NETWORK DEMAND FORECASTING WITH DATA SCIENCE

Lappeenranta–Lahti University of Technology LUT

Master's Programme in Business Analytics, Business Administration

Master's Thesis.

2022

Joonas Juvonen

Examiners: Associate Professor Jan Stoklasa, D.Sc. (Tech.)

Post-Doctoral Researcher Tomáš, Talášek, D.Sc. (Tech.)

ABSTRACT

Lappeenranta–Lahti University of Technology LUT

LUT School of Business and Management

Business Administration

Joonas Juvonen

Network demand forecasting with Data Science

Master's thesis

2022

133 pages, 42 figures, 14 tables and 12 appendices

Examiners: Associate Professor Jan Stoklasa, D.Sc. (Tech.) and Post-Doctoral Researcher Tomáš, Talášek D.Sc. (Tech.)

Keywords: Data Science, Econometrics Time series forecasting, Demand forecasting, Neural networks, Machine learning, Linear regression, Exponential smoothing, Integrated autoregressive moving average, Artificial intelligence

During 2020-2022 the aviation market has experienced the biggest demand disruption in its history. The global pandemic and traveling restrictions have seriously disrupted the seasonal and steadily growing aviation travel market.

The purpose of this thesis is to study the data science methods for forecasting networked aviation passenger demand under stable and disruptive conditions. Appropriate forecasting methods can improve profitability and sustainability of aviation operations in all conditions. During increased volatility, accurate demand forecasting can also enable dynamic responses to reduce losses and grasp new emerging business opportunities.

The research studied feasible forecasting methods for seasonal and non-stationary passenger demand. As a result, it was discovered that time series forecasting models such as exponential smoothing, autoregressive integrated moving average, neural networks, and linear regression models were suitable for this research. The performance of these methods changes significantly when compared between stable and disruptive conditions.

Under stable conditions, the forecasting methods that utilize the seasonality nature of the data have the best relative performance and were significantly more accurate than the standard seasonal naïve forecast that was used as a benchmark method. The forecasting performance of the linear regression and other explanatory variable models were proven to be dependent on the quality of the explanatory data and the relationship between the regression variables.

In disruptive conditions, the performance of all studied forecasting models deteriorates. However, the performance differences between the methods increase and quickly adapting models are more capable to operate under volatile conditions. This finding indicates that the different robustness characteristics of each forecasting method should be accounted when selecting appropriate model for time series forecasting.

TIIVISTELMÄ

Lappeenrannan–Lahden teknillinen yliopisto LUT

LUT-kauppakorkeakoulu

Kauppatieteet

Joonas Juvonen

Verkostokysynnän ennustaminen datatieteen keinoin

Kauppatieteiden pro gradu -tutkielma 2022

133 sivua, 42 kuvaa, 14 taulukkoa ja 12 liitettä

Tarkastajat: apulaisprofessori Jan Stoklasa, D.Sc. (Tech.) ja tutkijatohtori Tomáš, Talášek, D.Sc. (Tech.)

Avainsanat: Datatiede, Ekonometria, aikasarja ennustaminen, kysynnän ennustaminen, neuroverkot, koneoppiminen, lineaarinen regressio, eksponentiaalinen tasoitus, integroitu autoregressiivinen liukuva keskiarvo, tekoäly

Vuosina 2020–2022 ilmailutoimiala on kohdannut historiansa suurimman kysyntäshokin. Globaali pandemia ja matkustusrajoitukset ovat merkittävästi häirinneet normaalisti hyvin kausiluonteista ja vakaasti kasvavaa matkustajaliikennemarkkinaa.

Tämän opinnäytetyön tarkoituksena on tutkia matkustajaliikenteen kysynnän ennustamiseen tarkoitettuja datatieteen menetelmiä sekä stabiileissa että nopeasti muuttuvissa olosuhteissa. Tarkoituksenmukaiset ennustemenetelmät voivat parantaa lentoliikenteen kannattavuutta ja ympäristöystävällisyyttä kaikissa olosuhteissa. Tämän lisäksi nopeasti muuttuvissa olosuhteissa tarkka kysynnän ennustaminen mahdollistaa nopeat sopeutustoimet ja reagoinnin uusiin nouseviin liiketoimintamahdollisuuksiin.

Tutkimus perehtyi kausiluontaisen ja ajan myötä muuttuvan matkustajakysynnän ennustemenetelmiin. Tutkimuksen tuloksena huomattiin, että aikasarja ennustemenetelmät kuten eksponentiaalinen tasoitus, integroitu autoregressiivinen liikkuva keskiarvo, neuroverkko ja lineaarinen regressio olivat tarkoituksenmukaisia tässä tutkimuksessa. Näiden menetelmien ennustetarkkuus muuttuu merkittävästi siirryttäessä vakaasta toimintaympäristöstä nopeasti muuttuviin olosuhteisiin.

Vakaan kysynnän ympäristössä parhaan ennustetarkkuuden saavuttavat menetelmät, jotka hyödyntävät matkustajakysynnän kausiluontaista vaihtelua. Nämä menetelmät saavuttivat huomattavasti korkeamman tarkkuuden kuin tässä tutkimuksessa referenssimenetelmänä käytetty kausiluonteinen naiivi ennuste. Lineaarisen regressiomallin sekä muiden selittäviä muuttuja hyödyntävien mallien tapauksessa havaittiin, että ennustemallien suorituskyky riippui vahvasti muuttujadatan laadusta ja muuttujien välisestä suhteesta.

Nopeasti muuttuvissa olosuhteissa kaikkien tarkasteltujen ennustemallien suorituskyky heikkeni. Tutkimuksessa havaittiin, että mallien suorituskykyerot kasvavat ja osa malleista kykenee sopeutumaan paremmin muuttuvien olosuhteiden kysynnän ennustamiseen. Tämä havainto osoittaa, että ennustemallia valittaessa on syytä huomioida myös mallin mukautumiskyky ja luotettavuus nopeasti muuttuvissa olosuhteissa.

ACKNOWLEDGEMENTS

Foremost, I would like to express my sincere gratitude to my thesis advisor Associate Professor Jan Stoklasa for the continuous support for my econometric studies and during this research project. Your guidance is deeply appreciated.

Besides my academic advisor, I would like to thank my organizational mentor in data science for challenging me with new ideas and helping with the thesis process and direction. It has been wonderful to be able to use and develop more data skills and to have a meaningful project to work with. I am looking forward to future data science development projects.

Last but not least, I would like to thank my family and especially my wife Mari. I am blessed to have you on my side and your patience throughout the degree program is deeply appreciated and not forgotten.

“To stand up straight with your shoulders back is to accept the terrible responsibility of life, with eyes wide open. It means deciding to voluntarily transform the chaos of potential into the realities of habitable order. It means adopting the burden of self-conscious vulnerability and accepting the end of the unconscious paradise of childhood, where finitude and mortality are only dimly comprehended. It means willingly undertaking the sacrifices necessary to generate a productive and meaningful reality”

-Jordan B. Peterson

ABBREVIATIONS

ACF	Autocorrelation Function
ADF	Augmented Dickey-Fuller
AI	Artificial Intelligence
AIC	Akaike Information Criterion
ANFIS	Adaptive Neuro Fuzzy Inference System
ANN	Artificial Neural Network
AR	Autoregressive
ARCH	Autoregressive Conditional Heteroskedastic
ARIMA	Autoregressive Integrated Moving Average
ARMA	Autoregressive Moving Average
ARMSE	Adjusted Residual Mean Square Error
BLUE	Best Linear Unbiased Estimate
CLRM	Classical Linear Regression Model
CNN	Convolutional Neural Network
DBN	Deep Belief Network
DLNN	Deep Learning Neural Networks
GARCH	General Autoregressive Conditional Heteroskedastic
GRU	Gated Recurrent Unit
KPSS	Kwiatkowski-Phillips-Schmidt-Shin
LTSM	Long-term Short-term Memory
MA	Moving Average
MAD	Mean Absolute Deviation

MAE	Mean Forecast Error
MAPE	Mean Absolute Percentage Error
MLE	Maximum Likelihood Estimation
MLP	Multilayer Perceptron
MSE	Mean Squared Error
OLS	Ordinal Least Squares
PACF	Partial Autocorrelation Function
PP	Phillips-Perron
RNN	Recurring Neural Network
SARIMA	Seasonal Autoregressive Moving Average
SVM	Support Vector Machine
SVR	Support Vector Regression
TCN	Temporal Convolutional Network
VAR	Vector Autoregressive
VARMA	Vector Autoregressive Moving Average
2SLS	Two-Stage Least Squares Regression
3SLS	Tree-Stage Least Squares Regression

Table of contents

Appendices

1. Introduction

The aviation business is changing at an unprecedented pace due to the global pandemic (Abu-Rayash and Dincer 2020). In this change, the prediction of future demand has critical importance when aiming to exist and thrive in the future marketplace. This thesis aims to research and develop forecasting methods for predicting network demand for a European commercial airline operating in an intercontinental market. The airline needs to adapt for the current disruptive market by developing its network forecasting capabilities for both stable and disruptive conditions. To achieve this target, we are using 150 million rows of historical passenger demand data aggregated into 132 data points to understand driving factors and the underlying macroeconomic trends within the aviation competitive landscape. Using this data, we are evaluating different econometric, and machine learning-focused forecasting methods based on their feasibility and accuracy to predict network demand in both stable and disruptive environments.

From a business perspective forecasting future network demand in changing conditions has a critical importance for an airline since it enables profitable daily operations (UK, Aviation Demand Forecasting 2013; UN Statistics 2021). Regulation, airspace restrictions, and comparative travel route options all have a significant effect on the aviation business which makes it highly volatile in terms of route-specific demand and pricing (Alves and Caetano 2016, 22,26). The aviation supply and demand equation act as a system since airline capacity allocations have an interdependence relationship between competitors and affordable pricing acts as a stimulant, which accelerates consumers' leisure travel demand even for unexpected route combinations. (Boonekamp and Riddiough 2016, 3,4,7,23)

As a field, aviation is generating high volume and high-frequency data which creates favourable research setting for forecasting research. Even though aviation is a potential research field, the academic publications regarding global demand forecasting are limited due to the competitive landscape of the aviation industry. Prior studies in aviation suggest that case-specific iterations of time series and machine learning models can be used in aviation demand forecasting (Kanavos, Kounelis, Iliadis, and Makris 2021; Wang and Gao 2020). These studies are in most cases focusing on regional or national forecasting and therefore we have selected a study perspective including also demand forecasting research articles from supply chain and artificial intelligence research. The research available in these domains will reinforce the forecasting method literature review for this thesis. AI and supply chain demand forecasting research can also provide new perspectives since they have more active research regarding sophisticated models like fuzzy systems, multivariable regression models, and deep learning neural networks (Syntetos et al. 2016). The usability, limitations, and selection of these models in aviation demand forecasting are discussed in chapters 2 and 3. The shortage in public research regarding global aviation demand forecasting combined with good data availability and strong business case, makes network demand forecasting a prominent research subject.

In this research, we are using past customer flight purchases from origin-destination city selection including thousands of cities and this selection forms the core network for the airline. The core network includes both current and potential business opportunities for the airline and by generating an accurate forecast of future passenger demand in this network, the airline can plan flight frequencies, schedules, and equipment selections to capture that demand with the intention to generate a financially positive outcome.

The purpose of this thesis is to study aviation demand forecasting with econometrics and data science methods. Forecast method performance is evaluated in terms of forecasting accuracy. The main research questions of the thesis are described below.

1. **What demand forecasting methods are commonly used in airline network demand forecasting, and can we use these techniques in this research?**
2. **Can these popular forecasting methods outperform naïve forecasting in terms of accuracy under stable conditions with a 3-year forecasting horizon?**
3. **How do these popular forecast methods perform when they are used under disruptive operating conditions with 1-year forecasting horizon?**

The forecasting accuracy is measured with mean absolute percentage error (MAPE) and mean squared error (MSE) techniques utilizing in-sample/out-sample dataset splitting. The current forecast with an accuracy of 2.48% is based on the seasonal naïve forecasting method, which will be used as a benchmark accuracy in this research. Seasonal naïve forecasting uses the previous seasonal values of time series as forecasts when generating predictions.

MAPE is used as a primary accuracy measurement since in our research the forecasting error has a symmetrical effect in terms of business impact. Both under- and overestimating the future demand have an equally significant financial impact on the network demand planning. Overestimating demand leads to empty planes and financial loss, whereas underestimating demand results in lost revenue base and increased fixed costs per flight since the aircraft have a continuing lease and service agreements that translate into constant negative cash flow. MAPE as a measurement is used since its easily comparable and understandable by the business and is usable across the methods with different scaling and normalization. When possible, MSE is a good comparative measurement with MAPE since it is not sensitive for zero nominal values like MAPE, and it is a comparable measurement within a model family.

The main time horizon of this research is with 2011-2019 data with stable operating conditions. Data from 2011-2016 is used as model training data and 2017-2019 as a forecast testing data with research question's 3-year forecasting horizon. From a long-term business perspective, the current health-related travel restrictions are expected to be temporary and therefore, the research focuses on forecasting under normal operating conditions.

Disruptive conditions caused by the global pandemic during 2020-2021, are used as a robustness study perspective of the chosen forecasting techniques. The focus of the last research question is to evaluate forecasting methods during "Force majeure" conditions to discover, which forecasting models and methods are most robust under changing conditions.

Based on this research we can see that seasonality capturing models have strong forecasting capabilities in normal conditions. In disruptive conditions, quickly adapting models have relatively accurate forecasting capabilities among the researched methods, although the disruptions decrease the forecasting performance of all models on an absolute scale.

The scope of this research is focusing on utilizing the available data and determining the best possible forecasting methods based on this data. This also includes a comparative study regarding forecasting method accuracy with limited data availability. However, the scope of the research excludes the exploration and evaluation of new data sources which might have additional forecasting potential. This decision is made since the data available has been identified as significant for the target company by previous research projects. Currently, usable data sources are paid subscription services where a global service provider is gathering and enhancing data for the aviation industry. In addition to external sources, we will also use intra-company sources to fit the global data into the business context of the company.

The thesis structure composes of 6 sections. After the introductory part, we will deep dive into the literature review. The purpose of the review is to gain a holistic picture of the academic research related to demand forecasting. At first, we will study cardinal concepts regarding econometrics and data science forecasting. Then we will focus on the available demand forecasting research in aviation. Finally, we will conclude the review with comparative demand forecasting method research in related fields of supply chain and artificial intelligence. This will deepen our knowledge base of regarding demand forecasting and narrow down the usable method candidates for this research.

After a related research review, we will continue by building up the theoretical framework of usable forecasting methods. In this methodology part, we are going to define the forecasting performance measurement principles, forecast and data settings, and conduct academic research regarding the strengths and weaknesses of selected methods. We will begin by studying exponential smoothing-based methods. Autoregressive integrated moving average models are the second model approach after exponential smoothing. As a third approach, we will study neural networks in forecasting. Finally, the methodology is concluded by a linear regression model study.

The fourth part of the thesis is focusing on the implementation of these frameworks into practice by utilizing datasets to build and test forecasting models. This chapter will begin with an explanatory data analysis to understand the nature of our demand as a dependent variable. After analysing the demand, we will continue by testing forecasting methods in both stable and disruptive conditions, which provides us with quantitative performance data. Methods are also tested with limited data samples in both conditions to understand how the prediction changes if we have less data to use with the respectable methods.

In the fifth part of the thesis, we will evaluate the forecasting models and the quantitative results. We will evaluate how selected models are able to successfully predict demand in stable conditions and how the forecasting accuracy changes when we have a significant disruption in demand. This section will include both a mathematical evaluation of methods and a comparative discussion regarding method recommendations. In the final chapter, we will answer our research questions and compare research results with previous research.

Finally, we will end the thesis with the research limitations discussion and future research recommendations.

I want to conclude this introduction by talking about personal aspirations regarding the thesis topic. I have had a growing inspiration for analytics and data science over the past 5 years. During this time, I have worked in large organizations and seen how significantly a successful forecasting can affect into organization’s behaviour, resourcing, and performance. Still, it’s often the case that the forecasting potential has not been fully utilized or acknowledged by the organization. Although the methods offered by the econometric and data science domain are almost countless and the level of sophistication can be overwhelming, I firmly believe that this research can also provide effective findings and recommendations for relevant and cost-effective forecasting of everyday business problems.

2. Literature Review

In this chapter, we will study the background of variable relationship modelling and discuss demand forecasting from different perspectives. We will begin by defining the econometrics and data science as a research field since it enables us to build the forecasting models. After this, we will explore previous research specifically regarding network demand forecasting in aviation. This perspective is followed by previous research regarding demand forecasting supply chains. In the last part of the review, we will briefly cover the general trends and most used demand forecasting methods in artificial intelligence research.

2.1. Data Science and Econometrics

“The goal of data science is to improve decision making by basing decisions on insights extracted from large data sets” (Kelleher and Tierney 2018). Although this definition is not perfect, it captures the essence of data science by underlining the main target of problem-solving in the bigger picture of data science as defining trait. It can also be said that the boundaries of data science are context relative and difficult to define precisely (Provost and Fawcett 2013). As a definition, Provost and Fawcett also highlight the data science’s goal to enable and automate data-driven decision-making for organizations.

When defining a research field, a historical review can act as a valuable perspective. The field of data science has developed quickly over the last 20 years, but the foundational roots of data science are in mathematics, computer science, and statistics developed in the early and mid-twentieth century. Accelerating growth of computational power and memory coupled with the development of statistical and probabilistic modelling has been the enabler of advanced analytics and data science in the last two decades (Charles, Aparicio and Zhu 2020, 2,3,215, 221; Stahlbock et al. 2019, 3-4). First computer-based machine learning and time series forecasting methods were developed in the nineteen-fifties and sixties, but the current trend of exponential information growth rate and computing power has accelerated the growth of machine learning in the past 20 years. Machine learning focuses on providing algorithms that can automatically analyse large data sets to extract valuable patterns otherwise hardly understandable to human comprehension (Kelleher and Tierney 2018; Donoho 2017, 745-763).

As a definition, data science combines statistics, machine learning, data analysis, and computer science with domain-specific knowledge and usage case like business settings and aims to extract valuable information from data to enhance decision making (Kelleher and Tierney 2018; Stahlbock et al. 2019, 3-17; Donoho 2017). It can also be highlighted that the application of data science is not domain-specific, but it can be applied to all business and academic fields which are generating data (Longbing 2017; Mariani and Zenga 2021,2-4). From medical to aviation to revenue optimization and supply chain development, data-based decision-making has potential in all these domains. A visual presentation of data science's key expertise areas can be seen in figure 1.

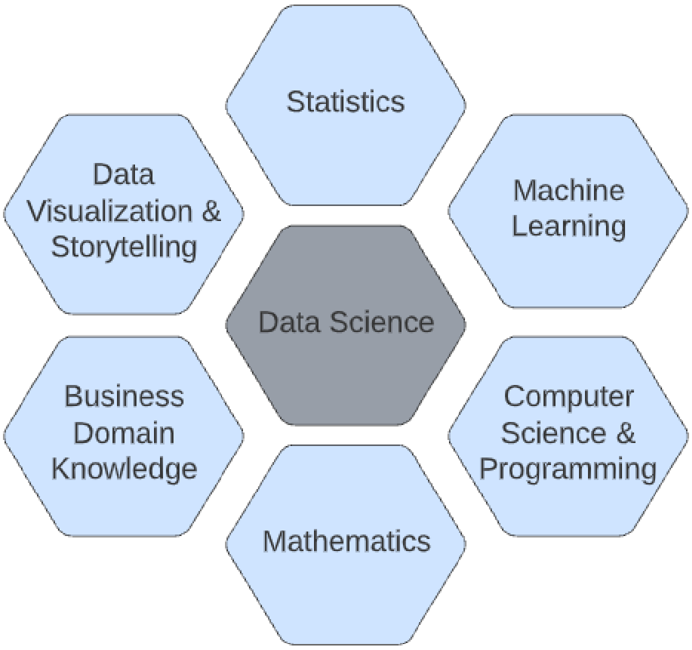

Statistics
Data Visualization & Storytelling
Machine Learning
Data Science
Business Domain Knowledge
Computer Science & Programming
Mathematics

Figure 1. *Data science key competence areas (Donoho 2017).*

Since this thesis includes a data science research project focusing on forecasting models, it is insightful to go through the data scientist's workflow which can be seen below in figure 2. It starts with the problem definition by the business or academic field. A significant part of the workflow is focused on data cleaning, extraction, and transformations required to the data into a readable format (Longbing 2017). After data cleaning an explanatory data analysis is conducted to understand the key parameters, statistics, and possibilities of the data. After explanatory data analysis, data can be enriched and modified using feature engineering.

In feature engineering data is modified into a format that enables the utilization of specific modelling techniques. In practice, this can be for example transforming data from a daily format into a weekly format and highlighting repeating seasonal shifts in data which can be captured by the modelling technique. In our research, these techniques can be simple as naïve forecasts or more complex like neural network algorithm fitting. Methods are often derived from machine learning, econometrics, or mathematics and they enable the insight extraction from data. Finally, the insights derived from the data need to be visualized or formed into an easily understandable data story. (Charles et al. 2020 1-5; Kelleher and Tierney 2018)

Data visualization and storytelling ensure that business stakeholders can understand the key insights and make the right decisions based on the algorithms and models developed throughout the process. It also needs to be noted that each step of the process can be iterative. This means that when during the workflow some parameters like data coverage or engineered features are proving to be insufficient for the problem solving, then these steps need to be revisited and enhanced to continue with the process (*Donoho 2017).*

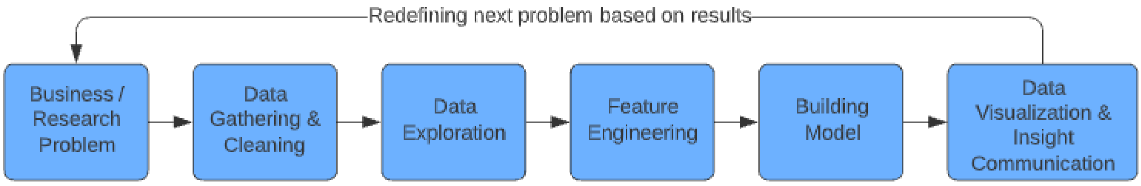

Redefining next problem based on results
Business / Research Problem
Data Gathering & Cleaning
Data Exploration
Feature Engineering
Building Model
Data Visualization & Insight Communication

Figure 2. *Data science workflow is iterative process from problem definition to insight communication.*

Econometrics can be defined as statistical and mathematical analysis and measurement of economic relationships (Brooks 2014, 2-6). Modelling of these relationships often serves as a base for economic forecasting (Geweke, Horowits and Pesaran 2008; 1-3, Tinbergen 2005, 1-7). Tinbergen also points out that econometrics can be viewed as mathematical economics working with measured data aiming to solve business or academic problems. It can also be said that econometrics methods are used to obtain the values of parameters which are essentially the coefficients of the mathematical form of the economic relationships (Satchell 2003,397-398; Shalabh 2016, 1-3).

Foundational statistical analysis in the early and mid-nineteen hundred has had a significant impact on economic measurement (Tinbergen 2005, 1-7). One of the culmination points for econometrics was the development of integrated autoregressive moving average (ARIMA) models by Box and Jenkins in the 1970s. (Alagad and Egrioglu 2012, 3-8; Stellwagen and Tashman 2013). Although economics has had a consistent development curve since the seventeenth century, computational advancements in the late nineteen hundred have also been a huge enabler for econometric measurement (Geweke et al 2008, 2-10). The accelerated computational development from the 1980s has enabled the utilization of computationally heavy and data-rich models such as neural networks (Alagad and Egrioglu 2012, 3-8).

Econometrics model development has its focuses on the theoretical formulation based on economic theory and model evaluation based on statistics and hypothesis testing (Brooks 2014, 1-6, Geweke 2010, 1-8). Theoretical problems can often include optimization related tasks in which iterative loops are highly effective when evaluating the validity of created solution (Fan, Pastorello, and Renault 2014) Continuous development loop in econometric research can be used for single problems like building a forecasting model, but it has also a larger meaning of iterating research results and re-evaluating validity of old theories and assumptions.

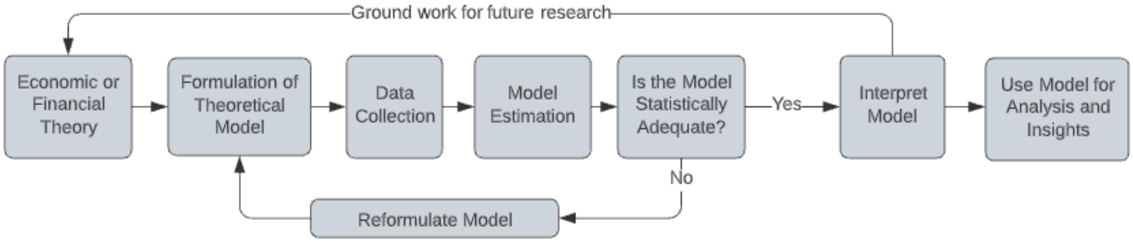

Ground work for future research
Economic or Financial Theory
Formulation of Theoretical Model
Data Collection
Model Estimation
Is the Model Statistically Adequate?
Yes
No
Interpret Model
Use Model for Analysis and Insights
Reformulate Model

Figure 3. *Econometric model development focuses on the evaluation of driven theory and hypothesis of that theoretical framework (Brooks 2014, 4-6)*

When comparing econometrics and data science we see a significant overlap among them. Both fields are utilizing statistics and mathematics principles with modern data-heavy information technology to model parameter relationships and behaviour. These models are simplified representations of more complex reality (Satchell 2003, 397-398). Historically, econometrics has focused more on variable relationship modelling within economics research. Outside of the economics research field, econometrics is replaced with the umbrella term data science, which shares many techniques and practices used in econometrics research. We could also say that the data science term has become associated with a wider scope of research subjects and is more heavily focused on software engineering principles and machine learning engineering with subjects like deep learning or recommendation engines (Longbing 2017).

The popularity and excitement associated with both data science and econometrics have increased significantly during the last decade. (Sousa et al. 2021, 14523–14531). As a conclusion, we can say that current data science as a research field already includes econometrics which uses similar methods but with a highlighted economic research focus. The utilization and development of these methods are increasing across different industries (Sousa et al. 2021, 14524; Longbing 2017). This is a positive contributing factor since it accelerates the data science method development and problem-solving capabilities across the international research community.

2.2. Demand Forecasting in Aviation

An aviation network can be defined as multi airplane operations where airplanes transport passengers from several origin cities to their travel destinations through one or more transfer hub airports. The network term can be viewed from either passenger or from airline perspectives which are the supply and demand parties of the aviation market. From a passenger perspective transfer airport hubs and individual flights are generating a networked route that enables a dynamic selection of different destinations by combining several flights into a single travel plan. Passenger’s origin and flight destination city combination is called a city pair.

From an airline perspective, the supply side of the aviation market is more complex. When an airline operates as a network it has typically several fleets of airplanes operating with different traveling ranges both intracontinental and intercontinental levels. A typical traveling pattern of the aviation network is that passenger combines two or more flights which can be either shorter regional flights or longer intercontinental flights. This combination creates a network that constantly gathers passengers from network cities and brings them to the transfer airport and from there the passengers are redistributed to their second flight which connects them to their destination. When an airline operates from tens to hundreds of different routes, it can offer thousands of city pair combinations for passengers. When a new route is added to this network, it increases the demand for all other city pairs since this enables multiple different city pair combinations. So, there is an interdependency between supply and the demand of the network where a new supplied route by one airline affects demand for that airline and for other airlines operating with overlapping networks. This is called network-based demand and it also has a dynamic effect on airline ticket pricing, since acceptable pricing is significantly affected by the traveling route availability and alternative offerings by competitor airlines. (Suh and Ryerson 2019)

Aviation network passenger demand is affected by multiple major components such as regulation, pricing, and seasonality (Boonekamp and Riddiough 2016, 3,4,7,23). It is also affected by regional micro and macro conditions such as holiday season weather or travel restrictions (Ryerson and Suh 2019). These changing conditions make the forecasting of future demand both business-critical and complex. Since market conditions change dynamically with changing market participants, it is important from an airline's perspective to conduct continuous research and measurement of emerging travel trends. This ensures that the top management of an airline can follow and predict changes in passenger flows which translate into profitable business when operated over the airline's complete route network.

Previous studies regarding aviation network demand forecasting can be divided into two subgroups: local market focusing on national or regional routes and global market forecasting (Wang and Gao 2020; Suh and Ryerson 2019, 1-7). These markets can have partially different demand drivers at both microscopic and macroscopic levels. According to Suh and Ryerson, public data availability in a highly competitive and regulated environment can be a limiting factor for transparent academic discussion regarding accurate forecasting methods for aviation network demand.

The concern highlighted by Suh and Ryerson is verified also by Kanavos et al. (2021, 16329-16342) who have studied the application of deep learning models for demand forecasting in aviation. Kanavos et al. have utilized both seasonal and non-seasonal integrated autoregressive moving average models (SARIMA & ARIMA) and deep learning neural networks (DLNN) in their study. In their research, they have also implemented an advanced forecasting model by combining regression and neural network models to forecast aviation demand. According to Kanavos et al. (2021, 16330) the utilization of advanced econometric and data science methods such as the implemented combination model, are partly unexploited research subject, especially in global aviation demand forecasting.

A data science literature review conducted by Wang and Gao (2020) partially challenges this observation developed by Kanavos et al. by underlying that there is more forecasting research available on an airport, national, and regional levels. The review includes the aviation demand studies published between 2010-2020 and it discusses the most common methods used for aviation demand forecasting. In this review seasonal naïve, exponential smoothing, ARIMA, SARIMA, and regression model variations are the popular choices. When we compare the review this with Kanavos et al.'s research, they align to point out that there is significant potential for improvement in the utilization of data-based forecasting techniques especially for global aviation market demand forecasting, since most of the public research is focused to airport, regional or national aviation travel demand perspectives.

When considering the forecasting accuracy in aviation, Ryerson and Suh (2019) point out that forecast makers have incentive bias where positive forecasts are likely to benefit the forecast creators financially, whereas negative predictions are less popularly perceived by upper management of business organizations. Forecasting models operating with limited data seem to be especially prone to this bias if we look at the local market forecasting of regional airlines or airports, which are channelling passenger flows from larger hubs (Wadud 2011, 59-62). The lack of global market data or analysing capabilities in local operations and significant dependency on single passenger routes are likely to affect local demand forecasting. Local demand forecasting can also be affected by the locally changing correlation relationships if global explanatory variable-based regression models are used demand forecasting. In these cases, simple time series forecasting methods can achieve higher consistency and accuracy than explanatory variable models (Sivrikaya 2013, 70-78).

On the other hand, local forecasting can operate in a more consistent and less competed market which can also enable better forecasting conditions and improved understanding of local customer demand drivers (Solvoll, Mathisen and Morten 2020, 1-3,7). This observation is also supported by Li (2019) who highlights in his research that with highly consistent data, regression models can also outperform ARIMA based time series forecasting models.

In terms of passenger purchasing behaviour, local markets seem to be relatively less similar themselves when compared with intercontinental markets but more homogenous regionally (Benoit, Pascal, and Julien 2011; TRP Aviation Demand Forecasting 2002). These homogenous markets are likely to perform more consistent over the long term and recover more quickly from the disruptions since the demand is less stimulated by leisure travel and more based on family relations and local business when compared with intercontinental markets (Li and Trani 2014; Solvoll et al. 2020, 1-3). In practice, this could mean that the forecasting of local aviation markets could differ from global markets in terms of forecasting consistency, driving factors, and recovery under disruptions. Although similar data are used to forecast local and global demand, it can be expected that anomalies like local incidents or annual holidays might project differently towards demand with the respectable markets. This differentiative finding supports the need for a global aviation demand forecasting research.

A significant part of aviation demand forecasting and its research is conducted by commercial market operators like Amadeus and Statistical Analysis Systems. These solutions are based on fundamental statistics and econometrics modelling and they also include machine learning methods to generate more accurate forecasting. Companies that have specialized in aviation demand forecasting also have extensive amounts of data gathered globally from collaborative relationships with dozens of airlines. For example, Amadeus has originally built its business around aviation ticket reservation systems. Its holistic overview of the aviation market has enabled it to utilize one of the best data sources in the industry to build its forecasting product services. Even Though these products claim to offer unparallel information advantages, their exact methods and principles are not publicly available and therefore open to academic review. (Winter 2021; SAS 2012)

In their research Kanavos et al. (2021) point out that short-term operational forecasting in aviation demand can often be effectively conducted by quickly adapting advanced data-intensive methods like deep learning neural networks since these models can better adapt to the spiking variation levels in daily operations. If advanced neural network’s structure is not used, then the research highlights 2 secondary observations. Firstly, the univariate time series models can often outperform multivariate explanatory models in a presence of increased demand volatility. Secondly, in long-term time series modelling simpler variance decomposition techniques like support vector machines (SVM) can outperform more complex models. (Kanavos et al. 2021)

Ghobbar and Friend (2003) support this observation in their research which is focused on the prediction of aviation spare part usage based on flight volumes and passenger demand. They also point out that different forms of exponential smoothing models are appropriate for different demand forecasting environments. Their research also evaluates that linear or seasonal regression models can reach good performance when forecasting demand in the aviation business. However, it must be pointed out that aviation spare part demand forecasting differs from passenger demand forecasting although they have partially the same factors affecting demand. In spare part demand forecasting, neural network base solutions are also successfully implemented by Sahin, Kizilaslan, and Demirel (2013). In this solution, both the feedforward-based artificial neural network (ANN) method and recurrent neural network (RNN) are used. Based on this research, both ANN and RNN outperform exponential smoothing and a simpler ANN network can outperform more sophisticated RNN network in cases where quick adaptability is a key feature.

Aviation demand can be seasonal by nature and this seasonality can be utilized in forecasting. This observation was pointed out by Shuojiang, Chan, and Zhang (2019, 169-178) who also points out that utilization of seasonal models like SARIMA and hybrid models like SARIMA-SVM can be effective. The seasonality importance observation is backed up by studies conducted by Shih-Yao, Shiau, and Chang (2010) and Xiao et al (2014) who have successfully generated promising regional aviation forecasting results with neural network-based and adaptive-network-based fuzzy inference system (ANFIS).

In their study, Shuojiang et al. (2019) also find out that when we exclude the network aviation demand forecasting as a research field, forecasting methods in general are quite active and well-researched field, especially in global supply chains and in machine learning research. When we evaluate this observation with the previous findings covered in this chapter, we can conclude that there is a need to extend the literature review and study also general forecasting methods research outside the field of aviation. This enables us to compare the methods commonly used in aviation with an alternative more general perspectives like supply chain demand forecasting and general Artificial intelligence-based demand forecasting. These alternative perspectives can propose potential methods which are more popular in other demand forecasting research but less used in aviation.

Based on the conducted literature review, the most common demand forecasting methods in the global aviation industry are regression models, neural networks, exponential smoothing, and ARIMA/SARIMA. The performance order of these models seems to be highly case specific. These results can be seen in table 1 in chapter 2.4. In this summary table we are comparing the usage frequency of each forecasting method across the different study perspectives reviewed in this literature review.

2.3. Demand Forecasting in Supply Chains

Supply chains demand forecasting has been a target of interest continuously over the last decades. Traditionally supply chains are described by the value stream model seen in figure 4. The produced goods and services flow from the manufacturing or service provider to the customers through the retailers. This flow is reversed when considering demand information and financial transactions. Although traditionally supply chains are focusing on material flow, they can also include immaterial products like software, know-how, and services.

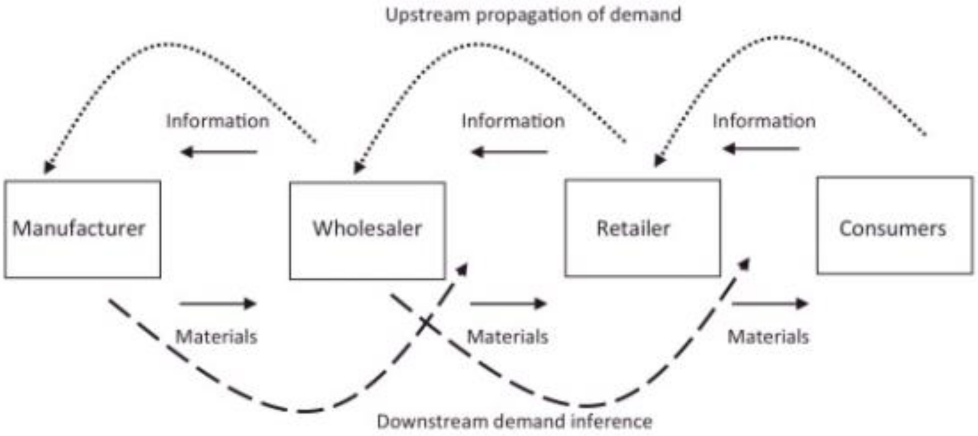


Figure 4. *Supply chain structure. The manufacturing part of the chain can include multiple subcontractors and dozens of steps increasing the actual process complexity (Syntetos et al. 2016,3).*

The impact of global supply chains on a modern economy is remarkable. Global manufacturing supply chains alone are responsible for 16% of the planet’s GDP generation. Therefore, the incentives and possibilities for improving forecasting methods regarding these operations seem plentiful (UN Statistics, 2021). Supply chain forecasting has also been one of the general discussion topics during 2020 and 2021 due to the shutdowns of electronic and high computing chip manufacturing in early 2020 during the first months of the global pandemic (Cooney 2022; Leary 2021).

A major miscalculation in demand forecasting made by large manufacturing corporations like car and electronic manufacturers generated a rippling effect into the supply chains when manufacturers cancelled their production backlogs when the global pandemic first started in Q1/2020. This rippling effect was multiplied by the Bullwhip-Forrester effect in which each link in the supply chain extrapolates the demand into a given change direction causing significant forecasting error when the original demand signal is compared with the supply chain generated independent estimate. This caused most supply chain semiconductor manufacturers to reduce production numbers significantly. Unlike forecasted by the corporations, the demand peaked due to changing consumer needs, and this shift in the demand-supply equation balance has caused a global semiconductor shortage (Kleinhans and Hess 2021, 2-13).

The research regarding supply chain demand forecasting is readily available with a decades-long track record, and the number of new publications emerging into academic portals suggests, that there is an active and dedicated research community working to solve prediction-based problems in supply chains (Perera et al. 2019). Forecasting in supply chains shares many mathematical and statistics-based principles also seen in aviation demand forecasting. Just like in aviation, also in supply chains demand is stimulated by price and there are interconnections between similar products and overlapping markets. In both cases data can also be reviewed and analysed as time series data, therefore enabling the utilization of univariate time series methods (Zhang and Zhao 2010, 463-465). However, there are also differences between the aviation network and supply chain network demand forecasting.

One special characteristic of the supply chain is the Bullwhip-Forrester effect which was described above. Forecasting error generated by the Bullwhip-Forrester effect is amplified by the length of the supply chain and the low frequency of communication and order placements between supply chain stakeholders (Syntetos et al. 2016, 1-6). The impact of the Bullwhip-Forrester effect can be reduced by the direct information loops within the supply chain and by the implementation of time series forecasting models such as ARIMA (Syntetos et al. 2016, 1-6). These models are effective to reduce demand signal amplification since autoregressive processes can utilize previous lagged values for future prediction and account for non-stationarity if present. Autoregressive integrated moving average methods can also be effective to forecast supply chain demand in periodically changing environments due to methods combined with stochastic and deterministic capability (Yanxin, Sujian, and Yongfang 2018, 6630-6631; Zakrytnoy 2021).

Another differentiating point to consider when evaluating supply chain forecasting possibilities seems to be the dimensional nature of the supply chain. According to the research conducted by Syntetos et al. (2016) supply chains include product, time, and channel dimensions. The product dimension includes the perspectives of inventory management, distribution, and warehousing. Channel dimension covers the need for multichannel supply in the case of multiple store locations both offline and online. In both dimensions, we also have multiple different planning time horizons which are affecting to the requirements of the forecast. In practice, this perspective means that upper-level demand forecast does not translate automatically as good planning across different detail levels included in supply chains. Therefore, this should be considered when evaluating distinct supply chain forecasting method potential in aviation network demand forecasting.

When reviewing commonly used methods for supply chain forecasting, exponential smoothing has a consistent track record. Exponential smoothing can achieve relatively good results, especially when joined with some seasonality accounting technique (Ferbar et al. 2009). These techniques can be a combination of data science-related methods like clustering or neural networks. They can also include more traditional techniques like triple exponential smoothing that adds seasonality and trend components to the forecasting model (Paldino 2021, 3-6).

In addition to exponential smoothing and neural networks, also ARMA models are used successfully with various iterations. These autoregressive moving average models can include stationarity, or they can be non-stationary. They can also have seasonal components included like in SARIMA, which can improve prediction capability with seasonal demand cycles (Pongdatu and Putra 2018; Dekker, Donselaar and Ouwehand 2004, 1-5). If there are appropriate multiple time series sources available also vector autoregressive models like VAR and VARMA can be utilized in supply chain forecasting (Sadeghi 2015, 44-47).

There is also a significant amount of forecasting solutions that are business case specific. These solutions can include machine learning techniques like K-means clustering with support vector machines to identify customer group-specific demand patterns to be utilized in forecasting (Lu, Chang, and Huang 2014). Also, classical regression models are commonly used in supply chain forecasting if there is explanatory variable data available. Regression models with one or multiple explanatory variables can be effectively utilized by modern statistical software where demand planner operators can utilize automated testing and model fitting. (Boone et al. 2019)

Less traditional demand forecasting techniques utilized in supply chains include for example deep learning neural networks. These methods can utilize vast amounts of data and are made to fit complex models. They can also be complemented by the utilization of cross-sectional hierarchical forecasting in which base forecasts are generated at a chosen level of the hierarchy, and after that, base demand forecasts are aggregated or disaggregated using algebra manipulations to fill the other levels of hierarchy (Punia, Singh and Madaan 2020, 1-3). Cross-sectional hierarchical forecasts can be combined with temporal hierarchy frameworks in which we can separate trend and seasonality components from the data to improve forecasting accuracy (Athanasopoulos, et al. 2017, 1-5).

In addition to simple neural networks also recurrent networks can be utilized for supply chain forecasting according to Carbonneau, Laframboise, and Vahidov (2008, 1140-1144,1153). Traditional neural networks typically operate with feedforward or back-propagation principles. In these networks, the individual elements are organized into layers in a way that output signals from the first neuron layer are passed as input for the next neuron layer if there are multiple hidden layers of neurons. In a back-propagating network, we calibrate the layer weights based on the following layer results when iterating the model accuracy. These networks are mainly operating in a single direction layer-by-layer principle. In recurrent networks, we are partially using outputs from the neuron layers as a recurring input for the same or previous neuron layer to enhance the network iterating capability when generating forecasts (Carbonneau et al. 2008, 1141-1144).

In a summary, it can be said, that in supply chains traditional forecasting techniques seem to be highly similar when compared with aviation demand forecasting. Classical time series techniques like exponential smoothing and ARIMA are highly common. These models are often enhanced with seasonal components to improve the forecasting capability. In addition to more traditional methods, also neural networks, support vector machines, and clustering techniques are used in demand forecasting. It is worth noticing, that there is an abundance of research papers available when it comes to supply chain forecasting and many forecasting methods are specific for supply chain operations. Constant research in supply chains ensures that when there are technological and operational changes in supply chain management, also the forecasting will adapt to accommodate these changes to form new research and capabilities around the supply chain ecosystem.

2.4. Forecasting in AI research

The focus of this chapter is to review demand forecasting methods outside aviation and supply chains and to draw a conclusion regarding potential forecasting methods for this research. Research around demand forecasting is constantly developing and some new methods might achieve high prediction performance with our research problem. Even though all methods reviewed in this chapter are not selected for model testing, this study might still provide valuable insights and ideas for future research or towards similar research problems.

When machines learn from the data and from the feedback that they receive and can mimic human decision-making by executing tasks similarly, this can be called as artificial intelligence (AI) (Wei, Bhardwaj, and Wei 2018, 7-8). Unlike science fiction movies, in practice this means mathematical and statistical modelling of variable relationships with modern computational power. Artificial intelligence is present in relationship modelling when the system can analyse incoming data and measure its current model capabilities to make decisions. Based on this measurement, make changes to the current model which improves the accuracy of decision-making models. This iterating feedback loop which enables learning from previous values is essential for artificial intelligence model building (Raza and Khosravi 2015).

When reviewing demand forecasting outside of aviation and previously discussed supply chains, the first methods emerging from academic journals and discussion papers are the same time series methods discussed previously. Although neural networks were commonly used for aviation forecasting, they also have several iterations which are less common in aviation forecasting. Adaptive neuro-fuzzy inference system (ANFIS) is one of these methods and it has been used successfully in hospital demand forecasting. AFNIS uses fuzzy inference system learning algorithms and may provide more accurate results than traditional neural network (Jebbor, Chiheb, and Abdellatif 2022).

Neural networks are a specialized machine learning research field that focuses on mimicking human brain neuron layer activities by learning hierarchical structures and levels of abstraction to understand and utilize data patterns (Wei et al. 2018, 7-12). Neural networks include several different sub-fields which have their special characteristics. Convolutional neural networks (CNN), deep belief networks (DBN), and long-short term memory (LTSM) are commonly used in forecasting (Sezer, Gudelek and Murat, 2020,1-7; Wei et al. 2018, 7-12). When neural networks have multiple hidden layers, they are called deep learning. In deep learning higher-level abstractions are defined as the composition of lower-level abstractions. The nonlinear feature transformations with multiple possible states within the hierarchical levels are the reason why the method is called deep learning. The biggest advantage of deep learning neural networks is the ability to learn feature representations in multiple levels of abstraction (Jatin and Durga 2019, 1313-1316). In practice, deep learning networks can adapt complex functions without human-crafted feature engineering. (Wei et al. 2018, 8-19)

In addition to deep learning neural networks, Jatin and Durga (2019, 1313-1316) discuss the use of several interesting methods in demand forecasting. Random forest classification can support demand forecasting models with predictive classification components. In stepwise regression, we fit the regression model with a selection of independent variables and iterate the model based on its explanatory forecasting power. Multilayer perceptron (MLP) and gated recurrent unit (GRU) are iterative versions of artificial neural networks that have proved significant potential in demand forecasting according to Jatin and Durga (2019, 1313-1315).

Fuzzy methods are a research field for precisely precenting unprecise information. In practice, this means that boundaries between variable states are not crisp and for example forecasted weather can be partly sunny and cloudy simultaneously unlike in binary logic where a variable can have only one defined state and a crisp boundary between states. According to Zakrytnoy (2021, 63-63,68), fuzzy time series forecasting proposes significant potential for time series forecasting. This logic first presented by Zadeh in 1965 enables time series forecasting with relative readability, simplicity, and scalability. In Fuzzy time series forecasting dependent variable range is divided into fuzzy sets and through these sets, we extract data behavioural patterns. These patterns can tell us how values change over time when moving from one fuzzy set to another, and this acts as a baseline enabler for fuzzy model training. (Reuter and Möller 2010, 363-373)

When comparing these techniques with the previous chapters and their conclusions we can select the methods for this research. Based on the similarity of research perspectives, the implementation frequency of forecasting methods in the reviewed articles, and the feasibility of the method application we have chosen the following methods seen in table below. Table 1 summarizes the usage quantity found from reviewed literature and the selected method variations. Complete reference table and listed search words and research portals for the literature review can be seen from Appendix 1.

Table 1. *In total, 4 model families and 20 variations were selected for more detail research and testing since they appeared commonly either in aviation demand forecasting or in general forecasting research. *After more detail research ARMA and CNN models were not feasible for this research and therefore we have 18 successful forecasting variations.*

Model Name	Aviation forecasting articles using method	Supply chain forecasting articles using method	AI research forecasting articles using method	Selected variations	Total count (success)
Exponential Smoothing	7	10	2	Simple, Double, Triple	7
ARMA/ ARIMA/ SARIMA	5	9	4	Seasonal, Non-seasonal	3 (2)*
Neural Networks (ANN, RNN, CNN)	8	7	6	Feedforward, Recurrent	7 (6)*
Linear Regression & Explanatory var model	13	6	4	Simple exp. Var model	3
Excluded forecasting models					
SARIMA-SVM, SVM, SVR	3	3	4	-	-
Regression variations (3SLS, 2SLS, Log. reg.)	6	1	2	-	-
ANFIS, Fuzzy system	2	1	4	-	-
Neural Networks based on MLP	2	0	3	-	-
Neural Networks based on DBN	0	0	1	-	-
GARCH, ARCH	2	0	1	-	-
VAR, VARMA	2	2	0	-	-

As seen from the table 1, we have selected 4 model families: exponential smoothing, ARIMA/SARIMA, neural networks, and linear regression based explanatory variable models. Based on the conducted literature review, these traditional time series methods are commonly used both in aviation demand forecasting and in other research fields as well. Our decision to limit the forecasting method quantity and focus on more popular methods aims to reduce the computational requirements and the implementation costs in this research.

3. Methodology

In this part of the thesis, we are going to discuss the settings, principles, and methods regarding the domain of demand forecasting methods and their performance measurement. We are starting with the forecasting performance evaluation framework for the applicable methods. After the evaluation framework, we are defining the data and forecasting settings, and benchmarking forecasting method. As the main part of the methodology chapter, we are researching the selected demand forecasting methods that were identified to be popular in the previous chapter.

3.1. Forecasting Performance Evaluation

Since this thesis focuses on the evaluation of forecasting methods for aviation network demand, we need a numeric and established way to evaluate and compare the performance of each forecasting method. For this purpose, we will use mean squared error (MSE) and mean absolute percentage error (MAPE). MSE measures forecasting accuracy in absolute terms and is not affected by forecasts error's front sign due to its squared formulation. MAPE is a relative percentage measure. It's more understandable from the reader's perspective and it works with normalized data scales unlike the MSE, but it does not work if the evaluated data point would be 0 since it's used as a divider in the formula. As seen in formulas 1 and 2, the forecasted parameter value $\hat{y_t}$ is deducted from the actual data point value y_t therefore producing the evaluated error term. In equations, N is the number of observations and as a divider, it is dividing the summed total squared error into a single datapoint relative mean error either in percentage terms or in a nominal scale (Brooks 2014, 283-287; Chicco, Warrens and Jurman 2021)

$$MSE = \frac{1}{N}\sum_{t=1}^{N}(y_t - \widehat{y_t})^2 \qquad (1)$$

$$MAPE = \frac{100}{N}\sum_{t=1}^{N}\left|\frac{y_t - \widehat{y_t}}{y_t}\right| \qquad (2)$$

In this research numeric evaluation is critical since we are testing different forecasting methods and their approaches to solving the problem differ substantially. Numerical evaluation is a concrete tool to compare methods within an easily understandable framework. In the method comparison MAPE is the primary measurement since it has easily understandable business context and it’s not effected by possible data scale normalizations.

3.2. Research Settings

The research settings in the thesis determine which methods can be used to answer the research questions. These settings include data and forecasting settings for this case study and general forecasting principles. Both data and forecast settings are adding limitations for the suitable demand forecasting methods. They are also a significant determining factor from the thesis result perspective. From general forecasting literature we can also highlight general principles for demand forecasting.

3.2.1. Data settings

Data quality is one of the most significant factors affecting forecasting accuracy (Paldino et al. 2021, 1-4). Data accuracy, coverage, and completeness combined with appropriate explanatory variables can cover a significant part of the variation seen in the dependent variable outcomes (Moran, Nono, Rherrad 2019).

The applied forecasting method significantly affects the dataset’s acceptable sampling and accuracy requirements. Different methods have a different number of estimated parameters. When parameter number increases so does the minimum sample size to define relationships between model parameters. Another important factor affecting forecast reliability is the variance level in the data (Moran, Nono, Rherrad 2019; Hyndman and Kostenko 2007). When noise or stochastic variance increases, the minimum sufficient sample size increases as well. The sufficient number of observations is also dependable on the accuracy or confidence value that we wish to have with our analysis. The basic rule for observation size is, that it will always need to exceed the number of parameters in the model (Hyndman and Kostenko 2007, 13).

Since the base idea of a statistical model is to describe how the data changes over time, a longer time series usually enables improved capture of seasonal and other patterns if these patterns are consistent over the time horizon (Verma and Verma 2020, 2-6). In our research, we have data points from the beginning of 2011 until the end of 2021 constituting 132 data points in total. For the primary forecasting evaluation, we utilize the stable 2011-2019 demand data including 108 observations and for the disruptive evaluation we will use the final 24 observations.

Our data is in a monthly format and includes several annual cycles that needs to be captured by the model. In practice, this means, that we need a minimum of 12 data points to go over the annual cycle. In addition to this, we need to capture the trend developing between each annual cycle which would require additional points as comparison values. The final value-adding to the minimum sampling size is the model parameter count. For the exponential smoothing methods, there are several smoothing parameters and components (5). This would bring the theoretical minimum observation size into the range of 17-20 samples.

For ARIMA (p,d,q) models there is the same requirement to capture seasonality and possible non-stationarity. This would bring the minimum sample size into the same range with exponential smoothing. Similar formulation repeats in the regression model where we will need to model the relationship between explanatory variables and dependent variables. When comparing the size range of these numbers with our 108 and 132 observations we can be reasonably sure that we have sufficient data size if the data variation is low, and we are not missing values from the dataset. With all methods selected from the literature review, our number of observations is larger than 10 times multiplied number of parameters. According to Verma and Verma (2020), this can be considered a generally safe region in model fitting since we have sufficient degrees of freedom in the model.

Dataset building for the thesis will follow the workflow seen in figure 5 below. First, the data is extracted from global data subscription services. These subscription services can provide a holistic overview picture of the aviation industry as a whole and cover data sources like financial planning and airplane manufacturing databases. In an essence, data-focused information portals can provide reliable and high-quality data sources with access also to reflecting factors that are indirectly affecting or describing the current state of the aviation market.

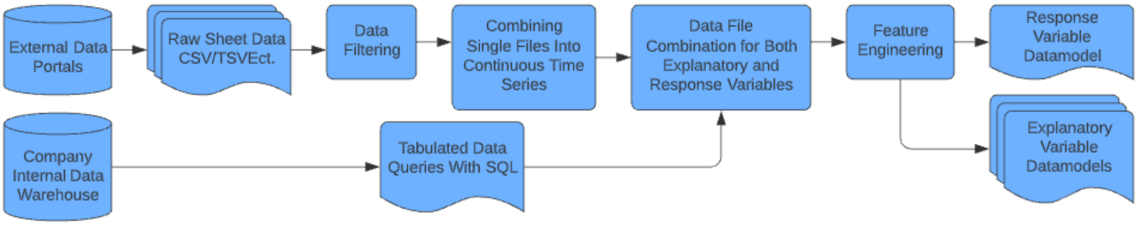

External Data Portals
Raw Sheet Data CSV/TSVEct.
Data Filtering
Combining Single Files Into Continuous Time Series
Data File Combination for Both Explanatory and Response Variables
Feature Engineering
Response Variable Datamodel
Company Internal Data Warehouse
Tabulated Data Queries With SQL
Explanatory Variable Datamodels

Figure 5. *Generation of dataset starts from individual databases and concludes as cleaned time series data files for both dependent and explanatory variables. These steps are a more detailed look from overall model development seen in chapter 2.1 figures 2 and 3.*

Raw data is collected from these portals and then combined with suitable internal company databases which enrich the dataset by bringing the data into the right context. Combining data from multiple sources requires joins executed in column tabulated databases and adding of features like calculated columns which are a necessity for data modelling. This feature engineering and data exploration discussed in data science workflow can also be described with explanatory data analysis term which is a classical statistics process step (Flachaire and Nuñez 2007; Brooks 2014, 5-6). Calculating data value means, quantiles and distributions and plotting data and its aggregations form the basis for engineering data into a format that captures its special nature and key characteristics.

This research is focusing on demand forecasting of a globally operating airline and therefore the dependent variable is sold passenger tickets for a specific route selection in a monthly level. The route in this context means a city pair connection between two cities like London and Sydney. Since the data is extracted from both paid subscription services and intra company databases we must be using an indexed and relative data format for visualization and evaluation in this research. This is a key action to ensure that we retain the data confidentiality in this research and still can enable transparent comparative evaluation of forecasting methods.

3.2.2. Forecast settings

As discussed in the research questions and in the introduction of this thesis, the primary forecasting horizon for this research is a fixed 3-year window from January 2017 to December 2019. In all cases, the forecasting model accuracy is tested with an in-sample & out-sample method where we split the dataset into training and testing sets. In stable conditions, the training set includes 72 data points from 2011 to 2016 and is used to fit the model. After fitting the model, we can predict the last 36 data points until the end of 2019 and measure their accuracy by comparing forecasted values with actual data values. If the method is not able to generate a forecast for the 2017 to 2019 period, it is not suitable for solving our research question and therefore is not evaluated in this research.

The forecast horizon is fixed in this research since 2017-2019 is used as the main reference period in company reporting as it was the latest stable operating period in aviation. Using the same forecasting period increases the value of this research for the sponsoring company since it enables research result comparison with financial forecasts and intra-company metrics. As discussed in data settings, using the 2017-2019 forecasting period also creates a setting where we utilize all the available data for forecasting method evaluation.

The same 3-year period is also the length of the company's strategic planning horizon. 3-year forecast window is used as a long-term planning horizon for aviation network planning that includes fleet and capacity planning. In practice, the implementation of our research results in aviation network demand forecasting requires the long-term capacity adjustments made in this 3-year planning and therefore it is essential from the network adaptability perspective. In other words, large-scale implementation of seasonal network changes like longer routes or more frequent operations requires the forecasting for the upcoming 3 seasons since significant capacity adjustments in the fleet, engineering, and flight crews take to years implement.

The secondary perspective for forecasting methods is to test the methods in disruptive conditions during 2020-2021. In these disruptive conditions, a fixed 1-year forecasting horizon is used since we have disruptive data only from 2 seasons and the first is used as a training season for the models. The second reason to use 1-year forecasting horizon in disruptive conditions is, that in a highly volatile environment the company network planning is focused on the upcoming and constantly changing operating season and the company's survival, and therefore the longer-term forecasting is not relevant in disruptions.

Addition to normal and disruptive conditions we will also test the forecasting model performance under limited data. In this setting we will keep original forecasting windows but reduce the available training data for the models. For the stable conditions we will reduce the training from 2011-2016 to 2014-2016. For disruptive cases we will reduce the training from 2011-2021 to 2017-2021. The main idea of this test is to understand how the models operate with less training data and how the training data period selection effects into the research results.

3.2.3. General Forecasting Principles

In this thesis research, we will use the naïve forecasting method as a benchmark forecasting technique to evaluate the forecasting method's viability. This is due to its easily understandable structure and current usage in company forecasting. In this research naïve forecast uses the last 12 measured actual data points in time T to forecast values for the next 12 data points until timepoint T+12. (Paldino et al. 2021, 3). This means that since data is gathered in a monthly tabulated format and we are predicting next season's values, we use the current target year's January data to forecast next year's January results.

Based on the classical econometric definition, forecasting models are either time series forecasting models or econometric forecasting models. Econometric forecasting uses explanatory variables to predict future values, whereas time series forecasting models predict the future using only information contained in the past values or the current and past values of the error term. In this research first 3 model families are pure time series models, and the final model family utilizes the explanatory variable to generate demand forecasts. (Brooks 2014, 245-247,274-277).

According to Brooks (2014), all forecasting models are essentially extrapolative. Forecasting models in general are prone to break around turning points or structural changes or regime shifts. It is also likely that forecasting accuracy declines with an expanding forecast horizon. Brooks (2014, 246-247,277-287) also highlights in his book, that most effective forecasting models are built on solid theoretical foundations and statistical testing and are supplemented by expert judgment and interpretation.

3.3. Exponential Smoothing

Exponential smoothing is a classical time series forecasting technique where recent observations have exponentially declining weight into the forecasted value when they slide into the past when time moves forward. Exponential smoothing does not use explanatory variables to predict future values. In equation 3 seen below F_{t+1} is the desired next forecasted value, F_t latest forecasted value, Y_t is the current actual value and α is the smoothing constant which varies in a range from 0 to 1 (Brooks 2014, 274-277; Paldino 2021, 2-6).

$$F_{t+1} = \alpha Y_t + (1 - \alpha)F_t \tag{3}$$

This so-called single smoothing can be optimized by iterating smoothing coefficient α in a way that the forecast in-sample achieves minimum error measured with MSE or similar forecast accuracy measure. The same iteration process towards minimum forecasting error can also be done with the more complex double and triple exponential smoothing methods and their respectable smoothing constants. The weakness of simple smoothing is that it does not adapt into including data trends or seasonality which is present in our dataset (Paldino 2021, 2-8). To remedy this, we can include a trend factor into the model, this method is called Holt’s method or double exponential smoothing.

The first component of Holt’s method is estimating the level of the data. It utilizes the same exponential smoothing constant alpha with the current actual value. It also includes the previous timestep level and trend components (Koushik and Ravindran 2016, 101-105). The second equation, trend includes a beta component, which is an exponential smoothing parameter for the trend which also varies in a range from 0 to 1. The *k* component seen within the bracket in equation 6 is the number of forecasts into the future which often has a value of 1 in a simple forecasting setting. The trend equation itself is evaluating the change between current and last level values and adds this with the previous trend value.

When implementing Holt’s method into practice we don’t often have the previous level and trend component levels to make the first forecast. Therefore, we start from the second data point and for this first step $L_2 = Y_2$, $T_2 = Y_2 - Y_1$ and $F_3 = L_2 + T_2$. After these initials setups which apply also to triple exponential smoothing, we can start systematically applying equations to generate a forecast.

$$L_t = \alpha Y_t + (1 - \alpha)(L_{t-1} + T_{t-1}) \tag{4}$$

$$T_t = \beta(L_t - L_{t-1}) + (1 - \beta)T_{t-1} \tag{5}$$

$$F_{t+1} = L_t + T_t(k) \tag{6}$$

Including a trend component in the model makes it a two-component model. It's important to understand that the components need to be calculated simultaneously since they have an interdependence equation relationship for each timestep. Holt's method can be further advanced to include seasonality. This method is called triple exponential smoothing or Holt-Winter's method (Koehler, Snyder, and Ord 2001, 269-272).

Holt-Winter's method includes trend, level, and seasonal components to maximize the capability of exponential smoothing methods. The added seasonal component can track annual variance in the data and improve the forecast based on this repeating pattern (Paldino et al. 2021). The model consists of 4 equations seen below. When comparing these two previous methods: Holt's method and simple smoothing, it is obvious that Holt-Winter's is building on top of these simpler methods (Koushik and Ravindran 2016, 101-105).

Both the trend and the seasonal components in exponential smoothing can be either additive or multiplicative. This means that there are 2 iterations with Holt's method and 4 with Holt-Winters. In this methodology part, we cover the additive approach for time series data where the series is the sum of its components. With multiplicative data, the time series is the product of its components, and this approach can be checked in Appendix 2 which goes over a similar multiplicative method. (Koehler et al. 2001)

$$L_t = \alpha(Y_t - S_{t-m}) + (1 - \alpha)(L_{t-1} + T_{t-1}) \tag{7}$$

$$T_t = \beta(L_t - L_{t-1}) + (1 - \beta)T_{t-1} \tag{8}$$

$$S_t = \gamma(Y_t - L_{t-1} - T_{t-1}) + (1 - \gamma)S_{t-m} \tag{9}$$

$$F_{t+1} = (L_t + T_t)S_{t-m(k+1)} \tag{10}$$

As seen above, adding the seasonal component to the model creates some differences from previous double smoothing. The Holt-Winter's additive method has additional smoothing parameter gamma which varies from 0 to 1. This parameter is seen in equation 9 and acts as a similar multiplier term to alpha and beta. In the new seasonal equation, the current actual value is deducted from previous trend and seasonal component values and then it is multiplied with gamma. This outcome is added to the gamma multiplied previous seasonal component. In the equation 10, k is the number of forecast periods into the future, and m is the number of seasons. (Weiheng et al. 2020)

3.4. ARMA & ARIMA & SARIMA

In addition to exponential smoothing, there are also other time series models. Based on the literature research autoregressive moving average process (ARMA) and its seasonal and integrated variants (ARIMA, SARIMA) are often used to generate highly accurate forecasting. These models are a combination of several pattern recognition algorithms. The first part of the model is autoregressive (AR) where the model values depend linearly on its previous values and the error terms of those values. Equation 11 below is describing the AR(p) process and in it, the current value is an additive sum of the previous values of the univariate series.

$$Y_t = \mu + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \phi_p Y_{t-p} + u_t \qquad (11)$$

In this equation, ϕ are the AR process coefficients, Y_t are the actual values in time *t*, μ is the constant, u_t is the error term and *p* describes the order of the AR process.

Another part of the ARMA and ARIMA models is the MA part. The moving average part of the model is a linear combination of previous error terms. The equation for the MA(q) process can be seen below. In this model, we are tracking the previous error terms u_t, ϴ are the coefficients for the MA (q) orders and μ is the constant.

$$Y_t = \mu + \Theta_1 u_{t-1} + \Theta_2 u_{t-2} + \Theta_p u_{t-p} + u_t \qquad (12)$$

When we combine the stochastic MA(q) part deterministic AR(p) part we will get ARMA (p,q) model which captures both stochastic and deterministic patterns from the data. For non-stationary processes, we can use the integrated model ARIMA (p,d,q). Non-stationarity means that there is a trend or drift in data that alters the mean and variance over time. So, this means that ARMA (p,q) is used for forecasting stationary processes whereas ARIMA (p,d,q) is suited for non-stationary processes. In essence ARIMA(p,d,q) is a linear regression model on previous values p, previous errors q, and differenced d times to make series stationary. (Brooks 2014, 246-269)

If the data also includes a seasonal cycle, this can be captured by the SARIMA model. SARIMA model includes the (p,d,q) order features present in the previous ARIMA model but it also adds seasonal order components (P,D,Q,M) to the overall model. In this model M is the seasonal cycle length. In overall this means that the SARIMA (p,d,q)(P,D,Q,M) model has a significant number of orders to be iterated and evaluated during the model estimation. On the other hand, it should have pattern capture capability to fit even more complex datasets including patterns otherwise undetectable. (Będowska-Sójka 2017, 92-96)

3.4.1. Box-Jenkins Approach

Box-Jenkins's approach is a systemic approach for optimizing ARMA model usage. The process has 3 steps: identification, estimation, and model validation through diagnostics. The first step of the model identification includes the dataset stationarity testing as a preparation step. Stationarity testing through tests like Augmented Dickey-Fuller (ADF) is conducted for all series. If testing shows non-stationarity data is transformed to make it stationary. This process is covered in more detail in chapter 3.4.2 below.

When the dataset is stationary, we can focus on the determination of model order and type by utilizing graphical or information criteria-based evaluation. This is the second step of Box-Jenkins's approach. By testing different models and measuring information criteria and autocorrelation function values, we can evaluate model fit for the data. The aim is to iterate towards a model which captures all the patterns from the data resulting a stable autocorrelation function plot without spiking values and the lowest possible information criteria value. The idea is also to utilize seasonality if was detected during stationarity testing since it can improve model fit in ARIMA models. (Brooks 2014, 269-272)

As a part of model estimation, we can evaluate optimal model structure by viewing autocorrelation function (ACF) and partial autocorrelation function (PACF) plots. If ACF declines gradually and PACF drops close to zero after a lag of zero, we have an autoregressive model structure. If PACF declines gradually and ACF drops close to zero after a lag of zero, the appropriate model is the moving average model. If both graphs decline exponentially without any spikes after lag 0, the data seems not to include patterns capturable by ARIMA model structures. (Brooks 2014, 262-264)

The numerical method to evaluate model fit and pattern capture is the so-called information criteria. The smaller the information criteria values are, the more its capturing dataset's patterns and variance. Information criteria evaluates the number of parameters included in the model and the amount of variance of which it tries to minimize. Information criteria is a relative measure that tries to balance between increasing complexity versus increasing model fit. In this research, we focus on using Akaike's information criterion (AIC) which can be seen in the equation below.

$$AIC = \ln\left(\hat{\delta}^2\right) + 2k/T \tag{13}$$

Akaike's information criterion evaluates residual variation $\hat{\delta}$ captured by the model. In the equation T is the sample size and k is an order combined factor ($k = p + q + 1$) which is calculated from proposed model orders to introduce a penalty from increasing model complexity. In addition to AIC there are also other information criterions such as Bayesian information criterion (BIC) and Hannan-Quinn criterion (HQIC). (Brooks 2014, 263-272)

In addition to information criterion and ACF plotting also logarithmic maximum likelihood estimation (MLE) can be used to evaluate model order fit. MLE aims to maximize the probability of including observation datapoint into the model and the model with the highest MLE value captures much as possible from dataset characteristics. Another technique to relatively evaluate model fit is to use non-linear least squares where the objective is to minimize numeric value to maximize model fit. (Pelgrin 2011)

The final step is to run diagnostics testing to check residual data after a model fit. The basic idea of diagnostics is to evaluate if the model is a bad fit for data or not. A properly fitting model can extract all patterns from data. If this succeeds, residual data should be so-called white noise process. This means a stochastic process with constant mean, variance, and covariance over time. Diagnostics is executed using tests like Breusch-Godfrey and Durbin-Watson which test model for autocorrelation structures. An alternative approach is to use plotting techniques like residual distributions and correlogram plotting. (Brooks 2014, 269-272)

3.4.2. Stationarity Testing

As a part of Box-Jenkins's approach stationarity testing is a key step during the first phase of model identification. Stationarity testing is required since the outcome of this testing can help to define if another statistical testing can be used for the dataset. Stationarity in its weak form means that variance, mean, and autocovariance structure remain stable over the time series. Strict stationarity means that series values distribution remains the same across the time series. In practice stationarity means that possible shock introduced to the series would have a decaying effect on the series when time passes on. (Brooks 2014, 263-270, 334-340)

From a non-stationarity perspective, this means that if the series is non-stationary shocks have an infinite influence on the series values. This means that the forecast reliability in methods that require stationarity is seriously affected by non-stationarity since statistical foundations are built using the stationarity assumption. This effect can be mitigated by using appropriate methods for non-stationary models. As an example, ARMA models assume stationarity whereas ARIMA models are used with non-stationary datasets. (Brooks 2014, 263-270, 334-340)

When conducting stationarity testing, we have several available methods. The most common methods are Augmented Dickey-Fuller (ADF), Phillips-Perron (PP), and Kwiatkowski-Phillips-Schmidt-Shin (KPSS). The test assumption or so-called null hypothesis for ADF and PP is that the series is non-stationary. For KPSS it's the opposite. Based on these tests we can select appropriate methods between ARMA and ARIMA before defining the right order of model using Box-Jenkins's approach. (Brooks 2014, 334-340)

When we use the ADF for stationarity testing we have a standard dual-sided hypothesis. With this hypothesis testing, we estimate the statistical significance to not reject the hypothesis which in the ADF case is that the data series is non-stationary. Since non-stationarity means that the series has time-dependent structure, ADF testing is often used to confirm visual trends from the data. The ADF result includes both hypothesis probability (p-value) and critical test values with different confidence intervals (1%, 5%, 10%). In the case of ADF we evaluate these critical values using the ADF table which uses sample size and t-distribution based values as guidance to evaluate test generated critical ADF values. ADF also provides a p-value which tells us if we can reject the null hypothesis. If the p-value is smaller than 0.05 we can reject the null hypothesis and say that series is stationary. (Brooks 2014, 334-340)

If the dataset is non-stationary, we can turn the dataset into stationarity by detrending the non-stationary series. Detrending or so-called differencing is conducted by subtracting previous observations from current observations. In a case of a seasonal pattern, this would translate also to previous season subtraction to transform series into stationarity. After differencing data stationarity testing should be repeated to ensure that the data has no higher unit-roots. If data is still non-stationary, this process can be repeated. After data is stationary, we can proceed with the second step of Box-Jenkin’s approach. (Brooks 2014, 334-340)

After these steps, ARMA, ARIMA and SARIMA models can be used to forecast future values of a series. Usually, forecasting is conducted with separate training and testing sets. This secures that model performance can be evaluated with the objective numeric method. Based on the literature research and the methodology study we confirm that these models are including significant potential for demand forecasting, and they are commonly used in forecasting modelling across different industries.

3.5. Artificial Neural Networks

This chapter focuses on the neural network structure usage in demand forecasting. At first, we are going to explore the general operating principles regarding neural networks. Operating principles help us to understand how the models are constructed and how parameter setting affects forecasting performance. After general structure, we will study the differences and potential of prominent neural network types. Finally, we will discuss the strengths and weaknesses of neural networks from our research problem perspective.

As discussed in chapter 2.4, AI methods like artificial neural networks (ANN) are increasing in popularity and they can effectively solve many forecasting problems. Like previous modelling techniques, also neural network operations include two stages: training and testing. Model training can also be supervised or unsupervised learning. In supervised learning, we utilize model iteration through error measurement whereas in unsupervised learning there are no right answers and error measurement forehand but only pattern recognition from current facts. (Boritz, and Kennedy 1995, 504)

Artificial neural networks are a machine learning subfield in which a computer learns to recognize patterns by analysing sample data through network layer structure. A single neural network layer is a combination of linear parameters. An artificial neural network with a single neuron layer consisting of nodes is similar to a linear regression model. The combination of multiple hidden layers in neural networks enables a recognition of also nonlinear relationships between parameters. Neural structures with multiple hidden layers are often called deep neural networks. A simplified neural network structure can be seen below in figure 6. (Zhang and Zhang 2018, 29-31)

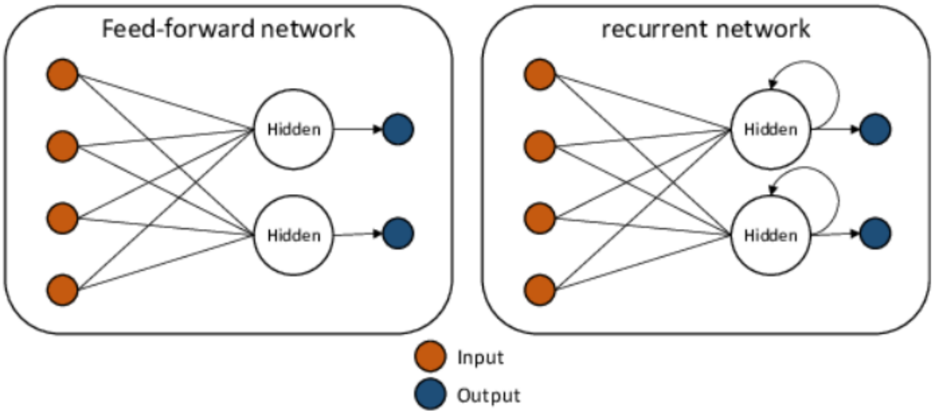


Figure 6. *Neural network differences between recurrent and feedforward in simplified form. When a model has multiple hidden layers, it is called a deep learning model. (Alfarraj and Alregib 2018)*

In more detail, a neural network consists of nodes that are interconnected into neural layers. These layers operate in a feedforward principle where the data flows through the layers only in one direction. Each node within a layer has a weight multiplier for each of its input connections from the previous node layer which can be seen in figure 7. These weights combined with input signals result as a single number which can be measured within a node to redeem if the resulting signal is sufficient to pass on to the next node layer. This evaluation is called threshold valuation. (Colin and Ruxton 2010, 7-10)

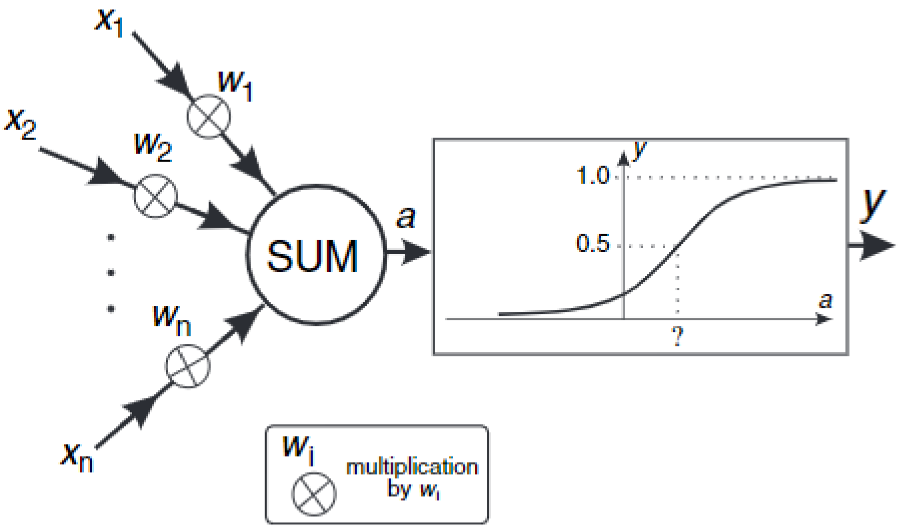


Figure 7. *Simple neuron node structure with 3 interconnections from the previous network layer. Summed input of previous layer nodes is evaluated by the squashing function (Colin and Ruxton 2010, 8,13).*

When the neural network is trained, the initial values for weight multipliers and threshold values are set to random. Training data is fed into the network input layer, and it passes through the hidden network layers, getting multiplied and added together within the nodes. Finally, the modified signal arrives at the network output layer as an outcome. This process is repeated, and weights and thresholds are adjusted until the network achieves a consistent result with the training data's known output values. (Colin and Ruxton 2010, 7-10)

In practice, the weight and threshold adjustment mean that we measure the absolute error between the network node's current output and correct output across the network. Minimizing this error seen in equation 14 over iterative rounds is the purpose of neural network training. In equation 14 we sum the squared differences between the current output y_p and the correct output t_p from each node.

$$E = \sum_{p \epsilon P}(y_p - t_p)^2 \tag{14}$$

In standard setting the initial threshold and weights are random. Therefore, there is a possibility that the iterative process trying to minimize *E* value from a random starting point does not achieve the best possible outcome. This problem is illustrated in figure 8. Since the error curve can include local minimums with suboptimal performance a simple iterative process can conclude as suboptimal performance. To solve this issue, we can utilize multiple random starting points, data normalization, gradient descent calculation, and backpropagating techniques which enable us to guide weight and threshold value training toward global minimum error. (Boritz, and Kennedy 1995, 504-506)

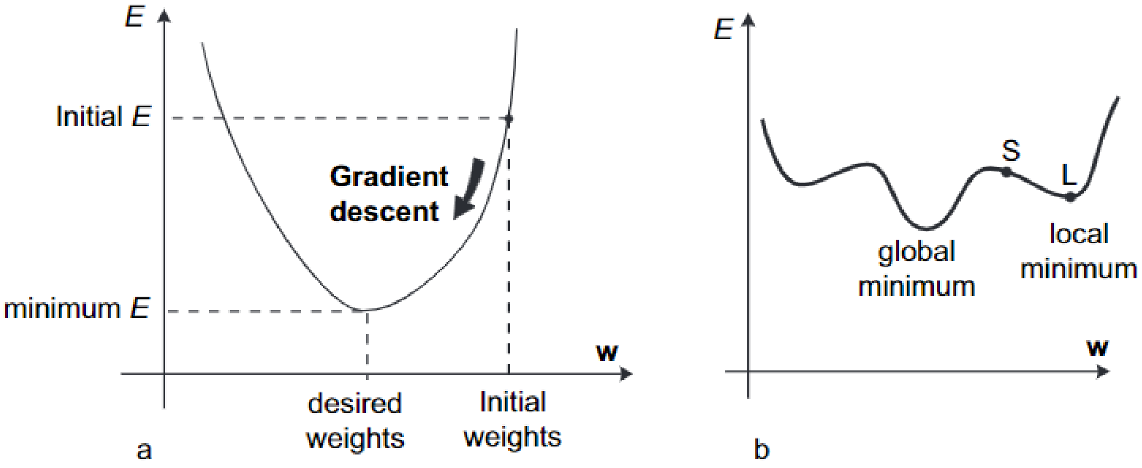


Figure 8. *Random weights and threshold value iteration can lead to sub-optimized performance when an error is minimized to local rather than the global minimum (Colin and Ruxton 2010, 14).*

The main purpose of neural network training is to adjust the weights and threshold functions of neural layers. This enables data pattern capture from input data. In modern neural network models, the system measures the change in error term after each iteration round. For some rounds, this E might increase through seeking of global minimum instead of a local minimum but in a longer horizon when the E reaches a long-term minimum point, we can end the neural network training. Training iteration length is therefore dependent on the model layer structure and the input data. After training this neural network model can be used to simulate future values. (Boritz, and Kennedy 1995, 504-506)

Most common neural networks can be classified into three main classes: feedforward, convolutional and recurrent networks, although there are numerous iterations developed continuously (Wei et al. 2018, 7-12). Feedforward networks with single-layer or multilayer are the most used network type in forecasting and they are often referred to as general artificial neural networks (ANN). They were initially developed to solve non-linear relationship problems which were not solvable through traditional regression modelling. Non-linear capabilities make network types like feedforward networks well suited for more complex forecasting problems. Additional usage cases for feedforward networks are found in computer vision, pattern recognition, and classification. (Brooks 2014, 385-389)

Convolutional neural networks (CNN) are commonly used in image and object recognition. They specialize in replacing highly entropic features in observations with a modelled relationship. CNN structures are based on the shared-weight architecture of the convolution kernels and filters that slide along input features and provide translation equivariant responses known as feature maps which allow the network to remember previously recognized features of the input data. This structure is explained in more detail in Appendix 3 and is also called as convolutional function layer. (Lin, Zhengbing, and Srinivas 2018, 259-263)

A special version of CNN networks called temporal convolutional network (TCN) uses causal convolution where output values of the convolutional layer are only dependent on the previous input values. This neural network structure first proposed by Lea et al. (2016) can also take a sequence of any length and map it to an output sequence with a matching length. However, TCN is limited by CNN operating principle of using surrounding patterns to predict a sequence (Appendix 3). In practice, this means that although CNN networks are a potential case for sequence forecasting in general, they are not able to forecast time period demand without knowing the time samples after the forecast period, and therefore cannot be used to solve the forecasting problem of our research. (Hewage et al. 2020)

Recurrent neural networks (RNN) have a recurring loop operation that enables the handling of complex pattern recognition. RNN retains a memory of what it has already processed and thus can learn from previous iterations during its training. Therefore, recurrent neural networks are commonly used in sequential input data such as time series data, text and speech recognition, and language generation. Generating natural language is a highly complex process enabled by the feedback structure. The feedback structure of RNN can iterate over a single neuron layer or over multiple layers depending on the detailed network type. There are multiple RNN subtypes like long short-term memory (LTSM) and gated recurrent units (GRU) which were mentioned in related research chapter 2.4 concerning forecasting models. (Bianchi et al 2017, 9-11, 23-28; Raminez-Montanez et al. 2021)

LTSM is RNN subfield that has both long- and short-term memory of previous iteration rounds and used weights and thresholds. Therefore, it can learn from patterns that have longer impact sequences in source data. Long-term pattern recognition increases the model complexity, but it can also capture trends that are not possible for normal short memory-equipped RNN. GRU models are LTSM variants with more efficient computation. 2014 developed GRU model has some modifications to the normal and much older LTSM model such as merged input signals and limited output gate updates. GRU simplifies the LTSM model to achieve better computational performance with often similar forecasting capability. Simplified generic structures of feedforward and recurrent network types can be seen in figure 6. (Raminez-Montanez et al. 2021)

Although neural networks sound like sophisticated modelling techniques with an easy answer for every problem this is not the case. Neural networks have both strengths and weaknesses in their implementation. These attributes are resulting in the core operating principle of network structure and therefore competence areas also differ based on the modelling type.

In general, we can say that neural network strengths include the ability to work with incomplete or non-linear data since neural network structures are not expecting only continuous data. The node structure of the network is also able to share the load among processing tasks which makes the network less vulnerable and enables distributed cloud computing to reduce processing times. This also enables simultaneous computations which greatly improves the modelling performance under continuous loads (Mijwel 2018). A neural network also stores the information it processes in a form of threshold and weight attributes which makes it more robust to operate in a case of partial memory loss. Finally, since neural networks are a machine learning subfield, they can learn from events and utilize these observations in future decision-making. (Tu 1996; Dumitru 2013, 444-448)

Disadvantages of neural networks are mostly related to lack of context understanding and computation net impact. (Thompson et al. 2020) Since networks apply to any context, they have no specific context understanding or reference which makes neural networks prone to overfitting and less usable in practical model implementation. Neural networks can also have unexpected and unexplainable behaviour due to capturing noise from the data. This can cause major trust decay from the user perspective when the network is outputting inconsistent results from a black box solution. A vast field of network types and configurable parameters also means, that there is no coordinated process or context-related principles to narrow down the optimum model for a given dataset. Model fit iteration needs to be done by testing and evaluating after each iterative parameter change. (Tu 1996; Dumitru 444-448)

The second disadvantage is related to network size. Neural networks vary from simple to vast deep learning networks. Network structure complexity increases the computing potential and the capability to learn from the data. This increased complexity will also have a significant impact on the required computing power. Computational requirements increase exponentially when the neural network adds new hidden layers. This is caused by the training process which iterates through hundreds of cycles with each parameter setting for the model. Computational intensity is further underlined by the training data requirements which increases when model complexity increases. (Thompson et al. 2020)

In practice this computational intensity means that complex models are calculated by utilizing distributed computation services like AWS or Google cloud computing. The utilization of distributed computing adds both time and cost factors to the net impact equation for neural networks. These costs can outweigh the benefits of neural network forecasting and therefore the computational power requirement is a dimension that needs to be considered in neural network model implementation.

Based on the strengths of each neural network type we can conclude feedforward networks (ANN) to be the most prominent candidate for time series forecasting, although RNN has versions that are capable to utilize univariate data to predict future values. Feedforward network popularity is based on a simple computationally efficient structure that is capable capturing numerical patterns. Recurring networks can capture patterns from continuous data and iterate model fit over memory functions. This cell-based memory in LTSM and GRU models enables the capture of long-term patterns from the data which can be utilized for future demand forecasting.

3.6. Linear Regression Model

Upon to this point, we have studied time series models which are utilizing previous values of the target variable or the error terms of previous values without external explanatory variables. The regression model evaluates the relationship between our given dependent variable and one or more other variables which are usually known as independent variables (Brooks 2014, 94-95). At the beginning of this chapter, we will focus on the classical linear regression model (CLRM) structure and basic principles. After this, we will cover the testing associated with regression model fit. Lastly, we will cover the strengths and weaknesses of linear regression models.

In regression models, we assume that the dependent variable (y) is stochastic, and it has a probability distribution. The Independent variable (x) is assumed to have fixed non-stochastic values in repeated samples. The main idea is to model these repeated samples from x variable to describe the behaviour of the dependent variable y values. The typical form of this relationship also includes a random disturbance term. This error term indicates the uncertainty of the model, and it presents the residual value between theoretical and actual observed relationships between variables. The equation for a basic linear regression model with a single explanatory variable can be seen in equation 15 (Brooks 2014, 94-102)

$$y_t = \alpha + \beta x_t + u_t \tag{15}$$

In this equation, alpha is the intercept term which indicates the value on the y-axis where the model crosses the x-axis value of zero. The beta value in the equation is the slope value which indicates the growth rate and direction of y values when x values change accordingly. The final term in the equation is the error term describing the model uncertainty level. This error term of the linear regression model includes the excluded explanatory variables, measurement errors of the dependent variable, and external random influences on the dependent variable. (Brooks 2014, 94-99)

The basic idea of CLRM is to fit a regression line into the data in a way where the linear line describes the data with minimum error. The line describes the relationship between dependent and one or more independent variables. The most used method to accomplish this is to use the ordinary least squares (OLS) method. In this method, we are iteratively calculating the distance between each data point and the current line in the y-axis dimension. This difference between the datapoint estimator value indicated by the regression line and the actual data point is squared to avoid omitting negative and positive distances from one another. Regression line is iterated by measuring and minimizing the total summed distance between the estimator line and actual data points. When this minimum OLS point is achieved, we can derive intercept and slope estimators from the regression line. The regression model fitting figure can be seen in chapter 4.4 figure 25. (Brooks 2014, 106-110)

Building the CLRM model includes assumptions regarding to error term. These assumptions can be called “Best Linear Unbiased Estimators” (BLUE). Firstly, we assume that error terms have zero mean and therefore no systematic error. Secondly, we assume that the variance of the error terms is constant and finite over all values of the independent variable. Thirdly, we assume that error terms are statistically independent and therefore have no autocovariance structure between them. Finally, we assume that error terms have no covariance structure between the independent variable values. An additional assumption for CLRM is that error terms would be normally distributed. (Brooks 2014, 105-110)

If the BLUE assumptions hold, we can say that the estimator values of true coefficients have minimum variance among the class of linear unbiased estimators. This is proved by the Gauss-Markov theorem. If these assumptions do not hold, the coefficient estimators generated by our model might be inconsistent, biased, or inefficient. Consistency in the context of CLRM means that model estimators $\hat{\alpha}, \widehat{\beta}$ will converge towards actual α, βvalues when the sample size of the model increases towards infinity. Estimators are said to be unbiased when they are equal to actual coefficients over a large sample size on average. The efficiency of estimators is evaluated to be true if they are unbiased and have a minimum variance when compared with other estimators of the CLRM model. (Brooks 2014, 105-110)

CLRM model validity can be questioned by the error term assumptions mentioned above. This means that we need to test these assumptions to see if we can trust model results and to see how well the model is capturing data patterns. The first assumption of zero mean error term can be checked by calculating residual value mean and by plotting the residual values and observing if they are normally distributed. Linear regression model residual mean should always be zero since OLS minimizes the distance symmetrically.

The second assumption of constant error term variance can also be called homoskedasticity. When this is not true the model has heteroskedastic error terms. This can be observed by tests like Goldfield-Quandt or White’s test. Since these tests are focused on finding heteroskedasticity it’s important to remember that a single confirming result is enough to redeem the model be heteroskedastic, but a negative result does not explicitly prove that there is no heteroskedastic point among error samples. Heteroskedasticity will bias the use of standard error calculation in model fit statistics, but model coefficients are not biased using ordinary least squares estimation. (Brooks 2014, 185-189)

The third assumption of the CRLM model was that there is no autocorrelation structure between disturbance terms, and they are therefore independent. Since actual real error terms are not known when building model estimate residual error estimate $\hat{u}_t$ are used. Autocorrelation can be positive or negative. Positive autocorrelation is indicated by a cyclical residual plot whereas negative autocorrelation structure is indicated by a spiky alternating pattern where residuals cross the time axis constantly. Autocorrelation is structure present also in the final CLRM assumption and it can be detected by using tests like Durbin-Watson or Breusch-Godfrey. In Durbin-Watson, we evaluate residual autocorrelation on a scale from 0 to 4 where 2 is considered to be neutral values above it are considered to move towards negative autocorrelation whereas values closer to zero are indicating positive autocorrelation. In most cases, autocorrelation can be reviewed as a possibility since this means that there is a pattern left in the model to be utilized by another structure. Capturing this autocovariance pattern can improve overall model fit. (Brooks 2014, 190-209)

We can also test models for their normality to see if the residual values of a model fit are skewed or normally distributed. Skewness or kurtosis of the residual values is indicating that there is a value shift that is not captured by the model. Non-normality is present especially if a dataset has significant outliers and this should be also evaluated when conducting normality testing. Normality testing can be conducted by using Jarque-Bera normality testing. (Brooks 2014, 209-2011)

When evaluating model fit, we can use the goodness of fit statistics as a guiding measure alongside MSE and MAPE calculations. The basic idea is to describe how much our model explains the total data variance. The most used statistic is R^2 which is the square of the correlation coefficient of y_t and $\hat{y}_t$ estimator. In other words, R^2 is describing the model's relative capability to explain variance out of total dataset variance. Equation of R^2 can be seen below and it subtracts the mean $\overline{y_t}$ from the actual and estimator values. In the equation, the numerator is describing the total sum of squares from the model. The denominator is the explained sum of squares. If the explained number is high when compared to the total sum of squares the model is a good presentation of the data points. If there is no unexplained residual sum of squares left the model fit is 1 and the minimum value of 0 is achieved when a model is not explaining any dependent value variation. (Brooks 2014, 159-162)

$$R^2 = \frac{\sum_t(\hat{y}_t - \bar{y}_t)^2}{\sum_t(y_t - \bar{y}_t)^2} \tag{16}$$

Addition to R^2 we can also use adjusted R^2 as a model fit measure. This takes into account the increasing number of model coefficients. Problem with R^2 utilization in complex models is that R^2 based model fit will never decrease if we add new estimators into the model even when they are not significant. This loss of degrees of freedom is compensated by adjusted R^2 which adds a penalty based on model complexity. (Brooks 2014, 159-162)

A holistic overview of a forecasting model also includes the critical evaluation of model strengths and weaknesses. The most significant downsides of linear regression models are that they are prone to be overfitted when utilizing multiple parameters and only few data samples without careful significance testing or dimensionality reduction. They can also underperform in complex relationships where variable relationships are not linear or if there is a complex pattern like autocorrelation present in the dataset. The most common fitting method of a linear regression model is OLS. Since OLS is measuring the squared distance between datapoint and regression line, its sensitive to outliers and missing values and this should be evaluated in explanatory analysis. (Brooks 2014, 217-230)

When considering the strengths of linear regression models, the simple model fitting is a significant advantage. Linear regression models can also provide simple relationship modelling with separable datasets making the CRLM widely adoptable and computationally efficient. With appropriate dimensionality reduction and regularization, it can be effective also as a sophisticated multivariate model. Overall, CLRM is a powerful and simple tool that enables quick execution but requires careful model fit and assumption alignment analysis.

4. Forecasting Method Implementation

The main idea of this chapter is to test the feasibility of different forecasting techniques. Testing includes first building the utilized methods with python programming modelling libraries. After building the method, we will use the dataset built according to the process that was described in chapter 3.2. as an input for the forecasting techniques. The dataset presents the real passenger demand values formed into an index “Pax” to simplify the presentation. Our research aims to find the methods which can predict these values with the highest accuracy.

The data and forecasting settings were covered in detail in chapter 3.2. As previously discussed, the forecasting model accuracy is tested with an in-sample & out-sample method where we split the dataset into training and testing sets. The training set includes 72 data points from 2011 to 2016 and is used to fit the model. After fitting the model, we can predict last 36 data points until end of 2019 and measure their accuracy by comparing forecasted values with actual data values.

As a first method implementation, we will do a brief exploratory data analysis to see how the dataset is suitable for model fitting. Based on the explanatory analysis we can see that the data has seasonal nature (figure 9, subplot 3). We can also see from subplot 2, that the dataset has an increasing trend present. After extracting the seasonal and trend components from the data, the residuals values seen in subplot 4 are constant with low variation and no significant outliers. This high consistency and low random noise level are positively affecting factors towards forecasting accuracy.

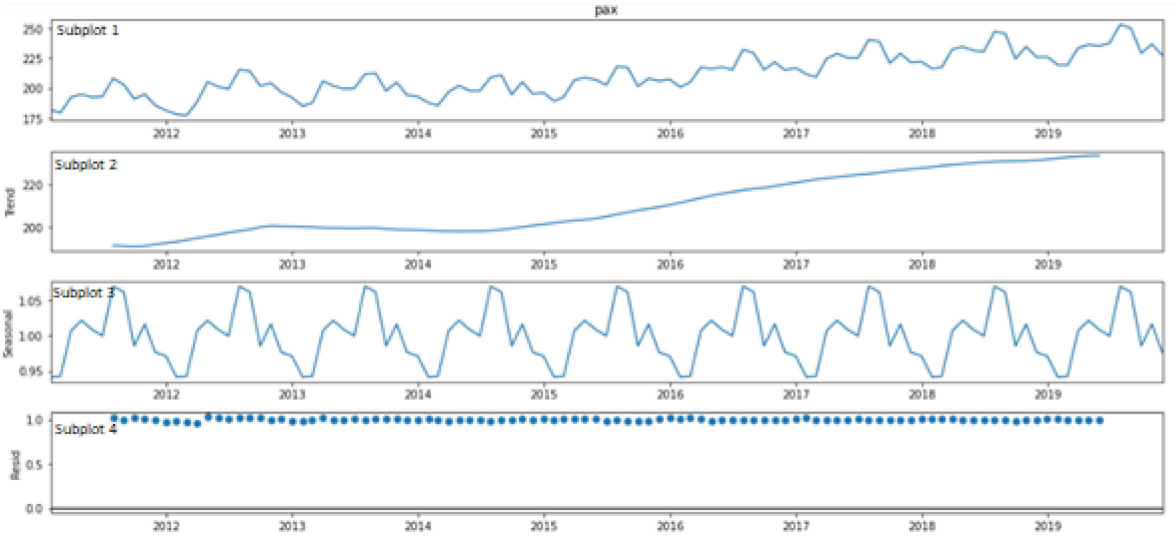


Figure 9. *Subplot 1 shows nominal dataset values. The second subplot focuses on trend and the third on seasonality components. The fourth subplot describes the residual values after seasonal and trend components are removed from the nominal dataset values.*

As stated previously in chapter 3.1, we evaluate model performance by measuring its accuracy with MSE and MAPE from which MAPE is the primary evaluation metric for this research. In addition to this we compare method accuracy to our benchmark forecasting method: naïve forecasting. The accuracy measured from 2017-2019 testing set for simple naïve forecast is (MSE: 39.82, MAPE 2.48%).

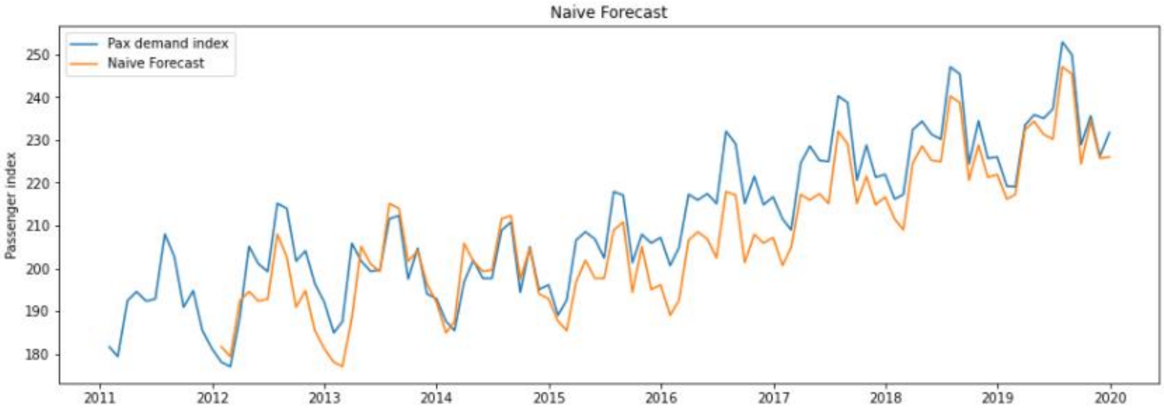


Figure 10. *Nominal values plotted with naïve forecasting estimates.*

In the figure 10 above we can see the naïve forecasting capability in practice with our dataset. The naïve model described with the orange line tracks actual demand development seen with the blue line. In these stable conditions the visual difference between the two lines is small.

4.1. Exponential Smoothing

As a first implemented forecasting method we test exponential smoothing. During the literature review and methodology, we researched the use and application of exponential smoothing. Based on this research conducted during this thesis we already know that simple exponential smoothing is not able to track accurately seasonality or trend development, although general average tracking is possible.

In exponential smoothing, we have tested different smoothing constants to understand the method capabilities. Since smoothing constants range from 0 to 1, we can evaluate from equation 3 that the larger the smoothing constant we use, the more the exponential smoothing technique gives weight to the previous value. With large smoothing constants this translates into similar data values as with the original dataset but with a lag of the length of the smoothing operation. Therefore, for forecast, we will optimize the model parameters and for plotting, we will use smoothing constants of 0.1 to view the exponential smoothing model actuation in practice. A comparison model fitting with maximum smoothing constant 1 can be seen in Appendix 4.

A simple smoothing model fit with our dataset can be seen in figure 11. The simple smoothing is only able to return the latest constant value as a model forecasted value across the forecasting horizon.

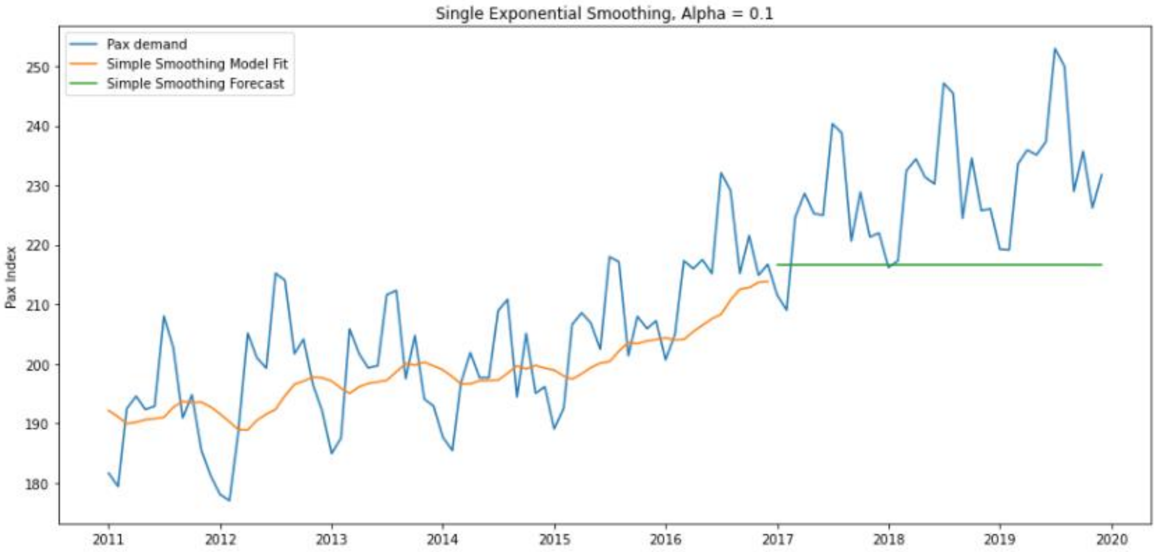


Figure 11. *Exponential smoothing tracking is based on prior values and since the forecast is just a constant value across the forecast horizon.*

When we move from simple smoothing to double smoothing which was also called Holt's method, the model can be trained to track a trend effectively. Double smoothing increases the model complexity when compared to simple smoothing but by utilizing equations 4-6 we can generate a model to include this trend component. This improves the model fit significantly since our data has a distinctive trend structure as stated in the explanatory analysis. When reviewing the following figures, it is good to note that accuracy between the actual and forecasted values is measured by calculating the absolute difference between the dataset and model forecasted values in the y-axis by each data point in the x-axis.

Both trend and seasonal components for double and triple smoothing can be either additive or multiplicative. This means there are two model iterations for Holt's method and four for triple exponential smoothing. In Additive we combine the factors inside of equations whereas in multiplicative equations we multiply the factors. During the research, we covered the additive calculation principles. The multiplicative equations can be seen in Appendix 2. We do the model fitting both ways and plot the model with a better fit for evaluation. Holt's method can generate a forecasted trendline based on model fit but it does not include seasonality prediction.

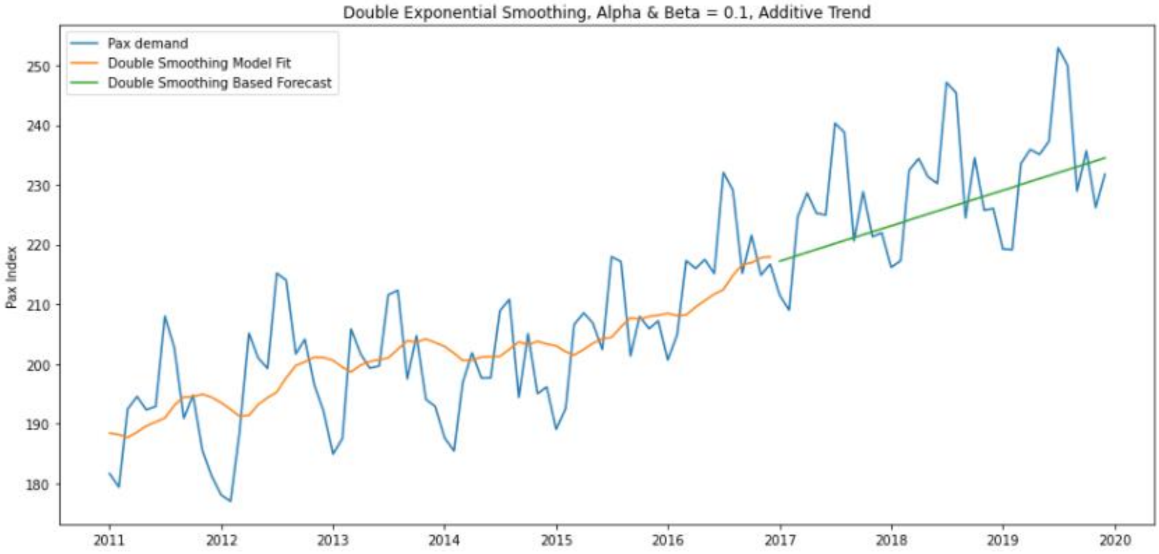


Figure 12. *Double Exponential smoothing can track a trend development over time.*

Finally, when building the model for the combined triple exponential smoothing which is also called as Holt-Winters method, we utilize equations 7-10 to include a seasonality component. Adding this parameter for the model equations enables the capture of seasonality in our dataset. We use the training dataset to train the model and measure model accuracy based on the differences between actual demand values and model forecasted values. We can see a visual improvement in the model forecast in figure 13 when compared to forecasts seen in figures 11 and 12.

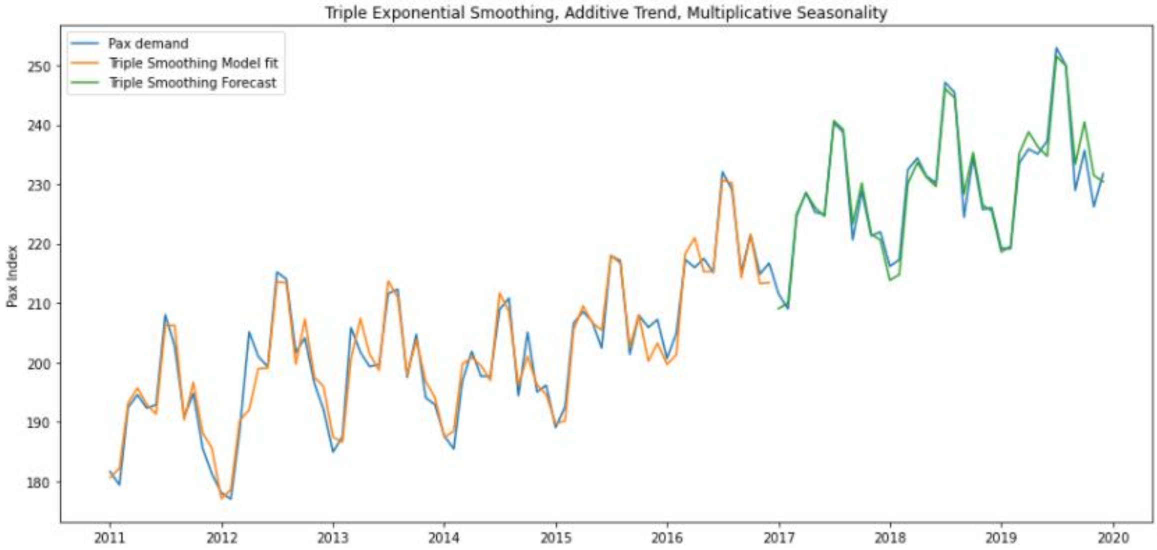


Figure 13. *Triple exponential smoothing.*

The procedure is repeated for other iterations of exponential smoothing. As expected, the method accuracy increases towards more complex smoothing techniques. The exact results of each method variation can be seen in table 2.

Table 2. *Exponential Smoothing method forecasting accuracies from simple to complex method.*

Method	MSE	MAPE
Naïve Forecasting (Benchmark)	**39.82**	**2.48%**
Simple Exponential Smoothing	269.53	5.84%
Double Smoothing, Trend = Multiplicative (Mul.)	124.46	3.70%
Double Smoothing, Trend = Additive (Add.)	94.72	3.33%
Triple Smoothing, Trend = Mul., Seasonal = Mul.	7.38	0.84%
Triple Smoothing, Trend = Mul., Seasonal = Add.	6.18	0.83%
Triple Smoothing, Trend = Add., Seasonal = Add.	5.19	0.79%
Triple Smoothing, Trend = Add., Seasonal = Mul.	4.18	0.67%

In the domain of exponential smoothing, triple smoothing is the only model which can generate a reasonable forecast considering both trend and seasonal components of our dataset. Triple exponential smoothing's forecasting accuracy is 0.67% in terms of mean absolute percentage error. This result is a significant improvement when compared with our benchmark forecasting method.

4.2. ARIMA & SARIMA

Forecasting models based on autoregressive and moving average processes requires stationarity testing. In the first step of the Box-Jenkins approach, we will need to understand if our dataset is stationary. Since we already know from exploratory analysis that our dependent variable dataset does not have constant distribution over time, we are testing weak stationarity and not strict stationarity. The definition of weak stationarity is that series has a constant mean, variance, and autocovariance structure over a time horizon. By looking at our exploratory data analysis and visual data presentations we can already say that it has a clear trend and changes over time. This observation is confirmed by doing Augmented Dickey-Fuller stationarity testing for our dataset. Based on the test results seen in table 3, we cannot reject the null hypothesis of ADF and so it is highly likely that the series is non-stationary.

Table 3. *ADF testing confirms that the dataset is non-stationary. Differencing data turns data into stationary.*

ADF (original data)		ADF (differenced data)
ADF Statistics:	-0.81	-3.11
p-value:	0.815	0.025
Critical val. (1%, 5%, 10%):	-3.50, -2.89, -2.58	-3.51, -2.89, -2.58

Since the dataset is non-stationary, this excludes the usage of simple ARMA models. However, we can still test the adequacy of ARIMA and SARIMA models for solving our research problem by including the differencing pre-processing step of the series to make the dataset stationary, called integration (I). When looking at the initial explanatory data analysis we can see that the data has both seasonal and trend components. We need to difference the data in terms of seasonality and trend components to make the data stationary. Differencing is conducted by subtracting previous observations from current observations. When we run the ADF testing with this altered data we see that now we can reject ADF null hypothesis, and so the data is stationary.

After introducing stationarity into the series, we can continue with the Box-Jenkins approach. This second step begins with the estimation of model structure, and it includes the utilization of autocorrelation and partial autocorrelation function plots and information criteria. When looking at the ACF and PACF plots in figure 14 we would estimate that the optimal model order could be ARIMA (18,1,2). ACF has two autocorrelation spikes with lag values of 1 and 12. PACF has 7 spikes at lags: 1, 2, 11, 12 ,13, 16 and 18. These observations combined with differencing the data once would indicate that by visual interpretation ARIMA (18,1,2) model would be suitable. However, model order identification from ACF and PACF plots is not exact or crisp, as described in chapter 3.4.1. To confirm the visual interpretation, we will need to evaluate model order with information criteria.

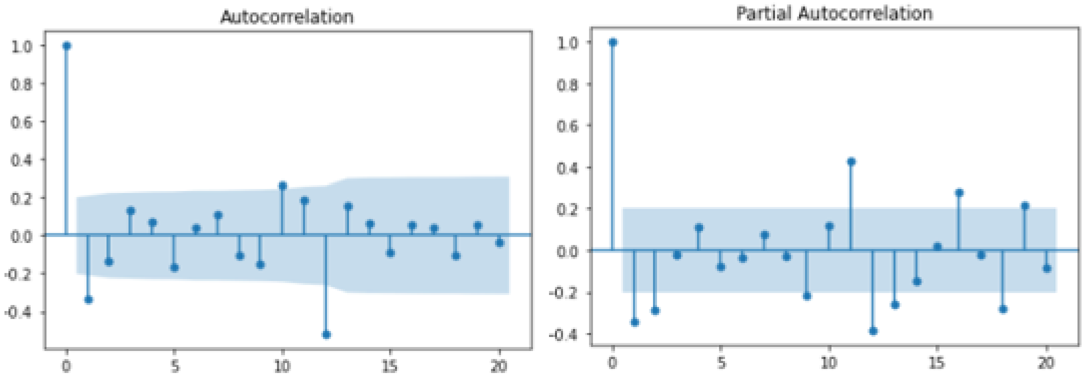


Figure 14. *PACF and ACF plots show that there is an autocorrelation structure that can be captured by the ARIMA model.*

The model identification is an iterative process where we try to minimize information criteria that evaluates the model fit. We try ARIMA(p,d,q) models with different orders to understand which model achieves the best data pattern captured from the data. This process is done for all order combinations up to AR lag 18 and MA lag 12. Based on this iterative process ARIMA(12,1,2) and ARIMA(10,1,12) models achieve the best model fit with the smallest Akaike's information criteria (489.6, 486.6). When we evaluate other model possibilities, we also see that ARIMA(12,1,12) and ARIMA(18,1,1) are close with their information criterion values (490.5, 493.9). To support this model evaluation, we can use python's optimization packages which test different model combinations to find the best model fit. Manual testing statistics of ARIMA models can be seen in Appendix 5.

This process is also repeated with the SARIMA (p,d,q)(P,D,Q,M) model where we also have the seasonal order values to optimize. For SARIMA models, we will test the model fit up to 12 lags with (p,d,q) part and up to 5 lags with seasonally multiplied (P, D, Q) orders. Based on testing, SARIMA(1,1,1) (2,1,1,12) model achieves best performance when fitted with our data. With this combination, AIC reaches a value of 469.4. Another viable model is SARIMA(1,1,1)(2,1,0,12) with AIC of 471.3. Based on this model identification it seems that SARIMA models could be more suitable for series forecasting for our research problem. From Appendix 5 we can see the model performances of each order iteration.

In comprehensive testing, we could repeat this optimization using another information criteria like Bayesian information criterion (BIC) or Hannan-Quinn information criterion (HQIC) and compare the test results to see if different information criteria methods achieve the same superiority order since the model complexity is penalized differently in different information criteria. However, since the focus of this research is on forecasting performance, evaluation by autocorrelation function plots and AIC is sufficient.

After we have estimated model parameters with ACF plots and AIC evaluation, we can evaluate the statistical significance of model coefficients. In ARIMA(18,1,1) only AR1, AR11, AR12, AR13 coefficients seem to be statistically significant. This pattern of majority insignificant coefficients is repeated with ARIMA(12,1,12) and ARIMA(10,1,12). When tested with ARIMA(12,1,2) only 5 coefficients out of 17 are insignificant, model is the simpler than comparative models, and the model achieves as good AIC and log-likelihood values in model fitting. Therefore, we will proceed with ARIMA(12,1,2) model. These AIC estimations and model coefficient table can be seen in Appendix 5.

This Process is duplicated with SARIMA models. Based on the statistical significance testing and max loglikelihood optimization we choose SARIMA(1,1,1)(2,1,0,12) as the most suitable model. SARIMA(1,1,1)(2,1,1,12) is also possible option, but since its more complex and has several insignificant coefficients, we prefer simple SARIMA(1,1,1)(2,1,0,12) model. All its coefficients in SARIMA(1,1,1)(2,1,0,12) are statistically significant and in overall perspective, it performs with low AIC and high max log-likelihood value. For comparison coefficients from both SARIMA models can be seen from Appendix 5.

As a third and final step for the Box-Jenkins model generation, we will do model validation. Validation is primarily focused on the residual study. As discussed in the methodology part, if the model fit is good the residual outcome should be a white noise process which means zero mean and variance over time and normal distribution of residuals around 0. ARIMA (12,1,2) model fit can be seen in the figure 15 below.

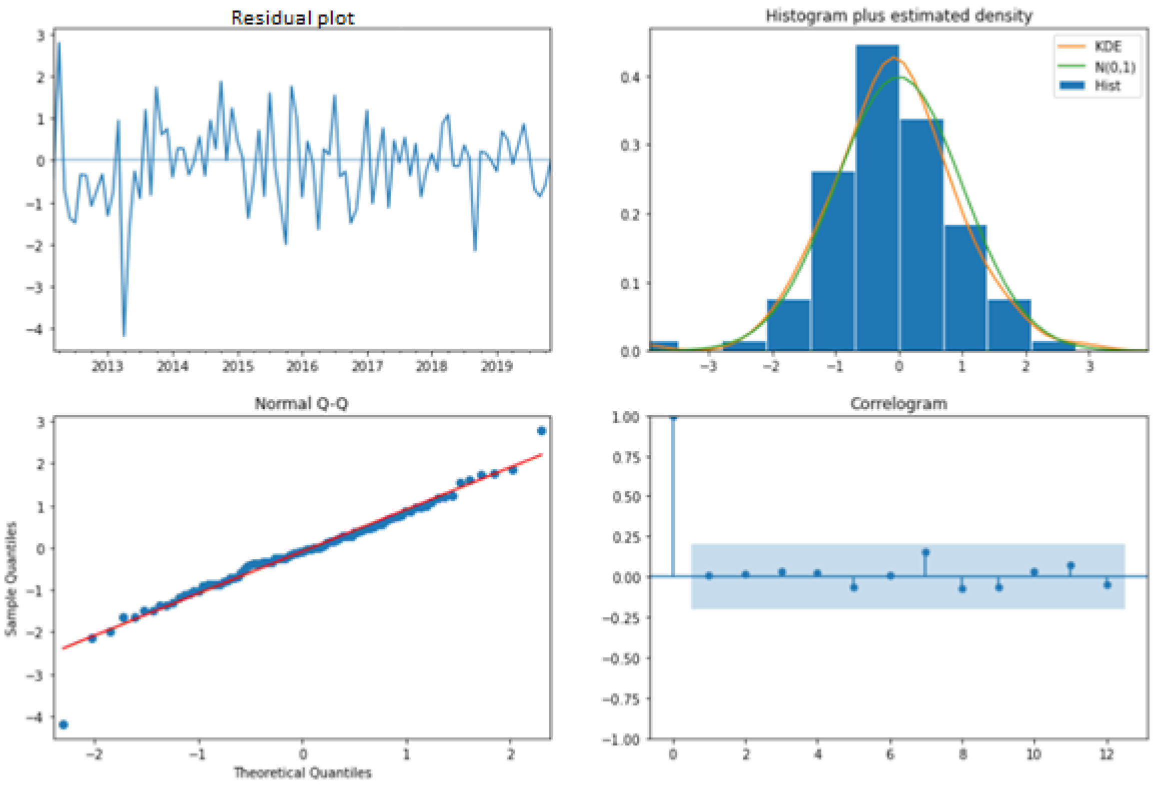


Figure 15. *ARIMA(12,1,2) model residual statistics. Residual statistics include a time series residual plot, residual distribution plot comparison with normal distribution, QQ-plot and a correlogram.*

Based on validation statistics model fit is good. In the upper left figure, we can see the residual values in a time series format for our model. In the upper right, we can see the residual value distribution. Residuals data seems to be close to normally distributed white noise process. The QQ-plot shows that the ordered distribution of residuals follows the linear trend taken from the standard normal distribution, suggesting residuals are normally distributed. The correlogram seen on the bottom left indicates that there is a low correlation between residuals and their lagged versions. This step is repeated for the SARIMA model. Based on figure 16 below the model fit is good. Both the residual distributions and correlogram look promising in terms of model fit.

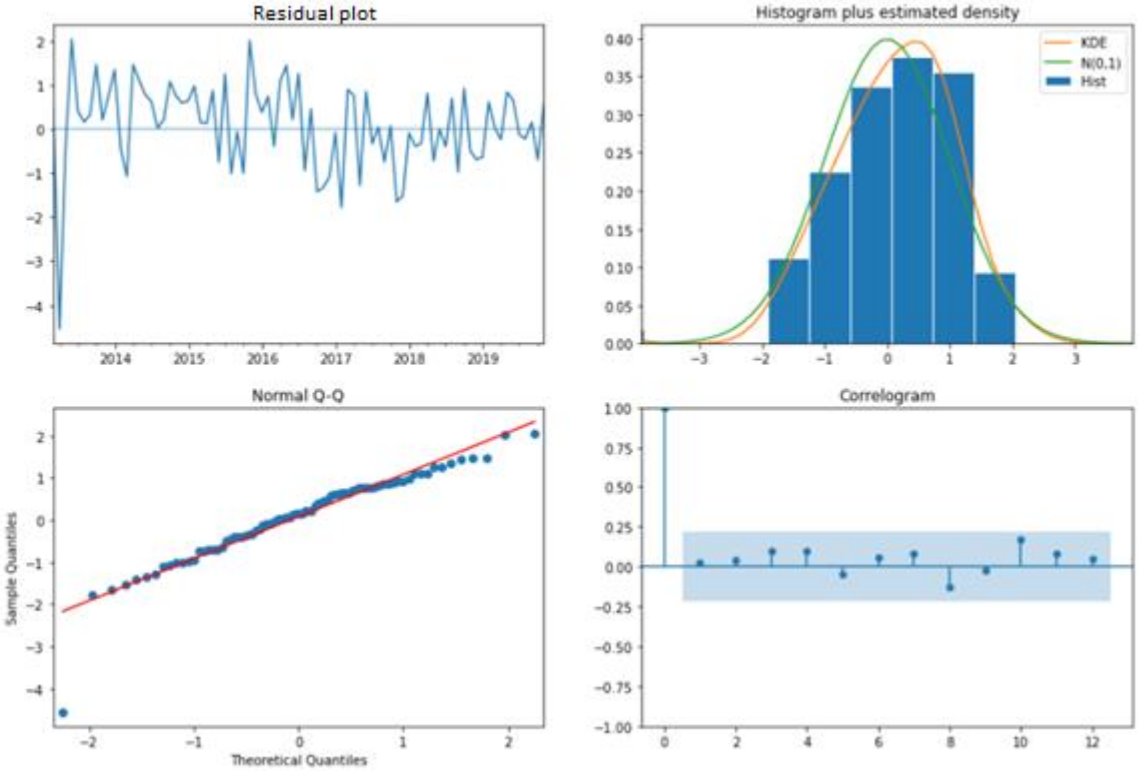


Figure 16. *SARIMA(1,1,1)(2,1,0,12) model diagnostics indicate that the model residual data is a white noise process.*

This observation if further confirmed by Durbin-Watson tests for residual autocorrelation. In Durbin-Watson, the test results range from 0 to 4, and value 2 means zero autocorrelation. Values below 2 are positive autocorrelation values and above 2 negative autocorrelations. ARIMA model has a test value of 2.02 whereas SARIMA achieves a value of 2.16 which tells us that there is no significant autocorrelation structure present in the model residuals.

As a result of Box-Jenkins's approach, we now have configured ARIMA and SARIMA models. Testing and optimization ensure that the model parameters are chosen to maximise the pattern recognition capabilities of these techniques. These models can now be utilized to solve our research problem and to test the accuracy of ARIMA and SARIMA models for our research problem.

Forecasting is conducted by splitting a dataset into training and testing sets where we use the last 3 years as forecasting accuracy measurement. After training the ARIMA(12,1,2) model we forecast the 2017-2019 data. In figure 17 we can see the modelling performance of the model. This process is replicated with the SARIMA model and the model fit is visible in figure 18.

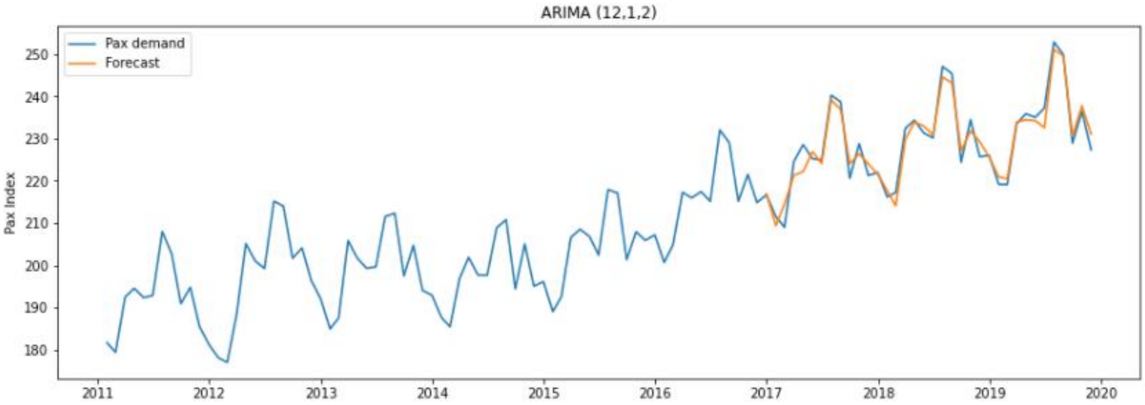


Figure 17. *ARIMA(12,1,2) demand forecast follows closely the actual demand index.*

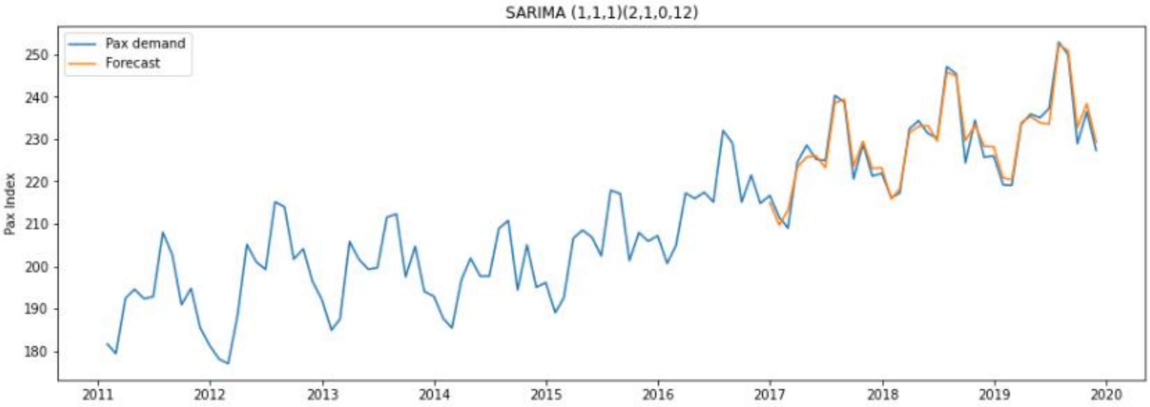


Figure 18. *SARIMA(1,1,1) (2,1,0,12) forecasted values are nearly identical to the fit seen above with ARIMA.*

The visual model results are nearly identical as seen in figures 17 and 18. When comparing the results in terms of MAPE and MSE, SARIMA achieves slightly higher model performance. When compared with our benchmark forecasting accuracy, both models are highly capable.

Table 4. *Model performance compared with benchmark forecast.*

Method	MSE	MAPE	Durbin-Watson test statistic
Naïve Forecasting (Benchmark)	**39.82**	**2.48%**	**-**
ARIMA (12,1,2)	6.47	0.92%	2.02
SARIMA (1,1,1) (2,1,0,12)	4.13	0.75%	2.16

In general, ARIMA and SARIMA models are capable of time series forecasting models. When compared to previous time series models, they are more complex and can adapt to more advanced pattern recognition. ARIMA and SARIMA models also require significant statistics understanding and validity testing which increases the total implementation complexity when considering recurring model utilization.

4.3. Artificial Neural Networks

Implementation of neural networks utilizes the data science workflow process introduced in chapter 2.1. Like with the earlier models, we will begin with a simple feedforward neural network model and increase the complexity of the model by including additional parameters in a hope that a more sophisticated model would yield improved forecasting results regarding our research problem. In the second part of this chapter, we will implement a recurrent neural network model and see if can match the performance of the most common feedforward neural networks.

4.3.1. Feedforward Artificial Neural network

For this research, we will use the NeuralProphet feedforward neural network framework developed by deep learning researchers at Stanford University and Facebook (Triebe et al. 2021). It is based Pytorch machine learning framework and trained with classical in-sample/out-sample method also used with previous methods. In addition to neural networks, NeuralProphet can also include statistical models like autoregressive processes, seasonality functions, and trend regression. The biggest advantage of NeuralProphet is that it offers an easily approachable and tested framework as an introduction to neural network forecasting where we can start with automated settings and disabled features. After a low barrier start, we can enable model functionalities in an iterative process where we seek to improve the model fit.

Forecasting is evaluated with the same 3-year forecasting horizon that we have used with previous techniques. In the standard setting, we are telling the neural network that data is in a monthly format, and we want to predict for the duration of our testing set. This standard-setting also includes that we are using additive seasonality with linear growth trend expectation over time since from explanatory analysis we know that the pax growth is more linear than exponential over time. Like with the previous models, the network also tries to achieve a model where residual data is normally distributed white noise process. In figure 19 we can see the forecasting capability of this crude network without additional information.

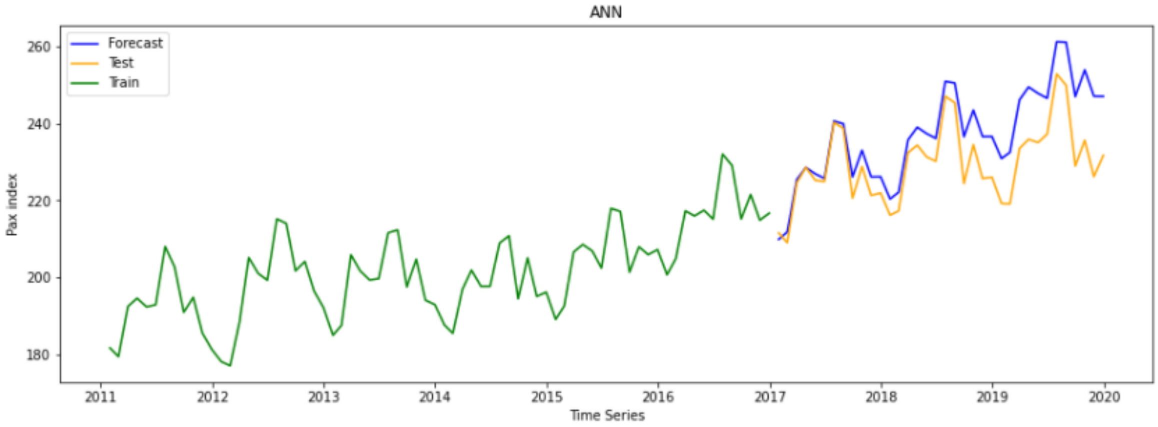


Figure 19. *Standard neural network achieves reasonably good accuracy without any additional information regarding the data characteristics.*

To combat the local minimum error iteration problem discussed in chapter 3.5 (figure 8) we are using model fitting cycles that will fit multiple times and each fitting uses 300-800 iteration cycles and random starting points to find the minimum error point for network node weights and threshold functions. The iteration cycle amount is dependent on the model complexity and the improvements seen in the model training error described in equation 14. The model performance is the mean of these model rounds, and it, therefore, describes the average performance of the network with standard parameters.

The second network training model is improved by telling the network in advance, that there is an annual seasonal pattern present in the data. This has a significant positive effect on the forecasting accuracy which can be seen in figure 20.

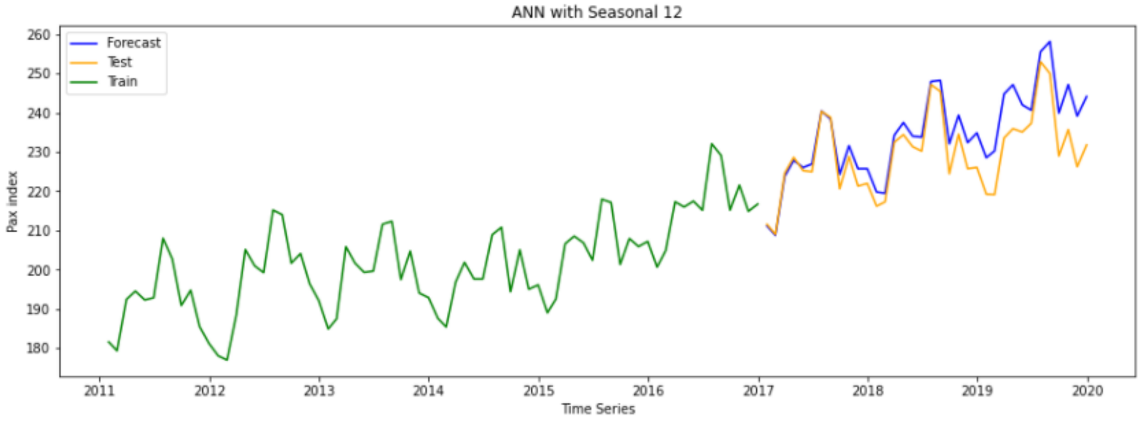


Figure 20. *The model with advised seasonality has a more accurate model fit.*

Finally, we add the neural network model with an autoregressive process which we saw to improve fit with ARIMA model in chapter 4.2. In this case, we test out the performance of different AR orders. Based on the testing, AR 1 is achieving consistently the best model fit. The use of AR1 can also be justified when looking at the residual autocorrelation plot of our first ANN model. We see a clear autocorrelation structure with lag value 1. This can be seen from Appendix 6 which also includes the residual plots of other ANN models in this chapter. Exact numerical accuracy results are reviewed in table 5 at the end of chapter 4.3.

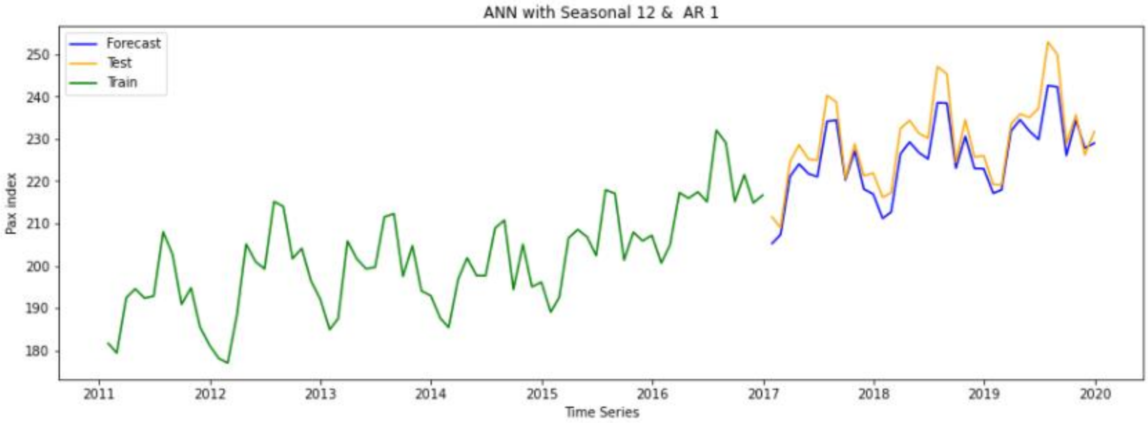


Figure 21. *ANN model with autoregressive component and informed seasonality.*

Based on these results artificial neural network has in an overall good performance. From residual plots, in Appendix 6 we can see that model fit in ANN models improves when we use the seasonal and autoregressive components. We can also see that the model outcome is not a white noise process since the residual data is not normally distributed around zero. A special note is that so-called blind performance is relatively good with neural networks when compared to benchmark forecasting. And this performance is achieved without any additional information regarding the forecasting data features. When we enhance this crude model with info regarding the data features, we see improvement in forecasting performance which can be compared in table 5.

4.3.2. Recurrent Neural Network Model

Recurrent neural networks differ from feedforward networks by having a feedback loop to the previous network node layer. This enables the constant recalibration of node weights and threshold function. Based on the research conducted in the methodology part we already know that for time series forecasting, RNN models are more complex, but they have a memory that enables the capture of long-term patterns. This highlights the question if RNN can match or outperform feedforward neural networks in forecasting accuracy in absolute terms. In this research, we are going to use 3 types of RNN models. Classical RNN (vanilla), LTSM, and GRU. Models are built using a python-based Darts machine learning library. We are going to start with the simplest one and increase the complexity with GRU and LTSM models.

Unlike the classical feedforward network, which was originally developed to solve non-linear parameter models, in RNN models we are using normalized datasets with passenger index values calibrated from 0 to 1. This enhances the network's pattern extraction capability and improves the network's capability to counter gradient problem described in chapter 3.5 (figure 8). In practice, normalization means that we are losing the standard MSE comparison between other models and therefore the relative MAPE measurement is highly useful to understand the comparative model performance. The performance of RNN models is largely affected by the utilization of RNN’s memory capability. If the model can only iterate for a shorter time, quickly adapting simple models are stronger in performance. Another option is to iterate the model longer, this means that the LTSM models are likely to show significant strength over the classical vanilla model. Like with all neural network models, the parameter tuning has a significant impact on the model accuracy.

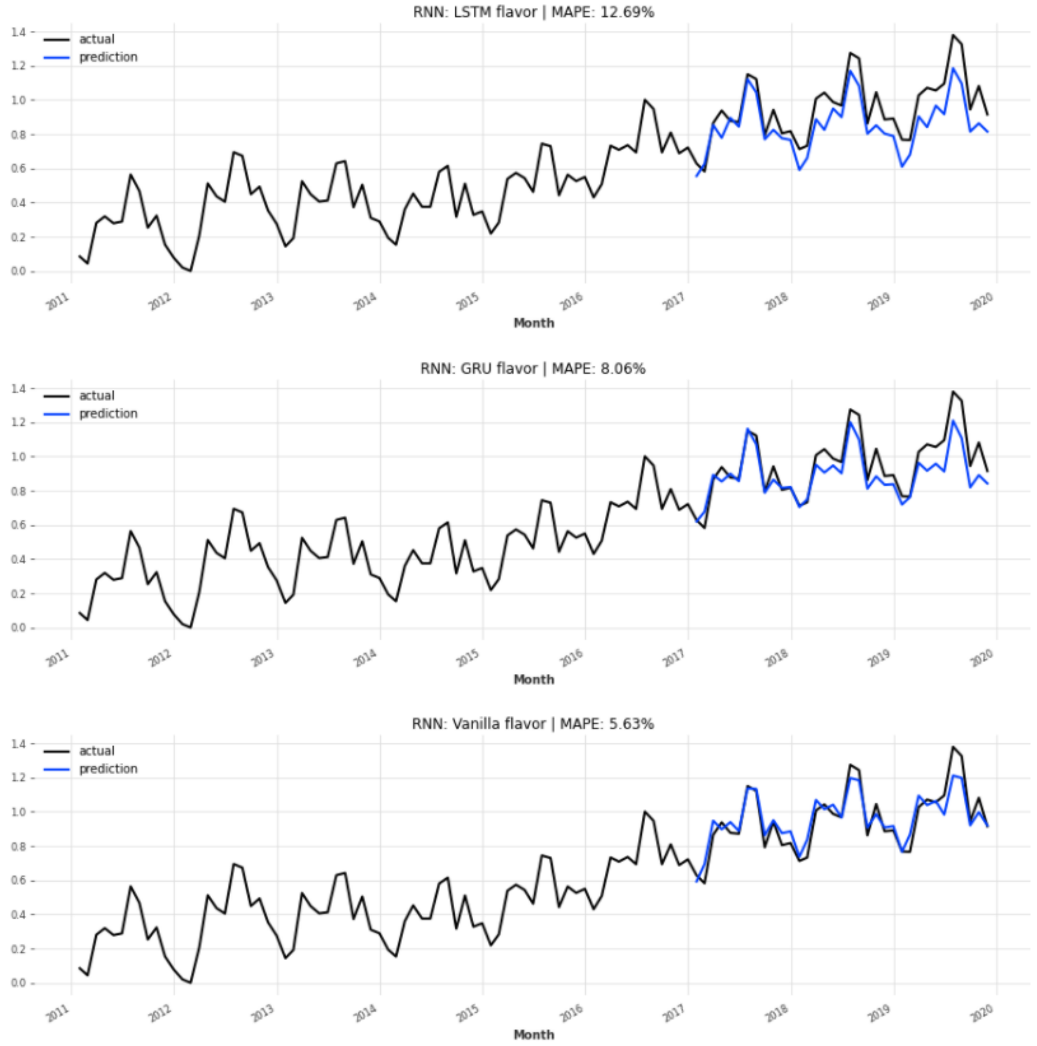


Figure 22. *RNN models from complex to simple.*

Based on the results parameter tuning has a significant effect on the RNN forecasting capability. Since this pattern is relatively consistent and it has clear trend and seasonality structures and a limited sample size, LTSM and GRU are not able to effectively use their memory capabilities in this research problem. Therefore, the classical RNN is outperforming these computationally more demanding variants by a significant margin. Exact numerical results can be seen from table 5 below.

From these results we can also see that the superiority order of the methods differs if we use MSE instead of MAPE when comparing naïve and ANN seasonal 12 forecasts. This occurs since MSE squares the errors as described in equation 1. MSE results as a better value if a measured error is more consistent over the measuring period whereas MAPE is determining average error in percentages and is not affected by the error squaring. This observation highlights why we need to be conscientious when selecting accuracy measurement framework for any research.

Table 5. *Neural network model iterations quantitative comparison including benchmark method.*

Method	MSE	MAPE
Naïve Forecasting (Benchmark)	**39.82**	**2.48%**
ANN (Standard Settings, No info)	85.52	3.26%
ANN (Seasonal 12)	45.62	2.30%
ANN (Seasonal 12, Autoregressive)	20.18	1.66%
RNN (Vanilla)	-	5.63%
RNN (Recurrent Gated Unit) (GRU)	-	8.06%
RNN (Long-Term Short-Term Memory) (LTSM)	-	12.69%

When comparing the results from different neural network iterations we can see why the simple feedforward neural network (ANN) is popular in forecasting literature. By using models like Neuralprophet we can quickly generate relatively accurate results without too many mandatory parameter tunings. ANN models can also be enhanced by additional data info like seasonality or autoregressive processes.

In addition to superior forecasting performance in similar time series forecasting, ANN models are also simpler and computationally cheaper than more complex RNN models. Since our research problem has a simple identifiable seasonal trend structure, these more complex models are not able to utilize their memory and feedback functionalities to their full extent. In conclusion, it can be said that based on this model testing, our research problem seems to have patterns that are well recognized by feedforward neural networks.

4.4. Linear Regression

In regression model implementation we are starting with a normal linear regression model utilizing a single explanatory variable. After the standard regression model, we will implement an alternative model utilizing the explanatory power of the independent variable. The independent variable is extracted from the external confidential subscription database, and it is one of the leading indicators for demand planning. The exact variable is confidential information for the company, but the variable describes the market growth which has a correlation relationship with aviation demand market. The explanatory variable has been proven to be effective in demand forecasting by other internal company studies and therefore the validity of this parameter is not evaluated forehand in this research. However, we will evaluate the model performance and compare it with other models in this implementation chapter and in the following results section of this research.

In the model implementation, we will utilize explanatory data variables to predict passenger demand index development. In these efforts, we use the term “exp_var1” as a work term with this independent variable. This modelling starts with explanatory analysis where we study the characteristics of the explanatory variable. Based on the numeric data we can see that the data update frequency differs from our dependent variable. The explanatory variable value changes every quarter and therefore we have the same monthly value over 3 consecutive data points across the dataset. We are matching the explanatory variable time horizon with the dependent variable to make the series datapoint comparison and calculation easier. In figure 23 we can see exp_var1 time series development.

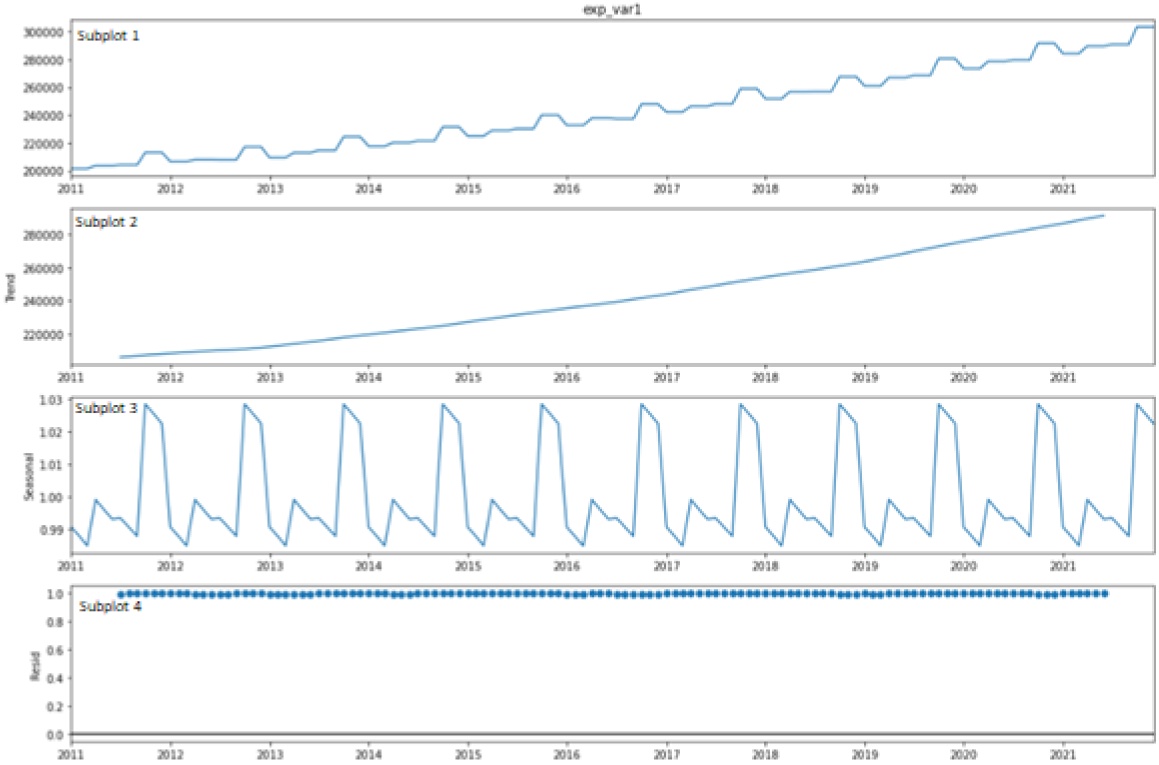


Figure 23. *Explanatory analysis shows that there is a clear seasonal trend structure present in the independent variable. We also see a similar increasing trend over the time series as we saw with the dependent variable.*

From figure 23 we can also see that the dataset has values for the 2011-2021 period. For the following linear regression model building we are filtering this down to the 2011-2019 timeframe, but the last 2-years will be utilized in the upcoming model robustness testing in chapter 4.5. The explanatory variable has clear seasonality and a similar increase that we saw with the dependent variable. It is good to note that the growth trend continues during 2020 and 2021. We can also see that the form is highly consistent and residual values do not include significant outliers. This is because exp_var1 is including a lot of data points concluded into this leading indicator and it, therefore, seems to have little noise included. If we normalize both dependent and independent variable and plot them into a single time series figure we can see that they have a similar increasing trend with a seasonal spike, although exp_var1 has a delay of 4 months when compared to the pax variable. Exp_var1 also has more consistent linear annual growth whereas the passenger index includes less growth during the years 2012-2014. This figure 24 can be seen below.

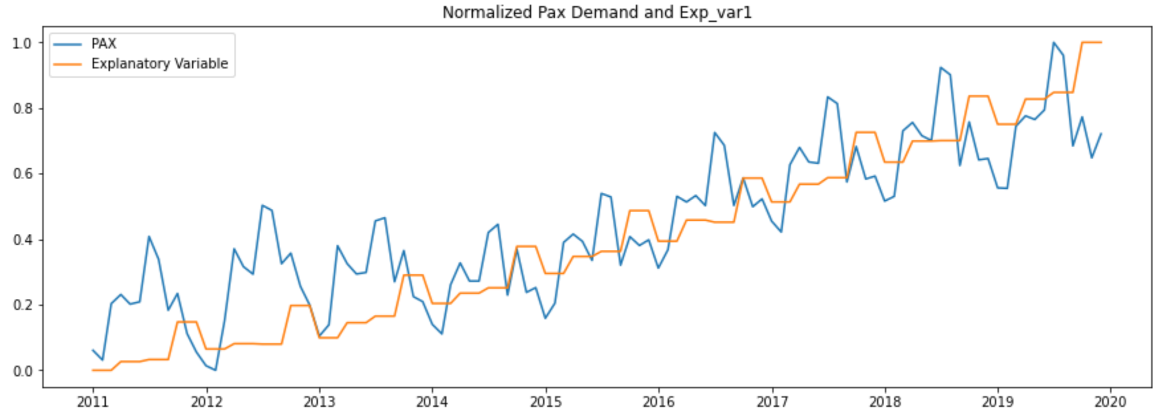


Figure 24. *When independent and dependent variables are normalized and plotted into the same graph, we can see that there is a 4-month delay shift in the explanatory variable, and they also have a similar increasing trend.*

We can tackle the delay seen in figure 24 with 2 different approaches. Firstly, we can leave it to the data. Since we are fitting a regression model, and the lag stays consistent over time it will affect coefficient and accuracy, but in a limited scale since the delay is consistent over the model. An alternative approach is to shift the exp_var1 data 4 months towards the beginning. This will align variables increasing the uniformity as seen from Appendix 7. For the following presented CLRM model, we are using aligned datasets. The second option with original not-aligned datasets is also calculated and modelled, and numerical results are presented in tables 6 and 7. A visual presentation of the non-aligned regression model can be seen from Appendix 7.

Initial model performance fitting is done by utilizing a python-based Pycaret library which includes the most common regression models for model cross-evaluation reasons. This enables us to simulate most combinations of regression models. Out of numerous regressions models such as ridge regression, Bayesian ridge, lasso, and k-nearest neighbour, we choose linear regression since it’s an easily understandable method with well-defined testing and evaluating process. The linear regression model also achieves one of the best-simulated regression fit performances for our research setting. This simulation statistic can be seen from Appendix 7. We are fitting the CLRM in accordance with OLS principles. In this approach, we minimize the y-axis difference between the model coefficient and the actual data sample aggregated over the entire dataset. The linear model fitting can be seen in the figure 25 below.

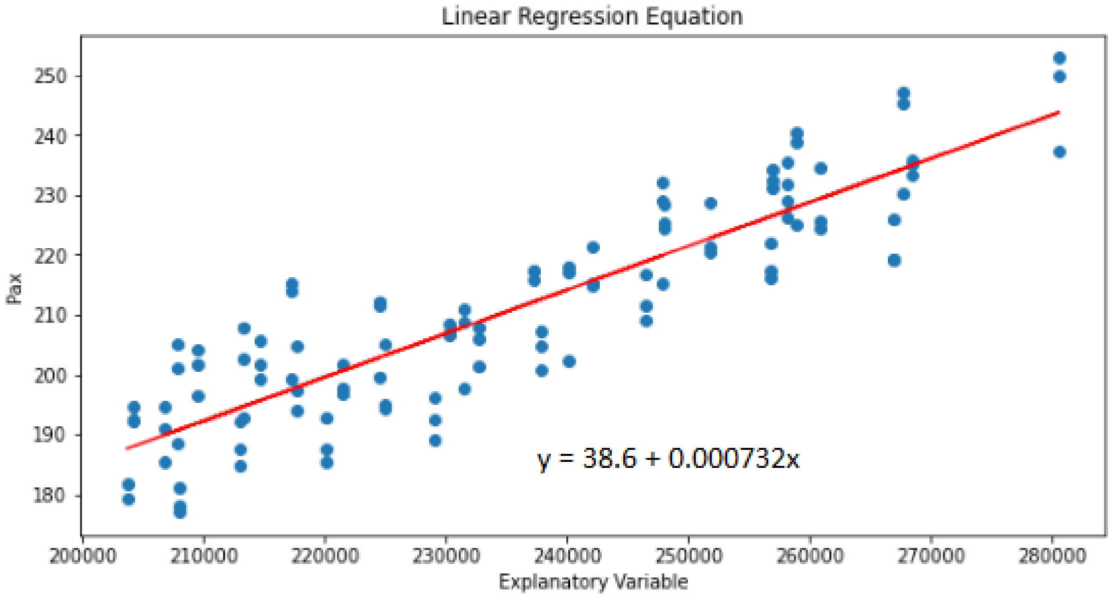


Figure 25. *Linear regression model between independent and dependent variables with the aligned explanatory variable. OLS-derived model equation can be seen at the bottom of the figure.*

We evaluate CLRM disturbance term assumptions as a part of the model fit evaluation. The first assumption of linear regression models was that the model error terms have a zero mean. Zero residual mean (1.22e-13) can be observed to be true from the residual plot seen in figure 27 and calculated from residual values.

The second assumption of constant variance can be evaluated by splitting the model residuals into two subgroups and conducting a Goldfield-Quandt test of heteroskedasticity. The splitting is done utilizing the visual level difference seen in the residual value plot. In splitting, we try to find the most extreme two subsamples of the series. We use the years 2014-2015 as the first subsample and 2016-2019 as the second subsample. Based on these samplings (24 and 48 observations), the test has a value of 1.37 which is significantly lower than the f-statistics based critical value (2.29-2.12) for these sampling sizes. We repeat Goldfield- Quandt test with different subsamples in the simulation 50 times and observe the results based on the subsample's degrees of freedom. In all simulated tests, the test result is more than 30% lower than the f-statistic based critical value. Based on the test results and the high value consistency seen in figure 27 we can say that the data is likely to be homoscedastic for 2011-2019 period.

The third and fourth assumptions are related to autocorrelation. From the residual plot seen in figure 27, we can already spot that the data has a clear seasonality-based positive autocorrelation structure which is not captured by the linear regression model. This is tested by using Durbin-Watson's test for autocorrelation. Since the Durbin-Watson test's scale is from 0 to 4 and we get a test result of 1.10 we can confirm that the residuals have a positive autocorrelation structure present which can be captured to improve the model forecasting capabilities. This autocorrelation structure of the dataset was also used in ARIMA and neural network models.

Additional to BLUE assumption we can also test model for non-normality in residuals. This is done by utilizing the Jarque-Bera test. When observing dataset outliers, we can say that the data has no significant outliers which would affect the normality testing. Jarque-Bera results with an outcome of 2.6 with a non-normality probability of 27.2%. This result is likely affected by the relatively low sampling size of the dataset. These test results indicate that there is non-normality present in the dataset and the model fit is not ideal. All the previous regression model test results are seen in table 6 below.

Based on this created CLRM we can now evaluate how do a change in the explanatory variable is reflected in the dependent variable. Based on modelled regression relationship we can generate predictions for passenger values. This is conducted with the same training and testing dataset splitting seen also with previous model families. In figure 26 we can see a standard linear model forecast with aligned peak structures. Since the exp_var1 is updated once during each quarter this is also visible in the linear regression model forecast with a fixed coefficient across the entire time series.

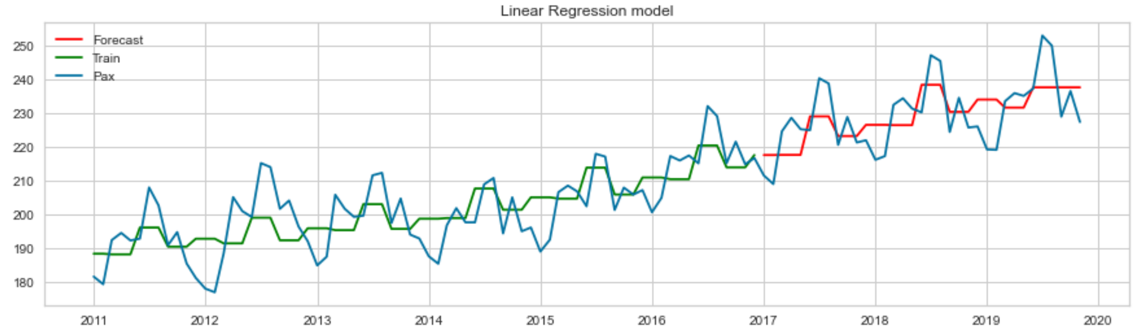


Figure 26. *Aligned explanatory variable based linear regression model forecast for demand.*

If we evaluate the model fit of the linear regression model and calculate the residual values between actual and model indicated values, it is evident that model is not able to present the seasonal nature of our dependent variable. Residual plot CRLM is presented in figure 27. From this plot, we can see that model residuals are centred around zero, but model fit is not ideal in terms of seasonality capture since there is a clear structure present in the residual data.

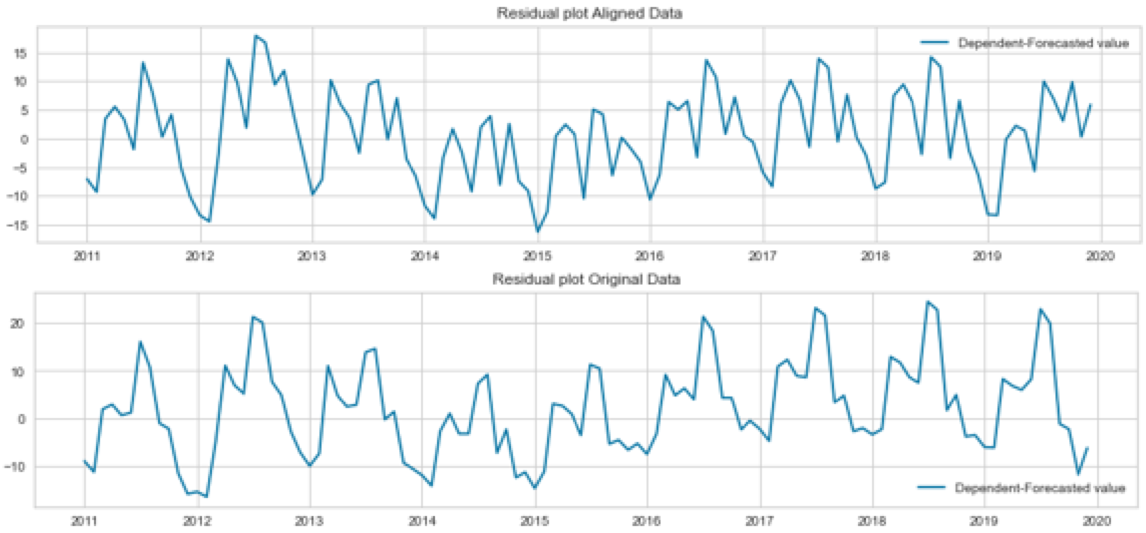


Figure 27. *Regression model residual plots shows clearly that simple CLRM is not able to fully use the seasonality pattern from the data. Aligned CLRM model residuals captures peaks better than non-aligned but still, there is a clear data pattern present in the residuals.*

Like seen from the table 6 below, CLRM model coefficients change if we shift the exp_var1 to the right to align the series with dependent variable seasonality. This affects both the intercept and the coefficient of the model. Based on the test statistics coefficients are statistically significant. When we compare the residual data with previous models, we can see that the variation of the model is quite high, and the seasonal trend of the original dataset overspills the linear model approximation on both sides of the residual distribution.

Table 6. *CLRM model coefficients and R^2 fit values.*

Model & Coefficient		Coef. Value	Std. errs.	p>|z|
CLRM	X	6.72 e-4	4.31 e-5	0.000
	Intercept	54.0	10.10	0.000
CLRM (Seq. shift 4 mo.)	X	7.32 e-4	3.62 e-5	0.000
	Intercept	38.6	8.56	0.000

Model	R^2	Residual Mean	Residual Std.	Jarque-Bera	Goldfield-Quandt	Durbin-Watson
CLRM (Unmodified)	0.696	-1.92 e-14	9.64	4.09 13.0%	1.54	0.75
CLRM (Seq. shift 4 mo.)	0.794	1-22 e-13	7.95	2.60 27.2%	1.37	1.10

The regression model forecasting results can be seen from table 7. Based on relatively good R^2 value, MAPE values of 3.76% and 2.84%, and aligned fit seen in figure 26, we can confirm that the regression model is an acceptable approximation for the dataset. However, there is improvement potential with the standard linear regression model. This observation was confirmed by the Durbin-Watson test which shows clear evidence of autocorrelation structure that could be used to improve the model. This observation can also be confirmed from figure 27 where we can see a clear seasonal structure that could be used to improve model fit.

4.4.1. Experimental explanatory variable models

To enhance the limited linear regression model capability, we can generate a combination model where we try to capture the trend from exp_var1 and augment this trend with a standardized seasonality pattern from a dependent variable training set. The seasonality pattern is captured by normalizing all dependent variable seasonal patterns and calculating the mean pattern out of these annual variations. This seasonal pattern is then combined with the annual growth rate derived from the 2011-2016 exp_var1 time series. The actual forecasting accuracy is calculated with the 2017-2019 period to match previous methods calculation. This combined model is presented in figure 28 below.

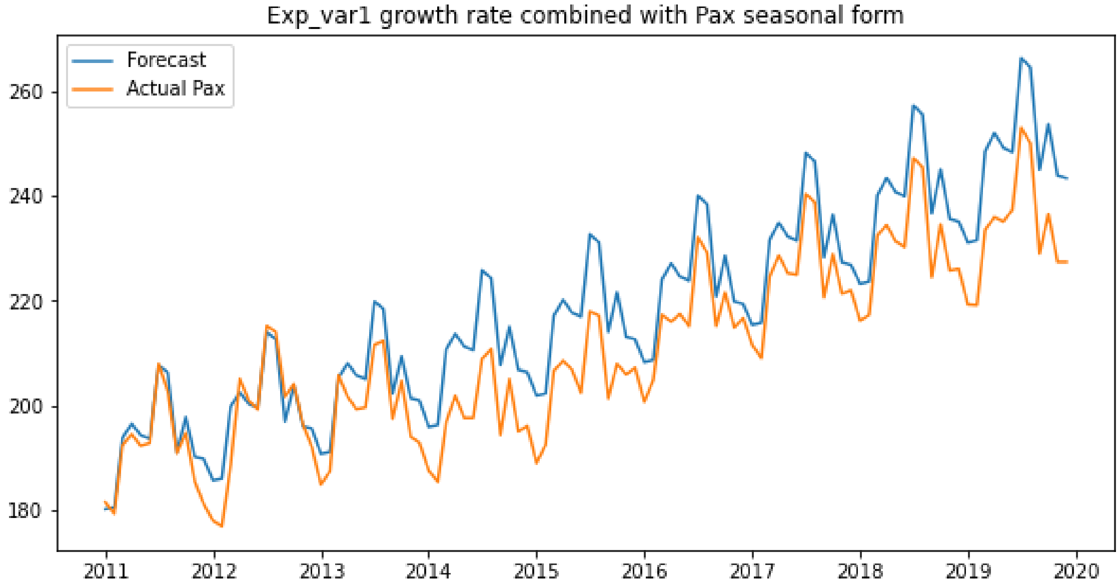


Figure 28. *Compounding growth over time increases the multiplicative nature of seasonality.*

From the figure 28 we can see, that over time the seasonal form captured from the dependent variable training set is compounding over continuing growth figures. This stretches the end part of the forecasting horizon decreasing the model accuracy. Exp_var1 has also a more consistent growth trend over time whereas 2012-2014 differs from this trend with our dependent variable. Observing this overestimating forecast in figure 28 raises the question that what would happen if we would use this growth tracking method from 2014 onwards with more consistent growth? This question is answered in chapter 4.5.5 when we limit the data availability and recalculate all the forecast models with a shorter training set.

The over estimation seen in figure 28 also points out that trend prediction is critical when evaluating forecasting methods since the original combination model underperforms the CLRM model derived as the first method in this chapter although it included a seasonal component. This can be observed from table 7 below.

If we evaluate the combined explanatory model from the residual data perspective, we can confirm that the residual data has no clear seasonal structure left like normal CLRM residuals had. The amplitude of residual plot is reduced and there is no clear repeating pattern left in the residuals (Appendix 8). However, the model residuals are not normally distributed, and residual have a mean value of 4.99. This is caused by the lack of regression optimization method like OLS that would align the seasonality pattern with the training set growth trend.

Table 7. *Regression model results*

Model	MSE	MAPE
Naïve Forecasting (Benchmark)	**39.82**	**2.48%**
CLRM (Unmodified)	134.05	3.76%
CLRM (Sequence shift 4 months)	60.28	2.84%
Norm. y seasonality with Explanatory trend tracking	96.91	4.12%

As a conclusion, we can confirm that the CLRM model is suitable for capturing even seasonal time series data, but its accuracy is highly limited since it is not able to capture the positive autocorrelation structure present in seasonal data. This disadvantage can be mitigated with alternative explanatory variable models where we use the seasonality structure with the explanatory trend prediction. However, from the alternative model, we must highlight that it is highly affected by the trend prediction tracking of the exp_var1 and the correlation continuity between dependent and explanatory variables. The regression models are also affected by the limited data frequency for the exp_var1 data. Based on these results we can say that the linear regression model shows its strengths by generating an easily understandable although not precise estimate of the time series development. The overview comparison between evaluated model families can be seen in chapter 5 table 13.

4.5. Forecasting Model Robustness in Turning Points

During previous chapters, we have implemented 4 families of models: exponential smoothing based, combined ARIMA models, neural networks, and linear regression models. The forecasting was based on 2011-2019 aviation demand data which was proven to be highly stable with clear seasonal and a trend component included in the data. In this chapter, we will deepen our understanding of forecasting models by including demand data from 2020-2021 which includes the biggest disruptions in aviation history due to the global COVID-19 pandemic. Dataset with the included 2020-2021 data is referred to as a "disruptive environment" with the following method implementation.

The focus of this chapter is to evaluate model performance from the perspective of robustness. We are especially interested to understand how our researched models adapt to disruptions and does this robustness perspective redefine our relative superiority order between prediction models. In this chapter, we will also seek an answer to our third research question related to forecasting capabilities in a disruptive environment with 1-year forecasting horizon.

We will start with a similar exploratory data analysis conducted at the beginning of chapter 4. This time the focus is on analysing the 2020-2021 period. Based on the data displayed in figure 30 we can see that aviation demand has collapsed in 2020. This disruption has continued in 2021 although the demand has partially recovered from the sudden decrease experienced during Q2/2020. We can also see that the seasonal trend can still be isolated from the data during 2020 and 2021, although its impact is heavily supressed in disruptive conditions. This indication supports the testing of seasonal models also in disruptive conditions.

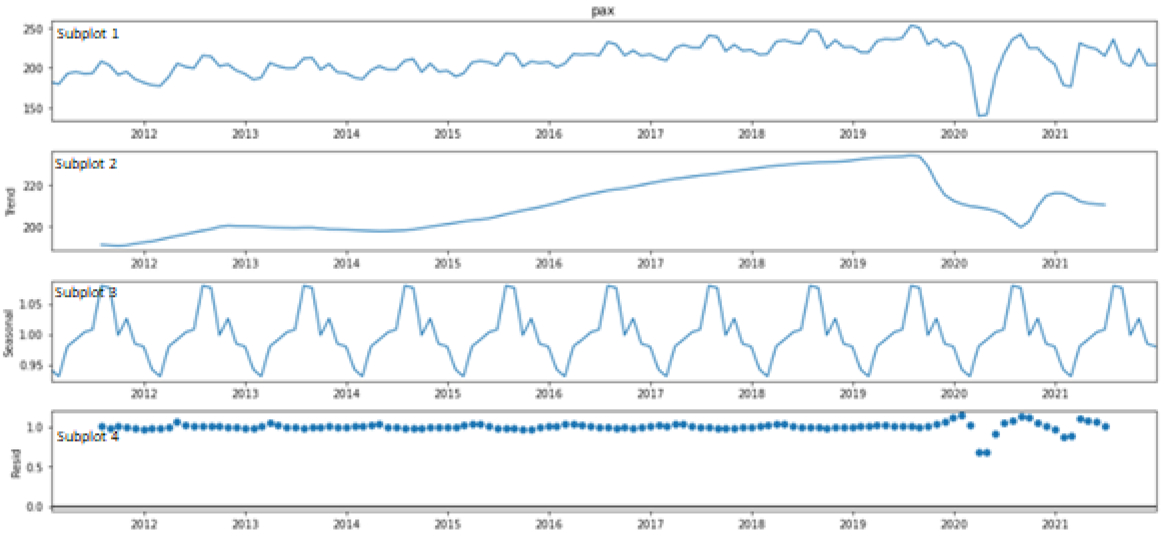


Figure 29. *Exploratory data analysis shows that normally stable aviation data has been disrupted seriously during 2020-2021.*

In the second research question, we used naïve forecast as a benchmark forecasting technique. This is repeated with disruptive data but with the original and a modified format. The original standard naïve forecast can be seen in figure 31. As seen from the figure, the standard naïve forecast does not perform well when it is using previous years disruption (2020) as the next years prediction (2021). From the decision-making perspective, it is also unreasonable to expect that if decision-maker experiences significant demand shock during global pandemic in 2020, that this exceptional situation is expected repeat in 2021 with similar demand shape and seasonal timing. Therefore, we can argue that in this kind of decision-making scenario we would use 2019 demand as a secondary naïve forecast to compare against in 2021. This scenario is visible from figure 32.

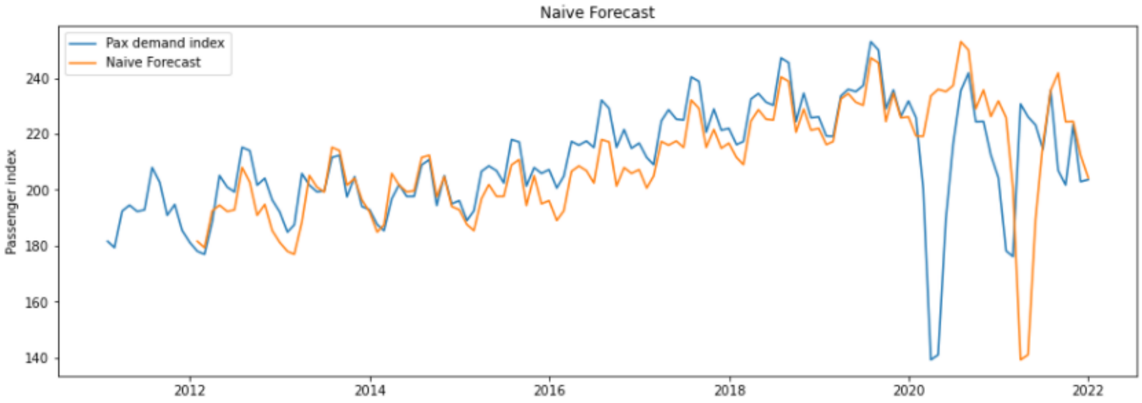


Figure 30. *Standard naïve forecast over the entire dataset.*

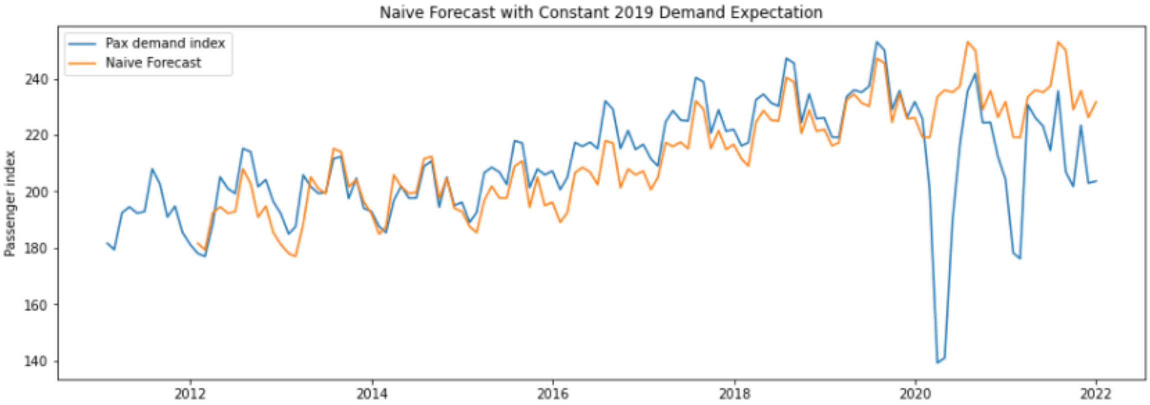


Figure 31. *Naïve forecast using 2019 demand for both 2020 and 2021 prediction.*

The main idea of the third research question is that we want to know which models would perform relatively well even in difficult conditions. Based on figures 31 and 32 and calculated accuracies we can say that naïve forecasting operates well under constant conditions but lacks capability to adapt to market changes during difficult conditions. MAPE accuracy for original naïve is 13.94% and for constant 2019 demand-based naïve 11.79%. These and MSE accuracies can be seen in table 7 below.

4.5.1. Exponential Smoothing

In this chapter, we are evaluating the forecasting performance of exponential smoothing-based methods in a disruptive demand environment. We will tune the model configurations used in chapter 4.1 with the same process and include all model variations. The first model to test in disruptive conditions is the simple exponential smoothing which can be seen in figure 33.

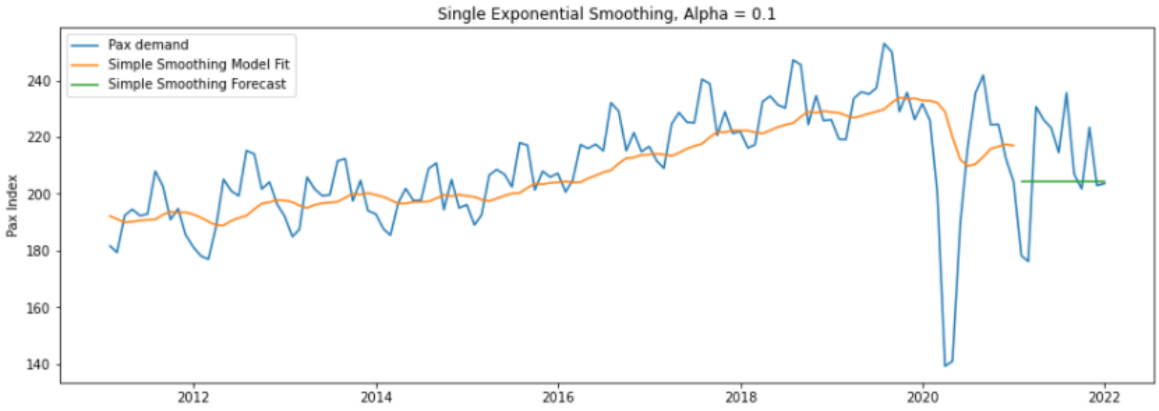


Figure 32. *Simple exponential smoothing forecast a constant average trend for 2021.*

Simple exponential smoothing can predict a constant average number across the forecasting horizon. This is a rough estimate and its largely dependent on the smoothing constant tuning during the model training. For the double exponential smoothing, that trend can have a changing slope like we saw in chapter 4.1. This forecast can be seen in figure 34.

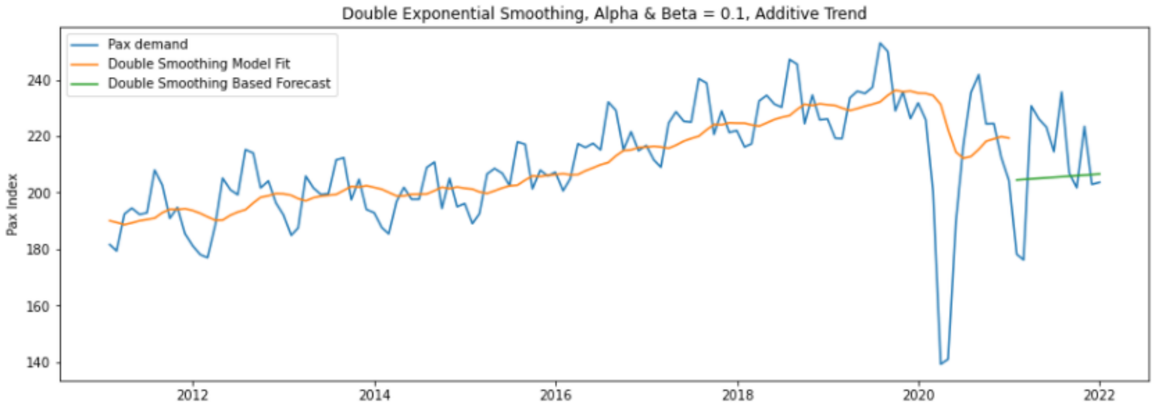


Figure 33. *Double exponential smoothing also called as Holt's method.*

For the double and triple smoothing, we have both additive and multiplicative iterations for the trend component. In addition to this, we have a seasonal component to adjust in triple exponential smoothing. Measured forecasting performance of these variations can be seen in table 7 below. In figure 35 we see the same model iteration that was the best performer in a constant operating environment. This was the triple smoothing with additive trend and multiplicative seasonality that were calculated using equations 7-10 described in chapter 3.3.

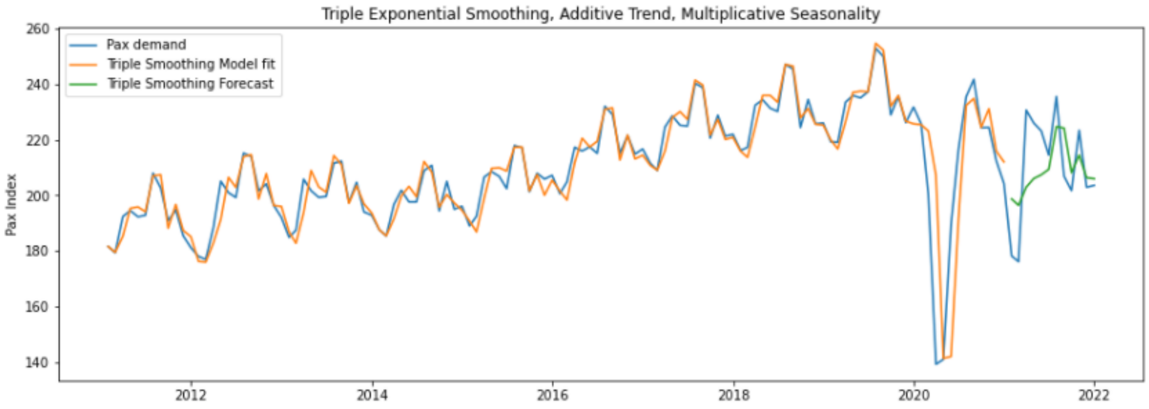


Figure 34. *Triple exponential smoothing is also called as Holt-Winters method.*

The performance of triple exponential smoothing is quite similar across all model iterations. In general, exponential smoothing performance is heavily affected by the selection of the smoothing constants in a disruptive environment. Usage of large smoothing constants makes smoothing results highly similar to the last values of the training set. In disruptive conditions, this generates forecasts which have larger variation level and are more random, since consecutive monthly demand values have larger differences between them unlike in normal conditions. This behaviour is especially visible in simple and double smoothing simulations due to their lack of capability to capture seasonal trends, and therefore for the testing we have used the same 0.1 smoothing constants.

Table 7. *Model performance in disruptive conditions including naïve model and all iterations of exponential smoothing.*

Method	MSE	MAPE
Naïve Forecasting (Benchmark)	**1783.73**	**13.94%**
Naïve Fore. constant 2019 prediction	**729.19**	**11.79%**
Simple Exponential Smoothing	372.51	7.53%
Double Smoothing, Trend = Multiplicative (Mul.)	389.81	7.68%
Double Smoothing, Trend = Additive (Add.)	354.57	7.50%
Triple Smoothing, Trend = Mul., Seasonal = Mul.	242.68	6.28%
Triple Smoothing, Trend = Mul., Seasonal = Add.	238.71	6.20%
Triple Smoothing, Trend = Add., Seasonal = Add.	234.09	6.29%
Triple Smoothing, Trend = Add., Seasonal = Mul.	235.92	6.36%

Triple exponential smoothing with multiplicative trend and additive seasonal component has the best exponential smoothing based performance in disruptive conditions, although the differences between triple exponential smoothing variants are small. In the results, we see variant superiority difference between methods if we evaluate with MSE instead of MAPE. Triple exponential smoothing with additive trend and additive seasonality has lower MSE but higher MAPE value when compared to triple exponential smoothing with multiplicative trend and multiplicative seasonality. This difference was resulting from error squaring in MSE calculation that we discussed in chapter 4.3.2 with table 5 results.

4.5.2. ARIMA & SARIMA

Utilizing ARIMA and SARIMA models under the disruptive environment requires retuning of the model parameters. This process is the same that we used with original ARIMA and SARIMA model implementation. First, we test the dataset for stationarity with the Augmented Dickey-Fuller test. After this, we will introduce stationarity be detrending the data and test with ADF again to make sure the data is stationary without higher unit-roots.

When data is stationary, we will plot the ACF and PACF plots to indicate possible model orders. This visual interpretation is confirmed with AIC testing with each model order combination. Based on this testing which can be reviewed in Appendix 9, we will choose ARIMA(12,1,12) and SARIMA(0,1,4)(1,1,2,12) models to be fitted. Model is trained with the dataset including 2011-2020 data and tested with 2021 data. Forecasting results based on the model fitting can be seen figures 36 and 37.

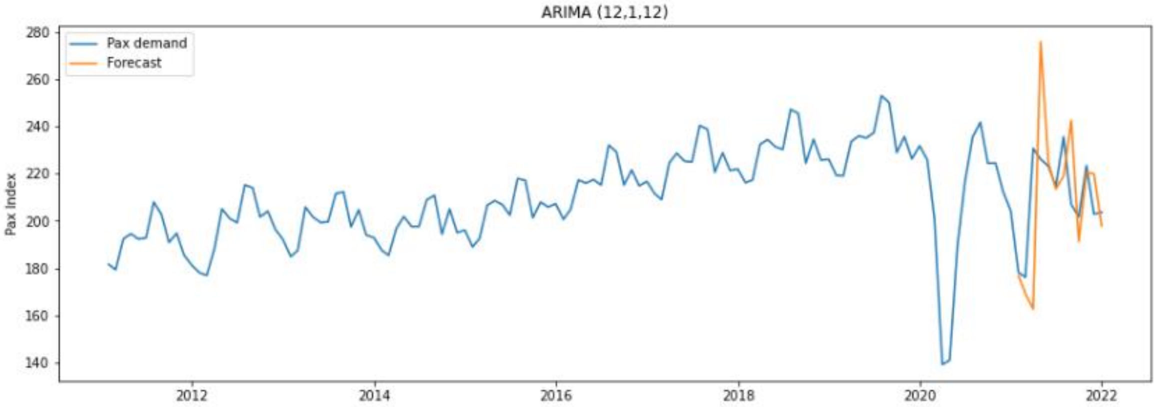


Figure 35. *ARIMA(12,1,12) model forecast under disruptive conditions*

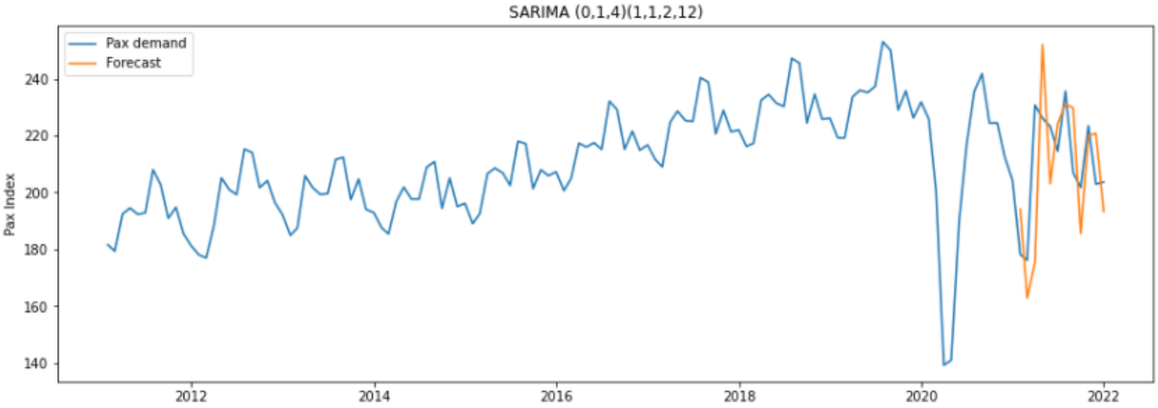


Figure 36. *SARIMA(0,1,4)(1,1,2,12) model forecast under disruptive conditions*

Visually model fitting is quite similar with both models. Model fitting can also be evaluated by conducting residual autocorrelation testing with Durbin-Watson. Model fit is also diagnosed using residual diagnostics which can be seen in Appendix 9. If we compare these residual diagnostics results with the original residual results seen in chapter 4.2, we can see that the model fit has decreased significantly. This increased amount of noise in the 2020-2021 data is also visible in the Durbin-Watson test results seen in table 8. We see that both models are less neutral in terms of autocorrelation structure when compared to the original ARIMA and SARIMA models.

Table 8. *Model forecasting accuracy and residual autocorrelation structure with disruptive data.*

Method	MSE	MAPE	Durbin-Watson test statistic
Naïve Forecasting (Benchmark)	**1783.73**	**13.94%**	0.93
Naïve Fore. constant 2019 prediction	**729.19**	**11.79%**	0.36
ARIMA (12,1,12)	760.33	8.27%	2.08
SARIMA (0,1,4)(1,1,2,12)	490.76	8.48%	2.42

From the selected models we can see that the model complexity increases, and the accuracy decreases with the use in the disruptive environment. This increased noise level is also visible with model coefficients where all ARIMA model coefficients are statistically insignificant. When reviewing model options, we can point out that ARIMA(12,1,14) achieves the lowest AIC value in initial model fit testing. However, when we generate predictions using this model and analyse residual diagnostics, we see that the simpler ARIMA(12,1,12) model has better overall performance. In general, we can say that the ARIMA and SARIMA models are showing the relative capability to adapt to the new conditions when compared with the naïve method, but the overall performance is still modest, especially when considering the increased complexity of the models.

4.5.3. Neural Networks

Neural network performance under disruptive conditions is our next evaluation case. In this part, we are reusing the RNN and ANN models that were created and tuned during the initial model fitting and forecasting. As with previous techniques, we will use 2011-2020 as a training set and try to predict the 2021 data. In plotting, we are displaying only training from 2017 onwards to increase the plot readability without size increase. We will start with the normal artificial neural networks that we created previously in chapter 4.3.1. We can see the models fitted and tested in figure 38.

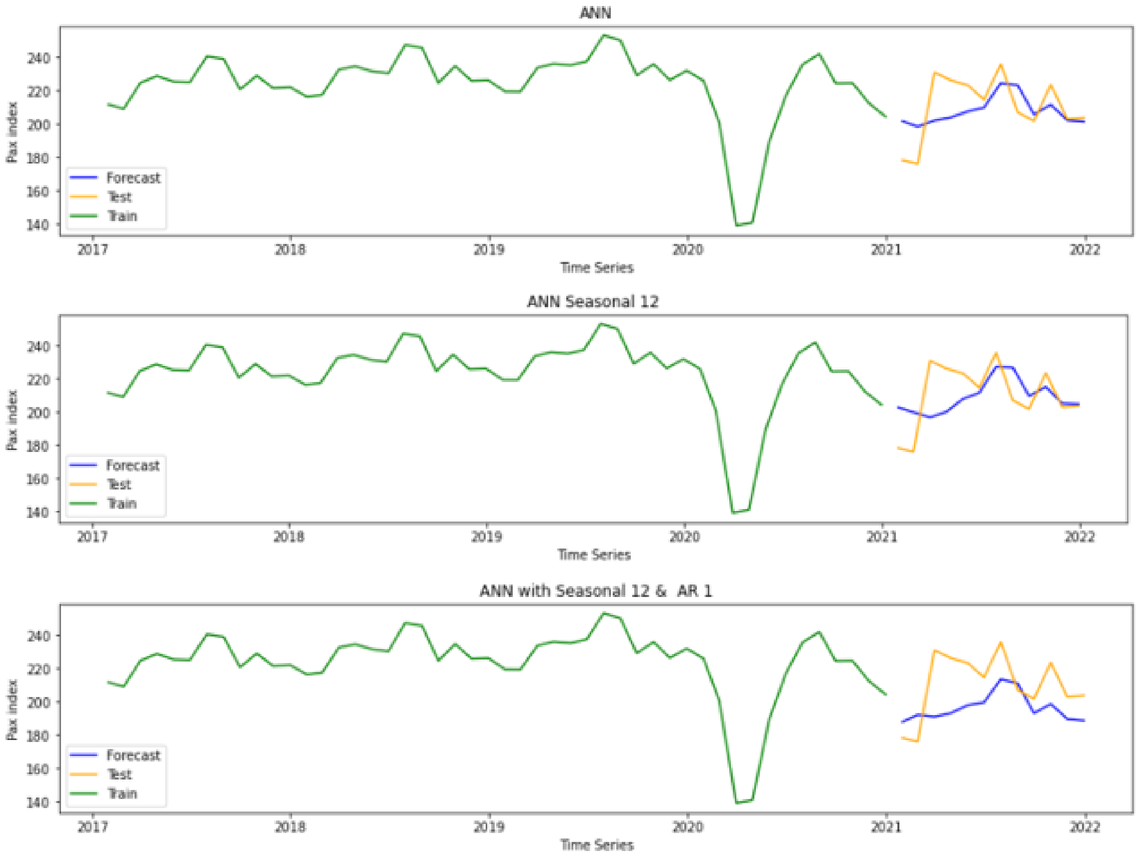
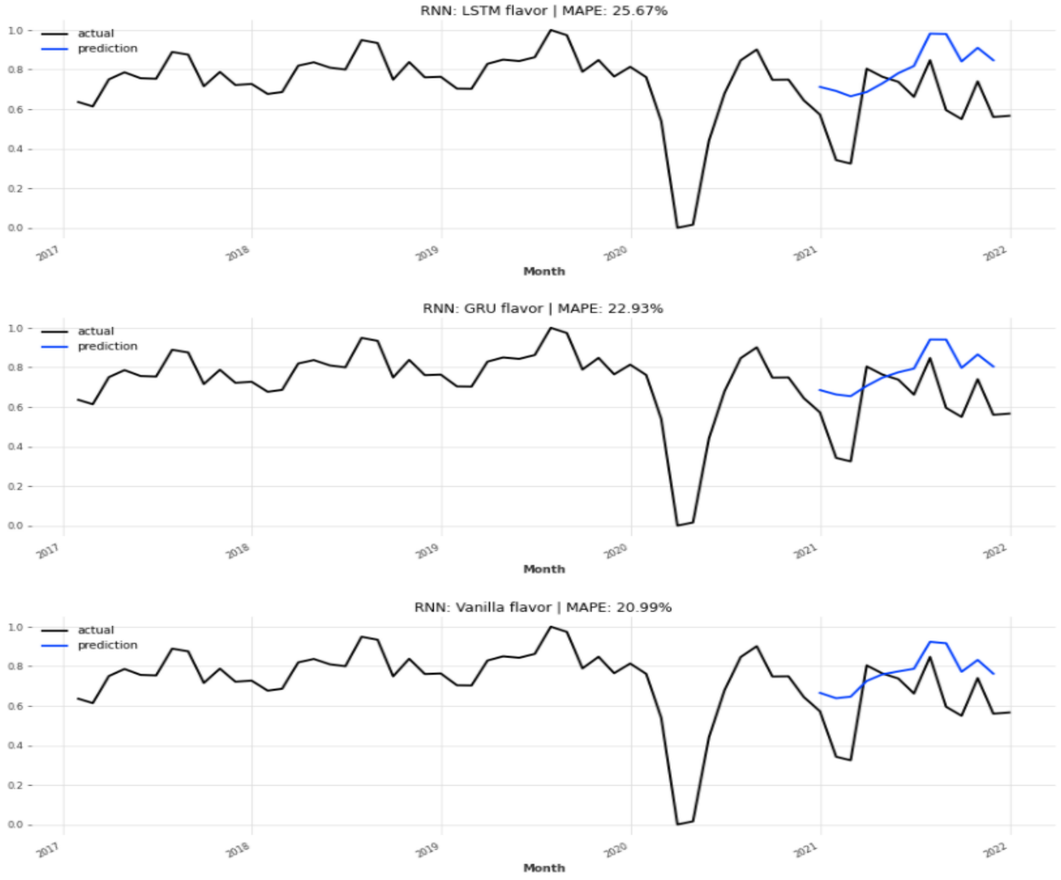


Figure 37. *Neural network iterations with disruptive data.*

From figure 38 we can see that unlike with the original data, providing information for the neural network does not improve the model fitting. This lack of seasonal or autocorrelation structure is also present in the residual data plots in Appendix 10. From residuals distributions, we can also see that the simple ANN without any characterized information or autocorrelation structure has the relatively best fit. This can be also reviewed from table 9 at the end of this chapter.

The model training is repeated with the recurrent neural network models. In the initial dataset, it was the case that LTSM and GRU were not able to benefit from their more complex memory capabilities due to the limited sample size and strong seasonality nature of the data. When a model is trained with the disruptive data and 1-year forecasting horizon RNN models are showing similar behaviour but with significantly lower accuracy which can be seen from figure 39.


RNN: LSTM flavor | MAPE: 25.67%

RNN: GRU flavor | MAPE: 22.93%

RNN: Vanilla flavor | MAPE: 20.99%

Figure 38. *Recurrent neural network performance under disruptive conditions.*

We were able to improve the model performance by increasing the learning rate of the network due to the short training period in the disruptive environment. However, the performance remains still modest compared to benchmark performance. Comparative results with neural network can be seen in table 9 below.

Table 9. *Neural network forecasting results in disruptive environment.*

Method	MSE	MAPE
Naïve Forecasting (Benchmark)	**1783.73**	**13.94%**
Naïve Fore. constant 2019 prediction	**729.19**	**11.79%**
ANN (Standard Settings, No info)	263.11	6.49%
ANN (Seasonal 12)	290.31	6.72%
ANN (Seasonal 12, Autoregressive)	519.71	9.33%
RNN (Vanilla)	-	20.99%
RNN (Recurrent Gated Unit) (GRU)	-	22.93%
RNN (Long-Term Short-Term Memory) (LTSM)	-	25.67%

Neural network performance with ANN and RNN variants differs significantly with disruptive dataset. Untrained ANN has relatively strong performance when compared to benchmark performance whereas RNN forecasting capability is highly limited when compared to ANN or with benchmark accuracy.

4.5.4. Linear Regression

Finally, we are going to use linear regression-based models with disruptive data forecasting. The initial dataset had a clear growth trend and a seasonality structure that our model was able to track with reasonable accuracy. Based on the theoretical understanding gathered in chapter 3.6. we know that OLS is heavily affected by outliers. This also means that CLRM coefficients are affected by 2020 data values which are part of the disruptive training set. In figure 40 we can see the new fitted linear regression line with the data points.

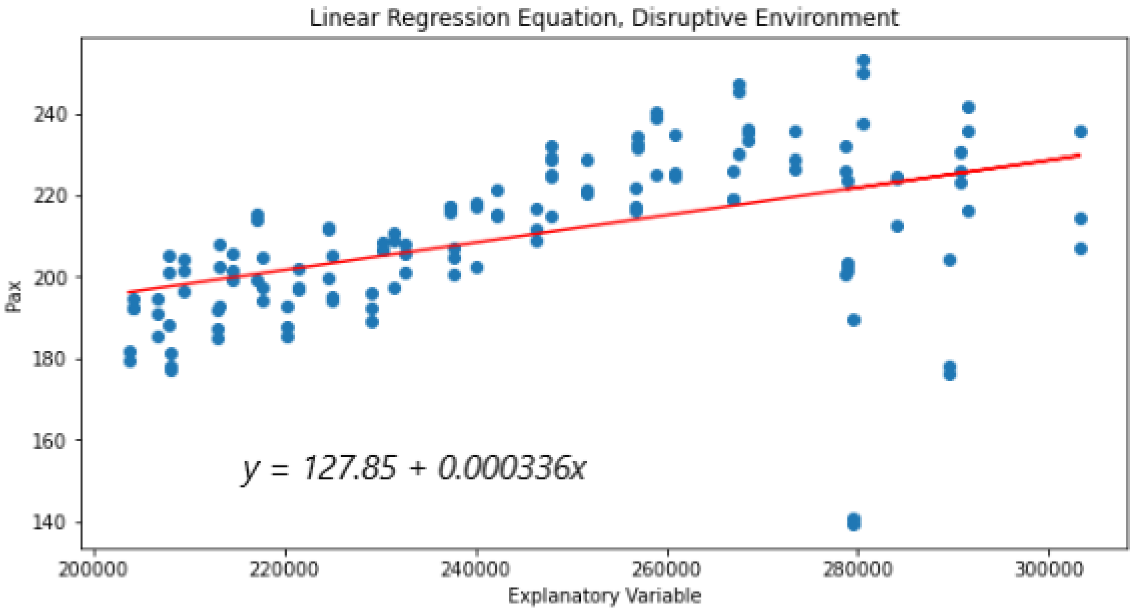


Figure 39. *Linear regression equation fitted into the disruptive dataset.*

As we see from figure 40, the training set is now underfitting the previous growth trend and tries to count in the significant demand drop which causes the 2020 data points to have a significant distance between the regression line and data points. Since we are using OLS based regression line fitting, these individual data points get significant weight in OLS optimization because of differencing distances are squared into a minimum sum. After deriving the coefficients based on this optimization, we are using them to forecast future demand based on exp_var1 values. In figure 41 we can see the model fit with aligned exp_var1 based CLRM.

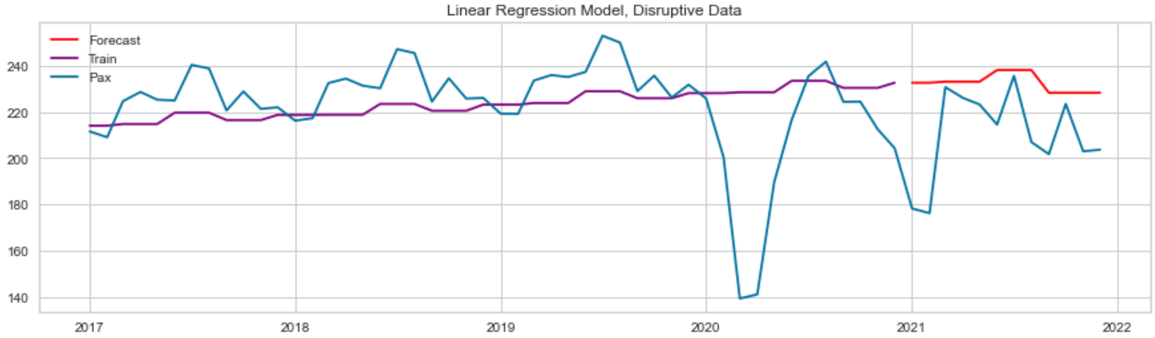


Figure 40. *Linear regression model fitted into disruptive data.*

We can see the test statistics regarding model fit in table 10. Based on these results we can see a significant decrease in model fit. Squared *R*-value is decreased with both models although aligned CLRM still achieves a better model fit. Based on the Goldfield-Quandt test there is also heteroskedasticity present in the dataset. Durbin-Watson test indicates that there is a strong autocorrelation structure included into the residual data. Jarque-Bera indicates that the data is strongly non-normally distributed. From the residual test statistics, we can also see that the variation in residual data has increased significantly. Based on these statistics we can conclude that the linear regression model does not provide an effective model fit in the case of a disruptive dataset.

Table 10. *CLRM results in disruptive conditions.*

<table>
<tr><th colspan="2">Model &amp; Coefficient</th><th colspan="2">Coef. Value</th><th>Std. err</th><th colspan="3">p&gt;|z|</th></tr>
<tr><td rowspan="2">CLRM</td><td>X</td><td colspan="2">2.95 e-4</td><td>5.36 e-5</td><td colspan="3">0.000</td></tr>
<tr><td>Intercept</td><td colspan="2">138.45</td><td>13.12</td><td colspan="3">0.000</td></tr>
<tr><td rowspan="2">CLRM
(Seq. shift 4 mo.)</td><td>X</td><td colspan="2">3.36 e-4</td><td>5.25 e-5</td><td colspan="3">0.000</td></tr>
<tr><td>Intercept</td><td colspan="2">127.85</td><td>12.96</td><td colspan="3">0.000</td></tr>
<tr><th colspan="2">Model</th><th>R^2</th><th>Residual Mean</th><th>Residual Std.</th><th>Jarque-Bera</th><th>Goldfield-Quandt</th><th>Durbin-Watson</th></tr>
<tr><td colspan="2">CLRM (Unmodified)</td><td>0.189</td><td>1.61 e-13</td><td>17.52</td><td>203.6
6.3e-45%</td><td>17.46</td><td>0.55</td></tr>
<tr><td colspan="2">CLRM (Seq. shift 4 mo.)</td><td>0.239</td><td>6.03 e-14</td><td>16.96</td><td>347.4
3.7e-76%</td><td>19.51</td><td>0.58</td></tr>
</table>

In figure 42 we can see the trend tracking explanatory variable-based forecast for the disruptive period. From the figure we can see that trend tracking is not able to follow dependent variable changes. Forecasted demand values which are based on explanatory variable growth keeps on growing, and therefore the forecasting error between forecasted and actual demand is substantial. The model accuracy is measured from the matching 2021 data period.

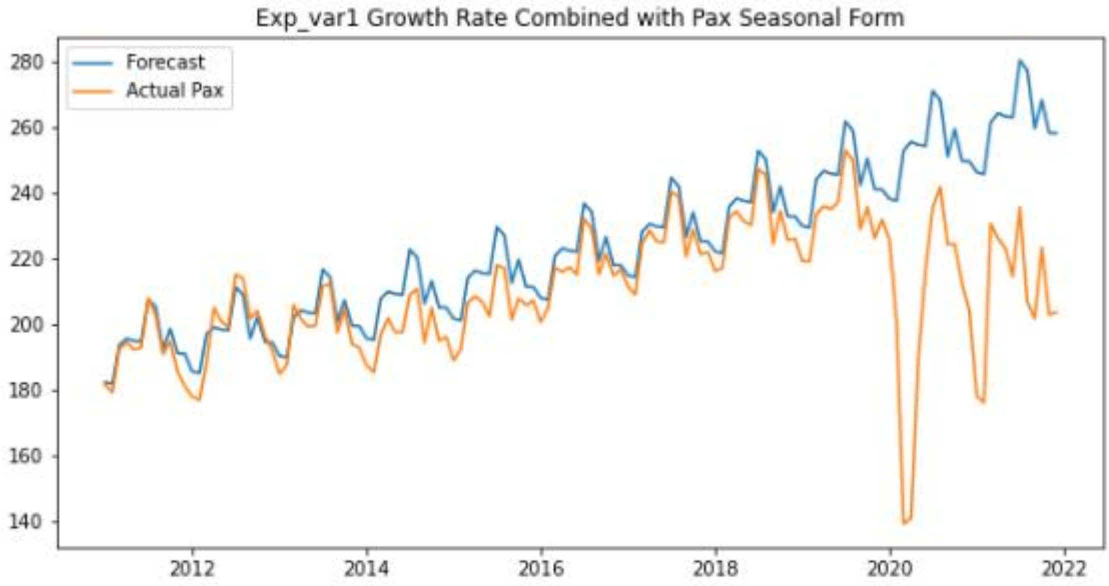


Figure 41. *Explanatory variable trend tracking-based forecasting fails to adapt to disruptive conditions.*

With this model, the residual values are not normally distributed. Residual values have a mean of 9.71 and a standard deviation of 20.86. The residual plot can be seen from Appendix 11. Based on the model fit we can say that trend tracking with a fixed seasonality format is not effective to operate in disruptive conditions. Model accuracies can be found from table 11.

Table 11. *CLRM forecast accuracies under disruptive conditions.*

Model	MSE	MAPE
Naïve Forecasting (Benchmark)	**1783.73**	**13.94%**
Naïve Fore. constant 2019 prediction	**729.19**	**11.79%**
CLRM (Unmodified)	859.98	12.07%
CLRM (Sequence shift 4 months)	817.55	11.51%
Norm. y seasonality with Explanatory trend tracking	2341.11	22.87%

In overall regression models seem to be reasonably good predictors under stable low outlier conditions. When these models are used in disruptive conditions the OLS model fitting skews the data towards the outliers, therefore decreasing the fit for the normal data points. It is also important to note that regression model accuracy depends on the continuity of the relationship between dependent and independent variables.

4.5.5. Forecasting Performance with Reduced Data Availability.

In chapter 3.2 we discussed that our initial dataset of 108 samples should be sufficient for utilizing our selected models from a method usability standpoint. This raises the question that what happens to our forecasting performance if we use these models with less data. To answer this question, we are going to retrain the models for both stable and disruptive conditions and review their accuracies when used with a smaller data sample. This chapter aims also to highlight strengths and weaknesses between methods since changing data availability and forecast training set is another perspective for forecast model robustness observation.

For stable conditions, we are using a dataset from 2014-2016 to train the models and the same 3-year forecasting horizon from 2017-2019. For the disruptive period we are using data between 2017-2020 to train the models with the 1-year forecasting horizon. In practice, this means 3-year or 50% reduction in training data with the original dataset and 6-year or 60% reduction in disruptive conditions. For both cases, we are using models which were selected based on their performance in earlier implementation chapters. To simplify the comparison, we are focusing on the MAPE accuracy of the methods. Method accuracies with limited subsamples can be seen in table 12. The table also includes a comparative column for both conditions where model performance is compared with original full dataset and limited dataset. In these columns results are presented in percentage point changes to MAPE.

Table 12. *All methods tested with limited data amount. Comparison column describes the forecasting performance change from full to limited data and is measured with MAPE percentage point change. Therefore, increasing figures are negative and signed with a plus and decreasing forecasting error with a minus. Significant changes are highlighted with colour, and changes caused by the subsampling are described in chapter 5.2.*

Method	MAPE Normal conditions (2014-2019)	MAPE pp-% diff. (2011-2019) vs. (2014-2019)	MAPE Disruptive conditions (2017-2021)	MAPE pp-% diff. (2011-2021) vs. (2017-2021)
Naïve Forecasting	**2.48%**	**0.00%**	**13.94%**	**0.00%**
Simple Exp. Smoothing	5.67%	(-0.17%)	7.53%	(+0.00%)
Double Exp. Smo. (Best var)	3.75%	(+0.42%)	7.68%	(+0.16%)
Triple Exp. Smo. (Best var)	2.40%	(+1.73%)	7.56%	(+1.20%)
ARIMA (12,1,2) / ARIMA (12,1,12)	1.02%	(+0.1%)	9.49%	(+1.22%)
SARIMA (1,1,1)(2,1,0,12) / (0,1,4)(1,1,2,12)	0.94%	(+0.19%)	10.15%	(+1.67%)
ANN (Standard, No info)	2.89%	(+0.37%)	7.34%	(+0.85%)
ANN (Seasonal 12)	2.04%	(+0.26%)	6.57%	(-0.15%)
ANN (Seasonal 12, AR 1.)	1.91%	(+0.25%)	7.58%	(-1.75%)
RNN (Vanilla)	11.20%	(+5,57%)	26.90%	(+5.91%)
RNN (GRU)	14.14%	(+6.08%)	26.88%	(+3.98%)
RNN (LTSM)	18.53%	(+5.84%)	26.91%	(+1.24%)
CLRM (Unmodified)	4.21%	(+0.45%)	8.67%	(-3.40%)
CLRM (Seq. shift 4 months)	4.63%	(+1.79%)	9.92%	(-1.59%)
Norm. y seasonality with Explanatory trend tracking	1.67%	(-2.45%)	25.40%	(+2.53%)

When comparing these results, we can see that some accuracies have remained stable even when the data sampling has decreased. Other methods have suffered significantly from decreasing dataset size. Breaking down each significant difference seen in the table helps us to understand the fundamental principles behind each modelling family and their possible future usage cases. Differences can also highlight any forecast limitations in terms of our selected fixed forecast horizon. These performance figures are evaluated in the next section of the research regarding result discussion.

5. Results

In this chapter, we are going to analyse the modelling results and discuss the insights that were produced by the modelling. First, we are going to focus on the results created by our main modelling implementation in chapters 4.1-4.4. The second priority of this chapter is to analyse the results from our forecasting robustness study under disruptive conditions and reduced data availability in chapter 4.5.

In table 13 we can see the collected results from all used forecasting model combined into a single table. To simplify the comparison, we have only included the best performing iterations from exponential smoothing. The table also includes the model accuracies from disruptive forecasting conditions. In the second column of the table, we have highlighted the best-performing methods based on the initial forecasting problem setting. When reviewing the forecasting accuracies, it is important to understand that these results are case-specific and achieved with sample sizes of 108 and 132 samples. This means that we cannot draw strong conclusions from small accuracy differences between methods since the ranking order between similarly performing methods might change with the exclusion or inclusion of few data points. However, the accuracies are indicating, which methods have the potential for similar demand forecasting cases.

Table 13. *Forecasting models and their forecasting error-based evaluation. The best models have the smallest forecasting error.*

Method	Normal Conditions MAPE	Disruptive Cond. MAPE
Naïve Forecasting (Benchmark)	**2.48%**	**13.94%**
Simple Exponential Smoothing	5.84%	7.53%
Double Smoothing, T= Add. (Best perf. Variate)	3.33%	7.50%
Triple Smoothing. , T= Add, S= Mul. (Best perf. Var.)	**(1.) 0.67%**	6.36%
ARIMA (12,1,2) / (12,1,12)	**(2.) 0.92%**	8.27%
SARIMA (1,1,1)(2,1,0,12) / (0,1,4)(1,1,2,12)	**(3.) 0.75%**	8.48%
ANN (Standard Settings, No info)	3.26%	6.49%
ANN (Seasonal 12)	2.30%	6.72%
ANN (Seasonal 12, Autoregressive)	1.66%	9.33%
RNN (Vanilla)	5.63%	20.99%
RNN (Recurrent Gated Unit) (GRU)	8.06%	22.93%
RNN (Long-Term Short-Term Memory) (LTSM)	12.69%	25.67%
CLRM (Unmodified)	3.76%	12.07%
CLRM (Sequence shift 4 months)	2.84%	11.51%
Norm. y seasonality with Explanatory trend tracking	4.12%	22.87%

When we observe these results from table 13, we can see that best performing models in normal conditions are triple exponential smoothing, ARIMA, and SARIMA models (0.67%, 0.92%, 0.75%). Out of these models exponential smoothing achieves the highest forecasting accuracy with additive trend and multiplicative seasonality components. Triple exponential smoothing also called as Holt-Winters method can effectively capture and predict both the long-term trend development and seasonal cycle of network passenger demand. The triple exponential smoothing model is also relatively simple to calculate when compared with other models.

When reviewing ARIMA and SARIMA models we see that they are quite equal in their performance and since the difference is so small, favouring the simpler ARIMA model can be argued. The performance equality between ARIMA and SARIMA is no surprise since our data was highly seasonal which favours SARIMA, and since ARIMA uses 12 autoregressive lags effectively making it a seasonal forecast model. The overall performance of the models is strong and the residual data that we reviewed in chapter 4.2. showed that the model fit was good and without significant autocorrelation, structures left in the data.

From the table 13 we can also see that artificial neural networks are also providing good forecasting results. (1.66-3.26%) From the results, we can also see that forecasting performance increases when we tell the network additional information regarding the data characteristics. ANN provides us with a simple black box solution that requires little tuning and is still able to achieve reasonably good results.

Although recurrent neural networks are an interesting concept, they are not especially suitable for small sample size-based seasonal forecasting. This is visible in the forecasting results (5.63-12.69%) where the recurrent structure of the GRU and LTSM variants are not able to fully utilize sequential memory with our research case. This difference in performance difference between ANN and RNN shows why ANN is more used in time series forecasting and RNN is highly popular in speech and handwriting recognition.

CLRM provides a reasonable demand estimate in normal conditions (2.84-4.12%). CLRM is not able to accurately track seasonal trends, but since the explanatory variable has also cyclical nature, aligning these seasons provides us with a good overall estimate under normal conditions. The experimental model combining seasonal structure from the dependent variable training set and explanatory variable growth trend can generate consistent yet not highly accurate forecasting results (4.12%). This performance like the CLRM performance in general is highly dependent on the constant correlation structure between our dependent and independent variables. Since the dependent variable growth saw a reduction in 2012-2014, this created a small trend change resulting as a forecasting error into the experimental combination model under normal conditions.

5.1. Forecast Robustness Under Disruptive Conditions

When we observe the forecasting model results from table 13 for disruptive conditions the ranking order changes. Triple exponential smoothing is still the most consistent and best-performing method (6.36%). It can adapt and use the 2020 disruption data combined with the underlying seasonal structure to generate a reasonable estimate in a highly volatile environment. In a disruptive environment triple exponential smoothing can adapt to the change by decreasing the overall weight of the seasonal component in equation and increasing the weight of the trend and level components. The smoothing constants within these components are also increasing in disruptive conditions, therefore increasing the overall weight of the recent observations in the smoothing equation.

When we trained and tuned the ARIMA and SARIMA models with disruptive data we noticed that the fit decreases and model complexity increases with both cases. The deterministic AR process in the models is not equally effective when the data has more volatility. To improve the model fit we increased the order of stochastic moving average processes in the models, which can capture partially the high data volatility, but the performance still saw a significant decrease when compared to stable conditions (8.27%, 8.48%).

ANN models can adapt relatively well with disruptive conditions and therefore achieve relatively good performance (6.49%-9.33%) . This is no surprise since the network node structure of ANN can learn under all conditions and be fitted to various data forms. With the ANN we can also see that when we provide ANN with additional info, its capability to freely adapt into disruptive data decreases. This means that supplementing the right information for ANN improves the performance, but it also makes the ANN less robust since if the data changes and info is no longer valid, its constraints the ANN adaptability with incorrect rules. Including an AR process into ANN that increased accuracy in stable conditions is a good example of this, since in stochastic conditions it decreases the model performance by constraining ANN’s learning capability.

The similar complexity-related decrease in performance can be seen with RNN methods (20.99%-25.67%). The short sampling and high volatility do not complement RNN methods in general and like with the stable condition the best performing variant is the simplest vanilla model. Stochastic conditions, low sampling size and long-term recurring memory functionality from recurring iterations are a combination that seems to provide particularly poor results in this forecasting case.

With CLRM we can see an interesting change when non-aligned dataset-based CLRM outperforms seasonality-aligned CRLM (11.51%, 12.07%). This is caused by the OLS-based calculation. When we align the seasonality, we are moving the seasonal peaks of the explanatory variable towards the y-axis origin in a regression plot (figures 24 and 25, chapter 4.4.) This action combined with an increasing time series trend means that the aligned model has a higher regression coefficient value when compared with the non-aligned model. When we introduce the model with disruptive data both regression lines are recalculated with OLS but since aligned dataset has higher point earlier in the time series its recalculated regression line slope remains larger, therefore generating a larger forecasting error under disruptive conditions when actual passenger demand declines sharply.

Our experimental explanatory model was based on the forecasting continuing seasonal trend. Since this seasonal trend is getting supressed under high volatility figures, the forecasting performance is poor in disruptive conditions (22.87%).

With both CLRM and the experimental model we also see that the correlation relationship between dependent and explanatory variables changes with the introduction of disruptive data. Explanatory variable continues to grow steadily over 2020-2021 whereas network demand trend collapses in a disruptive environment. The constant correlation structure between dependent and independent variables was one of the CLRM assumptions covered earlier in chapter 3.6. This change and the lower frequency data with the independent variable are the main reasons why the explanatory variable models perform so poorly in disruptive conditions.

5.2. Forecast Method Performance with Limited Data

The comparative accuracy difference stated in table 12 is the percentage point difference between the full dataset and the limited subsample dataset tested in chapter 4.5.5. In this measure, negative numbers are indicating a performance increase since MAPE decreases with a more limited dataset, whereas positive numbers are indicating an increase in forecasting error when compared to the model performance with full dataset. In this chapter, we are using (best, worst) format where we announce the best and worst comparative accuracy of that revied method with limited sampling. This is done to simplify the comparison between method performance changes with limited data.

As we see from the table 12. Simple and double exponential smoothing are not significantly affected by the subsampling of the dataset (-0.17%, +0.42%). This is caused by their simple smoothing structure which is driven by the selection of smoothing constants and last training set values. Since the last values of the exponential smoothing training set have most the weight depending on the smoothing constants, cutting the training set does not affect significantly. Triple exponential smoothing sees a more significant difference in performance since its third component seasonality smoothing is affected by the cutting of a limited number of seasons (+1.20%, +1.73%). When there are less seasonal samples in the training set, the effect of single sample increases in exponential smoothing. This increase in seasonal trend variation is seen as a forecasting performance decrease with all models which utilize seasonality in their prediction algorithm.

ARIMA and SARIMA model are also utilizing this seasonality parameter since both models use autoregressive structures. These models are also more dependable on parameter tuning since subsampling data with this already quite limited dataset might change the optimum model parameters decided with information criterion evaluation. Information criterion evaluation itself is not exact since different iterations of criterion might balance model fit and complexity with different weights and with the disruptive dataset, we see a performance decrease with both models (+1.20, +1.67%).

ANN models are automatically tuning model where we can add parameter info to increase the model learning. This augmentation of data info also means that if model changes and the info is not any more accurate, the false information might have an effect into the model forecasting accuracy. With the feedforward neural network models, we must also say that since the neural network is learning differently in every training round and there is no recurring iteration loop, the forecast results are not exact. Training and testing the model 20 times and evaluating performance as a mean of these rounds gives us an estimate regarding model performance in general. ANN models seem also to be learning quicker when data sampling size is reduced in a disruptive environment, therefore keeping the performance stable and even increasing it by a small margin (-1.75%, +0.85%).

RNN models are using recurrent memory-based learning in their modelling. This memory-based operation is significantly affected by a reduction in training samples which can be seen from all RNN model accuracies in both conditions (+1.24%, +6.08%). This effect can partially be mitigated by increasing the learning rate of the RNN, but this also means that with bigger learning steps RNN pattern recognition capability might be reduced in the process.

CLRM models are derived using OLS-based regression line fitting. This is highly dependent on the model outliers and therefore is heavily affected by the subsampling of the data. We can see model performance improvements with subsampled disruptive data since with a smaller dataset the 2020 training set values get higher overall weight in OLS and therefore the coefficient slope decreases (-3.40%, -1.59%). This provides lower forecasted pax demand estimates which have lower forecasted error. This does not mean that CLRM model would be generally more capable in limited data disruptive conditions. Performance increase is a simple outcome of the increased outlier weight in OLS-based regression line calculation.

When data availability is limited in stable conditions un-aligned CRLM outperforms aligned CLRM (4.21% vs 4.63%). This is caused by fitting the model with strong continuous growth training data in 2014-2016 which over-predicts pax demand for 2017-2019, especially for 2019. When we train the original model without the alignment, the OLS-based slope is smaller, and this creates a smaller forecasting error with this limited scenario. Even though aligned CLRM model performance deteriorated more than a non-aligned model, it provides a more information-rich forecast with aligned forecasting peaks (+0.45%, +1.79%). In general perspective, it does improve forecasting capabilities when we match two similar seasonal structures in regression model forecasting.

From table 12 we can also see that reducing data size does not always decrease model fit in stable conditions. This is the case when a smaller sample has higher consistency over the fitted method. Normalized seasonality combined with explanatory variable trend tracking improves its performance in this exceptional situation since data from 2014 onwards had more consistent trend development with explanatory variable and therefore the model is able to increase its performance in this limited case (-2.45%).

These results highlight the general observation from explanatory variable model performance. The performance is tied to 2 factors: continuity of correlation relationship between the dependent and independent variable and explanatory data availability. If the correlation relationship changes during disruptions, models created with another correlation relationship are incorrect. In practice, we saw this with both CLRM and the experimental trend tracking model when explanatory values remained stable under high demand variation. To take full advantage of explanatory variable models we would also need to have indicating explanatory variable data ready for the forecasted period forehand. This enables us to track explanatory variable progress and using it generate dependable variable forecast before the actual demand data is available.

Based on the results we can say that changing dataset sampling might change optimal model parameters from a tuning perspective. As we know from the training of the models, model tuning differs with each method and therefore subsampling affects differently among forecasting techniques. Methods like vanilla ANN or simple exponential smoothing can adapt quicker to new data subsamples whereas models like SARIMA require case-specific tuning and parameter selection. Some methods like CLRM are more prone to outliers and therefore are more affected by the changes in dataset sampling. From table 14 in chapter 6.1. we can see summarized key results from this research in a verbal form.

When evaluating these forecasting results in general we can say that since the data in this study was highly seasonal the forecasted method's ability to recognize and capture this seasonal nature is critical. This key observation from the model implementation means that relatively adaptable and generic methods like recurrent neural networks are not the first option for stable seasonal trend forecasting if want to focus purely into accuracy. With the highly seasonal data, we could use methods like ARIMA, SARIMA, triple exponential smoothing, or other methods which can use the seasonal nature to their full advantage. Alternative methods like ANN models, can provide us both relative adaptability and seasonal trend tracking capability under less stable conditions.

6. Conclusions

In this final chapter of the thesis, we are going to answer the research questions, evaluate the thesis results and conclude the main findings of the research. In addition to answering the research questions, we will compare the results to the previous research regarding aviation demand forecasting, discuss the limitations of this research and raise possible future research topics that have been highlighted by this research.

6.1. Answering the Research Questions

The first objective of this research was to define the most common forecasting methods used for aviation network demand forecasting and the validity of those methods for our research case. Based on the literature research we learned that the most common models for demand forecasting in aviation were exponential smoothing models, ARIMA models, neural network structures, and regression models.

Based on the research and testing that we conducted in chapters 3 and 4 we learned that these methods can be used for our network demand forecasting research. We also learned that not all methods covered by available articles and journals were suitable for our research setting. As an example, we researched the possibility to implement CNN and ARMA models, but these iterations were not suitable for this case with a non-stationary and highly seasonal dataset. However, we did find iterations of these models like classic artificial neural network and seasonal autoregressive integrated moving average models which were usable and effective with our dataset.

In the second research question of the thesis, we asked if we could find a suitable forecasting method that can outperform standard naïve forecast in terms of MAPE accuracy in stable conditions with a 3-year forecasting horizon. If we review the research results from tables 12 and 13, we can see that we did find models from each model family that can achieve this target benchmark performance.

The best performing model for stable conditions was triple exponential smoothing with MAPE-based forecasting error of 0.67% measured from 2017-2019 testing period. Also, the

ARIMA and SARIMA models (0.92%, 0.75%) were able to deliver highly performing forecasting results that were exceeding naïve forecast capabilities in stable conditions. In addition to these models also ANN model (1.66%) with added seasonality and the autoregressive process was able to outperform the naïve forecast. Reviewing the fourth model family: CLRM models we can say that the accuracy is not matching the naïve forecast with full dataset. As discussed in chapter 5 this was due to the small differences in trend development during the training period. When we decreased the training set sampling size in chapter 4.6, we saw that the experimental model combining dependent variable seasonality with explanatory variable growth rate (1.67%) was able to outperform the naïve forecast.

The final research question of the thesis was focused to explore the effects of a global pandemic or similar disruption to demand forecasting. In this question, we used training data from 2011 to 2020 and forecasted the demand using a 12-month forecasting horizon. Based on the results we can say that all the methods we covered in this research were affected by the disruptive demand in terms of forecasting accuracy. However, like in stable conditions, the relative performance difference between methods was significant. It is no surprise that the increase in demand volatility results as an increase in demand forecasting volatility.

When we review the method performance in disruptive conditions, we see that triple exponential smoothing was still the best performing method with a MAPE of 6.36%. ARIMA and SARIMA model performance deteriorates more, and their performance is 8,27% and 8.48% respectively. When we look at other potential models, we can see that classical neural network models (6.49%) have a relatively strong performance. Vanilla ANN's relatively strong performance is resulting from its highly adaptable structure of networked neural nodes. The performance of these models is strong when compared against naïve forecasting under disruptive conditions (13.94%).

When we review RNN models we can see that the method is not best suited to this kind of forecasting with relatively small datasets of time series data. The forecasting accuracy of RNN is poor in this usage case (20.99%-25.67%). With our research RNN models are not able to utilize their recurrent memory capable loop structure which is effective in sequenced speech and writing recognition.

Explanatory data and regression-based forecast models are also inaccuracy under disruptive conditions (11.51%-22.87%) since the correlation relationship between explanatory and dependent variable changes during disruptive conditions. The finding highlights that all models which utilize explanatory variables are affected if the correlation structure between dependent and independent variables changes during the disruption period.

From a general perspective, we need to acknowledge that even though CLRM was not capable in this case the model itself could be suitable with different explanatory variable data. If data consistency and frequency remain high and the correlation relationship between independent and dependent variables stays constant in a disruptive environment, the CLRM model could also be a highly capable forecasting model.

Based on these observations we can summarize research findings into a verbally formed table. Table 14 evaluates the model performance on a relative scale within the tested methods. The performance under the limited data sampling column is focusing on the stable condition's performance. The table also includes an evaluation of parameter tuning effort with each method. This column comprises both the method parameter selection and the complexity of the method. The evaluation criteria for this summary table are visible in Appendix 12.

Table 14. *Forecasting model performance key results that have been extracted using highly seasonal demand data. In columns 2-4 high value is positive, in the last column high effort for model implementation is negative. Table focuses on the models on a general family level.*

Model Forecasting Performance Under:	Normal Conditions	Disruptive Conditions	Normal Cond. Limited Data	Implementation Complexity
Triple Exp. Smoothing	High	High	High/Med.	Low Effort
ARIMA/ SARIMA	High	High/Med.	High	High Effort
ANN	High/Med.	High/Med.	High	Med./Low Effort
RNN	Low	Low	Low	Medium Effort
CLRM	High/Med.	Low	High/Med.	Medium Effort
Exp. Var Trend Tracking	High/Med.	Low	High	Low Effort

As an overall rule from table 14 results, we can say that the best forecasting method is highly dependent on the case and its conditions and disruptions decrease the performance of all forecasting methods. However, from these results, we conclude that triple exponential smoothing is relatively strong and adaptable in most cases. We have simplified the result table to include only triple exponential smoothing since it includes also the simple and double exponential smoothing components.

Based on the table results, we can also say that ARIMA and SARIMA have good overall performance, but they are heavier to implement into practice. ANN models are a low effort and robust although not the most accurate or transparent solution. RNN models are not best suited for time series forecasting when compared with other models. Regression model performance is largely dependent on the explanatory data quality and correlation relationship consistency over time but if these are consistent, regression models can be highly effective.

6.2. Result Comparison

When we compare our research results with the previous research, we can use the same 3 perspective that was discussed in chapter 2. The primary comparison can be made with previous research regarding network demand forecasting in aviation. In a secondary perspective, we can compare results with supply chain demand forecasting. Finally, we can also compare our results in the light of demand forecasting in general AI research.

Previous research in aviation demand forecasting highlighted that demand is affected both by macro and micro-level conditions. These affecting factors are often related to regulation, travel restrictions, pricing, and the seasonality nature of the demand (Boonekamp and Riddiough 2016, 3-4.7.23). In our research, we see the results of these global macro level pandemic events with the introduction of a high level of volatility to the normally highly seasonal and stable passenger demand.

Since the global fluctuation in passenger demand is a phenomenon experienced during 2020-2022 there is little previous research to compare our disruptive condition forecasts with. If we compare our results with the stable condition forecasting conducted by Kanavos et al. in 2021 our results are concluding with the similar outcomes. Kanavos et al. pointed out that simple neural networks with AR, ARIMA, or SARIMA structures can outperform more complex options like pure deep learning neural networks. Kanavos et al. also highlighted that explanatory variable model is often less accurate in long-term time series forecasting under disruptive conditions when compared with univariate time series forecasting models.

Our research results are supporting these findings since simple time series triple exponential smoothing delivered the best result under both conditions and it outperformed ANN in all conditions.

We can also evaluate our research findings by comparing them with supply chain demand forecasting research. In their research, Ferbar et al. (2009) suggested that forecasting models that include seasonality recognition are likely to be more successful in demand forecasting. ARMA and SARIMA models were also proposed by Pongdatu and Putra (2018) in their research. Our research confirmed the good performance of exponential smoothing, ARIMA and SARIMA models. However, when extracting other forecasting models into aviation from supply chains we will need to consider the special nature of the supply chain demand forecasting that includes both the Bullwhip-Forrester effect and product, time and channel dimensions suggested by Syntetos et al (2016) in chapter 2.3.

In previous AI demand forecasting research Sezer et al. (2020) suggested that advanced models like convolutional neural network could be a viable solution for sequential time series forecasting. As a result of the research, we concluded that in this research the TCN variation of the convolutional neural network was not usable due to its context understanding driven input sequence-based memory limitations. However, we were successful in the implementation of RNN networks with both LTSM and GRU variations that were also commonly recommended by Jatin and Durga (2019, 1313-1315). Based on the results described in tables 11, 12, and 13 we can confirm that the RNN solution was usable although the method is not best suited for time series forecasting and therefore should be used in other sequential machine learning domains such as speech and writing recognition.

Additional comparison with general AI research also points out that vanilla feedforward neural network models have the strongest performance in our case, but they are one of the least hyped neural network types in current AI research. Standard ANN models are not gathering attention which would be comparable to their relatively strong performance in both terms of accuracy and robustness. The current research seems to be more focused on advances in new frontiers with experimental network methods. Based on this finding we can recommend that ANN models are a prominent technique to start when seeking a quick and well-balanced neural network forecasting model with low tuning requirements.

This research contributes to the overall aviation network demand forecasting by providing transparent and comprehensive case study results regarding the model performance in both stable and disruptive conditions. This comparative perspective of forecast performance under disruptions is a new addition to the available research since the demand disruption experienced during 2020-2022 has been unprecedented in aviation history.

Disruptive condition forecasting comparison can also be used in a case example when evaluating current forecasting practices in supply chains and operations. This research has proven that the performance difference between stable and disruptive conditions can be significant and different methods can have equal performance in normal conditions but drastically different performance in disruptive conditions. Therefore, a key finding of this research is that when implementing a forecasting model, a risk assessment including a turning point performance evaluation should be included into the model development process. In this risk assessment we can ask the following questions:

1. What kind of forecast models does your organization use in its operations and how do these models behave under turning points, disruptive periods and with limited data?
2. Should there be alternative forecasting methods developed and simulated, that are more robust in terms of performance under disruptions or with limited data?

From a greater societal importance perspective, we can say that more accurate aviation demand forecasting helps to reduce aviation ticket prices and allocate capacity into routes where consumers want to use aviation transportation services. From an environmental perspective, it also means that on average the airplanes have less free seats which transfers into a fewer total flights and emissions and therefore more efficient air travel. From the governmental perspective we can also say that since during 2020-2022 majority of the airlines have received governmental loans and support from the taxpayers, the effective use of aviation capacity and enabling of profitable business is also a responsible decision from the taxpayer's perspective.

In conclusion from previous research comparison, we can say that this research confirms that simple time series forecasting methods can be highly effective in both stable and disruptive conditions when compared to more complex models like recurrent learning neural network structures. Simple structures can be more accurate, robust, and adaptable to changing conditions or data limitations than more complex models. These more complex models may fail to recognize the seasonal structure, learn from small datasets, or adapt to quick changes in conditions.

6.3. Research Limitations

This research also has several limitations which highlight the need for additional research. First, since the research is quantitative, we can argue that increasing sampling size or repeating the method implementation with different datasets could improve our understanding of model adaptability and possible shortcomings that were not highlighted by this research. An alternative dataset could highlight model weaknesses and change the overall recommendation made by this research. This is especially true for the disruption simulation since alternative demand volatility would return different results from robustness testing.

The dataset limitations can also be viewed from the perspective of accuracy measurement. Our primary testing set was 36 samples and secondary with disruptive data was 12 samples. Especially with 12 predicted samples, we could argue that this sample size is quite limited when a single observation point determines 8.3% of the total result. Increasing the data frequency with the same data period could reduce this concern significantly.

The dataset limitation related to forecasting method accuracy measurement could also be reduced by testing the models with several slightly altered forecasting windows. Like discussed in chapter 3.2.2 this was not practical from the perspective of research objectives since the fixed forecasting window was preferred from the company perspective in stable conditions and in disruptive conditions the window selection was limited by the data availability and network planning usage case. However, taking several forecasting windows in future research and testing the model accuracies in these slightly different selections would enable us to transform the research results into more generalizable format including the mean and variation for each forecasting method and scenario.

Another key limitation is that our implemented methods require a different amount of expertise. For example, ARIMA and SARIMA require parameter tuning, testing, and evaluation utilizing Box-Jenkins or similar methodologies. This makes the models less effective for an untrained user whereas ANN with automatic settings could be easily applicable even for relatively blind research settings without extensive method understanding. These user requirements are a notable point to consider when evaluating model implementation cost and reliability in other research cases.

In the initial literature research, we also learned that demand forecasting can be often biased to methods that are proposing increased demand due to the incentive bias. Since we are evaluating these methods with symmetrical accuracy evaluation and the direction of the forecasting error has no significance in this research, our conclusions regarding forecasting method recommendations are unbiased in this regard.

In the final modelling family, we used an explanatory variable. In the result section, we already highlighted that the correlation relationship between dependent and independent variables needs to remain predictable when using explanatory variable models. Another improvement possibility with explanatory variable-based forecasting is that we could have reliable data available from the explanatory variable before the forecasted period. In practice, this could mean that we have the company's annual traveling budget known before the next calendar year and we could use this as a known explanatory variable to predict the passenger travel demand generated by this company. Often it is the case that we are using forecasted explanatory variable values to forecast dependent variable values. In this case, the model’s weakness is that disruption might change both the correlation relationship between the variables and the actual values of the explanatory variable. This is a significant limitation that needs to be accounted in the future research.

Finally, we can conclude that this research and the previous research conducted within aviation demand forecasting are only including a limited number of methods and referenced articles. This could mean that we have overlooked methods that could prove to be potential in this research case. Since the scope of single research is always finite, we hope to remedy limitations by experimenting with new models and theories in future research.

6.4. Future Research

When considering possible future research perspectives, we have already mentioned that model testing with different datasets and different data quantity could provide us with interesting insights. The research could also be expanded to include new model families which were not tested in this research. Based on the results seen in table 1, we can highlight that potential future research could include SARIMA models combined with support vector machines since support vector machines were present in both aviation demand forecasting research and general research. Based on table 1, we could also increase the amount of neural network methods in this research. A multilayer perceptron network would be a viable candidate. A third research expansion could be to test fuzzy systems and a combination of fuzzy systems and neural networks (ANFIS) as a demand forecasting model.

Changing the methods, data and its volume can also be complemented by adding different measurement methods in terms of accuracy. In this research, we evaluated the measurement accuracy to be equally valuable regardless of the prediction error direction. In many practical applications, this might not be the case and in those cases, we should use an alternative method to measure accuracy. In non-symmetrical cases, we could also evaluate the forecasting error direction with a weight prioritization that corresponds to the business impact. Measuring forecasting performance with an alternative method could include techniques like mean forecast error (MAE), mean absolute deviation (MAD) or adjusted residual mean square error (ARMSE).

As a final research perspective, I want to propose research where using a financial scale we would implement, and test forecasting methods based on their accuracy and capability to enable profitable revenue growth. The research would include also cost factor from model building and utilization. In this framework, we could test how valuable each percentage point in accuracy is in practical terms and draw a guideline on how much should we invest into demand forecasting based on the revenue impact potential and the execution costs of the forecasting. In this research, we could also identify when and if the models that were recognized in our research to be effective, would become inefficient in terms of accuracy and impact when comparing to more complex and advanced forecasting models.

References

Abu-Rayash, A., & Dincer, I. 2020. Analysis of mobility trends during the COVID-19 coronavirus pandemic: Exploring the impacts on global aviation and travel in selected cities. Energy Research & Social Science, Vol. 68. DOI: 10.1016/j.erss.2020.101693

Aladag, C., & Eğrioğlu, Erol. 2012. Advances in Time Series Forecasting. Oak Park, Ill: Bentham eBooks, Vol. 1(1), pp. 3-10. DOI: 10.2174/97816080537351120101 0003.

Alfarraj, M., & Alregib, G. 2018. Petrophysical-property estimation from seismic data using recurrent neural networks. DOI: 10.1190/segam2018-2995752.1

Alves, U., & Caetano, M. 2016. Analysis of ticket price in the airline industry from the perspective of operating costs, supply and demand. Journal of Aeronautical Sciences. Vol. 7, pp. 21-28. DOI: 10.15448/2179-703x.2016.2.23185

Athanasopoulos, G., Hyndman, R.J., Kourentzes, N., & Petropoulos, F.2017. Forecasting with temporal hierarchies. European Journal of Operational Research. Vol. 262. pp. 60-74. DOI: 10.1016/j.ejor.2017.02.046.

Baltagi., B. 2008. Econometrics. Syracuse University Center for Policy Research. DOI: 10.1007/978-3-540-76516-5

Będowska-Sójka, B. 2017. Unemployment Rates Forecasts – Unobserved Component Models Versus SARIMA Models In Central And Eastern European Countries. Comparative economic research. Central and Eastern Europe. Vol. 20(2), pp. 91–107. DOI: 10.1515/cer-2017-0014

Benoit, C., Pascal, G., & Julien, C. 2011. Forecasting World and Regional Aviation Jet Fuel Demands to the Mid-Term (2025). Energy policy. Vol. 39(9), pp. 5147–5158. Web. DOI: 10.1016/j.enpol.2011.05.049

Bianchi, F. M., Maiorino, E., Kampffmeyer, M., Rizzi, A., & Jenssen R. 2017. Recurrent Neural Networks for Short-Term Load Forecasting An Overview and Comparative Analysis. Cham: Springer International Publishing. DOI: 10.1007/978-3-319-70338-1

Boone, T., Boylan J., Fildes, R., Ganeshan, R., & Sanders, N. 2019. Perspectives on Supply Chain Forecasting. International journal of forecasting Vol. 35(1), pp. 121–127. DOI: 10.1016/j.ijforecast.2018.11.002.

Boonekamp, T., & Riddiough, H. 2016. Market stimulation of new airline routes. SEO Amsterdam Economics. Discussion Paper nr. 88. Available: https://www.seo.nl/en/publications/market-stimulation-of-new-airline-routes-2/

Boritz E., & Kennedy, D. B. 1995. Effectiveness of neural network types for prediction of business failure. Expert systems with applications. Vol. 9(4), pp. 503–512. DOI: 10.1016/0957-4174(95)00020-8

Brooks, C. 2014. Introductory Econometrics for Finance. 3rd ed. Cambridge: Cambridge University Press.

Brooks, C. 2009. RATS Handbook to Accompany Introductory Econometrics for Finance. Cambridge: Cambridge University Press, 2008. pp 22.

Carbonneau, R., Laframboise, K., & Vahidov, R. 2007. Application of Machine Learning Techniques for Supply Chain Demand Forecasting. European journal of operational research. Vol.184(3), pp. 1140–1154. DOI: 10.1016/j.ejor.2006.12.004

Charles, V., Aparicio, J., & Zhu, J. 2020. Data Science and Productivity Analytics. International Series in Operations Research & Management Science. Vol. 290. DOI: 10.1007/978-3-030-43384-0

Chicco, D., Warrens, & M., Jurman G. 2021. The Coefficient of Determination R-Squared Is More Informative Than SMAPE, MAE, MAPE, MSE and RMSE in Regression Analysis Evaluation. PeerJ. Computer science. Vol. 7. DOI: 10.7717/peerj-cs.623

Cooney, M. 2022. Cisco faces a $14B backlog thanks to component scarcity: Cisco and competitors Juniper and Arista report they have been hit hard by the chip shortage plus supply-chain issues. Trade Journal, Network World.

Colin, T., & Ruxton, G. 2010. Modelling Perception with Artificial Neural Networks. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9780511779145

Dekker, M, Donselaar, K., & Ouwehand, P. 2004. How to Use Aggregation and Combined Forecasting to Improve Seasonal Demand Forecasts. International journal of production economics. Vol. 90(2), pp. 151–167. DOI: 10.1016/j.ijpe.2004.02.004

Dumitru, C., & Vasilescu, M. 2013. Advantages and Disadvantages of Using Neural Networks for Predictions, Ovidius University Annals, Economic Sciences Series, Ovidius University of Constantza, Faculty of Economic Sciences, Vol. 1, pp. 444-449. Available: https://ideas.repec.org/a/ovi/oviste/vxiiy2012i1p444-449.html

Donoho, D. 2017. 50 Years of Data Science. Journal of Computational and Graphical Statistics. Vol. 26, Issue 4. pp. 745-766. DOI: 10.1080/10618600.2017.1384734

Fan, Y., Pastorello S., & Renault, E. 2015. "Maximization by Parts in Extremum Estimation." The econometrics journal. Vol. 18(2), pp. 147–171. DOI: 10.1111/ectj.12046

Ferbar, L., Creslovnik, D., Mojskerc, B., & Rajgelj, M. 2009. Demand forecasting methods in a supply chain: Smoothing and denoising. International journal of production economics. Vol. 118(1), pp. 49–54. DOI: 10.1016/j.ijpe.2008.08.042.

Flachaire, E & Nuñez, O. 2007. Estimation of the Income Distribution and Detection of Subpopulations: An Explanatory Model. Computational statistics & data analysis. Vol. 51, pp. 3368–3380. DOI: 10.1016/j.csda.2006.07.004

Geweke, J.,Horowitz J., & Pesaran, H. 2008, "Econometrics: A Bird's Eye View," forthcoming in The New Palgrave Dictionary, Second Edition. DOI: 10.1007/978-1-349-20570-7_1

Geweke, J. 2010. Complete and Incomplete Econometric Models. Course Book. Princeton: Princeton University Press, pp. 1-6.

Ghobbar, A. A. & Friend, C. H. 2003. Evaluation of forecasting methods for intermittent parts demand in the field of aviation: a predictive model. Computers & operations research. Vol. 30(14), pp. 2097–2114. DOI: 10.1016/S0305-0548(02)00125-9

Hewage, P., Behera, A., Trovati, M., Pereira, E., Ghahremani, M., Palmieri, F., & Liu, Y. 2020. Temporal convolutional neural (TCN) network for an effective weather forecasting using time series data from the local weather station. Soft computing (Berlin). Vol. 24(21), pp. 16453-16482. DOI: 10.1007/s00500-020-04954-0.

Hyndman, R., & Kostenko, A. 2007. Minimum sample size requirements for seasonal forecasting models. Foresight International Institute of Forecasters. Issue 6. Available: https://robjhyndman.com/papers/shortseasonal.pdf

Jatin, B., & Durga T. 2019. Deep Learning Framework to Forecast Electricity Demand." Applied energy. Vol. 238, pp. 1312–1326. DOI: 10.1016/j.apenergy.2019.01.113

Jebbor, S., Chiheb, R., & Abdellatif, E. 2022. A Preliminary Study for Selecting the Appropriate AI-Based Forecasting Model for Hospital Assets Demand Under Disasters. Journal of humanitarian logistics and supply chain management. Vol. 12(1), pp. 1–29. DOI: 10.1108/JHLSCM-12-2020-0123

Kanavos, A., Kounelis, F., Iliadis, L., & Makris, C. 2021. Deep learning models for forecasting aviation demand time series. Neural Computing & Applications, Vol. 33, pp. 16329–16343. DOI: 10.1007/s00521-021-06232-y

Kelleher, J., & Tierney, B. 2018. Data Science. The MIT Press, Massachusetts Institute of Technology. Vol. 1. pp. 1-28.

Kleinhans, J., & Hess, J. 2021. Understanding the global chip shortages, why and how the semiconductor value chain was disrupted. Stiftung Neue Verantwortung. Available: https://www.stiftung-nv.de/sites/default/files/understanding_the_global_chip_shortages.pdf

Koehler A., Snyder, R., & Ord. K. 2001. Forecasting Models and Prediction Intervals for the Multiplicative Holt–Winters Method. International journal of forecasting. Vol. 17. pp. 269–286. DOI: 10.1016/S0169-2070(01)00081-4

Koushik, R., & Ravindran Sharan. 2016. R Data Science Essentials: Learn the Essence of Data Science and Visualization Using R in No Time at All. Birmingham: Packt Publishing. pp. 101-105.

Lea, C., Vidal, R., Reiter A., & Hager G. 2016. Temporal convolutional networks: A unified approach to action segmentation. European Conference on Computer Vision. Springer, Cham. DOI: 10.48550/arXiv.1608.08242

Leary, A., 2021. Supply-Chain Crunch, Chip Shortage Focus of White House Meeting; The Wall Street journal. Eastern edition. Available: https://www.wsj.com/articles/supply-chain-crunch-chip-shortage-focus-of-white-house-meeting-11632387600

Li, T., & Trani, A. 2014. A Model to Forecast Airport-Level General Aviation Demand. Journal of air transport management Vol.40, pp. 192–206. DOI: 10.1016/j.jairtraman.2014.07.003

Li, C. 2019. Combined forecasting of civil aviation passenger volume based on ARIMA-REGRESSION. International journal of system assurance engineering and management. Vol. 10(5), pp. 945–952. DOI: 10.1007/s13198-019-00825-6

Lin, L., Zhengbing, & H., Srinivas, P. 2018. Predicting station-level hourly demand in a large-scale bike-sharing network: A graph convolutional neural network approach. Transportation research. Part C, Emerging technologies. DOI: 10.1016/j.trc.2018.10.011.

Longbing C. 2017. Data science: A comprehensive overview. ACM Computer. Survey. Vol. 50(3), Article 43 pp.1-34. DOI: 10.1145/3076253

Lu, C.-J., Chang, C.-C., & Huang, K.-N. 2014. A Hybrid Sales Forecasting Scheme by Combining Independent Component Analysis with K-Means Clustering and Support Vector Regression. TheScientificWorld. DOI: 10.1155/2014/624017.

Lässig, F. 2021. Temporal Convolutional Networks and Forecasting. Unit8. Available: https://unit8.com/resources/temporal-convolutional-networks-and-forecasting/

Mariani, P., & Zenga, M. 2021. Data Science and Social Research II. DSSR International conference on Data Science and Social Research. ISBN 978-3-030-51221-7. DOI: 10.1007/978-3-030-51222-4

Mijwil, M. 2018. Artificial Neural Networks Advantages and Disadvantages. Computer science, college of science, University of Baghdad. Available: https://www.researchgate.net/publication/323665827_Artificial_Neural_Networks_Advantages_and_Disadvantages

Moran, K., Nono, S., & Rherrad, I. 2019. Does Confidence Data Help Forecast Business Cycles? New Evidence from Canada. Applied economics Vol. 51, pp. 2289–2312. DOI: 10.1080/00036846.2018.1542119

Paldino, G., Stefani., J., De Caro, F., & Bontempi, G. 2021. Does AutoML Outperform Naïve forecasting? Engineering proceedings Vol. 5(1), DOI: 10.3390/engproc2021005036.

Pelgrin, F. 2011. Lecture 4: Estimation of ARIMA models. University of Lausanne, Department of Mathematics (IMEA-Nice). Available: https://math.unice.fr/~frapetti/CorsoP/Chapitre_4_IMEA_1.pdf

Perera, N., Hurley, J., Fahimnia, B., & Reisi, M. 2019. The human factor in supply chain forecasting: A systematic review. European journal of operational research. Vol. 274(2), pp. 574–600. DOI: 10.1016/j.ejor.2018.10.028

Peterson, J. B. 2018. 12 Rules for Life: An Antidote to Chaos. Random House Canada, Chapter 1.

Pongdatu, G., & Putra. Y. 2018. Seasonal Time Series Forecasting Using SARIMA and Holt Winter's Exponential Smoothing. IOP conference series. Materials Science and Engineering Vol. 407(1), pp. 1-6. DOI: 10.1088/1757-899X/407/1/012153

Punia, S., Singh, S., & Madaan, J. 2020. "A Cross-Temporal Hierarchical Framework and Deep Learning for Supply Chain Forecasting." Computers & industrial engineering Vol. 149, pp. 1–9. DOI: 10.1016/j.cie.2020.106796

Provost, F., & Fawcett, T. 2013. Data Science and its Relationship to Big Data and Data-Driven Decision Making. Mary Ann Liebert, Inc. Vol. 1, pp. 51-59. DOI: 10.1089/big.2013.1508

Ramírez-Montañez, J. A., Aceves-Fernandez, M. A., Pedraza-Ortega, J. C., Gorrostieta-Hurtado, E., & Sotomayor-Olmedo, A. 2021. Airborne Particulate Matter Modeling: A Comparison of Three Methods Using a Topology Performance Approach. Applied sciences. Vol. 12(1), pp. 1-20. DOI: 10.3390/app12010256

Raza, M. Q., Khosravi, A. 2015. A review on artificial intelligence-based load demand forecasting techniques for smart grid and buildings. Renewable & sustainable energy reviews. Vol. 50. pp. 1352-1372. DOI: 10.1016/j.rser.2015.04.065

Reuter, U., & Möller, B. 2010. Artificial Neural Networks for Forecasting of Fuzzy Time Series. Computer-aided civil and infrastructure engineering. Vol. 25(5), pp. 363–374. DOI: 10.1111/j.1467-8667.2009.00646.x

Reynolds, A. 2022. Convolutional Neural networks. Available: https://anhreynolds.com/blogs/cnn.html

Sadeghi, A. 2015. Providing a Measure for Bullwhip Effect in a Two-Product Supply Chain with Exponential Smoothing Forecasts. International journal of production economics. Vol. 169, pp. 44–54. DOI: 10.1016/j.ijpe.2015.07.012

Sahin, M., Kizilaslan, R., & Demirel, Ö. 2013. Forecasting Aviation Spare Parts Demand Using Croston Based Methods and Artificial Neural Networks. Journal of economic and social research. Vol. 15(2), pp. 1-21.

SAS, Statistical Analysis Systems. 2012. How can we reduce inventory costs while increasing aircraft fleet readiness? Defence and Aerospace Solution Brief. Available: https://www.sas.com/content/dam/SAS/en_us/doc/solutionbrief/defense-and-aerospace-reduce-inventory-costs-105547.pdf

Satchell, S. 2003. Introductory Econometrics for Finance. The Economic Journal 2003. Vol. 113, pp. 411–413. Web. ISSN: 0013-0133. DOI: 10.1111/1468-0297.13911

Sezer, O., Gudelek, M., & Ozbayoglu, A. 2020. Financial Time Series Forecasting with Deep Learning: A Systematic Literature Review: 2005–2019. Applied soft computing. Vol. 90(5), pp. 5271-5280. ISSN: 1568-4946. DOI: 10.1016/j.asoc.2020.106181.

Shalabh, G. 2016. MTH 676: Econometric Theory. Indian Institute of Technology Kampur, Department of Mathematics & Statistics. https://home.iitk.ac.in/~shalab/econometrics/Chapter1-Econometrics-IntroductionToEconometrics.pdf

Shih-Yao, K., Shiau, L., & Chang, Y. (2010). Air transport demand forecasting in routes network by artificial neural networks. Hangkong Taikong ji Minhang Xuekan/Journal of Aeronautics, Astronautics and Aviation, Series B. 42. 67-72.

Shuojiang, X., Chan, H., & Zhang, T. 2019. Forecasting the demand of the aviation industry using hybrid time series SARIMA-SVR approach. Transportation research. Part E, Logistics and transportation review Vol. 122, pp. 169–180. DOI: 10.1016/j.tre.2018.12.005.

Sivrikaya, O. 2013. Demand Forecasting for Domestic Air Transportation in Turkey. Okan University, Institute of Social Sciences, Business Management. Available: https://acikbilim.yok.gov.tr/bitstream/handle/20.500.12812/673458/yokAcikBilim_10002738.pdf?sequence=-1&isAllowed=y

Sousa, M., Mele P., Pesqueira, A., Rocha, A., Sousa, M., & Noor, S. 2021 Data Science Strategies Leading to the Development of Data Scientists' Skills in Organizations. Neural computing & applications Vol. 33.21 14523–14531. DOI: 10.1007/s00521-021-06095-3

Solvoll, G., Mathisen T., & Morten, W. 2020. Forecasting Air Traffic Demand for Major Infrastructure Changes. Research in transportation economics Vol. 82, Web. ISSN: 0739-8859. DOI: 10.1016/j.retrec.2020.100873

Stahlbock, R., Weiss, G., & Abou-Nasr, M. 2019. Data Science. Proceedings of the 2018 International Conference on Data Science. CSREA Press.

Stellwagen, E., & Tashman, L. 2013. ARIMA: The Models of Box and Jenkins. International Institute of Forecasters. pp. 28-33. Available: https://www.researchgate.net/publication/285902264_ARIMA_The_Models_of_Box_and_Jenkins.

Suh, D., & Ryerson, M. 2019. Forecast to grow: Aviation demand forecasting in an era of demand uncertainty and optimism bias, Transportation Research Part E: Logistics and Transportation Review, Vol. 128, pp. 400-416. DOI: 10.1016/j.tre.2019.06.016

Syntetos, A., Badai, Z., Boylan, J., Kolassa, S., & Nikolopoulos, K. 2016. Supply Chain Forecasting: Theory, Practice, Their Gap and the Future." European journal of operational research. Vol. 252(1), pp. 1–26. DOI: 10.1016/j.ejor.2015.11.010

Thompson, N., Greenewald, K., Keeheon, L., & Manso, G. 2020. The Computational Limits of Deep Learning. MIT Iniative on the digital economy research brief. Massachusetts

Institute of Technology. Vol. 4. Available: https://ide.mit.edu/wp-content/uploads/2020/09/RBN.Thompson.pdf

Tinbergen, J. 2005. Econometrics. London: Routledge. pp. 1-23. DOI: 10.4324/9780203486610

TRP Aviation Demand Forecasting. 2002. A Survey of Methgodologies. Committee on Aviation Economics and Forecasting (AIJ02). Available: https://www.trb.org/publications/circulars/ec040.pdf

Tu, J. V. 1996. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. Journal of clinical epidemiology. Vol. 49(11), pp. 1225–1231. DOI: 10.1016/S0895-4356(96)00002-9

UK Aviation Demand Forecasting Discussion Paper 01. 2013. Govermental Airports Commission. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/73143/aviation-demand-forecasting.pdf

UN Statistics. 2021. Goal 9: Build resilient infrastructure, promote inclusive and sustainable industrialization and foster innovation. United Nations, Department of Economic and Social Affairs, Statistics Division. Available: https://unstats.un.org/sdgs/report/2021/goal-09/.

Verma, J.P., & Verma, P. 2020. Determining Sample Size and Power in Research Studies A Manual for Researchers. 1st ed. Singapore: Springer Singapore, DOI: 10.1007/978-981-15-5204-5

Wadud, Z. 2011. Modeling and Forecasting Passenger Demand for a New Domestic Airport with Limited Data: Aviation 2011." Transportation research record Vol. 2214, pp. 59–68. DOI: 10.3141/2214-08

Wang, S. & Gao, Y. (2021) A literature review and citation analyses of air travel demand studies published between 2010 and 2020. Journal of air transport management. DOI: 10.1016/j.jairtraman.2021.102135

Weiheng, J., Xiaogang, W., Yi, G., Wanxin, Y., & Xinhui, Zhong. 2020. "Holt–Winters Smoothing Enhanced by Fruit Fly Optimization Algorithm to Forecast Monthly Electricity Consumption." Energy (Oxford) Vol. 193, pp. 1-8. DOI: 10.1016/j.energy.2019.116779

Wei, D., Bhardwaj, A., & Wei, J. 2018. Deep Learning Essentials: Your hands-on guide to the fundamentals of deep learning and neural network modeling. 1st edition. Packt Publishing.

Winter, E. 2021. How SAS gained 30% forecast accuracy thanks to innovative revenue management tools. Case Study. Available: https://amadeus.com/en/insights/blog/how-sas-gained-forecast-accuracy-thanks-innovative-revenue-management-tools

Xiao, Y., Liu, J., Hu, Y., Wang, Y., Lai, K., & Wang, S. 2014. A neuro-fuzzy combination model based on singular spectrum analysis for air transport demand forecasting. Journal of air transport management. Vol. 39. DOI: 10.1016/j.jairtraman.2014.03.004

Yanxin Z., Sujian, L., & Yongfang P. 2018. "Research on Duplicate Combined Forecasting Method Based on Supply Chain Coordination." Cluster computing Vol. 22, pp. 6621–6632. Web. DOI: 10.1007/s10586-018-2356-z

Zakrytnoy, S. 2021. Comparative Study of Classic and Fuzzy Time Series Models for Direct Materials Demand Forecasting. Available: https://lutpub.lut.fi/bitstream/handle/10024/162547/MaterialForecasting_MSc_szakrytnoy_0305.pdf?sequence=1

Zhang, L., & Zhang, L. 2018. Blind equalization in neural networks: theory, algorithms, and applications. Berlin, De Gruyter. DOI: 10.1515/9783110450293-202

Zhang, X. & Zhao, Y. 2010. The Impact of External Demand Information on Parallel Supply Chains with Interacting Demand. Production and operations management. Vol. 19(4), pp. 463–479. DOI: 10.1111/j.1937-5956.2009.01114.x

Appendices

Appendix 1. Literature review references

Appendix 1.1. Reference article table for selected methods.

To be accepted into this list the reference article needs to discuss or implement the method in demand forecasting. Discussing means in our context that the method is mentioned more than one in the reference article and its operating principle and limitations are explained.

Model Name	Aviation forecasting articles using method	Supply chain forecasting articles using method	AI research forecasting articles using method
Exponential Smoothing	3,7,11,12,13,15,17	18,19,21,22,23,25, 26,27,29,31	36,38
ARMA/ ARIMA/ SARIMA	3,4,8,15,17	18,19,20,21,22,25, 26,29,31	34,35,36,37
Neural Networks (ANN, RNN, CNN)	3,4,5,6,14,15,16,17	19,21,22,24,28,30, 32	33,34,35,36,37,38
Linear Regression & Explanatory var model	2,3,5,6,7,8,9,10,11, 12,13,16,17	18,19,21,22,30,32	34,35,36,37
Excluded forecasting models			
SARIMA-SVM, SVM, SVR	3,4,15,	19,28,32	34,35,36,37
Regression variations (3SLS, 2SLS, Log. reg.)	1,3,4,13,16,17	24	34,36
ANFIS, Fuzzy system	5,17	22	34,35,36,37
Neural Networks based on MLP	4,17	-	34,36,37
Neural Networks based on DBN	-	-	36
GARCH, ARCH	3,17	-	36
VAR, VARMA	15,17	20,27	-

1.(Suh and Ryerson 2019), 2. (Boonekamp and Riddiough 2016), 3. (Wang and Gao 2020), 4. (Kanavos et al. 2021), 5. (Wadud 2011), 6. (Sivrikaya 2013), 7. (Solvoll, Mathisen and Morten 2020), 8. (Li 2019), 9. (Benoit, Pascal, and Julien 2011), 10. (Li and Trani 2014), 11. (TRP 2002), 12. (Winter 2021), 13. (Ghobbar and Friend 2003), 14. (Sahin, Kizilaslan, and Demirel 2013), 15. (Shuojiang, Chan, and Zhang 2019), 16. (Shih-Yao, Shiau, and Chang 2010), 17. (Xiao et al. 2014), 18. (Syntetos 2016), 19. (Perera 2019), 20. (Zhang and Zhao 2010), 21. (Yanxin et al. 2018), 22. (Zakrytnoy 2021), 23. (Ferbar 2009), 24. (Paldino 2021), 25. (Pongdatu and Putra 2018), 26. (Dekker, Donselaar and Ouwehand 2004), 27. (Sadeghi 2015), 28. (Lu, Chang, and Huang 2014), 29. (Boone et al. 2019), 30. (Punia, Singh and Madaan 2020), 31. (Athanasopoulos 2017), 32. (Carbonneau, Laframboise, and Vahidov 2008), 33. (Wei, Bhardwaj, and Wei 2018), 34. (Raza and Khosravi 2015), 35. (Jebbor, Chiheb, and Abdellatif 2022), 36. (Sezer, Gudelek and Murat, 2020), 37. (Jatin and Durga 2019), 38. (Reuter and Möller 2010)

Appendix 1.2. Search word table for literature review references

2.1 Data Science and Econometrics	2.2 Aviation demand forecasting	2.3 Supply chain demand forecasting	2.4 AI demand forecasting
Data science, data science history, data science development, econometrics, econometrics history, econometrics relationship modelling, variable relationship modelling	Aviation demand forecasting, aviation demand, aviation forecasting, passenger demand forecasting, aviation passenger forecasting, aviation seasonal demand forecasting	Supply chain demand forecasting, demand forecasting in supply-chains, demand forecasting, demand prediction, supply chain forecasting, explanatory variable forecasting in supply-chains	Demand forecasting in AI, AI demand forecasting, neural network demand forecasting, artificial intelligence demand forecasting, demand forecasting advanced methods, AI demand prediction

The research portals that were used to find the references for this thesis:

1. LUT primo research portal
2. Google Scholar
3. Google open search

Appendix 2. Multiplicative double and triple smoothing equations

(Koehler et al. 2001; Weiheng et al. 2020)

Double Smoothing Multiplicative

Double Smoothing with multiplicative method. Φ is the dampening term which in standard multiplicative exponential smoothing is 1. If dampening term is 0, double smoothing becomes simple smoothing.

$$L_t = \alpha Y_t + (1-\alpha)(L_{t-1} * T^{\Phi}{}_{t-1}) \tag{1}$$

$$T_t = \beta(L_t / L_{t-1}) + (1-\beta)T^{\Phi}{}_{t-1} \tag{2}$$

$$F_{t+1} = L_t + T_t^{\Sigma\Phi} \tag{3}$$

Triple Smoothing Multiplicative

$$L_t = \alpha \frac{Y_t}{S_{t-m}} + (1-\alpha)(L_{t-1} + T_{t-1}) \quad (4)$$

$$T_t = \beta(L_t - L_{t-1}) + (1-\beta)T_{t-1} \quad (5)$$

$$S_t = \gamma(\frac{Y_t}{L_{t-1} - T_{t-1}}) + (1-\gamma)S_{t-m} \quad (6)$$

$$F_{t+1} = (L_t + T_t)S_{t-m(k+1)} \quad (7)$$

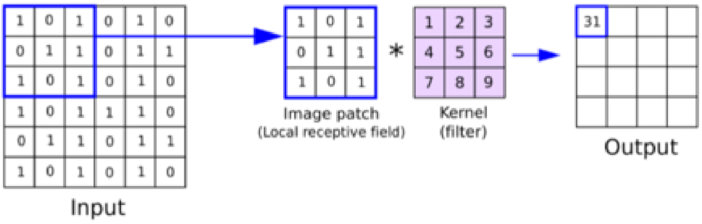
Appendix 3. Convolutional neural network structure and TCN principle

Appendix 3.1. CNN network limitations explained.

The convolutional neural network architecture is defined by a three-dimensional arrangement of neurons, instead of the standard two-dimensional array. The first layer in such neural networks is called a convolutional layer. Each neuron in the convolutional layer only processes the information from a small part of the visual field. The convolutional layers are followed by rectified layer units or ReLU, which enables the CNN to handle complicated information. From mathematics perspective convolution operation takes two functions, one of which one is shifted and reversed, and calculates the integral of their product. Convolution function describes how the first function affects the shape of the second function and therefore a combination of several convolutional functions generates constantly adjusting network that is capable for complex pattern recognition. (Reynolds 2022)

TCN variation of CNN process is divided into two stages: in the first stage CNN structure is used to calculate low-level features. CNN encodes time dependent spatial information (spatio-temporal). In the second stage low-level features are fed into the classifier which receives high-level temporal information using RNN structure. In practice TCN can use an input of any length and the output is matched by this length. These are called input and target-sequences. In forecasting this means that the prediction horizon is limited by the output length. (Hewage et al. 2020).

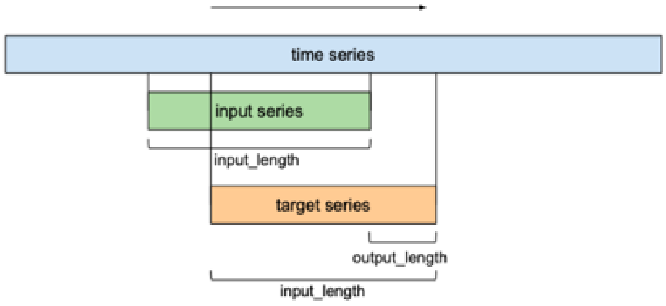
Appendix 3.2. Convolutional neural network structure and the limited forecasting horizon of temporal convolutional network (Lässig, 2021)



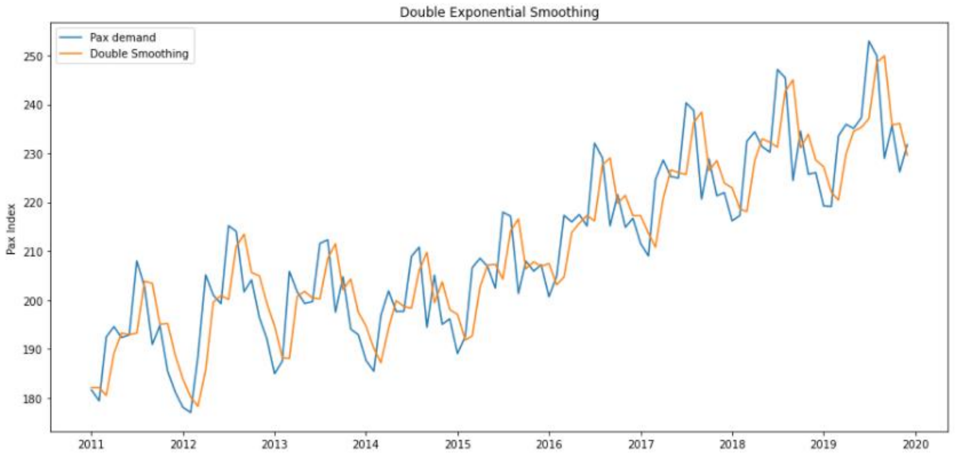
1	0	1	0	1	0
0	1	1	0	1	1
1	0	1	0	1	0
1	0	1	1	1	0
0	1	1	0	1	1
1	0	1	0	1	0

1	0	1
0	1	1
1	0	1

1	2	3
4	5	6
7	8	9

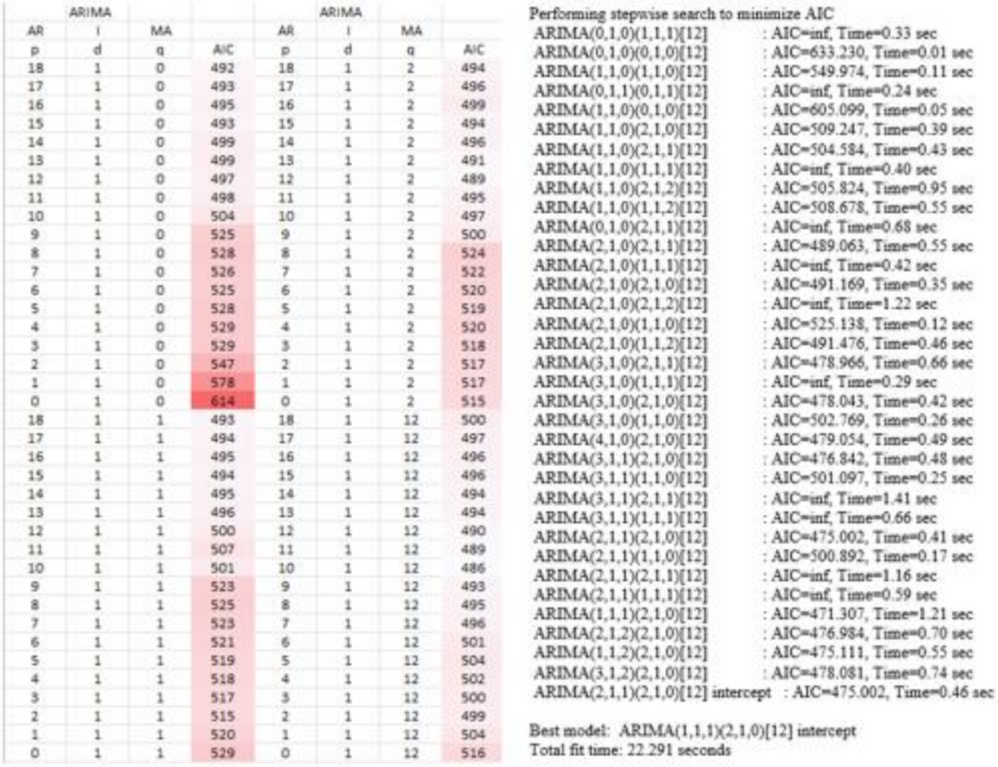


Appendix 4. Double exponential smoothing with smoothing constants of 1, tracks previous values with a lag.



Appendix 5. ARIMA/SARIMA model estimation through AIC and coefficients.

Appendix 5.1. Manual AIC Optimization Statistics for ARIMA and automatic model optimization for SARIMA models.



	ARIMA				ARIMA		
AR	I	MA		AR	I	MA	
p	d	q	AIC	p	d	q	AIC
18	1	0	492	18	1	2	494
17	1	0	493	17	1	2	496
16	1	0	495	16	1	2	499
15	1	0	493	15	1	2	494
14	1	0	499	14	1	2	496
13	1	0	499	13	1	2	491
12	1	0	497	12	1	2	489
11	1	0	498	11	1	2	495
10	1	0	504	10	1	2	497
9	1	0	525	9	1	2	500
8	1	0	528	8	1	2	524
7	1	0	526	7	1	2	522
6	1	0	525	6	1	2	520
5	1	0	528	5	1	2	519
4	1	0	529	4	1	2	520
3	1	0	529	3	1	2	518
2	1	0	547	2	1	2	517
1	1	0	578	1	1	2	517
0	1	0	614	0	1	2	515
18	1	1	493	18	1	12	500
17	1	1	494	17	1	12	497
16	1	1	495	16	1	12	496
15	1	1	494	15	1	12	496
14	1	1	495	14	1	12	494
13	1	1	496	13	1	12	494
12	1	1	500	12	1	12	490
11	1	1	507	11	1	12	489
10	1	1	501	10	1	12	486
9	1	1	523	9	1	12	493
8	1	1	525	8	1	12	495
7	1	1	523	7	1	12	496
6	1	1	521	6	1	12	501
5	1	1	519	5	1	12	504
4	1	1	518	4	1	12	502
3	1	1	517	3	1	12	500
2	1	1	515	2	1	12	499
1	1	1	520	1	1	12	504
0	1	1	529	0	1	12	516

```
Performing stepwise search to minimize AIC
 ARIMA(0,1,0)(1,1,1)[12]            : AIC=inf, Time=0.33 sec
 ARIMA(0,1,0)(0,1,0)[12]            : AIC=633.230, Time=0.01 sec
 ARIMA(1,1,0)(1,1,0)[12]            : AIC=549.974, Time=0.11 sec
 ARIMA(0,1,1)(0,1,1)[12]            : AIC=inf, Time=0.24 sec
 ARIMA(1,1,0)(0,1,0)[12]            : AIC=605.099, Time=0.05 sec
 ARIMA(1,1,0)(2,1,0)[12]            : AIC=509.247, Time=0.39 sec
 ARIMA(1,1,0)(2,1,1)[12]            : AIC=504.584, Time=0.43 sec
 ARIMA(1,1,0)(1,1,1)[12]            : AIC=inf, Time=0.40 sec
 ARIMA(1,1,0)(2,1,2)[12]            : AIC=505.824, Time=0.95 sec
 ARIMA(1,1,0)(1,1,2)[12]            : AIC=508.678, Time=0.55 sec
 ARIMA(0,1,0)(2,1,1)[12]            : AIC=inf, Time=0.68 sec
 ARIMA(2,1,0)(2,1,1)[12]            : AIC=489.063, Time=0.55 sec
 ARIMA(2,1,0)(1,1,1)[12]            : AIC=inf, Time=0.42 sec
 ARIMA(2,1,0)(2,1,0)[12]            : AIC=491.169, Time=0.35 sec
 ARIMA(2,1,0)(2,1,2)[12]            : AIC=inf, Time=1.22 sec
 ARIMA(2,1,0)(1,1,0)[12]            : AIC=525.138, Time=0.12 sec
 ARIMA(2,1,0)(1,1,2)[12]            : AIC=491.476, Time=0.46 sec
 ARIMA(3,1,0)(2,1,1)[12]            : AIC=478.966, Time=0.66 sec
 ARIMA(3,1,0)(1,1,1)[12]            : AIC=inf, Time=0.29 sec
 ARIMA(3,1,0)(2,1,0)[12]            : AIC=478.043, Time=0.42 sec
 ARIMA(3,1,0)(1,1,0)[12]            : AIC=502.769, Time=0.26 sec
 ARIMA(4,1,0)(2,1,0)[12]            : AIC=479.054, Time=0.49 sec
 ARIMA(3,1,1)(2,1,0)[12]            : AIC=476.842, Time=0.48 sec
 ARIMA(3,1,1)(1,1,0)[12]            : AIC=501.097, Time=0.25 sec
 ARIMA(3,1,1)(2,1,1)[12]            : AIC=inf, Time=1.41 sec
 ARIMA(3,1,1)(1,1,1)[12]            : AIC=inf, Time=0.66 sec
 ARIMA(2,1,1)(2,1,0)[12]            : AIC=475.002, Time=0.41 sec
 ARIMA(2,1,1)(1,1,0)[12]            : AIC=500.892, Time=0.17 sec
 ARIMA(2,1,1)(2,1,1)[12]            : AIC=inf, Time=1.16 sec
 ARIMA(2,1,1)(1,1,1)[12]            : AIC=inf, Time=0.59 sec
 ARIMA(1,1,1)(2,1,0)[12]            : AIC=471.307, Time=1.21 sec
 ARIMA(2,1,2)(2,1,0)[12]            : AIC=476.984, Time=0.70 sec
 ARIMA(1,1,2)(2,1,0)[12]            : AIC=475.111, Time=0.55 sec
 ARIMA(3,1,2)(2,1,0)[12]            : AIC=478.081, Time=0.74 sec
 ARIMA(2,1,1)(2,1,0)[12] intercept   : AIC=475.002, Time=0.46 sec

Best model:  ARIMA(1,1,1)(2,1,0)[12] intercept
Total fit time: 22.291 seconds
```

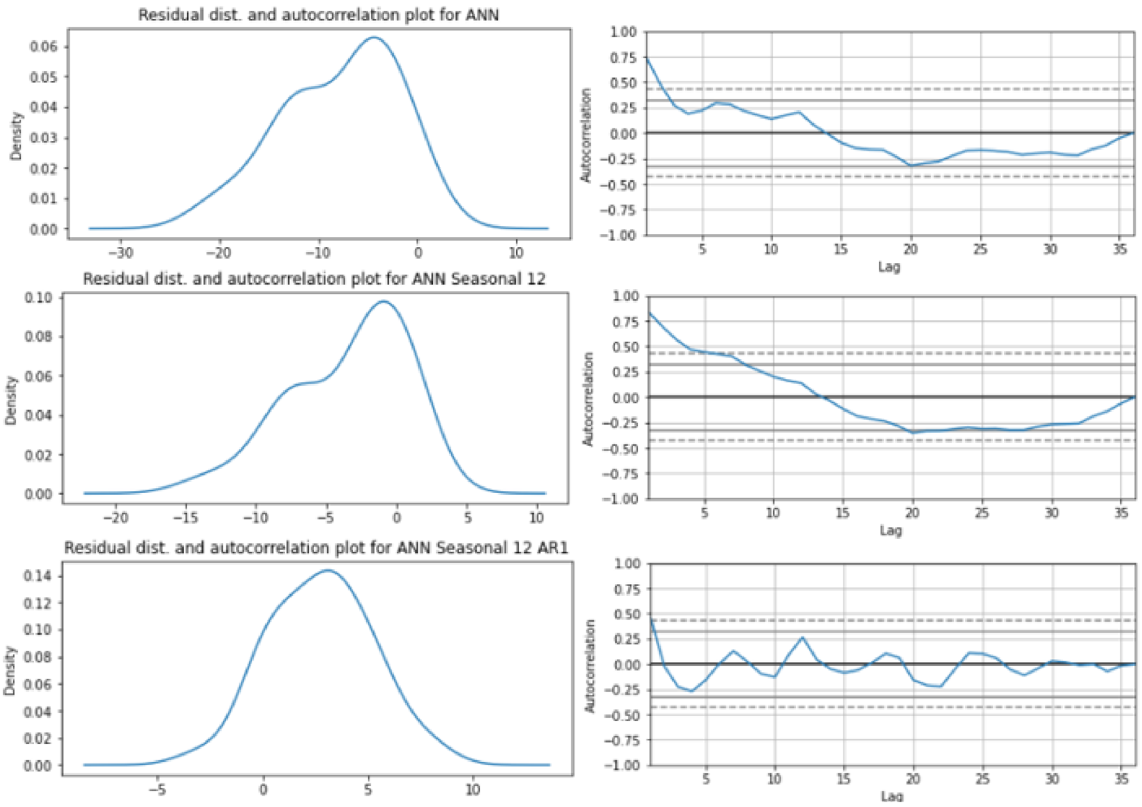
Appendix 5.2. ARIMA model coefficients for ARIMA(12,1,2) model.

ARIMA(12,1,2)	Coef.	Std err.	p>|z|
AR Lag 1	0.025	0.107	0.810
AR Lag 2	-0.179	0.115	0.120
AR Lag 3	-0.335	0.112	0.003
AR Lag 4	-0.424	0.119	0.000
AR Lag 5	-0.430	0.117	0.000
AR Lag 6	-0.272	0.114	0.017
AR Lag 7	-0.317	0.128	0.013
AR Lag 8	-0.311	0.136	0.023
AR Lag 9	-0.210	0.111	0.058
AR Lag 10	0.075	0.123	0.542
AR Lag 11	-0.004	0.129	0.977
AR Lag 12	-0.594	0.107	0.000
MA Lag 1	-1.458	0.109	0.000
Ma Lag 2	0.907	0.099	0.000
δ^2	15.68	1.393	0.000

Appendix 5.3. SARIMA model coefficients and their statistical significance.

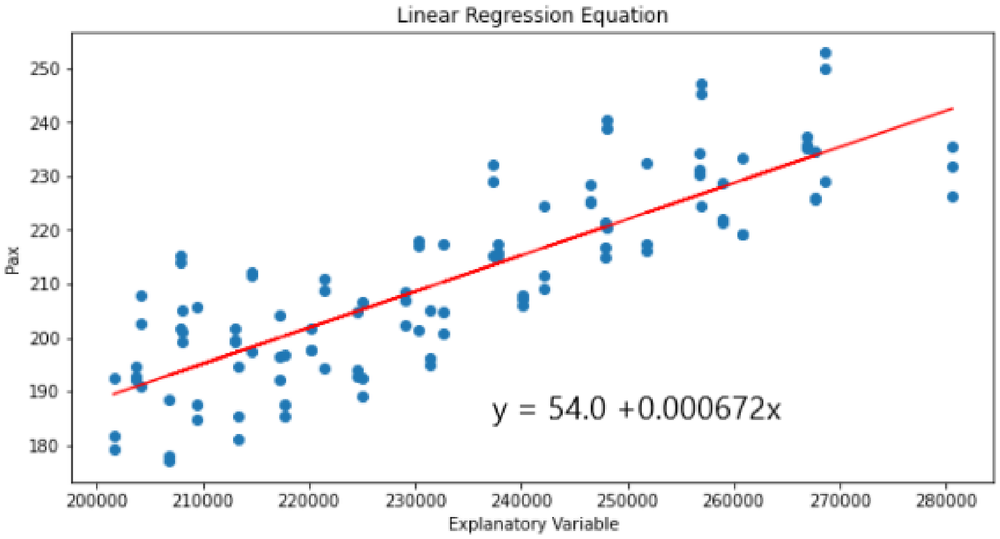
SARIMA(1,1,1)(2,1,0,12)	Coef.	Std err.	p>|z|
AR Lag 1	-0.43	0.11	0.00
MA Lag 1	-0.98	0.14	0.00
AR Seasonal Lag 12	-1.20	0.08	0.00
AR Seasonal Lag 24	-0.66	0.09	0.00
δ^2	12.31	2.64	0.00
SARIMA(1,1,1)(2,1,1,12)	Coef.	Std err.	p>|z|
AR Lag 1	-0.38	0.130	0.003
MA Lag 1	-0.98	0.223	0.000
AR Seasonal Lag 12	-0.89	0.221	0.000
AR Seasonal Lag 24	-0.43	0.218	0.051
MA Seasonal Lag 12	-0.53	0.342	0.118
δ^2	11.19	2.935	0.000

Appendix 6. ANN residual plots including residual distribution and autocorrelation plot

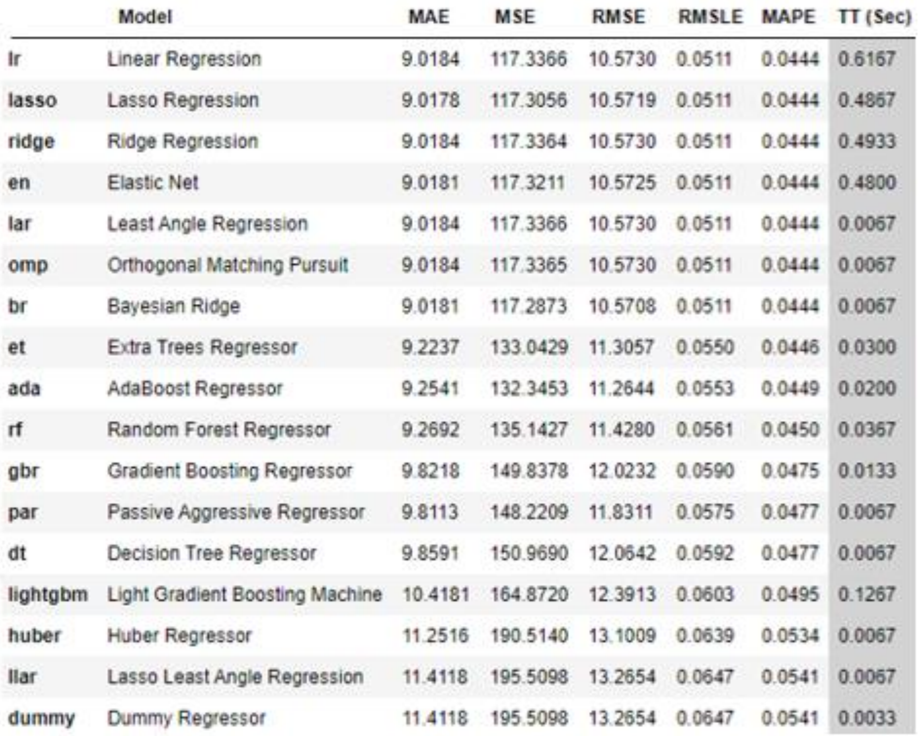


Appendix 7. Linear regression model fitting.

Appendix 7.1. Regression model fit for non-aligned original dataset.

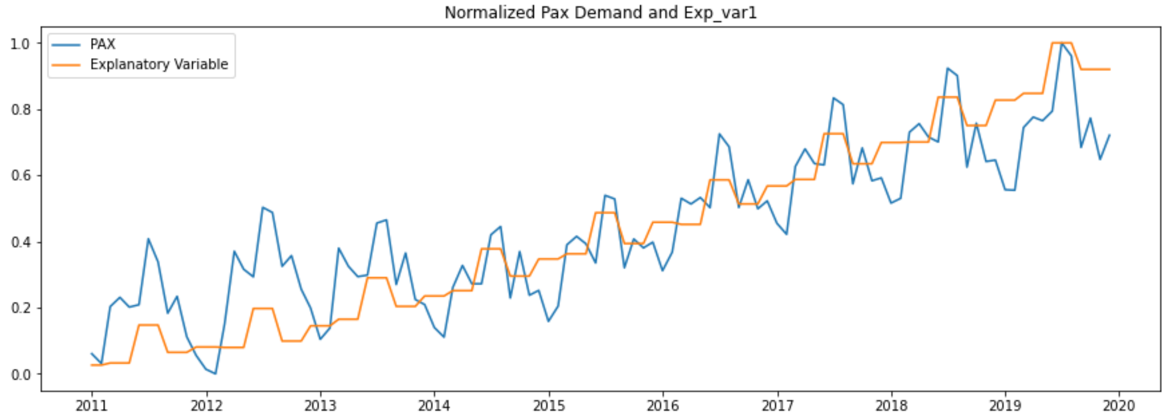


Appendix 7.2. Regression model selection simulation. Simulated accuracies are not exact but fit measures describe how well method is suited for model fitting for our dataset.

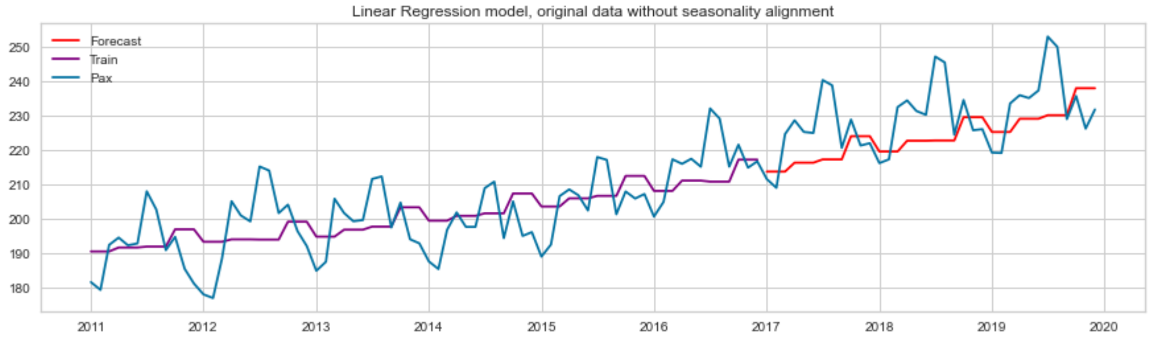


	Model	MAE	MSE	RMSE	RMSLE	MAPE	TT (Sec)
lr	Linear Regression	9.0184	117.3366	10.5730	0.0511	0.0444	0.6167
lasso	Lasso Regression	9.0178	117.3056	10.5719	0.0511	0.0444	0.4867
ridge	Ridge Regression	9.0184	117.3364	10.5730	0.0511	0.0444	0.4933
en	Elastic Net	9.0181	117.3211	10.5725	0.0511	0.0444	0.4800
lar	Least Angle Regression	9.0184	117.3366	10.5730	0.0511	0.0444	0.0067
omp	Orthogonal Matching Pursuit	9.0184	117.3365	10.5730	0.0511	0.0444	0.0067
br	Bayesian Ridge	9.0181	117.2873	10.5708	0.0511	0.0444	0.0067
et	Extra Trees Regressor	9.2237	133.0429	11.3057	0.0550	0.0446	0.0300
ada	AdaBoost Regressor	9.2541	132.3453	11.2644	0.0553	0.0449	0.0200
rf	Random Forest Regressor	9.2692	135.1427	11.4280	0.0561	0.0450	0.0367
gbr	Gradient Boosting Regressor	9.8218	149.8378	12.0232	0.0590	0.0475	0.0133
par	Passive Aggressive Regressor	9.8113	148.2209	11.8311	0.0575	0.0477	0.0067
dt	Decision Tree Regressor	9.8591	150.9690	12.0642	0.0592	0.0477	0.0067
lightgbm	Light Gradient Boosting Machine	10.4181	164.8720	12.3913	0.0603	0.0495	0.1267
huber	Huber Regressor	11.2516	190.5140	13.1009	0.0639	0.0534	0.0067
llar	Lasso Least Angle Regression	11.4118	195.5098	13.2654	0.0647	0.0541	0.0067
dummy	Dummy Regressor	11.4118	195.5098	13.2654	0.0647	0.0541	0.0033

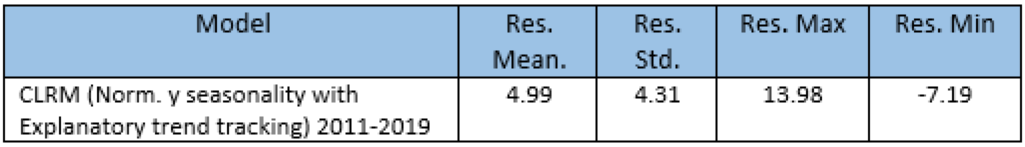
Appendix 7.3. Shifted explanatory variable with dependent variable in normalized range.



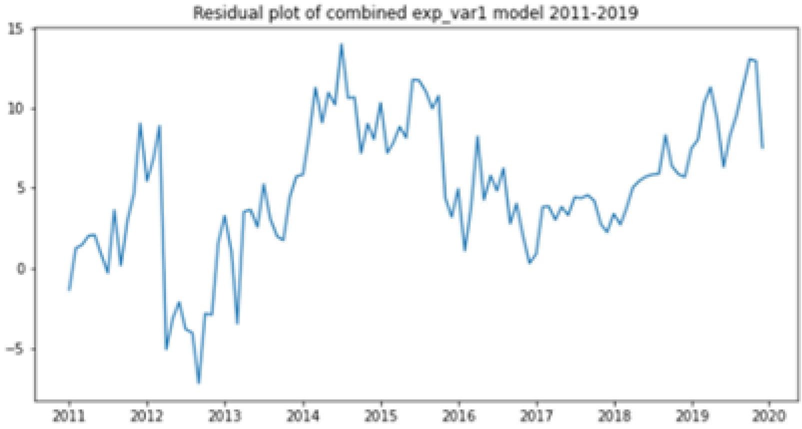
Appendix 7.4. Linear regression model forecast without seasonality alignment.



Appendix 8. Combined explanatory variable model residual plots and statistics.

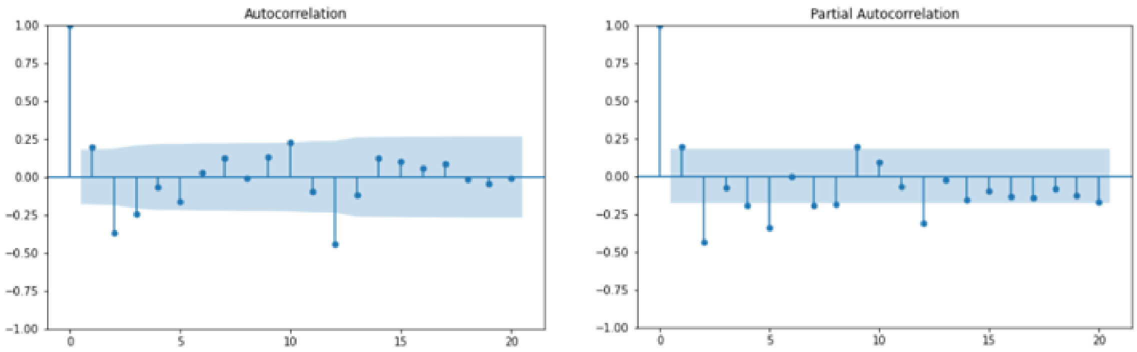


Model	Res. Mean.	Res. Std.	Res. Max	Res. Min
CLRM (Norm. y seasonality with Explanatory trend tracking) 2011-2019	4.99	4.31	13.98	-7.19

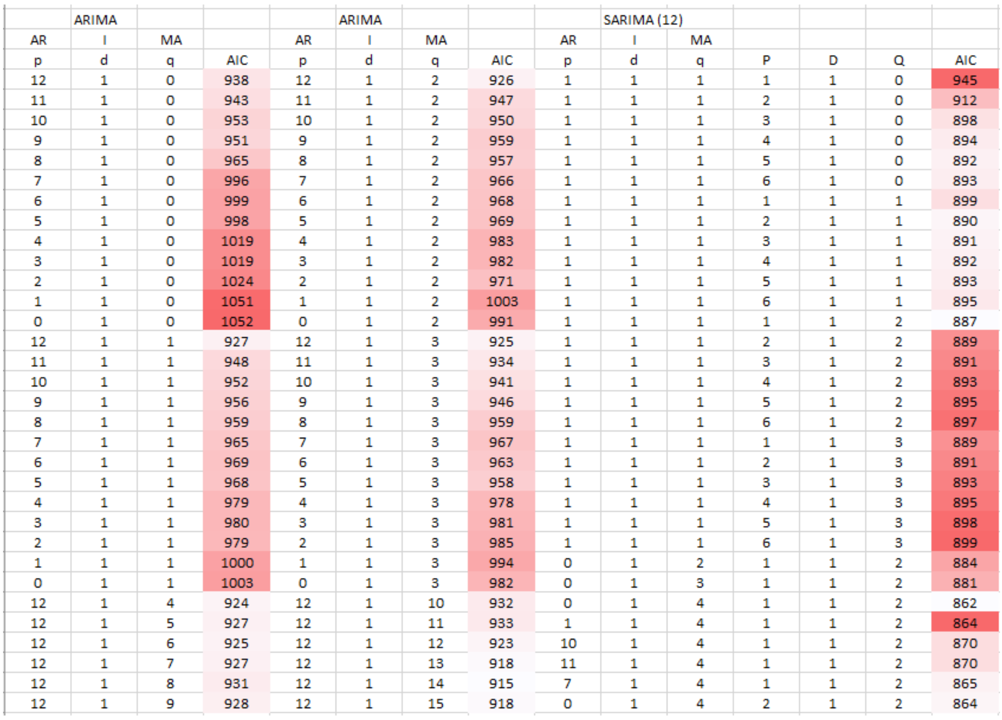


Appendix 9. ARIMA and SARIMA model selection, fitting, evaluation, residual diagnostics, and model coefficients for disruptive data.

Appendix 9.1. ACF and PACF plots for detrended disruptive dataset. Based on visual interpretation, ARIMA (12,1,12) could be appropriate model form based on ACF spikes seen on the plots.

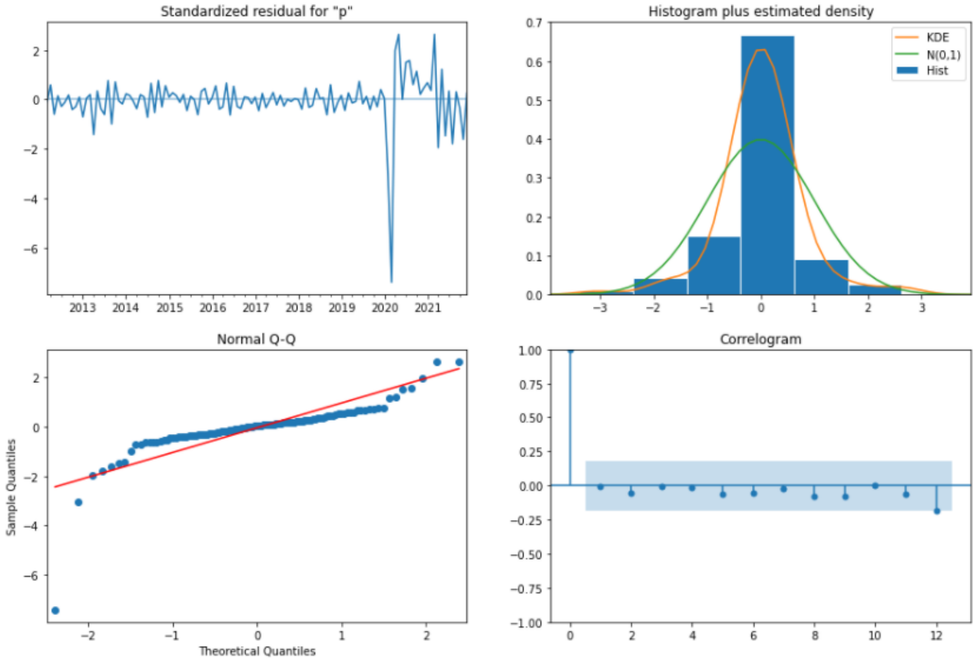


Appendix 9.2. Manual iterative model tuning for ARIMA and SARIMA models with AIC. ARIMA(12,1,14) and SARIMA (0,1,4)(1,1,2,12) have lowest AIC value.

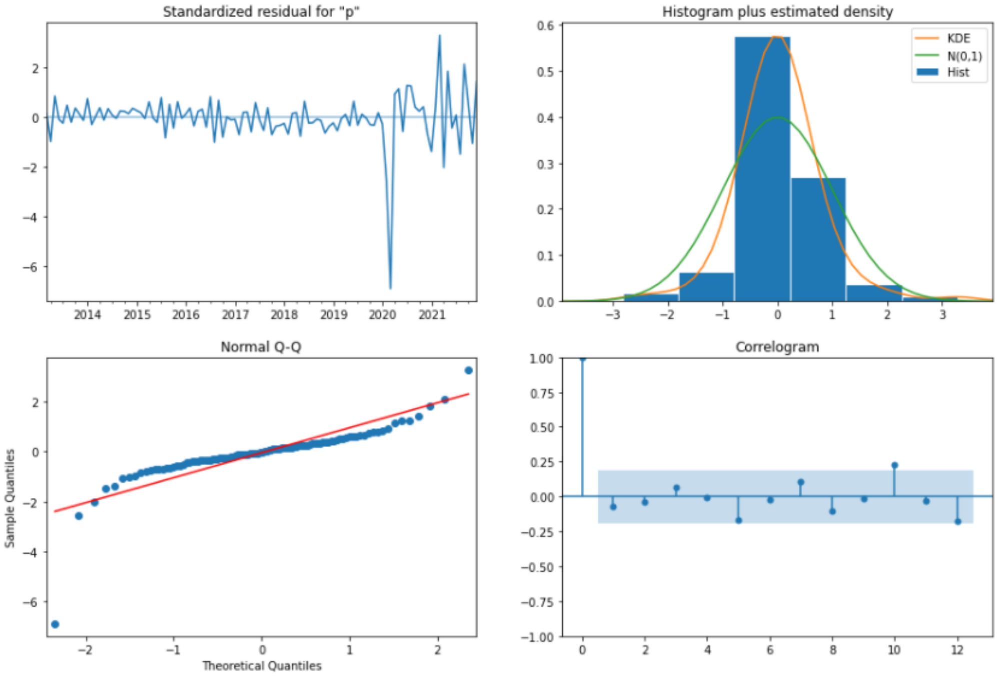


	ARIMA				ARIMA				SARIMA (12)					
AR	I	MA		AR	I	MA		AR	I	MA				
p	d	q	AIC	p	d	q	AIC	p	d	q	P	D	Q	AIC
12	1	0	938	12	1	2	926	1	1	1	1	1	0	945
11	1	0	943	11	1	2	947	1	1	1	2	1	0	912
10	1	0	953	10	1	2	950	1	1	1	3	1	0	898
9	1	0	951	9	1	2	959	1	1	1	4	1	0	894
8	1	0	965	8	1	2	957	1	1	1	5	1	0	892
7	1	0	996	7	1	2	966	1	1	1	6	1	0	893
6	1	0	999	6	1	2	968	1	1	1	1	1	1	899
5	1	0	998	5	1	2	969	1	1	1	2	1	1	890
4	1	0	1019	4	1	2	983	1	1	1	3	1	1	891
3	1	0	1019	3	1	2	982	1	1	1	4	1	1	892
2	1	0	1024	2	1	2	971	1	1	1	5	1	1	893
1	1	0	1051	1	1	2	1003	1	1	1	6	1	1	895
0	1	0	1052	0	1	2	991	1	1	1	1	1	2	887
12	1	1	927	12	1	3	925	1	1	1	2	1	2	889
11	1	1	948	11	1	3	934	1	1	1	3	1	2	891
10	1	1	952	10	1	3	941	1	1	1	4	1	2	893
9	1	1	956	9	1	3	946	1	1	1	5	1	2	895
8	1	1	959	8	1	3	959	1	1	1	6	1	2	897
7	1	1	965	7	1	3	967	1	1	1	1	1	3	889
6	1	1	969	6	1	3	963	1	1	1	2	1	3	891
5	1	1	968	5	1	3	958	1	1	1	3	1	3	893
4	1	1	979	4	1	3	978	1	1	1	4	1	3	895
3	1	1	980	3	1	3	981	1	1	1	5	1	3	898
2	1	1	979	2	1	3	985	1	1	1	6	1	3	899
1	1	1	1000	1	1	3	994	0	1	2	1	1	2	884
0	1	1	1003	0	1	3	982	0	1	3	1	1	2	881
12	1	4	924	12	1	10	932	0	1	4	1	1	2	862
12	1	5	927	12	1	11	933	1	1	4	1	1	2	864
12	1	6	925	12	1	12	923	10	1	4	1	1	2	870
12	1	7	927	12	1	13	918	11	1	4	1	1	2	870
12	1	8	931	12	1	14	915	7	1	4	1	1	2	865
12	1	9	928	12	1	15	918	0	1	4	2	1	2	864

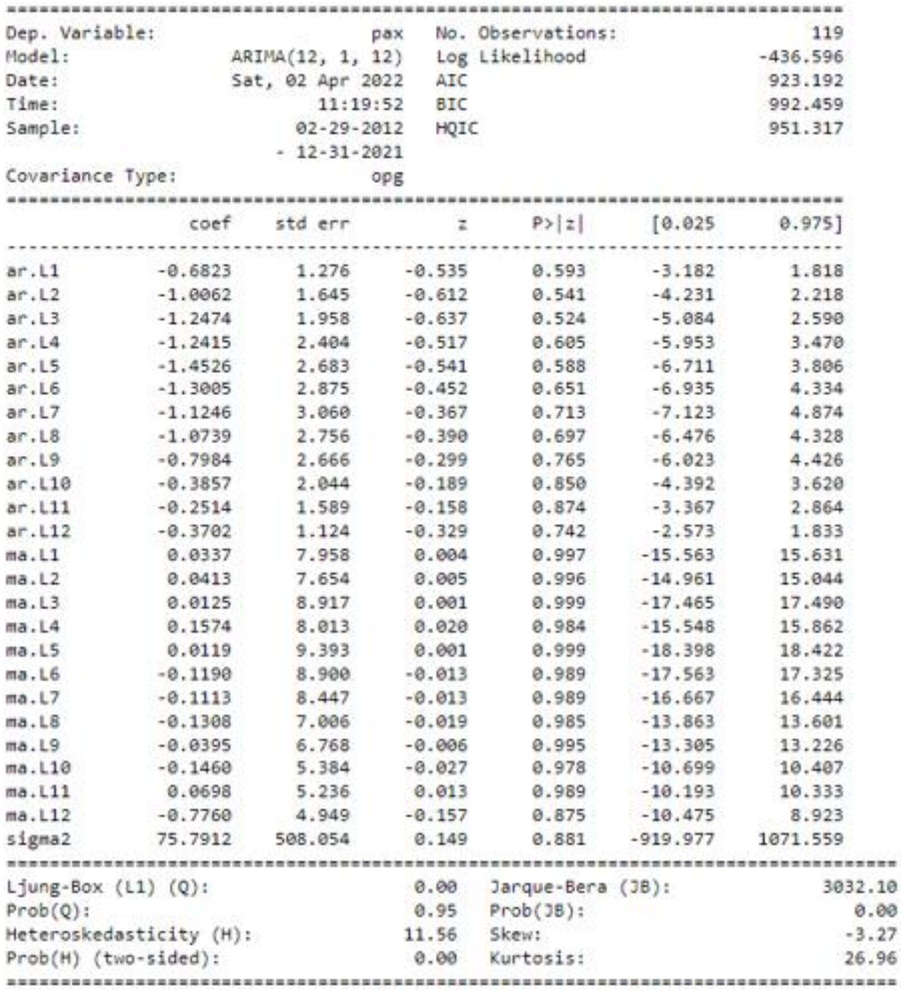
Appendix 9.3. ARIMA(12,1,12) model diagnostics.



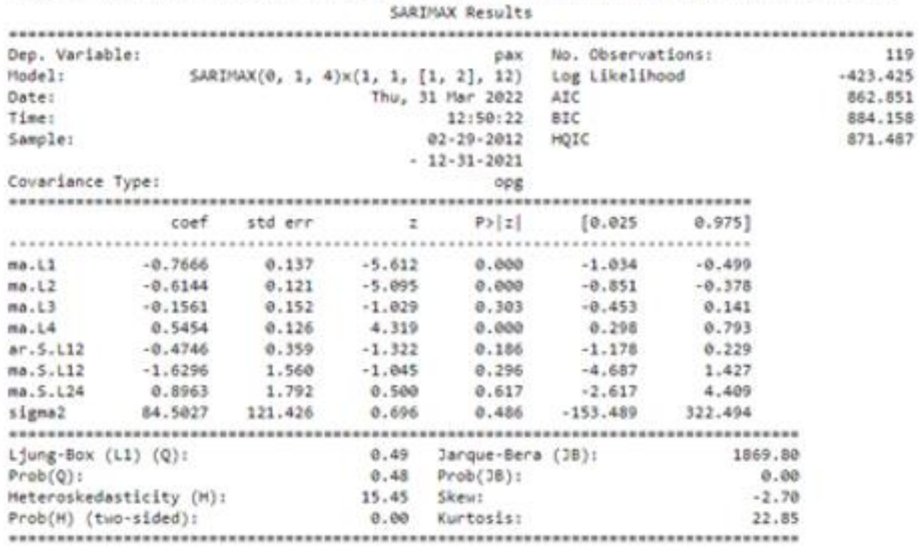
Appendix 9.4. SARIMA(0,1,4)(1,1,2,12) model diagnostics.



Appendix 9.5. ARIMA and SARIMA model coefficients and test results in disruptive conditions

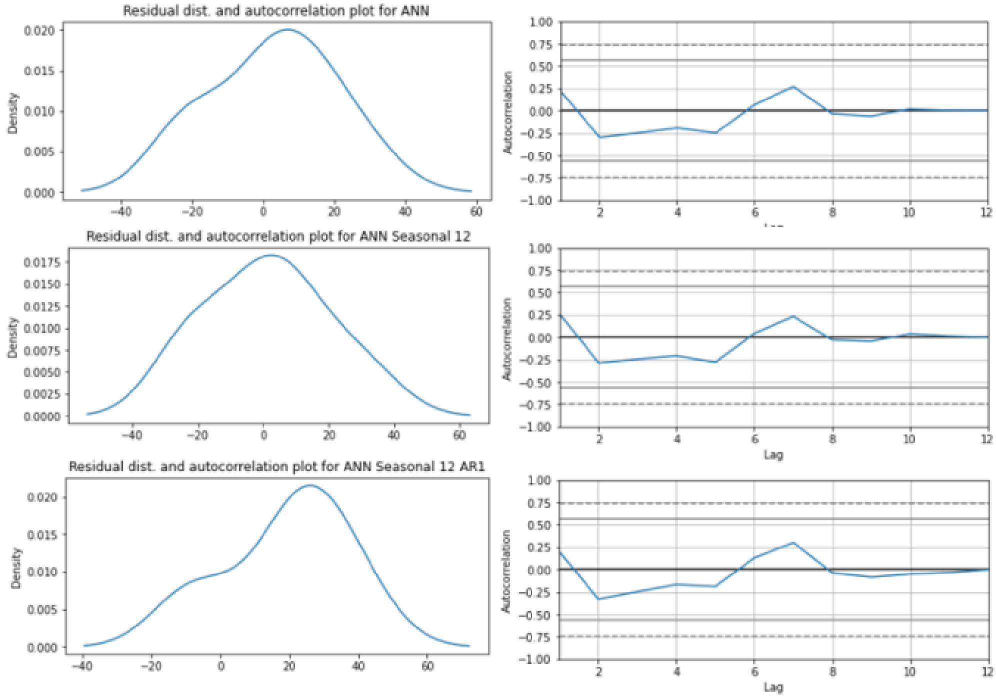


```
==============================================================================
Dep. Variable:                    pax   No. Observations:                  119
Model:                ARIMA(12, 1, 12)  Log Likelihood                -436.596
Date:                 Sat, 02 Apr 2022  AIC                            923.192
Time:                         11:19:52  BIC                            992.459
Sample:                     02-29-2012  HQIC                           951.317
                          - 12-31-2021
Covariance Type:                   opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.6823      1.276     -0.535      0.593      -3.182       1.818
ar.L2         -1.0062      1.645     -0.612      0.541      -4.231       2.218
ar.L3         -1.2474      1.958     -0.637      0.524      -5.084       2.590
ar.L4         -1.2415      2.404     -0.517      0.605      -5.953       3.470
ar.L5         -1.4526      2.683     -0.541      0.588      -6.711       3.806
ar.L6         -1.3005      2.875     -0.452      0.651      -6.935       4.334
ar.L7         -1.1246      3.060     -0.367      0.713      -7.123       4.874
ar.L8         -1.0739      2.756     -0.390      0.697      -6.476       4.328
ar.L9         -0.7984      2.666     -0.299      0.765      -6.023       4.426
ar.L10        -0.3857      2.044     -0.189      0.850      -4.392       3.620
ar.L11        -0.2514      1.589     -0.158      0.874      -3.367       2.864
ar.L12        -0.3702      1.124     -0.329      0.742      -2.573       1.833
ma.L1          0.0337      7.958      0.004      0.997     -15.563      15.631
ma.L2          0.0413      7.654      0.005      0.996     -14.961      15.044
ma.L3          0.0125      8.917      0.001      0.999     -17.465      17.490
ma.L4          0.1574      8.013      0.020      0.984     -15.548      15.862
ma.L5          0.0119      9.393      0.001      0.999     -18.398      18.422
ma.L6         -0.1190      8.900     -0.013      0.989     -17.563      17.325
ma.L7         -0.1113      8.447     -0.013      0.989     -16.667      16.444
ma.L8         -0.1308      7.006     -0.019      0.985     -13.863      13.601
ma.L9         -0.0395      6.768     -0.006      0.995     -13.305      13.226
ma.L10        -0.1460      5.384     -0.027      0.978     -10.699      10.407
ma.L11         0.0698      5.236      0.013      0.989     -10.193      10.333
ma.L12        -0.7760      4.949     -0.157      0.875     -10.475       8.923
sigma2        75.7912    508.054      0.149      0.881    -919.977    1071.559
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):              3032.10
Prob(Q):                              0.95   Prob(JB):                         0.00
Heteroskedasticity (H):              11.56   Skew:                            -3.27
Prob(H) (two-sided):                  0.00   Kurtosis:                        26.96
===================================================================================
```

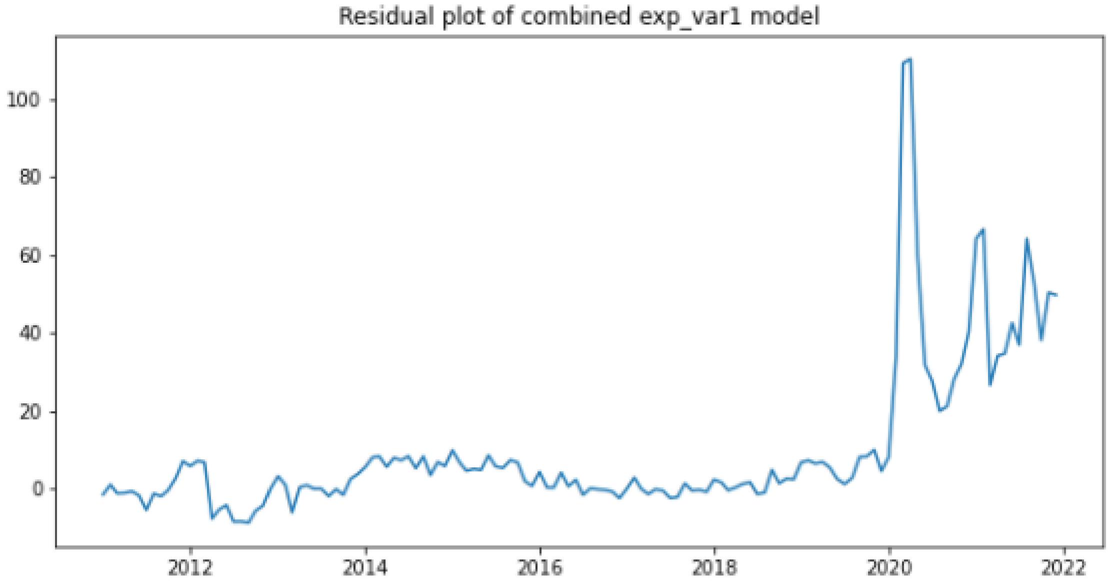


```
                                      SARIMAX Results
==========================================================================================
Dep. Variable:                                pax   No. Observations:                  119
Model:             SARIMAX(0, 1, 4)x(1, 1, [1, 2], 12)   Log Likelihood             -423.425
Date:                            Thu, 31 Mar 2022   AIC                            862.851
Time:                                    12:50:22   BIC                            884.158
Sample:                                02-29-2012   HQIC                           871.487
                                     - 12-31-2021
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -0.7666      0.137     -5.612      0.000      -1.034      -0.499
ma.L2         -0.6144      0.121     -5.095      0.000      -0.851      -0.378
ma.L3         -0.1561      0.152     -1.029      0.303      -0.453       0.141
ma.L4          0.5454      0.126      4.319      0.000       0.298       0.793
ar.S.L12      -0.4746      0.359     -1.322      0.186      -1.178       0.229
ma.S.L12      -1.6296      1.560     -1.045      0.296      -4.687       1.427
ma.S.L24       0.8963      1.792      0.500      0.617      -2.617       4.409
sigma2        84.5027    121.426      0.696      0.486    -153.489     322.494
===================================================================================
Ljung-Box (L1) (Q):                   0.49   Jarque-Bera (JB):              1869.80
Prob(Q):                              0.48   Prob(JB):                         0.00
Heteroskedasticity (H):              15.45   Skew:                            -2.70
Prob(H) (two-sided):                  0.00   Kurtosis:                        22.85
===================================================================================
```

Appendix 10. Residual distribution and autocorrelation plot of the disruptive environment ANN models.



Appendix 11. Combined trend tracking model residual plot in disruptive environment.



Appendix 12. Verbal result summary table evaluation criteria

Appendix 12.1. Relative Numerical results scale behind the verbal evaluation.

Limited scale column evaluates the performance deterioration under limited data conditions and focuses on performance change in normal conditions. Parameter tuning column evaluates the necessary steps tests and simulations needed to iterate an optimal forecasting model.

Model Performance Under:	Normal Conditions	Disruptive Conditions	Limited Data Stable Cond. Perf Change	Limited Data Disrup. Cond. Perf. Change	Overall complexity
Exp. Smoothing	0.67%	6.36%	1.73%	1.20%	Low
ARIMA / SARIMA	0.75/0.92%	8.27/8.48%	0.1/0.19%	1.22/1.67%	High
ANN	1.66/3.26%	6.49/9.33%	0.25/0.37%	-1.75/0.85%	Med./Low
RNN	5.63/12.69%	20.99/25.67%	5.57/6.08%	1.24/5.91%	Medium
CLRM	2.84/3.76%	11.51/12.07%	0.45/1.79%	*-3.40/-1.59%	Medium
Explanatory Var Trend Tracking	4.12%	22.87%	-2.45%	2.53%	Low

*CLRM performance in disruptive conditions is biased since the measured accuracy is focused only to last 12 observations and actual regression coefficient is heavily affected by the last training set values when we remove the earlier data samples with limited testing case. The starting level of CLRM accuracy is low, hence the performance increase.

Appendix 12.2. Accuracy evaluation scale, scales are relative to comparative method perf.

Model Forecasting Perf. Scale	Normal Conditions	Disruptive Conditions	Limited Data	Parameter Selection Required	
High	0-1.5%	0-7%	0-1.5%	Low	1-2
High/Medium	1.5-4.5%	7-9.5%	1.5-4.5%	Med/low	3-4
Medium	4.5-5.5%	9.5-12%	4.5-5.5%	Medium Effort	5-6
Low	> 5.5%	>12%	>5.5%	High Effort	>6

Appendix 12.3. Forecasting method verbal complexity evaluation.

Overall complexity is combination of model parameter selection count and complexity measure. Complexity measure includes the steps necessary to take for creation of full model and the mathematical complexity evaluation of each method. We simplify the mathematical evaluation to be binary, if model parameters are transparent and we have clear equations to calculate model results manually, then model is simple. If method usage requires computer simulation and computational packages to be used then it is complex.

Overall Complexity: (Parameter selection count + Model complexity)

Model complexity = (Iterative implementations steps + Computational complexity)

Example: **Method [Parameter selection count / (Iterative implementations steps + Computational complexity)]**

Methods sorted to ascending overall complexity order.

Exponential Smoothing [3/(2/Simple)] uses the equations 7-10 described in chapter 3.3. In the model we can select the smoothing constants when targeting for minimum forecasting error, but the overall modelling is simple and computationally light. Model fitting includes the typical residual evaluation that is conducted with all forecasting models.

With **Explanatory variable trend tracking** [**2/(3/Simple**)] model we first calculated the normalized annual pattern. This was complemented by calculating annual growth figures from both explanatory and dependent variable. Then the growth trend from explanatory variable is combined with seasonal pattern. There is no normal regression optimization included into the model fitting, but normal residual analysis is present. Model can be calculated by hand if necessary.

ANN [3/(2/Complex)] model is simple to implement but the fundamental operating principle described in chapter 3.5. is quite complex if compared to normal equation models like exponential smoothing or ARIMA. Model also has limited transparency due to its neural network node structure and iterative nature which results as accuracy range, not with exact result. Model needs to be simulated since its computationally intensive.

RNN [+6/(2/Complex)] model includes significantly more parameters like learning rates for the recurrent structures, number of hidden layers, input observation length, random starting point count, and iterative round counts.

CLRM [2/(4/Simple)] normal regression model optimization through OLS is quite simple and includes only 2 coefficients. However, in complete model selection we also simulated the performance of other regression variations to find out that linear regression vas reliably a highly performing option. CLRM model usage also includes the CLRM model assumption evaluation and testing which increases the complexity.

Classical **ARIMA/SARIMA [+6/(5/Simple)]** methods are not highly complex, but the overall relative high implementation effort is resulting from model evaluation process. The iterative Box-Jenkins approach with ARIMA/SARIMA requires stationarity testing to begin with. After this method implementation requires several iteration rounds, manual fit comparison, and evaluation parameters like AIC, HBIC or Max log. Likelihood comparison.

All evaluations are relative to this research and evaluated with research dataset. If we use different kind of data or include new methods this might change relative performance scale.