Erasmus University Rotterdam

Erasmus School of Economics

Master Thesis Econometrics and Management Science:
Business Analytics and Quantitative Marketing

Forecasting Airline Revenue using Historic Revenue Curves

Author:
Mats Bierhuizen
ID: 547030

Supervisor:
Prof. Dr. D. J. C. van Dijk

Second Assessor:
Dr. A. Pick

November 11, 2021

Abstract

The goal of this research was to construct a model that forecasts monthly airline revenue for the upcoming 12 months. The airline's current forecast is not yet able to do so, and it is based on a naive model that depends on a lot of subjectivity. An automated forecast that makes use of statistical forecasting models should take away the subjectivity. It will also allow the airline to control monthly revenue in a better way, as marketing campaigns can be adjusted to the forecasted progress of monthly revenue.

The construction of a new forecast is done by using both a short- and long-term model. The short-term model will predict the first 4 months, and the long-term model will extend this model to forecast up to 12 months. These models make use of a combination of time series models such as Exponential Smoothing and SARIMA(X) models, as well as a Dynamic Factor model. These models can also be combined with the aim of increasing forecast accuracy. The data of the airline can be disaggregated by applying the attributes Point of Sale, Booking Channel, Cabin Class, Carrier, and Line Group. By doing so, the airline is able to trace back more precisely how revenue is earned. The most important data that is used in this thesis are revenue curves, which are curves that show the progress of the earned revenue for a specific flight month, and monthly revenues.

From the results it can be concluded that the combination of a short-term model for the first 4 months and a long-term model for months 5-12 had a positive outcome. The forecast accuracy of the short-term model for the first 4 months is higher than when the long-term model is applied to the first 4 months. Hence, the two models complement each other. Moreover, it is found that the attributes that are used to disaggregate the data are Point of Sale dependent; the so-called large Point of Sales do not require the same attributes as the small Point of Sales. For the short-term model, a combination of the SARIMA and SARIMAX models shows good results for large Point of Sales, whereas for small ones the best results are obtained when Exponential Smoothing, SARIMA, and SARIMAX models are combined. For the long-term model, two out of four tested Point of Sales showed the best results when a combination of the Exponential Smoothing method and the Dynamic Factor model was used, whereas for the other two the naive model performed best. A combination of the Exponential Smoothing method and the Dynamic Factor model performed second-best for these two Points of Sales.

The airline is advised to replace the current forecast that is based on a naive model by the model combinations found in this research, since the forecast accuracy increases and the influence of subjectivity decreases. However, the long-term model needs some improvement as it does not always outperform the naive model.

Contents

Chapter 1

Introduction

Forecasting techniques have proved to be valuable for revenue management (RM) over the years. For the airline industry, a well-performing revenue forecast can lead to an increase in revenue as the forecast could positively influence tactics and decision-making. Lee (1990) has shown that if the forecast accuracy increases by 10 percent, the revenue on high demand flights can increase by 0.5-3.0 percent. The increase in revenue is, amongst others, an important factor that could lead to the investigation of such forecasting models.

Currently, the airline for which this research is being done is in need of a new advanced method to correctly predict the revenue per line group, booking channel, cabin, carrier, and flight month, based on historic data. Each month, the airline wants to know what their PaxKM, Revenue and Yield will be, while costs do not have to be taken into consideration. The PaxKM is defined as the number of kilometers all passengers together will fly. The Revenue is equal to the total earnings and the Yield is determined to be the number of earnings per flown kilometer, which is the Revenue divided by the PaxKM. By forecasting the PaxKM and the Yield, the Revenue can be predicted by multiplying these two values. Currently, a factor model is used for the prediction of the PaxKM. However, the airline has not been able to find a working model for the Yield or the Revenue yet.

Since the airline does not have an official automated and advanced forecasting model, it is possible for the different establishments of the airline all over the world to apply their own techniques and ideas. As a result, the current forecast is handmade every month by each establishment, requiring a lot of man-hours and being subjected to subjectivity. Thus, there is a great need for an automated model that provides a well-performing forecast, which can be implemented by all establishments.

In this research it will be investigated whether a model can be created that can forecast the monthly revenue for an airline. A forecast is needed that predicts the upcoming 12 months, as the current one only predicts up to December of the current year, no matter if the current month is January or November. The new forecast will be called a rolling forecast, and it will be a combination of a short-term and a long-term model. The short-term model will predict the first 4 months, and the long-term model extends this forecast to 12 months.

During this research, the data will be disaggregated using the attributes: point of sale, line group, booking channel, cabin class and carrier. The purpose of this disaggregation is that forecasts will be made for specific sets of attributes instead of for the airline as a whole, in order to find more precisely how revenue is obtained. These individual forecasts can then be aggregated

to form a forecast for the entire airline. By splitting the data into different attributes, the models will provide a clear picture of where the revenue is coming from. During this research, it will become clear which variables are of the most importance for the revenue forecast. The data that will be used is real airline data, consisting of e.g., revenue curves, monthly revenues and PaxKM data. Revenue curves show the progress of how much revenue has been obtained by selling tickets for a specific month. Tickets for the flights in this month can be bought approximately a year before the flight, which allows the airline to create a curve of the progress throughout the year. Given the data, models that are able to deal with time series, trends and seasonality will be used.

The most important reason the airline is in need of the outcome of this research is that a well-performing revenue forecast can inform the management regarding the position in the market, and it can keep the management up to date regarding developments in the monthly revenue. Also, it is necessary to decrease subjectivity and to see whether targets that are set are still achievable. Thus, when the forecast shows results that might or might not be satisfactory, fares and marketing tactics can be optimized to the situation. The shortcomings that come with not having a high-performing forecast are the motivation to investigate new models. For researchers outside of this airline, the research of this thesis can still be of importance, as it investigates the use of known models with new data from a different point of view. This research could function as a new inspiration to improve the future research of forecasting in revenue management.

L. Weatherford (2016) provides an overview of forecasting methods that have been used in the airline industry for RM. By using forecasting methods, issues as overbooking, seat availability and pricing can be solved and optimized. Some of the techniques that have been used are Exponential Smoothing (ES) models, linear regressions and Moving Average (MA) models. Also, the research of L. R. Weatherford et al. (2003) is mentioned, in which neural networks slightly outperform the aforementioned methods.

However, in collaboration with Lufthansa Airlines, Lemke et al. (2013) looked into forecasting the revenue with real airline data. It became clear that models that are simple and based on robust time series show a significant increase in performance compared to sophisticated methods. Such simple methods are regression, ES, or simple average models. The reason why these simple models outperform the sophisticated ones lies in the adaptability of the simple models. This is a different result than the one that was given in L. Weatherford (2016), where a neural network model was preferred. As the current research will also be done in collaboration with an airline, which allows the use of real airline data, the conclusion of Lemke et al. (2013) is of importance for this thesis.

Subsequently, Lemke et al. (2013) mention diversification procedures in which different models or models with diverse data are combined. By using different models, a diverse method pool can be used that considers a trade-off between individual accuracy and diversity in the pool. It is encouraged to combine complex (non-linear) and less complex (linear) models. Within these models, a diversification in data is also possible. For example, by using data with different levels of aggregation. By using a diverse method pool and combining these different methods, the combinations could lead to better forecasts. The use of such forecast combinations will be

applied to the current research as well.

The literature on revenue management leads to the investigation of time series models. Models used in RM research show that Structural Time Series Models (STSMs) are interesting models to investigate. These models, also known as Unobserved Component Models, have already successfully been used in time series forecasts. The elements of the STSMs can be separated into different components, describing for example trends, seasonal patterns, and cycles. Proietti (2002) describes the STSMs and their application for the aforementioned components, which will be present in the data that is used for the current research. Models that are fundamental for the STSMs are the State Space Models (SSMs).

With enough historical data, important patterns and components can be captured to create accurate forecasts. Examples of SSMs are Autoregressive Integrated Moving Average (ARIMA) models and ES methods, which were also mentioned in the literature on RM. These models are well known and often used for time series research. When using linear models that have normally distributed errors, these models can be called Linear Gaussian State Space Models. A. C. Harvey & Shephard (1993) already called these methods exceedingly useful time series methods.

Among others, the research of Jackman & Greenidge (2010), Song et al. (2011) and Athanasopoulos & Hyndman (2008) has shown that STSMs and SSMs can be successfully applied to forecasting economic time series. Therefore some representations of the SSMs, the ES method and the Seasonal ARIMA (SARIMA) method, will be described in more detail. Moreover, some literature of the Dynamic Factor Model (DFM) will be given, as the airline has previously used a factor model for the forecast of the PaxKM.

The first application of the SSMs that will be discussed is the ES method, given in Hyndman et al. (2002). Here the authors go into detail about the SSMs using multiple different ES methods. For the current research, the ES method that uses an additive trend and seasonal component will be of importance. The method that uses these additive components is called the Holt-Winter's Exponential Smoothing Method (HWESM).

Gardner Jr (2006) provides an overview of the research that has been done using ES methods up to the year 2005. A table with 65 different papers is presented, and from these 65 papers, 58 obtained a successful result in the form of forecast accuracy. Hence, since this is a large portion of the 65 papers, the ES methods have often proved to be valuable. Using data from Makridakis et al. (1982), forecasts were performed in Hyndman et al. (2002) that indicate that the use of ES models contributes to good results, especially on the short-term where a forecast is made for up to 6 periods. Both Gardner Jr (2006) and Hyndman et al. (2002) mention that the ES methods can no longer be seen as ad hoc forecasting approaches, and that these models are on the same level as ARIMA models.

Other representations of the SSMs that are often used, are the well-known ARIMA models. These models allow one to forecast time series by looking at autoregressive and moving average factors. In the research of Goh & Law (2002), a variant of the ARIMA model is tested to forecast

the demand of tourism where the data contained seasonality. The model that was used was the SARIMA model. The SARIMA models were used to forecast tourism demand and consistently outperformed naive, ARIMA, MA, and ES models.

A specific framework of ARIMA that has worked well for the airline industry is the so-called airline model. This model is described in Box et al. (2015), which is a fifth version from the original book of 1970. The optimal framework of the airline model is already determined, and therefore optimization of the number of parameters is no longer necessary. The framework was later modified for SARIMA as well. Since this study researches the revenue of an airline, this model will be interesting to investigate.

There are not many papers available in which SARIMA or SARIMAX (SARIMA with exogenous variables) models are used for (airline) revenue management. This is where the current research will contribute to the literature. Nonetheless, applying SARIMA(X) models to time series data with seasonal patterns has proved to lead to good results. This is for example the case in Goh & Law (2002), but Faraway & Chatfield (1998) also showed promising results. Here the SARIMA airline model outperforms neural networks when forecasting the number of monthly passengers for an airline.

Since the airline already uses a factor model for the PaxKM forecast, a DFM will also be considered in order to see what the impact of another type of model could be. The DFM was first discussed in Geweke (1977), where the DFMs were used as a time-series variant of the factor models. These models are not necessarily used in airline revenue management, but Breitung & Eickmeier (2006) and Stock & Watson (2011) have reviewed the use of DFMs in other papers. In the first paper, existing applications for macroeconomic problems are mentioned. The use of DFMs led to encouraging results and good performances. Stock & Watson (2011) describe applications of the DFM and review empirical findings, and thereby show the versatility of these models.

Combining models, as earlier mentioned by Lemke et al. (2013), has proved to lead to more accurate results. Among others, in Makridakis et al. (2018), Timmermann (2006) and Bates & Granger (1969) it is shown that combining models can increase forecast accuracy. Hence, during this research, the models will be combined using a simple average method with the aim of increasing the performance of the models. The use of a simple average method has been proved to be hard to beat (Timmermann, 2006). The effectiveness of combining models is shown in the M4 research of Makridakis et al. (2018), which continues on Makridakis et al. (1982). The authors asked forecasting experts to improve the forecast accuracy of time series methods. The results show that out of the 17 methods that had the highest accuracy, 12 consisted of a combination of methods. These combinations could consist of only statistical approaches such as the aforementioned ES or SARIMA models, but also machine learning methods. The best results were obtained when both machine learning and statistical approaches were combined. These models also resulted in the most accurate 95% prediction intervals (PIs).

As mentioned earlier, little research has been done for revenue management in the airline

industry using the models specified in the previous paragraphs. This makes this research more relevant and interesting, not only for this particular airline but also for other researchers and corporations who deal with the issues addressed in this thesis. This thesis will therefore contribute to the current research on revenue management, which makes it scientifically relevant.

Next to the fact that there are not many studies regarding this topic, this thesis has access to a large amount of real airline data. In competitive businesses such as the airline industry, companies do not necessarily share their data and findings. This research will therefore not only look at the right models to use, but also the data that is used is revealed. Since the accessibility to the results of such research is not common, new insights could be provided that can be helpful in future research.

The airline earns its revenue by selling tickets in countries all over the world. Each of these countries is called a Point of Sale (PoS). Based on the earned revenue per country, the airline divides these PoS's into two groups, small and large PoS's. It is believed that these two groups might need different models. By combining forecasting methods for both the short- and the long-term model and the two types of PoS's, the model combinations with the best performance must be found.

Besides the forecasting methods, it will also be important to choose the right attributes to disaggregate the data with. Because the airline wants to pinpoint the origin of the earned revenue, besides the Point of Sales, attributes such as the line group, cabin class, booking channel, and carrier can be used. The airline prefers the use of the line group, booking channel, and carrier attribute for large PoS's, and cabin class and carrier for small PoS's. These preferences will be tested.

Since booking curves and their corresponding revenue curves can hold a lot of information on what the revenue in a certain month will turn out to be, this data will be used in the short-term model. The curves of multiple years will form a time series with a sawtooth pattern, and modeling techniques will be applied that predict future values based on these curves. For the short-term forecast, the ES models, a SARIMA model with and without exogenous variables and a DFM will be used to forecast the first 4 months. This boundary of 4 months is chosen by the company, as this is the forecast horizon for which they actively react on the current state of bookings. The models will also be combined to increase forecast accuracy.

For the long-term forecast, the same ES model, SARIMA without exogenous variables, and the DFM will be used and combined. However, there is a difference in the data. The revenue curves 12 months before departure will contain too little information as too few tickets have been sold yet. Therefore, for this model, the time series of the monthly revenues will be used.

This research has shown that there is indeed a difference between large and small PoS's. Firstly, they both require different attributes to optimize forecast accuracy. For the large as well as the small PoS's, the attributes that result in the best performance are the ones that were preferred by the airline. Besides that, the results of the short- and long-term models show that the combination of the two models and their different forecast horizons outperforms the single use of the long-term model for the entire 12-month forecast.

For the short-term model, a combination of HWESM, SARIMA, and SARIMAX shows the best results for small PoS's, whereas a combination of SARIMA and SARIMAX has the best forecast accuracy for large PoS's. For both types of PoS's, a naive model is outperformed. Out of the tested models for the long-term model, the combination of HWESM and DFM shows the best results for one small and one large PoS, and for the other two PoS's the naive model outperforms all other models.

In the next chapter, the data of this research will be presented with the help of figures. In Chapter 3, the methods used in this research are described. Chapter 4 will consist of the results and a discussion of these results, followed by a conclusion in Chapter 5.

Chapter 2

Data

In this chapter the different types of data are discussed. As can be expected, airlines have a lot of data which can be disaggregated into various levels of detail. For example, data can be given for an entire country or for a specific flight. In order to obtain this detailed data, attributes must be selected. Depending on the level of detail of the data, the number of observations can vary. In the upcoming sections there will be focused on the visualization of the data using attributes that are of importance for the data disaggregation, as well as variables that can be used to predict the revenue. The data that will be used is monthly data, and depending on the variable this is available since January 2015 or January 2016. Besides that, the last data is from December 2019 and is not effected by the Covid-19 pandemic. Since the data is real airline data, for confidentiality reasons some attributes are anonymized and indicated by capital letters.

2.1 Attributes

All the data that the airline possesses can be categorized by using multiple attributes. This collection of data could consist of e.g., the number of bookings, yield, revenue, ticket prices, and any other (flight specific) information. By using these attributes, the data can be disaggregated, which could increase the interpretability of the data as it is divided into subgroups. Due to this increased interpretability, the airline is able to see more clearly where, when and how revenue is obtained. The attributes that are of most importance are:

- Point of Sale, countries denoted by W, X, Y and Z;
- Line group, destination areas;
- Cabin classes: Business, Economy, First and Premium Class;
- Booking channels: Direct Online, Direct Offline, Indirect Online, Indirect Offline and Unknown;
- Carrier, airline company;
- Ticketing date, date at which a ticket was bought.

Almost all data of the airline can be split up in different PoS's. The PoS filters the data on a specific country or subareas where tickets are sold, so the PoS is not necessarily the same as the location a plane departures from. Hence if you are in Berlin and you buy a ticket from London to New York City, the PoS will be Germany (Berlin) even though the flight departures from London. As airlines have establishments all over the world, PoS specific data can show how each establishment performs and if sales strategies should be altered.

Each PoS sells tickets to different parts in the world. The line group attribute divides the world in a number of groups that indicate what the destination will be. For the most part these line groups are large groups of countries, for example a whole continent. In the example given above the line group would be the United States of America.

When you buy a ticket from an airline, there is some variety in tickets you can choose from. The most well-known tickets are Economy and Business class tickets. Depending on the carrier, the cabin attribute divides the plane into two (Economy & Business) or all four different classes, each with their own fares and customer characteristics. The Economy and Premium class are also called the *low revenue* classes, whereas the Business and First class are known as the *high revenue* classes due to the average ticket prices in each cabin class. Within these low and high revenue classes, the tickets show some similar characteristics.

The booking channels show how tickets were purchased. These purchases can be done on the internet (online) or in person (offline), but also at the airline itself (direct) or through a travel agent (indirect). All channels show different behaviour when it comes to the prices and the moment when tickets are purchased.

The carrier attribute disaggregates the data into different airline carriers that are part of the airline. Lastly, in the airline industry people can buy their tickets months in advance. This means that an airline can already make an estimation on how many revenue they will make or how many passengers will travel in a certain period of time. Since people can still cancel their flights, revenue is only booked after a flight has departed. By using the ticketing date attribute, data such as the revenue earned in a certain month can be traced back to the moment a ticket was bought. This allows one to create a graph that shows how the revenue is distributed over the periods of time prior to the departure date. The ticketing date can be daily, weekly or monthly.

All these attributes can contribute to specifying the data. Note that the data is not on an individual flight level, as the aim of this research was to look at a more aggregate level of the revenue. This has consequences for the data that is available and methods that can be used. The figures that will be presented in this chapter will help in showing the importance of the data disaggregation.

2.2 Dependent variable

In this section the dependent variable, the monthly revenue, of the past years will be displayed in various ways in order to show the necessity of data disaggregation. This monthly revenue itself is available as of January 2005, however some important attributes were introduced later which causes the data to be sufficient since 2015. The sample frequency can be chosen to be daily, weekly, monthly, seasonally or yearly and the observations can be broken down into several attributes. Nonetheless, in this section the data is presented as the monthly data of all flights combined. In order to limit the amount of data that is presented here, only the data of PoS W is used in this section. This is the main PoS for this research and this data will provide enough information to understand the data of other PoS's as well. In case data of other PoS's needs to be visualized, this will be provided in Appendix A.

The total monthly revenue of PoS W from January 2015 - December 2019 is given in Figure 2.1. There can be seen a clear seasonal pattern as the revenue in the summer (July and August) is evidently higher than during the winter months (November-February). Also, there is an unmistakable trend over the years.

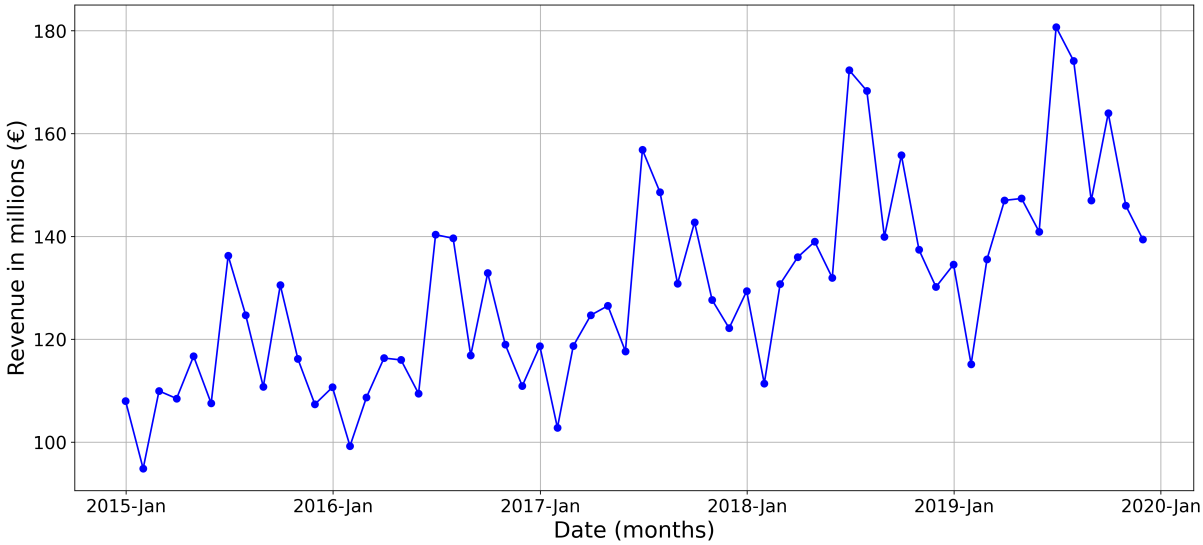


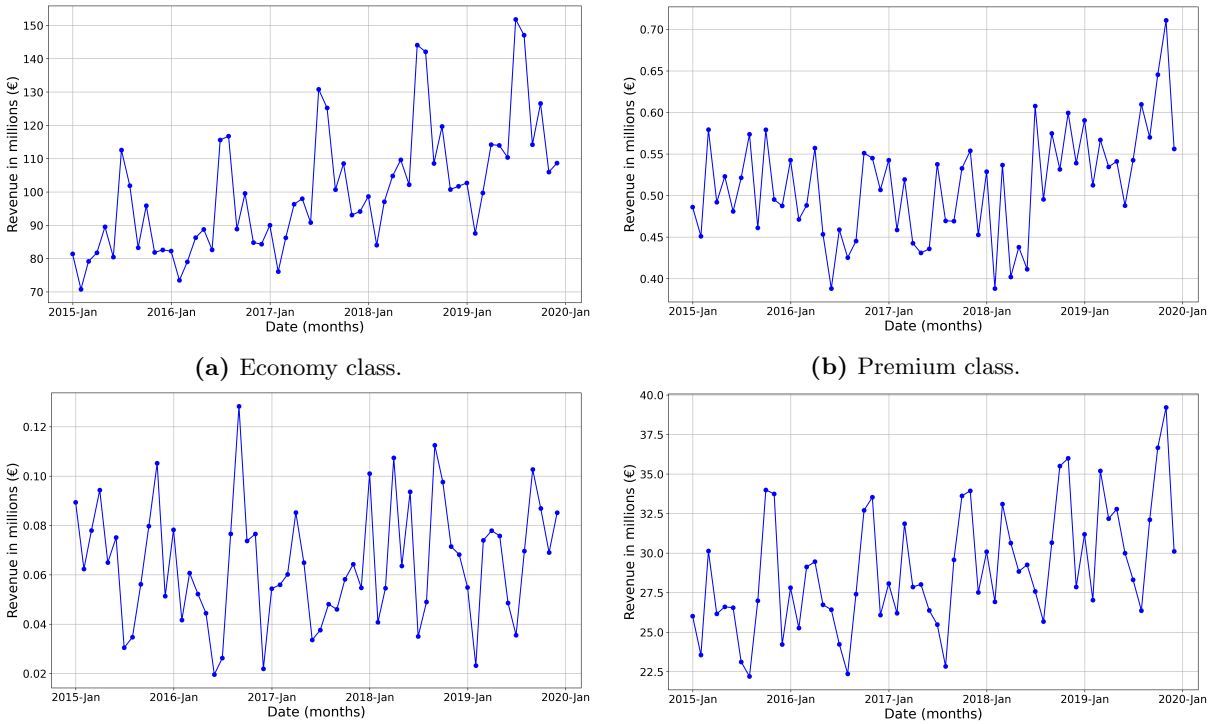
Figure 2.1: Total monthly revenue in millions of euros of PoS W from January 2015 - December 2019.

In Figure 2.2 the monthly revenue of PoS W in the period January 2015 - December 2019 is disaggregated using the cabin attribute, thereby showing how the revenue differs in both the amplitude and the seasonal pattern per cabin class.

It can be seen that not all classes show the same type of seasonality. For example, the revenue of the Economy class in Figure 2.2a shows obvious peaks during the summer months, as this is the moment where many people go on holiday, whereas the revenue drops during the winter. The seasonal pattern of the Premium class is very inconsistent, as displayed in 2.2b. When looking at the summer months, it can be seen that in 2016 the revenue is at a low point, whereas the peaks at July 2018 and August 2019 are some of the highest points. Also, the trend that can be seen in Figure 2.2a is not present for the Premium class. First class (Figure 2.2c) has a seasonal pattern that is somewhat inconsistent as well, even though overall the summer months have low revenues. The monthly revenue of Business class (Figure 2.2d) shows a clear seasonal pattern, with peaks at the March and at the end of the year, in October and November.

Moreover, for this PoS, as for the entire airline, it can be seen that the Economy and the Business classes have monthly revenues that are significantly higher than those of the Premium and First classes. This is because not all carriers sell First and Premium class tickets. As this PoS is largely represented by a single carrier that does not sell First and Premium tickets, Economy and Business are the most common tickets here.

In Figure A.1 in Appendix A the cabin classes of PoS Y are displayed. Even though this PoS is also largely represented by a carrier that does not sell Premium and First class tickets, the seasonal patterns of the Premium and First class are not as inconsistent as for PoS W. There



(a) Economy class.

(b) Premium class.

(c) First class.

(d) Business class.

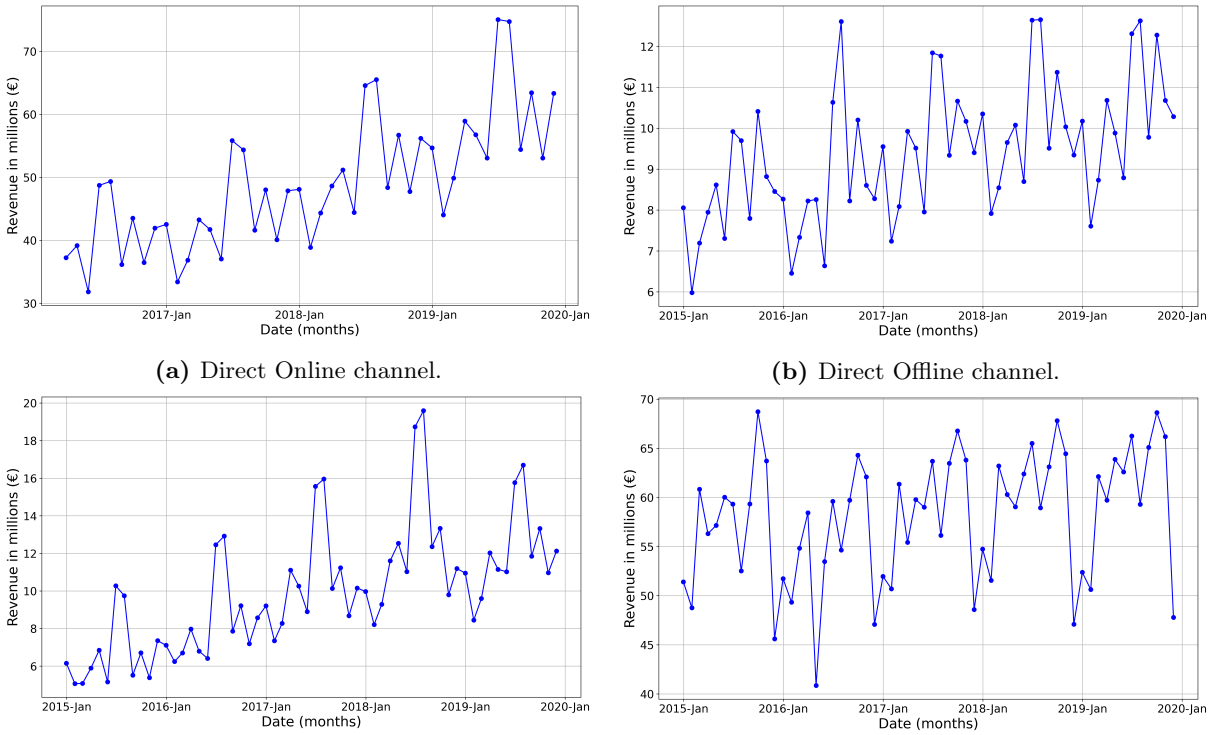
Figure 2.2: Monthly revenues in millions of euros of the cabin classes of PoS W from January 2015 - December 2019.

can even be seen similarities between the seasonality of the Economy and Premium class, and Business and First class. For the latter there is however a large difference in trend over time.

For the rest of this research the Economy and Premium classes will be combined, as will the Business and First class. From now on these two groups will be called *Economy* and *Business* respectively. This is done because there will not always be enough data of the Premium and First class tickets to forecast accurately. Also, the airline already divided the cabin classes in low and high revenue classes, and within these two groups the cabins show some similar characteristics. A characteristic of the Business class that might be of importance is the fact that most people who buy these tickets, buy them close to the flight date. If the purchase behaviour of Business class is compared to Economy class tickets, on average Business class tickets are purchased at a later stage.

In Figure 2.3 the total monthly revenue of PoS W of the period January 2015 - December 2019 is disaggregated into four booking channels.

There can be seen that the Direct Online and the Indirect Offline channels are the largest channels, whereas the Direct Offline and Indirect Online are the smallest. Overall there can be seen positive trends over the years 2015-2019 as well as clear seasonal patterns, which are different for each channel. An important observation is that in Figure 2.3a the monthly revenues of January 2015 until March 2016 are left out of the figure, because the values of these months were close to zero. This is due to the fact that for the Direct Online channel the monthly revenues were only successfully obtained as of April 2016 for this particular PoS. Before April, these rev-



(a) Direct Online channel.

(b) Direct Offline channel.

(c) Indirect Online channel.

(d) Indirect Offline channel.

Figure 2.3: Monthly revenue in millions of euros of the booking channels of PoS W from January 2015 - December 2019.

enues could have been given the label *Unknown*. Over the years, the amount of monthly revenue labeled Unknown has dropped because data is being processed better, and also the amount of monthly revenue that got this label is very inconsistent. Due to these reasons, the Unknown booking channel is not taken into account here. Because of the amount of revenue that is booked on this Unknown channel, not every set of data (PoS, channel, etc.) is able to use the same size of historic data.

Lastly, in Appendix A, a figure can be found that shows the monthly revenues when both the booking channel and cabin class attributes are used. Figure A.2 shows eight figures of the four channels in combination with the Economy and Business classes. Here it can be seen that the combinations of cabin classes and booking channels do not necessarily have the same trend or seasonal pattern over the years. For the two figures of the Direct Online channel, the first few observations were left out. This is due to the same reasons as for Figure 2.3a.

2.3 Independent variables

The independent variables can be used to predict the revenue. These variables can consists of lagged values of the dependent variable, but also exogenous variables that are different than the revenue. Just as for the dependent variable, the sample frequency and the number of observations depend on the level of detail of the data. The data that will be used will be monthly data. Furthermore, individual flight data is grouped together in the attributes that are

mentioned in Section 2.1 in order to prevent the loss of interpretability. Due to the necessary application of these attributes, not all the available airline data can be used as it is not specified on the level of the attributes. Also, since some data will be used to predict 12 months ahead, the data must be useful enough on the long term to be implemented in a model. Due to this, the selection of variables for a long-term forecast becomes limited.

2.3.1 Revenue curves

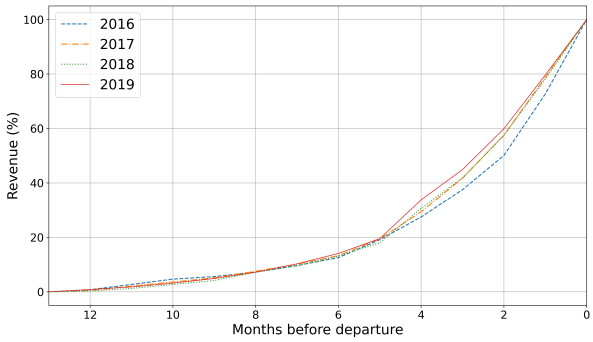
Per individual flight the booking curves are available. These curves show the progress of the number of tickets sold from the moment the tickets are up for sale until the flight date. This data can also be looked at on a more aggregate level in order to see the bigger picture. By using the ticketing date attribute and looking at the revenue per ticketing month, the booking curves are in a way transformed to revenue earned per month. Instead of knowing that for example 6 months before departure 20% of the tickets have been sold, the revenue per ticketing month displays the amount this 20% of tickets is worth. This way one can keep track of the progress of the revenue. Both the shape of the curves as the lagged values of the observations can be useful for predicting the final monthly revenue. As mentioned in Section 2.2, this data is also available as daily or weekly data since 2015, but considering that the dependent variable will be the monthly revenue the data will be presented as monthly data.

For PoS W and the months January and July of the years 2016-2019, the revenue per ticketing month of the booking channels is displayed in Figure 2.4. These figures show how the monthly revenue evolves over the months prior to the departure, but also how the shapes can differ between months and booking channels. Cancelled tickets are not taken into account in these figures, as in the end these have not brought in any revenue. As the revenue is not the same in each month, the curves in Figure 2.4 are scaled to show the difference in shape more clearly. These curves are hereinafter referred to as *revenue curves*. Figure A.3 in Appendix A displays these curves for the different cabin classes.

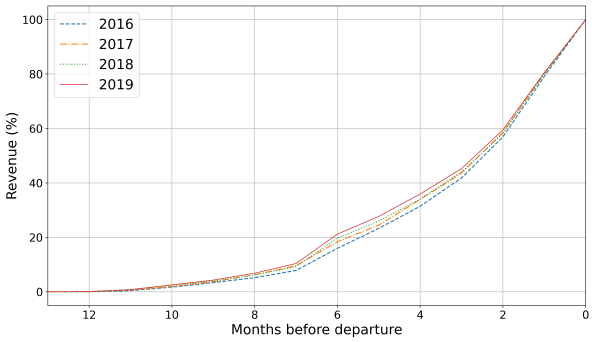
Figures 2.4a, 2.4c, 2.4e and 2.4g show the revenue curves of January, and the curves of July are displayed in Figures 2.4b, 2.4d, 2.4f and 2.4h. The biggest difference between the scaled curves are not necessarily found between months, but between channels. For example, the indirect offline channels have just past the 20% mark at 2 months before departure, whereas the other three channels are already at 40-60%. Also, for these three channels, the range of 40-60% is still quite a wide range. This shows that there is a different purchase behaviour in the last couple of months for the booking channels.

Figure A.3 in Appendix A shows the revenue curves of the cabin classes of January and July. The differences between these curves are not as big as for the booking channels, as at all times the curves of both the different cabin classes as the months are within a range smaller than 13%. The largest difference is found at 2 months before departure for July. Here, the economy class has obtained on average 46% of the total revenue whereas the business class is still at 34%.

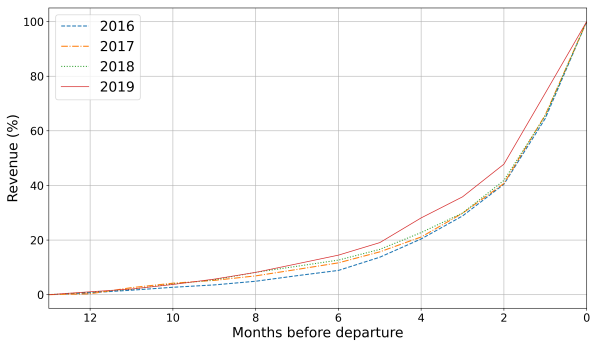
It must be kept in mind that the revenue curves visualized in Figures 2.4 and A.3 are made with the final monthly data, and they do therefore not contain any cancellations. The



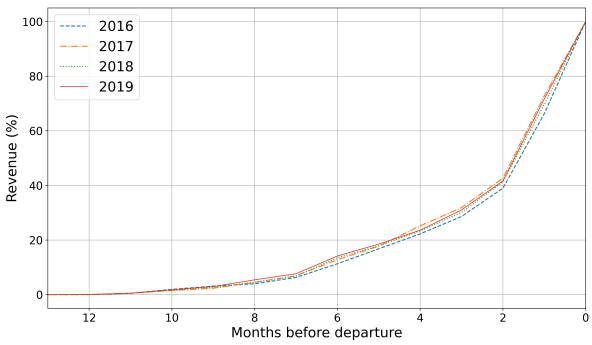
(a) Direct Online channel of January.



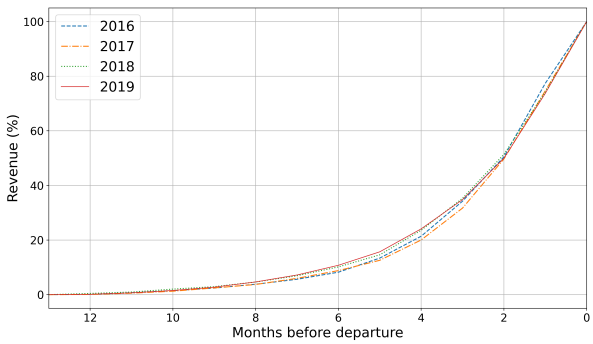
(b) Direct Online channel of July.



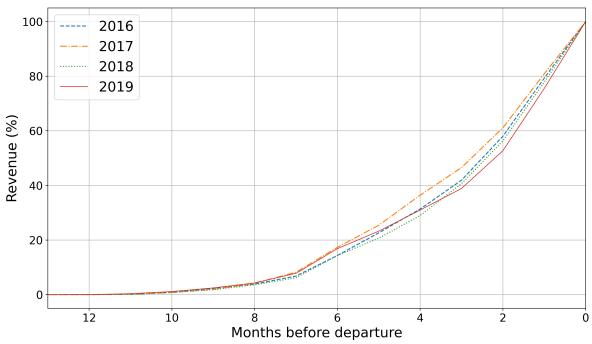
(c) Direct Offline channel of January.



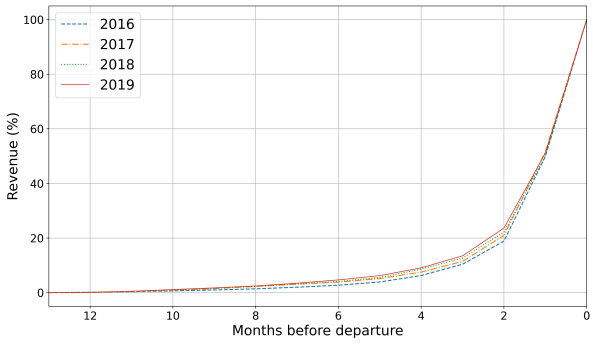
(d) Direct Offline channel of July.



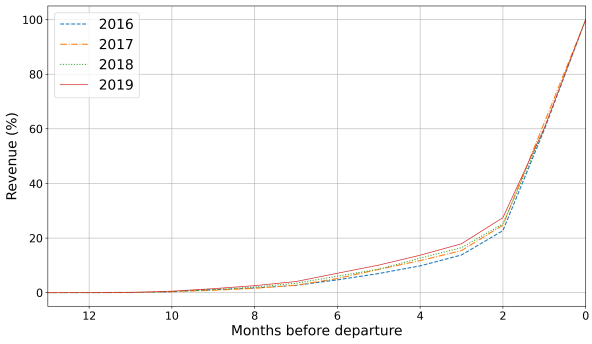
(e) Indirect Online channel of January.



(f) Indirect Online channel of July.



(g) Indirect Offline channel of January.



(h) Indirect Offline channel of July.

Figure 2.4: Channel revenue curves of all the tickets bought in PoS W for January and July 2016 - 2019. The cumulative percentage of the total earned revenue is given x months before departure.

cancellations are not displayed in the figures because cancelled tickets did not contribute to the final revenue. However, when present time data is used as an independent variable, there is still the possibility that the curves contain revenue of tickets that will be cancelled in the future, causing the revenue per ticketing month to drop again. Cancellation data for PoS's is

not directly available, as this data is only given for Point of Bookings. However, there is another way to handle cancellations. It is possible to look at the revenue curves at another point in time. For example, it enables one to look at the true revenue per ticketing months of Figure 2.4b while pretending to be in e.g. March 2019, instead of looking back at the data in for example August 2021. By doing so, it is possible to compare the revenue curves of July 2019 as seen in March 2019 to the final curves of July 2019. It enables one to see how the revenue per ticketing month decreases for a specific month as the departure months approaches, due to processed cancellations. This can be done for all the years that are available, enabling one to find a pattern in cancellations over time. However, this cancellation data is not available for years prior to 2016, which limits the data to the years 2016-2019.

2.3.2 Capacity, PaxKM and ASK

Besides the data that is presented above, there are some exogenous variables that could be relevant for this research. However, due to the expected relevance of the above-mentioned revenue data and the somewhat limited availability of other variables, the main focus lied on revenue data.

Unfortunately, not many variables can be linked to each attribute. For example, variables could be flight-specific, variables belong to a certain Point of Booking instead of a PoS which causes all data to be different, or variables are not disaggregated into all the necessary attributes. This last group of variables must be dealt with carefully in order to apply them to a PoS correctly. An example of such a variable is the Capacity. The Capacity variable shows how many seats there are available to the airline per flight. Since all four booking channels could sell the same amount of available seats for a specific flight, capacity is considered to be the same for each channel.

Also the PaxKM (total kilometers flown by passengers) can hold information on the revenue. When the expected PaxKM is higher, the chances are either that more people bought a ticket and therefore the revenue will be higher as well, or the overall length of the flights has increased. The PaxKM is correlated with the Capacity, as the maximum value of the PaxKM is determined by the flown kilometers of all planes times the capacity. A problem however is that, just as for the revenue, the exact value of the PaxKM depends on the number of customers that occupy a seat. Therefore this data is uncertain up to the moment a plane has departed. However, contrary to the revenue, the airline was already able to forecast the PaxKM. The last variable that could be of importance is the Available Seat Kilometers (ASK). The ASK data looks like the data of the PaxKM, but instead of showing how many kilometers passengers are expected to fly, it shows how many kilometers there can be flown by passengers. When this number gets higher, either more people can buy tickets and more revenue is expected, or there are more long flights. This variable is also correlated with both the PaxKM data and the Capacity.

The data of the PaxKM, Capacity and ASK variables can be divided over the PoS's, line groups and cabin classes. As the booking channels have no influence on e.g. the amount seats available in a plane, these variables can safely be used for all channels without further disaggregation.

Chapter 3

Methods

In this chapter the methods that will be used in this research are presented. In the first section, the use of the attributes and the splits in the data are explained. In Section 3.2 the modeling techniques of this research will be discussed. The application of the modeling techniques will be explained in the following two sections about the short- and the long-term model. Here it will also be clarified what data is used for which model. Lastly, information will be provided about prediction intervals and model performance.

3.1 Data disaggregation

The airline would like to have a 12-month rolling forecast overlooking different regions. As airlines have a lot of data, the airline wants to disaggregate the data in order to be able to trace back more precisely how the revenue is earned. The attributes that are going to be used are already introduced in Chapter 2, Section 2.1. It must be mentioned that one should be careful when using disaggregated data. When data is disaggregated too much, it could result in the loss of its informative power. Nonetheless, for the purpose of creating a useful model there must be zoomed in on some specific attributes.

In order to test whether the models are useful for different markets, four different PoS's have been chosen. The PoS's W and X are large PoS's with high revenues for each line group. These two countries will be used to create the models and to test them. For these large PoS's, the airline has appointed the line group, booking channel, and carrier attributes as the most important attributes. Therefore, these 3 attributes will be the main focus for the large PoS's, but it will also be tested whether the cabin attribute has any positive influence on the forecast accuracy.

PoS's Y and Z are a lot smaller and only have one major line group to focus on, as the others are smaller and of less importance. PoS Y and Z are used to see whether the models perform differently for large and small PoS's. The most important attribute for these smaller PoS's are the cabin and carrier. Because these PoS's have only one major line group, and the other line groups have only limited data, the airline prefers to not apply the line group attribute here. Hence, the focus will lie on applying the cabin class and the carrier attribute. The booking channel attribute will be tested as well, in order to see whether it could contribute to the forecast accuracy.

3.2 Modelling Techniques

In this section, the models and the ideas behind the choices that were made are explained. As shown in Section 2.3.1, twelve months before the departure date too few tickets have been sold to accurately predict the revenue. However, as the departure date comes closer, the tickets that have been sold already show how the revenue curve compares to those of previous years. Therefore, two different models will be necessary in order to provide an accurate forecast for the upcoming months in the short- and the long term. A short-term model will predict 4 months. This number is chosen as this is the furthest forecast horizon the airline actively acts upon when it comes to changes of tactics or promotions. The long-term model continues on the short-term model, up to 12 months. The short-term model will make use of the revenue curves discussed in Section 2.3.1 and variables presented in Section 2.3.2. The long-term model will use the monthly time series data given in Section 2.2. Hence, this is not the same data as the revenue curves. The use of data for the short and long-term models will be explained in more detail in Sections 3.3 and 3.4.

Different methods will be tested to see which performs the best. For each PoS, there will be many models because depending on the PoS, there will be numerous line groups, two cabins, four booking channels, and two carriers for each month in the year. The decision to disaggregate the data into so many levels is made because this would allow the airline to see more precisely where the revenue comes from, but it also allows the trends and seasonality of the different attributes to have an impact on the data. If the data would not be disaggregated, there is a risk of losing specific trends and seasonality in the data as these could be averaged out.

Moreover, by disaggregating the data it becomes possible to test whether some attributes systematically have higher predictability than others. For example, it could be the case that a certain booking channel always has a higher forecast accuracy than others, or that Economy class is easier to predict than Business class.

In the next subsections the techniques that are used for this research are introduced and explained.

3.2.1 Exponential smoothing

Holt-Winter's Exponential Smoothing Model is a version of the ES models that can take into account trends and seasonality. Since airline data is subjected to these two components, as seen in the figures in Section 2.2, these must be implemented in the model. As there are different forms of Holt-Winter's model, in this section the choice of the chosen model is explained.

The HWESM has an additive trend component and a seasonal component that can vary to be either additive or multiplicative. Also, the errors can be additive or multiplicative. When the seasonal component is chosen to be multiplicative, the model values could increase rapidly due to the characteristics of the multiplicative component. Given the type of data that is used for the short-term models, the use of additive models is advised. This way the chances that the model explodes or (heavily) over- or underpredicts become smaller.

For this model, that has an additive component for the trend, seasonality and the errors, the AAA (Additive trend, Additive seasonality, Additive errors) model is introduced. The AAA model can be seen in Equations 3.1 and 3.2 (Hyndman et al., 2001).

$$y_t = l_{t-1} + b_{t-1} + s_{t-m} + \varepsilon_t, \tag{3.1}$$

$$\begin{aligned} l_t &= l_{t-1} + b_{t-1} + \alpha\varepsilon_t \\ b_t &= \phi b_{t-1} + \alpha\beta\varepsilon_t \\ s_t &= s_{t-m} + \gamma\varepsilon_t. \end{aligned} \tag{3.2}$$

In Equation 3.1 the smoothing equation of y_t, the dependent variable at time t, is given, and Equation 3.2 shows the error-correction forms of l_t, b_t and s_t. These are elements of the State Space Vector that denote the level, trend and seasonal components at time t respectively. If these last three equations were written as a function of y_t, l_t, b_t and s_t, these would have been smoothing equations as well. However, it was chosen to write them in error-correction form, which means that the equation of y_t is used to rewrite l_t, b_t and s_t as a function of ε_t instead of y_t. ε_t is the error term at time t, which is i.i.d. $N(0, \sigma^2)$ distributed.

In Equation 3.2, α, β and γ are smoothing parameters that originate from the smoothing equations of the level, trend and the seasonal equations respectively. ϕ is a damping coefficient in case the trend must be damped and m is the length of a season within the data. For example, when you have monthly data, m will equal 12 when the season takes a year to reoccur. In the equation for b_t the factor before the error ε_t is written as separate coefficients α and β instead of one new coefficient. This is a result of the transition from the smoothing equation to the error-correction form. Moreover, these two coefficients will be used in the calculation of a prediction interval (PI).

It can be seen that the error term ε_t is present in all equations that make up the AAA model. The result is a Single Source of Error model (SSOE). When using this model, there is one single error sequence that drives all the observations as well as the variables (Hyndman et al., 2001). This causes the error terms to be perfectly correlated. Opposite of the SSOE model is the Multiple Source of Error model (MSOE). When using the MSOE, the error terms are independent and have different variances.

There are benefits to the SSOE model compared to the MSOE model, which are mentioned in Hyndman et al. (2002). The authors state that both linear and non-linear SSMs can be formulated when using the SSOE model, which is a practical benefit. Also, the perfectly correlated errors make it possible to correctly rewrite state space equations as functions of the only error term ε, as was done for the error-correction forms in Equation 3.2.

The coefficients in the HWESM model are being optimized by using the Limited-memory Broyden–Fletcher–Goldfarb–Shanno, or LBFGS, technique. This is an iterative line search algorithm related to the Broyden–Fletcher–Goldfarb–Shanno (BFGS) technique, that finds the best direction to move towards in order to find the point where the gradient equals 0 or close to zero. The difference between the LBFGS and the more known BFGS is that the BFGS has to store an approximation of the Hessian matrix whereas the LBFGS stores only some vectors of the approximation, and thereby requires less computational memory.

The variant of the HWESM model that is used here, the AAA model, is a linear representation of the State Space Models. For this linear representation, a method is suggested in Hyndman et al. (2001) to calculate a 95% confidence interval using the coefficients provided in Equations 3.1. The 95% PI will be given by $\zeta_h \pm 1.96\sqrt{v_h}$. Here ζ_h is the predicted value for h periods into the future and v_h is the corresponding variance. For the (damped) AAA model the expression for v_h is given by Equation 3.3 (Hyndman et al., 2001).

$$v_h = \begin{cases} \sigma^2 \left[1 + \alpha^2(h-1)\{1 + \beta h + \frac{1}{6}\beta^2 h(2h-1)\} + \gamma k\{\gamma + \alpha[2 + \beta m(k+1)]\}\right] & \text{if } \phi = 1 \\ \sigma^2 \left[1 + \sum_{j=1}^{h-1} \left\{\alpha^2(1 + \phi_{j-1}\beta)^2 + \gamma d_{j,m}[\gamma + 2\alpha(1 + \phi_{j-1}\beta)]\right\}\right] & \text{if } \phi \neq 1. \end{cases} \tag{3.3}$$

In this equation σ^2 is the variance estimated by $\sum_{t=1}^{n} \hat{\varepsilon}_t^2/n$, h is the forecast period for which holds that $h \geq 2$ or else $v_1 = \sigma^2$, and $d_{j,m}$ equals 1 if j=m (mod m) and 0 otherwise. Lastly, $k = \lfloor (h-1)/m \rfloor$. In the research of Hyndman et al. (2002) it was tested whether the PIs contained the actual values of the forecast. It was found that the 95% PIs tended to be too optimistic, as not always 95% of the predictions were contained by the interval. However, this phenomenon is well-known in forecasting, as stated by the authors. The coverage probability of PIs will be of importance for the current research, as the results must be usable for the airline. This means that the coverage probability must be high while the size of the interval remains small. However, the performance of the models will not be determined by using PIs. This will be done by using error measures on the point forecasts.

The performance of the models will be tested by using the Mean Squared Error (MSE), Weighted Mean Absolute Percentage Error (WMAPE), and Mean Absolute Error (MAE) error measures. The application of these error measures will be explained in more detail in Section 3.6. By comparing these scores to those of other models, the model that performs the best can be found.

3.2.2 SARIMA(X)

The SARIMA model is a Seasonal ARIMA model that has similar characteristics as the regular ARIMA model, as the dependent variable can consist of lagged values of the dependent variable or error terms. The difference between the two models is that the SARIMA model contains extra lag orders and an order of integration to include seasonality, whereas seasonality in the ARIMA models is implemented by using seasonal dummies. In Chapter 1 literature is provided that shows the successful use of the SARIMA model in the airline/tourism industry, but the application is not necessarily proved in revenue management. In order to test this application, during this research it was chosen to use SARIMA models instead of ARIMA models with seasonal dummies.

The SARIMA model has four lag orders, auto-regressive (AR) p, moving average (MA) q, seasonal AR sp and seasonal MA sq. It also has two orders of integration, integration d and seasonal integration sd, as well as the parameter m that indicates the length of a season in the

data.

Besides using the SARIMA model, it is possible to extend the SARIMA with exogenous variables, thereby creating the SARIMAX model. SARIMAX is almost completely the same as the SARIMA model, but the only difference is that for every time t, a value of exogenous variable X_t will be added. In preparation of the SARIMA(X) equations, the different components of the ARIMA model are given in Equations 3.4 and 3.5. In Equation 3.4 it is shown how the B operator works for the order of integration d, and Equation 3.5 shows the ARIMA model and its operators. Using these, it becomes easier to understand how the SARIMA(X) models, provided in Equations 3.6 and 3.7 respectively, are composed.

$$\begin{aligned} \Delta y_t &= y_t - y_{t-1} = y_t - By_t = (1-B)y_t \\ \Delta^d y_t &= y_t - y_{t-d} = y_t - B^d y_t = (1-B)^d y_t \end{aligned} \tag{3.4}$$

$$\begin{aligned} \Phi_p(B)(1-B)^d y_t &= \mu + \Theta_q(B)\epsilon_t \\ \Phi_p(B) &= 1 + \phi_1 B + \phi_2 B^2 + ... + \phi_p B^p \\ \Theta_q(B) &= 1 + \theta_1 B + \theta_2 B^2 + ... + \theta_q B^q \end{aligned} \tag{3.5}$$

$$\Phi_p(B)\Phi_{sp}(B^m)(1-B)^d(1-B^m)^{sd} y_t = \mu + \Theta_q(B)\Theta_{sq}(B^m)\varepsilon_t, \tag{3.6}$$

$$\Phi_p(B)\Phi_{sp}(B^m)(1-B)^d(1-B^m)^{sd} y_t = \mu + \psi X_t + \Theta_q(B)\Theta_{sq}(B^m)\varepsilon_t. \tag{3.7}$$

In these equations y_t is the dependent variable at time t, ϕ_q is the q^{th} order AR coefficient, θ_p is the p^{th} order MA coefficient, ε_t is the error at time t, μ is the constant mean, B is a difference operator, X_t is a (n×1) vector of exogenous variables at time t where n stands for the number of different exogenous variables, ψ is a (1×n) vector of coefficients and Φ_p and Θ_q are the collections of the p^{th} and q^{th} order coefficients, respectively.

Since there are four lag orders (p, q, sp and sq) and two orders of integration (d and sd) for which the values can be optimized for each model, it could be the case that for each booking channel and cabin different values are optimal. However, as it is desirable to have the same values for the lag orders and order of integration for all models, the SARIMA airline model is used. For this model the values of p, d, q, sp, sd and sq are set to 0, 1, 1, 0, 1 and 1 respectively. By using the same orders for all models, the SARIMA model can be tested on other data more easily and results can be compared. The SARIMA and SARIMAX airline models are given in Equations 3.8-3.9 and 3.10-3.11 respectively.

$$\Phi_0(B)\Phi_{m,0}(B^m)(1-B)(1-B^m)y_t = \mu + \Theta_1(B)\Theta_{m,1}(B^m)\varepsilon_t \tag{3.8}$$

$$y_t - y_{t-1} - y_{t-m} - y_{t-(m+1)} = \mu + \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_{1_m}\varepsilon_{t-m} + \theta_1\theta_{1_m}\varepsilon_{t-(m+1)} \tag{3.9}$$

$$\Phi_0(B)\Phi_{m,0}(B^m)(1-B)(1-B^m)y_t = \mu + \psi X_t + \Theta_1(B)\Theta_{m,1}(B^m)\varepsilon_t \tag{3.10}$$

$$y_t - y_{t-1} - y_{t-s} - y_{t-(s+1)} = \mu + \psi X_t + \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_{1_s}\varepsilon_{t-s} + \theta_1\theta_{1_s}\varepsilon_{t-(s+1)} \tag{3.11}$$

The coefficients in the SARIMA(X) Airline model are being optimized by using the LBFGS technique. As for the HWESM AAA model, the MSE, WMAPE, and MAE are used to compare the output. More information can be found in Section 3.6 on the model performance. The PIs of the SARIMA models are created based on the errors in the model and the variance of the error.

3.2.3 Dynamic Factor Model

The last modeling technique that will be tested is the Dynamic Factor Model (DFM). The DFM does not quite work in the same way as the aforementioned State Space Models, as these models were chosen because of their applications of trends and seasonal patterns. However, the DFM has been selected because it has successfully been used in other sectors, next to the fact that the airline has previously used a factor model to forecast the PaxKM. The idea behind the DFM is that there are latent dynamic factors f_t that drive the dependent variable.

A feature of the DFM that could be useful is that it can model multivariate data, for example different channels, where the channels have influence on each other. The DFM is displayed in Equation 3.12.

$$\begin{aligned} y_{t,n} &= \Lambda(L) f_{t,n} + \varepsilon_{t,n} \\ f_{t,n} &= \Psi(L) f_{t-1,n} + u_{t,n} \end{aligned} \tag{3.12}$$

Here, the time is displayed by $t \in T$, and the number of different data sets in case of multivariate data is given by N, where $n \in N$. For time $t \in T$ and $n \in N$, y is the dependent variable with size $T \times N$. $\Lambda(L)$ is a matrix of factor loadings with size $T \times L$ with lag L, f is the unobserved factor with size $L \times N$, ε is a $T \times N$ matrix of idiosyncratic errors, $\Psi(L)$ is a vector of autoregression coeffiecients with size $L \times L$, and u are the factor disturbances with size $L \times N$. The idiosyncratic errors $\varepsilon_{t,n}$ are specific for each data set, such that $\varepsilon_{t,n} \sim N(0, \sigma_n^2)$. These errors are independent of $u_{t,n} \sim N(0, Q)$, which means that $E[\varepsilon_{t,n} u_{t-k,n}] = 0 \; \forall \; k$.

In the case where 4 channels are modeled simultaneously, N equals 4. The amount of lagged factors that are added, L, can be optimized by using for example the Akaike Information Criterion (AIC). However, when the value of L is optimized for each individual data set, this can take a lot of time due to all the different PoS's and attributes. Since it is preferred to find a general model that performs well for each set of data, some cabins/channels will be optimized at random in order to find a value of L that has an AIC value that is the best overall. This value will then be used for all the data.

The factor loadings of the DFM will be optimized by using the Expectation-Maximization algorithm. The output of the model will be compared by using the MSE, WMAPE, and MAE error measures. More information can be found in Section 3.6 on the model performance. The PIs of the DFM will be created based on the variance of idiosyncratic errors $\varepsilon_{t,n}$

3.2.4 Combining methods

A recent article that discusses the combination of methods is the M4 research of Makridakis et al. (2018). In this article it is also mentioned that combining methods can increase the overall

accuracy of the model. Hence, this will be tested for the current research as well. The M4 research was not the first research of combining methods. In 1969 the authors of Bates & Granger (1969) already concluded that combining methods is beneficial for the model performance. Also, in Timmermann (2006) combining forecasts has been investigated on a theoretical level. It was stated that by combining different forecasts, accuracy can be increased compared to forecasts that are based on a single model.

In the M4 research, some researchers use Machine Learning techniques in order to find optimal weights for each of the applied methods. However, since this research focuses on many different cases due to the many PoS's, line groups, cabins, and channels, the chances of the weights being different for each one of these cases are likely to be high. Since the goal of this research is to find a model that can be implemented by all PoS's, it was chosen to use weights that are based on the simple average. This way, the weights will be the same for all cases. Moreover, the simple average has proved its worth (Timmermann, 2006). The goal of combining the methods is to increase forecast accuracy and decrease the values of the error measures.

3.3 Short-term forecast

For the short-term forecast the revenue curves will be the most important data in calculating the monthly revenue. A method is needed that can model the historical revenue curves, and by implementing the most recent data to this model, a short-term forecast will be created. As an example, there can be looked at the curves provided in Figures 2.4 and A.3. The data of 2016, 2017, 2018, and the available data of 2019 can be used to train a model, with which the last 4 months of 2019 will be predicted. This data will create a sawtooth-like time series with three full "seasons" of the years 2016-2018, and one incomplete one from 2019. Depending on the quality of the data, for each particular month, the revenue curves of the years 2016 to 2019 are available. The data of the curves is segmented into different attributes.

Training a model can be done by using adaptive models or non-adaptive models. The difference is that the coefficients of adaptive models change over time when new data is available, whereas non-adaptive models are trained by using historical data from a specific period in time, even though new data is available. Since airline data is not constant over the years, the adaptive models can take into account trends and seasonality better than non-adaptive models. Hence, adaptive models will be used to train the models. This is also advised by Kalekar et al. (2004). No complicated mechanism is necessary for the application of the adaptive model. When the current date, the number of years of training data one wants to use (lookback) and the month that needs to be predicted are given as input, the model will recognize which data to use as all data have a date-tag.

Besides using adaptive models, it can also be interesting to look at how monthly data can be used in the best way. Each month has month-specific data, but does a model perform better when using only this month-specific data, or is the performance higher when there is no distinction made between months? For example, when the data of 2016-2018 is used to predict the revenue of January 2019, either 3 series of observations of January can be used, or 36 series of observations of all the months in the previous four years. The downside is that the revenue curves of the different months do not necessarily have precisely the same shape, as shown in Figures 2.4

and A.3, and more importantly, the amount of revenue differs per month. On the other hand, the amount of data that becomes available could possibly result in a more reliable model. These two different models will be referred to as a *month specific model* when only the data of that particular month is used, and a *general monthly model* when all months are used to train the model.

The techniques that will be used to analyze the revenue curves in the short term are HWESM, SARIMA, SARIMAX, and DFM. These models will be tested using a month-specific model and a general monthly model. Lastly, the methods will be combined with the aim of improving performances. The individual predictions of the combined models will be a simple average of the predictions of the channels/cabins of the other models, as will the PIs (Grushka-Cockayne & Jose, 2020). The revenue curves will be used by each method, and the SARIMAX will use the PaxKM, Capacity, and ASK data as well.

Each complete revenue curve has 13 data points, from a specific month in the previous year up to that month in the current year. The length of the season, as mentioned in Section 3.2.1, is 13 in this case. It could be beneficial to remove the first couple of data points from a curve, as some combinations of attributes could cause the revenue curves to start with multiple zeroes, which will not benefit the models. This can be the case for smaller PoS's. For the SARIMAX model, there will be investigated which exogenous variables are of importance in the forecast, and for the DFM it will be analyzed whether univariate or multivariate data results in better forecast accuracy. The short-term (combined) models that will be tested are listed here:

- HWESM, both with month specific (SP) as general monthly (G) data
- SARIMA (SAR), both with month specific (SP) and general monthly (G) data
- SARIMAX (SARX), both with month specific (SP) and general monthly (G) data
- DFM, with month specific data
- All models combined except for DFM (All)
- HWESM SP + SAR SP + SARX SP (All SP)
- HWESM G + SAR G + SARX G (All G)
- SAR SP + SARX SP (SAREXO)
- All SP + SARX G (SP + SARX G)
- All SP + SAR G (SP + SAR G)

These models will be compared to a naive model, which predicts the revenue of a channel/cabin of a period to be the same as the revenue in the same period of the previous year. This also means that the prediction of the naive model will be the same for different forecast horizons, as they are all based on last year's results. Other combinations, such as the DFM with general monthly data, performed badly when testing some attributes at random, and were therefore not taken into account for the rest of the research.

All the cabins, channels, and carriers of each month in 2019 will be predicted 1-4 months ahead. Then, all the 1-month forecasts of the entire year will be combined to form a forecast overview of the overall 1-month prediction, and the same will be done for the 2-4 months predictions. This means that the overall 1-month prediction is a combination of predictions of January 2019, predicted in December 2018, up to December 2019, predicted in November 2019.

It is also important to process the cancellations of previous years into the data of the current

year. The revenue curves that are available from previous years, can also be obtained when using another *snapshot date*. This snapshot date means that it is possible to change the date on which the data is looked at. For example, when the revenue curve of December 2019 is displayed in 2021, there are 13 data points from December 2018 - December 2019. However, when using the snapshot date of September 1^{st} 2019 there will be 9 data points, from December 2018 - August 2019. This data has not yet processed all the cancellations, as some will take place after September 1^{st}. For the historical data both the final revenue curves as the revenue curves from other snapshot dates that include cancellations are available, enabling one to analyze the differences between these snapshot dates and the final curves.

The analysis of these differences can be applied to the current year, which only has the incomplete snapshot date data. As it is not precisely known how many tickets the revenue consists of or what these tickets' prices were, there will simply be looked at the percentage of revenue that was canceled in the previous years. This percentage will be used to deduct possible cancellations from the revenue curve in a simple way.

Lastly, it will be checked whether the chosen attributes are of importance or whether the performance will increase significantly when attributes are left out. The biggest difference will be between the two groups of PoS's (large and small), and not between the individual PoS's within these groups. Since the airline already has its preferences about which attributes to use (line group, carrier, and booking channel for large PoS's and cabin class and carrier for small PoS's), it will be investigated if these preferences are also the best attributes to apply to the data. To test whether the use of these preferences is optimal, the cabin class attribute will be included for large PoS's, and the booking channels will be added for small PoS's to compare the outcomes.

3.4 Long-term forecast

For the long-term forecast, the time series of the monthly revenue as given in Section 2.2 will be of importance as the revenue curves hold less information in the long term. As for the short-term model, the long-term model needs to be able to take into account trends and seasonality in order to correctly perform a rolling forecast of 12 months.

An adaptive model will prove to be more accurate than a non-adaptive model since the most recent data holds more information on the current trends than data that is not updated over the years. Hence, the use of an adaptive model is a more suitable approach. Just as for the short-term model, the data and therefore the parameters will automatically be updated each month when the current date and the desired lookback are given as input, since each data point has a date-tag.

A problem that is encountered for the long-term model is that a lot of data is not yet available for flights that depart in 12 months. This is also the reason that the revenue curves cannot be used. It is already known how many flights will depart approximately, but, for example, the ticket prices depend on the demand and the pricing tactics that will be decided in the upcoming months. Therefore it is hard to use reliable independent variables. The most reliable data is monthly revenue as displayed in Figures 2.1 - 2.3, and not the revenue curves as used in the short-term model. If other variables are used, the chances of using independent variables that

contain uncertainty increase, which increases uncertainty in the forecast.

The state space representations HWESM and SARIMA will be used again. Due to the adaptability of these models, they are believed to perform well. Nonetheless, the models are also combined using a simple weighted average between the models for both the point forecasts and the upper and lower bound of the PIs. Next to these two models, the DFM is used. The data that will be used is the same as for the HWESM and SARIMA models. There will be investigated whether the DFMs perform better with univariate or multivariate data and the DFM will also be taken into account when combining models. The long-term (combined) models that will be tested are listed here:

- DFM
- HWESM
- SARIMA (SAR)
- All three models (All)
- HWESM + SARIMA (HWESM-SAR)
- HWESM + DFM (HWESM-DFM)
- DFM + SARIMA (DFM-SAR)

As for the short-term model, these models will be compared to a naive model.

All the cabins, channels, and carriers of each month in 2019 will be predicted 5-12 months ahead. Then, as was also mentioned for the short-term model, all the 5-month forecasts of the entire year will be combined to form a forecast overview of the overall 5-month prediction, and the same will be done for the 6-12 months predictions. This means that the overall 5-month prediction is a combination of predictions of January 2019, predicted in August 2018, up to December 2019, predicted in July 2019.

3.5 Aggregated prediction intervals

As mentioned before, for each PoS, month, and line group, there will be multiple channels or cabins to predict. Each of these forecasts will have its own prediction and PI, but in order to summarize the data it will be beneficial to aggregate all the predicted values to obtain the total predicted revenue per month. The PI of the total predicted revenue per month can however not be obtained by adding the individual PIs due to possible covariance between the channels. For each individual forecast of a channel or cabin, the model's errors must be used to find the correlation between the errors of the channels or cabins. By using the correlations, it is possible to apply these to calculate the total covariance and also take into account the increased variance due to out-of-sample predictions. When a covariance matrix is calculated directly from the model's errors, the increased variance of a forecast horizon larger than 1 is not taken into account. In case a combination of models is used, the errors are a weighted average of the combined errors.

3.6 Model Performance

In order to choose the right method, for each forecast the absolute percentage error, squared error, and absolute error are calculated. These values can provide a quick overview of the performance of the different models. Since for each PoS and line group the monthly total revenue is composed of the individual forecasts of the other attributes, these values are aggregated and the total MSE, WMAPE, and MAE are calculated. Here the weight of an individual forecast is determined by the actual revenue of that forecast divided by the total actual revenue of that month. This is done because the MAPE could provide misleading insights when a forecast of a very low revenue has a high MAPE, and a forecast of a high revenue has a low MAPE. It is advised to use multiple error measures to determine the model performance, as a single one could provide one-sided information that does not necessarily portray the situation correctly.

To test whether combining methods has a significant impact on the results, the Diebold-Mariano (DM) test is applied (Diebold (2015), D. Harvey et al. (1997)). The DM-test compares for example MSE values of various forecast horizons for different models. Hereby the MSE values form a time series that can be compared to another model's MSE values. The DM-test looks like a t-test and is performed to see if the MSE values of the two models differ significantly or whether these can be considered the same. Here the null hypothesis is that the MSE values of the two models do not differ significantly from each other.

Lastly, there will be looked at the PIs. Since the outcome of this research will be applied in a very practical manner, the correctness of the PIs is of importance. An interval that is too small will not contain 95% of the predictions, whereas an interval that is too large will not be useful. As was already mentioned by Hyndman et al. (2002), for the ES methods, for example, it is a known phenomenon that not every 95% interval contains 95% of the predictions. Therefore, for each method, there will be looked at the coverage probability over a whole year of predictions. They also found that the ES models that found the best MAPE values, had the worst PI coverage probabilities. Hence it will be interesting to see if this holds for the other models as well. The models in this research will not necessarily be rejected based on a small PI coverage probability, as the opposing PIs that have a large(r) coverage probability can have a very large PI and can therefore be uninformative at the same time.

Chapter 4

Results & Discussion

In this chapter, the results of this research will be presented and discussed. At first, some general observations will be provided, followed by the results of both the short- and long-term models. Lastly, a discussion of the results will be given. In order to reduce the information given in this chapter, only the results of large PoS W and small PoS Y are given. The results of PoS's X and Z are given in Appendix B.

4.1 Results

As the results would be too extensive if the forecasts of the individual channels are presented, the values in this chapter are summarized. The total prediction of each month consists of multiple carriers and channels or cabin classes, which have their own squared error, absolute error, and absolute percentage error. The monthly values of the MSE, MAE, and WMAPE are the (weighted) means of the individual predictions. Besides that, the predictions of each month are aggregated as well such that the MSE, MAE, and WMAPE are given for the entire year of 2019. For the yearly WMAPE this means that monthly WMAPE values are weighted by the monthly revenues. The tables in this section will contain the error measures for all (combinations of) models and forecast horizons. In order to increase interpretability, the WMAPE and MSE results of the naive model are chosen to be a benchmark and the results of the other models are given relative to this naive model. This means that the WMAPE and MSE results of the other models are divided by the results of the naive model. Moreover, the results of the naive model are constant over time as for each forecast horizon the prediction is the same as last year's revenue.

4.1.1 General results

From the research it could be concluded that the airline's preferred attributes for both small and large PoS's resulted in the best results. When for the large PoS's the cabin class attribute was included, the models performed slightly worse. The difference in performance was not large, but since the addition of the cabin class did not enhance the model and the airline initially did not want to include this attribute, it will be left out. Thus, the attributes that are used for large PoS's are line groups, booking channels, and carriers.

When the booking channels were included for the small PoS's the results were worse compared to the case where only the preferred attributes were used. Because the small PoS's have less data, the extra disaggregation led to insufficient data for each channel-cabin combination. Also,

not every small PoS uses all booking channels consistently, which led to bad performances. Therefore, the attributes that will be used for small PoS's are the cabin classes and carriers, as proposed by the airline.

For the SARIMAX model, the variables ASK, PaxKM, and Capacity were tested. Since there are many models there cannot be given a single significance value of the coefficients. With the aim of using the same variables for all models, it was chosen to use the variables ASK and PaxKM, because the coefficient of the Capacity was insignificant on many occasions when testing some combinations of attributes. For the DFM it was tested whether the use of multivariate data would outperform univariate data. It was found that by using multivariate data, errors increased compared to the use of univariate data. Hence, only the results of the univariate model will be presented in the next subsections.

The predictability of channels or cabins can differ between the PoS's, as in each PoS there can be a preference for certain tickets or the way these tickets are bought. However, for the large PoS's it can be said that in general the larger channels, Direct Online and Indirect Offline, are more accurately forecasted than the others. Moreover, predictions of the line groups that bring in only a small share of the total revenue are less accurate than those of larger line groups. The results of the small PoS's show that in general the forecasts of Business class are more accurate than Economy class.

4.1.2 Short-term models

The short-term models calculated the revenue forecast for 1-4 months. The results that are about to be presented are summarized for all predictions in 2019, and for these models a training set of 3 years was used. For each of the months January till December in 2019, a forecast was made 1, 2, 3 and 4 months beforehand, and for each forecast horizon the predictions of all the months in 2019 are aggregated such that the results of the whole year are shown. This is also explained in Section 3.3. These results can be disaggregated into the different carriers, cabins and channels, but this will not be done here due to the extensiveness of the results.

Large PoS's

As the large PoS's have 10 different line groups, the results of the line groups will be combined by using a weighted mean where the weights are determined by the yearly revenue of the line groups divided by the yearly revenue of the entire PoS. The results of PoS W are given below in Tables 4.1 and 4.2 and the results of PoS X are given in the Tables B.1 and B.2 in Appendix B.

Tables 4.1 and 4.2 show the short-term results of the WMAPE, MSE, PI coverage probabilities and MAE error measures of PoS W. The results are given for a forecast horizon of 1, 2, 3 and 4 months. For example, this means that column *1 WMAPE* contains the weighted WMAPE values of all the 1-month predictions of that PoS.

For PoS W it can be seen that there are five models with similar results. These models are *SAR SP*, *SARX SP*, *SAREXO*, *SP + SARX G* and *SP + SAR G*. By applying the DM-test on the MSE values of these models, it is found that both *SAR SP* and *SAREXO* have MSE

Table 4.1: Relative error values for the aggregated short-term models of PoS W. For each forecast horizon in the short-term model and for all (combinations of) models, the values of WMAPE and MSE are given relative to the time-independent errors of the naive model, which are 8.80 % and 34×10^{10} respectively. Also the total PI coverage probabilities are provided per model.

Model	WMAPE				MSE				PI probability
	1	2	3	4	1	2	3	4	
HWESM SP	0.67	1.04	1.25	1.44	0.71	1.50	2.12	2.79	0.403
HWESM G	1.02	1.31	1.52	1.66	2.12	2.65	3.00	3.32	0.713
SAR SP	0.55	0.79	0.89	0.97	0.32	0.56	0.65	0.74	0.670
SAR G	0.82	1.10	1.25	1.44	1.71	2.35	2.65	3.12	0.764
SARX SP	0.58	0.76	0.85	0.94	0.62	0.62	0.65	0.82	0.495
SARX G	0.81	1.10	1.25	1.44	1.59	2.29	2.62	3.09	0.763
DFM	1.40	1.43	1.46	1.49	1.65	1.85	2.00	2.12	0.345
All	0.60	0.81	0.93	1.03	0.62	0.97	1.09	1.32	0.748
All SP	0.55	0.80	0.91	1.01	0.38	0.68	0.85	1.06	0.607
All G	0.83	1.10	1.28	1.42	1.56	2.18	2.47	2.91	0.753
SAREXO	0.55	0.76	0.86	0.94	0.38	0.56	0.62	0.74	0.656
SP + SARX G	0.58	0.81	0.94	1.04	0.50	0.79	0.94	1.15	0.676
SP + SAR G	0.53	0.75	0.85	0.94	0.41	0.68	0.79	1.00	0.720

Table 4.2: MAE values for the aggregated short-term model of PoS W. The error values are displayed for each forecast horizon in the short-term model and for all (combinations of) models. The errors of the naive model are time-independent.

Model	MAE (€1000)			
	1	2	3	4
Naive	257	257	257	257
HWESM SP	206	308	362	409
HWESM G	324	392	436	466
SAR SP	153	213	236	252
SAR G	258	331	366	414
SARX SP	170	208	227	246
SARX G	254	329	366	416
DFM	370	386	397	408
All	181	233	262	285
All SP	159	223	252	273
All G	261	333	372	406
SAREXO	158	208	230	243
SP + SARX G	171	227	257	277
SP + SAR G	155	215	239	259

values that are significantly lower than those of the other three methods, but they do not differ significantly from each other at a 95% level. A DM-test on the WMAPE and MAE values shows that the WMAPEs of *SAREXO* are significantly lower than those of *SAR SP* at a 95% level, and the MAEs of *SAREXO* are significantly lower than those of *SAR SP* at a 90% level.

Tables B.1 and B.2 in Appendix B show that for PoS X approximately the same results are found and also the *SAREXO* model gives the best results. For both PoS's the *SAREXO* models outperform the naive model.

Small PoS's

In Tables 4.3 and 4.4 the short-term results of the error measures of PoS Y are given.

Table 4.3: Relative error values for the aggregated short-term models of PoS Y. For each forecast horizon in the short-term model and for all (combinations of) models, the values of WMAPE and MSE are given relative to the time-independent errors of the naive model, which are 15.79 % and 323 $\times$ 10^{10} respectively. Also the total PI coverage probabilities are provided per model.

Model	WMAPE				MSE				PI probability
	1	2	3	4	1	2	3	4	
HWESM SP	0.17	0.36	0.50	0.55	0.02	0.11	0.20	0.24	0.990
HWESM G	0.41	0.69	0.90	1.08	0.14	0.33	0.57	0.83	0.859
SAR SP	0.35	0.88	1.09	1.28	0.12	0.66	1.02	1.43	0.990
SAR G	0.23	0.46	0.66	0.85	0.03	0.13	0.26	0.40	0.984
SARX SP	0.38	1.09	0.80	0.89	0.26	3.75	0.63	0.67	0.849
SARX G	0.23	0.45	0.67	0.86	0.03	0.12	0.27	0.41	0.979
DFM	1.50	1.74	1.77	1.80	2.28	3.10	3.22	3.37	0.657
All	0.19	0.42	0.42	0.47	0.03	0.19	0.15	0.19	1.000
All SP	0.22	0.61	0.52	0.59	0.05	0.55	0.24	0.29	0.984
All G	0.24	0.44	0.62	0.77	0.03	0.13	0.26	0.40	0.964
SAREXO	0.31	0.91	0.82	0.90	0.11	1.39	0.55	0.67	0.943
SP + SARX G	0.20	0.49	0.41	0.44	0.03	0.33	0.15	0.18	0.990
SP + SAR G	0.20	0.49	0.41	0.43	0.03	0.33	0.15	0.17	0.990

Table 4.4: MAE values for the aggregated short-term model of PoS Y. The error values are displayed for each forecast horizon in the short-term model and for all (combinations of) models. The errors of the naive model are time-independent.

Model	MAE (€1000)			
	1	2	3	4
Naive	1,109	1,109	1,109	1,109
HWESM SP	181	394	537	590
HWESM G	449	742	970	1,167
SAR SP	380	958	1,178	1,387
SAR G	246	498	717	916
SARX SP	416	1,185	861	962
SARX G	252	487	724	932
DFM	1,620	1,889	1,918	1,948
All	209	459	451	513
All SP	238	662	566	635
All G	262	479	673	830
SAREXO	339	981	884	979
SP + SARX G	212	536	444	477
SP + SAR G	211	534	444	464

There is no method that performs the best for all measures and forecast horizons. At the forecast horizon of 3 and 4 months, the combined models *SP + SARX G* and *SP + SAR G* have the lowest error measures, whereas *HWESM SP* has the lowest errors for a forecast horizon of 1 and 2 months. The *All* combination regularly comes in with the second or third best results. When applying the DM-test it is found that the MSE values of these four models differ significantly from the other models, but not necessarily from each other. The same holds

for the values of the WMAPE and MAE error measures. It is only found that for all 3 measures, the values of *SP + SAR G* are significantly less than those of the *SP + SARX G* combination at a 90% level.

In Tables B.3 and B.4 in Appendix B the results are given for PoS Z. For this PoS the lowest errors are found for *SP + SARX G* and *SP + SAR G*. When the DM-test is used to compare their MSE values, it was found that those of *SP + SARX G* are significantly less than those of *SP + SAR G*. The WMAPE and MAE values do not differ significantly at a 95% level. For both PoS Y and Z the naive model is being outperformed by the *SP + SAR G* and *SP + SARX G* combinations.

4.1.3 Long-term model

The results of the long-term models, as mentioned in Section 3.4, can be separated in the large PoS's W and X, and the small PoS's Y and Z. These PoS's are all compared to a naive model. For these models a training set of 4 years of data performed better than 3 years. The forecast horizon consists of 5-12 months. The forecast horizons 1-4 months were tested as well but they were all outperformed by the short-term model. For that reason only the results of 5-12 months are displayed. Just as for the short-term model, the results per forecast horizon are summarized for all the months in the year. This is also explained in Section 3.4.

Large PoS's

The results presented here are a weighted mean of the results of the individual line groups of PoS's W and X. The results of PoS W are shown in this paragraph, and the results of PoS X are shown in Figures B.5 - B.8 in Appendix B. Tables 4.5 - 4.8 display the results of PoS W.

Table 4.5: Relative WMAPE values for the aggregated long-term models of PoS W. Values are given relative to the time-independent naive model, with error 8.80 %. The error values are displayed for each forecast horizon in the long-term model and for all (combinations of) models.

Model	**WMAPE**							
	5	**6**	**7**	**8**	**9**	**10**	**11**	**12**
DFM	2.01	2.09	2.09	2.16	2.24	2.32	2.40	2.53
HWESM	1.58	1.64	1.61	1.60	1.55	1.46	1.38	1.38
SARIMA	1.54	1.64	1.81	1.94	2.25	2.46	2.69	2.96
All	1.46	1.50	1.51	1.50	1.58	1.61	1.66	1.76
HWESM-SAR	1.42	1.45	1.49	1.49	1.56	1.57	1.51	1.56
HWESM-DFM	1.57	1.57	1.54	1.48	1.44	1.40	1.36	1.40
DFM-SAR	1.62	1.70	1.76	1.89	2.08	2.22	2.38	2.59

By looking at Tables 4.5 - 4.7 it becomes very clear that the error measures of the naive model are always the lowest. This means that for this PoS, the naive model is the best model. When looking at the other models, the results of PoS W show that *HWESM-SAR* and *HWESM-DFM* are the two next best models, without there being model that directly stands out. By applying the DM-test on the MSE it is found that the MSE values of *HWESM-SAR* are significantly less than those of *HWESM-DFM* at a 95% level. The same test for the WMAPE and MAE values show no significant difference between the two models at a 95% significance level. When looking

Table 4.6: Relative MSE values for the aggregated long-term models of PoS W. Values are given relative to the time-independent naive model, with error 34×10^{10}. The error values are displayed for each forecast horizon in the long-term model and for all (combinations of) models.

Model	MSE							
	5	**6**	**7**	**8**	**9**	**10**	**11**	**12**
DFM	5.65	6.12	6.38	6.38	6.56	7.00	7.56	8.82
HWESM	2.50	2.76	2.29	2.21	2.21	1.74	1.74	1.74
SARIMA	1.97	2.12	2.59	2.85	3.94	4.65	5.88	7.65
All	2.35	2.35	2.32	2.12	2.15	2.06	2.18	2.65
HWESM-SAR	1.82	1.82	1.68	1.59	1.74	1.56	1.53	1.76
HWESM-DFM	2.85	2.82	2.53	2.18	1.82	1.53	1.38	1.56
DFM-SAR	3.18	3.41	3.79	3.88	4.44	4.91	5.79	7.21

Table 4.7: MAE values for the aggregated long-term models of PoS W. The error values are displayed for each forecast horizon in the long-term model and for all (combinations of) models. The errors of the naive model are time-independent.

Model	MAE (€1000)							
	5	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Naive	257	257	257	257	257	257	257	257
DFM	519	547	557	570	589	612	641	690
HWESM	400	423	408	406	393	360	356	355
SARIMA	400	426	466	497	583	645	720	802
All	374	384	386	379	398	411	431	464
HWESM-SAR	362	366	373	375	389	390	383	403
HWESM-DFM	396	395	391	374	351	339	339	352
DFM-SAR	423	444	461	495	546	586	637	706

Table 4.8: Average PI coverage probability of the long-term models of PoS W.

	DFM	**HWESM**	**SAR**	**ALL**	**HWESM-SAR**	**HWESM-DFM**	**DFM-SAR**
PI probability	0.781	0.829	0.911	0.912	0.912	0.854	0.915

at the results of PoS X in Appendix B, it can be seen that *HWESM-DFM* and the naive model show the best results. *HWESM-DFM* performs better than the naive model if we look at the WMAPE and MAE measures, except for a forecast horizon of 11 and 12 months. However, the naive model shows better results for the MSE measure. By applying a DM-test on all three error measures it can be shown that WMAPE values of *HWESM-DFM* are significantly less than those of the naive model, the naive model has MSE values that are significantly less than those of *HWESM-DFM*, and the MAE values do not differ significantly at a 95% level.

Small PoS's

In Tables 4.9 - 4.12 the long-term results of PoS Y are displayed.

There can be seen that all the WMAPE values, all the MAE values, and all but one MSE values of the *HWESM-DFM* model are the lowest, hence the best. By using the DM-test, it can be said that the error values of *HWESM-DFM* for PoS Y are significantly less than the error values of the other models at a 95% level. The coverage probability of the PIs is the lowest of

Table 4.9: Relative WMAPE values for the aggregated long-term models of PoS Y. Values are given relative to the time-independent naive model, with error 15.79 %. The error values are displayed for each forecast horizon in the long-term model and for all (combinations of) models.

Model	WMAPE							
	5	**6**	**7**	**8**	**9**	**10**	**11**	**12**
DFM	0.88	1.01	1.08	1.08	1.04	0.96	1.01	1.08
HWESM	1.24	1.29	1.23	1.10	1.03	0.99	1.21	1.39
SARIMA	1.14	1.25	1.33	1.37	1.30	1.54	2.10	2.75
All	0.70	0.83	0.93	0.98	0.92	0.89	1.19	1.52
HWESM-SAR	1.17	1.20	1.18	1.15	1.02	1.11	1.61	1.94
HWESM-DFM	0.60	0.72	0.87	0.91	0.85	0.71	0.81	0.95
DFM-SAR	0.61	0.74	0.89	1.00	1.02	1.04	1.34	1.76

Table 4.10: Relative MSE values for the aggregated long-term models of PoS Y. Values are given relative to the time-independent naive model, with error 323×10^{10}. The error values are displayed for each forecast horizon in the long-term model and for all (combinations of) models.

Model	MSE							
	5	**6**	**7**	**8**	**9**	**10**	**11**	**12**
DFM	0.79	1.16	1.35	1.30	1.23	1.02	1.10	1.09
HWESM	1.51	1.53	1.35	1.10	0.86	0.75	1.28	1.69
SARIMA	1.21	1.41	1.56	1.79	1.63	2.06	0.46	0.82
All	0.50	0.63	0.82	0.96	0.89	0.76	1.50	2.42
HWESM-SAR	1.31	1.36	1.3	1.25	1.00	1.07	2.41	4.02
HWESM-DFM	0.36	0.52	0.73	0.81	0.72	0.46	0.69	0.91
DFM-SAR	0.34	0.55	0.87	1.15	1.20	1.13	2.02	3.25

Table 4.11: MAE values for the aggregated long-term models of PoS Y. The error values are displayed for each forecast horizon in the long-term model and for all (combinations of) models. The errors of the naive model are time-independent.

Model	MAE (€1000)							
	5	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Naive	1,109	1,109	1,109	1,109	1,109	1,109	1,109	1,109
DFM	977	1,124	1,201	1,199	1,155	1,061	1,117	1,198
HWESM	1,380	1,427	1,369	1,223	1,139	1,094	1,339	1,537
SARIMA	1,262	1,383	1,474	1,522	1,437	1,707	2,330	3,051
All	781	917	1,033	1,083	1,020	983	1,315	1,682
HWESM-SAR	1,293	1,334	1,308	1,272	1,130	1,228	1,790	2,153
HWESM-DFM	664	804	962	1,010	941	785	899	1,058
DFM-SAR	680	818	990	1,113	1,129	1,154	1,489	1,951

Table 4.12: Average PI coverage probability of the long-term models of PoS Y.

	DFM	**HWESM**	**SAR**	**All**	**HWESM-SAR**	**HWESM-DFM**	**DFM-SAR**
PI probability	0.990	0.927	0.990	1.00	0.990	0.982	1.00

all the models, however this value is still 98.2 % while the PIs are 95% intervals. Hence, this is a good result.

The results of PoS Z are a bit different, as can be seen in Tables B.9 - B.12 in Appendix B. Here the naive model outperforms every other model for every error measure. By using a DM-test, the *HWESM-DFM* combination can be called the second best model, as all its error

measures are significantly less than the other six models at a 95% level. Only the MSE values of *DFM* did not differ significantly from *HWESM-DFM* on a 95% level.

4.2 Discussion

First, the used attributes will be discussed. No results have been presented that show the outcome when other attributes have been used, as this would be too extensive. However, there can be said that for the small PoS's the data was not consistent and sufficient enough to be disaggregated into line groups, channels, cabins, and carriers. By dropping the line group and the channel attribute, the overall data contained fewer missing values and proved to be more useful. Also, the airline was not interested in the line group and channel attribute for small PoS's. For the large PoS's the cabin attribute was left out of consideration. The results of the large PoS's with and without the cabin attribute were similar, but since the airline was not interested in the cabin attribute for large PoS's it was not included.

For the SARIMAX model the variables Capacity, ASK and PaxKM were tested. Since there are many models for all the different attributes, there cannot be given a single significance value for the variable coefficients. However, the inclusion of the capacity variable often led to insignificant coefficients when selecting some attributes at random. Because of these results, this variable was not included in the models. Also, the variables are all correlated, as an increase in capacity, which is an increase in available seats, makes it possible for the PaxKM and ASK to increase as well. Besides the SARIMAX model, it was expected that the possibility of using multivariate data in the DFM would lead to better results as it could take into account interactions between carriers, channels or cabins directly. However, the multivariate data unexpectedly underperformed considerably compared to the univariate data, which led to the use of the univariate data only.

Even though there was a lot of data available, not all the data could be used to test the models. For example, as the necessary attributes could not always be applied to the historical data, the usable data set decreased in size from the years 2005-2019 to 2015-2019. Besides that, for the revenue curves and the cancellation data, only data from 2016-2019 was available, which resulted in 3 years of training data and 1 year of test data. It would have been preferred to have more years of data to train the models with. This would have also made it possible to test the adaptive model on other years than 2019. This was not possible during this research, but it is very much advised to do this in the future. The models already showed good results over a large variety of PoS's, line groups, channels, cabins, carriers, and months, but the addition of extra years would show the performance of other years as well and it would have allowed the current optimized lookback values to be tested on e.g. the year 2018.

Nonetheless, over the years flights schedules have been adjusted which resulted in new flight routes or even the removal of routes within line groups. These changes can have a large effect on the revenue, which could lead to increased inaccuracies when (outdated) data is used of routes that are no longer flown. Hence, the lack of more years of data is not necessarily a huge issue. Lastly, there must be mentioned that some revenue curves without cancellation data were tested,

which enlarged the data set to the years 2015-2019. These results showed that the addition of older data did not lead to better results, as the accuracies dropped. Since this data did not include cancellation data, it was not sufficient for this research. However, it clearly showed that the addition of extra data did not lead to better results for the 2019 forecast.

During this research, the focus lied on high-level revenue forecasting instead of individual flights, as preferred by the airline. This led to the research as it is now and its minor issues regarding for example the data. If the revenue of individual flights was supposed to be forecasted, a lot of data and new variables had become available, such as ticket prices, competition for specific destinations, and recurrent flights. However, due to this increase in the amount of available data, it would have taken a lot more time to find the right data and models for all the different flights. Techniques that would have become more interesting to use in this case are Neural Networks, as used in L. R. Weatherford et al. (2003), and Machine Learning methods. For further research, this might be interesting to look into and possibly combine such methods with the current time series methods to improve the forecast accuracy. As mentioned in Chapter 1 regarding Makridakis et al. (2018), a model that consisted of a combination of time series methods and Machine Learning outperformed other models.

Regarding the results of the short- and long-term models, it can be seen that the results of the large PoS's are not necessarily the same as for the small PoS's, which shows that it was helpful to look at these PoS's separately. Within these sets of small and large PoS's, there can be seen some consensus in which models perform the best. Also, the model that comes out to be the best is always either a combination of methods or the naive model. This shows that combining forecasts is also profitable in revenue management for airlines.

For the short-term models, different models turned out to perform well. For both large PoS's the *SAREXO* combination showed the best results. For the small PoS's, the results were not so definite. PoS Y has four models for which the performance does not differ significantly, and PoS Z has two best models for which only the MSEs differ significantly from each other, in favor of *SP + SARX G*. For PoS Y, *SP + SAR G* and *SP + SARX G* would have scored better if the results of *SARX SP* would not have been this bad for the 2 month forecast. Out of the 48 predictions for the whole year, 3 predictions caused errors that led to a monthly MSE 300 times the size of other months. Without these few bad predictions, the scores of *SARX SP* and therefore *SP + SARX G* and *SP + SAR G* would have been a lot better for a 2 month forecast. When this problem is dealt with, for both small PoS's the model combinations *SP + SARX G* and *SP + SAR G* will perform well for all forecast horizons. From the results of these two small PoS's none of these two models can be appointed to be the best model conclusively.

The best models outperform the naive model. This is very important because the naive model resembles the current general idea of the forecasting methods of the airline. The short-term models now forecast up to 4 months, however, during this research some combinations of attributes were tested at random to see how they would perform if 5 or 6 months were forecasted. These showed similar results as the 1-4 month forecasts and were therefore promising. Due to the time in which this research had to be conducted, it was not possible to test these extended forecast horizons in full, but it is believed that the short-term model can outperform the long-

term model at a forecast horizon of 5 and 6 months, as it already outperformed the long-term model for 1-4 months. For larger forecast horizons the revenue curves probably hold too little information to predict accurately as only a small portion of tickets has been sold yet.

In this research revenue curves were used for the short-term model, which was a new approach in revenue management. However, when using this data as a sawtooth pattern, the models that are used in this research might not all be perfectly suited for the data. The SARIMA model, for example, is best used when using time series as presented in Section 2.2, which was done for the long-term model. However, to prevent the use of even more models for both short- and long-term models, it was chosen to test the models on both the time series of Section 2.2 and the revenue curves. The use of diffusion models was considered as these models can look at the adoption of products, which are sold tickets in this case. However, since the revenue earned per ticketing month only increases over time until the departure date, the data did not fit these models either. Even though the short-term models showed good results, in order to optimize the use of revenue curves extra research should be done to find models that fit the curves better. Continuing on the use of the SARIMA model, it would have been informative if also an ARIMA model with seasonal dummies was tested. For both the short- and long-term models, this would have shown whether the use of the SARIMA model was the right choice. Unfortunately, there was no time to test this at the end of the research.

The long-term models do not always outperform the naive models. This was not as expected, as the models take into account both seasonality and trends, whereas the naive model is only a simplified seasonal model. Nonetheless, when the forecast horizon increases to 12 months, the uncertainties and error measures are bound to go up, which is something the naive model is not affected by. Apart from the naive model, the model that almost always shows the best results is *HWESM-DFM*. For PoS's X and Y this combination often outperforms the naive model, whereas it performs worse for PoS's W and Z. For large PoS W also the *HWESM-SAR* combination performs well, but since the *HWESM-DFM* outperforms *HWESM-SAR* significantly for the other large PoS, PoS X, the *HWESM-DFM* is chosen to be the better combination out of the two for large PoS's. In the future, the airline wants to add business knowledge to the models in order to make the models of the individual PoS's more PoS-specific. This research focused on creating a general model, but the addition of business knowledge might overcome the outperformance by the naive model.

As mentioned in Chapter 1, Hyndman et al. (2002) stated that the ES models show good results, especially for a forecast horizon up to 6 periods, and also that the ES models were on the same level as ARIMA models. Looking at the short-term results, the SARIMA models outperform the ES models most of the time. However, when looking at the long-term models, the ES models seem to outperform the SARIMA models when the forecast horizon increases. This is not perfectly in line with the research of Hyndman et al. (2002). Also interesting to mention is the output of the DFMs. The DFM performs badly on the short-term model but is part of the best combination for the long-term model. It was found that the error measures of the DFM did not decrease as much when the forecast horizon decreased as the other models. This led to poor performances on the short-term model, but relatively better performances on

the long-term model. This was also the reason the DFM was not included in the *All*, *All SP*, or even the *All G* model, as the error measures were simply too high to be useful in model combinations.

Also, in Hyndman et al. (2002) it was mentioned that the models that obtained the best MAPE values had the worst PI coverage probabilities. This statement was based on the results of ES methods. When looking at the results of this research, similar patterns can be found. If the short-term results are compared, especially between the *SP* and *G* models, there can be said that the *SP* models have better WMAPE values than the *G* models of the same method, but their PI coverage probability is much lower. The cause of this was the higher variance of the errors in the *G* models, which resulted in PIs up to two times the size of those of the *SP* models, which led to a higher coverage probability. Now the practical applications must be kept in mind. A high coverage probability is preferred, but sometimes the PIs can become so large that a PI becomes useless for the airline. That is why the PIs are mentioned but the best model is not determined by them. Overall the best models hardly ever obtained PI coverage probabilities equal to or larger than 95%, but combining forecasts increased the coverage probability compared to the individual models.

Chapter 5

Conclusion

During this research, it is investigated whether a forecasting model could be created that accurately predicts monthly revenue for 12 months in the future in the airline industry. The data is disaggregated into different small and large PoS's, all with specific attributes such as cabins, booking channels, line groups and carriers. These attributes divide the forecasts into subgroups. Multiple time series methods are tested, for instance, HWESM, SARIMA and SARIMAX, as well as the DFM. Next, these models are combined to improve performance. Also, a distinction is made in the forecast horizon. A short-term model will forecast 1-4 months and a long-term model will forecast 5-12 months.

It is found that the forecast accuracy of the short-term model for the first 4 months is higher than when the long-term model is applied to the first 4 months. This shows the importance of using different techniques for the two forecast horizons. Moreover, the attributes that are used to disaggregate the data are PoS dependent. Large PoS's benefit from using the line group, booking channel and carrier attributes, whereas small PoS's only require the cabin class and carrier attributes. This indicates that a distinction between large and small PoS's contributes to the results. Also, for these small and large PoS's, the models with the best accuracy consist of different combinations of methods. For the short-term model, a combination between SARIMA and SARIMAX performs best for the large PoS's, whereas a combination of HWESM, SARIMA, and SARIMAX leads to better results for the small ones. Both these combinations outperform the naive model, which resembles the airline's current model the most. For the long-term forecast, one large and one small PoS showed the best results when HWESM and DFM were combined, whereas the other two showed the best results when the naive model was used. For these two PoS's, the combination of HWESM and DFM performed second best.

The airline is advised to put the combined forecasting models found in this research into use and thereby replacing their current naive model. Nonetheless, the long-term model will need some improvements as it does not always outperform the naive model.

For future research it is recommended to extend the forecast horizon of the short-term model as long as it outperforms the long-term model. Due to the period in which the research had to take place, this was not investigated in full, but because of some random tests with the extended forecast horizon, it is believed that the preferred 4-month forecast can be extended to 6 months. Moreover, it is advised to use more years of historical data to test the optimal value for the lookback for the adaptive model, even though some tests pointed out that using older data will not improve the results. The absence of more data limited this research in testing the models and the current optimal value of the lookback on years other than 2019.

References

Athanasopoulos, G., & Hyndman, R. J. (2008). Modelling and forecasting australian domestic tourism. *Tourism Management*, *29*(1), 19–31.

Bates, J. M., & Granger, C. W. (1969). The combination of forecasts. *Journal of the Operational Research Society*, *20*(4), 451–468.

Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: forecasting and control*. John Wiley & Sons.

Breitung, J., & Eickmeier, S. (2006). Dynamic factor models. *Allgemeines Statistisches Archiv*, *90*(1), 27–42.

Diebold, F. X. (2015). Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of diebold–mariano tests. *Journal of Business & Economic Statistics*, *33*(1), 1–1.

Faraway, J., & Chatfield, C. (1998). Time series forecasting with neural networks: a comparative study using the air line data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *47*(2), 231–250.

Gardner Jr, E. S. (2006). Exponential smoothing: The state of the art—part ii. *International journal of forecasting*, *22*(4), 637–666.

Geweke, J. (1977). The dynamic factor analysis of economic time series. *Latent variables in socio-economic models*.

Goh, C., & Law, R. (2002). Modeling and forecasting tourism demand for arrivals with stochastic nonstationary seasonality and intervention. *Tourism management*, *23*(5), 499–510.

Grushka-Cockayne, Y., & Jose, V. R. R. (2020). Combining prediction intervals in the m4 competition. *International Journal of Forecasting*, *36*(1), 178–185.

Harvey, A. C., & Shephard, N. (1993). 10 structural time series models.

Harvey, D., Leybourne, S., & Newbold, P. (1997). Testing the equality of prediction mean squared errors. *International Journal of forecasting*, *13*(2), 281–291.

Hyndman, R. J., Koehler, A. B., Ord, J. K., Snyder, R. D., et al. (2001). *Prediction intervals for exponential smoothing state space models* (Tech. Rep.). Monash University, Department of Econometrics and Business Statistics.

Hyndman, R. J., Koehler, A. B., Snyder, R. D., & Grose, S. (2002). A state space framework for automatic forecasting using exponential smoothing methods. *International Journal of forecasting*, *18*(3), 439–454.

Jackman, M., & Greenidge, K. (2010). Modelling and forecasting tourist flows to barbados using structural time series models. *Tourism and Hospitality Research*, *10*(1), 1–13.

Kalekar, P. S., et al. (2004). Time series forecasting using holt-winters exponential smoothing. *Kanwal Rekhi school of information Technology*, *4329008*(13), 1–13.

Lee, A. O. (1990). *Airline reservations forecasting: Probabilistic and statistical models of the booking process* (Tech. Rep.). Cambridge, Mass.: Flight Transportation Laboratory, Dept. of Aeronautics and ....

Lemke, C., Riedel, S., & Gabrys, B. (2013). Evolving forecast combination structures for airline revenue management. *Journal of Revenue and Pricing Management*, *12*(3), 221–234.

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., ... Winkler, R. (1982). The accuracy of extrapolation (time series) methods: Results of a forecasting competition. *Journal of forecasting*, *1*(2), 111–153.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). The m4 competition: Results, findings, conclusion and way forward. *International Journal of Forecasting*, *34*(4), 802–808.

Proietti, T. (2002). Forecasting with structural time series models. *A companion to economic forecasting*, *105132*.

Song, H., Li, G., Witt, S. F., & Athanasopoulos, G. (2011). Forecasting tourist arrivals using time-varying parameter structural time series models. *International Journal of Forecasting*, *27*(3), 855–869.

Stock, J. H., & Watson, M. (2011). Dynamic factor models. *Oxford Handbooks Online*.

Timmermann, A. (2006). Forecast combinations. *Handbook of economic forecasting*, *1*, 135–196.

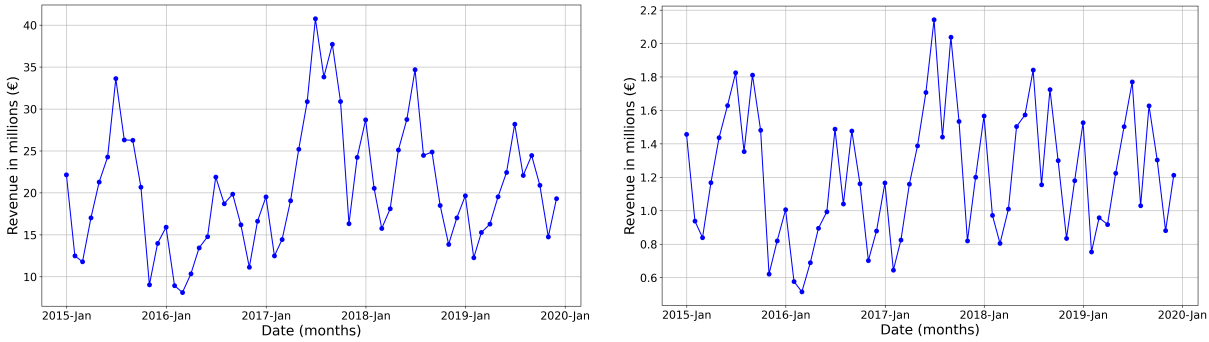
Weatherford, L. (2016). The history of forecasting models in revenue management. *Journal of Revenue and Pricing Management*, *15*(3-4), 212–221.

Weatherford, L. R., Gentry, T. W., & Wilamowski, B. (2003). Neural network forecasting for airlines: A comparative analysis. *Journal of Revenue and Pricing Management*, *1*(4), 319–331.

Appendix A

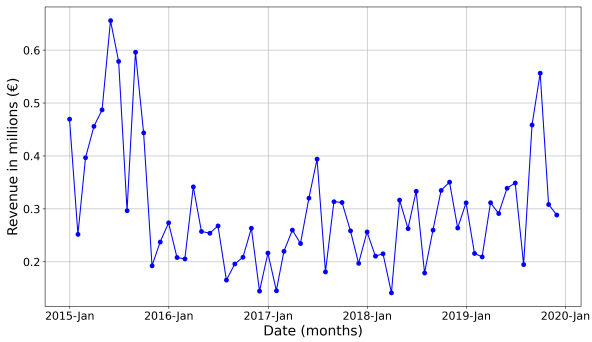
Appendix: Figures

A.1 Data Figures

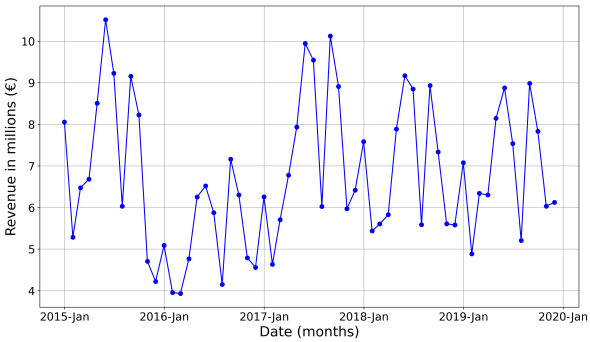


(a) Economy class.

(b) Premium class.

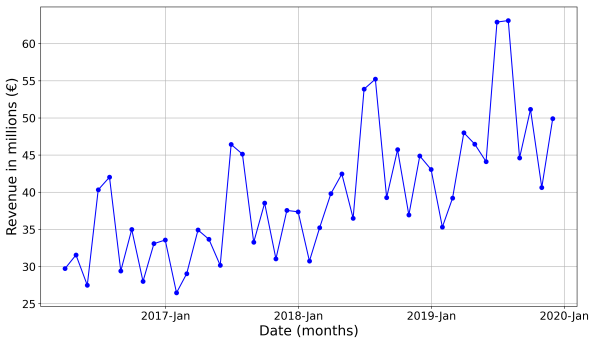


(c) First class.

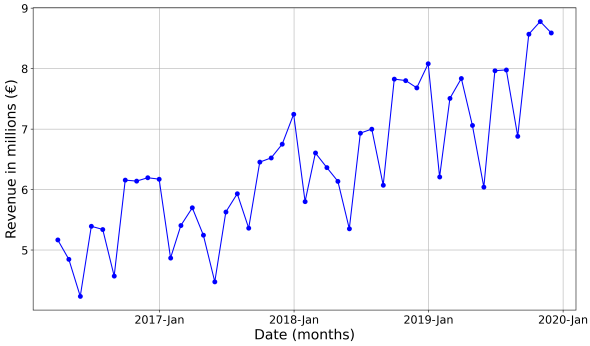


(d) Business class.

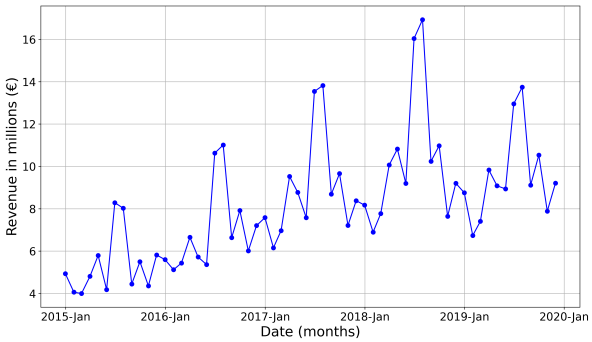
Figure A.1: Monthly revenues in millions of euros of the cabin classes of PoS Y from January 2015 - December 2019.



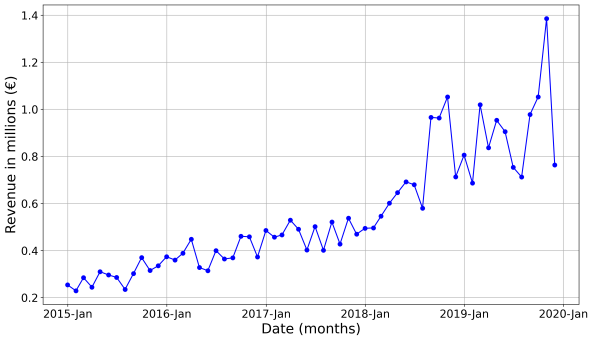
(a) Economy class and Direct Online channel.



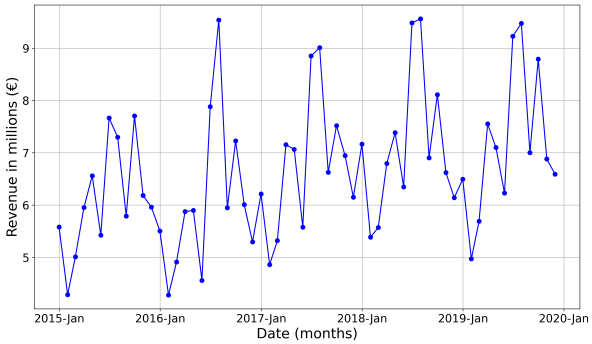
(b) Business class and Direct Online channel.



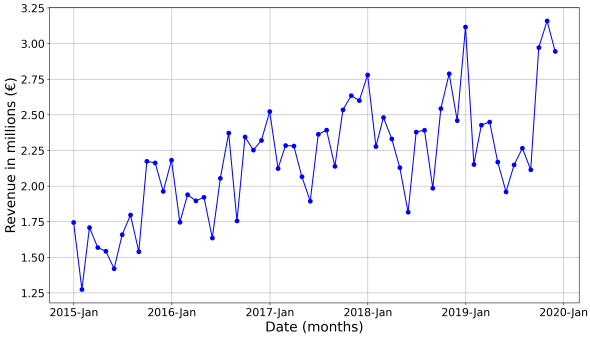
(c) Economy class and Indirect Online channel.



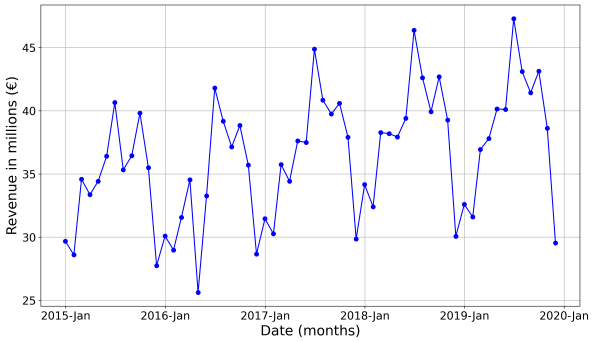
(d) Business class and Indirect Online channel.



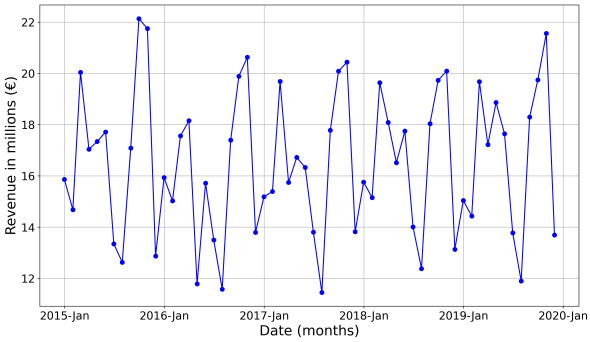
(e) Economy class and Direct Offline channel.



(f) Business class and Direct Offline channel.

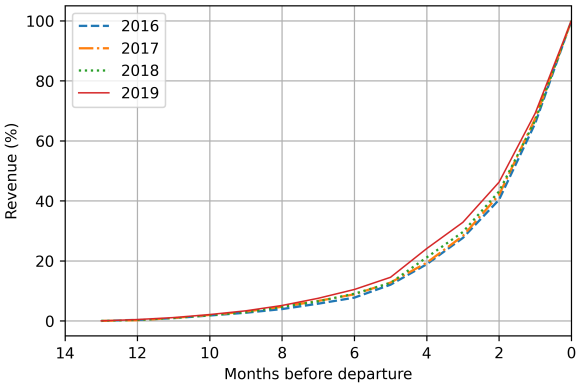


(g) Economy class and Indirect Offline channel.

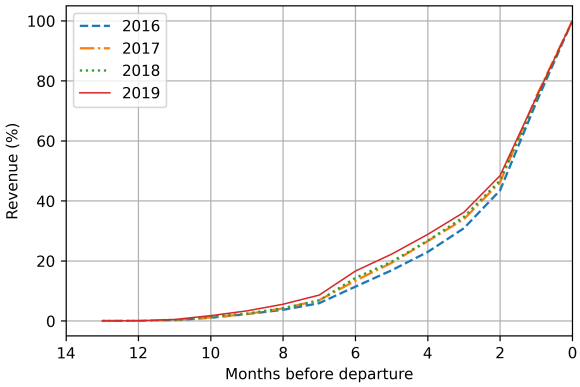


(h) Business class and Indirect Offline channel.

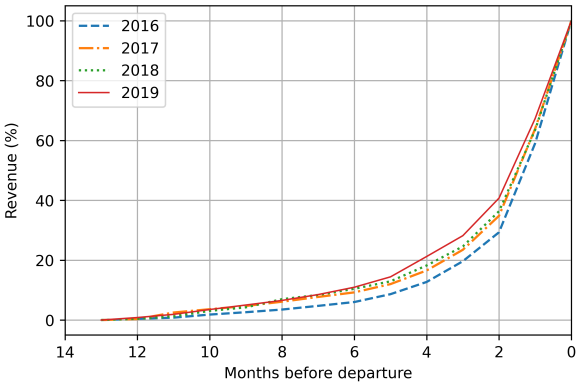
Figure A.2: Monthly revenue in millions of euros of PoS W when both the cabin class and the booking channel attributes are used to disaggregate the data, from January 2015 - December 2019.



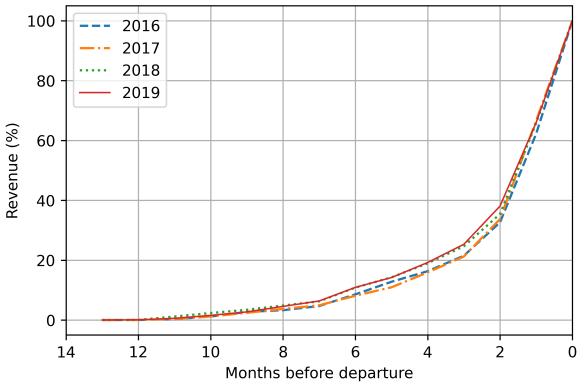
(a) Economy class revenue curves for flights departing in January.



(b) Economy class revenue curves for flights departing in July.



(c) Business class revenue curves for flights departing in January.



(d) Business class revenue curves for flights departing in July.

Figure A.3: Cabin class revenue curves of all the tickets bought in PoS W for the years 2016 - 2019. The cumulative percentage of the total earned revenue is given x months before departure.

Appendix B

Appendix: Results

B.1 Short-term models

B.1.1 Point of Sale X

Table B.1: Relative error values for the aggregated short-term models of PoS X. For each forecast horizon in the short-term model and for all (combinations of) models, the values of WMAPE and MSE are given relative to the time-independent errors of the naive model, which are 8.04 % and 124 $\times$ 10^{10} respectively. Also the total PI coverage probabilities are provided per model.

Model	**WMAPE**				**MSE**				**PI probability**
	1	**2**	**3**	**4**	**1**	**2**	**3**	**4**	
HWESM SP	0.66	1.00	1.17	1.29	1.29	2.14	2.22	2.37	0.476
HWESM G	0.98	1.39	1.57	1.73	2.59	3.31	3.65	3.94	0.756
SAR SP	0.52	0.77	0.80	0.90	0.55	0.85	0.70	0.79	0.726
SAR G	1.00	1.42	1.61	1.82	2.96	3.87	4.21	4.6	0.782
SARX SP	0.50	0.70	0.80	0.90	0.58	0.69	0.73	0.81	0.561
SARX G	0.99	1.41	1.60	1.82	2.92	3.8	4.15	4.56	0.783
DFM	1.18	1.25	1.29	1.31	1.16	1.24	1.29	1.33	0.477
All	0.59	0.87	0.98	1.09	0.77	1.19	1.29	1.41	0.809
All SP	0.51	0.76	0.85	0.94	0.52	0.96	0.99	1.06	0.683
All G	0.93	1.33	1.52	1.70	2.66	3.43	3.73	4.05	0.782
SAREXO	0.49	0.73	0.79	0.89	0.40	0.69	0.70	0.77	0.712
SP + SARX G	0.53	0.77	0.87	0.95	0.44	0.8	0.87	0.95	0.768
SP + SAR G	0.49	0.73	0.80	0.89	0.41	0.77	0.83	0.91	0.795

Table B.2: MAE values for the aggregated short-term model of PoS X. The error values are displayed for each forecast horizon in the short-term model and for all (combinations of) models. The errors of the naive model are time-independent.

Model	MAE (€1000)			
	1	**2**	**3**	**4**
Naive	628	628	628	628
HWESM SP	503	708	781	837
HWESM G	688	909	991	1,055
SAR SP	336	478	490	537
SAR G	731	961	1,051	1,148
SARX SP	343	452	489	534
SARX G	727	952	1,044	1,144
DFM	666	699	716	725
All	411	570	621	670
All SP	358	511	550	591
All G	678	896	990	1,072
SAREXO	316	460	485	530
SP + SARX G	351	497	535	572
SP + SAR G	333	478	511	549

B.1.2 Point of Sale Z

Table B.3: Relative error values for the aggregated short-term models of PoS Z. For each forecast horizon in the short-term model and for all (combinations of) models, the values of WMAPE and MSE are given relative to the time-independent errors of the naive model, which are 9.62 % and 5.90×10^{10} respectively. Also the total PI coverage probabilities are provided per model.

Model	WMAPE				MSE				PI probability
	1	**2**	**3**	**4**	**1**	**2**	**3**	**4**	
HWESM SP	0.70	1.43	1.70	1.66	0.47	2.02	2.88	2.85	0.620
HWESM G	0.99	1.67	2.05	2.23	0.83	2.64	4.24	5.19	0.901
SAR SP	0.61	0.94	1.07	1.14	0.34	0.73	0.93	1.24	0.776
SAR G	0.82	1.42	1.75	1.98	0.61	1.93	3.19	4.31	0.901
SARX SP	0.63	0.99	1.31	1.30	0.37	0.81	2.25	1.76	0.703
SARX G	0.81	1.34	1.67	1.90	0.58	1.78	2.95	4.07	0.896
DFM	1.52	1.47	1.47	1.45	1.92	1.78	1.76	1.73	0.474
All	0.55	0.92	1.10	1.12	0.27	0.73	1.17	1.20	0.927
All SP	0.58	0.95	1.15	1.11	0.29	0.75	1.17	1.03	0.886
All G	0.76	1.30	1.59	1.81	0.51	1.61	2.63	3.49	0.907
SAREXO	0.58	0.91	1.18	1.13	0.31	0.68	1.27	1.08	0.880
SP + SARX G	0.51	0.85	1.00	0.94	0.25	0.63	0.95	0.83	0.907
SP + SAR G	0.50	0.85	1.01	0.96	0.24	0.63	0.97	0.85	0.917

Table B.4: MAE values for the aggregated short-term model of PoS Z. The error values are displayed for each forecast horizon in the short-term model and for all (combinations of) models. The errors of the naive model are time-independent.

Model	MAE (€1000)			
	1	**2**	**3**	**4**
Naive	195	195	195	195
HWESM SP	130	264	314	307
HWESM G	183	310	380	413
SAR SP	114	175	198	211
SAR G	152	262	324	367
SARX SP	116	184	242	240
SARX G	150	248	309	352
DFM	282	272	272	269
All	101	170	203	207
All SP	107	176	213	205
All G	140	241	295	335
SAREXO	107	168	218	210
SP + SARX G	94	158	184	174
SP + SAR G	93	157	186	178

B.2 Long-term models

B.2.1 Point of Sale X

Table B.5: Relative WMAPE values for the aggregated long-term models of PoS X. Values are given relative to the time-independent naive model, with error 8.04 %. The error values are displayed for each forecast horizon in the long-term model and for all (combinations of) models.

Model	WMAPE							
	5	**6**	**7**	**8**	**9**	**10**	**11**	**12**
DFM	1.16	1.15	1.15	1.15	1.12	1.11	1.12	1.13
HWESM	1.10	1.15	1.16	1.18	1.20	1.21	1.21	1.18
SARIMA	1.21	1.29	1.36	1.46	1.60	1.80	1.98	2.19
All	0.96	0.98	0.99	1.00	1.04	1.10	1.18	1.23
HWESM-SAR	1.06	1.10	1.13	1.18	1.25	1.34	1.42	1.48
HWESM-DFM	0.95	0.96	0.97	0.95	0.96	0.95	1.00	1.02
DFM-SAR	1.01	1.03	1.05	1.09	1.13	1.22	1.32	1.44

Table B.6: Relative MSE values for the aggregated long-term models of PoS X. Values are given relative to the time-independent naive model, with error 124×10^{10}. The error values are displayed for each forecast horizon in the long-term model and for all (combinations of) models.

Model	MSE							
	5	**6**	**7**	**8**	**9**	**10**	**11**	**12**
DFM	1.41	1.44	1.48	1.50	1.40	1.37	1.37	1.50
HWESM	1.60	1.85	1.90	1.73	2.15	2.06	2.21	1.69
SARIMA	1.37	1.40	1.85	2.51	3.69	5.29	6.83	8.07
All	0.93	0.92	1.02	1.01	1.28	1.59	2.04	2.11
HWESM-SAR	1.20	1.26	1.40	1.51	2.16	2.73	3.31	3.37
HWESM-DFM	1.02	1.05	1.14	1.05	1.15	1.17	1.35	1.21
DFM-SAR	0.97	0.93	1.06	1.15	1.38	1.85	2.52	2.99

Table B.7: MAE values for the aggregated long-term models of PoS X. The error values are displayed for each forecast horizon in the long-term model and for all (combinations of) models. The errors of the naive model are time-independent.

Model	MAE (€1000)							
	5	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Naive	628	628	628	628	628	628	628	628
DFM	736	735	740	743	721	717	725	746
HWESM	700	722	730	732	751	757	779	720
SARIMA	743	781	827	911	1,043	1,209	1,368	1,537
All	591	594	601	621	653	722	783	819
HWESM-SAR	649	663	690	730	788	881	949	986
HWESM-DFM	595	592	602	597	612	610	665	662
DFM-SAR	627	631	643	677	724	798	891	971

Table B.8: Average PI coverage probability of the long-term models of PoS X.

	DFM	**HWESM**	**SAR**	**ALL**	**HWESM-SAR**	**HWESM-DFM**	**DFM-SAR**
PI probability	0.783	0.787	0.898	0.903	0.896	0.842	0.910

B.2.2 Point of Sale Z

Table B.9: Relative WMAPE values for the aggregated long-term models of PoS Z. Values are given relative to the time-independent naive model, with error 9.62 %. The error values are displayed for each forecast horizon in the long-term model and for all (combinations of) models.

Model	WMAPE							
	5	**6**	**7**	**8**	**9**	**10**	**11**	**12**
DFM	1.23	1.30	1.32	1.33	1.30	1.23	1.31	1.29
HWESM	1.59	1.77	1.98	2.13	2.23	1.85	1.58	1.29
SARIMA	1.42	1.56	1.85	2.14	2.33	2.47	2.80	2.96
All	1.08	1.15	1.28	1.49	1.52	1.47	1.54	1.39
HWESM-SAR	1.37	1.48	1.64	2.00	2.07	1.97	1.99	1.98
HWESM-DFM	1.03	1.10	1.18	1.35	1.41	1.22	1.24	1.14
DFM-SAR	1.07	1.16	1.30	1.4	1.47	1.54	1.71	1.57

Table B.10: Relative MSE values for the aggregated long-term models of PoS Z. Values are given relative to the time-independent naive model, with error 5.90×10^{10}. The error values are displayed for each forecast horizon in the long-term model and for all (combinations of) models.

Model	**MSE**							
	5	**6**	**7**	**8**	**9**	**10**	**11**	**12**
DFM	1.64	1.78	1.85	1.97	1.95	1.75	1.98	1.97
HWESM	4.69	5.27	6.19	6.98	7.00	4.78	3.12	2.15
SARIMA	2.95	3.10	4.97	7.05	7.93	8.00	10.12	11.66
All	1.53	1.61	2.22	3.08	3.24	2.51	2.95	2.83
HWESM-SAR	3.08	3.22	4.49	6.00	6.39	5.25	5.32	5.56
HWESM-DFM	1.42	1.47	1.71	2.27	2.32	1.53	1.63	1.34
DFM-SAR	1.25	1.49	2.03	2.66	2.8	2.47	3.51	3.42

Table B.11: MAE values for the aggregated long-term models of PoS Z. The error values are displayed for each forecast horizon in the long-term model and for all (combinations of) models. The errors of the naive model are time-independent.

Model	**MAE (€1000)**							
	5	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Naive	195	195	195	195	195	195	195	195
DFM	241	254	257	260	253	241	256	252
HWESM	311	345	387	416	435	361	308	252
SARIMA	277	304	360	418	456	482	546	577
All	210	225	250	290	298	287	301	272
HWESM-SAR	269	289	320	390	404	386	389	387
HWESM-DFM	202	214	231	263	275	238	242	223
DFM-SAR	210	227	253	274	287	301	333	307

Table B.12: Average PI coverage probability of the long-term models of PoS Z.

	DFM	**HWESM**	**SAR**	**ALL**	**HWESM-SAR**	**HWESM-DFM**	**DFM-SAR**
PI probability	0.904	0.930	0.997	0.987	0.979	0.948	1.00