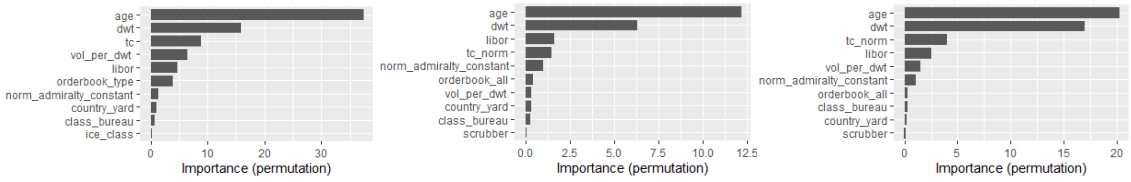
most basic way to present the marginal effect selected variable(s) on overall average of the predicted outcome. PDP is similar with GAM smooth-term plots, except this is given without the confidence intervals. Differently, *Individual Conditional Expectation (ICE)* presents the variable effect on vessel price for *each instance*. Thus, ICE plot consists of multiple lines where each line represent one observation (in this case one vessel)[81].

Model		All				Handysize				Handymax				Panamax				Capesize			
		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
RF	R-sq.	0.72	**0.73**	0.72	0.72	0.61	0.61	**0.61**	0.61	0.65	0.65	**0.69**	0.68	0.66	**0.66**	0.66	0.66	0.70	0.70	**0.71**	0.71
	RMSE	3.70	**3.69**	3.73	3.74	2.70	2.70	**2.70**	2.70	3.32	3.33	**3.12**	3.15	3.97	**3.94**	3.98	3.97	5.18	5.17	**5.09**	5.10
GBM	RMSE	3.26	3.32	**3.22**	3.24	2.80	2.80	2.78	**2.76**	3.21	**3.20**	3.24	3.24	**3.47**	3.49	3.49	3.48	**5.570**	5.583	5.575	5.629

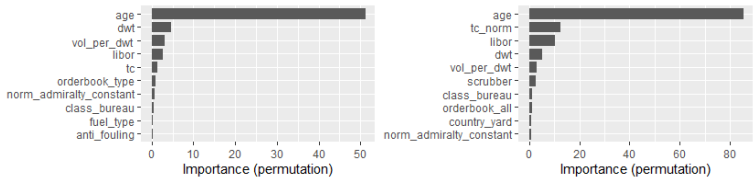
Figure 8.4: Machine Learning - Best Model Selection.

In this case, both plots are presented where ICE plot is given by the black-thin lines where PDP plot is is represented by the red thick line. One can also observe that PDP gives a single line which is the average value of multiple ICE lines. *A comparable plots will be obtained from Random Forest and GBM since both of them are tree-based models.* To conclude, although machine learning has an advantage over GAM in terms of visualization, it also has one weak point, namely its disability to extract data points from the plots. Therefore, it is not possible to obtain the explicit mathematical formulations of the machine learning result. The next section presents the machine learning results for each vessel type.

8.2. Random Forests Results



((a)) All ((b)) Handysize ((c)) Handymax



((d)) Panamax ((e)) Capesize

Figure 8.5: Significance Rating of All Features - Random Forests models.

From Random Forests, one can obtain the ranking of influence of variables used in model. Among all combinations, the best model with highest R-squared and the lowest RMSE value is selected. From each best model, one can obtain the ranking of the variable's effect on vessel price. One can observed from figure 8.5 that age consistently has the highest influence on sales price for every vessel types. Next to vessels' age, other important variables which consistently rank in the Top 3 variables are *DWT, TC-rate (average or normalized)*, and *LIBOR*. This result is in accordance with the findings suggested by Pruyn[5] and others[42][23].

However, since other new variables are suggested in this research, one can observe the importance of these new variables on vessel price. This finding is quite different when compared to GAM because *admiralty constant is proven to be insignificant in all cases when data is modelled using GAM*. On the other hand, by

looking at table 8.18 *GAM models indicate that* $\frac{Volume}{DWT}$ *Ratio is a significant variable for All, Handyize and Panamax vessels types.* This is in agreement with Random Forests result. Moreover, the top-3-features for every vessel types are discussed.

Essential Features for All Vessels

Refer back to subsection 8.1.3, the think red line shows in the plot indicates the average value while the thin black line refers to single observations. One can observe that *age has a negative on price*, while *DWT, TC-average and LIBOR have a positive impact on price.* This is in agreement with GAM single significance test presented in section 7.2.

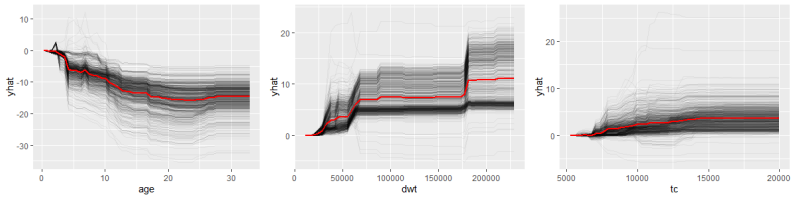


Figure 8.6: Top 3 Features - Random Forests - All Vessels.

Essential Features for Handysize Vessels

From figure 8.7 one can observe that *age has negative influence on sales price*, whereas *DWT, LIBOR and normalized TC-rate have positive influence.* Although not completely the same, this result is in accordance to GAM-result in section 7.4.

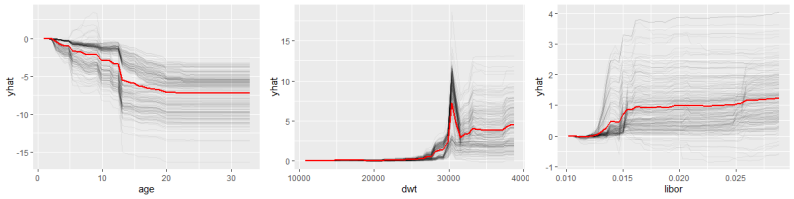


Figure 8.7: Top 3 Features - Random Forests - Handysize Vessels.

Essential Features for Handymax Vessels

One can observe in figure 8.8 that *age has a negative on price*, while *DWT, TC-average and LIBOR have a positive impact on price.* This is in agreement with GAM single significance test presented in section 7.2.

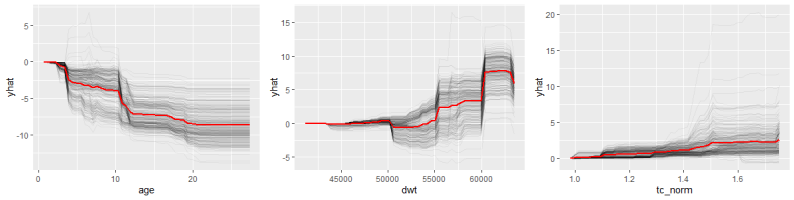


Figure 8.8: Top 3 Features - Random Forests - Handymax Vessels.

Essential Features for Panamax Vessels

From figure 8.8 one can observe in that *age is the only variable with negative effect.* Furthermore *DWT and LIBOR have the same trend,* namely an increase in the beginning and then stabilize. These trends are comparable with handysize and handymax vessels, as well as for GAM which is presented in section 7.4.

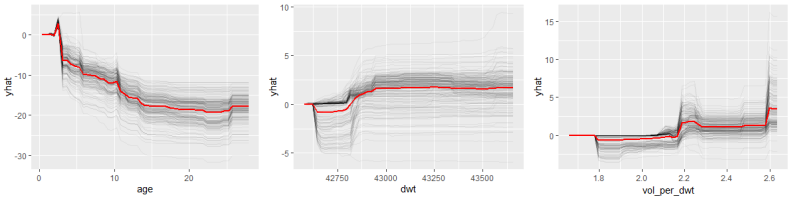


Figure 8.9: Top 3 Features - Random Forests - Panamax Vessels.

Essential Features for Capesize Vessels

One can observe in figure 8.10 that *age, normalized TC-rate and LIBOR have the same trends as Random Forests results for handysize, handymax and panamax vessels.* However, this is different when compared to GAM result in section 7.4. Random Forest result might be more reliable since it has no collinearity issue[54]. To conclude, according to Random Forests results, age, TC and LIBOR has the same trend for the entire types. However, there is slight variations for other variables. The GBM result will be discussed in section 8.3.

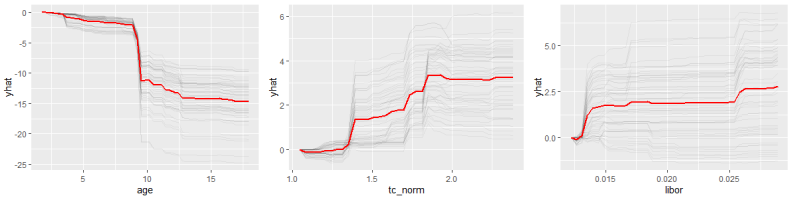
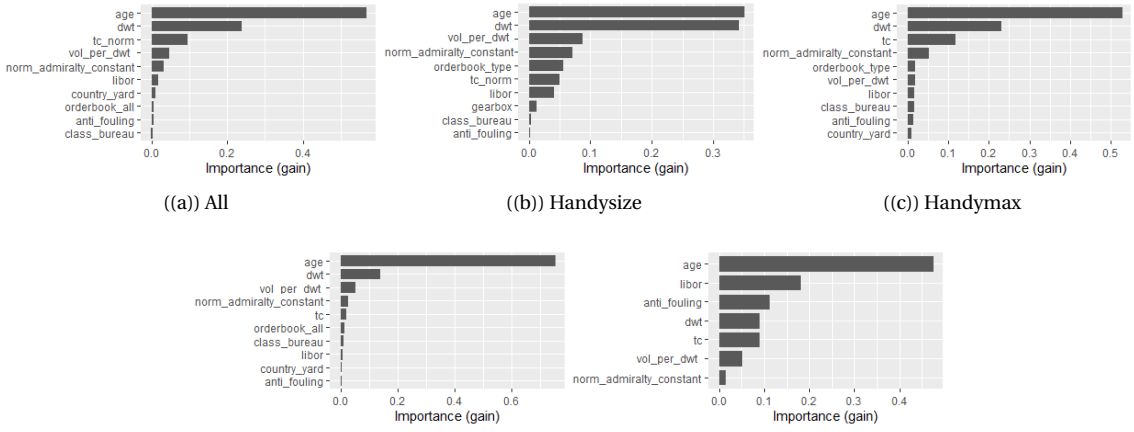


Figure 8.10: Top 3 Features - Random Forests - Capesize Vessels.

8.3. Gradient Boosting Machine Results



((a)) All ((b)) Handysize ((c)) Handymax

((d)) Panamax ((e)) Capesize

Figure 8.11: Significance Rating of All Features - GBM models.

For each vessel type, the ranking of top-10 variable's effect is given in figure 8.11. Reflecting on that, one can conclude age has the most prominent influence in all cases. Furthermore, DWT and TC-rate have consistently a predominant effect on price when being compared to other variables. This finding is quite similar with that of Random Forests. Moreover, normalized admiralty constant and $\frac{Volume}{DWT}$ also gives meaningful contribution in most cases. Lastly, the influence of orderbook is particularly important to handymax while the effect anti-fouling and LIBOR are particular to capesize. This result is quite different than Random Forests since LIBOR has an important effect in all (Random Forests) cases. In the next subsections, the effect of top-3 variables will be individually discussed for each vessel type.

Essential Features for All Vessels

To clarify, one can refer back to subsection 8.1.3 where the red line indicates the average value while the black lines refer to the effect single observations. There are many black-lines in the plots which correspond to the amount of observations. From figure 8.12, one can see that *age, DWT and normalized TC-rate have comparable trends to Random Forest result,* which is presented figure 8.6. This indicates that older and smaller vessels are expected to be cheaper. TC-rate indicates the market status, thus vessels are expected to be more expensive in high market.

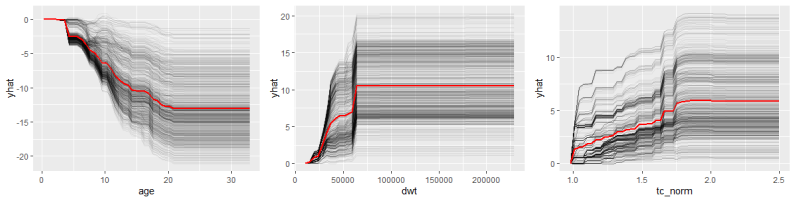


Figure 8.12: Top 3 Features - GBM - All Vessels.

Essential Features for Handysize Vessels

One can observe in figure 8.13 that *age has a negative effect whereas DWT has an increasing trend.* It is because the older and smaller vessels are expected to be cheaper. This finding is comparable with GAM results presented in figure 7.13, 7.15 and 7.17. On the contrary, particular to this case, $\frac{Volume}{DWT}$ *is shown to have a positive influence* with a peak at around 1.2 and 1.5. This might correspond to higher data density.

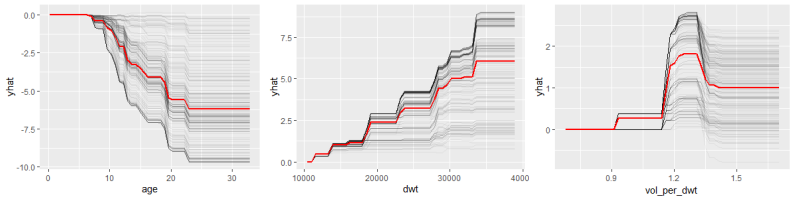


Figure 8.13: Top 3 Features - GBM - Handysize Vessels.

Essential Features for Handymax Vessels

From figure 8.14 one can observe that *age influences price negatively whereas DWT gives a positive influence.* This indicates that older and smaller vessels are expected to be cheaper. These results are similar to handysize results in figure 8.13. Furthermore, *TC-rate also positively influences price* which is similar to its influence for all vessels type (in figure 8.12). TC-rate indicates the market status. Thus vessels are expected to be more expensive in high market.

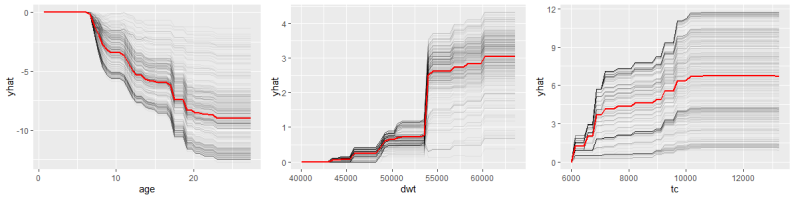


Figure 8.14: Top 3 Features - GBM - Handymax Vessels.

Essential Features for Panamax Vessels

Similar to previous cases, one can observe in figure 8.15 that *age influences price negatively while DWT influences it positively.* It is because older and smaller vessels are expected to be cheaper. In addition, one can also observe a 'ramp function' like increase for the $\frac{Volume}{DWT}$ which is unique to panamax type. This might happen due to lower data density in the beginning until the $\frac{Volume}{DWT}$ is equal to 2.0.

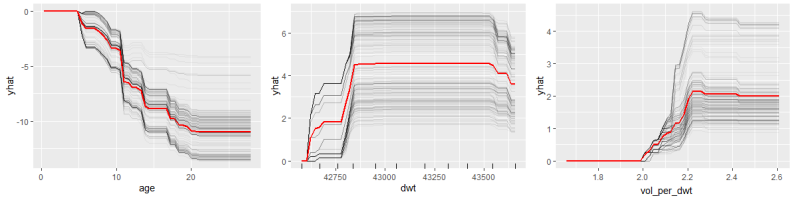


Figure 8.15: Top 3 Features - GBM - Panamax Vessels.

Essential Features for Capesize Vessels

Similar to previous cases, one can see in figure 8.16 that *age has a negative on price.* However, the other variables exhibit a unique behavior. Namely, they give an increasing step-function for LIBOR and decreasing step-function for anti-fouling. Capesize has a unique market characteristics compared to other vessel types; which might cause difference in plot shape. Lastly, it worth to mention that anti-fouling is a categorical variable whose effect has been quantified by GBM algorithm. Thus, the discussion about variables' effect on capesize sales price wraps up the discussion of GBM result. In the next section, the general result of GAM, Random Forests and GBM will be compared.

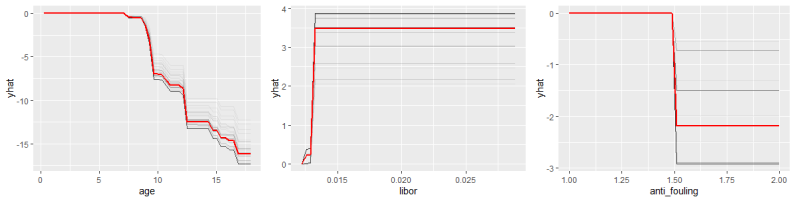


Figure 8.16: Top 3 Features - GBM - Capesize Vessels.

8.4. Chapter 8 Discussion and Conclusion

Fitting Quality

Higher R-value and lower RMSE & GCV indicate a better fitting quality[54]. Considering R-value, one can conclude that GAM scores better than Random Forests in most cases. The difference is normally insignificant except for the Panamax and Capesize where GAM scores 0.1 and 0.2 higher, respectively. This Implies that

GAM generally gives better performance in comparison to Random Forests. By considering the RMSE-value, one can notice that, Random Forests produces lower RMSE for Handysize, Handymax and Capesize vessel types. On contrary, GBM predicts more accurately for All and Panamax vessels. Thus, Random Forests and GBM performances are comparable qua fitting performance.

Model / Type	GAM - initial		GAM - new		Random Forests		GBM
	R-sq.	GCV	R-sq.	GCV	R-sq.	RMSE (mil. $)	RMSE (mil. $)
All	0.62	12.46	0.75	11.88	0.73	3.69	3.22
Handysize	-	-	0.65	8.87	0.61	2.70	2.76
Handymax	-	-	0.68	11.52	0.69	3.12	3.20
Panamax	-	-	0.75	12.56	0.66	3.94	3.47
Capesize	-	-	0.92	11.55	0.71	5.09	5.57

Figure 8.17: Comparison of All Models.

Since GAM scores better than Random Forest and GBM in most cases; on the face-value one might simply conclude that GAM is the best algorithm. However, one has to consider the amount of effort and complexity of building the GAM models. Because GAM is coupled with Backward Elimination where few variables are eliminated during each iteration. Thus many iterations have to be made. In the same manner, considerable effort is required to tune the machine learning parameter in this case. Thus, in this case, building GAM & machine learning model requires comparable effort and they yield comparably good outcomes. Furthermore, employing GAM for this problem is feasible because relatively low number of variables (15) are used.

However, when much more variables are used, performing the Backward Elimination could be very toilsome. It is very common for the machine learning algorithms to be applied in a dataset with few hundred variables. It is proven to work well for such a problem[64][54]. One only has to tune several hyperparameters and perform few more steps in data preparation when machine learning is used. The required effort will be much less than performing Backward Elimination for few hundred variables. Thus, this is the trade-off of using advanced statistics like GAM or tree-based models like Random Forests and GBM. Lastly, one can conclude that GAM, Random Forest and GBM can be used to build secondhand market model. Each of them has their own strengths and weaknesses.

Significant Variables

Furthermore, one can observe the significant variables across all models in the table in figure 8.18.

GAM-Variables

By looking at GAM-result, one can conclude that, Age, DWT TC-rate have significant influence on sales price for most vessel type[6][36][35]. Furthermore, the Cost of Finance (LIBOR) and Orderbook percentage are also significant for All and Capesize types. Previous publication signifies the importance of LIBOR and Orderbook when being tested for the All type[5][42]. However, the research on size-based types had not been done yet. In this study, the same conclusion is obtained for all vessels, although the sample size is different and they are taken from different time-period. This evidence might suggest the validity of Structural approach for modelling the secondhand market.

On contrary to machine learning, GAM also considers the influence of each member within the Classification Society and Country of Yards individually. Thus, one can conclude which specific country or classification society are important. One can observe the table above that country of yard Japan and South Korea have considerable influence on vessels price across most vessel types. Lastly, one can also see that scrubber (according to GAM) has some influence on price for Panamax and anti-fouling indicator (according to GBM) to capesize type. The significance of these (new) variables has not yet been tested in the previous research.

Machine Learning-Variables

From Random Forests part, one can observe that across all vessel types, *age, DWT, LIBOR and TC-rate* are

always the top-5 features. This confirms what is suggested by various literature[5][42]. A comparable conclusion is obtained from GBM models across most of vessel types; where *Age, DWT and TC-rate* are considered as the most important variables. Furthermore, one can also observe another variable who plays quite important roles, namely $\frac{Volume}{DWT}$. This variable is creatively chosen by analyzing the MBG feature. According to Random Forests, this variable is top-5 feature for every vessel types except for handymax. It is also considered to be a top-5 feature for All and Panamax vessels types according to GBM. The exact same results are obtained by GAM, where $\frac{Volume}{DWT}$ is also significant for Panamax and Capesize type.

Model	GAM			Random Forests	GBM
	***	**	*	Top 5 Features	Top 5 Features
Initial	Age, DWT, TC-norm., LIBOR, Orderbook Type,	-	-	-	-
All	Age, DWT, TC-norm., LIBOR Vol/DWT, Japan, Korea, Class Soc. Rest, N.K.K., C.C.S.	B.V., Yard Country Rest, Orderbook All	-	Age, DWT, TC-av., Vol/DWT, LIBOR	Age, DWT, TC-norm., Vol/DWT, Norm. Ca
Handysize	Age, DWT, TC-av., Korea	-	-	Age, DWT, LIBOR, TC-norm., Norm. Ca	Age, DWT, Vol/DWT, Norm. Ca, Orderbook Type
Handymax	Age, DWT, Korea, Japan, B.V. C.S.S., N.K.K., Class Soc. Rest,	TC-norm	LIBOR, Orderbook Type, China, Country Rest	Age, DWT, TC-norm., LIBOR, Vol/DWT	Age, DWT, TC-av., Norm. Ca, Orderbook Type
Panamax	Age, Vol/DWT, DWT	Japan, Scrubber	-	Age, Vol/DWT DWT, LIBOR, TC-av.	Age, DWT, Vol/DWT Norm. Ca, TC-av.
Capesize	Age, TC-av., DWT, LIBOR, Vol/DWT, Orderbook All,	B.V.	N.K.K., Class Soc. Rest	Age, TC-norm., LIBOR, DWT, Vol/DWT	Age, LIBOR, Anti-Fouling DWT, TC-av.

TC-norm. = TC-normalized |TC-av. = TC-average | Norm. Ca = Normalized Admiralty Constant| C.C.S.= China Corporation Register| Class Soc. = Class Society | N.K.K. = Nippon Kaiji Kyokai | B.V. = Bureau Veritas

Figure 8.18: Important Features of All Models.

Moreover, one can observe that according to GBM result, normalized Admiralty Constant (Ca) is ranked as an important (top-5) variable for every vessel types except for capesize. This variable represents the speed's influence, and this outcome confirms what is suggested by Chen et al[70]. In his research, they suggested that speed influences the operational expenses, thus it might influence the sales price. Furthermore, one can also see that Ca is ranked as important for handysize sales price according to Random Forests. On the contrary, *Normalized Ca* is not considered to be important according to GAM results. This is because Norm. Ca is among the first to be eliminated during the Backward Elimination.

The main reason of eliminating Normalized Ca early during the GAM iteration is because its lowest score during the PCA test which is done prior to the modelling (see table 6.4). Hastie et al suggest that PCA has a main weakness namely its inability to consider multivariate (higher order) interactions between variables since it treats all variables as linear[77]. This might explain why Ca has scored the lowest during this analysis. On the contrary, Hastie et al also suggest that Random Forests and GBM have an advantage over GAM, namely their (built-in) ability to assess the multivariate interactions between variables[77].The higher order effects of Ca might be the reason why it is important according to machine learning result. This also indirectly imply the advantage of machine learning above statistical methods.

Project Goals

Lastly, by reflecting the initial improvement goals which are presented in chapter 3 in subsection 3.3, one can conclude that GAM is the only algorithm that can achieve all of these goals. This is because the machine learning algorithm cannot give the explicit mathematical expressions which are desirable in this project. By this, the last subquestion, *"What is the final result of applying the suggested improvement points in the current pricing model?"*, is answered. Also, the main research question, ' ***What is the most practical way to improve the current secondhand bulk carriers pricing model used in Maritime Business Game based on the available data and what is the implementation result?***', that is proposed in the beginning is answered.

9
Conclusion and Recommendation

千里之行始于足下.

Chinese Proverbs

9.1. Conclusion

In this project, an improved pricing model has been proposed. To do this, several steps have been taken. First, the theory of secondhand shipping market are discussed in chapter 2. The relationships between four shipping market segments were explained. Moreover, the bulk market analysis and various methods for secondhand vessels valuation are discussed in chapter 3. Here, the scientific valuation method is used, specifically the structural modelling approach. This method is considered to be more reliable than time-series approach. The initial model uses five variables, namely ***age***, ***DWT***, ***TC-rate***, ***LIBOR*** and ***orderbook percentage***. They are re-used in new models. Afterwards, three improvement points are identified, namely:

1. Evaluating the effect of $\frac{Volume}{DWT}$ & features which are typical for smaller vessels like ***ice class*** & ***crane***.
2. Integrating relevant factors such as ***builder reputation***, ***efficiency indicator*** & ***sustainability indicator***.
3. Compacting the mathematical expressions for the smooth terms.

In chapter 4, various data mining algorithms are assessed using ***Analytic Hierarchy Process***. Among the algorithms, ***Generalized Additive Models (GAM)***, ***Gradient Boosting Machines (GBM)*** and ***Random Forests*** are voted as the most suitable algorithms. In chapter 5, the variable selection is elaborated. Initially, 19 variables are considered based on the literature review. In chapter 6, two statistical tests, ***collinearity tests*** and ***principal component analysis***, are conducted. Consequently, two variables are eliminated to prevent collinearity. Information is extracted from 1227 sales from July 2016 to July 2019. They are differentiated based on their size. In addition, four variable-based combinations are set up and summarized in table 6.5.

The results obtained from this research generally validate the results of previous research. ***Age, DWT, TC-rate & LIBOR*** are among the most influential variables concerning the vessels price. The results have been consistent across five different vessel types and three different algorithms. A comparable conclusion has been obtained by Pruyn, regardless of the different sample size sampling period. This evidence might suggest the validity of structural approach for modelling the secondhand market.

In addition, the importance of two variables has been discovered. These are $\frac{Volume}{DWT}$ ***& Normalized Admiralty Constant (Ca)***. According to Random Forests outcomes, $\frac{Volume}{DWT}$ is top-5 important for the entire types except for handymax. Furthermore, GBM results suggest that normalized Ca is a top-5 variable for the entire types except for Capesize. Lastly, among all variables, ***age and DWT*** have the most influence on price.

Likewise, this study discovers several 'occasionally important' variables. These variables are specific ***Country of Yards*** (especially Japan and South Korea) & particular ***Classification Society***. Afterwards, ***Scrubber*** is influential for Panamax type according GAM. Additionally, ***anti-fouling*** feature is considered to be influential for Capesize according to GBM. Scrubber and Anti-fouling indicator are suggested as part of 'Sustainability Indicator' following the literature review. Their significance on secondhand vessel price have not yet been investigated before.

Regarding the applicability of data mining algorithms, one can conclude that two machine learning algorithms, Random Forests and GBM, can be used to decently model the secondhand shipping market. Their performance has been tested again GAM, a proven algorithm. When the fitting quality is judged, one can conclude that the performance of three algorithms are comparably good. However, GAM scores slightly higher in most cases compared to the other two. However, GAM can be used in more limited context in comparison to machine learning; namely when the number of variables used in the model is low. Otherwise, it will be very toilsome to perform the Backward Elimination as the number of variables increases.

To conclude, considering the strengths and weaknesses of algorithms along with the improvement goals; all algorithms are useful, but GAM is the most applicable one for this project. By this, the main research question, ' ***What is the most practical way to improve the current secondhand bulk carriers pricing model used in Maritime Business Game based on the available data and what is the implementation result?***', that is proposed in the beginning of the project is answered

9.2. Recommendation

The further research can be done by considering following points of improvement:

- Examining the effect of other variables which potentially affect the price. This can be suggested by literature study or creative consideration based model analysis.
- Exploring other smoothing techniques such as *loess smoother, kernel smoother*. An also exploring other regression splines alternatives such as *smoothing splines* or *natural splines* along with other elimination procedure (such as *Forward Elmininination*) when GAM is used.
- More extensive tuning to find more optimum hyperparameters when GBM or Random Forests algorithms are used.
- Exploring other potential branches of machine learning for predictive analytic, such as *Naïve Bayes*, *k-nearest Neighbors* or *Regularized Regression*.

Bibliography

[1] Salomons, B. (2016).*The Dutch East India Company was richer than Apple, Google and Facebook combined.* Available from: <https://dutchreview.com/culture/history/how-rich-was-the-dutch-east-india-company/>. Accessed at 20 February 2019.

[2] Hattendorf, J. B. (2007). *Oxford Encyclopedia of Maritime History, volume 1, introduction.* Oxford: Oxford University Press.

[3] International Chamber of Shipping. (2018). *Shipping and World Trade.* Available from: <http://www.ics-shipping.org/shipping-facts/shipping-and-world-trade>. Accessed at 20 February 2019.

[4] Stopford, M. (2009). *Maritime economics 3e.* Routledge.

[5] Pruyn, J. (2013). *Shipping and Shipbuilding Scenario Evaluations through Integration of Maritime and Macroeconomic Models.* Delft: TU Delft doctoral dissertation.

[6] Pruyn, J., Van de Voorde, E., and Meersman, H. (2011). *Second hand vessel value estimation in maritime economics: A review of the past 20 years and the proposal of an elementary method.* Springer Journal, 13(2), pg:213-236.

[7] Clarksons Research. (2019). *Shipping Intelligence Network & World Fleet Register.* Available from: <https://www.clarksons.net/portal/>. Last Accessed at 6 August 2019.

[8] United Nations Conference on Trade and Development. (2018). *Review of Maritime Transport 2018.* New York: United Nations Publications.

[9] Investopedia. (2019). *Law of Supply and Demand.* Available from: <https://www.investopedia.com/terms/l/law-of-supply-demand.asp>. Accessed at 7 May 2019.

[10] World Bank. (2019). *GDP timeline.* Available from: <https://databank.worldbank.org/data/source/world-development-indicators>. Accessed at 10 May 2019.

[11] Investopedia. (2019). *Market & Economy.* Available from: <https://www.investopedia.com/terms/m/market.asp>. Accessed at 11 May 2019.

[12] study.com Business Course. (2003). *Introduction to Business: Heteregenous Product.* Available from <https://study.com/academy/lesson/heterogeneous-products-definition-lesson-quiz.html>. Accessed at 15 May 2019.

[13] Chron. (2018). *What Are the Characteristics of a Competitive Market's Structure?* Available from <https://smallbusiness.chron.com/characteristics-competitive-markets-structure-23832.html>. Accessed at 15 May 2019.

[14] Karatzas, B. M. (2009). *What's in the Value of a Vessel?* Tanker Operator Magazine.

[15] International Valuation Standards Council. (July 2016).*VS 105: VALUATION APPROACHES AND METHODS.* Available from: <https://www.ivsc.org/files/file/view/id/648>. Accessed at 15 May 2019.

[16] LTAV Association for Promotion of Long Term Asset Value Method. (2009). *Hamburg Ship Evaluation Standard.* Available from: <http://www.long-term-asset-value.de/>. Accessed at 15 May 2019.

[17] Pruyn, J. (2016). *"Will the Northern Sea Route ever be a viable alternative?."* Journal of Maritime Policy & Management, 43(6), pg:661-675.

[18] Pruyn, J. (2009). *"A different approach to modelling Maritime transport Demand."* Conference Paper presented at EWGT, At Padova.

[19] Evans, J. J. (1994). *An analysis of efficiency of the bulk shipping markets.* Maritime Policy and Management, 21(4), 311-329.

[20] Engelen, Steve, et al. (2005) *"A business game for the maritime sector."* Proceedings of the 12th European Concurrent Engineering Conference.

[21] Beenstock, M. (1985). *A theory of Ship Prices.* Maritime Policy Management Journal, 12, pg: 215-225.

[22] Beenstok, M. & Vengotis, A. (1989). *An econometric Model of the World Shipping Market for Dry Cargo, Freight and Shipping.* Applied Economics Journal, 21(3), pg: 339-356.

[23] Haralambides, H. E., Tsolakis, S. D., & Cridland, C. (2005). *Econometric modelling of newbuilding and secondhand ship prices.* Research in Transportation Economics, 12, pg: 65-105.

[24] Hale, C., & Vanags, A. (1992). *The market for second-hand ships: some results on efficiency using cointegration.* Maritime Policy & Management, 19(1), pg: 31-39.

[25] Glen, D. R. (1997). *The market for second-hand ships: Further results on efficiency using cointegration analysis.* Maritime Policy and Management, 24(3), pg: 245-260.

[26] Kavussanos, M. G. & Alizadeh, A. (2002). *Efficient pricing of ships in the dry bulk sector of the shipping industry.* Journal of Transport, Economic & Policy, 29(3), pg: 303-330.

[27] Tsolakis, S. D., Cridland, C., & Haralambides, H. E. (2003). *Econometric modelling of second-hand ship prices.* Maritime Economics & Logistics, 5(4), pg: 347-377.

[28] Kavussanos, M. G. & Alizadeh, A. (2002). *The Expectations Hypothesis of the Term Structure and Risk Premiums in Dry Bulk Shipping Freight Markets.* Journal of Transport, Economic & Policy, 36, pg: 267-304.

[29] Strandenes, S. R. (1986). *Norship: A Simulation Model of Market in Bulk Shipping.* Discussion Paper Chapter 11. Norwegian School of Economics and Business Administration, Bergen, Norway.

[30] Thalassinos, E. I., & Politis, E. (2014). *Valuation Model For a second-hand vessel: econometric analysis of the dry bulk sector.* Journal of Global Business and Technology, 10(1).

[31] Kavussanos, M. G. (1996). *Price risk modelling of different size vessels in the tanker industry using autoregressive conditional heterskedastic (ARCH) models.* Logistics and Transportation Review, 32(2), pg: 161.

[32] Veenstra, A. W., & Haralambides, H. E. (2001). *Multivariate autoregressive models for forecasting seaborne trade flows.* Transportation Research Part E: Logistics and Transportation Review, 37(4), pg: 311-319.

[33] Kavussanos, M. G., & Nomikos, N. K. (1999). *The forward pricing function of the shipping freight futures market.* Journal of Futures Markets: Futures, Options, and Other Derivative Products, 19(3), pg: 353-376.

[34] Glen, D. R., & Martin, B. T. (1998). *Conditional modelling of tanker market risk using route specific freight rates.* Maritime policy and management, 25(2), pg: 117-128.

[35] Adland, R., & Koekebakker, S. (2007). *Ship valuation using cross-sectional sales data: A multivariate non-parametric approach.* Maritime Economics & Logistics, 9(2), pg: 105-118.

[36] Koehn, S. (2008). *Generalized additive models in the context of shipping economics.* (Doctoral dissertation, University of Leicester).

[37] Geomelos, N. D., & Xideas, E. (2017). *Econometric estimation of second-hand shipping markets using panel data analysis.* SPOUDAI-Journal of Economics and Business, 67(1), pg: 7-21.

[38] Tsolakis, S. (2005). *Econometric Analysis of Bulk Shipping: implications for investment strategies and financial decision-making.* (Doctoral dissertation, Erasmus University Rotterdam).

[39] Bendall, H. B., & Stent, A. F. (2005). *Ship investment under uncertainty: Valuing a real option on the maximum of several strategies.* Maritime Economics & Logistics, 7(1), pg: 19-35.

[40] Liapis, K. J., Christofakis, M. S., & Papacharalampous, H. G. (2011). *A new evaluation procedure in real estate projects.* Journal of Property Investment & Finance, 29(3), pg: 280-296.

[41] Hu, Q., & Zhang, A. (2015). *Real option analysis of aircraft acquisition: A case study.* Journal of Air Transport Management, 46, pg: 19-29.

[42] Merika, A., Merikas, A., Tsionas, M., & Andrikopoulos, A. (2019). *Exploring vessel-price dynamics: the case of the dry bulk market.* Maritime Policy & Management, 46(3), pg: 309-329.

[43] Adland, R., Cariou, P., & Wolff, F. C. (2018). *Does energy efficiency affect ship values in the second-hand market?* Transportation Research Part A: Policy and Practice, 111, pg: 347-359.

[44] Warstilla Encyclopedia. (2017). *Admiralty constant.* Available from: <https://www.wartsila.com/encyclopedia/term/admiralty coefficient-admiralty-constant>

[45] Raucci, C., Prakash, V., Rojon, I., Smith, T., Rehmatulla, N., & Mitchell, J. (2017). *Navigating Decarbonisation: An approach to evaluate shipping's risks and opportunities associated with climate change mitigation policy.* UMAS: London, UK.

[46] International Maritime Organisation. (2018). *Emission Control Areas.* MARPOL Annex VI. Available from: <http://www.imo.org/en/OurWork/Environment/PollutionPrevention/AirPollution/Pages/Air-Pollution.aspx> Accessed at 20 June 2019.

[47] Cullinane, K., & Bergqvist, R. (2014). *Emission control areas and their impact on maritime transport.* Transportation Research Part D: Transport and Environment. 28. pg: 1-5.

[48] Adland, R., & Jia, H. (2018). *Dynamic speed choice in bulk shipping.* Maritime Economics & Logistics, 20(2), pg: 253-266.

[49] Corbett, J. J., Wang, H., & Winebrake, J. J. (2009). *The effectiveness and costs of speed reductions on emissions from international shipping.* Transportation Research Part D: Transport and Environment, 14(8), pg: 593-598.

[50] Riska, K. (2005). *Ship–ice Interaction in Ship Design: Theory and Practice.* Encyclopedia of Life Support Systems (EOLSS), Developed under the Auspices of the UNESCO. Available from: <>.

[51] Rupp, K.H. (2012). *Ship Operation In Winter And In Ice Conditions.* Encyclopedia of Life Support Systems (EOLSS). Available from: <http://www.eolss.net/Sample-Chapters/C05/E6-178-72.pdf>.

[52] Solakivi, T., Kiiski, T., & Ojala, L. (2018). *The impact of ice class on the economics of wet and dry bulk shipping in the Arctic waters.* Maritime Policy & Management, 45(4), pg: 530-542.

[53] Kana, A. (2018). *Analytic Hierarchy Process.* Course-slides MT44035 'Design of Complex Specials' TU Delft.

[54] Nisbet, R., Elder, J., & Miner, G. (2009). *Handbook of statistical analysis and data mining applications.* Academic Press.

[55] Daobin, P. (1999). *STUDIES ON THE HISTORY OF STATISTIGAL ANALYSIS IN ANCIENT CHINA.* Conference Paper, presented at Proceedings of the 52nd Conference of the International Statistical Institute, Finland.

[56] Sayad, S. (2015). *Support Vector Machine - Regression (SVR)*. Rutgers University, New Jersey. Available from: <http://www.saedsayad.com/supportvectormachinereg.htm>. Accessed at 11 July 2019.

[57] Boehmke, B. (2018). *Multivariate Adaptive Regression Splines*. University of Cincinnati Business Analytics. Available from: <http://uc-r.github.io/mars>. Accessed at 28 June 2019.

[58] BCCVL. (2015). *MULTIVARIATE ADAPTIVE REGRESSION SPLINES*. Available from: <https://support.bccvl.org.au/support/solutions/articles/6000118097-multivariate-adaptive-regression-splines>. Accessed at 9 July 2019.

[59] BCCVL. (2015). *RANDOM FOREST*. Available from: <https://support.bccvl.org.au/support/solutions/articles/6000083217-random-forest>. Accessed at 10 July 2019.

[60] Boehmke, B. (2018). *Random Forests*. Available from: <http://uc-r.github.io/randomforests>. Accessed at 10 July 2019.

[61] Clark, M. (2019).*Generalized Additive Models*. University of Michigan. Available from: <https://m-clark.github.io/generalized-additive-models/>. Accessed at 28 June 2019.

[62] BCCVL. (2015). *GENERALIZED BOOSTING MODEL*.
Available from: <https://support.bccvl.org.au/support/solutions/articles/6000083212-generalized-boosting-model>. Accessed at 10 July 2019.

[63] Boehmke, B. (2018). *Gradient Boosting Machines*. Available from: <http://uc-r.github.io/gbmregression>. Accessed at 10 July 2019.

[64] Clark, M. (2019).*Machine Learning*. University of Michigan. Available from: <https://m-clark.github.io/introduction-to-machine-learning/>. Accessed at 28 June 2019.

[65] Boehmke, B. (2018). *Artificial Neural Network Fundamentals*. Available from: <http://uc-r.github.io/annfundamentals>. Accessed at 10 July 2019.

[66] BCCVL. (2015). *ARTIFICIAL NEURAL NETWORK*.
Available from: <https://support.bccvl.org.au/support/solutions/articles/6000083200-artificial-neural-network>. Accessed at 11 July 2019.

[67] Karn, U. (2016). *A Quick Introduction to Neural Networks*.
Available from: <https://ujjwalkarn.me/2016/08/09/quick-intro-neural-networks/>. Accessed at 11 July 2019.

[68] Stuart, G., Spruston, N., Sakmann, B., & Häusser, M. (1997). *Action potential initiation and backpropagation in neurons of the mammalian CNS*. Trends in neurosciences, 20(3), pg: 125-131.

[69] Scikit-Learning. (2016). *Support Vector Regression (SVR) using linear and non-linear kernels* Available from: <https://scikit-learn.org/stable/autoexamples/svm/plotsvmregression.html>. Accessed at 12 July 2019.

[70] Chen, S., Frouws, K., & Van de Voorde, E. (2010). *Technical changes and impacts on economic performance of dry bulk vessels*. Maritime Policy & Management, 37(3), pg: 305-327.

[71] Geertsma, R., Vollbrandt, J., Negenborn, R., Visser, K., and Hopman, H. (2017, August). *A quantitative comparison of hybrid diesel-electric and gas-turbine-electric propulsion for future frigates*. Presented in 2017 IEEE Electric Ship Technologies Symposium (ESTS) (pp. 451-458). IEEE.

[72] Jain, K. (2017). *Improving the competitiveness of green ship recycling*. Delft: TU Delft doctoral dissertation.

[73] Naval Surface Treatment Center. (2017). *Underwater Hull, Appendages, and Boottop.* Available from: <http://www.nstcenter.biz/navy-product-approval-process/approved-exterior-ship-coatings/coating-systems-underwater-hull-surfaces/>. Accessed at 8 Augusts 2019.

[74] Shi, W., Grimmelius, H. T., & Stapersma, D. (2010). *Analysis of ship propulsion system behaviour and the impact on fuel consumption.* International shipbuilding progress, 57(1-2), pg: 35-64.

[75] Hair Jr, J. F., Hult, G. T. M., Ringle, C., & Sarstedt, M. (2016). *A primer on partial least squares structural equation modeling (PLS-SEM).* Sage publications.

[76] International Schroders (2017). *The story of the global economy in pictures - April 2017* Available from: <https://www.schroders.com/pl/uk/tp/economics2/economics/the-story-of-the-global-economy-in-pictures—april-2017/>. Accessed at 3 September 2019.

[77] Hastie, T., Tibshirani, R., Friedman, J., & Franklin, J. (2005). *The elements of statistical learning: data mining, inference and prediction.* The Mathematical Intelligencer, 27(2), pg: 83-85.

[78] Wood, S. N. (2017). *Generalized additive models: an introduction with R.* Chapman and Hall/CRC.

[79] MarinerSpotted - Mariner Resource Center (2017). *Comparison of Slow, Medium and High Speed Engines.* Available from: <https://marinerspotted.com/2017/04/comparison-slow-medium-high-speed-engines/>. Accessed at 25 September 2019.

[80] Ship Insight (2017). *The technology and developments in ship transmissions.* Available from: <https://shipinsight.com/articles/the-technology-and-developments-in-ship-transmissions/>. Accessed at 25 September 2019.

[81] Molnar, C. (2018). *Interpretable Machine Learning.* Available from: <https://christophm.github.io/interpretable-ml-book/>. Accessed at 27 September 2019.

[82] Boehmke, B. (2018). *Preparing for Regression Problems.* University of Cincinnati Business Analytics. Available from: <http://uc-r.github.io/regression_preparation>. Accessed at 28 June 2019.

[83] Stephacking (2016). *Backward Elimination.* Available from: <http://stephacking.com/multivariate-linear-regression-python-step-6-backward-elimination/>. Accessed at 28 September 2019.

[84] RapidMiner (2019). *Backward Elimination (RapidMiner Studio Core).* Available from: <https://docs.rapidminer.com/latest/studio/operators/modeling/optimization/feature_selection/optimize_selection_backward.html>. Accessed at 28 September 2019.

[85] Frost, J. (2018). *How To Interpret R-squared in Regression Analysis.* Statistics By Jim. Available from: <https://statisticsbyjim.com/regression/interpret-r-squared-regression/>. Accessed at 28 June 2019.

[86] Wright, M. N., & Ziegler, A. (2015). *ranger: A fast implementation of random forests for high dimensional data in C++ and R.* arXiv preprint arXiv:1508.04409.

[87] Chen, T., & Guestrin, C. (2016, August). *Xgboost: A scalable tree boosting system.* In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794). ACM.

[88] Veronesi, F. (2017). *Assessing the Accuracy of our models (R Squared, Adjusted R Squared, RMSE, MAE, AIC).* R-blogger. Available from: <https://www.r-bloggers.com/assessing-the-accuracy-of-our-models-r-squared-adjusted-r-squared-rmse-mae-aic/>. Accessed at 28 August 2019.

[89] Yu, W., Xu, W., & Zhu, L. (2019). *A combined p-value test for the mean difference of high-dimensional data.* Science China Mathematics, 62(5), 961-978.

[90] Economicshelp. (2016). *Stagflation.* Available from: <https://www.economicshelp.org/blog/glossary/stagflation/>. Accessed at 9 October 2019.

[91] Schott, D. L. (2017). *Discontinuous transport.* Course-slides ME44000 Introduction to Transport Engineering and Logistic' TU Delft.

[92] Workman, D. (2019). *Coal Exports by Country.* Available from: <http://www.worldstopexports.com/coal-exports-country/>.

[93] Dillinger, J. (2019, August 28). *The Top 10 Coal Producers Worldwide.* Retrieved from: <https://www.worldatlas.com/articles/the-top-10-coal-producers-worldwide.html>.

[94] Workman, D. (2019). *Iron Ore Exports by Country.* Available from: <http://www.worldstopexports.com/iron-ore-exports-country/>.

[95] Sawe, B. E. (2017). *"Top Iron Ore Producing Countries In The World."* Retrieved from: <https://www.worldatlas.com/articles/top-iron-ore-producing-countries-in-the-world.html>.

[96] Workman, D. (2019). *Iron Ore Imports by Country.* Available from: <http://www.worldstopexports.com/iron-ore-imports-by-country/>.

[97] Humphreys, M., Stokenberga, A., Herrera Dappe, M., Hartmann, O., & Iimi, A. (2019). *Port Development and Competition in East and Southern Africa: Prospects and Challenges.* United Nations.

[98] Schmitz, T. (2015). *The Dry Bulk Freight Market.* Company Presentation of The European Energy Exchange (EEX) Group.

[99] International Maritime Organisation. (2019). *INTERNATIONAL CODE FOR SHIPS OPERATING IN POLAR WATERS (POLAR CODE).* Available from: <http://www.imo.org/en/MediaCentre/HotTopics/polar/Documents/POLAR%20CODE%20TEXT%20AS%20ADOPTED.pdf>.

[100] Staalesen, A. (2019). *As ice shrinks to year's low, a powerful fleet of tankers sail Arctic route to Asia.* Available from: <https://thebarentsobserver.com/en/arctic/2019/10/ice-shrinks-years-low-powerful-fleet-tankers-sail-arctic-route-asia>.

[101] DNV-GL (2019). *Supporting safe steel coil transport.* Available from: <https://www.dnvgl.com/expert-story/maritime-impact/Supporting-safe-steel-coil-transport.html>.

[102] TheNavalArch Team (2019). *STEEL COILS LOADING – ITS CHALLENGES AND WAYS TO OVERCOME.* Available from: <https://www.thenavalarch.com/steel-coils-loading-its-challenges-and-ways-to-overcome/>.

[103] Isbester, J. (1993). *Bulk Carrier Practice.* Available from: <http://www.harbour-maritime.com/uploads/1/2/9/8/12987200/bulk_carrier_practice.pdf>.

[104] Lehmann, E., Bockenhauer, M., Fricke, W., & Hansen, H. J. (1970). *Structural design aspects of bulk carriers.* WIT Transactions on The Built Environment, 27.

[105] Schøyen, H., & Bråthen, S. (2011). *The Northern Sea Route versus the Suez Canal: cases from bulk shipping.* Journal of Transport Geography, 19(4), 977-983.

[106] Jiang, S., & Wang, Y. T. (2012). *On Cargo Operation at Capesize Vessel.* In Software Engineering and Knowledge Engineering: Theory and Practice (pp. 281-285). Springer, Berlin, Heidelberg.

[107] CyberLogitec. (2019). Online *Port Charges Calculation Methods for Bulk Carriers' Outstanding Data Management in Bulk Shipping.* Available from <https://www.cyberlogitec.com/zh/news/port-charges-calculation-methods-for-bulk-carriers-outstanding-data-management-in-bulk-shipping/>.

[108] Wang, H. (2000). *Shipping pools in bulk shipping markets.* Master Thesis, World Maritime University (Malmö, Sweden)

[109] CMU Statistics. (2006). *Regression Trees.* Lecture Note. Available from: <http://www.stat.cmu.edu/ cshalizi/350-2006/lecture-10.pdf>

Appendices

A

Chapter 2 Appendix

.

A.1. Supply and Demand of Sea Transport

Demand Key Influences

Shipping demand is measured in ton miles of cargo. In general, the demand for sea transport changes quickly. This makes predicting the freight rate movement becomes a challenging task. In extreme cases, ship demand could change by 10% – 20% in a year. Longer-term changes happens primarily due to technological advancement. They are usually less extreme and not as obvious as the short-term changes. The combination of these two influences the dynamics of ship demand. To sum up, four biggest influencing factors that dictate the ship demand dynamics are discussed individually with main focus lies in the dry bulk market.

1. **The World Economy**
   Most of the demand for sea transport is driven by the world economy. They largely come from the needs to import and export the raw materials for manufacturing purpose and to to trade the resulting products. Maritime economists suggest that the economic growth fluctuation influences the seaborne trade. And as the world economics has a cyclical pattern, it will also create a cyclical patter for the ships demand. For more insight, this trend can be seen from figure **??** which outlines the rate of economics growth and sea trade between 1966–2006. A very close relationship between the world economy and the sea trade fluctuation can be observed. Thus the growth of world economies will propel the growth of shipping demand.

2. **Seaborne Commodities Trade**
   The sort of commodities that is shipped influences the transport demand in two ways. These are the short-term and long-term influences. Some particular commodities have seasonality characteristics. For example the agricultural products such as grains and fruits which have specific harvesting season. The seasonality creates short term volatility in ship demand. Other commodities such as crude oil, has a completely different characteristic. For many years during 1960s, crude oil created the largest demand for the sea transport. At that time, crude oil demand was up to three times higher than the world economic growth rate.

   This was because many developed countries had switched from coal to oil as their new major energy source. However, as the oil prices kept increasing, the crude oil demand got stagnant in the 1970s. At the end of 1970, oil demand start decreasing until the mid of 1980. Meanwhile, the demand for tankers followed closely the crude oil demand. This is an example of how the commodity's demand influences the long-term sea transport demand.

3. **Random Shocks**
   Random shocks refer to various unexpected events which disturb the economic stability of a certain region or globally. As the economic system is disturbed, the economic cycles move irregularly. Several examples of randoms shocks are the climate changes, wars, the discovery new resources and certain political events. As economic shock is created, it might indirectly yet significantly influence the shipping demand.

For instance, the depression after the Wall Street Crash of 1929 has caused the global trade to decline. It has significantly brought disadvantage to global shipping demand in 1929. Another example is the political decision to nationalize and temporarily closed the Suez Canal by the Egyptian government in 1956. It has created a sudden increase in ship demand at that time because ships had to take a much longer Asia - Europe route.

4. **Transport Cost**
   The last important factor is the transport cost. It is important because a company will decide to import certain raw material from a distant source only if the shipping cost is reasonable enough for them to generate profit. In other words, as the shipping cost decreases, the shipping demand is expected to increase. Over the last century, the shipping cost has been gradually declining due to ships' efficiency improvement and better organization. In addition, the shipping service has also been gradually increasing. These lead to a non-dramatic yet longer-term positive effect on ships demand.

Supply Key Influences

There are four primary stakeholders who control the total ships supply. These are *shipowners*, *charterers*, *financial institutions* and *maritime regulators* such as International Maritime Organization. Shipowners decides whether new ships are ordered or old vessels are scrapped. Another option that shipowners can make is whether the ships are laid up during the recession time. Charterers influence the freight rate which indirectly affect the ships supply and demand. Sometimes, charterers can influence the shipowners to acquire an additional vessel by offering a long-term contract.

Bankers finance the shipbuilding or in bad situation, they may pressure shipowners to scrap their ships early. Lastly, regulators make new regulation which might influence the transport supply. One important example is Regulation 13G introduced in 2003 introduced by IMO has all forced single hull tankers to be scrapped by 2010. Lastly, four main key influences that control the fleet supply are examined, with main focus lies in the dry bulk market.

1. **World Fleet**
   The total number of merchant fleet is the most important factor regarding the sea transport supply. The more ships means higher transport supply. The fleet growth rate is determined by the sum of newbuilding deliveries and old-vessels scrapping. Since the average lifetime of a vessel is between 25-30 years, there are usually more vessels being built than scrapped yearly. In shipping market, ships supply is continually adjusted according to the transport demand. As an instance, main dry bulks were transported at the bottom of passenger liners in the past. Shippers started switching from combined carriers to the bulk carriers in the late 1950s.

   This happened mainly due to economical reason. The economic of scale has enabled transport cost to become considerably cheaper by using the large bulk carriers. As the result, the global demand for bulk transport has grown significantly. It is recorded that the total global demand for bulk transport has grown from 448 millions tons in 1970 to 3196 millions tons in 2007 [8]. This substantial growth of bulk carriers implies the rise of global sea transport supply.

2. **Fleet Productivity**
   The fleet size is given in deadweight or DWT. While fleet size remains the same, the vessels' productivity and cargo capacity could still be adjusted. There has been a trend of increasing vessels productivity due to technological advancement. Given that the number of fleets worldwide is constant, higher productivity will positively influence the transport supply. Ships productivity is usually measured in ton miles per deadweight. In general, there are 4 main factors which determine vessels productivity. These are ships speed, port time, deadweight utilization and number of days at sea. There has been an increasing trend in ships operating speed.

Furthermore, containerization has significantly reduced the port time to load and unload the cargo. Better bulk handling has also positively increased the bulk carriers efficiency. For bulk carriers, 95% of deadweight utilization is generally applied. Regarding the days at sea, there has been an increasing trend of "the productive days". In addition to it, modern ships are designed with flexibility which allow them to switch cargoes in their return journey. All these changes have been contributed to the increasing productivity which signifies higher transport supply.

3. **Shipbuilding Production**
   The shipbuilding actively influences the number of ship supply. Since shipbuilding is a long process (around 1- 4 years, depend on the orderbook size), number of fleets adjustment do not happen immediately. newbuilding orders are placed based on the expectation of future demand. In some countries, the shipbuilding are politically supported by the government to prevent laying-off. In shipyards' perspective, the type of vessels which they produce matter; since a peak or trough periods are different for various type of vessels. As a new strategy to survive the bad period, some shipyards produce a vast variety of vessels type.

4. **Scrapping and Losses**
   The sum of new ships deliveries and removal of older vessels is equal to the fleet growth rate. Fleets removal mainly takes place in the form of ships scraping and sometimes due to ships lost at sea. There are many reasons which influence a decision to scrap a vessel. The most important factors are the vessels' age, technical obsolescence, scrap prices, current earnings and market expectations.

   Older ships are less efficient and have higher operational expenses; thus shipowners tend to scrap them, early during the bad market. Additionally, ships that are lost at sea for different reasons. Several primary causes are the bad weather, fire and collision which led to accidental sinking. Human error, navigation error and poor design also played an important roles here. These have led to the reduction in global ships supply.

A.2. Shipping Cycles

The content is summarized from the book *Maritime Economics* by Stopford[4].

Long Shipping Cycles

The most substantial component of cyclical mechanism is unarguably the long-term cycles. They are driven by major technical, economic or regional change. Despite of their importance, they are the most difficult to identify because of their very long period. Regarding to the world economy, Kondratief was the first economist who has observed and developed the long-cycle theory. According to his statistical analysis, there were in total three long cycles of the world's economy. They took place between 1790 and 1916 with average period of 50 to 60 years.

Correspondingly, another economist, Schumpeter explained that these world's economy cycles were primarily driven by the technological advancement. The invention of steam machine was the major driver of the upturn of the first cycle (1790–1813). The major railway construction was the main factor of upturn of the second cycle (1844–74). Lastly, the massive usage of motor car and electricity has induced the upswing of the third cycle (1895–1916).

The shipping cycles are mainly influenced by economic cycles. However, man can make a clear distinction between the world's economic and shipping cycles. Among all, the long-term shipping cycles are more difficult to observe and there is no certain rule of thumb related to it. However, one can recognize the trend of freight rates decline which took place between 1869 and 1914. The main cause was the technological advancement which led to the increasing efficiency of steamships.

Thereafter, industrialization of shipping businesses by using bigger bulkers and more efficient cargo-handling technology led to an even lower recession of freight rates between 1945 and 1995 [4]. During that period, new advancement such as the triple expansion engine or containerization set a higher bar of efficiency standard in the shipping business. Lastly, the intense global digitization starting from year 2000 onward has made shipping business even more efficient which triggered even more decline in freight rates.

Short Shipping Cycle

Unlike the long cycles, the short cycles are much easier to observe, both in the context of the world economy and shipping market. The study about short economic cycles started in the early nineteenth century. These short cycles moves quickly and easily. While the period of short cycle is not regular, the cycle development always consist of four main stages. The first stage is the ***market trough*** where there are much shipping capacities surplus in comparison with demand. This cause the freight rates revenue to decrease below ships' operational expenses. In this condition, the price of second hand vessels fall and many would prefer to scrap their old ships.

Scrapping leads to the reduction of ships supply. As the supply keeps decreasing, the market will soon enter the second stage, namely the ***market recovery***. This is when supply and demand of shipping capacities move toward point of equilibrium. As the transport supply keeps decreasing, freight rates revenue keeps increasing until it soars above ships' operational cost. In this case, the cash flows positively into shipowners who now have more money to invest in more assets. After that point, the old and new vessels price increases. More new-buildings are coming to the market and less old vessels are scraped.

This will soon lead to the third stage which is the ***market peak***. At peak, freight rates become really high and many shipowners order new ships. Furthermore, as these new ships are ready, transportation supply becomes once more higher than the demand. This will lead to the downfall of freight rate which mark the ***market collapse***. This is the last stage of shipping cycle. Similar to every cycle, the short shipping cycle will repeat itself after some time. While each cycle is unique, every cycle typically consists of these four stage. These cycles naturally force the weak and inefficient shipping companies out of the market.

Seasonal Cycle

Seasonal cycles occur regionally in shipping market (in specific shipping routes). These are characterized by the fluctuations in freight rates which take place at specific seasons of the year. This is caused by seasonal patterns of sea transport demand. Seasonal cycles are influenced by external factors such as specific holiday season (like Christmas and Chinese New Year) when specific holiday products are produced and shipped only for a limited time period.

Another example is the tendency of dry bulk market to decline around July and August. The explanation is because relatively less grain is being shipped around the summer since it is not the harvest season. The seasonal cycle also occurs in the reefer trade since there are fresh fruit is shipped to the Northern Hemisphere during the specific harvest season. The last important example is the tendency of four season countries to stock up oil in the winter. This leads to seasonal peak period of tanker demand.

B

Chapter 3 Appendix

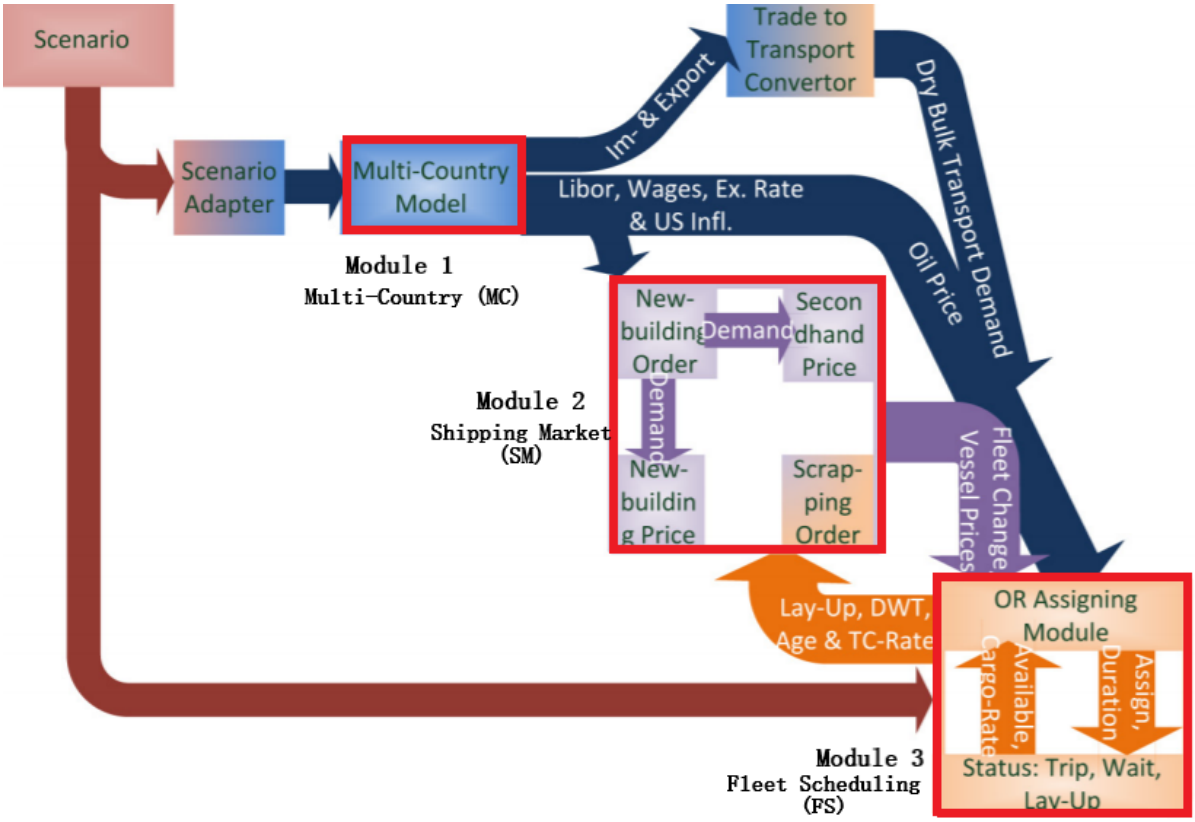


Figure B.1: Model Interaction and Scenario Generation. Source: Pruyn [5]

C
Chapter 4 Appendix

C.1. Bias & Variance Trade-off

The bias-variance trade-off is one of the central problems in supervised machine learning. The total error in machine learning model comes from three main sources, namely *bias*, *variance* and *irreducible error*. These errors are also known as *noise*; a noisy system can also mathematically be written as following:

$$y = f(X) + \epsilon$$

Where *f(x)* is the expected value of *y* given *X*, or *f(x) = E(y|X)*. In general, following assumptions can be taken: the expected value of the error, $E(\epsilon) = 0$ and variance of error $\text{Var}(\epsilon) = \sigma_\epsilon^2$. Moreover, three error components can be further explained as following:

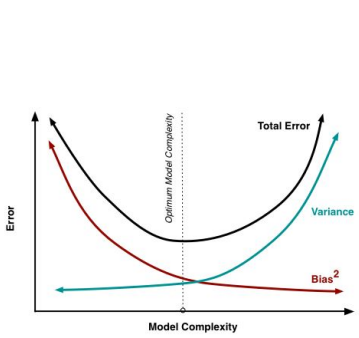
- **Irreducible error** - The variance of the (new test) target (σ_ϵ^2). This error is unavoidable, since our *y* is measured with error.
- **Bias**^{2} - The amount the average of our estimate varies from the true (but unknown) value $\left(E(\hat{f}) - f\right)$. This is often the result of trying to model the complexity of nature with something much simpler that the human brain can understand. While the simpler might make us feel good, it may not work very well.
- **Variance** - The amount by which our prediction would change if we had estimated it using a different training data set $\left(Var(\hat{f})\right)$. Even with unbiased estimates, we could still see a high mean squared error due to high variance.

The Error_{x_*} is the average, or expected value of the prediction error in this scenario, or $E[(y - \hat{f}(x))^2 | X = x_*]$, with $\hat{f}$ as estimates and f as true (unknown) values. Furthermore, these errors can be re-written as following:

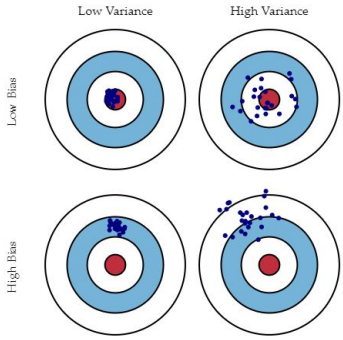
$$\text{Error}_{x_*} = \text{IrreducibleError} + \text{Bias}^2 + \text{Variance}$$
$$\text{Error}_{x_*} = \text{Var}(\epsilon) + (\text{E}[h_*] - f(x_*))^2 + \text{Var}(h_*)$$

The last two errors components make up the mean squared error. While they are controllable, they compete with one another. It means that reducing one type of error will increase the other since bias and variance are not independent. Generally speaking, the more complex the model, the lower the bias, but the higher the variance will be. On contrary, less complex model will have more bias and less variance. Figure C.1 illustrates the relationship between model complexity, variance-error and bias-error. To better understand the difference between bias and variance, the illustration of the nature of these two errors are depicted in figure C.1.

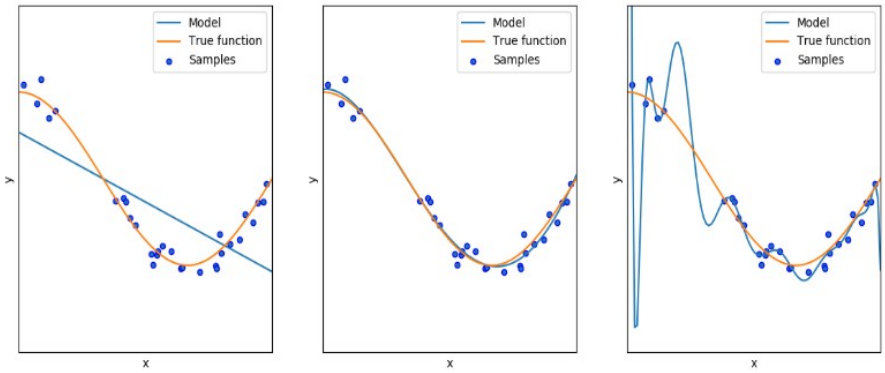
Models with high variance are susceptible to overfitting and overtraining. Overtraining model might fit one data set properly, however it is not general enough to be applicable to other data set. On the other hand, model with low variance are susceptible to underfitting, a condition where when a statistical model cannot adequately capture the underlying structure of the data. An illustration of overfitting and underfitting model is given in figure C.1. None of both conditions are desirable. Thus, the main modelling challenge is to find the balance between these two. Tuning should be done to find the proper balance of robust model.



((a)) Relationship between variables



((b)) Trade-off



((c)) Comparison between Models

Figure C.1: Model Complexity, Bias and Variance. Source: Clark[64]

C.2. GAM Comparison

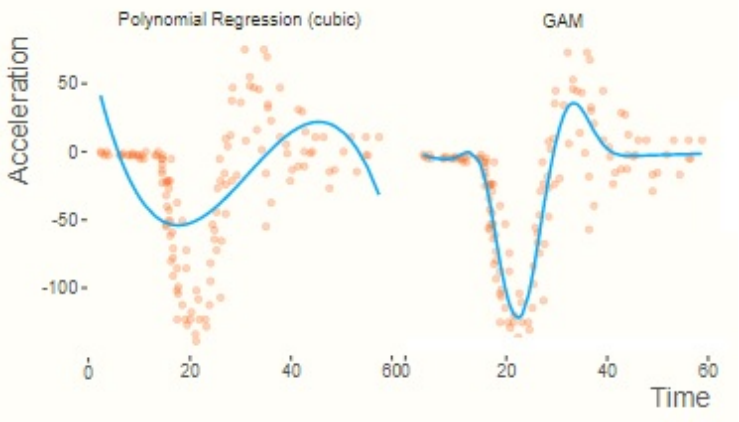


Figure C.2: Comparison of Result between GAM and Cubic Regression. Source: Clark [61].

C.3. MARS Comparison

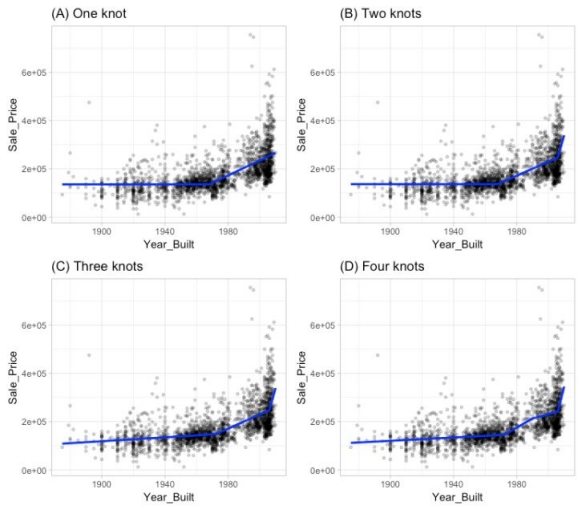


Figure C.3: MARS Simulation with Different BF Types and Number of Knots. Source: Boehmke [57]

C.4. Gradient Boosting Machine Illustration

C.4.1. Algorithm

1. Fit a decision tree to the data - $F_1(x) = y$.
2. Fit the next decision tree to the residuals of the previous - $h_1(x) = y - F_1(x)$.
3. Add new tree to previous algorithm - $F_2(x) = F_1(x) + h_1(x)$
4. Fit the next decision tree to the residuals of F_2 - $h_2(x) = y - F_2(x)$
5. Add new tree to previous algorithm - $F_3(x) = F_2(x) + h_2(x)$
6. Iterate this process until a (pre-)specified algorithm (for example *cross-validation*) command it to stop.

C.4.2. Loss Function

Different type of Loss functions are used for different types of outcomes. Two main type of outcomes are numerical and categorical. In this section, list of suitable loss functions for these two outcomes types are given. Each function is the most suitable in particular situation, however further discussion about these specific situations are outside the scope of this section. One can consult the book *The Elements of Statistical Learning (2009)* by Hastie, Tibshirani, and Friedman about more information about the specific situations in which those functions work best. Lastly, the comparison between the performance of these functions is given in figure C.4.

Numerical Outcomes

Squared Error

$$L(Y, f(X)) = \sum (y - f(X))^2$$

Absolute Error

$$L(Y, f(X)) = \sum |(y - f(X))|$$

Negative Log-likelihood

$$L(Y, f(X)) = n \ln \sigma + \sum \frac{1}{2\sigma^2} (y - f(X))^2$$

Categorical Outcomes

Misclassification

$$L(Y, f(X)) = \sum I(y \neq \text{sign}(f))$$

Binomial log-likelihood

$$L(Y, f(X)) = \sum \ln(1 + e^{-2yf})$$

Exponential

$$L(Y, f(X)) = \sum e^{-yf}$$

Hinge Loss

$$L(Y, f(X)) = \max(1 - yf, 0)$$

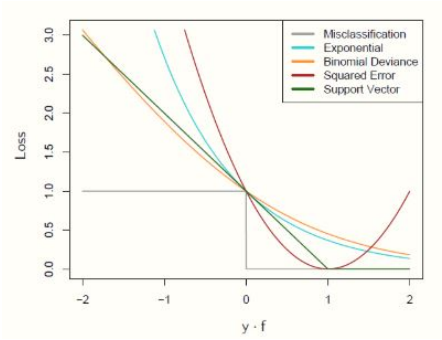
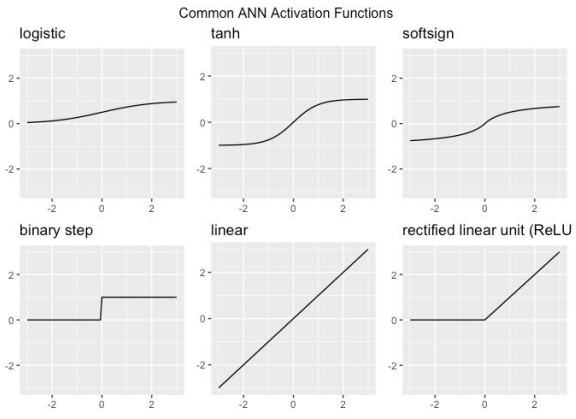
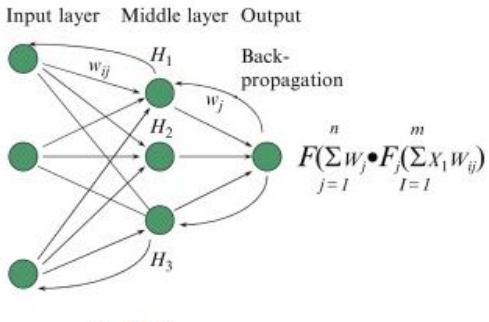


Figure C.4: Comparison of Performance between Different Loss Functions. Source: The Elements of Statistical Learning (2009)

C.5. Artificial Neural Netowrk



((a)) Activation Functions for ANN



((b)) Back-propagation ANN

Figure C.5: ANN Illustration

C.6. SVM Comparison

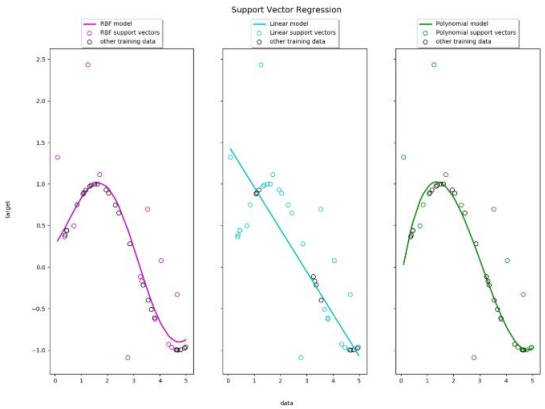


Figure C.6: Comparison of SVM Result Using Different Kernel Function. Source: Scikit [69]

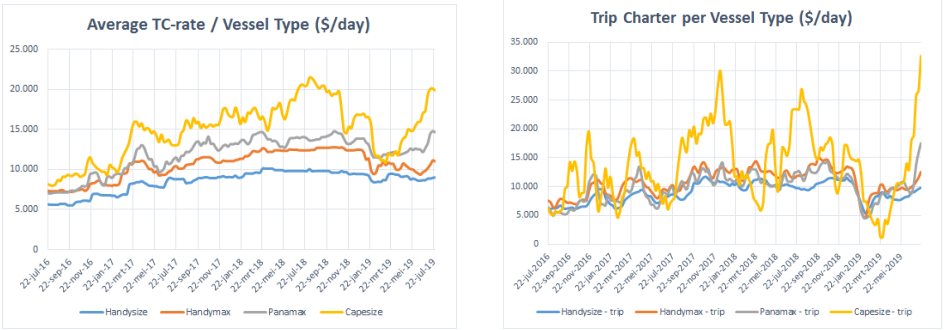
D

Chapter 5 Appendix

D.1. Time Series Variables

The content is the processed information from the Clarkson Database[7].

D.1.1. Time & Trip Charter Rate Summary



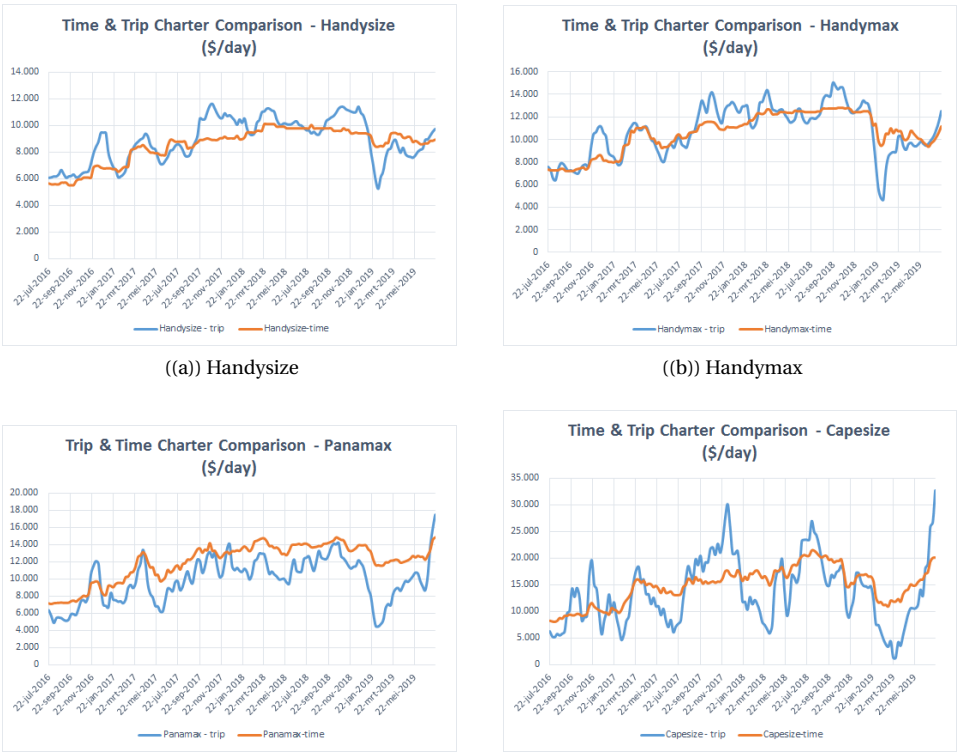
((a)) Average Time Charter

((b)) Trip Charter

Figure D.1: Income per Vessel Type

From figure D.1, one can observe that the capesize has the most dynamic fluctuation in both cases. In addition, one can also observe that time-charter is relatively more stable than trip charter. It might happen because time-charter contracts are more commonly used in comparison to trip-charter. For this reason, using the time charter rate instead of trip charter might be a better idea since it would give a better income representative.

D.1.2. Comparison between Time Charter and Trip Charter Rate



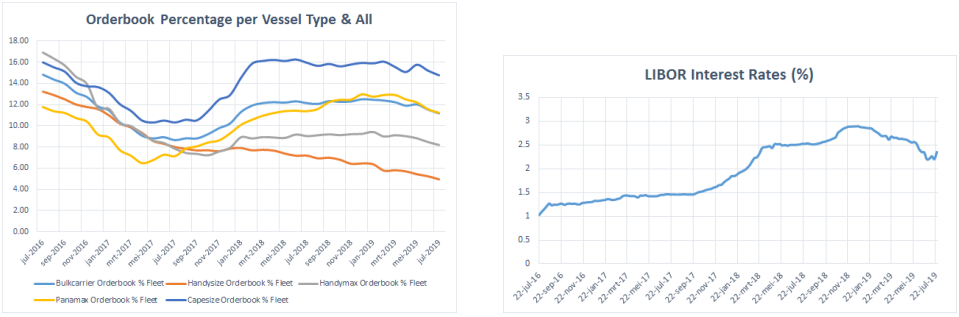
((a)) Handysize ((b)) Handymax

((c)) Panamax ((d)) Capesize

Figure D.2: Comparison Time Charter & Trip Charter

When one zooms into the vessel type, a comparison between trip and time-charter per vessel type can be made. From figure D.2 one can observe that, although both follows the same trend, trip charter rate is considerably unstable in every vessel type. This confirms that choosing the time-charter rate would be a better option.

D.1.3. Orderbook Percentage & LIBOR Summary



((a)) Orderbook ((b)) LIBOR

Figure D.3: Time-series Indicator Summary

From figure D.3 one can observe that the ordebook of all and specific vessel types (except handysize) follow the same trend. Handysize has a slightly different trend, namely the orderbook consistently declines while other types experiences a short "increase" at the end of 2017. It might be explained due to an already high number of handysize vessels, and without significant increase in demand, the orderbook kept going down. In addition to this, one can observe a relatively stable LIBOR rate with a "small increase" in 2018.

D.2. Builder Reputation

The content is the processed information from the Clarkson Database[7].

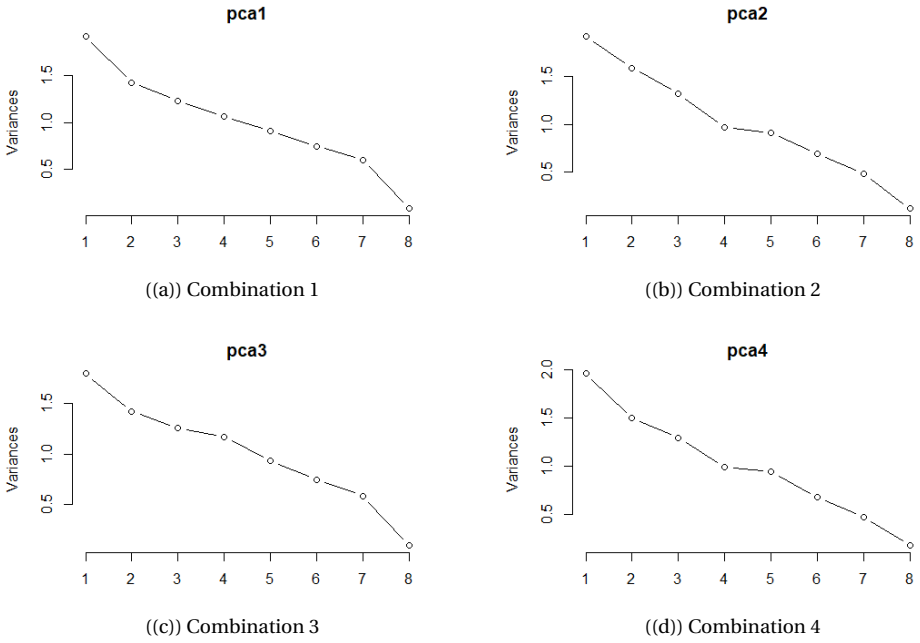
Country of Build	**Number of Vessels**	**Percentage**
Argentina	1	0.08%
P.R. China	367	29.91%
Croatia	2	0.16%
Indonesia	1	0.08%
Italy	2	0.16%
Japan	675	55.01%
Philippines	42	3.42%
South Korea	112	9.13%
Taiwan	13	1.06%
Vietnam	12	0.98%
Total	1227	100.00%

Table D.1: Country of Yard Summary.

Classification Society	Number of Vessels	Percentage
American Bureau of Shipping	98	8.0%
Biro Klasifikasi Indonesia	2	0.2%
Bureau Veritas	139	11.3%
China Classification Society	169	13.8%
DNV GL	79	6.4%
Dromon Bureau of Shipping	2	0.2%
Hellenic Register of Shipping	1	0.1%
Indian Register of Shipping	15	1.2%
Intermaritime Certification Services	2	0.2%
China Corporation Register	1	0.1%
Columbus American Register	1	0.1%
International Naval Surveys Bureau	1	0.1%
Isthmus Bureau of Shipping S.A.	1	0.1%
Korean Register of Shipping	50	4.1%
Lloyd's Register	98	8.0%
Nippon Kaiji Kyokai	442	36.0%
Overseas Marine Certificate	1	0.1%
Phoenix Register of Shipping S.A.	5	0.4%
Polish Register of Shipping	6	0.5%
Registrano Italiano Navale	61	5.0%
Russian Register	6	0.5%
Vietnamese Register	5	0.4%
Zhong Chuan	25	2.0%
Unknown	17	1.4%
Total	1227	100.0%

Table D.2: Classification Society Summary.

Chapter 6 Appendix



((a)) Combination 1

((b)) Combination 2

((c)) Combination 3

((d)) Combination 4

Figure E.1: Proportion Variance for Various Combinations of Numerical Variables.

Chapter 7 Appendix

F.1. Single Significant Test for Eliminated Variables

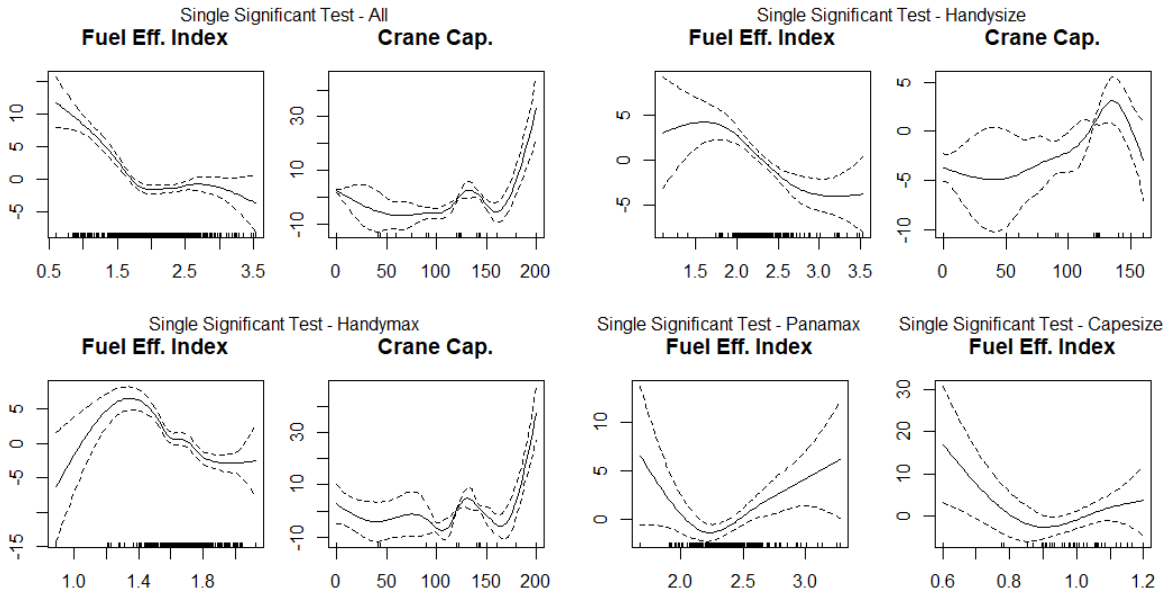


Figure F.1: Single Significant Test per Vessel Type.

The content is the processed information from the Clarkson Database[7].

F.2. Intermediate Result

F.2.1. All Vessels

Iteration	Second											
Model	All1			All2			All3			All4		
Parametric	Estimate	p-value		*Estimate*	p-value		Estimate	p-value		Estimate	p-value	
(Intercept)	7.31	< 2e-16	***	7.30	< 2e-16	***	7.32	< 2e-16	***	7.31	< 2e-16	***
scrubberY	2.01	0.00	***	2.01	0.00	***	2.13	0.00	***	2.09	0.00	***
anti_foulingY												
fuel_typeMDO												
japanY	2.31	< 2e-16	***	2.33	< 2e-16	***	2.32	< 2e-16	***	2.32	< 2e-16	***
chinaY	0.31	0.14		0.33	0.12		0.25	0.25		0.28	0.19	
southkoreaY	3.15	< 2e-16	***	3.15	< 2e-16	***	3.13	< 2e-16	***	3.12	< 2e-16	***
rest_countryY	1.54	0.00	***	1.50	0.00	***	1.63	0.00	***	1.59	0.00	***
bvY	1.67	0.00	***	1.67	0.00	***	1.67	0.00	***	1.66	0.00	***
ccsY	2.05	0.00	***	2.01	0.00	***	2.06	0.00	***	2.02	0.00	***
nkkY	1.71	< 2e-16	***	1.72	< 2e-16	***	1.67	< 2e-16	***	1.70	< 2e-16	***
rest_classY	1.89	< 2e-16	***	1.90	< 2e-16	***	1.93	< 2e-16	***	1.93	< 2e-16	***
Smooth-term	edf	p-val.	sig.	edf	p-val.	sig.	edf	p-val.	sig.	edf	p-val.	sig.
s(age)	7.04	< 2e-16	***	7.05	< 2e-16	***	7.05	< 2e-16	***	7.12	< 2e-16	***
s(tc)	6.92	0.00	***	6.68	0.00	***						
s(norm_tc)							4.04	0.00	***	3.86	0.00	***
s(dwt)	8.39	< 2e-16	***	8.41	< 2e-16	***	8.42	< 2e-16	***	8.39	< 2e-16	***
s(libor)	1.79	0.00	***	5.93	0.00	**	2.01	0.00	***	7.70	0.00	**
s(orderbook_all)	1.00	0.00	**				1.00	0.00	**			
s(orderbook_type)				1.00	0.36					1.00	0.85	
s(norm_fuel_consumption)												
s(vol_per_dwt)	8.55	< 2e-16	***	8.55	< 2e-16	***	8.58	0.00	***	8.58	0.00	***
s(norm_admiralty_constant)												
Result	R-sq.	GCV	n	R-sq.	GCV	n	R-sq.	GCV	n	R-sq.	GCV	n
	0.768	11.976	1172	0.768	12.032	1172	0.769	11.883	1172	0.769	11.957	1172

Figure F.2: Second Iteration for All Vessels

F.2.2. Handysize Vessels

Iteration	Second						Third						Fourth					
Model	Handysize1			Handysize2			Handysize1			Handysize2			Handysize1			Handysize2		
Parametric	Est.	p-val.	sig.	Est.	p-val.	sig.	Est.	p-val.	sig.	Est.	p-val.	sig.	Est.	p-val.	sig.	Est.	p-val.	sig.
(Intercept)	5.81	< 2e-16	***	5.81	< 2e-16	***	5.79	< 2e-16	***	5.88	< 2e-16	***	5.84	< 2e-16	***	5.86	< 2e-16	***
ice_classY																		
anti_foulingY	0.01	0.99		0.01	0.99													
gearboxY																		
japanY	0.91	0.01	*	0.91	0.01	*	0.93	0.28		0.79	0.36							
chinaY	-0.47	0.22		-0.47	0.22		-0.93	0.28		-0.94	0.28							
southkoreaY	5.15	0.00	***	5.15	0.00	***	4.21	0.00	***	4.15	0.00	***	4.26	0.00	***	4.29	0.00	***
rest_countryY	0.22	0.75		0.22	0.75													
bvY	1.74	0.00	***	1.74	0.00	***	1.66	0.00	***	1.59	0.00	***	1.81	0.00	***	1.73	0.00	***
ccsY	0.82	0.23		0.82	0.23		1.36	0.03	*	1.53	0.01	*	0.71	0.23		0.88	0.13	
nkkY	1.23	0.00	***	1.23	0.00	***	1.08	0.00	**	1.16	0.00	***	1.65	0.00	***	1.67	0.00	***
rest_classY	2.01	0.00	***	2.01	0.00	***	1.69	0.00	***	1.60	0.00	***	1.66	0.00	***	1.58	0.00	***
Smooth-term	edf	p-value		edf	p-value		edf	p-value		edf	p-value		edf	p-value		edf	p-value	
s(age)	2.85	< 2e-16	***	2.88	< 2e-16	***	4.83	< 2e-16	***	4.92	< 2e-16	***	6.07	< 2e-16	***	6.26	<2e-16	***
s(tc)	5.49	0.01	**				6.01	0.00	**				5.69	0.00	***			
s(norm_tc)				1.00	0.00	***				1.00	0.00	***				1.00	<2e-16	***
s(dwt)	1.43	0.00	***	1.39	0.00	***	1.75	< 2e-16	***	1.79	< 2e-16	***	1.00	< 2e-16	***	1.00	<2e-16	***
s(libor)	1.00	0.16		1.00	0.07	.	1.00	0.35		1.00	0.07	.						
s(orderbook_all)																		
s(orderbook_type)																		
s(norm_fuel_consumption)																		
s(vol_per_dwt)	1.00	0.43		1.00	0.26													
s(norm_admiralty_constant)	1.00	0.07	.	1.00	0.12													
Result	R-sq.	GCV	n	R-sq.	GCV	n	R-sq.	GCV	n	R-sq.	GCV	n	R-sq.	GCV	n	R-sq.	GCV	n
	0.631	9.425	316	0.626	9.397	247	0.664	8.637	355	0.658	8.658	355	0.648	8.968	355	0.658	8.658	355

Figure F.3: Second, Third and Fourth Iteration for Handysize Vessels

F.2.3. Handymax Vessels

Iteration	Second											
Model	Handymax1			Handymax2			Handymax3			Handymax4		
Parametric	Est.	p-val.	sig.	Est.	p-val.	sig.	Est.	p-val.	sig.	Est.	p-val.	sig.
(Intercept)	6.85	< 2e-16	***	6.87	< 2e-16	***	6.88	< 2e-16	***	6.89	< 2e-16	***
scrubberY	1.31	0.05	.	1.28	0.06	.	1.14	0.09	.	1.13	0.09	.
anti_foulingY												
fuel_typeMDO												
japanY	2.95	< 2e-16	***	2.92	< 2e-16	***	2.89	< 2e-16	***	2.87	< 2e-16	***
chinaY	0.14	0.64		0.12	0.69		0.12	0.68		0.10	0.74	
southkoreaY	1.99	0.00	***	2.03	0.00	***	1.99	0.00	***	2.04	0.00	***
rest_countryY	1.77	0.00	***	1.80	0.00	***	1.87	0.00	***	1.88	0.00	***
bvY	1.58	0.00	***	1.61	0.00	***	1.51	0.00	***	1.54	0.00	***
ccsY	1.97	0.00	***	1.95	0.00	***	2.01	0.00	***	2.02	0.00	***
nkkY	1.64	0.00	***	1.65	0.00	***	1.65	0.00	***	1.64	0.00	***
rest_classY	1.66	0.00	***	1.66	0.00	***	1.71	0.00	***	1.70	0.00	***
Smooth-term	edf	p-val.	sig.	edf	p-val.	sig.	edf	p-val.	sig.	edf	p-val.	sig.
s(age)	1.87	< 2e-16	***	1.91	< 2e-16	***	1.87	< 2e-16	***	1.88	< 2e-16	***
s(tc)	3.10	0.00	***	2.99	0.03	*						
s(norm_tc)				2.90	0.00	***	1.63	0.01	**	1.44	0.03	*
s(dwt)	2.88	0.00	***	2.90	0.00	***	2.92	0.00	***	2.93	0.00	***
s(libor)	1.00	0.03	*	1.00	0.06	.	1.56	0.01	**	1.09	0.00	**
s(orderbook_all)	1.00	0.05	.				1.00	0.14				
s(orderbook_type)				1.00	0.04	*				1.00	0.03	*
s(norm_fuel_consumption)												
s(vol_per_dwt)	2.32	0.12		2.31	0.11		2.36	0.09	.	2.32	0.09	.
s(norm_admiralty_constant)												
Result	R-sq.	GCV	n	R-sq.	GCV	n	R-sq.	GCV	n	R-sq.	GCV	n
	0.755	8.159	437	0.755	8.150	437	0.754	8.182	437	0.754	8.173	437

Figure F.4: Second Iteration for Handymax Vessels

Iteration	Third											
Variable	Handymax1			Handymax2			Handymax3			Handymax4		
Parametric	Est.	p-val.	sig.	Est.	p-val.	sig.	Est.	p-val.	sig.	Est.	p-val.	sig.
(Intercept)	7.19	< 2e-16	***	7.19	< 2e-16	***	7.19	< 2e-16	***	7.19	< 2e-16	***
scrubberY	1.09	0.18		1.04	0.20		0.87	0.28		0.85	0.28	
anti_foulingY												
fuel_typeMDO												
japanY	2.47	0.00	***	2.47	0.00	***	2.42	0.00	***	2.43	0.00	***
chinaY												
southkoreaY	1.42	0.13		1.52	0.10		1.52	0.10		1.61	0.08	.
rest_countryY	1.24	0.05	*	1.32	0.03	*	1.40	0.02	*	1.47	0.02	*
bvY	1.61	0.00	***	1.64	0.00	***	1.51	0.00	***	1.54	0.00	***
ccsY	1.78	0.00	***	1.80	0.00	***	1.85	0.00	***	1.87	0.00	***
nkkY	1.73	0.00	***	1.71	0.00	***	1.72	0.00	***	1.70	0.00	***
rest_classY	2.06	0.00	***	2.04	0.00	***	2.12	0.00	***	2.08	0.00	***
Smooth-term	edf	p-val.	sig.	edf	p-val.	sig.	edf	p-val.	sig.	edf	p-val.	sig.
s(age)	1.97	< 2e-16	***	1.91	< 2e-16	***	1.88	< 2e-16	***	1.84	< 2e-16	***
s(tc)	3.35	0.00	***	3.35	0.02	*						
s(norm_tc)							1.78	0.00	***	1.00	0.00	**
s(dwt)	2.71	0.00	***	2.78	0.00	***	2.80	0.00	***	2.83	0.00	***
s(libor)	1.00	0.23		1.00	0.11		1.45	0.02	*	1.21	0.01	**
s(orderbook_all)				1.00	0.16							
s(orderbook_type)										1.00	0.03	*
s(norm_fuel_consumption)												
s(vol_per_dwt)												
s(norm_admiralty_constant)												
Result	R-sq.	GCV	n	R-sq.	GCV	n	R-sq.	GCV	n	R-sq.	GCV	n
	0.686	11.515	438	0.687	11.501	438	0.684	11.122	438	0.684	11.540	438

Figure F.5: Third Iteration for Handymax Vessels

F.2.4. Panamax Vessels

Iteration	Second						Third					
Model	Panamax1			Panamax2			Panamax1			Panamax2		
Parametric	Est.	p-val.	sig.	Est.	p-val.	sig.	Est.	p-val.	sig.	Est.	p-val.	sig.
(Intercept)	7.90	0.00	***	7.85	0.00	***	9.96	< 2e-16	***	9.86	< 2e-16	***
scrubberY	2.38	0.03	*	2.36	0.03	*	2.60	0.00	**	2.63	0.00	**
anti_foulingY	-0.50	0.46		-0.48	0.47		-0.27	0.70		-0.19	0.79	
fuel_typeMDO	-5.83	0.11		-5.76	0.11		-2.92	0.43		-2.63	0.48	
japanY	1.40	0.22		1.57	0.16		1.24	0.01	**	1.34	0.00	**
chinaY	-0.56	0.65		-0.43	0.73							
southkoreaY	0.95	0.43		1.12	0.34							
rest_countryY												
bvY												
ccsY	2.20	0.01	**	2.11	0.01	*	1.81	0.03	*	1.88	0.02	*
nkkY	1.57	0.05	*	1.38	0.08	.	0.82	0.25		0.80	0.27	
rest_classY	2.11	0.01	**	1.99	0.01	**	1.12	0.10	.	1.20	0.08	.
Smooth-term	edf	p-val.	sig.	edf	p-val.	sig.	edf	p-val.	sig.	edf	p-val.	sig.
s(age)	2.56	0.00	***	2.59	< 2e-16	***	7.04	0.00	***	7.07	< 2e-16	***
s(tc)	2.40	0.49					3.88	0.22				
s(tc_norm)				2.53	0.00	**				2.89	0.13	
s(dwt)	2.25	0.21		1.52	0.17		2.26	0.06	.	2.04	0.10	.
s(libor)	1.00	0.52		1.00	0.15							
s(orderbook_all)												
s(orderbook_type)												
s(norm_fuel_consumption)	2.76	0.38		2.47	0.41							
s(vol_per_dwt)	7.67	0.00	***	7.45	0.00	***	8.28	0.00	***	8.18	0.00	***
s(norm_admiralty_constant)												
Result	R-sq.	GCV	n	R-sq.	GCV	n	R-sq.	GCV	n	R-sq.	GCV	n
	0.742	10.417	243	0.748	10.143	243	0.751	12.684	328	0.748	10.143	328

Figure F.6: Second Iteration for Panamax Vessels

F.2.5. Capesize Vessels

Iteration	Second								
Model	Capesize1			Capesize2			Capesize3		
Parametric	Est.	p-val.	sig.	Est.	p-val.	sig.	Est.	p-val.	sig.
(Intercept)	17.31	< 2e-16	***	17.31	< 2e-16	***	17.64	< 2e-16	***
scrubberY	0.45	0.75		2.10	0.19		2.56	0.11	
anti_foulingY	-1.55	0.30		-3.18	0.05	*	-4.69	0.00	**
japanY	1.90	0.11		0.57	0.64		-0.14	0.91	
chinaY									
southkoreaY									
bvY	6.74	0.00	***	6.77	0.00	***	6.95	0.00	***
ccsY	1.82	0.12		0.55	0.70		0.58	0.69	
nkkY	3.92	0.00	***	4.35	0.00	***	4.57	0.00	***
rest_classY	4.83	0.00	***	5.63	0.00	***	5.54	0.00	***
Smooth-term	edf	p-val.	sig.	edf	p-val.	sig.	edf	p-val.	sig.
s(age)	1.00	< 2e-16	***	1.00	0.00	***	1.00	<2e-16	***
s(tc)	4.98	0.00	**						
s(norm_tc)									
s(dwt)	3.934	0.00	**	1.000	0.90		1.000	0.99	
s(libor)	1.000	0.00	***	5.60	0.00	***	2.983	0.00	**
s(orderbook_all)	5.67	0.00	***				5.43	0.03	*
s(orderbook_type)				1.00	0.81				
s(vol_per_dwt)	8.51	0.00	***	5.322	0.03	*	5.56	0.03	*
s(norm_admiralty_constant)	1.00	0.20		1.00	0.11		1.00	0.15	
Result	R-sq.	GCV	n	R-sq.	GCV	n	R-sq.	GCV	n
	0.92	12.365	78	0.85	18.588	78	0.86	17.996	78

Figure F.7: Second and Third Iteration for Capesize Vessels

Iteration	Third								
Model	Capesize1			Capesize2			Capesize3		
Parametric	Est.	p-val.	sig.	Est.	p-val.	sig.	Est.	p-val.	sig.
(Intercept)	19.44	<2e-16	***	17.81	0.00	***	18.19	0.00	***
scrubberY	0.59	0.67		1.72	0.28		3.11	0.04	*
anti_foulingY	-1.32	0.38		-2.75	0.08	.	-4.75	0.00	**
japanY									
chinaY									
southkoreaY									
bvY	5.35	0.01	*	6.35	0.01	**	5.58	0.03	*
ccsY									
nkkY	3.32	0.05	*	4.98	0.01	*	4.63	0.02	*
rest_classY	3.50	0.02	*	5.40	0.00	**	4.97	0.01	**
Smooth-term	edf	p-val.	sig.	edf	p-val.	sig.	edf	p-val.	sig.
s(age)	1.00	< 2e-16	***	1.00	< 2e-16	***	1.00	< 2e-16	***
s(tc)	5.63	0.00	**						
s(norm_tc)									
s(dwt)	4.683	0.01	**	1.000	0.75		1.000	0.37	
s(libor)	1.582	0.00	**	6.31	0.00	***	2.742	0.00	***
s(orderbook_all)	5.166	0.00	***				5.895	0.00	**
s(orderbook_type)				1.00	0.97				
s(vol_per_dwt)	7.99	0.00	***	5.19	0.05	*	2.448	0.07	.
s(norm_admiralty_constant)									
Result	R-sq.	GCV	n	R-sq.	GCV	n	R-sq.	GCV	n
	0.921	12.419	78	0.860	17.691	78	0.862	17.072	78

Figure F.8: Third Iteration for Capesize Vessels

Iteration	Fifth			Sixth			Seventh (Final)		
Model	Capesize3			Capesize3			Capesize3 (Final)		
Parametric	Est.	p-val.	sig.	Est.	p-val.	sig.	Est.	p-val.	sig.
(Intercept)	18.22	0.00	***	18.12	0.00	***	17.70	0.00	***
scrubberY									
anti_foulingY	-3.67	0.02	*	-2.84	0.07	.			
japanY									
chinaY									
southkoreaY									
bvY	7.25	0.00	**	6.73	0.00	**	6.70	0.00	**
ccsY									
nkkY	4.76	0.02	*	4.88	0.01	*	5.20	0.01	**
rest_classY	5.30	0.00	**	5.33	0.00	**	5.38	0.00	**
Smooth-term	edf	p-val.	sig.	edf	p-val.	sig.	edf	p-val.	sig.
s(age)	1.00	< 2e-16	***	1.00	0.00	***	1.00	0.00	***
s(tc)									
s(norm_tc)									
s(dwt)									
s(libor)	5.52	0.00	***	6.74	0.00	***	7.03	0.00	***
s(orderbook_all)	4.443	0.28							
s(orderbook_type)									
s(vol_per_dwt)	5.323	0.01	*	5.174	0.01	*	5.047	0.01	**
s(norm_admiralty_constant)									
Result	R-sq.	GCV	n	R-sq.	GCV	n	R-sq.	GCV	n
	0.871	16.529	78	0.863	16.63	78	0.86	16.753	78

Figure F.9: Fourth until Sixth Iteration for Capesize Vessels

F.3. Smooth Terms Fitting Functions

Variable	*All*	*Handysize*	*Handymax*	*Panamax*	*Capesize*
Age	exp1	exp1	poly1	exp1	poly1
TC-Average	-	-	fourier1	-	-
TC-normalized	gauss2	fourier1	-	-	-
DWT	-	exp1	exp2	fourier1	poly2
LIBOR	poly1	-	poly1	-	poly3
Orderbook All	-	-	-	-	gaus2
Orderbook Type	-	-	-	-	-
Floor Strength	-	-	-	gauss3	-

Figure F.10: Summary of MATLAB Function used.

F.4. Initial Model Formulation

Element	Validity	Estimated equation	R^2
Intercept	-	263.1989	-
Age	>0	-6.090227E-01*Age^3 + 1.334744E+01*Age^2 - 1.038118E+02*Age + 4.999802E+02	0.9971
Age	>12	5.768023E-03*Age^4 - 5.161690E-01*Age^3 + 1.658816E+01*Age^2 - 2.538768E+02*Age + 1.544850E+03	0.9982
DWT	>0	-2.083713E-12*DWT^3 + 2.626840E-07*DWT^2 - 1.703817E-02*DWT + 4.610558E+02	0.9987
DWT	>72000	7.279959E-05*DWT - 1.889205E+02	0.0092
DWT	>138000	-1.015294E-02*DWT + 1.167099E+03	0.9943
DWT	>160000	1.242740E-03*DWT - 6.710918E+02	0.6718
TC-Rate	>0	6.747669E-03*TC	1.0000
LIBOR	>0	1.863405E+01*LIBOR	1.0000
Demand	>0	1.397652E-04*Demand^4 - 7.614145E-03*Demand^3 - 2.692865E-01*Demand^2 + 2.722780E+01*Demand - 3.682488E+02	0.9998
Demand	>40	-1.865088E+00*Demand + 2.286382E+02	0.8017

Figure F.11: Smooth-terms Formulation for Initial Models - All Vessels. Source: Pruyn[5]