Norwegian School of Economics
Bergen, Fall 2019

Demand Forecasting of Antarctic Krill Meal

An automatic model for comparison of time series methods

Miriam Slagnes Takseth & Tove Fotland Newermann

Supervisor: Mario Guajardo & Jonas Andersson

Master thesis, Economics and Business Administration

Major: Business Analytics

NORWEGIAN SCHOOL OF ECONOMICS

Acknowledgements

This Master's thesis was written as part of a Major in Business Analytics at the Norwegian School of Economics and concerns demand forecasting of Antarctic krill meal. Our choice of topic emerged as a result of our mutual interests for sustainability and programming.

Working on this thesis for the past months has been both challenging and rewarding. We have acquired knowledge about sustainability in the krill industry and have been able to apply the knowledge and experience we have accumulated throughout our years of studies.

We would like to thank our supervisor, Mario Guajardo, for valuable advice and guidance. We sincerely appreciate his close cooperation and availability in this research. We would also like to thank our co-supervisor Jonas Andersson for sharing valuable knowledge and expertise. His feedback and consultation has been crucial in working with this thesis. Lastly, we would like to thank Mats Tristan Tjemsland and his colleagues at Aker BioMarine for providing us with the data used in this thesis, as well as valuable insight about the krill market.

Norwegian School of Economics

Bergen, December 2019

Tove Fotland Newermann

Miriam Slagnes Takseth

Abstract

The world's population is growing faster than ever. As a consequence, it is challenging to maintain a sustainable food production to satisfy all needs. In recent years, krill has emerged as a viable and effective supplement, especially for fish- and animal feed. In an industry characterized by increasing demand and harvesting limitations, it is particularly interesting to investigate whether time series forecasting can be a useful tool to aid effective decision making and long-term strategic planning. Demand forecasting in the krill market is an area in which little previous research is attributed. However, research within related areas such as fisheries harvesting and food production have shown positive results from applying ARIMA and exponential smoothing models. This thesis therefore considers univariate demand forecasting of krill meal for twelve months ahead, applying both of these methods, as well as a combination of decomposition and exponential smoothing. We use historical sales data over a seven-year period from Aker BioMarine as a case study to test the accuracy of the proposed methods. This is done through an automatic model built using R, which chooses the best model from each method based on a variety of criteria. The performance of the models is evaluated using the mean absolute error and the mean absolute scaled error and compared to simple benchmarks. According to our results, the benchmarks seem to perform better than the more complex methods. However, the chosen models from the automatic modeling procedure generally yield a high forecasting error. The provided forecasts should therefore be interpreted by someone with expert knowledge about the krill market and the specific customer, in order to be useful for resource allocation and strategic planning purposes. Since the chosen models do not give satisfying results in terms of forecast error, this opens an opportunity for further research within demand forecasting of krill meal.

Keywords **– Demand forecasting, time series, krill, krill meal, ARIMA, exponential smoothing, ETS, decomposition, STL**

Contents

List of Figures

List of Tables

1 Introduction

1.1 Motivation

Today, the world's population is growing faster than ever. This is mainly due to medical advancements and increased agricultural productivity, and by 2050 the population will most likely have reached 10 billion (United Nations, 2019). This means that, in order to meet all needs, we need to significantly increase our food production. At the same time, food production imposes serious environmental consequences for our planet. Depletion and contamination of natural resources occur throughout the agricultural food chain (Baldwin, 2015).

Krill has high nutritional value and positive effects on both the growth and health of fish. This implies that krill may be invaluable as the demands on food production continue to increase. Krill fishery is the only reduction fishery in the world with a biomass rated as in very good condition (Aker BioMarine, 2018). Krill has an estimated biomass of 379 million tons (Atkinson, Siegel, Pakhomov, Jessopp & Loeb, 2009), which makes it one of the species with the largest total biomass. The massive biomass makes it possible to harvest a large amount of krill while still ensuring sustainable utilization of this resource.

The krill industry is relatively young, and demand has been increasing the past decade (Bender, 2006). Since demand forecasting is an important tool for effective decision making, it is especially interesting to investigate whether it could be applied to the krill industry. This is reinforced by the fact that krill is a sustainable alternative to fish meal in feed production for aquaculture. This is mainly due to the large biomass, in combination with early implementation of harvesting regulations that ensure that commercial harvesting does not have a negative impact on either krill as a species, or other parts of the Antarctic marine ecosystem.

The krill market has been developed by Aker BioMarine the past ten years. This makes it challenging to forecast demand, as the market is rising and developing in line with the company. This is especially emphasized in Aker BioMarine's customers, as they offer a premium product in a global aquaculture industry, and are thus able to sell everything they produce in the long run. Aker BioMarine wants to take part in solving the problems

that follow increased food production and has a mission to improve human and planetary health. They are continuously working to lower their CO_2 emissions and act as an environmentally responsible producer of marine ingredients (Aker BioMarine, 2018). In order to keep supplying krill in a sustainable manner, demand forecasting can therefore be a useful tool.

Considering their global position, it is especially interesting to forecast demand at a disaggregated customer level, as the various customers may have different demand patterns. Aker BioMarine has a variety of customers, all from sole proprietorships to large global companies all over the world, which can result in different purchasing patterns that may interfere with patterns at an aggregated level. In addition, demand forecasting can contribute to higher quality on sales- and financial forecasts and can be used as supplementary guidance to the sales force. This emphasizes the need for forecasts at a disaggregated customer level. Knowledge and information about future demand per customer are useful and important for allocation of resources and harvested volumes, as well as for both tactical and strategic planning. The global spread in the customer portfolio also motivates our choice of building an automatic forecasting model, as this makes it easy to extract forecasts for a certain customer and use this information to make better and effective decisions, both with regards to the respective customer and the company in question.

Another challenge is that the krill population is very variable from year to year (Atkinson et al., 2009), and there is scientific uncertainty about the size of the biomass and also the effect of krill harvesting on the biomass (Bender, 2006). Considering the critical role of krill in the Antarctic ecosystem, this makes the development of good forecasting tools important for this industry. Krill meal is an attractive product because it can contribute to more efficient utilization of food resources, hence improving the productivity and environmental performance of aquaculture. Thus, krill meal is a sustainable, nutritional solution for the aquaculture industry. Aquaculture has had an impressive growth rate for the past decades (Msangi et al., 2013), which makes the use of time series relevant, as trend is a time series feature that can be extracted through time series forecasting methods.

1.2 Research Question

Based on the above discussion of the need for demand forecasting in the krill meal market, we have formulated the following research question:

> *To what extent can common time series forecasting methods, implemented in an automatic model, produce accurate forecasts of future demand for krill meal at a disaggregated customer level?*

In order to answer this research question, we will explore some common methods for time series demand forecasting, hereunder exponential smoothing, a combination method of decomposition and exponential smoothing, and ARIMA. These methods will be implemented in an automatic model in order to produce forecasts of future demand for krill meal per customer for Aker BioMarine. We will use data from Aker BioMarine as a case to discuss the performance of these methods in the krill market. With this data as a basis, we will try to determine to what extent the different methods are able to produce reasonable forecasts for Aker BioMarine and the industry.

First, we will give a brief introduction on krill harvesting and the krill industry, followed by motivation for the choice of forecasting methods based on previous research relevant to the industry. Following this, we choose to elaborate on literature relevant to the chosen forecasting methods. Thereafter we will explain the automatic modeling procedure where these methods are applied, for then to present the results. Finally, we will discuss the findings and limitations of the modeling procedure, before we provide our conclusion. There is, to our knowledge, no previous research on demand forecasting of krill meal, which amplifies the relevance of this thesis.

2 Background

2.1 Krill and Krill Harvesting

Krill are small shrimp-like crustaceans found in all the oceans. Krill is near the bottom of the food chain and is hence an important trophic level connection. They feed on phytoplankton and some zooplankton and are a suitable form of nourishment for many larger species. This makes it all the more important to ensure sustainable utilization of this biological resource. There are different Arctic and Antarctic species of krill. Antarctic krill is among the species with the largest total biomass and is an important part of the Antarctic marine food chains. Antarctic krill is mostly eaten by whales, seals, penguins, squid, birds and fish (Støp−Bowitz & Sømme, 2017). Krill are packed full of the essential fatty acids omega-3 EPA and DHA (eicosapentaenoic acid and docosahexaenoic acid, respectively). These fatty acids are some of the most researched nutrients and provide health benefits for the heart, eyes, liver and brain, to name a few. The omega-3s in krill are mainly bound to phospholipids which helps the fatty acids integrate into the cell membranes; an advantage compared to e.g. fish oil, where the omega-3s are bound to triglycerides (Burri, Hoem, Banni & Berge, 2012). Further, krill is packed with protein and works as a growth accelerator for shrimp and fish (Aker BioMarine, 2016).

In the last decade, almost 60 percent of total catch has been done by Norway, followed by Korea and China with 17 and 12 percent, respectively. Since the start of commercial krill fishery in the early 1960s, the location of fishing has moved from being mainly in the Indian Ocean to being almost entirely in the Southern Ocean. The last decade, the fishery has become focused in the areas around the South Antarctic (CCAMLR, 2018).

In order to prevent fishing that will negatively impact krill or other species in the ecosystem, all catches of Antarctic krill must be reported to the Commission on the Conservation of Antarctic Marine Living Resources (CCAMLR). The catch and effort reporting occurs on a monthly basis until 80% of the permitted seasonal catch is harvested. Upon reaching this limit, the reporting occurs more frequently for the remainder of the triggered season (CCAMLR, 2018). The management of krill fishery is very robust; the consensus of 25 governments is needed to change any of the fishery regulations in the Antarctic (Aker

BioMarine, 2016).

Krill harvesting is currently concentrated in the South Antarctic where the estimated krill biomass is approximately 60 million metric tons. The total allowable catch is 620,000 metric tons annually, which corresponds to around 1 percent of the stock biomass in this area. For the past years, the annual amount of harvested krill has been around 300,000 metric tons. For 2018/2019 (until September 2019), the total catch reported was 380,000 (CCAMLR, 2018, 2019). This leaves over 99 percent of the biomass for other predators. Harvesting far below the precautionary limits is one of many important measures to make krill harvesting a sustainable alternative to meet the present and future environmental challenges. This makes it all the more interesting to investigate the possibilities of demand forecasting within this industry, in order to ensure a sustainable harvest and preserve stock biomass.

2.1.1 The Krill Industry

Krill is a much used ingredient for aquaculture and animal feed, among others. A large part of the krill industry is therefore included in the aquaculture industry, also comprising products made from fish for reduction caught in the sea, as well as fish waste from the fish industry (Nielsen & Olesen, 2003). Both fish meal and fish oil suffer from price issues along with sustainability concerns. Krill meal and oil can therefore be a good supplement and substitution, as a study shows improved fish growth when krill is added to the feed (Dalsegg, 2018). Further, only a small part of the harvested krill is used in products for human consumption. The majority is therefore used in aquaculture, and krill is just a minor part of all the ingredients used to produce various types of feed. Feeds containing krill give many health benefits for fish and other pets and solve challenges faced in aquaculture. For example, studies have shown that fish develop stronger heart muscles and healthier circulatory systems by eating krill. This can again result in lower mortality and less disease, in addition to improved fillet quality (Aker BioMarine, 2016), which increases the demand for krill. The krill industry is relatively young and small, and there is uncertainty regarding the biomass and the environmental effects of krill harvesting (Bender, 2006). Further, krill is a biological resource, which makes the amount of krill harvested constrained by the amount of krill in the ocean at the time of harvest. However, the future prospects of demand are positive (United Nations, 2019), which emphasizes

the importance of developing good models to forecast future demand, in order to allocate the available resources and thus increase profit.

2.1.2 Aker Biomarine

Aker BioMarine, hereafter denoted ABM, was established as an independent enterprise in 2006 on the basis of Aker ASA's krill and fishing operations (Aker ASA, 2018). ABM's core business involves harvesting, production, sales and marketing of krill-based products for aquaculture, animal feed applications, dietary supplements and pharmaceutical markets (Aker BioMarine, 2018). This places them in the fish meal market, where they offer a premium product and hold a small piece of total market share. Therefore, it can be assumed that all harvested krill are sold in the long-term. At the same time, they have always been concerned with protecting the krill biomass as well as the many species that ultimately depend on krill as a food source (Aker BioMarine, 2016). Due to this, they always harvest within precautionary limits. Krill harvesting has traditionally relied on trawl nets, which has resulted in unwanted by-catch of other species. This is and has been a significant challenge for a fragile marine ecosystem in the Antarctic. Over the last decade, ABM has therefore made major investments in order to develop their Eco-Harvesting technology. This is a trawl system that conveys krill onboard the vessels for processing while a submerged trawl module minimizes by-catches (Aker BioMarine, 2018). As discussed in section 1.1, demand forecasting of krill meal can be a useful tool for ABM, and similar actors, in allocation of harvested volumes as well as allocation of company resources in order to maintain a stable supply and adhere to sustainability targets.

2.2 Literature Review

There is a tremendous amount of research conducted within the field of demand forecasting. We will therefore use this section to present a brief overview of literature with focus on demand forecasting particularly relevant to the krill industry. In section 3.1 we will elaborate on literature relevant to the different forecasting methods used in this thesis.

A quick search for "demand forecasting" on Google Scholar gives more than two million search results and the same search term gives more than one hundred thousand research

articles at ScienceDirect. The energy sector, emergency resources, tourism and the food industry are just some of the areas where studies on demand forecasting have successfully been applied. Several different forecasting techniques have been used, where ARIMA and exponential smoothing models are very popular in many different areas. In the energy market, there has for example been done a substantial amount of research on several forecasting techniques in order to forecast future energy needs. Among these are time series regression, ARIMA and neural networks. For example, research shows that ARIMA models can contribute to improved accuracy of both short- and long-term energy demand forecasting (Suganthi & Samuel, 2012).

Holguín-Veras & Jaller (2012) also show that it is possible to estimate robust ARIMA models to forecast resource needs after disasters. In the work by Da Veiga, Da Veiga, Catapan, Tortato & Da Silva (2014), the performance of ARIMA and Holt-Winters models are compared when forecasting demand for dairy products. Their research concludes that the preferred method is the Holt-Winters method, which is a popular exponential smoothing method. However, for this method, they recommend to not exceed the seasonal cycle of the series for the forecast horizon. Further, Barbosa, Christo & Costa (2015) used some versions of exponential smoothing methods for demand forecasting for production planning in a food company. They concluded that the Holt-Winters method was effective for forecasting demand for products that present trend and seasonality patterns in sales history. In addition, they highlighted the method's simplicity and accessibility due to its low cost and easiness. Another research in the food industry was conducted by Tirkes, Güray & Celebi (2017), who compared performance between trend analysis, decomposition and Holt-Winters models to forecast demand for jam and sherbet products. Holt-Winters models obtained good results in this case as well. The decomposition models performed satisfactorily.

If a forecasting method's performance is not better than a simpler alternative, the method is not worth considering (Hyndman & Athanasopoulos, 2018). Simple forecasting methods are therefore often used as benchmarks when using more complex methods like ARIMA and exponential smoothing. The research of Athiyaman & Robertson (1992) is one example where simple forecasting methods outperformed the more complex ones. They used the simple forecasting method, naïve, as well as moving average and some versions

of exponential smoothing, to forecast international tourist arrivals from Thailand to Hong Kong. They concluded that simple forecasting techniques often outperform more complex ones in terms of accuracy, time- and cost-effectiveness.

The krill industry is a quite unique industry which is hard to compare to other, larger industries. Krill is harvested and processed and then used as an ingredient in several products like fish food, dietary supplements and various animal foods. Therefore, it might seem suitable to investigate previous research done on forecasting demand within fields like dietary supplements, animal foods and aquaculture, hereunder especially salmon farming, to see if demand forecasting methods have successfully been applied as a tool within these. There has been attributed a lot of research on forecasting to the fish industry, especially fisheries forecasting, i.e. forecast of fish harvesting. For example, Stergiou (1991) forecast catches of Trachurus from the eastern Mediterranean (Greek waters) by using the Winters seasonal exponential smoothing method, ARIMA and monthly averages corrected for linear trend. He used the naïve method as a benchmark. The study resulted in the conclusion that ARIMA was far superior compared to both the benchmark and the other more complex forecasting methods. However, Stergiou (1991) pointed out that, in the short-term, Winters seasonal exponential smoothing method may be of potential in fisheries forecasting. Stergiou (1989) also performed a study on ARIMA models for forecasting the fishery for pilchard in Greek waters, and came to the same conclusion: ARIMA models result in good forecasts for this industry. In addition, the work by Saila, Wigbout & Lermit (1980) showed that the ARIMA method is preferable when forecasting monthly average catch per day fished for rock lobster. Here, ARIMA was compared to the monthly averages method and harmonic regression analysis.

Most of the research within the fishery industry has been applied to fisheries with long data sets. However, Prista, Diawara, Costa & Jones (2011) did a study on the use of seasonal ARIMA models to assess data-poor fisheries. They only had a sample size of 60 observations and found that seasonal ARIMA models may provide better forecasts than many multivariate models. They therefore suggest that seasonal ARIMA models "should be more widely considered to extend the coverage of monitoring to all exploited marine resources" (Prista et al., 2011, p. 171). On the other hand, when Czerwinski, Gutiérrez-Estrada & Hernando-Casal (2007) evaluated short-term catch per unit effort

capacity forecast for Pacific halibut, the ARIMA model's performance was insufficient, while the neural network model provided far superior forecasts. In addition, the work by Tsai & Chai (1992) showed that other methods performed better than ARIMA when forecasting striped bass commercial harvest in the Maryland portion of Chesapeake Bay. However, none of the methods in this study were satisfying in terms of forecast error.

The above research of fishery forecasting is not quite comparable with demand forecasting of krill meal. However, they are somehow related because of their aquaculture similarities. In ABM's supply chain, krill harvesting is the step prior to sales of krill products. Even though ABM is restricted by catch limitations, accurate demand forecasting is interesting and important for resource allocation and strategic planning, as well as financial planning and risk reduction. It could therefore be interesting to see if some of the models above can also be successfully implemented in demand forecasting of krill meal. So far, there are, to our knowledge, little research on demand forecasting within these fields, which makes the contribution of this thesis all the more relevant.

Today, the importance of handling a great amount of data for accurate analysis has become more significant in terms of survival in a global market. Estimating models for forecasting can be a time consuming and complex procedure. More automatic forecasting procedures can therefore lead to lower costs in a company that faces such challenges. Anvari, Tuna, Canci & Turkay (2016) have developed a framework that has shown to be both effective and accurate in forecasting time series regardless of the application sector. This framework is automated and uses a number of statistical tests to substitute human judgment and applies comprehensive tests to select an accurate model. Their research finds that their proposed framework gives higher accuracy than many other models. There have also been done a large amount of research on the demand structure for fish and seafood products (Asche, Bjørndal & Gordon, 2007). This does however not extend to forecasting demand for such. Within seafood production, seasonal forecasting has proven useful, and Hobday, Spillman, Paige Eveson & Hartog (2016) also argues that the use of seasonal forecasting can be extended to other areas. This makes it especially interesting to investigate the possibilities for automatic demand forecasting within the krill industry.

3 Methodological Framework

Forecasting can be defined as "predicting the future as accurately as possible, given all of the information available, including historical data and knowledge of any future events that might impact the forecasts" (Hyndman & Athanasopoulos, 2018, Ch. 1.2). Forecasting is used to help inform decisions and can be useful in long-term strategic planning. The time horizons could be anything between a few seconds and decades ahead. Forecasting can be extremely difficult in many cases and several factors affect the predictability of an event or quantity. Among these are how well we understand the factors that affect it, the amount of data available and whether the forecasts can affect what we are trying to forecast (Hyndman & Athanasopoulos, 2018).

Good forecasts are those who can capture the essence of historical data in terms of genuine patterns and relationships, while not replicating past events that are unlikely to occur again (Hyndman & Athanasopoulos, 2018). There are a variety of different forecasting methods that can be used. Choice of method depends on the purpose of the forecast and the importance of forecast accuracy. Some methods are simple, such as using the most recent observation as a forecast for the next period. Others are highly complex, like neural networks that capture patterns that may be hard or impossible to detect for the human eye. Sometimes there are plenty of historical data, while other times there are no data at all.

For demand forecasting of krill meal, time series data will be used, since the ordering of the observations conveys important information, and patterns over time may be important to forecast what is going to happen next. In this thesis, we will therefore look further into some time series forecasting methods. Throughout this thesis, we will denote the forecast by $\hat{y}$, while y will denote realized demand. In section 3.1 we will take a closer look at some common forecasting methods. Further, we will discuss some important data features in section 3.2, and then elaborate on a selection of evaluation criteria that can be used to compare alternative models and how the most appropriate one can be chosen in section 3.3.

3.1 Forecasting Methods

Forecasting methods can be divided into two main categories: qualitative and quantitative. Qualitative forecasting is used when there are no data available or the available data is not relevant. Qualitative forecasting often implies judgmental forecasts, which can be both useful and accurate when the forecaster has important domain knowledge and a lot of available information. On the other hand, quantitative forecasting can be used when there exists numerical information about the past, at the same time as it is reasonable to assume that some aspects of the past patterns will continue into the future. Quantitative forecasting implies statistical methods based on historical data, e.g. time series data. When data are available, it is preferable to use quantitative and statistical methods, as these are generally superior to generating forecasts using only human judgment (Hyndman & Athanasopoulos, 2018). Statistical methods will therefore be the focus of this thesis.

Time series refers to observations on a variable that is observed sequentially over time (Pankratz, 1983). When forecasting time series we aim to estimate how the sequence of observations will continue into the future (Hyndman & Athanasopoulos, 2018). There are simple and more complex forecasting methods. Some methods only use information on the variable to be forecast, and disregard factors that affect its behavior. These methods extrapolate trend and seasonal patterns, but ignore other information about surrounding factors (Hyndman & Athanasopoulos, 2018). In the following sections, we will describe simple forecasting methods, exponential smoothing methods, a combination method of decomposition and exponential smoothing and the ARIMA method.

3.1.1 Simple Forecasting Methods

Some forecasting methods are simple, but effective, and are often used as benchmarks. Among these are the average method, the naïve method and the seasonal naïve method. The average method produces forecasts that, as the name indicates, are equal to the average or mean of the historical data. The naïve method sets all forecasts to be equal to the value of the last observation. An extension of the naïve method is the seasonal naïve method, which is useful for highly seasonal data. Since we denote the forecast by $\hat{y}$, the forecast for time $T + h$ can be written as

$$\hat{y}_{T+h|T} = y_{T+h-m(k+1)}, \tag{3.1}$$

where m is the frequency of the seasonal period, k is the number of complete years in the forecast period prior to time $T + h$, and h is the forecast horizon. Each forecast is set to be equal to the last observed value from the same season of the year. For monthly data this could for example mean that the forecast for all future February values is equal to the last observed February value (Hyndman & Athanasopoulos, 2018).

One of these three methods will often be the best forecasting method available. However, in many cases, they will serve as benchmarks rather than the method of choice. The more advanced forecasting methods will therefore be compared to these simple methods to ensure that the new method is better than the simple alternatives. If the more complex methods are not better, they are not worth considering (Hyndman & Athanasopoulos, 2018).

3.1.2 Exponential Smoothing

As discussed in section 2.2, several exponential smoothing methods have successfully been applied in e.g. the food industry, for tourist arrival and to some extent in fisheries forecasting. In this section, we will therefore elaborate on how such methods work.

Exponential smoothing methods use weighted averages of past observations to forecast new values. The weights decrease exponentially as the observations get older, which means that more recent observations are assigned higher weight. The advantage of exponential smoothing is that it generates reliable forecasts quickly and for a wide range of time series (Hyndman & Athanasopoulos, 2018). Exponential smoothing is especially useful when long-term forecasting is desired and it is unlikely to be worthwhile to fit a complicated model (Chan, 2002).

The simplest of the exponential smoothing methods is called *simple exponential smoothing* and is suitable for forecasting data with no clear trend or seasonal pattern. To illustrate how the forecasts are calculated using weighted averages, we look at the following equation

$$\hat{y}_{T+1|T} = \alpha y_T + \alpha(1-\alpha)y_{T-1} + \alpha(1-\alpha)^2 y_{T-2} + ..., \tag{3.2}$$

where α is the smoothing parameter and has a value between 0 and 1. The forecast for the next period, $T+1$, is a weighted average of all previous observations in the time series. α controls the rate at which the weights decrease, where a small value gives more weight to observations from the distant past. If α is close to 1, more weight is given to more recent observations (Hyndman & Athanasopoulos, 2018).

There are two equivalent forms of simple exponential smoothing: *weighted average form* and *component form*. Both lead to forecast equation (3.2). We will continue with the *component form*, which can be written as follows for simple exponential smoothing,

$$\begin{aligned} &\text{Forecast equation} && \hat{y}_{t+h|t} = \ell_t \\ &\text{Level equation} && \ell_t = \alpha y_t + (1-\alpha)\ell_{t-1}, \end{aligned} \tag{3.3}$$

where ℓ_t is the level of the series at time t. The level component is the only component included in simple exponential smoothing. However, more complex models can also include a trend component and/or a seasonal component. When looking at the forecast equation, we see that the forecast value at time $t+h$ is the estimated level at time t. The estimated level of the series at each period t is given by the level equation (Hyndman & Athanasopoulos, 2018).

When applying exponential smoothing methods, the smoothing parameters and the initial values must be chosen. The most reliable and objective way to obtain these is to estimate them from the observed data. For any exponential smoothing method, this can be done by minimizing the sum of squared residuals. Alternatively, the parameters can be estimated by maximizing the likelihood. The likelihood is the probability of the data arising from the specified model (Hyndman & Athanasopoulos, 2018). Maximum likelihood estimation can therefore be defined as estimating parameters from sample data such that the probability of obtaining the observed data is maximized. It is common to work with the logarithm of the likelihood function. As a general principle, the maximum of the log-likelihood function can be found with pretty much any valid approach for identifying the arguments of the maximum, as this is an unconstrained non-linear optimization problem (Harvey, 1993).

3.1.2.1 ETS

Exponential smoothing models combine *error, trend and seasonal components* in a smoothing calculation, and are therefore often referred to as ETS models. An ETS model is a *state space model*, which means that it "consists of a measurement equation that describes the observed data, and some state equations that describe how the unobserved components or states (level, trend, seasonal) change over time" (Hyndman & Athanasopoulos, 2018, Ch. 7.5). This means that an ETS model includes both a forecast equation and some smoothing equations for each component.

Each component in an ETS model has different possibilities. There can be no trend(N), an additive trend(A) or a damped additive trend(A_d), no seasonality(N), additive seasonality(A) or multiplicative seasonality(M). An additive trend indicates an increasing or decreasing trend, while a damped additive trend "dampens" the trend so that it diminishes in the long-run forecasts. The errors can either be additive(A) or multiplicative(M). Models with multiplicative errors are not numerically stable when the data is not strictly positive. This means that when the data contains zeros or negative values, multiplicative models should not be considered.

For a model with additive seasonality, the seasonal component is expressed in absolute terms in the scale of the observed series. The series is seasonally adjusted by subtracting the seasonal component in the level equation, which causes the seasonal component to add up to approximately zero each year. However, for a model with multiplicative seasonality the seasonal component is expressed in relative terms. The series is seasonally adjusted by dividing through by the seasonal component. This results in a seasonal component that adds up to the frequency of the seasonality m each year (Hyndman & Athanasopoulos, 2018). The following model illustrates an ETS-model including all the components

$$\begin{aligned} y_t &= (\ell_{t-1} + b_{t-1})s_{t-m}(1 + \varepsilon_t) \\ \ell_t &= (\ell_{t-1} + b_{t-1})(1 + \alpha\varepsilon_t) \\ b_t &= b_{t-1} + \beta(\ell_{t-1} + b_{t-1})\varepsilon_t \\ s_t &= s_{t-m}(1 + \gamma\varepsilon_t), \end{aligned} \tag{3.4}$$

where ℓ_t is the level of the series, b_t is the slope, s_t is the seasonal component of the series,

and ε_t is the residual, all at time t. α, β and γ are the smoothing parameters (Hyndman & Athanasopoulos, 2018). This is an ETS(MAM) model, which includes multiplicative error, additive trend and multiplicative seasonality.

3.1.2.2 Combination Method: STL + ETS

ETS can also be combined with other methods. For example, Hyndman & Athanasopoulos (2018) states that a combination of STL decomposition and ETS, usually produce quite good forecasts for seasonal time series. Some advantages of STL are that it can handle any type of seasonality, not just monthly and quarterly. It can also be robust to outliers so that occasional unusual observations will not affect the estimates of the trend-cycle and seasonal components. We have therefore chosen to investigate this method as well.

STL is a decomposition method, and is an acronym for *"Seasonal and Trend decomposition using Loess"* (Hyndman & Athanasopoulos, 2018). Loess is a modeling method for estimating flexible nonlinear relationships, which is done by utilizing the simplicity of linear least squares regression. The method was originally proposed by Cleveland (1979), who gives a detailed explanation of this method in his paper *"Robust Locally Weighted Regression and Smoothing Scatterplots"*. Loess fits simple models to localized subsets of the data to build a function that describes the deterministic part of the variation in the data. This is done for each point in the data by using explanatory variable values near the point whose response is being estimated, and fitting a low-degree polynomial to a subset of the data (Guthrie, Filliben & Heckert, 2003).

Since STL is a decomposition method, we will do a brief explanation of classical decomposition. When decomposing a time series, we divide it into three components: a trend-cycle component, a seasonal component and a remainder component containing anything else in the times series. After decomposition, there should be no pattern in the error term. A trend exists if there is a long-term increase or decrease in the data. Seasonality is present when a time series is affected by seasonal factors like the time of the year or day of the week. Seasonality is always of a known and fixed frequency (Hyndman & Athanasopoulos, 2018).

There are two forms of decomposition: additive and multiplicative, which can be written as follows

$$\begin{aligned} &\text{Additive} && y_t = S_t + T_t + R_t \\ &\text{Multiplicative} && y_t = S_t \times T_t \times R_t, \end{aligned} \tag{3.5}$$

where y_t is the time series, S_t is the seasonal component, T_t is the trend-cycle component, and R_t is the remainder component, all at time t. The variation in the seasonal pattern and around the trend-cycle determines whether to use additive or multiplicative decomposition (Hyndman & Athanasopoulos, 2018).

Additive decomposition consists of computing the trend-cycle component $\hat{T}_t$ by averaging the values within the frequency of the time series. For monthly series, the series is divided into subsets that each includes 12 observations, e.g. one observation for each month. Then the average of each subset is calculated. This eliminates some of the randomness in the data because observations nearby in time are likely to be close in value (Hyndman & Athanasopoulos, 2018).

After computing the trend-cycle component, you must calculate the detrended series $y_t - \hat{T}_t$. The third step is to estimate the seasonal component $\hat{S}_t$ by averaging the detrended values for that season. These seasonal component values are then adjusted to add to zero. Then the monthly values are stringed together, and this sequence replicated for each year of data to obtain the seasonal component. Lastly, the remainder component, $\hat{R}_t$, is calculated by subtracting the estimated seasonal and trend-cycle components: $y_t - \hat{T}_t - \hat{S}_t$.

For multiplicative decomposition, the process is similar, except that all subtractions are replaced by divisions. Also, for the seasonal component, the monthly indexes are stringed together to add to m. The remainder component for multiplicative decomposition is calculated by dividing out the estimated seasonal and trend-cycle components: $\hat{R}_t = y_t/(\hat{T}_t \times \hat{S}_t)$. The decomposed time series for the additive and multiplicative time series can therefore be written as

$$y_t = \hat{S}_t + \hat{A}_t \tag{3.6}$$

$$y_t = \hat{S}_t \times \hat{A}_t, \tag{3.7}$$

where $\hat{S}_t$ is the seasonal component and $\hat{A}_t = \hat{T}_t \times \hat{R}_t$ the seasonally adjusted component for multiplicative decomposition, and $\hat{A}_t = \hat{T}_t + \hat{R}_t$ for additive decomposition. When

forecasting a decomposed time series, we forecast the seasonal component and the seasonally adjusted component separately. Usually, we assume that the seasonal component is either unchanging or changing extremely slowly. The seasonal component is therefore forecast using the seasonal naïve method ($\hat{y}_{T+h|T} = y_{T+h-m(k+1)}$), where each forecast is equal to the last observed value from the same season of the year. The remaining components, trend and error, constitute the seasonally adjusted component. This component is used to fit and forecast a non-seasonal ETS model. Finally, the forecasts from the seasonal component and the seasonally adjusted component are combined (Hyndman & Athanasopoulos, 2018).

3.1.3 ARIMA

The ARIMA method aims to describe the autocorrelations in the data (Hyndman & Athanasopoulos, 2018). A good ARIMA model therefore describes how observations in a single time series are statistically related to past observations in the same series (Hyndman & Athanasopoulos, 2018). In section 2.2, we argued that ARIMA models perform well in several cases of fisheries forecasting, even in some cases of poor data, and that ARIMA models also can contribute to improved accuracy of energy demand forecasting. We will therefore elaborate on non-seasonal and seasonal ARIMA models in this section.

ARIMA models are another name for "Univariate Box-Jenkins" or UBJ models. Univariate means "one variable" and refers to that UBJ or ARIMA forecasts are based on only one variable: past values of the variable being forecast. ARIMA models are more suitable for short-term forecasting because they place more emphasis on observations in the recent past rather than the distant past. When building ARIMA models, it is necessary to have an adequate sample size. This will be further discussed in section 3.2.2.

3.1.3.1 Stationarity and Differencing

The first step when applying the ARIMA method is to check for stationarity in the data. If the data are non-stationary, differencing is applied to make it stationary. A stationary time series can be defined as "one whose properties does not depend on the time at which the series is observed" (Hyndman & Athanasopoulos, 2018, Ch. 8.1). This means that the time series has a mean, variance and autocorrelation that are constant through time (Pankratz, 1983). A stationary time series should look pretty much the same at any point

in time and have no predictable patterns in the long-term (Hyndman & Athanasopoulos, 2018).

An example of a non-stationary time series is a random walk process, where the slope coefficient ϕ equals 1 and y_t is a function of the previous values y_{t-1}. A random walk can be written as

$$y_t = c + \phi y_{t-1} + \varepsilon_t = c + y_{t-1} + \varepsilon_t, \tag{3.8}$$

where c is some constant and ε_t is the error term at time t. This implies uncertainty because of non-constant variance, hence the series is non-stationary. One common and simple transformation that can render a non-stationary series stationary, is differencing. Differencing involves calculating the successive changes in the values of a time series. Differencing can therefore stabilize the mean of a series by removing changes in the level of the times series, and in that way remove or reduce trend and seasonality (Hyndman & Athanasopoulos, 2018). A differenced time series can be written as

$$y_t^{'} = y_t - y_{t-1}. \tag{3.9}$$

Since it is not possible to calculate a difference for the first observation, the differenced time series will have $T - 1$ observations (Hyndman & Athanasopoulos, 2018). This series is called the first differences of y_t. If the series does not have a constant mean, we redefine $y_t^{'}$ as the first differences of the first differences. The series $y_t^{'}$ is now referred to as the second differences of y_t. Often it is sufficient with one difference to get a constant mean (Pankratz, 1983). In practice, it is rarely necessary with more than second differences (Hyndman & Athanasopoulos, 2018).

Another method is seasonal differencing, which works in a similar way as first- and second-order differencing. However, a seasonal difference is between an observation and the previous observation from the same season, and not between successive observations. Seasonal differencing can be written as

$$y_t^{'} = y_t - y_{t-m}, \tag{3.10}$$

where m is the number of seasons. Sometimes, a combination of first differences and

seasonal differences are necessary to achieve stationary data. There is some subjectivity in selecting which differences to apply, but if both differences first are applied, it does not matter which is done first. However, if the data have a strong seasonal pattern, it is recommended to do seasonal differencing first. This is since the resulting time series after seasonal differencing will sometimes be stationary and thus there will be no need for further first differencing (Hyndman & Athanasopoulos, 2018).

Whether differencing is required can either be determined by visual inspection of the estimated autocorrelation function (ACF) and partial autocorrelation function (PACF) or objectively through a *unit root test*. The estimated ACF and PACF measure the correlation between the observations within a single time series and are graphical tools used to identify patterns in the underlying data. They are used as guides when choosing one or more ARIMA models that seem appropriate as a starting point (Hyndman & Athanasopoulos, 2018).

3.1.3.2 Unit Root Tests

A unit root test checks if a time series is non-stationary and possesses a unit root, hence the name (Zivot & Want, 2006). The Dickey-Fuller (DF) test and the Augmented Dickey-Fuller (ADF) test are commonly used unit root tests. If we consider equation 3.8, the data are stationary as long as $|\phi| < 1$. However, if $|\phi| = 1$, the data are a random walk, hence there is a unit root, no pattern and the data are non-stationary. The null-hypothesis is therefore $H_0 : |\phi| = 1$, which is tested against the alternative hypothesis $H_1 : |\phi| < 1$. The regression model for the DF test can for example be written as

$$\begin{aligned} y_t - y_{t-1} &= c - (1-\phi)y_{t-1} + \varepsilon_t \\ \triangle y_t &= c + \delta y_{t-1} + \varepsilon_t. \end{aligned} \tag{3.11}$$

If $\delta = 0$, there is a unit root and the data are non-stationary. The hypotheses are therefore as follows

$$\begin{aligned} H_0 &: \delta = 0 \\ H_1 &: \delta < 0. \end{aligned} \tag{3.12}$$

The DF test then applies the ordinary least squares (OLS) method to find the estimator for ϕ, and the test statistic is given by (Maddala & Kim, 1998)

$$t_{\phi=1} = \frac{\hat{\phi} - 1}{SE(\hat{\phi})}. \tag{3.13}$$

The ADF test has the same basis as the DF test, but can also test for unit root for higher order processes. The regression model for the ADF test is defined by

$$\triangle y_t = c + \delta y_{t-1} + \sum_{i=1}^{p} \beta_i \triangle y_{t-i} + e_t, \tag{3.14}$$

where β is the lagged delta terms. The hypothesis is the same as for the DF test (3.12). The question is how many lags should be added? The more complicated the process, the more lags are needed. We therefore continue adding lags until we have no serial correlation in our error term ε_t (Maddala & Kim, 1998). We can use the same distribution as for the DF test; if the absolute value of the test statistic is lower than the DF critical value, we reject the null-hypothesis and differencing is necessary to make the time series stationary.

Another commonly used test for stationarity is the Kwiatkowski-Phillips–Schmidt–Shin (KPSS) test. It is worth noting that the KPSS test is in fact a *stationarity test*, opposed to *unit root tests* (Zivot & Want, 2006). In a stationarity test, the null hypothesis is that the data are stationary (Maddala & Kim, 1998). A small p-value therefore suggests that differencing is required (Hyndman & Athanasopoulos, 2018).

For seasonal time series, seasonal differencing might be necessary. For this purpose, there are some generalizations of the DF and KPSS framework from zero frequency to seasonal frequencies: The Hylleberg-Engle-Granger-Yoo (HEGY) test and the Canova Hansen (CH) test, respectively.

3.1.3.3 Non-Seasonal ARIMA

When the time series has been transformed to be stationary, we can proceed to fit an ARIMA model. ARIMA is an acronym for Autoregressive Integrated Moving Average (Hyndman & Athanasopoulos, 2018). Autoregression indicates regression of the variable against the variable itself. An autoregressive model of order p can be written as

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p} + \varepsilon_t. \tag{3.15}$$

We refer to this as an AR(p) model. This model is like multiple regression, except that the predictors are *lagged* values of y_t. ε_t represents white noise. Changing the parameters $\phi_1, ..., \phi_p$ will result in different time series patterns, while the variance for the error term ε_t will only change the scale of the time series. Autoregressive models are very flexible and can handle a wide range of different time series patterns (Hyndman & Athanasopoulos, 2018).

A non-seasonal ARIMA model is the combination of differencing, an autoregressive model and a moving average model. Moving average models are linear regressions on the current value of the time series and previously observed white noise error terms (Cowpertwait & Metcalfe, 2009). A moving average model of order q can be written as

$$y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + ... + \theta_q \varepsilon_{t-q}. \tag{3.16}$$

We refer to this as an MA(q) model. Each value of y_t can be seen as a "weighted moving average of the past few forecast errors" (Hyndman & Athanasopoulos, 2018, Ch. 8.4). As with autoregressive models, changing the parameters will result in different time series patterns, while the variance in the error term only changes the scale of the series. Since the lagged error terms in MA models are not observable, parameter estimation for an MA model is more difficult than for an AR model (Maddala & Kim, 1998).

The full ARIMA model can be written as a combination of an autoregressive model and a moving average model,

$$y_t^{'} = c + \phi_1 y_{t-1}^{'} + ... + \phi_p y_{t-p}^{'} + \theta_1 \varepsilon_{t-1} + ... + \theta_q \varepsilon_{t-q} + \varepsilon_t, \tag{3.17}$$

where $y_t^{'}$ is the differenced time series, which can have been differenced more than once. The right-hand side consists of both lagged values of y_t and lagged errors. This model is called an ARIMA(p,d,q) model where p represents the order of the autoregressive part, d represents the degree of first differencing and q the order of the moving average part.

Autoregression and moving average are actually just special cases of ARIMA models and can be written as ARIMA(p,0,0) and ARIMA(0,0,q), respectively (Hyndman & Athanasopoulos, 2018). To choose appropriate values for p,d and q is a difficult task. An important aspect when searching for a good model is however that we want a model with "the smallest number of estimated parameters needed to adequately fit the patterns in the available data" (Pankratz, 1983, p. 17). This means that if we have two models that perform equally well in terms of error, we prefer the simpler model with fewer parameters. The simpler ARIMA model is expected to be better because "it seems to be closer to the truth, has less probability of parameter redundancy, and is easier to fit and understand" (Anvari et al., 2016, p. 39).

The estimated coefficients of the model must satisfy certain mathematical inequality conditions, or else the model is rejected. The AR coefficients must satisfy some stationarity conditions: If $p = 1$, then $|\phi_1| < 1$. While if $p = 2$, then three conditions must be satisfied for the model to be stationary. First, $|\phi_2| < 1$, second $\phi_2 + \phi_1 < 1$ and lastly $\phi_2 - \phi_1 < 1$. Since we do not know ϕ_1 and ϕ_2 in practice, these conditions are applied to the estimates $\hat{\phi_1}$ and $\hat{\phi_2}$. Further, the MA coefficients must satisfy similar conditions of invertibility. Where $|\theta_1| < 1$ if $q = 1$. While if $q = 2$, then $|\theta_2| < 1$ and $\theta_2 + \theta_1 < 1$ and $\theta_2 - \theta_1 < 1$. The reason for the invertibility condition is that larger weights should be attached to more recent observations, while a non-invertible ARIMA model implies that weights put on past observations do not decline as we move further into the past (Pankratz, 1983).

3.1.3.4 Seasonal ARIMA

ARIMA models can also be useful in modeling seasonal data. The ARIMA method is based on the idea that by fitting an ARMA model to differenced observations, one can implicitly capture the non-stationary trend movements. This idea can be extended by supposing that evolving seasonality can be handled by the use of seasonal differencing, thus *seasonal ARIMA models* can be used to model seasonal data (Harvey, 1993). A seasonal ARIMA model is formed by including additional seasonal terms and can therefore be written as *ARIMA(p,d,q)*$(P, D, Q)_m$, where m is the number of time steps per seasonal period. The first parenthesis represents the non-seasonal part, while the last represents the seasonal part of the model. The seasonal part consists of similar terms as the non-seasonal part, but involves backshifts of the non-seasonal part (Hyndman & Athanasopoulos, 2018).

To remove additive seasonal effects, a seasonal ARIMA model includes differencing at a lag equal to the number of seasons (Cowpertwait & Metcalfe, 2009). In the same manner as lag one differencing is applied to remove trend, lag s differencing introduces a moving average term to the seasonal model. The modeling procedure for a seasonal ARIMA model is similar to the one for a non-seasonal ARIMA model, but we must also determine seasonal AR and MA terms, as well as the non-seasonal components of the model. If we consider a quarterly time series ($m = 4$) without a constant, a seasonal $ARIMA(1,1,1)(1,1,1)_4$ model can be written using backshift notation as

$$(1 - \phi_1 B)(1 - \phi_1 B^4)(1 - B)(1 - B^4)y_t = (1 + \theta_1 B)(1 + \Theta B^4)e_t, \tag{3.18}$$

where $By_t = y_{y-1}$ and $B^4 y_t = y_{t-4}$. The non-seasonal difference is represented in the third parenthesis in the equation and the seasonal difference is represented by the fourth. Further, the non-seasonal AR(1) is represented in the first part of the equation and the seasonal AR(1) by the second parenthesis. The MA(1) part is on the right-hand side of the equation, where the non-seasonal part is in the first parenthesis and the seasonal part in the other.

Since the AR- and MA components and the order of differencing all operate across multiple lags of s (number of seasons), seasonal ARIMA models can potentially have a large number of parameters. This makes it especially important to try out a wide range of models, and use an appropriate criterion to choose the best model (Cowpertwait & Metcalfe, 2009).

3.2 Data Features

The features of input data to any forecasting method can be crucial for the performance and accuracy of that method on given data. In the following two sections we will therefore discuss possible transformations of the data that can make the forecasting task simpler, as well as the importance of an adequate sample size.

3.2.1 Data Transformation

In many cases, adjustment of the historical data can lead to a simpler forecasting task. There are several types of possible adjustments, and the purpose of them all is to remove

known sources of variation or making the pattern more consistent across the whole data set. This is useful since simpler patterns usually lead to more accurate forecasts (Hyndman & Athanasopoulos, 2018). For example, if the variation in the data increases or decreases with the level of the series, a mathematical transformation may be useful.

Box-Cox transformation is a commonly used transformation method, which includes both logarithmic transformations and power transformations. Type of transformation to use is determined by the value of λ. To compute the appropriate λ for the data, one can use different methods. One possibility is Guerrero's method, which is "a model-independent method that is useful to select a power transformation that best stabilizes the variance of a time series variable" (Guerrero & Perera, 2004, p. 357). The Box-Cox transformation is defined as

$$w_t = \begin{cases} log(y_t) & \text{if } \lambda = 0; \\ (y_t^{\lambda} - 1)/\lambda & \text{otherwise} \end{cases} \tag{3.19}$$

Hyndman & Athanasopoulos (2018, Ch. 3.2) states that "If $\lambda = 1$, then $w_t = y_t - 1$, so the transformed data is shifted downwards, but there is no change in the shape of the time series. But for all other values of λ, the time series will change shape". This means that there is no need for a transformation of the data if λ is close to 1.

3.2.2 Sample Size

As mentioned in section 3.1.3, sufficient training data is essential for constructing good models. This is even more important when a large number of parameters must be estimated. Box and Jenkins (1976), referred in Pankratz (1983), suggests a minimum of 50 observations. However, Hyndman & Athanasopoulos (2018) argues that there is no "magic number" of minimum observations, and that number of observations required to fit a model depends on factors like the number of parameters to be estimated and the amount of randomness in the data. However, the fewer observations we have in the training data, the more likely we are to encounter overfitting. When the number of parameters to be estimated is high, overfitting is more likely (Quinn, McEachen, Fullan, Gardner & Drummy, 2019). Overfitting means that "the model performs well on the training data, but it does not generalize well" (Géron, 2019, p. 27). This happens when the model is too

complex for the data, in which a simpler model might be better.

First of all, statistically speaking, one should always have more observations than parameters to be estimated. Secondly, when estimating a model with data containing a lot of random variation, it is necessary to have a lot of data, while if the data have little variation, fewer observations may be sufficient (Hyndman & Kostenko, 2007). Further, Hyndman & Kostenko (2007) argues that exponential smoothing models require estimation of up to three parameters (smoothing parameters) for the level, trend and seasonal components of the data, as well as starting values for these. When dealing with seasonal data, there are also two parameters associated with the initial level and trend values and eleven parameters associated with the initial seasonal components. This means that with monthly data the theoretical minimum of observations is 17. With m seasons, one could therefore say that there are $m + 1$ initial values and three smoothing parameters, which means that there are a minimum of $m + 4$ parameters to be estimated. Thus, $m+5$ observations are the theoretical minimum of observations to estimate an exponential smoothing model. However, this is only sufficient when there is almost no randomness in the data, and realistically it is therefore necessary with substantially more data for most problems. For ARIMA models, the reasoning is similar; to estimate a seasonal ARIMA model, at least $p + q + P + Q + d + mD + 1$ observations are required (Hyndman & Kostenko, 2007).

3.3 Evaluation Criteria and Selection

When determining which model, within the same forecasting method, that is most appropriate for forecasting a given time series, several information criteria can be used. These criteria are used to compare models before forecasting and do not evaluate the actual forecasts. The information criteria can also not be used to compare models from different forecasting methods. To determine which of the above methods that produce the best forecasts, we must therefore evaluate forecast accuracy. Forecast accuracy must be calculated by evaluating model performance on new, unseen data. This means that the data that were used when fitting the model can not be used when evaluating forecast accuracy (Hyndman & Athanasopoulos, 2018). The time series is therefore divided into training and test data. The training data is used to estimate the model, while the test

data is used to measure the model's accuracy after forecasting.

3.3.1 Information Criteria

There are several information criteria that can be used to identify which model that performs best on a given time series. Three popular criteria are AIC, AIC_c and BIC. Hyndman & Athanasopoulos (2018) defines AIC, or *Akaike's Information Criterion*, as

$$AIC = T \times log\left(\frac{SSE}{T}\right) + 2(k+2). \tag{3.20}$$

Here, T is the number of observations used for estimation and SSE is the fit of the model. The $k+2$ part of the equation represents the number of parameters in the model. k is the number of predictors, while the other two parameters are the intercept and the variance of the residuals. The idea is to penalize the fit of the model with the number of parameters that need to be estimated (Hyndman & Athanasopoulos, 2018). When the sample size is small, i.e. T is small, AIC tends to select too many predictors and thus overfit. Therefore, the bias-corrected version, AIC_c has been developed. Minimizing one of these measures allows both the number of parameters and the amount of noise to be be taken into account. Hyndman & Athanasopoulos (2018) defines AIC_c as

$$AIC_c = AIC + \frac{2(k+2)(k+3)}{T-k-3}. \tag{3.21}$$

AIC_c is particularly useful for short time series and often leads to simpler models being chosen, since more than one or two parameters will produce poor forecasts due to estimation error (Hyndman & Athanasopoulos, 2018).

The third information criterion, BIC or *Schwarz's Bayesian Information Criterion*, imposes a stronger penalty for each additional parameter added to the model, than AIC and AIC_c. Further, BIC is a consistent criterion, which means that it determines the true model asymptotically. This means that BIC will select the true underlying model if the true underlying model is among the candidate models considered. AIC is not consistent under those circumstances. AIC is however efficient if the true model is not among the candidate models considered, in that it will asymptotically choose the model

which minimizes the error. Likewise, BIC is not consistent under those circumstances (Vrieze, 2012). Hyndman & Athanasopoulos (2018) defines BIC as

$$BIC = T \times log(\frac{SSE}{T}) + (k+2)log(T). \tag{3.22}$$

Choice of evaluation criterion is not always obvious and one could argue for one or the other. Several evaluation criteria can even be used simultaneously.

When the order of differencing is decided, the values of p and q are chosen by minimizing AIC, AIC_c or BIC, since all these have forecasting as their objective (Hyndman & Athanasopoulos, 2018). No one criterion will always outperform the others, and selection of evaluation criterion will therefore affect the final choice of model, as the different criteria may select different models. Box, Jenkins & Reinsel (1994) therefore suggested to only use the information criterion as supplementary guidance in addition to the visual inspection of the ACFs and PACFs.

When comparing different ARIMA models using AIC, AIC_c or BIC, all models must have the same order of differencing. Also, when selecting between models from the same method, AIC_c can be used to select an ARIMA model between candidate ARIMA models, but not to e.g. compare ARIMA and ETS models.

3.3.2 Time Series Cross-Validation

A commonly used split between the training and test data is to use 80% of the total data for training and 20% for testing. However, which distribution to use depends on the data and the number of observations. Since the test data is used to evaluate forecast accuracy, its size should be at least as large as the forecast horizon (Hyndman & Athanasopoulos, 2018).

However, when evaluating forecast accuracy, time series cross-validation can be useful for more accurate evaluation. This is another, more complex method for splitting the data into training- and test data. The estimated model can then be tested on several different distributions of training- and test data. However, when working with time series, the data can not be divided by random choice. The training data must include sequential observations, and the test data must include the consecutive observations or observations

several steps ahead. Many general cross-validation techniques, like k-fold, shuffles the data, which is not applicable to time series. One way of splitting the time series using cross-validation is the *expanding window method*. This method is illustrated in figure 3.1.

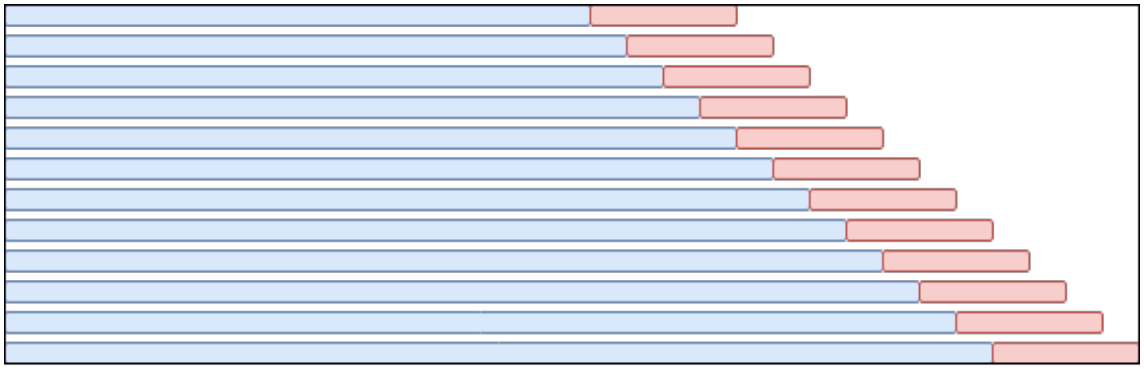

Figure 3.1: Illustration of expanding window method for time series cross-validation

Cross-validation can either be used when fitting the model or when evaluating forecast accuracy. The blue lines in figure 3.1 represent the training data, which is expanded by one or more observations every time. It is not possible to obtain a reliable forecast based on short training data, which is why the first training set must contain a sufficient amount of data. The red lines represent the test data, which always have the same length, but take one step forward each time. The test data can consist of one or more observations. For multi-step forecasts, the test set is not the consecutive observations from the train set, but n-steps ahead. For the 12-step-ahead forecast, the test set will be the 12^{th} observation following the train set. When used for fitting the model, the model is fitted for all the different lengths of training data. With short series, this may not be appropriate because there is not enough data to hold out for testing purposes (Hyndman & Athanasopoulos, 2018). When cross-validation is used for evaluating forecast accuracy, the same model is used to estimate the coefficients for every length of training data. The forecasts are then computed for all models and compared to the following test data. The forecast accuracy can then be computed by averaging over all the different test sets. If doing one-step-ahead forecasts, only one accuracy measure is computed for each training and test split. However, if for example doing 12-step forecasts, the accuracy is computed for 12 horizons for every training and test split. It is then possible to see how good the model performs for different forecast horizons.

3.3.3 Performance Measures

Since the information criteria discussed in section 3.3.1, can not be used to compare different forecasting methods, we must use some measure of forecast performance. The difference between an observed value and its forecast is called a forecast error. Forecast errors can be calculated in many different ways, and different methods are suitable in different situations. Several performance measures can also be used simultaneously.

Among the most common performance measures we find *Mean Error(ME)*, *Mean Absolute Error(MAE)* and *Root Mean Squared Error(RMSE)*, as well as percentage versions of these. There are also different kinds of scaled versions of these performance measures, that are supposed to account for differences in scale when comparing forecast accuracy across series with different units (Diebold, 2004).

ME measures average model bias, which is one component of accuracy. ME can be calculated by taking the mean of the difference between the observed value, y_{T+h}, and the forecast value, $\hat{y}_{T+h|T}$, for all t,

$$ME = \frac{1}{T}\sum_{t=1}^{T}(y_{t+h} - \hat{y}_{t+h|t}) = \frac{1}{T}\sum_{t=1}^{T} e_{t+h|t}. \tag{3.23}$$

We generally prefer a forecast with a small bias (Diebold, 2004). A positive bias indicates that the forecasts are underestimated, as the actual values are mostly higher than the forecasts. Vice versa, a negative bias indicates that the forecasts are generally overestimated. The ME should however be interpreted with caution, since positive and negative errors cancel each other out, and this measure is therefore often just used to decide if any measures must be taken to reduce the model bias. However, the MAE takes the absolute value of the errors and is therefore indifferent to whether the error is negative or positive. MAE can be calculated by taking the mean of the absolute value of the difference between the observed value and the forecast value (Diebold, 2004),

$$MAE = \frac{1}{T}\sum_{t=1}^{T}\left|e_{t+h|t}\right|. \tag{3.24}$$

RMSE also measures the average magnitude of the errors, without considering their

directions. It is calculated by taking the square root of the squared errors (Diebold, 2004),

$$RMSE = \sqrt{\frac{\sum_{t=1}^{T} e_{t+h|t}^2}{T}}. \tag{3.25}$$

For both MAE and RMSE low values are better. MAE gives all individual errors equal weight, while since the errors are squared before they are averaged for RMSE, higher weight is given to larger errors. RMSE is therefore especially useful when large errors are undesirable, while MAE is not that sensitive to outliers. A disadvantage of RMSE is that if we have noisy and random data, a single very bad forecast may skew the metric to overestimating how bad the model is. On the other hand, if all the errors are small, this measure may underestimate the model's badness. Further, minimizing MAE will lead to forecasts of the median value, while minimizing RMSE will lead to forecasts of the mean (Hyndman & Athanasopoulos, 2018).

MAE and RMSE are scale-dependent error measures. This does not entail any issues when comparing the different methods for the same time series. However, it causes some complications when comparing the forecast error between different time series. An alternative could therefore be the *Mean Absolute Percentage Error (MAPE)*, which is a percentage version of MAE and can be calculated as

$$MAPE = \frac{1}{T} \sum_{t=1}^{T} \left| \frac{y_{t+h} - \hat{y}_{t+h|t}}{y_{t+h}} \right| \tag{3.26}$$

The problem of MAPE is however that when the actual observation y_{t+h}, which is the denominator, is zero, MAPE can not be calculated (Gilliland, Sglavo & Tashman, 2015). Hyndman & Koehler (2006) therefore proposes an alternative to replace MAPE: the *Mean Absolute Scaled Error (MASE)*. MASE overcomes many problems related to other measures. This method scales the error based on the *in-sample*, or training, MAE from the naïve forecast method. MASE can be computed as follows

$$MASE = \frac{1}{J} \sum_{j=1}^{J} |q_j| = \left| \frac{y_{t+h} - \hat{y}_{t+h|t}}{\frac{1}{T-1} \sum_{t=2}^{T} |y_t - y_{t-1}|} \right|, \tag{3.27}$$

where J is the number of forecasts. If the scaled error is less than one, it arises from a

better forecast than the average naïve forecast computed on the training data. Conversely, it is greater than one if the forecast is worse than the average naïve forecast computed on the training data (Hyndman, 2006). MASE will only be infinite or undefined if all historical observations are equal, e.g. zero.

4 Data Analysis and Modelling

To assess the potential of demand forecasting in the krill industry, we have built an automatic forecasting model using the statistical programming language R. This model will (1) fit a model to the data by using three different forecasting methods, and (2) select the best method for 12-step forecasting. The automatic model takes several factors into account, and through a series of tests checks for stationarity, as well as autocorrelation and other factors that may have a decisive impact on the forecasting model selected. In this chapter, we will explain how we have built the first part of this modeling procedure, which tests we have chosen to include and why we have made these choices. This part of the model is illustrated in figure A1.1 in Appendix (7), with yellow boxes.

4.1 Data

For the purpose of this thesis, ABM have provided us with internal data of sales volume of krill meal from January 2012 to August 2019. We aim to use this data to provide forecasts for one year ahead, i.e. 12-step-ahead forecasts. Forecasts are desired at a 12-month horizon, mainly since the harvesting season is 12 months; from December 1 to November 30 the following year (CCAMLR, 2010). But also since 12 months is a common tactical planning horizon and is the desired forecast horizon for ABM.

4.1.1 Descriptive Statistics

Our initial data includes information about sales volume of four different products for almost 200 customers. The products are $QRILL^{TM}$ *Aqua*, $QRILL^{TM}$ *High Protein*, $QRILL^{TM}$ *Pet* and $QRILL^{TM}$ *Astaxanthin Oil*. The sales are registered on the dates the sales occur, hence dates with no sale are not included. Some of the customers often or regularly buy one or several products from ABM, while others buy less frequently. For some customers, no sales are registered for several months, while others have only made purchases the most recent months (see examples in figure 4.5).

The data consist of 8 variables and 3775 observations. The variables include *Year*, *Quarter*, *Month*, *MT* (sales volume in metric tons), *Date*, *ProductType*, *Invoiced* (invoiced customer name) and *Company*.

The data are quite clean and only need some simple adjustments before further analysis. All sales with a volume below ten metric tons are set equal to zero in our data, since these do not represent real sales. We also remove the product type $QRILL^{TM}$ *Astaxanthin Oil*, mainly because of its relatively low sales volume, but also because this product is produced, used and distributed in quite a different manner than the other product types. The other products, which are all different types of krill meal, will be summarized and forecast as total sales volume together. Further, since we want to forecast monthly sales, the time series should be on a monthly basis. Therefore, we combine the *Year* variable and the *Month* variable and exclude all other variables except *MT* and *Company*. This leaves us with a data set that contains 3 variables. In order to get complete monthly data, we add new rows for each month a given customer has not made any purchases, with sales volume equal to zero. With this data we can make separate time series for all 183 customers; 16 836 observations in total.

Before investigating the data per customer, we compute some descriptive statistics for total sales volume, see table 4.1. We observe that the maximum sales volume of 1,667 metric tons is much higher than the mean. This indicates that most of the observations are very low. In addition, we notice that the median equals zero sale. When we investigate this further, we find that 15,038 observations are zero, which means that only 1,798 observations have positive values. This indicates that there are far more months without a sale per customer than months with sales. Lastly, the standard deviation is somewhat high, which indicates a high dispersion in the observations.

Mean	**SD**	**Min**	**Max**	**Median**
11	71	0	1667	0

Table 4.1: Descriptive statistics

In general, the data contains a substantial amount of zeros, with several periods of zero sale. When a sale first occurs it is often small, and can also be highly variable in size. These are characteristics of what is often referred to as *intermittent demand*. Forecasting intermittent demand entails several challenges, one of which is that many zero values may render usual forecasting methods difficult to apply (Croston, 1972). The simple exponential smoothing method will for example give an upward forecast bias in periods directly after non-zero demand. Further, sparse data with periods with positive values

separated by a number of periods with zero values makes it difficult to identify trends and patterns. It also makes it difficult to estimate which periods in the future that will register activity, and which will not.

The monthly time series for every customer consists of 92 observations. For practical reasons, this thesis will focus on the 20 customers with the highest total sales volume for the rest of this thesis. As displayed in figure 4.1, the two customers with the highest sales volume account for a majority of total sales volume. Thereafter, the sales volumes decrease steadily for the remaining 18 customers.

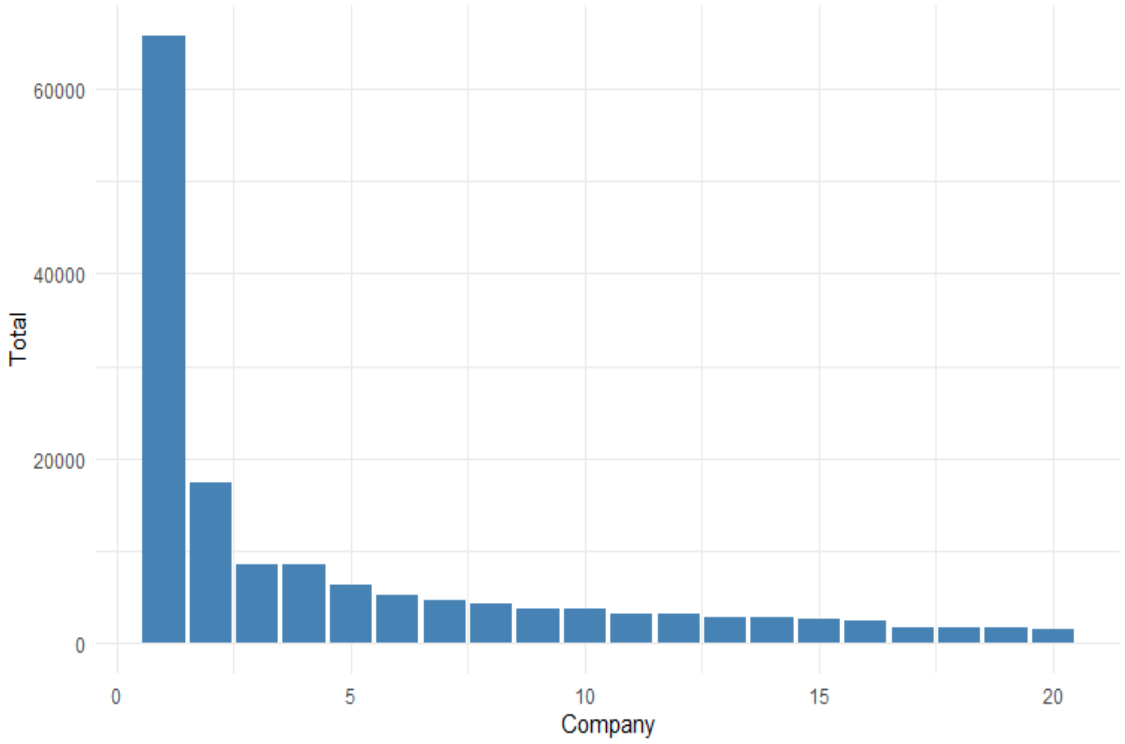

Figure 4.1: Total MT per customer (Company 1-20)

For illustrative purposes, figure 4.2 shows the time series of summarized total MT for all customers from January 2012 to August 2019.

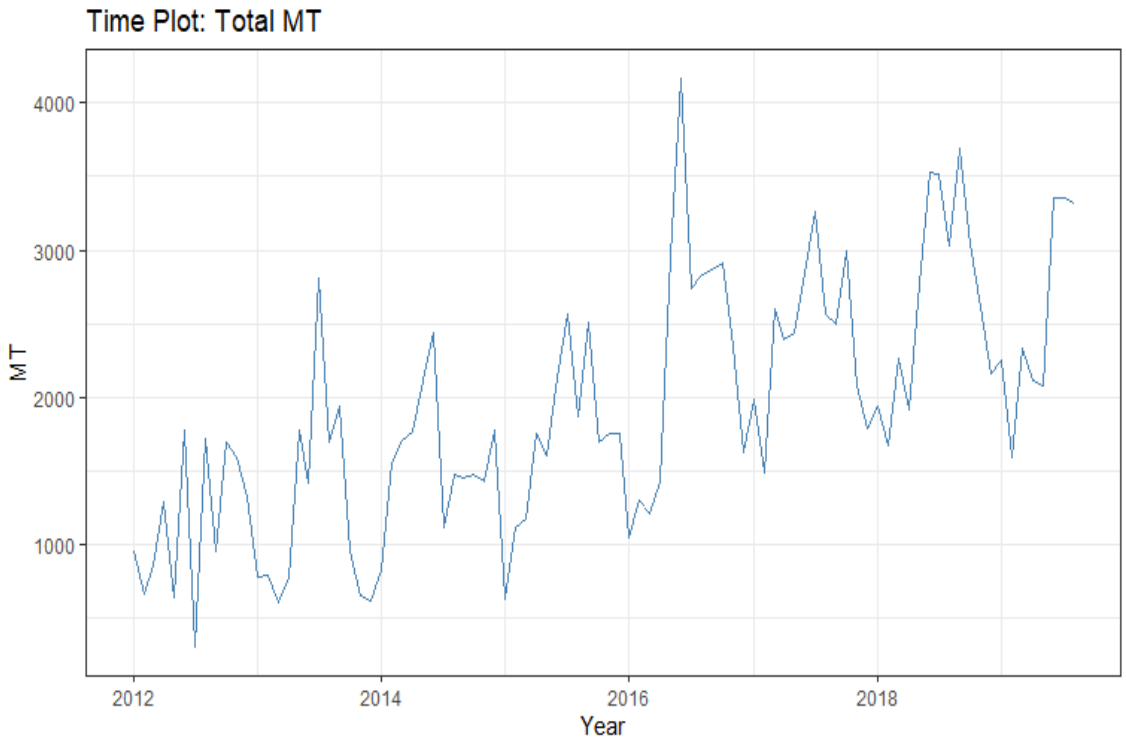

Figure 4.2: Summarized MT for all customers

We observe an increasing trend, which indicates that the sales volume has been increasing steadily since 2012. There also seems to be a spike around the middle of each year with a higher sales volume than the rest of the year.

4.1.2 Initial Plots

This thesis concerns forecasts of demand per customer. In this section, we will therefore investigate the time series of the four customers with the highest total sales volume. For practical reasons, we will only show the plotted time series of these customers and carefully investigate their patterns.

For Company 1, we observe an increasing trend. A weak positive trend can also be seen for Company 2.

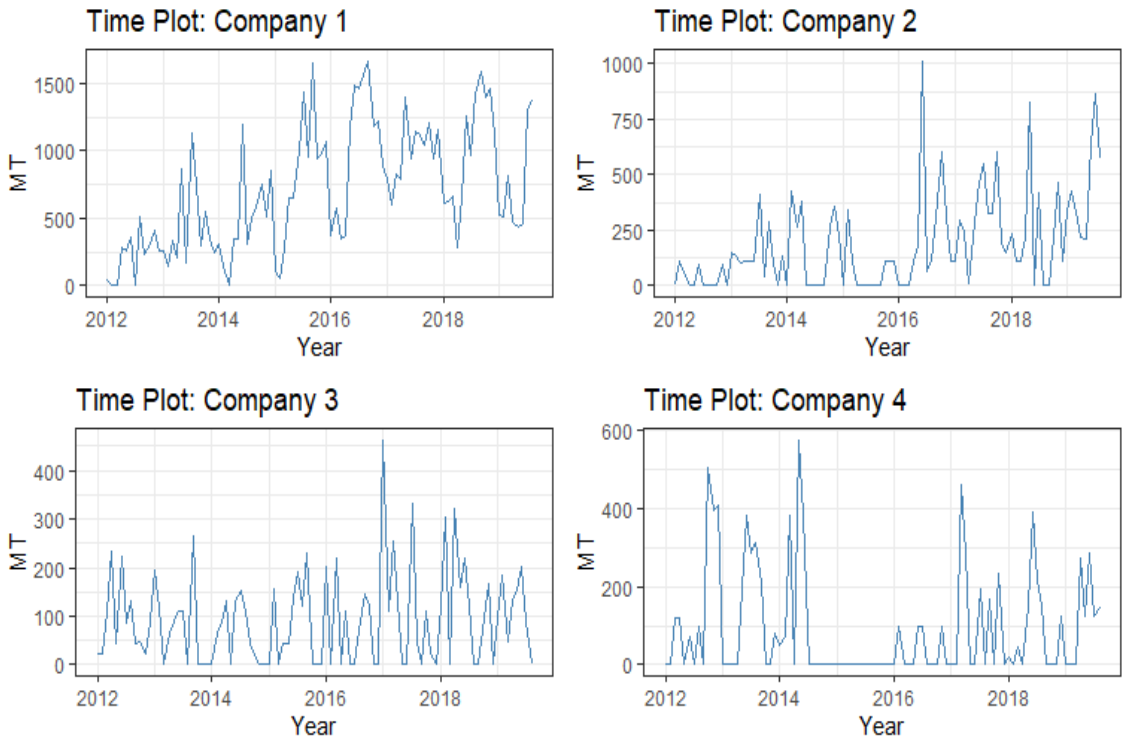

Figure 4.3: Time series of total MT (Company 1-4)

In addition, we notice that for Company 1, there may be a seasonal pattern with an increase in sales in the middle of each year and decrease at the end of each year, similar to the summarized MT seen in figure 4.2. This is confirmed in the seasonal subseries plot in figure 4.4. Company 2 might have a similar pattern. However, the time series for Company 3 and 4 appear random and do not indicate the same seasonal pattern. This concurs with our initial thoughts of quite different characteristics between the customers, which implies that it may be difficult to find a general model that will be suitable for all customers.

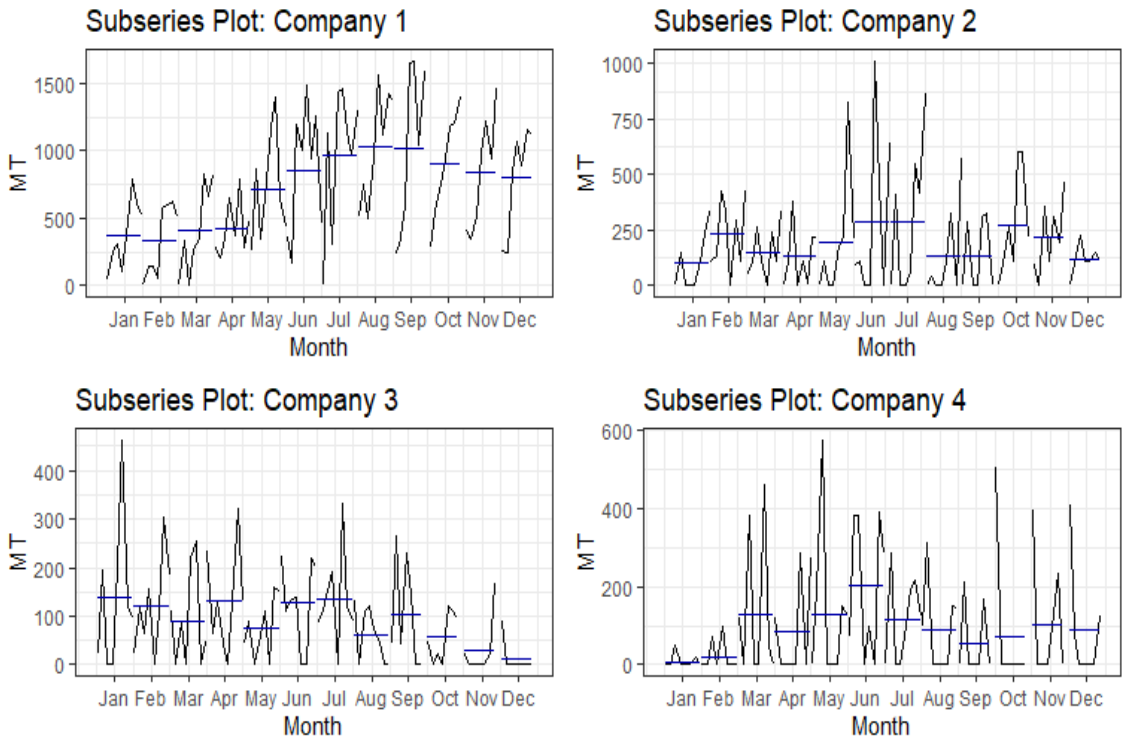

Figure 4.4: Seasonal subseries plot (Company 1-4)

For additional illustration, figure 4.5 shows the time series for some customers with many zero values.

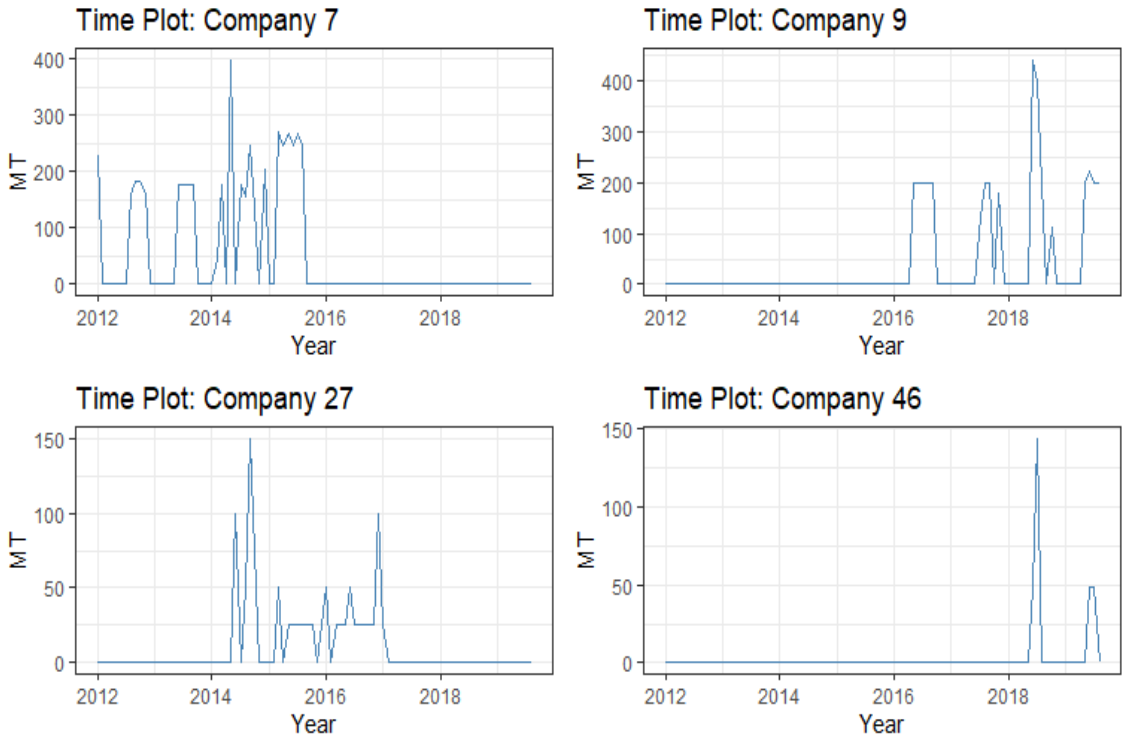

Figure 4.5: Time series of total MT (Company 7, 9, 27 and 46)

Company 7 has not purchased anything since 2015. Conversely, Company 9 has not purchased anything before 2016. In addition, Company 27 combines both of the above cases, while Company 46 only has 4 positive values. For the majority of the remaining customers, the time series appear quite random. There are also few customers with an apparent trend.

4.1.3 Data Transformation

As discussed in section 3.2.1, mathematical transformations of historical data can be useful to get more accurate forecasts. If transformation is suitable, it will be applied before fitting a model to the data. However, Box-Cox and other transformation methods are most useful if the variation in the data increases or decreases with the level of the series, as discussed in section 3.2.1. Such patterns do not seem prominent in our data, as can be seen in section 4.1.2, and data transformations are therefore not likely to be useful. We still do some experiments with Box-Cox transformation for some customers, but find that it generally gives worse results. We will therefore not apply transformations to our time series, or go further into the procedures of doing such transformations. In the following sections of this chapter, we will elaborate on how we have fitted the different models and built an automatic model without transformations of the data.

4.1.4 Training and Test Data

The automatic model takes a customer name as input. It then extracts the respective time series for this customer to initiate the modeling procedure. The first step is to divide the time series into training and test data. Here, the user can specify how many months to include in the training and test data. For our data, we set the last 24 observations as test data. These will be used to validate the forecasts. The test data is subtracted from the total data, 92 observations for our time series, which gives training data consisting of 68 observations. This satisfies the minimum sample size discussed in section 3.2.2. Since the forecast-horizon is 12 months, a 12-month test data is the minimum we should have, with reference to section 3.3. We choose to have 24 months in our test data in order to have sufficient data to validate the models, while still having enough training data.

4.2 Selection of Information Criterion

When fitting models for the different forecasting methods, we need a measure of predictive accuracy in order to select the best predictors and compare the different models. As mentioned in paragraph 3.3.1, there is a range of different measures that can be used to find the most suitable model, and no one criterion will always outperform the others. We have chosen to minimize AIC_c as measure of predictive accuracy for all the forecasting methods. This is mainly because AIC_c is a bias-corrected version of AIC, since AIC tends to select too many predictors for short data. Considering that we have relative short data, AIC_c seems more suitable. Further, BIC is a popular alternative among statisticians, as this will select the true underlying model if it exists. However, the true underlying model rarely exists, and even if it does, selecting that model will not necessarily give the best forecasts because the parameter estimates may not be accurate (Hyndman & Athanasopoulos, 2018).

Box et al. (1994) suggests that information criteria should only be used as supplementary guidance to the visual inspection of the ACF and PACF plots for ARIMA modeling, discussed in section 3.3.1. However, visual inspection is comprehensive to implement in an automatic modeling procedure and we have therefore chosen to exclude this part of the modeling process.

4.3 ETS Modeling

When fitting an ETS model, we use the *ets()* function from the "forecast" package (Hyndman et al., 2019), which includes methods and tools that are useful when doing exponential smoothing and ARIMA modeling. As mentioned in section 4.1.1, some customers only first purchased krill after 2016, and thus have few to no positive values in the training data - i.e. little or no observations to train the model on. To handle this issue, we set a limit of 20 sales (minimum 20 sales above zero) in the training data for estimating all models in this thesis.

4.3.1 ETS

The *ets()* function returns the model with highest predictive accuracy for the training data. The possible inputs to the *ets()* function are "N" for none, "A" for additive and "M" for multiplicative. The model also include a damped or non-damped trend. As discussed in section 3.1.2.1, when the data contain zeros or negative values, multiplicative error models should not be considered. Since our data contain zeros, only the fully additive models will be considered in the modeling procedure.

For illustrative purposes, we show the output when fitting an ETS-model for ABM's customer with the highest sales volume in the output seen in figure 4.6.

```
ETS(A,N,A)

Call:
 ets(y = train, model = best.ets.order, ic = "aicc")

  Smoothing parameters:
    alpha = 0.2654
    gamma = 1e-04

  Initial states:
    l = 369.3022
    s = 2.6593 46.5702 101.6731 292.8617 246.5948 267.7116
           280.5615 94.1892 -249.9388 -381.7127 -383.336 -317.8339

  sigma:  297.9597

     AIC     AICc      BIC
1076.037 1085.268 1109.330

Training set error measures:
                    ME     RMSE      MAE MPE MAPE      MASE       ACF1
Training set 32.21144 265.5217 209.9699 NaN  Inf 0.6196419 -0.139667
```

Figure 4.6: Chosen ETS model (Company 1)

The model selected is ETS(ANA). This model has additive errors, no trend and additive seasonality:

$$\begin{aligned} y_t &= l_{t-1} + s_{t-m} + \varepsilon_t \\ l_t &= l_{t-1} + \alpha\varepsilon_t \\ s_t &= s_{t-m} + \gamma\varepsilon_t, \end{aligned} \tag{4.1}$$

where the smoothing parameters are $\alpha = 0.2654$, $\gamma = 0.0001$. α controls how much weight

is given to old observations. Here, α is low, which means that more weight is given to old observations. γ is very low which means that the seasonal pattern change very little over time (Hyndman & Athanasopoulos, 2018). This can also be seen by looking at the graph of the seasonal component, figure 4.7.

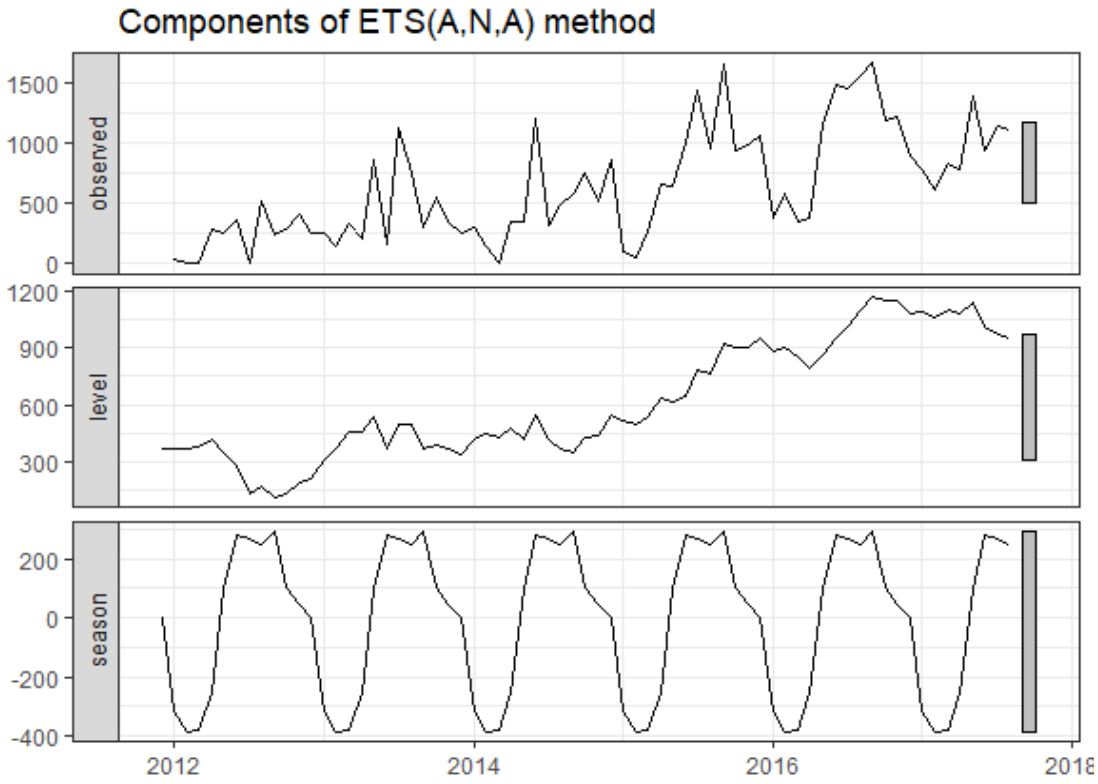

Figure 4.7: Components of ETS model (Company 1)

4.3.2 STL + ETS

As mentioned in section 3.1.2.2, the combination of STL decomposition and ETS modeling can handle any type of seasonality and is robust to outliers. We will therefore consider this combination in addition to regular ETS modeling. This can be done by using the *stlm()* function (also from the "forecast" package) (Hyndman et al., 2019), which first applies STL decomposition, then forecasts the seasonally adjusted series and finally returns the reseasonalized forecasts. When fitting the model, the desired method, in our case ETS, is specified in the function. The ETS method will then be applied to the seasonally adjusted series. Since the seasonal component already has been forecast from STL, as explained in section 3.1.2.2, only ETS models without a seasonal component is considered. In addition, multiplicative errors will not be considered because of zero values in the data (see section

3.1.2.1). The only models that can be considered are therefore fully additive models without a seasonal component: (ANN), (AAN), (AA_dN).

When applying the STL+ETS method to the time series of ABM's customer with the highest sales volume, the model selected is ETS(ANN):

$$\begin{aligned} y_t &= l_{t-1} + \varepsilon_t \\ l_t &= l_{t-1} + \alpha\varepsilon_t. \end{aligned} \tag{4.2}$$

Figure 4.8 below shows the components of STL decomposition and of the chosen ETS model. Here we clearly see a positive trend until 2017. We also see a clear seasonal pattern in the seasonal component. Unlike the ETS model in figure 4.7, we see that the chosen ETS model, ETS(ANN), does not include a seasonal component.

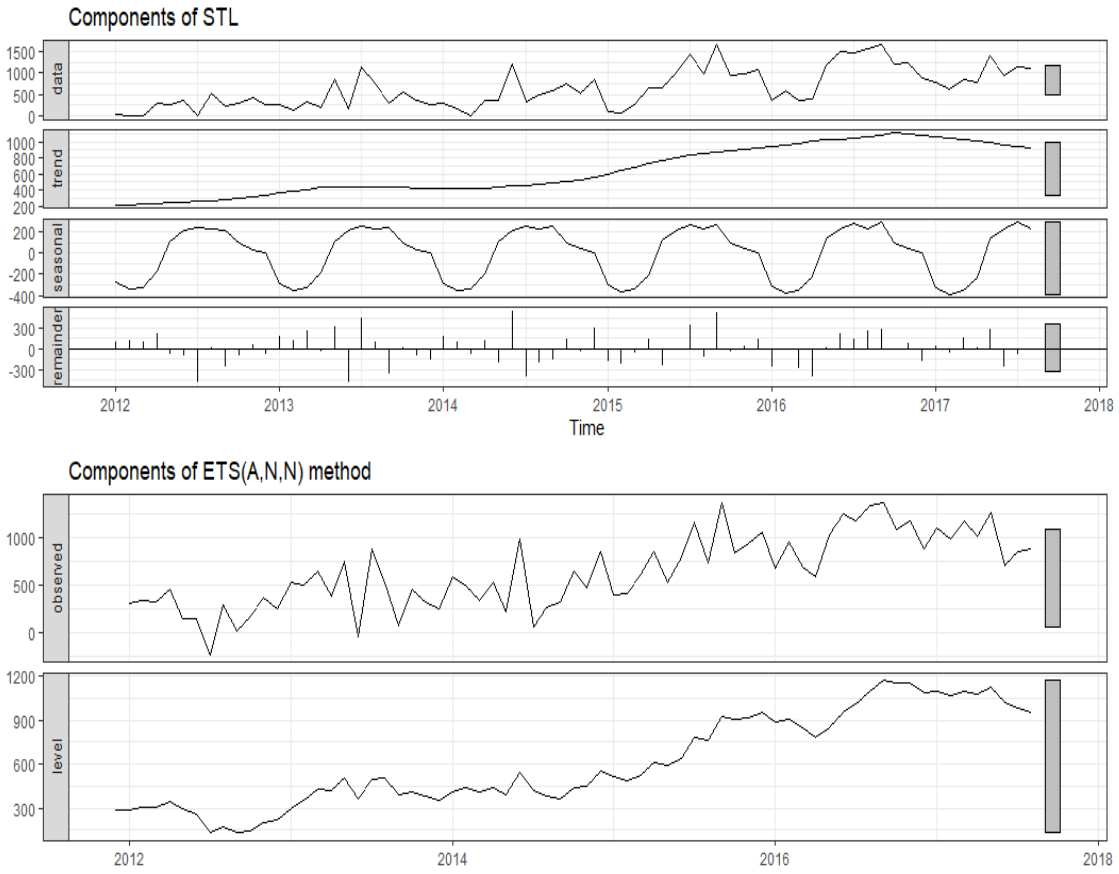

Figure 4.8: Components of STL+ETS model (Company 1)

4.4 ARIMA Modeling

When fitting an ARIMA model, we use the *Arima()* function from the "forecast" package (Hyndman et al., 2019). Another function, *arima()*, could also be used, but it does not return everything required for other functions in the "forecast" package to work. In addition, it does not allow the estimated model to be applied to new data (Hyndman & Athanasopoulos, 2018).

In section 3.1.3.4, we discussed how a seasonal ARIMA model contains a non-seasonal and a seasonal part. These are *(p,d,q)* and $(P, D, Q)_m$, respectively. When fitting an ARIMA model, we set restrictions for the value of all these components. First of all, if the non-seasonal and seasonal terms contain too many parameters, it is likely that the model is prone to over-parameterization, and as discussed in section 3.1.3.3, we often prefer a simpler model with fewer parameters. In addition, in section 3.2.2, we argued that when the training data is short and the number of parameters to be estimated is high, the models are prone to overfitting. Further, if there are too many parameters the model crashes. We therefore set the following restrictions, see table 4.2.

Non-seasonal term	$0 \leq \mathrm{p} \leq 2$	$0 \leq \mathrm{d} \leq 2$	$0 \leq \mathrm{q} \leq 2$
Seasonal term	$0 \leq \mathrm{P} \leq 2$	$0 \leq \mathrm{D} \leq 2$	$0 \leq \mathrm{Q} \leq 2$

Table 4.2: Component restrictions for ARIMA

The autoregressive (AR) component of both the non-seasonal and seasonal term is restricted to maximum 2 since a higher value does not occur often in practice (Pankratz, 1983). The differencing component is restricted to 2 for both the non-seasonal and seasonal term since according to Hyndman & Athanasopoulos (2018), more than second differences is rarely necessary, also discussed in section 3.1.3.1. Lastly, the moving-average (MA) component for both terms is restricted to 2, since according to Pankratz (1983) ARIMA models with an MA-component higher than 2 rarely occur.

4.4.1 Fitting Method

The *Arima()* function gives three alternatives when choosing fitting method. The CSS-method minimizes the conditional sum of squares, while the ML-method maximizes the

log-likelihood function of the ARIMA-model. The CSS-ML-method is the third and default method, which mixes both methods by first using CSS to find the starting values and then ML to fit the model (Hyndman et al., 2019). We have tested all the methods and choose to use the CSS-method because it is the only method that works in our modeling procedure.

To illustrate how the CSS-method works, we consider an MA(1) model

$$y_t = \varepsilon_t + \theta\varepsilon_{t-1}. \tag{4.3}$$

y_t conditional on the error term in the previous time period is normally distributed with mean $\theta\varepsilon_{t-1}$ and variance σ^2. When computing ε_t, we need to set ε_0 equal to zero because ε_{t-1} is not directly observable (Harvey, 1993). All the errors can then be computed by re-arranging the above equation,

$$\varepsilon_t = y_t - \theta\varepsilon_{t-1}. \tag{4.4}$$

When minimizing the conditional sum of squares function, we therefore use the following formula

$$S(\theta) = \sum_{t=1}^{T}(y_t - \theta\varepsilon_{t-1})^2 = \sum_{t=1}^{T}\varepsilon_t^2. \tag{4.5}$$

4.4.2 Differencing

As discussed in section 3.1.3.1, the data must be made stationary when applying the ARIMA method. This is done by differencing the data. We use statistical tests to determine how many differences that are required and apply the respective number of differences to the data. This can include both seasonal differencing and first differencing. An important aspect of this way of checking for stationarity is that there is no visual inspection of neither the data nor ACF and PACF. We therefore assume that the statistical tests will accurately determine this.

The functions we use to determine the required number of differences include several alternatives for choice of stationarity tests or unit root tests. One thing worth noting

when applying several different tests is that these "are based on different assumptions and may lead to conflicting answers" (Hyndman & Athanasopoulos, 2018, Ch. 8.1). We will therefore give a brief explanation of our choice of tests for stationarity.

We will start by explaining our choice of tests for first differencing, but in the automatic model we have chosen to apply seasonal differencing first. This is based on the argumentation in section 3.1.3.1, where we discussed how seasonal differencing should be applied first if strong seasonal patterns are present. This may not be the case for our data, as argued in section 4.1.2. However, since we in section 3.1.3.1 also argued that it is indifferent whether first or seasonal differencing is applied first when both differences are applied, we find it reasonable to start with seasonal differencing.

4.4.2.1 First Differencing

The *ndiffs()* function from the "forecast" package "estimates the number of first differences necessary" (Hyndman et al., 2019, p. 95). The output "1" implies that the data have a unit root, thus are non-stationary. Conversely, the output "0" implies no unit root in the data. When using this function, one can choose between three different tests; The Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test, the Augmented Dickey-Fuller (ADF) test, and the Phillips-Perron (PP) test. The function uses the KPSS test as default.

As discussed in section 3.1.3.2, the ADF test is a unit root test where the null hypothesis is that a unit root is present in the data. The PP test has the same null hypothesis. For these types of tests *ndiffs()* "returns the least number of differences required to **fail** the test at the level alpha" (Hyndman et al., 2019, p. 95). Generally, this means that if the absolute value of the test-statistic is higher than the absolute critical value for the significance level set as alpha, the null hypothesis is rejected, thus the data are stationary. Vice versa, the data are non-stationary if the test-statistic is lower than the critical value. Unlike the ADF and the PP test, the KPSS test is a stationarity test where the null hypothesis is that the data are stationary. In practice, when using the *ndiffs()* function, the KPSS test "returns the least number of differences required to **pass** the test at the level alpha" (Hyndman et al., 2019, p. 95).

For stationarity tests like KPSS, it is less likely to conclude that the data are non-stationary than for unit root tests. When doing stationarity tests, it is more likely that the data

are truly non-stationary if the null hypothesis is rejected. However, a nonrejected unit root test can not exclude that the data are stationary. We therefore argue that the risk of over-differencing is higher when using unit root tests, while the risk of under-differencing is higher when using stationarity tests. The question is therefore which of these scenarios are worse? If a stationary time series is differenced, MA unit roots can occur and there is over-differencing. It then becomes a non-invertible MA process, which is not desired, as discussed in section 3.1.3.3. It is possible to test for an MA unit root with MA unit root as the null-hypothesis. However, "tests for the MA unit root as null and tests for stationarity as null are related" (Maddala & Kim, 1998, p. 120). The KPSS test can therefore be used as a complementary test when testing for unit roots. Furthermore, Arltová & Fedorová (2016) have investigated which test to choose based on the length of the time series and the value of the AR(1) parameter. They found that the ADF and PP tests have the highest power for shorter time series, while the KPSS test is suitable for very small values of the AR(1) parameter. However, they do not recommend to only use the KPSS test, but in combination with a unit root test. Arltová & Fedorová (2016) argued that "the ADF test is and will be one of the most commonly used unit root test since its crucial advantage lies in its simple construction and feasibility". Therefore, we choose to use the ADF test, as well as the KPSS test. We will first consider the ADF test. If the ADF test rejects the null-hypothesis, we can be quite sure that the data are truly stationary and do not run the KPSS test in addition. However, if the ADF test keeps the null-hypothesis, we also consider the KPSS test. If the KPSS test rejects the null-hypothesis, it matches the ADF test and thus differencing is necessary. In the special case where both the ADF and the KPSS test keep the null-hypothesis, the time series do not provide enough information. However, in such cases we choose to difference because of Arltová & Fedorová (2016)'s conclusion that it is better to over-difference than to under-difference.

4.4.2.2 Seasonal Differencing

The *nsdiffs()* function from the "forecast" package "estimates the number of seasonal differences necessary" (Hyndman et al., 2019, p. 98), and outputs "1" and "0" for seasonal unit root and no seasonal unit root, respectively. Similar to the *ndiffs()* function, one can choose between several different tests: measure of seasonal strength (SEAS), the Canova-Hansen (CH) test, the Hylleberg-Engle-Granger-Yoo (HEGY) test and the Osborn-

Chui-Smith-Birchenhall (OCSB) test. Measure of seasonal strength is the default test.

As discussed in section 3.1.3.2, the CH test is a generalization of the KPSS framework, while the HEGY test is a generalization of the Dickey-Fuller framework from zero frequency to seasonal frequencies. We therefore use these tests when determining the order of seasonal differences to be consistent with the tests we use for first differencing.

The CH test has the null hypothesis that the seasonal pattern is deterministic, i.e. that the data are stationary. This means that the null hypothesis is rejected if the time series' seasonality is not constant, i.e. non-stationary. The HEGY test is a seasonal unit root test with the null hypothesis that the data are non-stationary. Since these tests are similar to the ones described in section 4.4.2.1, we will not go into further details of these tests.

4.4.3 Including Deterministic Trend or Drift

When fitting an ARIMA model, one option is to include trend or drift. As mentioned in section 2.1.2, ABM's sales volume is mostly determined by how much krill it is possible to harvest. We therefore do not believe that a deterministic trend is present in our data. In addition, we have investigated the alternative of including drift when fitting our model. Forecasts using the drift method is equivalent to drawing a line between the first and last observations and extrapolating it into the future. We have compared the error when forecasting with and without drift for the 20 customers with the highest sales volume and found that a model without drift gives better forecasts for a majority of these customers. Trend or drift is therefore not included in the automatic modeling procedure.

4.4.4 Ljung-Box Test

The Ljung-Box test is a test for autocorrelation where the autocorrelation for lag k, r_k, is grouped into groups of size h lags. Then "we test whether the first h autocorrelations are significantly different from what would be expected from a white noise process" (Hyndman & Athanasopoulos, 2018, Ch. 3.3). A white noise process is an example of a stationary time series. The test is defined by

$$Q^* = T(T+2)\sum_{k=1}^{h}(T-k)^{-1}r_k^2, \tag{4.6}$$

where h is the maximum lag being considered and T is the length of the time series. Hyndman & Athanasopoulos (2018) suggests using $h = 10$ for non-seasonal data and $h = 2m$ for seasonal data, where m is the period of seasonality. In our model for automatical selection of the best ARIMA-model, we use the *checkresiduals()* function from the "forecast" package (Hyndman et al., 2019) when conducting the Ljung-Box test. In this function, we use the default setting for h: For seasonal data $h = min(2m, n/5)$, and for non-seasonal data $h = min(10, n/5)$. High values of Q^* indicate that the autocorrelation do not come from a white noise series (Hyndman & Athanasopoulos, 2018). We compute the p-value for every possible model and remove the alternatives with p-value below 0.05. In this way, a model that passes the Ljung-Box test is always chosen.

4.5 Modeling Results

So far in this chapter, we have presented the modeling procedure for the automatic model. This model is, based on a series of different tests and fitting procedures, supposed to select a good forecasting model for each customer from the three different forecasting methods: ETS, STL+ETS and ARIMA. The input to the model is the respective time series for the customer, and the output is the best model, in terms of lowest AIC_c, out of all the candidate models from the different methods. For illustrative purposes, the chosen models for all methods and for the 20 customers with the highest sales volume are displayed in table A2.1 in the Appendix (7). In the following chapter, we will compare the different forecasting methods to select the best model for each customer and evaluate the results.

5 Forecasting and Evaluation

Evaluation is important to demonstrate the value of our modeling. Choice of performance measure can have a significant impact on the evaluation. In this chapter, we will therefore first discuss our choice of performance measure. Further, we will present our forecasting results. As mentioned in section 3.1.1, new and more complex methods are not worth considering if they do not lead to better forecasts. We will therefore present some simple benchmarks and do a comparison between them and the results from our model presented in chapter 4, in which time series cross-validation will be used. This part of the model is illustrated in figure A1.1 in the Appendix (7), with orange boxes.

5.1 Choice of Performance Measure

The data per customer are on the same scale, and all models are fitted based on the same training and test data. For our data, the most relevant alternatives for choice of performance measure are therefore ME, MAE and RMSE. RMSE is generally a more popular metric for statisticians than for forecasters and is a bit more challenging to interpret (Gilliland et al., 2015). Moreover, the RMSE gives more weight to larger errors, while the MAE gives equal weight to all errors, as discussed in section 3.3.3. For the purpose of demand forecasting of krill meal, small errors are desirable, but larger errors do not necessarily have more impact than smaller errors. MAE may therefore be more suitable than RMSE for our data.

As discussed in section 3.3.3, minimizing MAE will lead to forecasts of the median value, while minimizing RMSE will lead to forecasts of the mean. Our data contains a significant amount of zeros. We therefore find it more reasonable to believe that a forecast of the median will concur with a random observed value, than a forecast of the mean. We will therefore use MAE to quantify the accuracy of the methods we use, even though this implies that the forecasts in many cases will be zero. In section 4.1.1 we discussed that some customers buy regularly, while others buy less frequently. For the customers with a persistent purchasing pattern, it is unlikely that the forecasts will be zero, while for the customers who buy less frequently, zero forecasts are more likely to occur. This is however often consistent with reality, which is why we will minimize MAE when searching for the

best model for each customer.

Since we also want to compare the relative model performance across customers, and thus across different time series, we will also compute MASE. We choose MASE as a relative metric since it overcomes many problems related to other measures, as discussed in section 3.3.3. In addition, Hyndman (2006) recommends this measure when dealing with intermittent demand. A relative metric can both tell us something about how our method improves on a certain benchmark and provide a needed perspective for situations with bad data. When dealing with bad or little data, forecast models usually give high forecast errors, but that does not necessarily mean that the forecast method has failed. By comparing the errors to a benchmark, we may find that we have made progress, and hence that the source of the high error rate is not bad forecasting but bad data (Gilliland et al., 2015).

5.2 Forecasting Results

When forecasting, we use the *forecast()* function from the "forecast" package (Hyndman et al., 2019). This function produces negative forecasts for some of the customers. Since the demand for krill meal will never be negative, all negative forecasts are set to zero. We will first present our benchmarks, and then the resulting forecasts from the models chosen as the best for each method in our modeling procedure. Thereafter we will evaluate the results by applying cross-validation to these models.

5.2.1 Benchmark

The simple forecasting methods mentioned in section 3.1.1 are often used as benchmarks when evaluating new forecasting methods. Since we have many different time series (one for each customer), we will look at three different methods for our benchmark. We will use two of the simple forecasting methods: the naïve method and the seasonal naïve method. In addition, an AR(1) model is often used as benchmark when doing ARIMA modeling. The method with the lowest MAE for each customer will be used as benchmark for the customer in question.

Figure 5.1 shows illustrations of all three methods for ABM's biggest customer. The plots show the time series from January 2016 until the end of the forecast period.

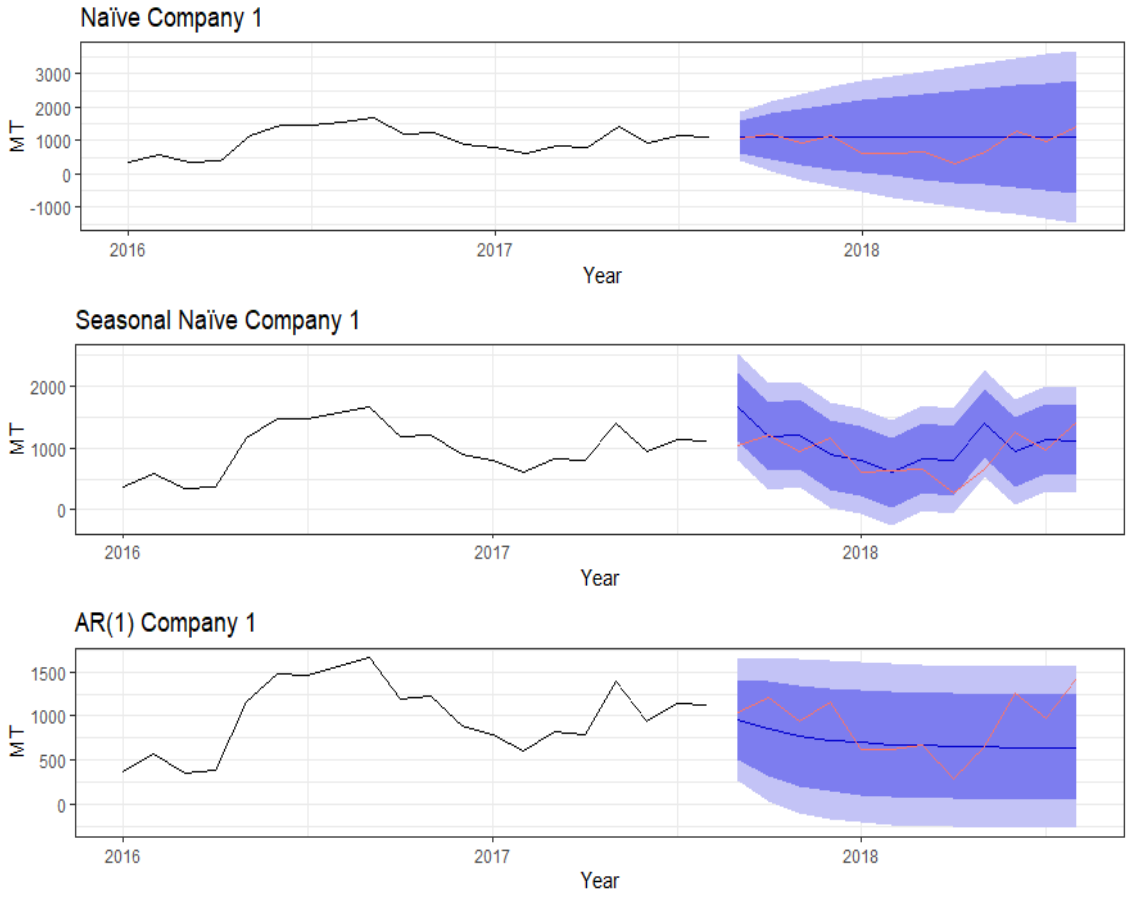

Figure 5.1: Alternative benchmarks (Company 1)

The blue line represents the forecasts, while the red line is the test data. The blue shaded area is the prediction interval, which expresses the uncertainty in the forecasts. The dark shaded area is the 80% prediction interval, in which a forecast is expected to lie within with 80% certainty. The light shaded area shows the 95% prediction interval. For the benchmarks illustrated above, both the 80% and the 95% prediction intervals are wide. This means that the forecasts are expected to be inaccurate. The prediction interval can contain negative values, but all these can be interpreted as zero, i.e. no sale.

By looking at the above plots, the seasonal naïve method seems to follow the data more closely than the other two methods. However, to determine which is the best benchmark for this particular customer, we look at the calculated MAE for the three methods, see table 5.1.

Method	MAE
Naïve	311.73
Seasonal Naïve	302.05
AR(1)	274.33

Table 5.1: Benchmark MAE (Company 1)

The MAE for the three methods are quite similar, but the AR(1) model actually has the lowest value and is therefore chosen as benchmark for this customer. As shown in table 5.2, the AR(1) model is chosen for the six customers with the highest total sales volume, but is not the most appropriate benchmark for all ABM's customers.

	Benchmark
Company 1	AR(1)
Company 2	AR(1)
Company 3	AR(1)
Company 4	AR(1)
Company 5	AR(1)
Company 6	AR(1)
Company 7	Naïve
Company 8	Seasonal naïve
Company 11	Seasonal naïve
Company 12	Naïve
Company 13	Naïve
Company 14	Seasonal naïve
Company 15	Naïve
Company 16	Seasonal naïve
Company 17	Seasonal naïve
Company 18	Seasonal naïve
Company 19	AR(1)
Company 20	Naïve

Table 5.2: Chosen benchmarks (Company 1-20)

As mentioned in section 4.3 we set a minimum limit of 20 sales (minimum of 20 sales above zero) in the training data for any models to be estimated. Company 9 and 10 do not satisfy these requirements and are therefore not displayed in table 5.2. They will not be included in any further analysis.

The final benchmark for the four customers with the highest total sales volume is illustrated in figure 5.2 below.

Figure 5.2: Chosen benchmark (Company 1-4)

5.2.2 ETS

In this and the following two sections, we will illustrate the forecasts from all the different methods for the four customers with the highest total sales volume. These forecasts are for the 12 first months of the test data. For ETS, the forecasts are displayed in figure 5.3.

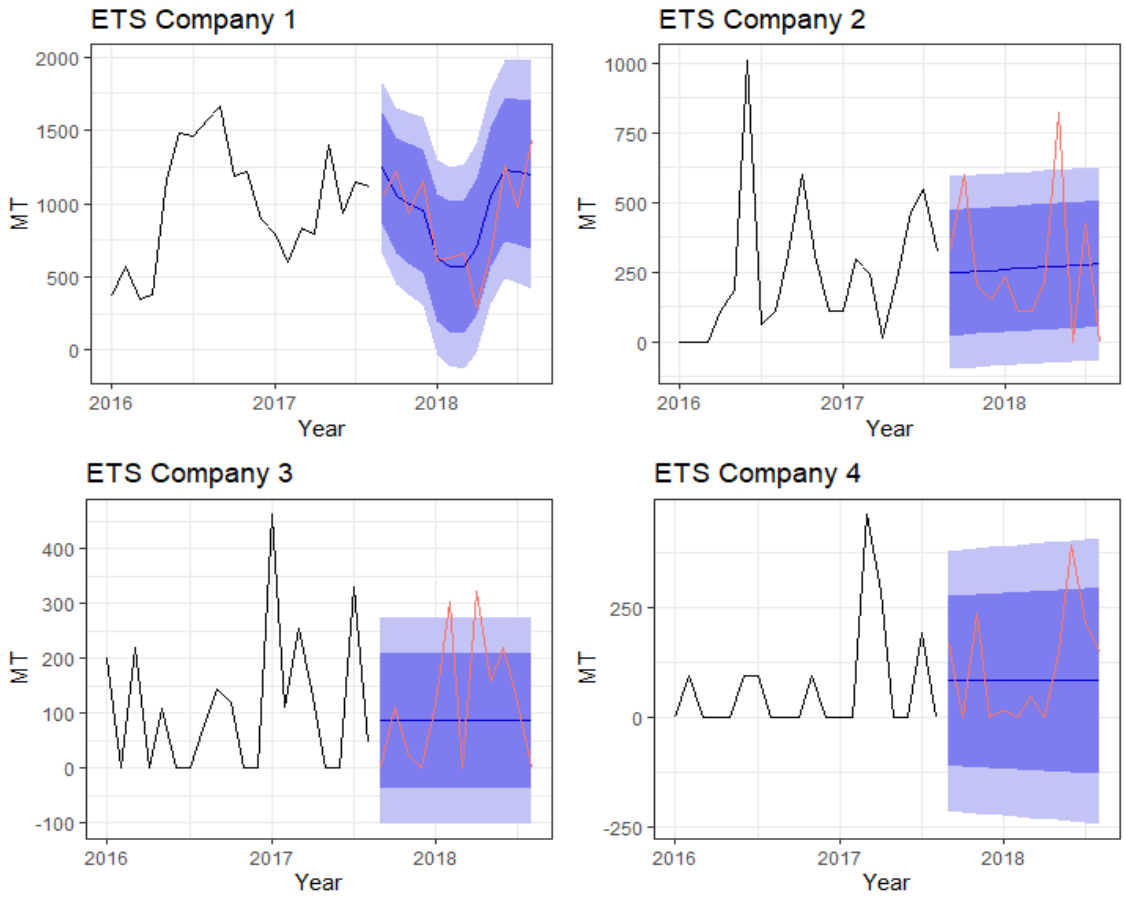

Figure 5.3: Forecasts from ETS method (Company 1-4)

From these plots, we see that for Company 1, the forecasts from the fitted model follow the test data to some extent. The seasonal pattern in the forecasts is similar to the seasonal component of the chosen model, ETS(ANA), illustrated in figure 4.7 on page 41. For the remaining three customers, the ETS method does not find a pattern to extrapolate into the future, hence the forecasts flatten out. This is since the models chosen for Company 2, 3 and 4 are AAN, ANN and ANN, respectively (see table A2.1 in the Appendix). For these models, no seasonal component is included. For Company 2, there is a small positive trend in the forecast, which implies that the model has found some positive trend in the historical data. This concurs with the chosen AAN-model, which contains an additive trend component. The forecasts from the ETS models for Company 2, 3 and 4 do however not give much value.

5.2.3 STL + ETS

For the combination method, STL + ETS, the forecasts for the four customers with the highest total sales volume are displayed in figure 5.4.

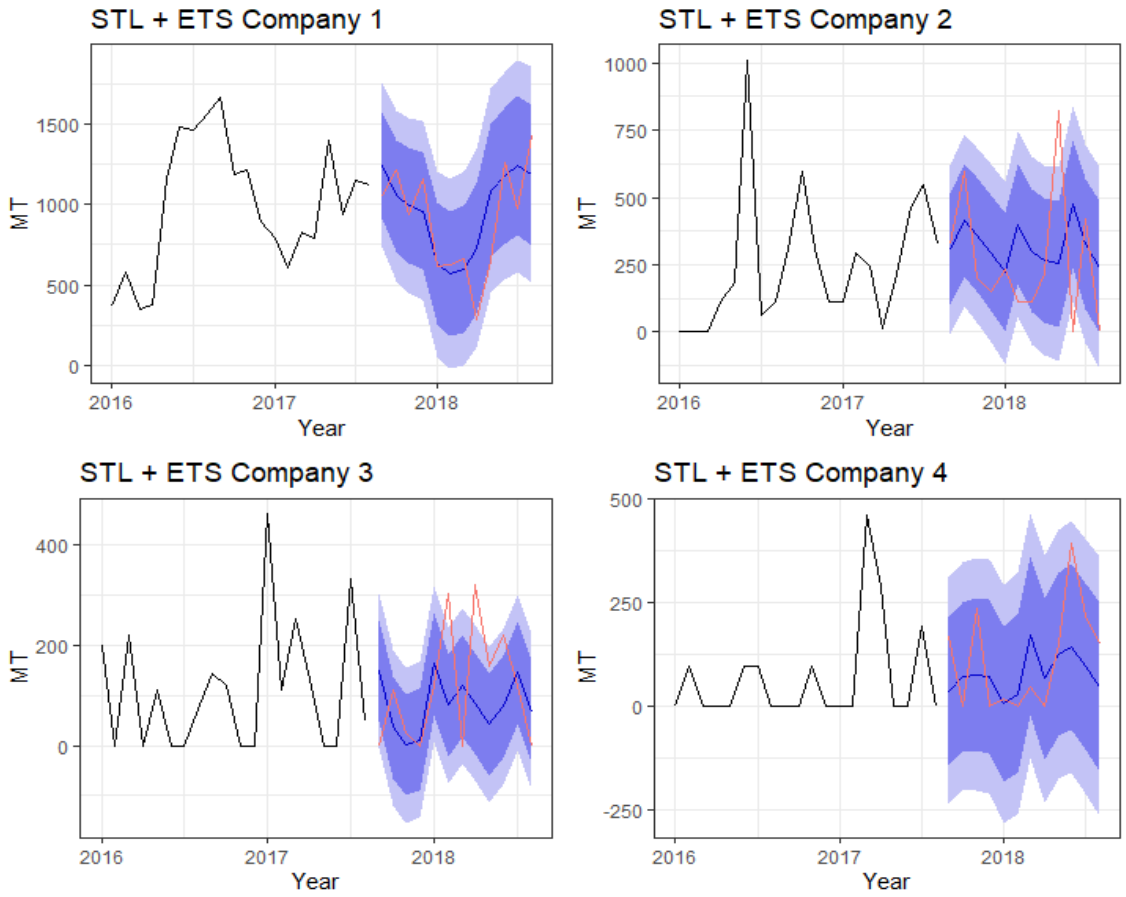

Figure 5.4: Forecasts from STL+ETS method (Company 1-4)

Unlike the ETS method, the models from the STL+ETS method are able to capture some sort of pattern for all customers. This is since the ETS method is applied to the seasonally adjusted data and reseasonalized using the last year of the seasonal component from STL decomposition, as discussed in section 3.1.2.2. For Company 1, we see that the seasonal pattern in the forecasts is similar to the seasonal component after STL decomposition, seen in figure 4.8 on page 42, similar to the ETS model chosen for Company 1. For Company 2, 3 and 4, we see that the models are not able to capture the large, random variations in the data, thus the forecasts have less variation than the test data.

5.2.4 ARIMA

For ARIMA, the forecasts for the four customers with the highest total sales volume are displayed in figure 5.5.

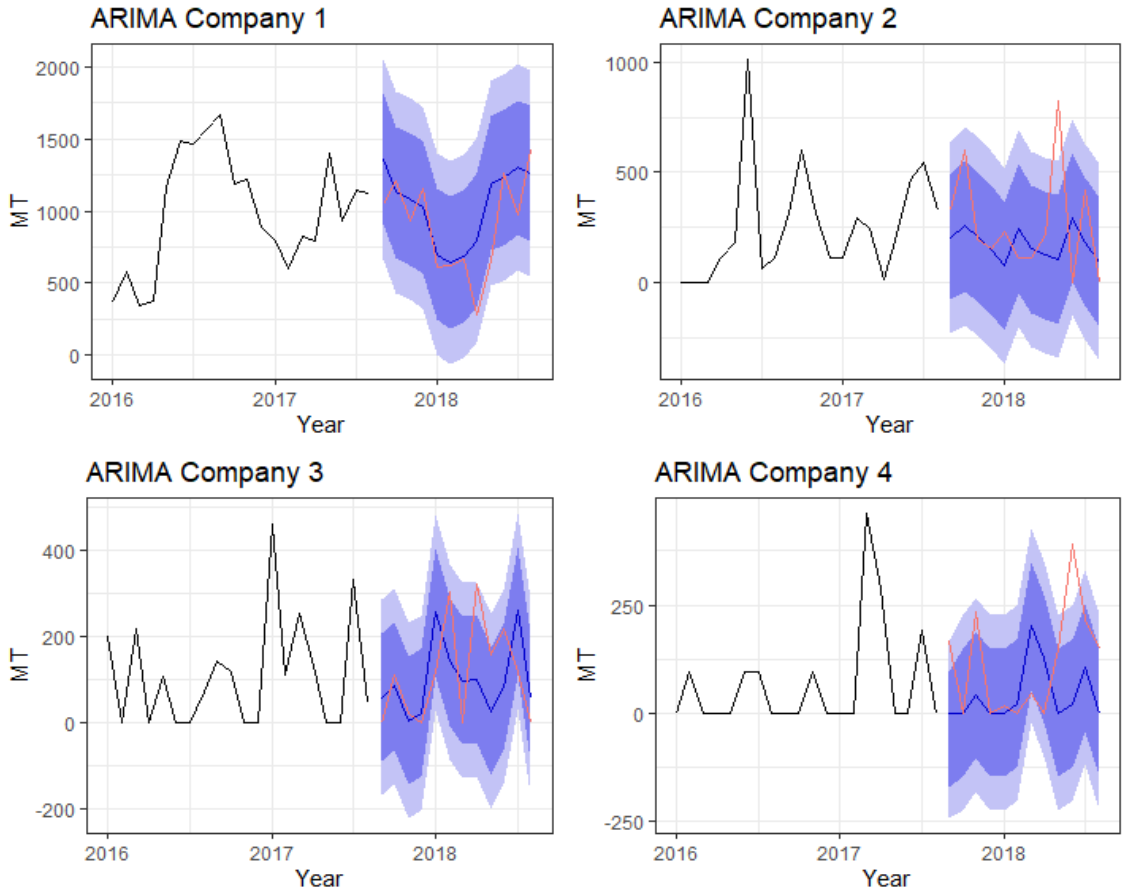

Figure 5.5: Forecasts from ARIMA method (Company 1-4)

It seems like the ARIMA model follows the pattern in the test data to some extent for Company 1. For Company 2 and 4, the forecasts from the ARIMA models are quite far off from the observed values. Similar to the models chosen for Company 2, 3 and 4 for the STL+ETS method, there is a pattern in the forecasts for these companies, but the variation is much smaller than in the test data. For Company 3, the forecasts are similar to the observed values at the start of the forecast period, but deviates after the beginning of 2018.

5.3 Time Series Cross-Validation

In order to assess which of the different methods that are best for each company, we have applied cross-validation. The model that yields the lowest 12-step-ahead MAE is chosen as the best model for the individual customers. This means that the MAE is calculated for forecasts of y_{T+12}, when the training data has T observations and averaged over all cross-validation runs.

Cross-validation has been applied by splitting the whole data set into twelve different training and test sets and averaged the 12-step-ahead error from all these, as explained in section 3.3.2. We keep the test set at 12 months for all the runs (12-step cross-validation), but increase the training set by 1 month for each run. We start with a training set of 68 months and the test set as the 12 consecutive months. The next training set will then be the first 69 months, and the test set the 12 consecutive months after this. We do this up until the training set consists of 80 months and the test set includes the last 12 months. The MAE is calculated for each horizon per test set. MAE is then averaged for each horizon over the different test sets, i.e. we get one MAE for $T + 1$, one for $T + 2$, up until $T + 12$, which is the 12-step-ahead MAE. Even though our target is to forecast demand 12 months ahead, we have chosen to compute MAE for all the 12 consecutive months after the training set. This is in order to investigate how well the chosen models perform at other forecast horizons, e.g. short-term. The MAE computed for the 12^{th} month ahead is however the criterion that determines which model that is chosen for each customer. Figure 5.6 illustrates all the 12-step MAE for the four customers with the highest total sales volume.

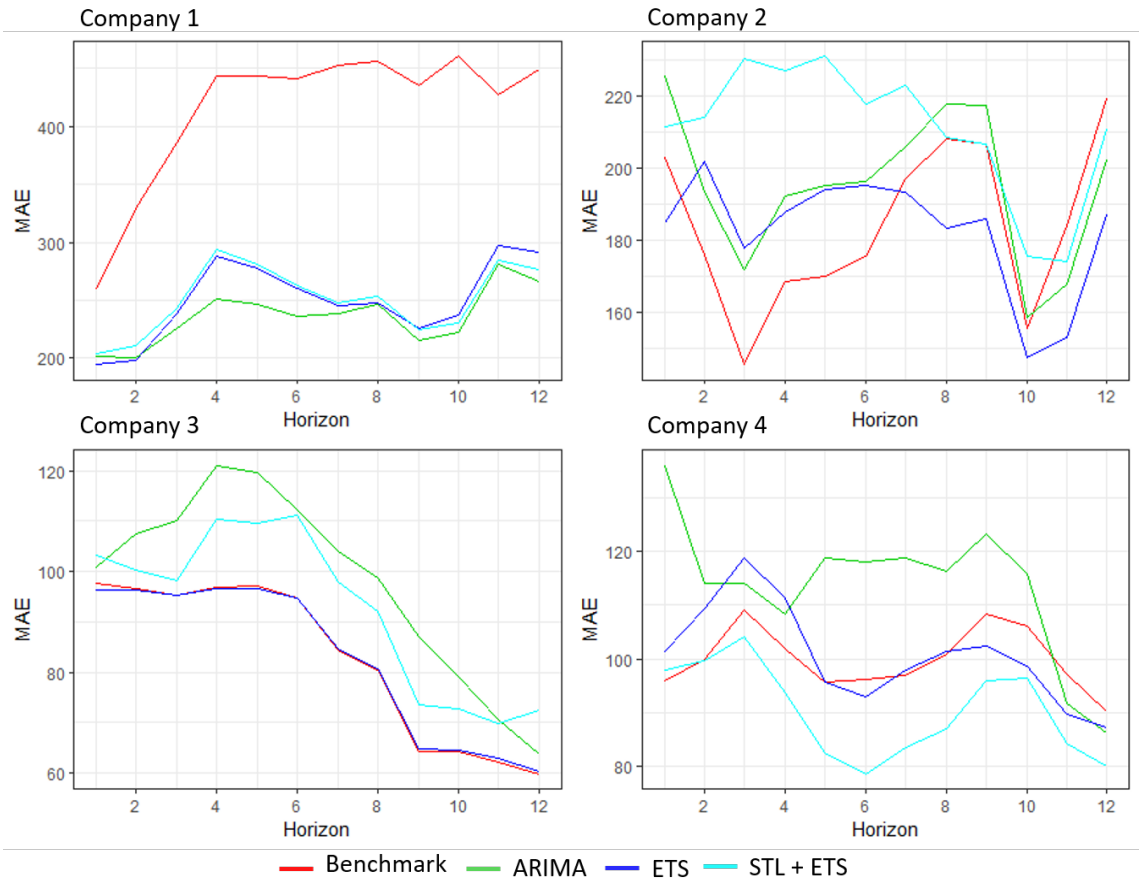

Figure 5.6: Cross-validation MAE for forecast horizon 1-12 (Company 1-4)

We see that the benchmark performs significantly worse than the other methods for Company 1. However, the benchmark performs on an equal level as the other methods for the three other customers. The benchmark yields the lowest MAE for a few forecast horizons for Company 2, as well as for Company 3 together with the ETS method. In addition, it seems like the three other methods have the lowest MAE interchangeably for the four customers.

For Company 1, the ARIMA model gives the lowest MAE for all forecast horizons, except for the 1- and 2-step-ahead horizon. Further, the ETS model yields the lowest MAE for all forecast horizons after seven months for Company 2, with the lowest error for a 10-step-ahead horizon. All the methods seem to perform better at a 10-step-ahead horizon for this customer. Thereafter, the MAE increases for the following horizons for all methods. For Company 3, the MAE is steadily declining for longer forecast horizons, with all the methods performing better for the 12-step-ahead horizon. An interesting aspect is that the benchmark and the ETS model seem to have almost identical MAE for all forecast

horizons for this customer. This is likely since the forecasts illustrated for Company 3 in figure 5.2 and 5.3, both flatten out just below 100 MT, which indicates that simpler models are more appropriate for this time series for all forecast horizons. The model chosen for the STL+ETS method is clearly performing better than the other methods for Company 4. The benchmark and the ETS model perform at a similar level, while the ARIMA model clearly performs worse than the other methods for both short-term and long-term forecasts. This is interesting, since the ARIMA method generally performs better in the short-term, as mentioned in section 3.1.3.

The object of this thesis is a 12-step-ahead forecast at a disaggregated customer level. From figure 5.6, we see that the chosen models perform relatively good at this horizon for Company 3 and 4. However, since this does not apply to Company 1 and 2, the automatic model does not select models that are generally suitable for 12-step-ahead forecasts for all customers.

The MAE for the 12^{th} month represents our final cross-validation MAE per method. The final 12-step-ahead MAE for all the methods and for the 20 customers with the highest total sales volume are displayed in table 5.3. By doing cross-validation, we are able to capture how the different methods perform on different parts of the data, and thus hopefully identify the method that generalizes best to the time series for each customer.

	Benchmark	ARIMA	ETS	STL + ETS
Company 1	448.87	265.82	291.45	276.42
Company 2	219.30	202.27	186.94	210.79
Company 3	59.83	63.63	60.28	72.56
Company 4	90.29	86.27	87.09	80.06
Company 5	155.85	166.34	156.34	156.20
Company 6	79.78	85.39	96.71	81.69
Company 7	0	32.53	0	20.18
Company 8	51.94	41.36	23.51	33.49
Company 11	54.57	51.61	41.60	40.68
Company 12	26.00	18.76	22.85	18.22
Company 13	20.02	15.20	14.00	12.86
Company 14	40.33	43.11	42.17	36.45
Company 15	31.04	42.16	40.45	44.76
Company 16	34.41	31.91	25.06	32.56
Company 17	16.82	22.09	22.70	16.40
Company 18	37.54	35.37	29.95	34.44
Company 19	20.48	14.79	16.51	15.10
Company 20	8.33	21.24	21.06	20.49

Table 5.3: 12-step-ahead MAE (Company 1-20)

First of all, we see that the MAE ranges from 0 to almost 450 for the different customers. This is since the customers make purchases in different ranges. The lowest MAE for each customer is highlighted in blue, where the STL+ETS method yields the lowest error most often. Thereafter, the chosen benchmark performs better than the other methods for several customers. The forecast error is relatively high for all methods, but the benchmarks and the STL+ETS method seem to be the preferred methods for the majority of the customers. An interesting aspect is that the ARIMA method rarely outperforms the other methods. For the majority of the customers, there are however only marginal differences between the different methods.

Regarding Company 20, the benchmark seems to perform quite a lot better than the other methods. By looking at the data for Company 20, figure 5.7, we see that the whole test set consists of zeros. Since the benchmark forecasts zeros for the whole forecast horizon, the forecasts and the observations in the test set are exactly alike. From the plots, it looks like the other methods have chosen models that give more complex forecasts, which explains why the benchmark is superior for this customer.

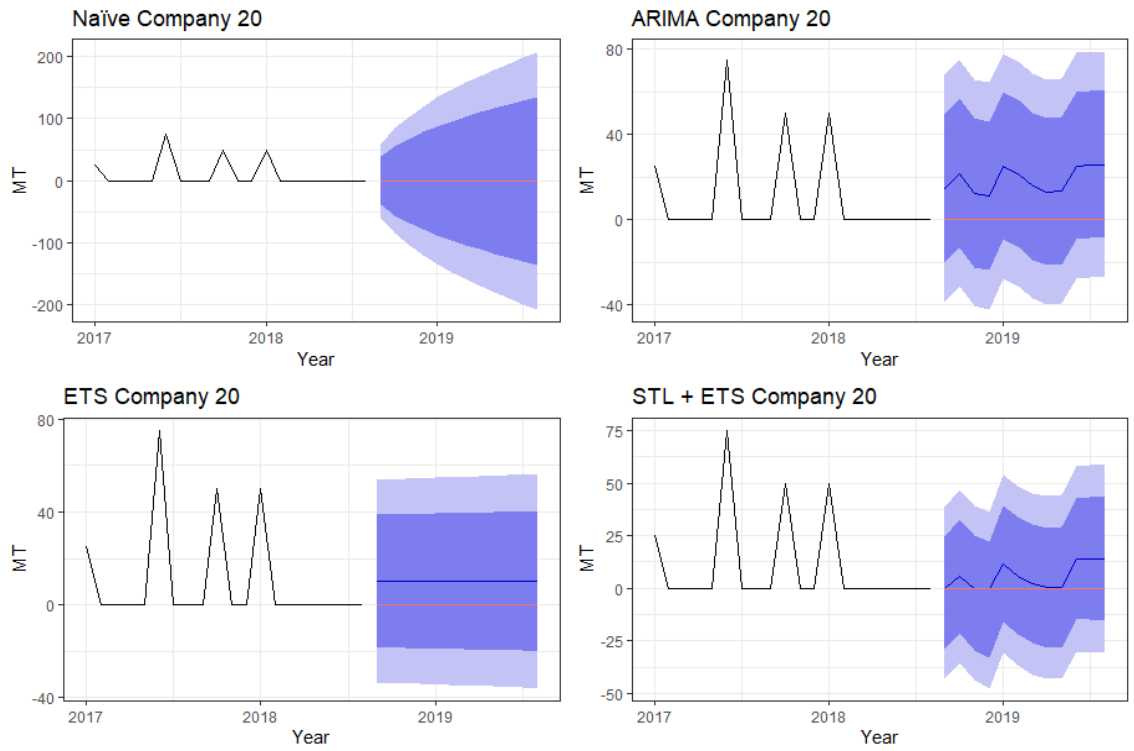

Figure 5.7: Forecasts from all methods (Company 20)

In figure 5.7, we can clearly see some of the characteristics of the different methods explained in the subsections under section 3.1 and section 5.2.1. Firstly, the naïve method simply forecasts zero for the whole forecast period. This is since this method sets all forecasts to be equal to the value of the last observation. Further, the forecasts from the ETS method flatten out at a level just above zero, which is likely since this method use weighted averages of past observations to forecast new values. The weights decrease exponentially as the observations get older, and the newest observations are all zero, which is likely why the fluctuations earlier in the training data are not extrapolated into the forecast period. For the ARIMA method and the STL+ETS method, which are more complex methods, it looks like the chosen models try to recreate some of the fluctuations in the forecast period, but that these are adjusted to a lower level due to the long period of zero values at the end of the training data. This indicates that the more complex models are too advanced for this customer and that simpler models who are easier to compute are preferred.

Regarding Company 7, we see that both the benchmark and the ETS model are chosen as the best model with a forecast error of zero. By looking at the plot of the time series for

this customer, see the top-left plot in figure 4.5 on page 37, we clearly see that Company 7 has not made any purchases since the end of 2015, and thus the last three years of the training data for this customer only consists of zero values. This explains why the chosen models, the benchmark and the ETS model, just forecast zeros for all the different test sets and hence yield a cross-validation error of zero. Since the first 68 months include all the positive values for this time series, there will never occur any positive values for any of the test sets when doing cross-validation for this customer. The benchmark or the ETS model is therefore the best model for this time series. As discussed in section 3.1.1, we prefer the simpler methods, and the benchmark would therefore be chosen here. This does however not give much value for ABM, as they are probably already aware of the fact that this is an inactive customer, and therefore do not expect a sale to this customer.

5.4 Model Evaluation

5.4.1 Overfitting

In the previous section, 5.3, we displayed the resulting forecasts for Company 1-20. The models are chosen based on the automatic modeling procedure and do not include any human discretion. Due to the given restrictions for the ARIMA components as well as our choice of information criterion, the automatic model may choose more complex models than necessary. This makes the models prone to overfitting. Therefore, we will take a closer look at the training and test MAE for all the customers. The training MAE is computed on the 68 observations in the training data, while the test MAE is the same cross-validation MAE displayed in table 5.3. The training and test MAE for the chosen models are therefore highlighted in blue in table 5.4 as well.

	Benchmark		ARIMA		ETS		STL + ETS	
	Train	Test	Train	Test	Train	Test	Train	Test
Company 1	274.08	448.87	192.94	265.82	209.97	291.45	203.67	276.42
Company 2	124.58	219.30	105.20	202.27	125.58	186.94	109.14	210.79
Company 3	75.02	59.83	31.86	63.63	74.27	60.28	57.89	72.56
Company 4	97.33	90.29	46.32	89.12	110.46	87.09	101.28	80.06
Company 5	76.09	155.85	59.68	166.34	60.03	156.34	59.16	156.20
Company 6	43.61	79.78	35.15	85.39	35.19	96.71	36.88	81.69
Company 7	54.72	0	22.42	37.39	66.26	0	68.35	20.18
Company 8	27.34	51.94	14.18	39.76	24.54	23.51	18.61	33.49
Company 11	32.41	54.57	12.21	51.61	26.62	41.60	21.11	40.68
Company 12	41.01	26.00	17.34	18.76	34.14	22.85	28.30	18.22
Company 13	17.34	20.02	4.74	15.20	15.20	14.00	11.64	12.86
Company 14	32.27	40.33	7.59	43.11	29.64	42.17	24.71	36.45
Company 15	40.12	31.04	16.69	38.65	30.67	40.45	27.04	44.76
Company 16	28.93	34.41	7.03	31.91	27.84	25.06	21.23	32.56
Company 17	26.88	16.82	6.73	22.09	19.74	22.70	17.81	16.40
Company 18	15.07	37.54	11.36	35.37	11.01	29.95	11.46	34.44
Company 19	13.48	20.48	5.28	14.79	13.24	16.51	9.60	15.10
Company 20	22.00	8.33	9.34	21.24	17.35	21.06	16.89	20.49

Table 5.4: Training MAE and 12-step-ahead test MAE (Company 1-20)

From table 5.4, we see that for many of the customers, the test MAE for the chosen method (in blue) is much higher than the training MAE. This indicates that the model performs well on the training data, but does not generalize well, as discussed in section 3.2.2. For Company 2 and 14, the test MAE is almost 50% higher than the training MAE, which indicates that the models for these time series are likely overfitting, but not to the same extent as for Company 5, 6, 11, 18 and 19. As an example, the best model for Company 19 gives a test MAE that is almost three times as high as the training MAE. This implies that the chosen ARIMA model will not generalize well to new data, and is thus not a good model for forecasting this time series. These models are therefore too complex for the data, and simpler models would likely be more appropriate. This is evident in the chosen ARIMA model for Company 19: ARIMA(2,0,2)(1,1,0) (see table A2.1 in the Appendix), which has a significant amount of parameters and is clearly overparameterized for this time series.

For some of the other customers, the test MAE is however much lower than the training MAE, especially for Company 7 and 20. For these two particular customers, this is a result of many zeros in the data as explained in section 5.3. For other customers, it could

be that by random chance the model chosen fits better to the test data than the training data. Generally, it is difficult to interpret the behavior of the models when the test error is lower than the training error.

5.4.2 Relative Model Performance

As discussed in section 5.1, we compute MASE for the same 20 customers in order to compare the relative model performance across customers. The MASE is computed by scaling the test MAE relative to the respective training MAE from the naïve method for each cross-validation run. The MASE for each customer can be seen in table 5.5.

	Benchmark	**ARIMA**	**ETS**	**STL + ETS**
Company 1	0.885	0.580	0.625	0.589
Company 2	0.980	0.948	0.809	0.980
Company 3	0.591	0.666	0.594	0.706
Company 4	0.805	0.745	0.797	0.716
Company 5	1.109	1.306	1.323	1.222
Company 6	1.490	1.555	1.787	1.482
Company 7	0.000	0.505	0.000	0.316
Company 8	1.465	1.180	0.699	1.052
Company 11	1.874	1.642	1.222	1.259
Company 12	0.748	0.527	0.637	0.527
Company 13	1.391	1.025	0.930	0.833
Company 14	1.207	1.296	1.250	1.083
Company 15	1.008	1.262	1.290	1.339
Company 16	1.178	1.069	0.821	1.072
Company 17	0.945	1.239	1.262	0.886
Company 18	1.880	1.811	1.533	1.779
Company 19	1.163	0.808	0.876	0.793
Company 20	0.385	0.958	0.951	0.934

Table 5.5: 12-step-ahead MASE (Company 1-20)

In section 3.3.3, we discussed that a MASE lower than one indicates that the model gives forecasts that improve on the average naïve forecast computed on the training data. From table 5.5, we see that this applies to eleven of the customers, in addition to Company 7 and 20, which are not representative, since the naïve method is chosen as the best model for these customers. For Company 1, 3, and 12, the MASE is around 0.5, which is a significant improvement. This also applies to Company 4 and 8, with a MASE of approximately 0.7. This implies that the models chosen for these customers are preferred over the average naïve method. Company 2, 13, 16, 17 and 19 all have a MASE around

0.8. These models therefore produce forecasts that also improve some on the average naïve method.

As discussed in section 5.4.1, for the models that give a relatively low MASE, the one for Company 2 and 19 seem to overfit to the data. However, this does not regard the models chosen for Company 3, 4, 8, 12, 13, 16 and 17. Company 8, 13, 16 and 17 all have quite similar training and test MAE, which can indicate that these models generalize well to new data. The models chosen for Company 3, 4 and 12 have much lower test MAE than training MAE, which makes it hard to interpret the behavior of these models. For the models with $MASE > 1$, the simple naïve method would be preferred over the models chosen through the automatic modeling procedure.

5.4.3 Model Bias

In section 3.3.3, we briefly mentioned how the mean error, or ME, can be used to measure model bias. Since ME is computed by subtracting the forecast from the observed value for all t, a positive ME indicates that the model in general underestimate the true value, while a negative ME indicates a model that mostly overforecasts the observed value. An ME of zero indicates that the model is unbiased. The calculated ME for the chosen models are shown in table 5.6.

	Benchmark	ARIMA	ETS	STL+ETS	Mean
Company 1	308.74	-22.38	39.92	38.79	639.19
Company 2	163.86	144.26	23.74	23.74	143.74
Company 3	9.81	10.91	10.22	10.22	86.93
Company 4	3.25	18.27	4.62	4.62	87.88
Company 5	68.54	13.10	1.41	1.41	57.21
Company 6	27.15	8.85	12.42	12.42	50.04
Company 7	0.00	-19.71	0.00	0.58	69.06
Company 8	-36.20	-31.16	-18.66	-18.51	45.74
Company 11	3.70	7.53	11.74	17.42	32.28
Company 12	22.00	2.94	-5.38	14.21	39.19
Company 13	12.00	8.25	7.26	4.8	30.76
Company 14	18.33	18.49	15.72	16.96	30.11
Company 15	13.90	11.29	15.01	15.01	23.33
Company 16	-20.54	-19.00	-17.31	-18.44	28.23
Company 17	13.10	15.72	20.12	3.23	19.22
Company 18	4.21	6.16	5.79	5.45	13.12
Company 19	13.50	1.38	5.51	4.29	13.08
Company 20	-8.33	-21.22	-21.06	-20.49	21.06

Table 5.6: 12-step-ahead ME (Company 1-20)

From table 5.6 it can be seen that the ME is mostly positive for the different customers. This means that the chosen models more often underestimates the true value. Company 7 has an ME of 0 due to the zero values in the data. Many customers have quite high ME compared to their respective in-sample mean sales volume. In general, it also seems like the chosen models often under- or overforecasts demand, since the majority of the customers have an ME that deviates significantly from zero. However, the ME is averaged over 12 cross-validation runs, and to investigate whether the models are truly biased, we calculate a 95% confidence interval on ME for a selection of customers, see table 5.7.

	ME	Lower bound	Upper bound
Company 1	-22.38	-239.47	194.71
Company 2	23.74	-133.33	108.81
Company 3	9.81	-35.89	55.51
Company 4	4.62	-66.29	66.80
Company 5	68.54	-96.16	233.25

Table 5.7: 95% confidence interval on ME for the chosen method (Company 1-5)

Estimates of the confidence intervals are useful since the estimate of the ME varies from sample to sample (Filliben & Heckert, 2003), and is as discussed in section 5.3 computed

for 12 different test sets. The narrower the interval, the more precise the estimate of ME. For all the five customers, the confidence interval includes zero, which means that the ME for these customers is not significantly different from zero (Coolidge, 2006). This indicates that the models are unbiased. The confidence interval also gives an indication of how much uncertainty there is in our estimate of ME. For the five customers whose confidence intervals have been computed, the interval is quite wide. This means that the estimated MEs are very uncertain.

One important note on the evaluations done in this chapter is that the test data is relatively small, which means that the performance estimates are less reliable. This is since the more test data, the more accurate the error estimate will be. With more test data we would have been able to get more accurate estimates of the different performance measures discussed.

In this chapter, we have seen that the forecasts produced for the different customers in general give a high forecast error. As discussed in section 1.1, good demand forecasts can contribute to higher quality on sales- and financial forecasts, and can be used as supplementary guidance to the sales force. For ABM, the object of the forecasts produced through this thesis was to be used for these purposes. However, since the accuracy is quite low, the results must be interpreted with caution. This means that the sales force could use the forecasts as a tool for visualizing potential future purchasing patterns and get an indication of when sales are more likely to occur. The forecast values alone should however not be used for financial and strategic planning and resource allocation.

6 Discussion

The main target of forecasting time series is to predict an uncertain future. This can be done by constructing a suitable model based on an analysis of the historical development in a time series. In this thesis, we have explored three types of forecasting methods: exponential smoothing (ETS), a combination of STL decomposition and ETS, as well as ARIMA. These methods recreate patterns from the time series in different ways and project these for a specified period of time into the future. The purpose of the fitted model determines whether it is more or less appropriate than other models built for the same data. The comparison between our candidate models in section 5.3 is an attempt to choose the most appropriate model for demand forecasting of krill meal based on time series data over a seven-year period. However, as circumstances alter, the choice may change, and additional data and knowledge of the industry could therefore yield a different conclusion.

6.1 Overall Findings

Our analysis shows that none of the investigated methods perform well on time series data of krill meal. In section 3.2.2, we discussed how an adequate sample size is necessary for constructing good models, especially when a large number of parameters need to be estimated. As emphasized in section 4.1.1, the time series used in this thesis are quite scarce, with only 92 historical observations for each customer, out of which only 68 are used to fit the models. This is probably the main explanation as to why the common forecasting methods yield a high forecasting error for this data. Through this thesis, we have however found that the simple benchmarks often outperform the more advanced methods in terms of forecast accuracy. In section 5.4.2, we also discussed how the very simple average naïve method is preferred over the chosen models for the time series of several customers. This indicates that simple forecasting methods may be more suitable to model time series data of krill meal.

Even though the forecasting errors are generally quite high, the STL+ETS models and the benchmarks seem to often produce forecasts that yield a lower forecasting error than the other candidate models. The simple models are likely performing better due to their

simplicity, considering the high variance in the data. The relative performance of the STL+ETS method can be rooted in the fact that the ETS method, and thus the STL+ETS method, generates reliable forecasts for a wide range of time series, as mentioned in section 3.1.2. Further, the STL+ETS method can handle any type of seasonality and is also robust to outliers, as discussed in section 3.1.2.2. This can explain why the STL+ETS models often outperform the ETS and ARIMA models in terms of forecast accuracy. However, this is to some extent contradicting, since the ARIMA models are also supposed to handle seasonal data, but produces quite poor forecasts for our data.

ARIMA generally performs worse than the other methods for all the customers investigated. This could be explained by the fact that the time series investigated in this thesis in some ways contradicts the prerequisites for building a good ARIMA model. The ARIMA method, as the other methods, requires a sufficient amount of data, especially when estimating many parameters. As mentioned in section 3.1.3.4, seasonal ARIMA models can potentially have a large number of parameters, which amplifies the importance of trying out a wide range of models when fitting to data and also of using an appropriate criterion to choose the best model. Another information criterion, e.g BIC or AIC, could have given different results and thus perhaps better models for the time series investigated in this thesis. Further restrictions of the ARIMA components in order to yield smaller models might also give better forecasts, since simpler ARIMA models may perform better considering our short data, with reference to the discussion in section 3.1.3.3 and 3.2.2. On the other hand, with richer data one could benefit from larger models.

As discussed in section 5.3, the different methods seem to perform relatively well for a 12-step-ahead forecast for some customers, and better for one-step-ahead forecasts for others. It is therefore not obvious whether any of the methods work better in the short- or long-term for our data. For the customer with the highest total sales volume, Company 1, the forecasts are better in the short-term. However, the 12-step-ahead forecasts give the lowest MAE for both Company 3 and 4. The methods investigated in this thesis, for the given time series, do therefore not give accurate forecasts at the desired horizon.

6.2 Limitations

In our analysis, the forecast errors have only been calculated for a selection of the customers (approximately 10 %). We therefore implicitly assume that the results computed for the time series of that selection also apply to the remaining customers. In general, the MAE of the researched customers is relatively high. This indicates that the examined models may not give reasonable forecasts of future demand for krill meal. There are several reasons why these models may not perform sufficiently well for this purpose.

6.2.1 Zero Values and Scarce Data

First of all, as emphasized at the end of section 4.1.1, the data contains a lot of zero values. This may be problematic in the parameter estimation and the model selection processes. Since the data contains a lot of sudden drops to zero, it may lead to forecasts that are upward biased in the period directly after a non-zero demand. On the other hand, if zero values are the main reason for the poor forecasts, the model chosen for Company 1 should be able to produce more accurate forecasts, since this time series contains almost no zero values. This indicates that the length of the time series could play a more significant role in terms of poor forecasts, or that time series of sales volume of krill meal is not suitable for demand forecasting. The data examined in this thesis are somewhat scarce and hence observed over a relatively short time period. As a result, the data may be insufficient to capture underlying patterns. This makes it more challenging to estimate good models using common forecasting methods. As a result of this, the forecasts from the automatic modeling procedure may not give an accurate picture of future demand at a disaggregated customer level.

Short time series complicates both the development and verification of models that aim to produce seasonal forecasts. For example, the choice of starting values becomes critical (Hyndman & Kostenko, 2007). As some of the customers in our data may have seasonal behavior, it is not unlikely that we have insufficient data to produce good forecasts. It could therefore be reasonable to use other available information in addition to the data itself. Moreover, when there is a lot of randomness in the data, the minimum statistical requirements discussed in section 3.2.2 will be insufficient to estimate seasonal models. As discussed in section 4.1.1 and 4.1.2, there are several customers with no clear pattern,

and there is likely a lot of randomness in our data, which implies that many of our models may be estimated on insufficient data. Since our data are quite scarce and also contain random variation, we may be estimating models with too many parameters compared to the amount of data available, as seen in the discussion in section 5.4.1. This can again lead to overfitting, as discussed in 3.2.2, and hence models that fit well to the training data, but do not generalize well to the test data. This could be one explanation as to why our forecasts perform quite poorly on test data.

One measure to avoid overfitting is to apply cross-validation in the model selection process. This could potentially have led to fewer parameters being estimated for the chosen models. However, to maintain sufficient data for both training and testing, we did not find it reasonable to do an additional partitioning of the data and chose to only apply cross-validation when selecting between the candidate models from the different forecasting methods. Another possible measure to avoid overfitting could be to use BIC as information criterion instead of AIC_c. In section 3.3.1, we discussed that BIC imposes a stronger penalty for each additional parameter added to the model, which would likely have led to estimation of fewer parameters in the chosen models, and thus probably less overfitting.

6.2.2 Measuring Demand Through Sales Volume

In section 1.1, we emphasized the utility of demand forecasting within the krill industry. One important note on the modeling procedure of this thesis is that historical sales volume is used to represent demand. This implies that there must always be inventory, or krill available. If ABM sells all their harvest and have nothing stored, sales volume will not necessarily represent true demand, as it will not capture if demand is higher than what is actually sold. As a consequence, our forecasts may be underestimating true demand. Further, our models may fail to capture underlying demand patterns like trend or seasonality, if such patterns are not represented by the sales volume. However, even though ABM sells all their harvest, they do store krill and distribute it between harvests, which makes it more likely that sales volume captures the essence in the underlying demand. On another note, the krill industry is restricted by harvesting limitations, which makes it hard to use sales volume to represent demand, as there likely is a higher demand than the sales volume suggests. This is especially prominent in ABM, since they offer a

premium product in a large market and only have a small market share. Despite this, it is still interesting to apply demand forecasting to this industry, for strategic planning and resource allocation, e.g. allocation of a harvest between the relevant customers, as mentioned in section 1.1. One interesting aspect for further research could be to include qualitative forecasting with the quantitative forecasts from this thesis. In that way, one could implement expert knowledge from e.g. highly experienced employees to provide insights into future outcomes, and thus produce better forecasts.

6.2.3 Excluding Internal and External Factors

In this thesis, we have chosen to only use the variable itself as a predictor and excluded external factors that may affect the demand for krill meal. One of the main reasons for this is since statistical forecasting methods extrapolate time series features like seasonality and trend, and use these to forecast future demand, as discussed in section 3.1. Further, ABM operates all over the world and has customers with very different attributes. This makes it difficult to identify the most relevant leading indicator variables for the demand of a particular customer. However, Sagaert, Aghezzaf, Kourentzes & Desmet (2017) successfully built a model that automatically identified the key leading indicators that drive sales from a massive set of macro-economic indicators for the tire industry. It would therefore be interesting to investigate the effects of building a similar model for the krill industry. Since krill is a much used ingredient in fish farming, it would be both interesting and relevant to include time series of e.g. demand and price of farmed salmon, or even factors that affect the living conditions of krill like ocean temperature and other climatic conditions. Despite this, we decided to restrict the work related to our thesis to investigate the potential of univariate statistical methods on time series of historical sales volume.

6.3 Implications of Automatic Modeling

Another important aspect of the modeling procedure in this thesis is that the aim is to build an automatic model that can be applied to all the different customers separately. As a result, we let a lot of decisions be made based on certain criteria, to remove the human judgment aspect of the procedure. This means that there is no room for discretionary assessment. The data contains a variety of customers with different attributes and behavior.

This makes it extremely difficult to construct a model that will be applicable to each of the customers separately. For some customers, we find both seasonality and trend, while others seem to buy completely arbitrary. Some have not purchased anything for several years, while others have only recent purchases. When evaluating these differences over the same time period and by the same evaluation criterion, it leads to generalizations of the customers that may not be justified.

The many challenges of building an automatic forecasting model are not exclusively related to customers with different behavior and scarce data. The very automatic aspect is causing obstacles itself. When all choices must be based on different criteria, we must choose which criterion that will capture the essence of the time series most accurately. One of these choices is to check for stationarity in the data. As discussed in section 4.4.2.1 and 4.4.2.2 we choose to use the KPSS- and ADF test, as well as the CH- and the HEGY test, when determining the number of differences to apply. The different tests may yield different answers and are more or less suitable for different data. Our choice of such tests can therefore affect the parameter estimation and can potentially result in exclusion of what could have been a good model for a specific customer.

6.4 Potential Improvements

6.4.1 Clustering Customers

In section 4.1.1, it was briefly mentioned that the different customers have different purchasing patterns, and buy in quite a different range of quantities. There are also customers from all over the world, operating in different markets and under varying circumstances and environments. Due to this, it would have been interesting to group or cluster the different customers together. This could have solved some of the issues related to building an automatic model, as it would have been possible to set different criteria for the different clusters, to more accurately customize the modeling procedure according to the group's characteristics. However, as discussed in section 4.1.2, many of the time series for the different customers seem random and there are few clear similarities across them. This makes it hard to group similar customers. On the other hand, this could potentially be explored through machine learning by doing clustering analysis. In

addition, the customers could perhaps have been clustered based on external factors like origin, market share, etc. However, this would require significant knowledge of both the different customers and the markets they operate in. Considering the scope of this thesis, we did not see collection of such information for almost 200 customers as reasonable. It could however be an interesting area for further research on demand forecasting of krill meal.

6.4.2 Inputting Richer Data

We have used time series for a seven-year period as input in our automatic model to produce forecasts of demand for krill meal for 12 months ahead. As discussed in section 6.2.1, some of the reason for the poor results can likely be linked to the scarce data. We also amplified in section 5.1 that when dealing with bad or little data, the source of a high error rate could be bad data, and not necessarily bad forecasting. It would therefore be especially interesting to investigate how our automatic model would perform on richer data. By inputting data for a longer time period, one could be able to more accurately determine whether the forecasting methods investigated in this thesis are not suitable to forecast demand for krill meal, or if the data used in this thesis are too scarce to represent the demand in the industry. On the other hand, the krill industry is a very young industry, which results in short data in general. It is also likely that the krill industry is characterized by random variations and changing demand. This could be rooted in the harvesting restrictions discussed in section 2.1, as well as the possibilities to store krill over time, which may affect the customers' purchasing patterns. In that case, inputting richer data to our automatic model would likely not give more valuable results than those already provided in this thesis.

6.4.2.1 Aggregated sales volume

In section 1.1 we discussed how forecasts at a disaggregated level are desired both due to different demand patterns for the customers, as well as for planning purposes. On the other hand, sometimes aggregating different sources of uncertainty can result in better forecasts. It could therefore be useful to investigate the results of implementing the total aggregated sales volume of krill meal to the automatic model. The aggregated sales volume does not contain any zeros, which eliminates the issues related to sparse

data. As a result, the forecasting methods applied in this thesis might produce better forecasts at an aggregated level. However, the time series for Company 1 contains few zeros and has a similar time series pattern as the aggregated sales volume. Since the chosen model for Company 1 does not produce accurate forecasts, this may indicate that using aggregated sales volume will not give better forecasts. Considering the target of our automatic modeling procedure; to produce forecasts at a disaggregated customer level, we have not investigated the possibilities at an aggregated level.

6.4.3 Other Forecasting Methods

So far we have seen that the forecasting methods investigated in this thesis seem to have trouble modeling sparse data with many zero values. This may indicate that these methods are not suitable to forecast demand for krill meal. As discussed at the end of section 4.1.1, forecasting intermittent demand entails several challenges, one of which is that many zero values may render common forecasting methods difficult to apply. It is therefore likely that forecasting methods that are specifically developed to deal with intermittent demand, or that for other reasons are more robust to such data, will perform better on our type of time series. An example of such a method is Croston's method (1972), developed to forecast products with intermittent demand, and overcome difficulties related to forecasting intermittent demand using general forecasting methods. Further, bootstrapping and temporal aggregation, as well as discrete- and integer-valued ARIMA models, have also been proposed. On the other hand, as argued in section 2.2, the methods investigated in this thesis have shown to perform well in fishery harvesting and other industries in which we have seen similarities to the krill industry. Since these methods are also popular forecasting methods for a variety of purposes, we found it reasonable to initialize the research within demand forecasting of krill meal with these common methods.

7 Conclusion

Throughout this thesis we have investigated the effect of applying different forecasting methods to time series of historical sales volume of krill meal. We have used exponential smoothing (ETS), decomposition with ETS (STL+ETS) and ARIMA, and compared these to simple benchmarks: naïve, seasonal naïve or AR(1). Our main findings are that in general, none of these methods are able to produce accurate 12-step-ahead forecasts for demand for krill meal. However, the STL+ETS models produce forecasts with the lowest error more often than the other methods investigated. The benchmarks perform well approximately equally often as the STL+ETS method, and since simpler models are preferred, this could indicate that the benchmarks are the most appropriate for a majority of the customers investigated. ARIMA models seem to perform poorly overall, which is likely since the chosen models include too many parameters. There is no clear pattern as to whether the methods work better for shorter or longer forecast horizons. Therefore, it would be interesting to investigate shorter forecast horizons for the krill market in further depth.

In conclusion, it therefore seems like time series data of historical sales volume can not be used to produce reasonable forecasts of demand for krill meal for neither ABM nor the krill industry. For ABM, this implies that the forecasts alone should not be used for resource allocation and strategic planning. Careful interpretation by employees with domain knowledge for a specific customer is therefore necessary, for the forecasts to be somewhat useful. Further, the data used in the modeling procedure of this thesis may not be representative for the krill industry, and we can therefore not conclude on an overall industry level. We also found that there may be other forecasting methods that will produce better forecasts with the same data as input, which would be an interesting area for future research within demand forecasting of krill meal.

It is hard to determine whether the poor forecasting results are due to the scarce data, a non-suitable automatic modeling procedure or if historical sales volume of krill meal simply can not be used as input to make sensible forecasts. As discussed, none of the classical forecasting methods, implemented in the automatic model, work well for demand forecasting of krill meal. Through this thesis, we therefore highlight the opportunity to

develop methods that can work for this industry. Moreover, we accentuate the fact that there are many unexplored areas within demand forecasting of krill meal. We emphasize that several adjustments could be done to our automatic model which could improve the performance of the chosen forecasting methods. Among these, we consider judgmental qualitative forecasts, investigating the possibilities of including several forecast variables, clustering companies and experimenting with other forecasting methods as most relevant. Considering the value krill has in handling the food production challenges, it is important to do more research within demand forecasting of krill meal.

References

Aker ASA (2018). About Aker. Last accessed October 23, 2019: https://eng.akerasa.com/About-Aker/History.

Aker BioMarine (2016). Annual report 2016.

Aker BioMarine (2018). Annual report 2018.

Anvari, S., Tuna, S., Canci, M., & Turkay, M. (2016). Automated Box–Jenkins forecasting tool with an application for passenger demand in urban rail systems. *Journal of Advanced Transportation*, *50*(1), 25–49. https://doi.org/10.1002/atr.1332.

Arltová, M. & Fedorová, D. (2016). Selection of Unit Root Test on the Basis of Length of the Time Series and Value of AR(1) Parameter. *Statistika*, *96*(3), 47–64. https://www.czso.cz/documents/10180/32912822/32019716q3047.pdf/09710b90-e1d0-4bb1-816e-5b83faad686b?version=1.0.

Asche, F., Bjørndal, T., & Gordon, D. V. (2007). Studies in the demand structure for fish and seafood products. In A. Weintraub, C. Romero, T. Bjørndal, & R. Epstein (Eds.), *Handbook of Operations Research in Natural Resources* chapter 15, (pp. 295–314). New York: Springer.

Athiyaman, A. & Robertson, R. (1992). Time Series Forecasting Techniques: Short-term Planning in Tourism. *International Journal of Contemporary Hospitality Management*, *4*(4), 8–11. https://doi.org/10.1108/09596119210018864.

Atkinson, A., Siegel, V., Pakhomov, E., Jessopp, M., & Loeb, V. (2009). A re-appraisal of the total biomass and annual production of Antarctic krill. *Deep Sea Research Part I: Oceanographic Research Papers*, *56*(5), 727–740. doi:10.1016/j.dsr.2008.12.007.

Baldwin, C. J. (2015). *The 10 principles of food industry sustainability.* Chichester, West Sussex, UK: John Wiley & Sons.

Barbosa, N., Christo, E., & Costa, K. A. (2015). Demand forecasting for production planning in a food company. *ARPN Journal of Engineering and Applied Sciences*, *10*(16), 7137–7141.

Bender, P. (2006). THE PRECAUTIONARY APPROACH AND MANAGEMENT OF THE ANTARCTIC KRILL. *Journal of Environmental Law*, *18*(2), 229–244. http://www.jstor.org/stable/44248546.

Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (1994). *Time Series Analysis; Forecasting and Control* (3 ed.). Englewood Cliffs, New Jersey: Prentice Hall.

Burri, L., Hoem, N., Banni, S., & Berge, K. (2012). Marine omega-3 phospholipids: metabolism and biological activities. *International journal of molecular sciences*, *13*(11), 15401–15419.

CCAMLR (2010). Conservation Measure 51-01 (2010) Precautionary catch limitations on Euphausia superba in Statistical Subareas 48.1, 48.2, 48.3 and 48.4. https://www.ccamlr.org/en/measure-51-01-2010.

CCAMLR (2018). Krill Fishery Report 2018. Last accessed October 25, 2019: https://www.ccamlr.org/en/document/publications/krill-fishery-report-2018.

CCAMLR (2018). REPORT OF THE THIRTY-SEVENTH MEETING OF THE SCIENTIFIC COMMITTEE. https://www.ccamlr.org/en/system/files/e-cc-xxxvii.pdf.

CCAMLR (2019). Preliminary Report of the Thirty-eighth meeting of the Scientific Committee.

Chan, N. H. (2002). *Time series: Applications to Finance*. New York: John Wiley & Sons.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, *74*(368), 829–836. http://www.jstor.org/stable/2286407.

Coolidge, F. L. (2006). *STATISTICS: A Gentle Introduction* (2 ed.). Sage Publications, Inc.

Cowpertwait, P. S. & Metcalfe, A. W. (2009). *Introductory Time Series with R*. New York: Springer.

Croston, J. D. (1972). Forecasting and Stock Control for Intermittent Demands. *Journal of the Operational Research Society*, *23*(3), 289–303. https://doi.org/10.1057/jors.1972.50.

Czerwinski, I. A., Gutiérrez-Estrada, J. C., & Hernando-Casal, J. A. (2007). Short-term forecasting of halibut CPUE: Linear and non-linear univariate approaches. *Fisheries Research*, *86*(2-3), 120–128. https://doi.org/10.1016/j.fishres.2007.05.006.

Da Veiga, C. P., Da Veiga, C. R. P., Catapan, A., Tortato, U., & Da Silva, W. V. (2014). Demand forecasting in food retail: A comparison between the Holt-Winters and ARIMA models. *WSEAS Transactions on Business and Economics*, *11*, 608–614.

Dalsegg, M. (2018). New Study Shows Potential Future of Aquaculture Feed. AkerBioMarine.com. Last accessed October 31, 2019: https://www.akerbiomarine.com/news/new-study-shows-potential-future-of-aquaculture-feed.

Diebold, F. X. (2004). *Elements of Forecasting* (3 ed.). Ohio, US: Thomson South-Western.

Filliben, J. J. & Heckert, A. (2003). Exploratory data analysis. In J. J. Filliben (Ed.), *NIST/SEMATECH e-Handbook of Statistical Methods* chapter 1. NIST/SEMATECH. https://www.itl.nist.gov/div898/handbook/eda/section3/eda352.htm.

Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems* (2 ed.). Sebastopol, California: O'Reilly Media.

Gilliland, M., Sglavo, U., & Tashman, L. (2015). *Business Forecasting: Practical Problems and Solutions*. Hoboken, New Jersey: John Wiley & Sons.

Guerrero, V. M. & Perera, R. (2004). Variance Stabilizing Power Transformation for Time Series. *Journal of Mordern Applied Statistical Methods*, *3*(2), 357–369. https://doi.org/10.22237/jmasm/1099267740.

Guthrie, W., Filliben, J. J., & Heckert, A. (2003). Process modeling. In W. Guthrie (Ed.), *NIST/SEMATECH e-Handbook of Statistical Methods* chapter 4. NIST/SEMATECH. https://www.itl.nist.gov/div898/handbook/pmd/section1/pmd144.htm.

Harvey, A. C. (1993). *Time Series Models* (2 ed.). Hemel Hempstead, United Kingdom: Harvester Wheatsheaf.

Hobday, A. J., Spillman, C. M., Paige Eveson, J., & Hartog, J. R. (2016). Seasonal forecasting for decision support in marine fisheries and aquaculture. *Fisheries Oceanography*, *25*(S1), 45–56. https://doi.org/10.1111/fog.12083.

Holguín-Veras, J. & Jaller, M. (2012). Immediate Resource Requirements after Hurricane Katrina. *Natural Hazards Review*, *13*(2), 117–131. https://doi.org/10.1061/(ASCE)NH.1527-6996.0000068.

Hyndman, R., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., Petropoulos, F., Razbash, S., Wang, E., Yasmeen, F., R Core Team, Ihaka, R., Reid, D., Shaub, D., Tang, Y., & Zhou, Z. (2019). *Package "forecast"*. CRAN. Last accessed December 14, 2019: https://cran.r-project.org/web/packages/forecast/forecast.pdf.

Hyndman, R. J. (2006). Another Look at Forecast Accuracy Metrics for Intermittent Demand. *Foresight: The International Journal of Applied Forecasting*, *4*(4), 43–46.

Hyndman, R. J. & Athanasopoulos, G. (2018). *Forecasting: Principles and Practice* (2 ed.). Melbourne, Australia: OTexts. OTexts.com/fpp2.

Hyndman, R. J. & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*(4), 679–688. https://doi.org/10.1016/j.ijforecast.2006.03.001.

Hyndman, R. J. & Kostenko, A. V. (2007). Minimum sample size requirements for seasonal forecasting models. *Foresight*, (6), 12–15.

Maddala, G. S. & Kim, I.-M. (1998). *Unit Roots, Cointegration, and Structural Change*. Cambridge, United Kingdom: The Press Syndicate of the University of Cambridge.

Msangi, S., Kobayashi, M., Batka, M., Dey, M., Vannuccini, S., & Anderson, J. (2013). FISH TO 2030: Prospects for Fisheries and Aquaculture. *Agriculture and environmental services discussion paper*, *83177-GLB*(03). http://www.fao.org/3/i3640e//i3640e.pdf.

Nielsen, P. H. & Olesen, E. (2003). Fishmeal and oil production. Last accessed October 31, 2019: http://www.lcafood.dk/processes/industry/fishmealproduction.htm.

Pankratz, A. (1983). *Forecasting with univariate Box-Jenkins models: Concepts and cases*. John Wiley & Sons.

Prista, N., Diawara, N., Costa, M. J., & Jones, C. M. (2011). Use of SARIMA Models to Assess Data-Poor Fisheries: A Case Study With A Sciaenid Fishery Off Portugal. *Fishery Bulletin*, *109*(2), 170–185. https://digitalcommons.odu.edu/cgi/viewcontent.cgi?article=1039&context=oeas_fac_pubs.

Quinn, J., McEachen, J., Fullan, M., Gardner, M., & Drummy, M. (2019). *Dive Into Deep Learning: Tools for Engagement*. Thousand Oaks, California: Corwin Press.

Sagaert, Y., Aghezzaf, E.-H., Kourentzes, N., & Desmet, B. (2017). Temporal Big Data for Tire Industry Tactical Sales Forecasting. *Interfaces*. https://doi.org/10.1287/inte.2017.0901.

Saila, S. B., Wigbout, M., & Lermit, R. J. (1980). Comparison of some time series models for the analysis of fisheries data. *ICES Journal of Marine Science*, *39*(1), 44–52. https://doi.org/10.1093/icesjms/39.1.44.

Stergiou, K. I. (1989). Modelling and forecasting the fishery for pilchard (Sardina pilchardus) in greek waters using arima time-series models. *ICES Journal of Marine Science*, *46*(1), 16–23. https://doi.org/10.1093/icesjms/46.1.16.

Stergiou, K. I. (1991). Short-term fisheries forecasting: comparison of smoothing, ARIMA and regression techniques. *Journal of Applied Ichthyology*, *7*(4), 193–204. https://doi.org/10.1111/j.1439-0426.1991.tb00597.x.

Støp−Bowitz, C. & Sømme, L. S. (2017). krill. *Store Norske Leksikon*. Last accessed October 23, 2019: https://snl.no/krill.

Suganthi, L. & Samuel, A. A. (2012). Energy models for demand forecasting—A review. *Renewable and Sustainable Energy Reviews*, *16*(2), 1223–1240. https://doi.org/10.1016/j.rser.2011.08.014.

Tirkes, G., Güray, C., & Celebi, N. (2017). DEMAND FORECASTING: A COMPARISON BETWEEN THE HOLT-WINTERS, TREND ANALYSIS AND DECOMPOSITION MODELS. *Tehnicki Vjesnik-Technical Gazette*, *24*(S2), 503–510.

Tsai, C.-F. & Chai, A.-L. (1992). Short-term forecasting of the striped bass (Morone saxatilis) commercial harvest in the maryland portion of chesapeake bay. *Fisheries research*, *15*(1-2), 67–82. https://doi.org/10.1016/0165-7836(92)90005-E.

United Nations (2019). World Population Prospects 2019: Highlights. https://population.un.org/wpp/Publications/Files/WPP2019_10KeyFindings.pdf.

Vrieze, S. I. (2012). Model selection and psychological theory: A discussion of the differences between the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). *Psychological Methods*, *17*(2), 228–243. https://doi.org/10.1037/a0027127.

Zivot, E. & Want, J. (2006). *Modeling Financial Time Series with S-PLUS* (2 ed.). New York: Springer.

Appendix

A1 Illustration of the Automatic Model

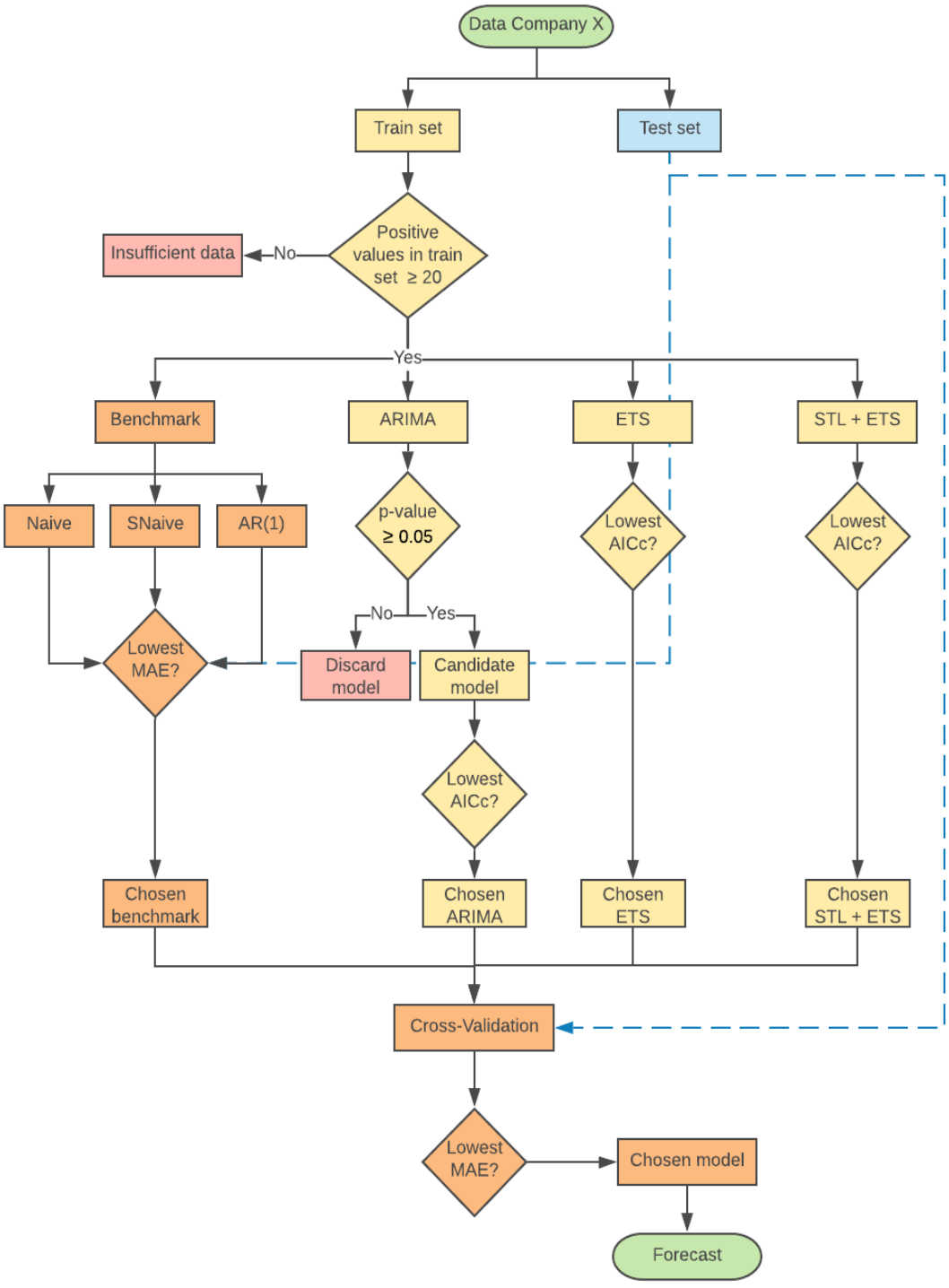

Figure A1.1: Illustration of the automatic model

A2 Modeling Results

Company 9 and 10 have insufficient training data to estimate any models and is therefore not displayed.

	ARIMA	**ETS**	**STL + ETS**
Company 1	(1,0,1)(0,1,1)	ANA	STL + ANN
Company 2	(1,0,2)(0,1,1)	AAN	STL + ANN
Company 3	(2,0,2)(1,1,2)	ANN	STL + ANN
Company 4	(0,0,1)(2,1,0)	ANN	STL + ANN
Company 5	(2,0,1)(0,1,1)	ANN	STL + ANN
Company 6	(1,0,1)(0,1,1)	ANN	STL + ANN
Company 7	(2,0,2)(2,1,1)	ANN	STL + ANN
Company 8	(0,0,0)(1,1,0)	ANN	STL + ANN
Company 11	(2,0,1)(1,1,0)	ANN	STL + AA_dN
Company 12	(2,0,0)(1,1,2)	ANN	STL + ANN
Company 13	(0,0,0)(2,1,0)	ANN	STL + AA_dN
Company 14	(2,0,2)(2,1,2)	ANN	STL + ANN
Company 15	(0,0,0)(2,1,0)	ANN	STL + ANN
Company 16	(2,0,2)(2,1,0)	ANN	STL + ANN
Company 17	(1,0,2)(2,1,0)	AA_dN	STL + AAN
Company 18	(1,0,0)(0,1,1)	ANN	STL + ANN
Company 19	(2,0,2)(1,1,0)	ANN	STL + ANN
Company 20	(2,0,2)(1,1,0)	ANN	STL + ANN

Table A2.1: Chosen models for all methods (Company 1-20)