DEGREE PROJECT IN MATHEMATICS,
SECOND CYCLE, 30 CREDITS
STOCKHOLM, SWEDEN 2021

Forecasting Volume of Sales During the Abnormal Time Period of COVID-19

An Investigation on How to Forecast, Where the Classical ARIMA Family of Models Fail

CHRISTINA GHAWI

KTH ROYAL INSTITUTE OF TECHNOLOGY
SCHOOL OF ENGINEERING SCIENCES

Forecasting Volume of Sales During the Abnormal Time Period of COVID-19

An Investigation on How to Forecast, Where the Classical ARIMA Family of Models Fail

CHRISTINA GHAWI

Degree Project in Mathematical Statistics (30 ECTS credits)
Master's Programme, Applied and Computational Mathematics
(120 ECTS credits)
KTH Royal Institute of Technology, School of Engineering Sciences
Supervisor at KTH: Pierre Nyquist
Examiner at KTH: Pierre Nyquist
External partner: Klarna Bank AB
Supervisor at external partner: David Zenkert
Stockholm, Sweden 2021

Abstract

During the COVID-19 pandemic, customer shopping habits have changed. Some industries experienced an abrupt shift during the pandemic outbreak while others navigate in new normal states. For some merchants, the highly-uncertain new phenomena of COVID-19 expresses as outliers in time series of volume of sales. As forecasting models tend to replicate past behavior of a series, outliers complicates the procedure of forecasting; the abnormal events tend to unreliably replicate in forecasts of the subsequent year(s).

In this thesis, we investigate how to forecast volume of sales during the abnormal time period of COVID-19, where the classical ARIMA family of models produce unreliable forecasts. The research revolved around three time series exhibiting three types of outliers: a level shift, a transient change and an additive outlier. Upon detecting the time period of the abnormal behavior in each series, two experiments were carried out as attempts for increasing the predictive accuracy for the three extreme cases. The first experiment was related to imputing the abnormal data in the series and the second was related to using a combined model of a pre-pandemic and a post-abnormal forecast.

The results of the experiments pointed at significant improvement of the mean absolute percentage error at significance level $\alpha = 0.05$ for the level shift when using a combined model compared to the pre-pandemic best-fit SARIMA model. Also, at significant improvement for the additive outlier when using a linear impute. For the transient change, the results pointed at no significant improvement in the predictive accuracy of the experimental models compared to the pre-pandemic best-fit SARIMA model. For the purpose of generalizing to large-scale conclusions of methods' superiority or feasibility for particular abnormal behaviors, empirical evaluations are required.

The proposed experimental models were discussed in terms of reliability, validity and quality. By residual diagnostics, it was argued that the models were valid; however, that further improvements can be made. Also, it was argued that the models fulfilled desired attributes of simplicity, scaleability and flexibility. Due to the uncertain phenomena of the COVID-19 pandemic, it was suggested not to take the outputs as long-term reliable solutions. Rather, as temporary solutions requiring more frequent updating of forecasts.

Keywords

Time series forecasting. ARIMA family of models. Abnormal time period. Extreme events. COVID-19 pandemic. Time series outliers.

Sammanfattning

Estimering av försäljningsprognoser under den abnorma tidsperioden av coronapandemin

Under coronapandemin har kundbeteenden och köpvanor förändrats. I vissa branscher upplevdes ett plötsligt skifte vid pandemiutbrottet och i andra navigerar handlare i nya normaltillstånd. För vissa handlare är förändringarna så pass distinkta att de yttrar sig som avvikelser i tidsserier över försäljningsvolym. Dessa avvikelser komplicerar prognosering. Då prognosmodeller tenderar att replikera tidsseriers tidigare beteenden, tenderas det avvikande beteendet att replikeras i försäljningsprognoser för nästkommande år.

I detta examensarbete ämnar vi att undersöka tillvägagångssätt för att estimera försäljningsprognoser under den abnorma tidsperioden av COVID-19, då klassiska tidsseriemodeller felprognoserar. Detta arbete kretsade kring tre tidsserier som uttryckte tre avvikelsertyper: en nivåförskjutning, en övergående förändring och en additiv avvikelse. Efter att ha definierat en specifik tidsperiod relaterat till det abnorma beteendet i varje tidsserie, utfördes två experiment med syftet att öka den prediktiva noggrannheten för de tre extremfallen. Det första experimentet handlade om att ersätta den abnorma datan i varje serie och det andra experimentet handlade om att använda en kombinerad pronosmodell av två estimerade prognoser, en pre-pandemisk och en post-abnorm.

Resultaten av experimenten pekade på signifikant förbättring av ett absolut procentuellt genomsnittsfel för nivåförskjutningen vid användande av den kombinerade modellen, i jämförelse med den pre-pandemiskt bäst passande SARIMA-modellen. Även, signifikant förbättring för den additiva avvikelsen vid ersättning av abnorm data till ett motsvarande linjärt polynom. För den övergående förändringen pekade resultaten inte på en signifikant förbättring vid användande av de experimentella modellerna. För att generalisera till storskaliga slutsatser giltiga för specifika avvikande beteenden krävs empirisk utvärdering.

De föreslagna modellerna diskuterades utifrån tillförlitlighet, validitet och kvalitet. Modellerna uppfyllde önskvärda kvalitativa attribut såsom enkelhet, skalbarhet och flexibilitet. På grund av hög osäkerhet i den nuvarande abnorma tidsperioden av coronapandemin, föreslogs det att inte se prognoserna som långsiktigt pålitliga lösningar, utan snarare som tillfälliga tillvägagångssätt som regelbundet kräver om-prognosering.

Acknowledgments

Throughout the conduction of this thesis, I have received a great deal of support and assistance.

I would like to acknowledge my supervisor at KTH, Associate Professor Pierre Nyquist in the Department of Mathematics. Your continuous support and insightful feedback throughout the project helped me strengthened the quality of my work.

I would like to thank the Data Science competence at Klarna, for giving me the opportunity to write my thesis at the competence and to take part in an amazing journey the past five months. Specifically, I would like to thank the Financial Modeling team, for the warm welcome and the developing and joyful discussions.

I would like to acknowledge my supervisor at Klarna, Senior Data Scientist in the Financial Modeling team, David Zenkert. Thank you for always making time for me to discuss problems, theories and potential avenues. Your continuous encouragement pushed me to sharpen my thinking and your constructive feedback brought my work to a higher level.

In addition, I would like to thank my family and my friends, for always believing in and encouraging me, throughout my journey at KTH. Last, but certainly not least, I would like to thank my partner, Navid, for your endless love, patience and support these past years. Thank you for your continuous encouragement and for the late-night discussions throughout my thesis.

Stockholm, May 2021
Christina Ghawi

Contents

List of Figures

List of Tables

List of Acronyms

ACF Autocorrelation Function

ADF (test) Augmented Dickey-Fuller (test)

AIC Akaike Information Criterion

AICc Corrected Akaike Information Criterion

AO Additive Outlier

AR Autoregressive

ARCH Autoregressive Conditional Heteroskedasticity

ARIMA Autoregressive Integrated Moving-Average

ARMA Autoregressive Moving-Average

COVID-19 Coronavirus Disease 2019

GARCH Generalized Autoregressive Conditional Heteroskedasticity

HLN (test) Harvey-Leybourne-Newbold (test)

KPSS (test) Kwiatkowski–Phillips–Schmidt–Shin (test)

LS Level Shift

LSTM Long Short-Term Memory

MA Moving-Average

MAPE Mean Absolute Percentage Error

MBE Mean Bias Error

PACF Partial Autocorrelation Function

PI Prediction Interval

Q-Q (plot) Quantile-Quantile (plot)

SARIMA Seasonal Autoregressive Integrated Moving-Average

TC Transient Change

WHO World Health Organization

Chapter 1

Introduction

In this thesis, we will investigate how to forecast volume of sales during the abnormal time period of COVID-19, where the classical ARIMA family of models fail to produce reliable forecasts.

1.1 Background

A time series is a series of observations made sequentially through time. That is, a set of observations indexed in a specific time order. Time series analysis is a field of statistics used in investigating the internal structure in a time series, by identifying patterns in the data such as trends, seasonal variations, autocorrelations and cyclic behaviours (Chatfield 2003).

There are several objectives in analysing a time series; one of these is *forecasting*. Time series forecasting is the use of a model to predict future estimates based on past observations and is often incorporated in making data-driven decisions. For the past time, forecasting has become more of an integrated part in decision-making activities for many industries (Hyndman and Athanasopoulos 2018).

The predictability of an event depends on multiple factors. According to Chatfield (2003): how well we understand the factors that contribute to it, how much data are available and whether or not forecasts can affect the thing one is trying to forecast. Although, even if a forecasting model mimics observed data accurately, future predictions could become inaccurate. An abnormal or unexpected event, e.g. the COVID-19 pandemic, could majorly influence the forecasts, as it increases inhomogeneity and uncertainty in the time series.

This thesis is conducted as a collaboration with Klarna Bank AB, henceforth

Klarna. Klarna operates within the financial technology sector and offers transaction services for hundreds of thousands merchants worldwide. For some of these, the abnormal event of the COVID-19 pandemic has increased inhomogeneity in data collections, complicating the procedure of forecasting.

1.2 Problem space

During the COVID-19 pandemic, customer shopping habits have changed. Some industries experienced an abrupt shift at the time point of the pandemic outbreak while others navigate in highly-uncertain new normal states. For some merchants, the abrupt change expresses as outliers in time series of volume of sales, complicating the procedure of forecasting.

As forecasting models tend to replicate past behavior of a time series in predictions of future estimates, the abnormal event related to the COVID-19 outbreak tends to be unreliably replicated in forecasts of the subsequent year(s). Thus, on the one hand, models obtained prior to the pandemic outbreak may fail to forecast accurately due to replication of abnormal events. Further, on the other hand, models obtained during the pandemic may fail to forecast accurately due to small sample sizes post the pandemic outbreak. The problem complexity is clear.

Based on the assumption that the abnormal event will not recur in the subsequent year(s), it is of interest to investigate how to improve the forecasting procedure of predicting volume of sales for the few yet important extreme cases. That is, to investigate how we can tackle the problem of unreliable forecasts replicating abnormal events that most probable will not recur. The formal research question is stated in Subsection (1.2.1).

1.2.1 Research question

The research question of this thesis is as follows: How can we forecast volume of sales, to a greater extent of reliability and accuracy, during the abnormal time period of the COVID-19 pandemic?

1.3 Purpose

The aim of this thesis is to tackle the problem of unreliable forecasts replicating certain abnormal events that most probable will not recur in the subsequent year(s). Consequently, the purpose of this thesis is to improve the forecast reliability and predictive accuracy for some particular extreme cases of time series outliers. The objectives and scope of this thesis are stated in Subsections (1.3.1-1.3.2).

1.3.1 Objectives

The goal of this thesis is to find a method for tackling the problem of unreliable forecasts replicating certain abnormal events. In particular, the goal is to find a method that could work as a complement to existing building blocks of forecasting models, that could be used for increasing the forecast reliability and predictive accuracy for specific extreme cases of time series outliers.

Desired quality attributes of sought method

The sought method is desired to satisfy specific quality attributes.

Primarily, the attribute of simplicity. As the method should be used as a complement to existing building blocks of forecasting models, it is desired to retain a comprehensible approach of low cost-complexity. Also, a low cost-complexity is crucial if the method should be applicable to a large-scale number of data sets.

Secondarily, the attribute of scaleability. It is desired to scale the method to several data sets. Thus, the method should be automatable and of low cost-complexity. In addition, it is of importance to not necessarily overfit the method to a specific data set.

Tertiarily, the attribute of flexibility. In the setting of this thesis, the time series consists of volume of sales for some specific merchants. However, time series outliers can arise in all kind of time series regardless of context. Thus, it is desired to find a method adaptable to external changes such as changing the context of the series.

1.3.2 Scope

The scope and delimitations of this thesis are as follows.

First, rather than finding a (new) forecasting model, the research will be centered around improving the prerequisites for already existing building blocks of forecasting models. The (potential) improvements will be made by investigating methods that could work as a complement for handling particular extreme cases.

Second, the research will be centered around the framework of time series analysis. In particular, in order to delimit the research, it will be centered around the ARIMA family of models.

Third, the research will be centered around three types of abnormal behavior, also called time series outliers. Namely, the level shift (LS), the transient change (TC) and the additive outlier (AO). The classifications of time series outliers are stated in Subsection (4.1.4). In order to delimit the research, each abnormal behavior will be illustrated by one specific time series.

Fourth, the focus of the research is not to find a method that is optimised and overfit to a specific time series, but rather to find a method that could be deemed feasible for the abnormal behavior the specific series exhibits.

1.4 Research method

This section summaries the research method used for investigating the research question in a mathematical modeling manner. Subsection (1.4.1) summaries the experimental setup and design. Subsection (1.4.2) summaries the evaluation framework.

1.4.1 Experimental design

To begin with, the experimental setup consists of three time series. Each time series exhibits a distinct abnormal behavior. Further, the experimental design is divided in three steps as follows. First, illustrating the problem. Second, pinpointing the problem. Third, tackling the problem. The mathematical modeling procedure, described in detail in Chapter (5), is as follows.

First, for each time series, a benchmark forecasting model is chosen. The benchmark model will illustrate the problem of interest and function as a foundation throughout the thesis. For the purpose of choosing a valid and reasonable benchmark model, model selection is performed in accordance to a classical time series framework.

Second, for each time series, the abnormal time period assumed related to the COVID-19 outbreak is detected. The detected abnormal time period will pinpoint the time period of outliers and the problem of interest. Consequently, it will enable handling the particular time period with additional attention.

Third, for each time series, two experiments are performed. The two experiments will attempt to tackle the problem of interest by increasing the forecast accuracy and forecast reliability compared to the benchmark forecast. The first experiment, Experiment A, will relate to imputing data of the abnormal time period. The second experiment, Experiment B, will relate to combining forecasting models pre- and post the abnormal time period.

1.4.2 Evaluation framework

For each time series, forecasts will be estimated by the benchmark model and the two experimental models. The resulting forecasts will be evaluated in terms of two metrics: primarily, the mean absolute percentage error and secondarily, the mean bias error. In addition, as the metrics can yield misleading conclusions, forecasts will be visually analysed. A Harvey-Leybourne-Newbold hypothesis test will be performed with the purpose of testing a significant difference in the predictive accuracy of the metric superior forecast compared to the benchmark forecast.

1.5 Structure of the thesis

The structure of this thesis is as follows. Chapter (2) presents a literature review. Chapter (3) presents theoretical background to time series analysis. Chapter (4) presents data collection and software environment. Chapter (5) presents the methods used in this thesis. Chapter (6) presents the results. Chapter (7) presents discussions on the results and the methods of this thesis. Chapter (8) presents the major conclusions and suggestions of future work.

Chapter 2

Literature review

This chapter presents a literature review of related work to this thesis. Section (2.1) presents related work on time series forecasting models. Section (2.2) presents related work on anomaly detection in time series. Section (2.3) presents related work on time series forecasting during abnormal time periods.

2.1 Forecasting models

Methods used for time series forecasting have been well studied. In addition to time series regression models, e.g. the ARMA family of models presented in Section (3.1), classical forecasting approaches include exponential smoothing models, e.g. the Simple Exponential Smoothing and Holt-Winter's Exponential Smoothing (Hyndman and Athanasopoulos 2018). According to Hyndman and Athanasopoulos (ibid.), these two families of models are the most commonly used approaches to time series forecasting. However, unlike the ARIMA family that aims to describe the autocorrelations in a time series, the exponential smoothing models are based on a description of the series' trend and seasonality.

Further, the autoregressive conditional heteroskedasticity model, ARCH, is a time series model used for modeling changing volatility or volatility clustering* (Brockwell and Davis 2002). The ARCH process may be appropriate for a series if the error variance of the series follows an AR model. As an extension to this, the generalized ARCH process, GARCH, may be appropriate for a series if the error variance follows an ARMA process (Bollerslev 1986). Unlike the ARMA family that imposes a certain structure of the conditional mean of a series, the ARCH family imposes a certain structure

* That is, time periods of oscillations followed by time periods of calm.

of the conditional variance of a series.

As the field of machine learning has grown the past decade, neural networks have started to challenge the classical time series models. The recurrent neural network Long Short-Term Memory, LSTM, is recurring in papers on forecasting. For example, see Lin et al. (2021) on forecasting stock price index using LSTM and Yan et al. (2021) on forecasting air quality index using LSTM. However, in contrast with the models presented in this thesis, neural networks are not designed explicitly for time series data.

In this thesis, the ARIMA family of models will be used as a foundation. The ARIMA family can replicate the problem of interest and is suitable in terms of simplicity and interpretability.

2.2 Anomaly detection in time series

Time series outliers manifest in many ways and are often classified by their impact on a series. Common classifications of time series outliers include the level shift, the transient change and the additive outlier. For further elaboration on outlier classifications, see Subsection (4.1.4).

Tsay (1988) proposed a method for detecting additive outliers, level shifts and variance changes in time series, based on a least squares technique and residual variance ratios. The approach uses an iterative procedure of a specification-estimation-detection-removal cycle, to handle the most distinct disturbance in a series. The approach shows promising results in identifying observations that require additional attention. Quality attributes of the method include simplicity and robustness. A similar method could be useful in the setting of this thesis.

Yu et al. (2014) proposed an approach to anomaly detection based on detecting deviations from historical patterns. The approach is to build a forecasting model on previous observations, predicting future values and constructing prediction confidence intervals. Then, concluding if anomalies are present, based on if the observed values fall outside the prediction confidence interval. The method scales well and does not require pre-classification of anomalies. Consequently, quality attributes of the method include scaleability and robustness.

In this thesis, an approach similar to Yu et al. (ibid.) was used with unsatisfying results. Consequently, another approach, described in Chapter (5), was deemed sufficient for this thesis.

Furthermore, more complex methods for detecting time series anomalies include Isolation Forests, LSTM encoders and backpropagating networks[*]. However, due to complexity, these are outside the scope of this thesis.

2.3 Forecasting during abnormal events

Summarily, approaches for abnormal time period forecasting include feature engineering, imputation of abnormal data and using a combination of several time series models.

Chen, Yang, and Zhang (2020) address the issue of forecasting electrical loads during the COVID-19 pandemic by including new features in addition to the existing building block. As a result of social distancing restrictions related to the pandemic, power consumption profiles around the world have changed, complicating the procedure of forecasting electrical loads. For this purpose, the authors introduce a new feature, mobility, as a measure of economic activities on a population level. Due to small sample sizes of mobility data associated with the pandemic, the authors use a method of knowledge transfer between several geographical regions. In regards to forecasting volume of sales during the COVID-19 pandemic, a similar approach could be suitable. Here, features representing population movement patterns or dates of introducing and lifting social distancing restrictions could be appropriate to include. However, due to confidentiality related matters, the approach is not feasible for this paper; features as such could enable deriving confidential information about the data[†].

Further, Akouemo and Povinelli (2014) proposed a method for detecting and imputing anomalies in time series. The approach uses an iterative procedure of hypothesis testing on the most distinct residuals of an ARIMAX[‡] model in order to detect anomalies. Then, anomalies are replaced by a naive impute of the mean of past and previous timesteps. Further, the ARIMAX model is re-trained on the imputed, cleaner data set. Last, the re-trained model forecast a value of the anomaly to replace the naive impute with. The idea of imputing anomalies could be useful in forecasting during abnormal time periods. In this thesis, a similar idea to Akouemo and Povinelli (ibid.) will be used.

[*] For example, see Ding and Fei (2013), Kieu et al. (2019) and Vishwakarma, Paul, and Elsawah (2020).

[†] For example, confidential information such as the operating country and specific industry related to each data set.

[‡] That is, an ARIMA model that can include exogenous variables, e.g. other time series, as input variables.

Furthermore, for the purpose of increasing forecast accuracy, the approach of combining forecasting models can be used. A combination of forecasting models can be made in different ways. For example, in an additive manner of two similar models or in a complementary manner of two complementary models. In the additive manner, the approach is to use several methods on the same time series. Then, to obtain a resulting forecast by adding weights to the methods. Weighting can be done by a simple average (Hyndman and Athanasopoulos 2018) or by using more complex methods; e.g. Zou and Yang (2004) use a convex combination of sequentially updated weights and Bates and Granger (1969) determine weights based on past errors of each included forecast. In addition, time dependent weights are commonly used. The approach of using an additive combination of models could be appropriate in abnormal time period forecasting.

Another approach for combining forecasting models is to use methods of complementary nature, here referred to as using a hybrid model. Hybrid models often combine low-volatility ARIMA models with high-volatility ARCH models. For a time series showing periods of irregular behavior, a combination of ARIMA modeling of the series' mean and ARCH modeling of the series' variance could be appropriate. For example, Güngör, Ertuğrul, and Soytaş (2021) address the issue of forecasting Turkish gasoline consumption during the COVID-19 pandemic by adding volatility to the pre-pandemic best-fit ARMA model. First, the authors investigated volatility dynamics using various ARCH models. Then, the gasoline consumption volatility was chosen as the variance of the best-fit model. Last, the volatility variable was added to the pre-pandemic best-fit ARMA model.

In this thesis, two attempts to tackle abnormal time series forecasting will be outlined. First, an approach of imputing anomalies, similar to Akouemo and Povinelli (2014). Second, an approach of using an additive combination of time series models. The approach of feature engineering and the method of using a volatility combined model is also of interest. However, both are excluded in this thesis due to confidentiality related matters and to time limitations.

Chapter 3

Theoretical background

This chapter provides background information on time series analysis and is intended for the reader without prior knowledge on time series. Section (3.1) presents the foundations in time series analysis. Section (3.2) presents a framework to model selection. Section (3.3) presents methods used in forecast evaluation.

3.1 Time series analysis

This sections presents the most relevant ideas and tools for time series analysis at a high-level, as these constitute the foundation of the modeling done in this thesis. For more details on the theory, see Brockwell and Davis (2002).

3.1.1 Stochastic processes

A time series is a set of observations x_t of which each observation is being recorded at a specific time $t \in \mathbb{Z}$. The definition of a time series model is stated in Definition (3.1).

Definition 3.1. A time series model for the observed data $\{x_t\}$ is a specification of the joint distributions (or possibly only the means and co-variances) of a sequence of random variables $\{X_t\}$, of which $\{x_t\}$ is postulated to be a realization.

Throughout this paper, the term time series is used for both the data and the process of which it is a realization.

A time series can include several components such as trends, seasonal behaviour and cyclic behaviour. In order to split a series into components, one can use a decomposition model as in Definition (3.2).

Definition 3.2. The classical decomposition model of a time series $\{X_t\}$ is

$$X_t = m_t + s_t + Y_t, \tag{3.1}$$

where m_t is a trend component; s_t a seasonal component with known period d, i.e. $s_{t+d} = s_t, \sum_{j=1}^{d} s_j = 0$; Y_t a random noise component, $\mathbb{E}[Y_t] = 0$.

The classical decomposition model represents a time series with additive seasonality. In the case of multiplicative seasonality, i.e. a seasonal component of increasing magnitude, one can use a multiplicative decomposition of $\{X_t\}$. See Brockwell and Davis (2002) for further reference. Depending on the series, further components can be relevant to include in the decomposition; e.g. a cyclic component and a holiday component.

It is beneficial to start by plotting the time series in order to visually identify trends, seasonalities and cyclic behaviour. Also, to check for outliers and sharp changes of behavior in the series. In order to fit a time series model to $\{X_t\}$, a common approach is to first remove the trend and seasonal component (and cyclic or holiday component, if present) and then fit a model on the random noise component $\{Y_t\}$. After a model has been fitted on $\{Y_t\}$, the remaining components is added in order to represent the original time series $\{X_t\}$.

Time series models often assume the input data to be so called stationary. The decomposition of a series $\{X_t\}$ into residuals $\{Y_t\}$ can result in a stationary time series.

3.1.2 Stationarity of a time series

An assumption of several time series models is stationarity in the input data. A time series $\{X_t\}$ is stationary if it has similar statistical properties to the shifted time series $\{X_{t+h}\}, \forall h \in \mathbb{Z}$.

In order to formally define stationary, we first define the mean function and the covariance function of a series.

Definition 3.3. Let $\{X_t\}$ be a time series with finite second moment, Var$(X_t) < \infty$. The mean function of $\{X_t\}$ is

$$\mu_X(t) = \mathbb{E}[X_t]. \tag{3.2}$$

The covariance function of $\{X_t\}$ is

$$\gamma_X(r,s) = \text{Cov}(X_r, X_s) = E[(X_r - \mu_X(r))(X_s - \mu_X(s))], \quad \forall r, s \in \mathbb{Z}. \tag{3.3}$$

Definition 3.4. A time series $\{X_t\}$ is stationary if the following conditions are satisfied

$$1. \quad \mu_X(t) \text{ is independent of } t, \tag{3.4}$$

$$2. \quad \gamma_X(t+h,t) \text{ is independent of } t \text{ for each } h. \tag{3.5}$$

In other words, a time series is said to be stationary if the mean and covariance are independent of time. Several approaches can be used in order to transform a time series into a stationary time series. In addition to using the classical decomposition model in Definition (3.2), Subsection (3.1.7) elaborates on useful transformations and approaches for the purpose of achieving stationarity.

Several time series models can be fitted to a stationary time series. The subsequent subsections summarises some of them.

3.1.3 The Moving-average process, MA(q)

Definition 3.5. The process $\{X_t\}$ is said to be a moving-average process of order q, MA(q), if

$$X_t = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}, \tag{3.6}$$

where $\theta_1, .., \theta_q$ are constants and the process $\{Z_t\}_{t\in\mathbb{Z}} \sim WN(0, \sigma^2)$ is white noise. That is, Z_t is independent for all t and identically distributed with a mean of zero.

3.1.4 The Autoregressive process, AR(p)

Definition 3.6. The process $\{X_t\}$ is said to be an autoregressive process of order p, AR(p), if it is stationary and

$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = Z_t, \tag{3.7}$$

where $\{Z_t\}_{t\in\mathbb{Z}} \sim WN(0, \sigma^2)$ and $\phi_1, .., \phi_p$ are constants.

3.1.5 The Autoregressive-Moving-average process, ARMA(p, q)

Definition 3.7. The process $\{X_t\}$ is said to be an autoregressive-moving-average process of order p and q, ARMA(p, q), if it is stationary and

$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}, \tag{3.8}$$

where $\{Z_t\} \sim WN(0, \sigma^2)$ and $(1 - \phi_1 z - \cdots - \phi_p z^p) \neq (1 + \theta_1 z + \ldots \theta_q z^q), \forall z \in \mathbb{R}$.

Further, we will refer to $\phi(z) = (1 - \phi_1 z - \cdots - \phi_p z^p)$ as the autoregressive polynomial and $\theta(z) = (1 + \theta_1 z + \cdots + \theta_q z^q)$ as the moving-average polynomial.

For a note on existence and uniqueness, we state Theorem (3.1).

Theorem 3.1. A unique stationary solution $\{X_t\}$ to Equation (3.8) exists if and only if

$$\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p \neq 0, \qquad \forall |z| = 1. \tag{3.9}$$

3.1.6 Autocovariance and autocorrelation function

Time series have several important functions and characteristics. Two of them is the autocovariance function and the autocorrelation function.

Definition 3.8. Let $\{X_t\}$ be a stationary time series. The autocovariance function, ACVF, of $\{X_t\}$ at lag h, is

$$\gamma_X(h) = \text{Cov}(X_{t+h}, X_t). \tag{3.10}$$

The autocorrelation function, ACF, of $\{X_t\}$ at lag h is

$$\rho_X(h) = \frac{\gamma_X(h)}{\gamma_X(0)} = \frac{\text{Cov}(X_{t+h}, X_t)}{\text{Var}(X_t, X_t)}. \tag{3.11}$$

Both the ACVF and the ACF provide a measure of the degree of dependence of the values in a time series at different lags. When considering time series forecasting, predictions are based on previous and present values. It can thus be beneficial to understand the ACVF and ACF of the time series, as these indicate how many previous data points one should include in order to produce reasonable forecasts. The partial autocorrelation function, a slightly changed version of the ACF, is commonly used for the purpose of forecasting and determining model parameters.

Definition 3.9. The partial autocorrelation function, PACF, of an ARMA(p, q) process $\{X_t\}$, is the function $\alpha(\cdot)$ defined by

$$\begin{cases} \alpha(0) = 1, \\ \alpha(h) = \phi_{hh}, h \geq 1, \end{cases} \tag{3.12}$$

where ϕ_{hh} is the last component of

$$\phi_h = \Gamma_h^{-1} \gamma_h, \tag{3.13}$$

$\Gamma_h = [\gamma(i-j)]_{i,j=1}^h$ and $\gamma_h = [\gamma(1), .., \gamma(h)]'$.

For an elaboration on this definition, see Brockwell and Davis (2002).

3.1.7 Transforming a series to stationary

The use of the classical decomposition model of $\{X_t\}$ does not per se guarantee a stationary series of the residuals $\{Y_t\}$. The exclusion of the trend and seasonal component can be made in several ways; the appropriateness is highly dependent on the data set. In addition to estimating the components and subtracting them from the time series, two approaches for achieving stationarity is presented here: *differencing* and using a *logarithmic transformation* in prior to removing the trend and seasonal component. Last, a paragraph on stationary tests is presented.

Differencing

Differencing is the approach of applying a differencing operator repeatedly on a time series until the differenced observations resemble a realisation of a stationary time series. The lag-d differencing operator is stated in Definition (3.10).

Definition 3.10. The lag-d differencing operator ∇_d is defined by

$$\nabla_d X_t = X_t - X_{t-d}, \qquad d \in \mathbb{Z}_+. \tag{3.14}$$

Note that the differencing operator is a linear operator.

Example 3.1. If a time series has a linear trend component, $m_t = c_0 + c_1 t, c_0, c_1 \in \mathbb{R}$, de-trending can be made using the lag-1 differencing operator, as

$$\begin{aligned} \nabla_1 X_t &= \nabla_1 m_t + \nabla_1 s_t + \nabla_1 Y_t = && (3.15)\\ &= c_0 + c_1 t - (c_0 + c_1(t-1)) + \nabla_1 s_t + \nabla_1 Y_t = && (3.16)\\ &= c_1 + \nabla_1 s_t + \nabla_1 Y_t. && (3.17) \end{aligned}$$

That is, by applying the lag-1 differencing operator, the linear trend component transforms to a constant, independent of time.

Example 3.2. If a time series has a seasonal component of period d, de-seasonalising can be made using the lag-d differencing operator, as

$$\begin{aligned} \nabla_d X_t &= X_t - X_{t-d} = && (3.18)\\ &= m_t + s_t + Y_t - (m_{t-d} + s_{t-d} + Y_{t-d}) = && (3.19)\\ &= m_t - m_{t-d} + Y_t - Y_{t-d}, && (3.20) \end{aligned}$$

where the last equality holds as $\{s_t\}$ has period d. That is, by applying the lag-d differencing operator, the seasonal component of period d vanishes.

Then, the original time series is replaced by the de-trended and de-seasonalised series obtained through differencing. This enables us to use the theory of stationary time series upon modeling and analysis.

Logarithmic transformation

For time series showing a roughly linear increasing magnitude of the fluctuations with the level of the series, it is of interest to apply a logarithmic transformation to the series. That is,

$$\{X_1, \dots, X_n\} \xrightarrow{\ln} \{\ln X_1, \dots, \ln X_n\}.$$

The log-transformed time series will resemble a more constant magnitude of fluctuations in the series, which can result in a stationary time series.

Although, it is not necessarily sufficient to just apply a transformation to the series. In some cases, it is necessary to apply a log-transformation of the series in prior to performing de-trending and de-seasonalising through differencing.

Tests for stationarity

Hypothesis tests can be used to infer stationary behavior in time series. When de-trending and de-seasonalising a time series by differencing, it is of interest to test for the presence of a unit root in the series. A unit root in a series arises whenever the autoregressive or moving-average polynomial of an ARMA process has a root on or close to the unit circle. A root in the AR-polynomial suggests that the series should be differenced before fitting an ARMA model; a root in the MA-polynomial suggests that the series is overdifferenced. Two tests used for inferring stationarity are the augmented Dickey–Fuller (ADF) test and the Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test.

The ADF test is testing for the presence of a unit root in the autoregressive polynomial. The test is applied to a model on the form

$$\Delta x_t = \alpha + \beta t + \gamma x_{t-1} + \delta_1 \Delta x_{t-1} + ... + \delta_{p-1} \Delta y_{t-p+1} + \epsilon_t, \qquad (3.21)$$

where $\alpha, \beta \in \mathbb{R}$ and p is the lag order of the autoregressive polynomial. The ADF tests the null hypothesis of the presence of a unit root, $H_0 : \gamma = 0$; with the alternative hypothesis of the contrary, $H_a : \gamma < 0$. A test statistic $\text{DF} = \widehat{\gamma}/\text{SE}(\widehat{\gamma})$ is compared to the critical value of the test. If the test statistic is less than the critical value, the null hypothesis of a unit root is rejected. Equivalently, a p-value can be used. For a significance level α, if $p < \alpha$, the null hypothesis is rejected. If the null hypothesis is failed to be rejected, the ADF test indicates that the series is non-stationary and suggests further differencing. For further references, see Box et al. (2016).

As a complement to the unit root test, the KPSS test is introduced. The KPSS tests the null hypothesis that a time series is trend-stationary, against the alternative hypothesis of the presence of a unit root in the series. A trend-stationary series is stationary around a deterministic trend. That is, the series can be expressed as a sum of the deterministic trend, random walks and stationary error. The KPSS test is consequently the test of the hypothesis that the random walk process has zero variance. For further reference, see Kwiatkowski et al. (1992).

By testing for a unit root in ADF and trend-stationary behavior in KPSS, one can infer whether or not the series appear to be stationary. If we reject H_0 in the ADF test and fail to reject H_0 in the KPSS test, we infer that the time series is stationary.

3.1.8 The Autoregressive-Integrated-Moving-average process, ARIMA(p, d, q)

The ARIMA(p, d, q) model is a generalisation of the ARMA(p, q) model and enables modeling a wide range of non-stationary time series. In prior to defining the ARIMA process, we define the backward shift operator.

Definition 3.11. The backward shift operator B is defined as

$$B(X_t) = X_{t-1}. \tag{3.22}$$

The backward shift operator simplifies notation. For example, $B^j(X_t) = X_{t-j}, j \geq 1; \nabla X_t = X_t - X_{t-1} = (1 - B)X_t$.

Now, the ARIMA(p, d, q) process is stated in Definition (3.12).

Definition 3.12. The process $\{X_t\}$ is said to be an autoregressive-integrated-moving-average process, ARIMA(p, d, q), if $d \in \mathbb{Z}_+$ and

$$\tilde{X}_t = (1 - B)^d X_t \tag{3.23}$$

is a ARMA(p, q) process, fulfilling the condition

$$\phi(z) = 1 - \phi_1 z - \ldots \phi_p z^p \neq 0, \qquad \forall |z| \leq 1. \tag{3.24}$$

The ARIMA process enables us to model non-stationary time series that attains stationary upon applying the lag-1 differencing operator, d times. In other words, the ARIMA(p, d, q) process reduce to an ARMA(p, q) process when applying the lag-1 differencing operator, d times.

The Seasonal-Autoregressive-Integrated-Moving-average process, SARIMA(p, d, q)(P, D, Q)s

Example (3.2) illustrates how differencing of a time series $\{X_t\}$ at lag s vanishes a seasonal component of period s. Using the backward shift operator, the lag-s differenced series,

$$Y_t = (1 - B^s)X_t, \tag{3.25}$$

can then be fitted to an ARMA(p, q) model. Based on this idea, we extend the definition of the ARIMA process to include a seasonal part: the seasonal auto-regressive-integrated-moving-average process, SARIMA(p, d, q)(P, D, Q)s.

This is the final time series model we will introduce as this process will be used in modeling in Chapter (5). The SARIMA$(p,d,q)(P,D,Q)s$ process is stated in Definition (3.13).

Definition 3.13. The process $\{X_t\}$ is said to be a seasonal-autoregressive-integrated-moving-average process, SARIMA$(p,d,q)(P,D,Q)s$ with period s, if $d, D \in \mathbb{Z}_+$ and

$$\tilde{X}_t = (1-B)^d(1-B^s)^D X_t \tag{3.26}$$

is an ARMA(p,q) process, defined by

$$\phi(B)\Phi(B^s)\tilde{X}_t = \theta(B)\Theta(B^s)Z_t, \quad \{Z_t\} \sim WN(0,\sigma^2). \tag{3.27}$$

Here,

$$\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p, \qquad \Phi(z) = 1 - \Phi_1 z - \cdots - \Phi_P z^P, \tag{3.28}$$

$$\theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q, \qquad \Theta(z) = 1 + \Theta_1 z + \cdots + \Theta_Q z^Q \tag{3.29}$$

and

$$\phi(z) = 1 - \phi_1 z - \ldots \phi_p z^p \neq 0, \qquad \forall |z| \leq 1, \tag{3.30}$$

$$\Phi(z) = 1 - \Phi_1 z - \ldots \Phi_P z^P \neq 0, \qquad \forall |z| \leq 1. \tag{3.31}$$

The SARIMA$(p,d,q)(P,D,Q)s$ process reduce to an ARMA(p,q) process as stated in Definition (3.27) when applying the lag-1 differencing operator d times and the lag-s differencing operator D times.

3.2 Model selection

This section aims to present the procedure of model selection based on the Box-Jenkins method. The Box-Jenkins method applies to finding the best fit ARMA- and ARIMA model to a time series. For references, see Brockwell and Davis (2002), Hyndman and Athanasopoulos (2018), Box et al. (2016). This section is formed on modeling a SARIMA$(p,d,q)(P,D,Q)s$ process and further used upon choosing benchmark models in Chapter (5).

3.2.1 Model identification

The first step in model selection is to detect if the time series is non-stationary and to detect seasonality in the series. Also, if possible, to stationarize the series through differencing.

The period of seasonality, s, is often chosen by domain knowledge. Regarding the differencing parameters, d is chosen as the number of times the lag-1 differencing operator is applied to the time series in order to de-trend the series without over-differencing. The seasonal differencing parameter D is chosen as the number of times the lag-s differencing operator is applied to the series in order to de-seasonalise the series. Consequently, the series should appear stationary after applying the lag-1 differencing operator d times and the lag-s differencing operator D times.

3.2.2 Choice of parameters

The second step in model selection aims at identifying the order of lags of the autoregressive and moving-average polynomials; parameters p, q for the non-seasonal terms and parameters P, Q for the seasonal terms.

For the purpose of identifying these parameters, authors use different approaches. For example, in Brockwell and Davis (2002) an information criterion (AIC with correction) is used as the prime criterion for parameter selection of p, q. Another commonly used approach is to use the autocorrelation function, ACF, and partial autocorrelation function, PACF, as stated in Definitions (3.8-3.9). These functions exhibit a decaying behavior as the correlation in the series decreases with increased lag. Depending on how quickly the function value tends to zero, one obtains an indication of the number of terms to include in the autoregressive- and moving-average polynomials. A rule of thumb is to choose the order of the AR-polynomials, p and P, as the number of positive significant terms in the ACF and PACF plots. Further, to choose the order of the MA-polynomials, q and Q, in a corresponding manner, although as the number of negative significant terms.

Parameter criterion

Depending on the purpose of a forecast, a trade-off between reliability and interpretability is often inevitable. It may be more reliable to include higher order lagged terms in the AR- and MA-polynomials. However, this complicates interpretability. For this purpose, authors sometimes define a parameter criterion. For the SARIMA process, a commonly used parameter criterion is $p + d + q + P + D + Q \leq 6$.

3.2.3 Akaike Information Criterion, AIC

The Akaike Information Criterion, stated in Definition (3.14), can be used in parameter selection.

Definition 3.14. The Akaike Information Criterion, AIC, is defined as

$$\text{AIC} = 2k - 2(\hat{L}), \tag{3.32}$$

where k is the number of estimated parameters in the model; $\hat{L}$ the maximum likelihood for the model.

In the case of a time series model, the likelihood function is assumed to be Gaussian. Moreover, both k and $\hat{L}$ are functions of the model parameters p, d, q, P, D, Q. The AIC rewards goodness of fit in terms of likelihood and penalizes a higher number of estimated parameters. Given a set of candidate models, the preferred model is the one with the minimum AIC value.

The corrected Akaike Information Criterion, AICc, is AIC with correction for small sample sizes. Namely, $\text{AICc} = \text{AIC} + \frac{2k^2+2k}{n-k-1}$, where n is the sample size. However, it follows the same idea as AIC.

Neither AIC nor AICc tell anything about the absolute quality of a model; rather the relative quality in a set of models. After choosing a model by AIC or AICc, one should validate the absolute quality by testing the models predictions and by residual diagnostics.

3.2.4 Model estimation

The third step, after choosing model parameters $\{p, d, q, P, D, Q\}$, is to fit the model to the series by estimating the coefficients in the autoregressive and moving-average polynomials: $\{\phi_i\}_{i=1}^{p}, \{\Phi_j\}_{j=1}^{P}, \{\theta_k\}_{k=1}^{q}, \{\Theta_l\}_{l=1}^{Q}$. Statistical software can be used for this purpose. Common approaches to fitting time series models include maximum likelihood and least-squares estimation.

3.2.5 Residual diagnostics

In order to validate a time series model, residual diagnostics can be used. Diagnostics as such can include visualising the *standardized residuals* and checking if they behave as white noise; plotting a *histogram of the standardized* *residuals* along their estimated kernel density and the standard Gaussian distribution; checking a *normal Q-Q plot* of the residuals; and visualising the *ACF of the residuals* and checking for unpleasant residual correlation.

3.3 Forecast evaluation

This section aims to summarise techniques used in forecast evaluation, further used in Chapter (6). First, two concepts of forecasts are presented: the long-term forecast and the walk-forward forecast. Then, approximate prediction intervals are introduced along evaluation metrics for time series predictions. Last, hypothesis tests for inferring statistically significant difference between two forecasts are presented.

3.3.1 Long-term forecast

To begin with, in order to fit and evaluate a time series model, the series is split into a training and test set. As a time series has a structured order, it is of importance not to use a random split that shuffles the observations. Rather, it is necessary to split the series in chronologically order.

A long-term forecast is obtained by training a model on a (constant) training set, followed by predicting a forecast of a specified horizon. Hence, the model is trained only once upon producing a long-term forecast although more recent observations can become available to include in the training set.

3.3.2 Walk-forward forecast

In contrary to a long-term forecast, a walk-forward forecast does not use a constant training set. Rather, an expanding or sliding window of observations, for the purpose of including more recent data points in the training. The walk-forward expanding window fits a model on the training data; forecasts one-step-ahead; expands the training window to include the most recent observation; and then, iterates the procedure until the penultimate point of the test set has been included in the training set. The sliding window approach is similar: however, when including the most recent observation in the training set, the most preceding observation is excluded. Consequently, the size of the training set remains constant while the set is sliding through the series.

The walk-forward method is often used in evaluating forecasts as it tends to produce more accurate results compared to the long-term method.

In particular, walk-forward testing can be thought of as a cross-validation procedure for time series.

3.3.3 Prediction intervals

A prediction interval is an interval in which a future observation is expected to lie with a certain probability. As Hyndman and Athanasopoulos (2018) state, if assuming Gaussian forecast errors, a 95%- and 80% prediction interval for an h-step-ahead forecast is

$$\hat{y}_{t+h|t} \pm 1.96\hat{\sigma}_h, \qquad 95\% \text{ prediction interval,} \tag{3.33}$$
$$\hat{y}_{t+h|t} \pm 1.28\hat{\sigma}_h, \qquad 80\% \text{ prediction interval.} \tag{3.34}$$

Here, $\hat{y}_{t+h|t}$ is an estimated prediction of y_{t+h} given past observations; $\hat{\sigma}_h$ is an estimate of the standard deviation of the h-step-ahead forecast distribution.

As a prediction interval indicates the uncertainty in a forecast, it typically increases in amplitude with forecast horizon; σ_h increases with h. According to Hyndman and Athanasopoulos (ibid.), the residual standard deviation can provide a good estimate of the forecast standard deviation in the case of a one-step-ahead forecast. For a multi-step-ahead forecast, the authors state that more complicated method of calculation is required.

3.3.4 Evaluation metrics

Mean bias error, MBE

The mean bias error of a forecast is stated in Definition (3.15).

Definition 3.15. The bias error, e_t, of a forecast prediction $\hat{y}_t = \hat{y}_{t|t-1}$ of an observation y_t, is defined as

$$e_t = y_t - \hat{y}_t. \tag{3.35}$$

The mean bias error, MBE, of forecast predictions $\hat{y}_t = \hat{y}_{t|t-1}$, $t = 1, \ldots, h$, of observations y_t, is defined as

$$\text{MBE} = \frac{1}{h}\sum_{t=1}^{h} y_t - \hat{y}_t. \tag{3.36}$$

Mean absolute percentage error, MAPE

The mean absolute percentage error of a forecast is stated in Definition (3.16).

Definition 3.16. The mean absolute percentage error, MAPE, of forecast predictions $\hat{y}_t = \hat{y}_{t|t-1}$, $t = 1, \dots, h$, of observations y_t, is defined as

$$\text{MAPE} = \frac{1}{h} \sum_{t=1}^{h} \left| \frac{y_t - \hat{y}_t}{y_t} \right|, \quad y_t \neq 0. \tag{3.37}$$

3.3.5 Statistically comparing forecast accuracy

A hypothesis test can be used for inferring statistically significant difference between two forecasts' predictive accuracy.

The Diebold-Marino test, proposed by Diebold and Mariano (1995), can be used for comparing the predictive accuracy of two forecasts. To begin with, for each forecast $\hat{y}_i, i = 1, 2$, the forecast errors are defined as in Equation (3.35),

$$e_{it} = y_t - \hat{y}_{it}, \quad i = 1, 2. \tag{3.38}$$

Then, let $g(\cdot)$ denote a loss function associated to each forecast. For example, in the case of the absolute error loss, $g(e_{it}) = |e_{it}|$. Using this notation, a loss differential d_t is defined as

$$d_t = g(e_{1t}) - g(e_{2t}). \tag{3.39}$$

The Diebold-Marino test tests the null hypothesis $H_0 : \mathbb{E}[d_t] = 0, \forall t$; with the alternative hypothesis $H_a : \mathbb{E}[d_t] \neq 0$. In other words, the null hypothesis of equal accuracy of two forecasts; with the alternative hypothesis of the opposite. The test statistic DM is defined as

$$DM = \frac{\bar{d}}{\sqrt{\frac{\gamma_d(0) + 2\sum_{k=1}^{l-1} \gamma_d(k)}{h}}}, \tag{3.40}$$

where $h \geq 1$ is the forecast horizon; $\bar{d} = \sum_{t=1}^{h} d_t$; $\gamma_d(k)$ is the autocovariance function of d_k; $l = h^{1/3} + 1$. The test statistic is asymptotically standard Gaussian under the null hypothesis. Consequently, H_0 is rejected at a significance level α if

$$|DM| > z_{\alpha/2}, \tag{3.41}$$

where $z_{\alpha/2}$ is obtained by a z-table.

The Diebold-Marino test is commonly used; although, it tends to reject H_0

too often for small sample sizes. Improvements of small sample properties, proposed by Harvey, Leybourne, and Newbold (1997), resulted in a corrected test. Namely, the HLN test.

The Harvey-Leybourne-Newbold, HLN, test

Harvey, Leybourne, and Newbold (ibid.) proposed two improvements of the Diebold-Marino test resulting in the HLN test. First, to make a bias correction of the test statistic. Second, to compare the corrected statistic to a Student's t-distribution of $(n-1)$ degrees of freedom rather than the standard Gaussian. Using previous notation, the corrected test statistic HLN is defined as

$$HLN = DM\sqrt{\frac{h+1-2l+l(l-1)}{h}}. \tag{3.42}$$

Further, the null hypothesis is rejected at a significance level α if

$$|HLN| > t_{\alpha/2}, \tag{3.43}$$

where $t_{\alpha/2}$ is obtained by a t-table.

Chapter 4

Data and Software

This chapter aims to present the data and software environment used in this thesis. Section (4.1) presents the data. Section (4.2) presents the software environment.

4.1 Data

This section summarises the data collection used in this thesis including data wrangling and data anonymization. In addition, the three types of time series outliers observed in the data collection is presented.

4.1.1 Data collection

The data collection was provided by Klarna. The collection consisted of three data sets, each set constituted a time series of volume of sales for a merchant using the transaction services of Klarna. Due to the nature of the data, confidential information of each data set e.g. the merchant name, operating country of origin, industry of merchant, actual values of sales et.c. can not be provided. Furthermore, in a confidentiality related manner, the sample size of each data set can not be provided. Consequently, the start date of each time series is not stated.

A description of relevant columns of the data is presented in Table (4.1). Here, prior to data wrangling, the date column was of a daily frequency.

Column	**Description**
1	Name of merchant
2	Date [daily freq.]
3	Volume [currency]
...	...

Table 4.1: Description of each data set prior to data wrangling. Actual column names as well as irrelevant columns are excluded.

4.1.2 Data wrangling

For each data set, the data wrangling was performed as follows. To begin with, the first column as well as the irrelevant columns were removed. Then, the third column was aggregated on a weekly level. The weekly aggregation of volume was made in order to exclude cyclic weekly behavior, as it was assumed to be irrelevant for the purpose of this thesis. By following the procedure of data wrangling, each data set was transformed to two columns of a weekly frequency: one column for the date and one column for the aggregated volume of sales. See Table (4.2).

As each data set constitute a time series (a logged value and corresponding date) it will be referred to both as data set i and time series i, $i = 1, 2, 3$.

Column	**Description**
1	Date [weekly freq.]
2	Volume [currency]

Table 4.2: Description of each data set after data wrangling. Actual column names are excluded.

Training and test set

In prior to the modeling procedure described in Chapter (5), each time series was chronologically split in a training and test set in accordance to Subsection (3.3.1). As the sample size of each time series not will be stated in this paper, it is excluded to state the size of each training set. However, for each series, the training set contained observations until 31 January 2021. The corresponding test sets were chosen as observations between 01 February 2021 until 04 April 2021.

4.1.3 Data anonymization

In order to anonymize the data used in this thesis, each data set has been re-scaled by a constant. In addition, the magnitude and currency of the re-scaled volume of sales have been removed.

Also, as mentioned, the sample size of each time series is not stated. Figures throughout this paper only provide a (anonymized) subset of the total amount of data, for each data set.

4.1.4 Time series outliers in data collection

The three data sets used in this thesis illustrate time series exhibiting abnormal behaviors. Time series outliers as such are often classified on their impact to the series. The three time series used in this thesis exhibit three types of anomalies: namely, a level shift (LS), a transient change (TC) and an additive outlier (AO). Each set exhibits a distinct abnormal behavior.

First, a level shift represents an abrupt change in the mean level of a series, resulting in a new mean state after the outliers' occurrence. Figure (4.1) illustrates the impulse of a level shift on a series. Specifically, the first data set exhibits a level shift, further seen in Figure (5.1).

Second, a transient change represents a change in the series' trend: a trend-change to reaching a maximum/minimum, followed by a trend-change back to the pre-outlier level. Figure (4.2) illustrates the impulse of a transient change. Specifically, the second data set exhibits a transient change, further seen in Figure (5.2).

Third, an additive outlier represents an isolated short-lived spike in a time series. Note that this is a type of transient change, however of high magnitude and short time duration. Figure (4.3) illustrates the impulse of an additive outlier. Specifically, the third data set exhibits an additive outlier, further seen in Figure (5.3).

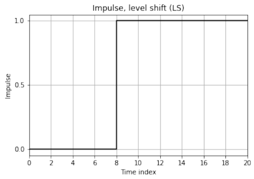

Figure 4.1: Impulse of a level shift.

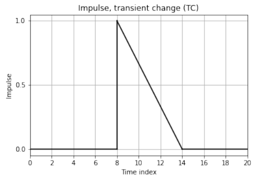

Figure 4.2: Impulse of a transient change.

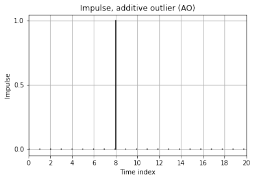

Figure 4.3: Impulse of an additive outlier.

4.2 Software environment

The programming language, libraries and packages used in this thesis are summarised in Tables (4.3-4.4).

Name	**Type**	**Area of use/Comment**
Python 3	Language	Core use
pandas	Library	Python data analysis
NumPy	Library	Numerical python
statsmodels	Package	Time series analysis
pmdarima	Library	Time series analysis

Table 4.3: Software environment used in this thesis.

Specific import	**Area of use/Comment**
statsmodels.tsa.statespace .sarimax.SARIMAX	Chapter (5): Models, forecasts, 95% prediction intervals
pmdarima.arima.auto_arima	Subsection (5.1.1): Iterative fitting a SARIMA and computing AICc

Table 4.4: Specific software imports used in this thesis.

The specific imports in Table (4.4) fits ARIMA- and SARIMA models by maximum likelihood estimation. See Subsection (3.2.4) for clarification on model estimation.

Chapter 5

Methods

This chapter presents the methods used for investigating the research question. Section (5.1) presents the modeling procedure of choosing benchmark models. Section (5.2) presents the approach used for detecting abnormal time periods. Section (5.3) presents Experiment A, imputing abnormal data. Section (5.4) presents Experiment B, combining two forecasting models.

The performance of the benchmark- and two experimental models are evaluated in Chapter (6).

5.1 Benchmark forecasting models

For each time series, the SARIMA process was used for the purpose of modeling a benchmark model. As the benchmark models will be used as a foundation throughout the thesis, is of importance to choose a valid model for each time series. In addition, a model that can replicate the problem of interest and the specific abnormal behavior.

The procedure of choosing benchmark SARIMA$(p, d, q)(P, D, Q)s$ models is presented in Subsection (5.1.1) and performed in accordance to the Box-Jenkins methodology outlined in Section (3.2). Illustrations of the problem of interest are presented in Subsection (5.1.2).

5.1.1 Modeling procedure of choosing benchmark models

To begin with, each time series was visualised against its rolling mean and rolling standard deviation. Each series manifested a non-stationary behavior as the mean and variance were time dependent. By using an ADF and KPSS test, described in Subsection (3.1.7), non-stationary behavior was inferred. Furthermore, the third data set expressed an increasing variance; consequently, a logarithmic transformation was used in prior to de-trending and de-seasonalising the series. The idea behind using a transformation as such is described in Subsection (3.1.7). Further, by domain knowledge, the seasonality in the series was chosen as yearly; consequently, $s = 52$.

The procedure of differencing, described in Subsection (3.1.7), was used for the purpose of transforming each series to stationary. Examples of de-trending and de-seasonalising a series by differencing are illustrated in Examples (3.1-3.2). An explanation on choosing differencing parameters d and D is outlined in Subsection (3.2.1). To each data set, the lag-1 differencing operator, as well as the lag-52 seasonal differencing operator, was applied one time. A visualisation of each differenced series indicated stationary behavior. By testing the ADF and KPSS again, stationary behavior was inferred. Consequently, it was concluded that $d = 1$ and $D = 1$ was sufficient for transforming each series to stationary.

In order to find appropriate parameters p, q, P, Q, the autocorrelation function, ACF, and partial autocorrelation function, PACF, of each stationary series were computed. The definitions of the ACF and PACF are stated in Definitions (3.8-3.9); the importance of them is described in the subsequent paragraph. By following the rule of thumb of parameter selection described in Subsection (3.2.2), initial parameters of p, q, P, Q were chosen.

To find an appropriate combination of parameters, it is of interest to iterate several parameter combinations and compute an information criterion of the model. Here, the corrected Akaike Information Criterion, stated in Subsection (3.2.3), was used. For each time series, an iterable range of each parameter p, q, P, Q was chosen in accordance with the initial parameter obtained by the ACF and PACF. For each parameter p, q, P, Q, the range was chosen as $\{0, 1, 2\}$. Then, in an iterative procedure, a SARIMA($p, 1, q$)($P, 1, Q$)52 model was fitted and AICc computed. For each data set, model selection was based on lowest AICc in addition to the parameter criterion $p + d + q + P + D + Q \leq 6$, described in Subsection (3.2.2).

Last, upon selecting a model by AICc, residual diagnostics were checked. As explained in Subsection (3.2.3), it is necessary to validate the absolute quality of a model chosen by AICc. Residual diagnostics were checked in accordance to Subsection (3.2.5). For each data set, the standardized residuals indicated approximately white noise behavior and the estimated residual density exhibited a Gaussian looking distribution. In addition, a normal Q-Q plot indicated that the sample quantiles approximately followed theoretical quantiles and a correlogram of residuals indicated no or low residual correlation. Consequently, residual diagnostics indicated valid model selection for each data set.

By following the previously outlined modeling procedure, a benchmark SARIMA$(p, d, q)(P, D, Q)s$ model was chosen for each time series. The models are presented in Table (5.1).

	Benchmark model
Data set 1	SARIMA(1,1,0)(2,1,0)52
Data set 2	SARIMA(1,1,2)(1,1,0)52
Data set 3	SARIMA(2,1,0)(1,1,0)52

Table 5.1: Benchmark SARIMA model for each time series.

Although model selection are formed on a famous theoretical framework, the benchmark models fail to produce reliable forecasts for the subsequent year(s). Illustrations of the problem of interest are presented in Subsection (5.1.2).

5.1.2 Illustration of problem

Figures (5.1-5.3) illustrate long-term forecasts estimated by the benchmark SARIMA models in Table (5.1). As seen in these figures, the benchmark models yield unreliable forecasts; they replicate abnormal behavior related to the COVID-19 outbreak in the subsequent year(s).

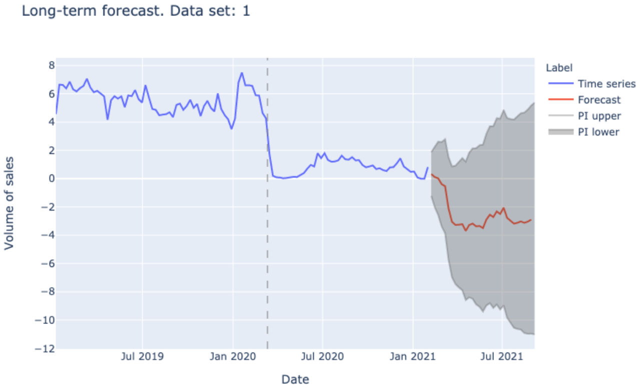

Figure 5.1: Illustration of problem: Data set 1. Benchmark SARIMA model forecasts a level shift related to the COVID-19 outbreak. Vertical line: 11 March 2020 (WHO declares a pandemic). Grey area: 95% prediction interval.

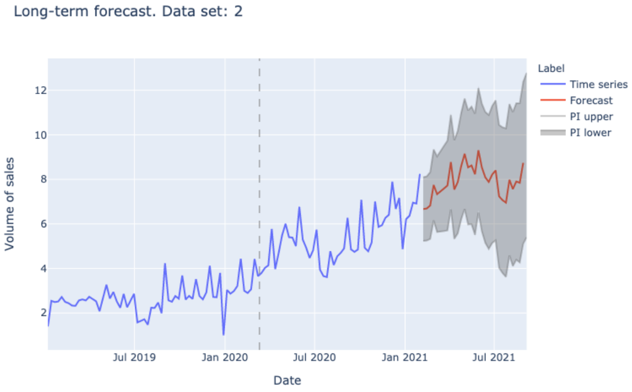

Figure 5.2: Illustration of problem: Data set 2. Benchmark SARIMA model forecasts a transient change related to the COVID-19 outbreak. Vertical line: 11 March 2020 (WHO declares a pandemic). Grey area: 95% prediction interval.

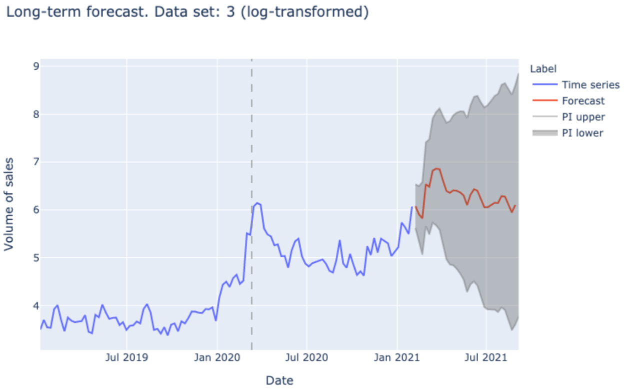

Figure 5.3: Illustration of problem: Data set 3. Benchmark SARIMA model forecasts an additive outlier related to the COVID-19 outbreak. Vertical line: 11 March 2020 (WHO declares a pandemic). Grey area: 95% prediction interval.

5.2 Anomaly detection

In order to tackle the problem of unreliable forecasts, it is of interest to detect the time period of anomalies assumed related to the COVID-19 pandemic. The approach of abnormal time period detection is presented in Subsection (5.2.1). The chosen approach, although it could be made more robust, is simple and sufficient for the purpose of this thesis.

Several approaches of less success were also tested. First, a similar approach to Yu et al. (2014), obtaining a forecast based on pre-pandemic observations and predicting the abnormal time period. Then, comparing the predictions to actual observations and inferring anomalies based on the predictions' deviation to the series. Second, an approach of using a rolling Tukey filter, by transforming the series to Gaussian and using interquartiles to detect outliers on a rolling basis. Third, an approach based on computing a difference of two subsequent observations in the series, then comparing the differences throughout the series. Fourth, an approach based on Isolation Forest.

Illustrations of the detected abnormal time periods are presented in Subsection (5.2.2).

5.2.1 Modeling procedure of detecting abnormal time period

To begin with, it is of importance to recognise that the time series in Figures (5.1-5.3) not necessary return to pre-pandemic normal levels. This complicates finding a time point corresponding to the end of the abnormal time period; it is not clear when an abnormal time period has transitioned to a (new) normal state. Consequently, it is necessary that the detection algorithm considers the previous level in the series as well as the current. These indicate when the series starts deviating globally from prior behavior and when the series reaches an assumed (new) normal state as it stops deviating locally from nearby observations. Relying on this idea, the method of abnormal time period detection used for each series was designed as follows.

For notation, let each time series be denoted $\{(x_i, y_i)\}_{i=1}^{n}$, where n is the length of the series, x_i the index of the series at point i and y_i the series value at index i. Note that, using this notation, $x_i = i, \forall i$.

First, a rolling mean μ_i was computed at each index i. The rolling mean was computed on w weeks, where w was chosen according to domain knowledge as $w \in \{4, 8\}$ dependent on the specific series. Note that w is somewhat related to how quickly an assumed (new) normal state can be achieved for the specific merchant. The rolling mean μ_i at index i was computed as

$$\mu_i = \frac{1}{w} \sum_{j=i-w+1}^{i} y_j, \qquad i \in [w, n]. \tag{5.1}$$

The mean constitutes an indicator of the level in the series. As it is computed on a rolling basis of w weeks, it reflects both the prior and current level in the series. Figures (5.4, 5.6, 5.8) present the time series with corresponding rolling mean.

Then, for each time series observation y_i, the distance to the rolling mean μ_i was defined as

$$|y_i - \mu_i|, \qquad i \in [w, n]. \tag{5.2}$$

The absolute value is of importance; as the rolling mean is slightly horizontal shifted, sign shifts of the difference $y_i - \mu_i$ occur ambiguously.

Further, it is crucial to compute the distance in Equation (5.2) in relation to the series' level, as the level in the series fluctuate in magnitude over time. Thus, for each observation y_i, the relative distance to the rolling mean μ_i was defined as

$$\frac{|y_i - \mu_i|}{\mu_i}, \qquad i \in [w, n]. \tag{5.3}$$

Last, to get a percentage relative distance, the quantity in Equation (5.3) was multiplied by 100. To this end, the quantity

$$cv_{approx,i} = 100 \cdot \frac{|y_i - \mu_i|}{\mu_i}, \qquad i \in [w, n], \tag{5.4}$$

was computed at each index $i \in [w, n]$ for each time series. The quantity cv_{approx} will be referred to as the approximate coefficient of variation, as it resembles a classical coefficient of variation $cv = \sigma/\mu$. Although, in contrary to cv, that express the extent of variability in relation to the mean of a population; $cv_{approx,i}$ aims to capture the extent of variability of observation i in relation to the nearby region. This quantity was used as an indicator for finding a start- and end date of the abnormal time period in each time series. Figures (5.5, 5.7, 5.9) present the approximate coefficients of variation for the three series.

Note that, in order for this approach to be feasible, the rolling mean μ_i should not tend to zero; then $cv_{approx,i}$ will tend to infinity. This is a concern of the first data set as some observations tend to zero in succession. In this case, the outlined approach was tested several times: both on the original data

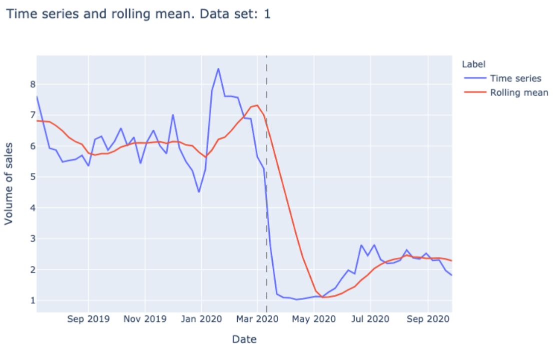

Figure 5.4: Time series and rolling mean, data set 1. Rolling mean is computed on $w = 8$ weeks. Grey line: 11 March 2020 (WHO declares a pandemic).

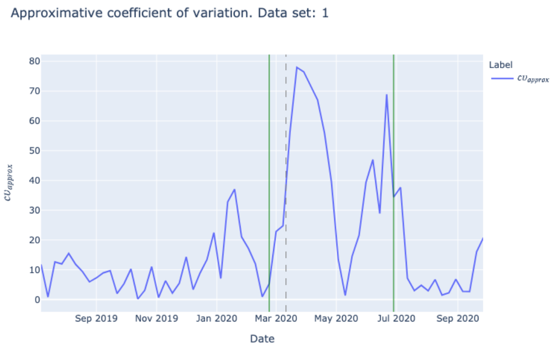

Figure 5.5: Approximate coefficient of variation, data set 1. Green lines: Start- and end date of abnormal time period. Grey line: 11 March 2020.

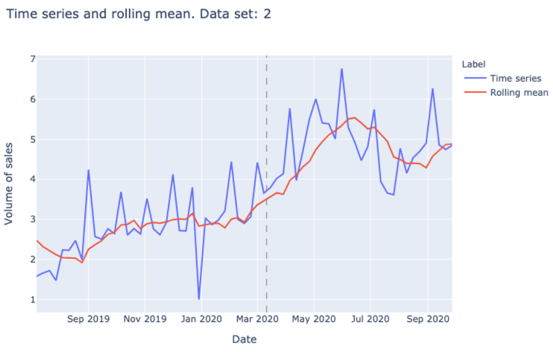

Figure 5.6: Time series and rolling mean, data set 2. Rolling mean is computed on $w = 8$ weeks. Grey line: 11 March 2020 (WHO declares a pandemic).

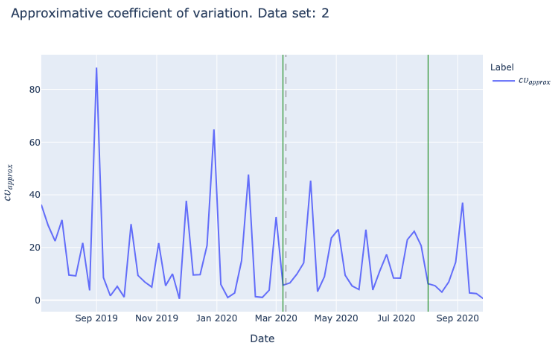

Figure 5.7: Approximate coefficient of variation, data set 2. Green lines: Start- and end date of abnormal time period. Grey line: 11 March 2020.

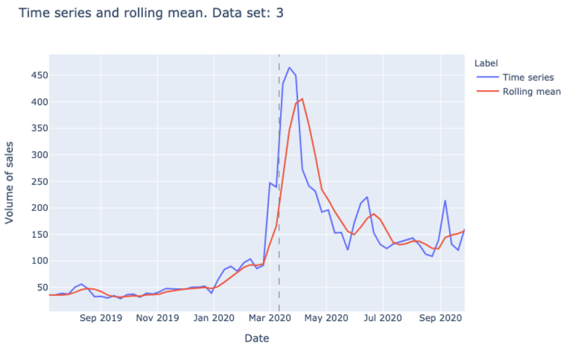

Figure 5.8: Time series and rolling mean, data set 3. Rolling mean is computed on $w = 4$ weeks. Grey line: 11 March 2020 (WHO declares a pandemic).

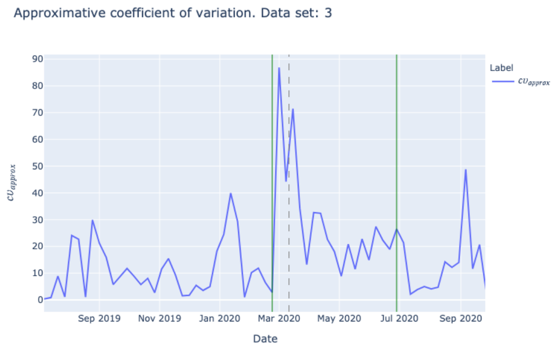

Figure 5.9: Approximate coefficient of variation, data set 3. Green lines: Start- and end date of abnormal time period. Grey line: 11 March 2020.

set and on a set slightly increased in magnitude (such that μ_i did not tend to zero, rather to one). The small increase in magnitude can be seen if comparing Figure (5.4) and Figure (5.1). However, the conclusions of the start- and end date of the abnormal time period were consistent in the different tests.

Choosing start- and end date of abnormal time period

The approximate coefficient of variation was used as an indicator for choosing start- and end date of each abnormal time period.

For the first time series in Figure (5.5), cv_{approx} expresses an increasing behavior in the end of February 2020 followed by high magnitude several weeks in succession until the beginning of July 2020. This indicates that an abrupt behavior occurs in the end of February; followed by high variability in succession until July when the series reaches an assumed (new) normal state. Consequently, it was concluded that the abnormal time period starts around the end of February and ends around the beginning of July.

For the second series in Figure (5.7), cv_{approx} is not as efficient. By domain knowledge of this series, it is reasonable to expect the variability to peak monthly. In Figure (5.7), cv_{approx} expresses an assumed irregular behavior in the beginning of March 2020. Even though it is not very distinct, it could be argued that the variability should tend to zero at this time point. Further, the irregular behavior continues as the variability peaks tend to increase in width until the beginning of August 2020. The width-increased peaks indicate high variability several weeks in succession; contradicting the assumed monthly pattern of variability peaks. Consequently, it was concluded that the abnormal time period starts around the beginning of March and ends around the beginning of August.

For the third series in Figure (5.9), cv_{approx} expresses an increasing behavior in the end of February 2020 followed by high variability in succession until the beginning of July 2020. Consequently, it was concluded that the abnormal time period starts around the end of February and ends around the beginning of July.

The outlined approach enabled pinpointing the abnormal time period assumed related to the COVID-19 outbreak in each time series. The chosen start- and end dates of the abnormal time periods are presented in Table (5.2).

	Start date	**End date**
Data set 1	23 Feb 2020	28 Jun 2020
Data set 2	08 Mar 2020	02 Aug 2020
Data set 3	23 Feb 2020	28 Jun 2020

Table 5.2: Start- and end date of each abnormal time period assumed related to the COVID-19 outbreak.

Illustrations of the detected abnormal time period of each time series are presented in Subsection (5.2.2). Upon pinpointing the time periods, the two experiments are carried out.

5.2.2 Illustration of detected abnormal time period

Figures (5.10-5.12) illustrate the time series and corresponding abnormal time periods.

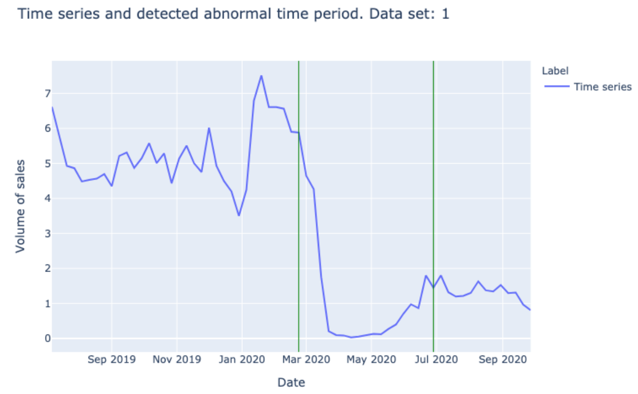

Figure 5.10: Time series and corresponding abnormal time period, data set 1. Green lines: Start- and end date of abnormal time period.

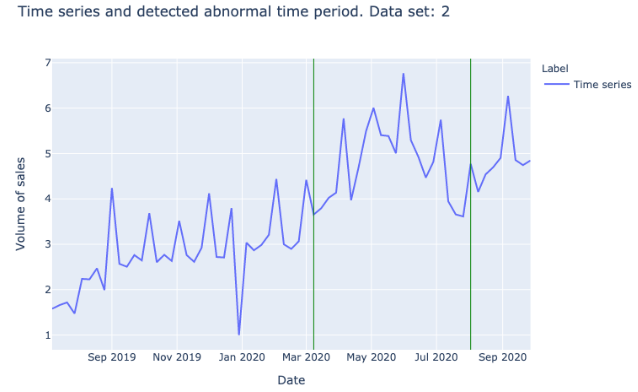

Figure 5.11: Time series and corresponding abnormal time period, data set 2. Green lines: Start- and end date of abnormal time period.

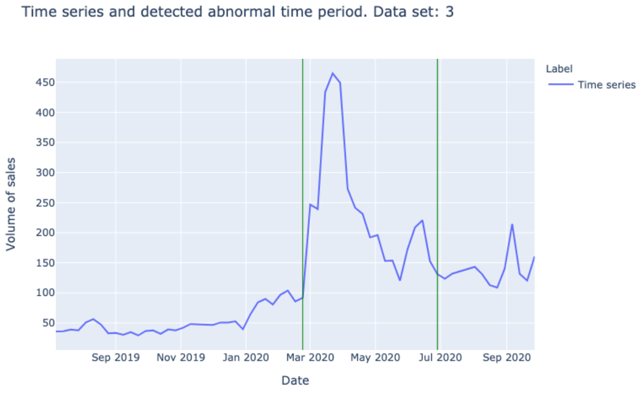

Figure 5.12: Time series and corresponding abnormal time period, data set 3. Green lines: Start- and end date of abnormal time period.

5.3 Experiment A: Imputing data

The first experiment in tackling unreliable forecasts relates to imputing the observations of the abnormal time period in each time series. The imputation is performed in two ways. First, by a linear interpolation, presented in Subsection (5.3.1). Second, by a combination of a linear interpolation and time series predictions, presented in Subsection (5.3.2).

5.3.1 Linear interpolation

The first attempt of imputing abnormal observations was by a linear interpolation.

Following previous notation, let each series be denoted $\{(x_i, y_i)\}_{i=1}^{n}$ and the corresponding abnormal time period $\{(x_k, y_k)\}_{k=j}^{l}, 1 < j, l < n$. By this notation, (x_j, y_j) denotes the start point of the abnormal time period and (x_l, y_l) the end point. The linear impute $\tilde{y}_k$ at index k was computed as

$$\tilde{y}_k = y_j + (x_k - x_j)\frac{y_l - y_j}{x_l - x_j}, \quad \forall k \in [j+1, l-1]. \tag{5.5}$$

Then, for each time series, the observations within the abnormal time period were replaced by the linear imputes,

$$y_k = \tilde{y}_k, \qquad \forall k \in [j+1, l-1]. \tag{5.6}$$

Figures (5.13, 5.15, 5.17) illustrate the linear impute of each time series.

5.3.2 Linear interpolation and time series predictions

The second attempt of imputing abnormal observations was by a combination of a linear interpolation and time series predictions.

First, the linear impute $\tilde{y}_k$ was computed as in Equation (5.5). Then, by forecasting the abnormal time period using the benchmark SARIMA model, time series predictions of the abnormal time period were estimated. Following previous notation, let $\{x_k, p_k\}_{k=j+1}^{l-1}$ denote the predictions, where p_k is the prediction at index k. Last, the amplitude a at the start of the abnormal time period was computed, $a = a(x_j) = y_j$.

Further, a combination of the linear interpolation and time series predictions was computed. The combined impute $\hat{y}_k$ at index k was computed as

$$\hat{y}_k = \tilde{y}_k + p_k - y_j, \qquad \forall k \in [j+1, l-1], \tag{5.7}$$

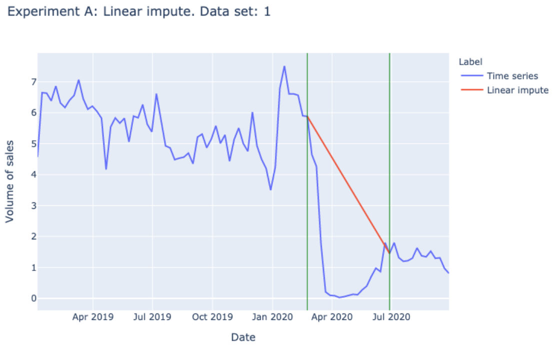

Figure 5.13: Experiment A: Linear impute, data set 1. Green lines: Start- and end date of abnormal time period.

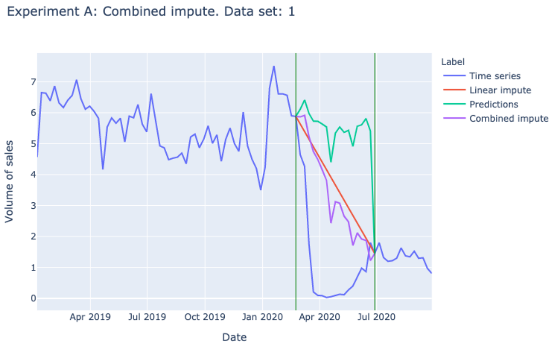

Figure 5.14: Experiment A: Combined impute, data set 1. Green lines: Start- and end date of abnormal time period.

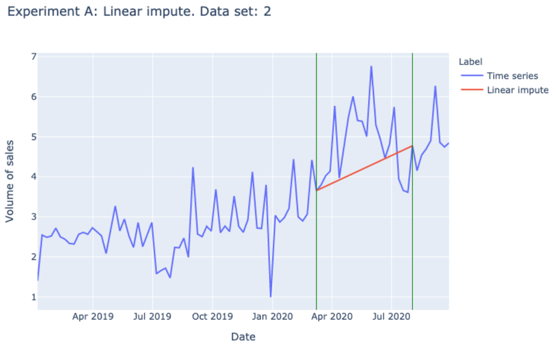

Figure 5.15: Experiment A: Linear impute, data set 2. Green lines: Start- and end date of abnormal time period.

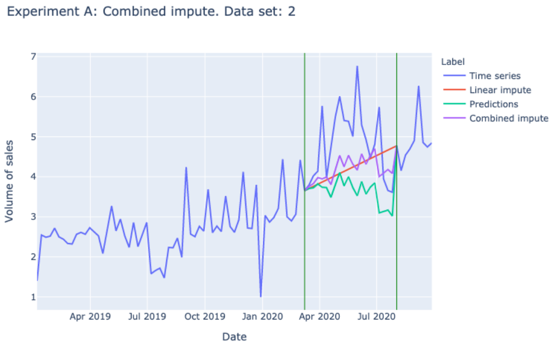

Figure 5.16: Experiment A: Combined impute, data set 2. Green lines: Start- and end date of abnormal time period.

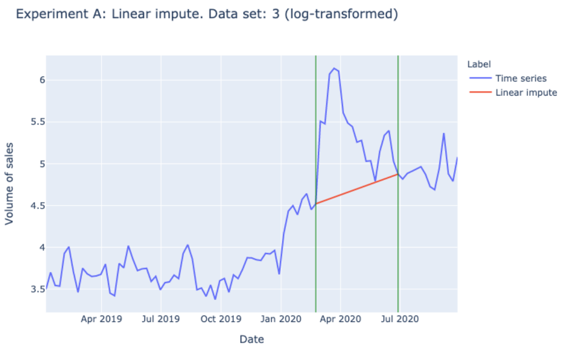

Figure 5.17: Experiment A: Linear impute, data set 3. Green lines: Start- and end date of abnormal time period.

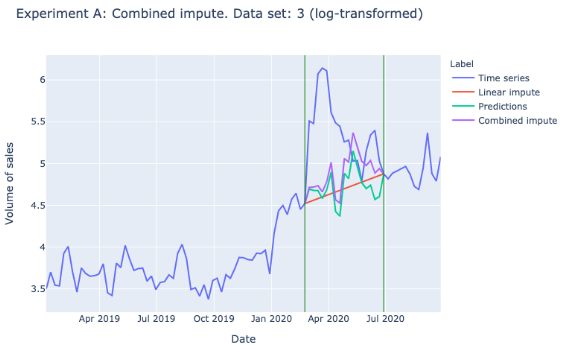

Figure 5.18: Experiment A: Combined impute, data set 3. Green lines: Start- and end date of abnormal time period.

where $\tilde{y}_k$ is the linear impute, p_k is the time series prediction and y_j is the amplitude of the abnormal time period. Consequently, $\hat{y}_k$ resembles a combination of a linear interpolation, time series predictions and an amplitude; an attempt to project the predictions on the linear interpolant.

Then, for each time series, the observations within the abnormal time period were replaced by the combined imputes,

$$y_k = \hat{y}_k, \qquad \forall k \in [j+1, l-1]. \tag{5.8}$$

Figures (5.14, 5.16, 5.18) illustrate the combined impute of each time series.

5.4 Experiment B: Combining forecasting models

The second experiment in tackling unreliable forecasts relates to combining forecasting models. For the purpose of combining models, the idea is to use several methods on a series and obtain a resulting forecast by adding weights to the methods. When using two forecasting methods, a combined model can be written as

$$a \cdot p_1 + b \cdot p_2, \qquad a, b \in \mathbb{R}, \tag{5.9}$$

where a, b are the weights and $p_i, i = 1, 2$ the forecast of the ith model.

In this experiment, the combined model constitutes two models of different time periods: one in prior to the abnormal time period and one subsequent the abnormal time period. The purpose of using a combination as such is to capture both the historical and current behavior in a series, without including the abnormal event.

The procedure of combining forecasting models is presented in Subsection (5.4.1). The procedure of deriving corresponding approximate prediction intervals is presented in Subsection (5.4.2).

5.4.1 Combining forecasting models

Following previous notation, let each series be denoted $\{(x_i, y_i)\}_{i=1}^{n}$ and corresponding abnormal time period $\{(x_k, y_k)\}_{k=j}^{l}$. Further, assume that the forecast horizon is $h \geq 1$. In other words, let the estimated predictions of interest be denoted $\{(x_i, y_i)\}_{i=n+1}^{n+h}$.

Forecast model of historical behavior

For each time series, a forecast aimed to capture the historical behavior of the series was computed. The forecast was obtained by the benchmark SARIMA model fitted to observations in prior to the abnormal time period $\{(x_i, y_i)\}_{i=1}^{j}$. By this procedure, a forecast of horizon $n-j+h$ was computed. Following previous notation, let the forecast be denoted $\{x_i, p_{1,i}\}_{i=j+1}^{n+h}$, where $p_{1,i}$ denotes the estimated forecast prediction at index i. Then, the pre-abnormal forecast, referred to as p_1, is denoted

$$p_1 = \{(x_i, p_{1,i})\}_{i=j+1}^{n+h}. \tag{5.10}$$

Figures (5.19, 5.21, 5.23) illustrate the pre-abnormal forecasts. Here, it can be seen that the forecasts reflect the historical behavior of the series that occurred in prior to the abnormal time period.

For further notation, the (benchmark) SARIMA$(p, d, q)(P, D, Q)m$ model used in obtaining the pre-abnormal forecast will be referred to as the first model in the combined model of forecasts.

Forecast model of current behavior

For each time series, a forecast aimed to capture the current behavior of the series was computed. Due to a small sample size of observations post the abnormal time period, the benchmark SARIMA$(p, d, q)(P, D, Q)s$ model could not successfully be fitted to the observations. As the sample size $n - l$ is less than the inferred seasonality m, the seasonal term of the SARIMA model was excluded. Consequently, the forecast was obtained by the corresponding ARIMA(p, d, q) model fitted to observations subsequent the abnormal time period $\{(x_i, y_i)\}_{i=l}^{n}$. By this procedure, a forecast of horizon h was computed. Following previous notation, let the forecast be denoted $\{x_i, p_{2,i}\}_{i=n+1}^{n+h}$, where $p_{2,i}$ denotes the estimated forecast prediction at index i. Then, the post-abnormal forecast, referred to as p_2, is denoted

$$p_2 = \{(x_i, p_{2,i})\}_{i=n+1}^{n+h}. \tag{5.11}$$

Figures (5.20, 5.22, 5.24) illustrate the post-abnormal forecasts. Here, it can be seen that the forecasts fail to reflect the current pattern of the series subsequent the abnormal time period. Due to a small sample size, they rather reflect the current level of each series.

For further notation, the ARIMA(p, d, q) model used in obtaining the post-abnormal forecast will be referred to as the second model in the combined model of forecasts.

Assumption on seasonal pattern post the COVID-19 outbreak

Due to a small sample size of observations subsequent the abnormal time period, it is impossible to infer a (new) seasonal pattern for each series post the pandemic outbreak. For this purpose, an assumption is stated: for each series, the seasonal pattern in prior to the outbreak resembles the seasonal pattern subsequent the abnormal time period.

As seen in Figures (5.19, 5.21, 5.23), the pre-abnormal forecasts reflect the historical pattern in each series, constituting the seasonal pattern in prior to the outbreak. As seen in Figures (5.20, 5.22, 5.24), the post-abnormal forecasts reproduce the current level in each series, constituting the level subsequent the outbreak. Now, a combination of the two forecasts is constructed.

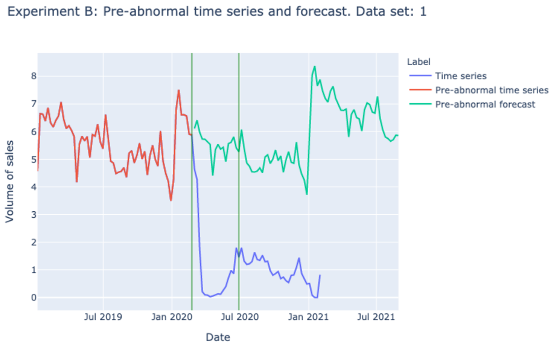

Figure 5.19: Experiment B: Pre-abnormal forecast, p_1, data set 1. Green lines: Start- and end date of abnormal time period.

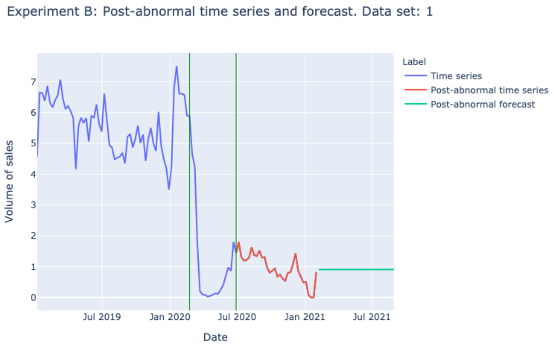

Figure 5.20: Experiment B: Post-abnormal forecast, p_2, data set 1. Green lines: Start- and end date of abnormal time period.

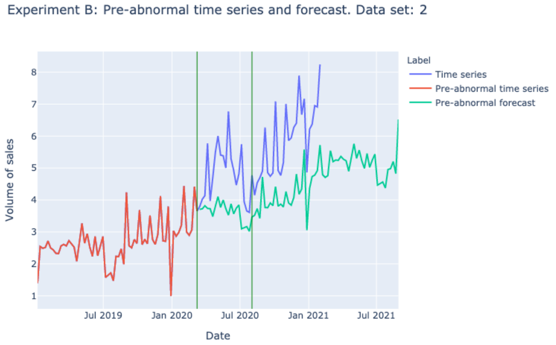

Figure 5.21: Experiment B: Pre-abnormal forecast, p_1, data set 2. Green lines: Start- and end date of abnormal time period.

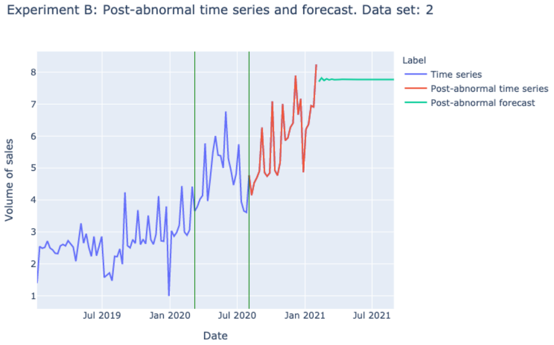

Figure 5.22: Experiment B: Post-abnormal forecast, p_2, data set 2. Green lines: Start- and end date of abnormal time period.

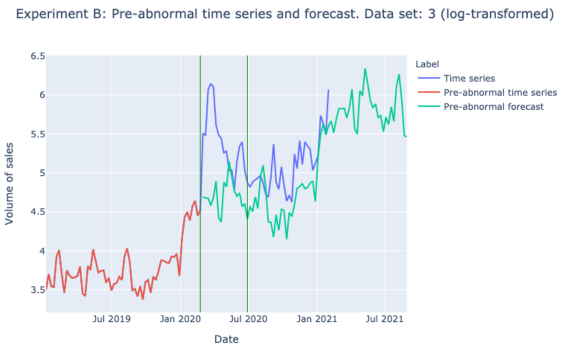

Figure 5.23: Experiment B: Pre-abnormal forecast, p_1, data set 3. Green lines: Start- and end date of abnormal time period.

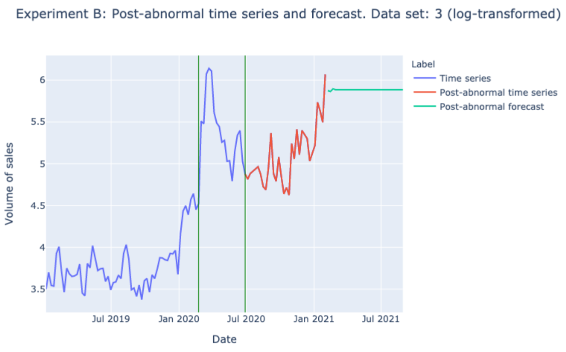

Figure 5.24: Experiment B: Post-abnormal forecast, p_2, data set 3. Green lines: Start- and end date of abnormal time period.

Convex combination of forecasting models

The first attempt of combining forecasting models was by a convex combination of forecasts.

For each time series, assuming the forecast horizon h, the convex combination of forecasts $\hat{y}_{convex,i}$ at index i was computed as

$$\hat{y}_{convex,i} = a \cdot p_{1,i} + (1-a) \cdot p_{2,i}, \qquad \forall i \in [n+1, n+h], \quad a \in [0,1]. \tag{5.12}$$

Here, a is the scaling parameter. For a convex combination, $a \in [0,1]$. Figures (5.25, 5.27, 5.29) illustrate the combination of forecasts for some values of a. In these figures, it can be seen that the combined forecast reflects both the historical and current behavior of a series. Though, the resulting forecast level is not necessarily appropriate.

Corrected combination of forecasting models

The second attempt of combining forecasting models was by a corrected combination of forecasts.

As the convex combination of forecasts can yield an inappropriate resulting forecast level, the combination was extended to include a projection term. For notation, let μ_m denote the mean of the forecast obtained from model m,

$m = 1, 2$, defined as

$$\mu_m = \frac{1}{h} \sum_{i=n+1}^{n+h} p_{m,i}, \qquad m = 1, 2. \tag{5.13}$$

For each time series, assuming the forecast horizon h, the corrected combination of forecasts $\hat{y}_{corrected,i}$ at index i was computed as

$$\hat{y}_{corrected,i} = a \cdot p_{1,i} + (1 - a) \cdot p_{2,i} + (-1)^{I\{\mu_1 > \mu_2\}} \cdot a \cdot |\mu_1 - \mu_2|, \tag{5.14}$$

for each $i \in [n + 1, n + h]$*. Here, $I\{\cdot\}$ is an indicator function,

$$I\{a > b\} = \begin{cases} 1, & a > b, \\ 0, & a \leq b. \end{cases} \tag{5.15}$$

Figures (5.26, 5.28, 5.30) illustrate the corrected combination of forecasts for some values of a. In these figures, it can be seen that the corrected combined forecast results in a more appropriate forecast level.

As the corrected combined forecast not necessarily is a convex combination, a is not limited to $a \in [0, 1]$. To this end, a static parameter $a \in \mathbb{R}_{>0}$ was experimented with. The choice of a static parameter is elaborated in Subsection (7.2.1) along a discussion on a potential time dependent parameter.

Experimenting with scaling parameter

As stated, the corrected combination of forecasts in Equation (5.14) aims to reflect the historical behavior in p_1 and current behavior in p_2. Consequently, the scaling parameter a resembles a measure of the degree of extent to weight the historical (seasonal) behavior in future estimates. For the purpose of choosing a, the standard deviation of each series pre- and post the abnormal time period are stated.

The standard deviation of a time series post the abnormal time period $\{(x_i, y_i)\}_{i=l}^{n}$, is defined as

$$\sigma_{post} = \sqrt{\frac{1}{n - l + 1} \sum_{i=l}^{n} (y_i - \mu_{post})^2}. \tag{5.16}$$

* Note that, in the case of a long-term forecast of the second model: $\mu_2 \approx p_{2,i}, \forall i$. Thus, Equation (5.14) reduce to $\hat{y}_{corrected,i} = a \cdot p_{1,i} + p_{2,i} - a \cdot \mu_1, \forall i$. However, in the case of a walk-forward forecast: $\mu_2 \neq p_{2,i}, \forall i$. Consequently, Equation (5.14) does not reduce to a simpler expression.

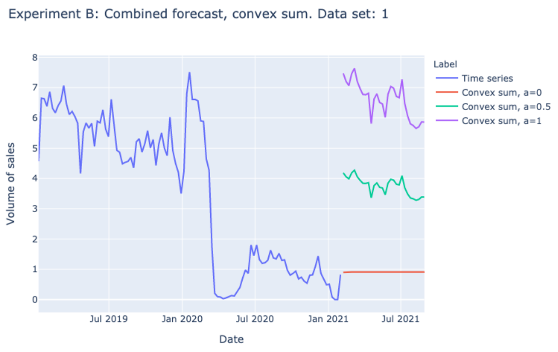

Figure 5.25: Experiment B: Combined model using a convex combination of forecasts, data set 1.

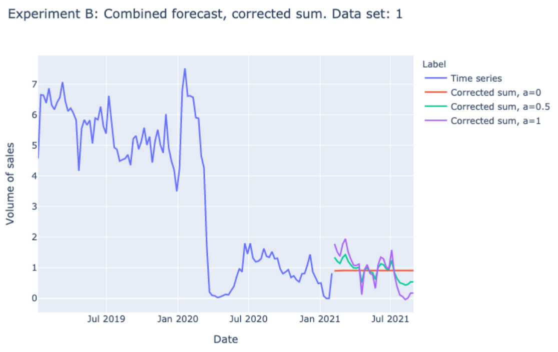

Figure 5.26: Experiment B: Combined model using a corrected combination of forecasts, data set 1.

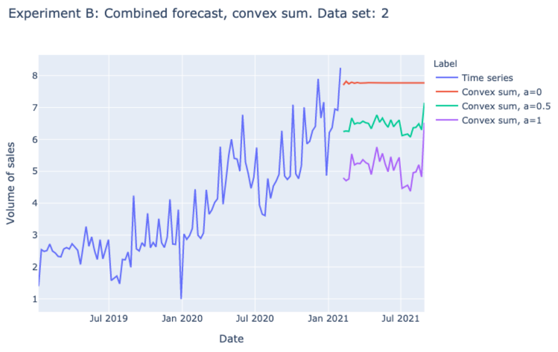

Figure 5.27: Experiment B: Combined model using a convex combination of forecasts, data set 2.

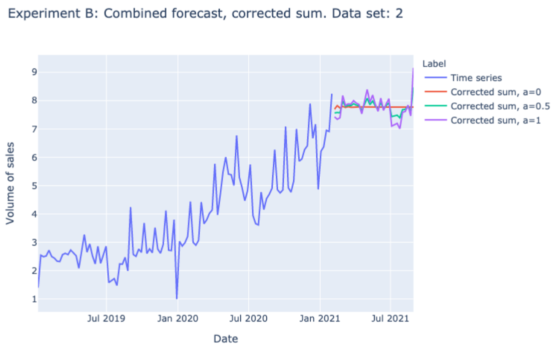

Figure 5.28: Experiment B: Combined model using a corrected combination of forecasts, data set 2.

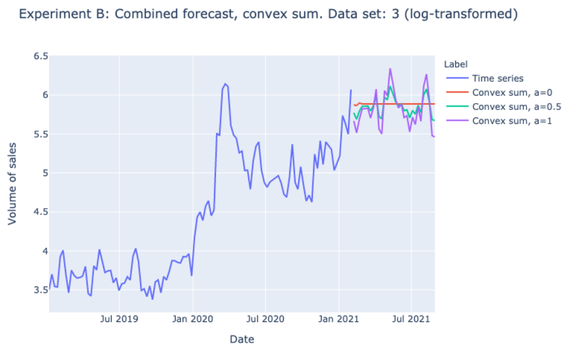

Figure 5.29: Experiment B: Combined model using a convex combination of forecasts, data set 3.

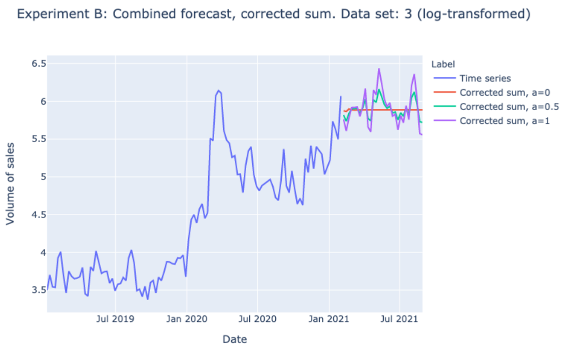

Figure 5.30: Experiment B: Combined model using a corrected combination of forecasts, data set 3.

Here, $\mu_{post} = \frac{1}{n-l+1} \sum_{i=l}^{n} y_i$ is the mean of the series post the abnormal time period. Now, to define deviation of a series in prior to the abnormal time period, an equally-sized computing window of $(n-l+1)$ observations is used. The standard deviation of a time series in prior to the abnormal time period $\{(x_i, y_i)\}_{i=1}^{j}$, is defined as

$$\sigma_{pre} = \sqrt{\frac{1}{n-l+1} \sum_{i=j-(n-l)}^{j} (y_i - \mu_{pre})^2}. \tag{5.17}$$

Here, $\mu_{pre} = \frac{1}{n-l+1} \sum_{i=j-(n-l)}^{j} y_i$ is the mean of the $(n-l+1)$ observations in prior to the abnormal time period.

Further, for choosing a, the two standard deviations are used as follows. Based on the assumption that a series transition to a (new) normal state post the abnormal time period, the series' current extent of variability is assumed to continue in future estimates. For this purpose, a was chosen as a function of the standard deviation of the series post the abnormal time period. That is,

$$a = a(\sigma_{post}). \tag{5.18}$$

Further, the historical forecast reflects the standard deviation of the series in prior to the abnormal time period. Consequently, a was chosen to also take

σ_{pre} into account. That is,

$$a = a(\sigma_{post}, \sigma_{pre}). \tag{5.19}$$

In particular, as a resembles a measure of the degree of extent to weight the historical behavior, a was chosen to reflect the eligible deviation of the series in relation to the previous deviation. That is,

$$a = \frac{\sigma_{post}}{\sigma_{pre}}. \tag{5.20}$$

To this end, the corrected combination of forecasts in Equation (5.14) was used with a as in Equation (5.20).

5.4.2 Approximate prediction intervals

For a combined model of forecasts, prediction intervals are complicated to compute. As stated in Subsection (3.3.3), if assuming Gaussian forecast errors, a 95% prediction interval for an h-step-ahead forecast is

$$\hat{y}_{t+h} \pm 1.96\hat{\sigma}_h. \tag{5.21}$$

Here, $\hat{y}_{t+h}$ is an estimated prediction of y_{t+h} and $\hat{\sigma}_h$ an estimate of the standard deviation of the h-step-ahead forecast distribution. Hyndman and Athanasopoulos (2018) state, for $h = 1$, that the residual standard deviation can provide a good measure of the forecast standard deviation. Though, that more complicated computing methods are required for $h > 1$.

In contrary to Hyndman and Athanasopoulos (ibid.), a simplifying approximation is made. Namely, that the residual standard deviation can provide a measure of the forecast standard deviation $\forall h \geq 1$. Consequently, naively approximating that $\hat{\sigma}_h$ does not increase with h. Now, approximate prediction intervals were computed as follows.

To begin with, in order to use Equation (5.21), the forecast errors of the combined model should be Gaussian. For the corrected combination of forecasts, the forecast errors arise from the first and second model*. Residual diagnostics of the two models indicated Gaussian residuals. By considering model residuals as estimates of forecast errors, the forecast errors of the models were assumed Gaussian. Further, as a sum of two Gaussian distributions is Gaussian, the corrected combination of forecasts was assumed to have

* That is, from the SARIMA$(p, d, q)(P, D, Q)s$- and ARIMA(p, d, q) model, respectively.

Gaussian residuals and thus Gaussian forecast errors. Consequently, Equation (5.21) could be used for computing approximate prediction intervals.

Further, as the standard deviation was approximated as $\hat{\sigma}_h \approx \hat{\sigma}, \forall h \geq 1$, an estimate of the standard deviation was computed as follows. First, assuming that the variance of the combined model occur from the first- and second model. Second, assuming that the first- and second model are independent, as they are fitted to different observations. Then, an estimate was obtained as

$$\hat{\sigma}^2 = a^2\hat{\sigma}^2_{mod_1} + (1-a)^2\hat{\sigma}^2_{mod_2}. \tag{5.22}$$

Here, $\hat{\sigma}_{mod_m}, m = 1, 2$, is the residual standard deviation of model m. To this end, the approximate 95% prediction interval was computed as

$$\hat{y}_{corrected,i} \pm 1.96\sqrt{a^2\hat{\sigma}^2_{mod_1} + (1-a)^2\hat{\sigma}^2_{mod_2}}, \quad i \in [n+1, n+h]. \tag{5.23}$$

Chapter 6

Results

This chapter presents the results obtained from the methods in Chapter (5). Sections (6.1-6.3) present the resulting long-term forecasts. Section (6.4) presents the evaluation metrics of the resulting forecasts. Section (6.5) presents the statistical conclusions of the HLN tests.

6.1 Resulting forecasts, benchmark models

The resulting long-term forecasts obtained by the benchmark models are illustrated in Figures (6.1, 6.5, 6.9).

6.2 Resulting forecasts, Experiment A

The resulting long-term forecasts obtained by Experiment A are presented here. Forecasts obtained by the linear impute are illustrated in Figures (6.2, 6.6, 6.10). Forecasts obtained by the combined impute are illustrated in Figures (6.3, 6.7, 6.11).

6.3 Resulting forecasts, Experiment B

The resulting long-term forecasts obtained by Experiment B are illustrated in Figures (6.4, 6.8, 6.12).

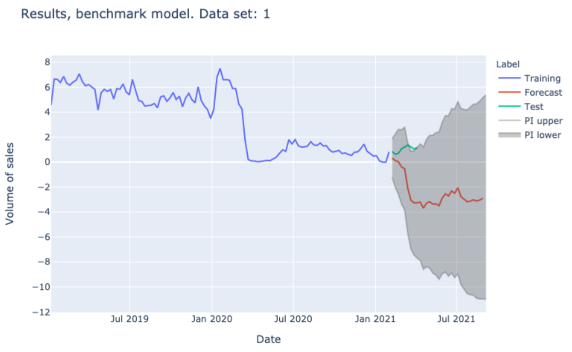

Figure 6.1: Resulting forecast of benchmark model, data set 1. Grey area: 95% prediction interval.

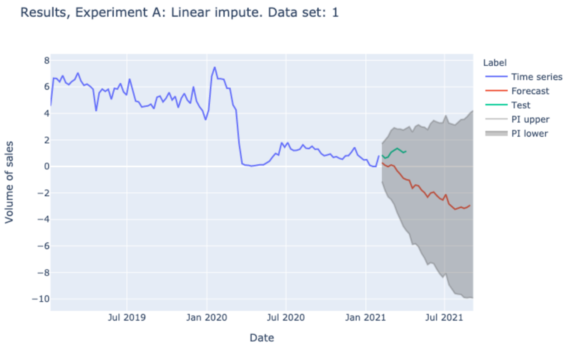

Figure 6.2: Resulting forecast of Experiment A: Linear impute, data set 1. Grey area: 95% prediction interval.

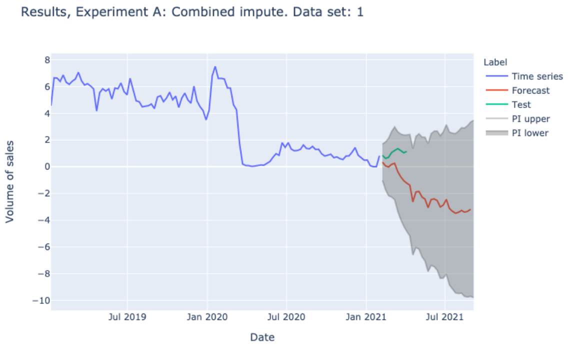

Figure 6.3: Resulting forecast of Experiment A: Combined impute, data set 1. Grey area: 95% prediction interval.

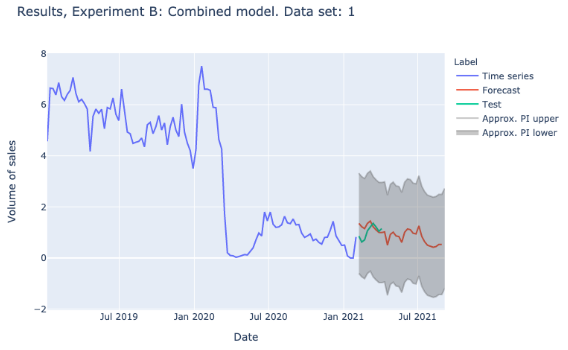

Figure 6.4: Resulting forecast of Experiment B, data set 1. Scaling parameter $a \approx 0.52$. Grey area: Approx. 95% prediction interval.

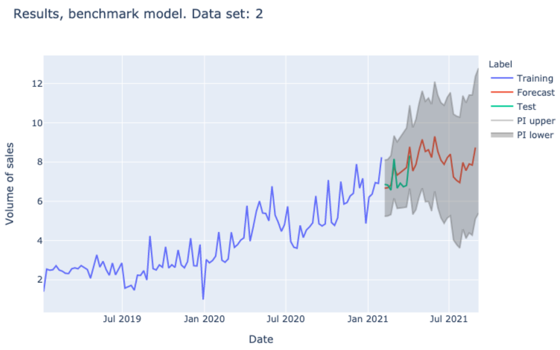

Figure 6.5: Resulting forecast of benchmark model, data set 2. Grey area: 95% prediction interval.

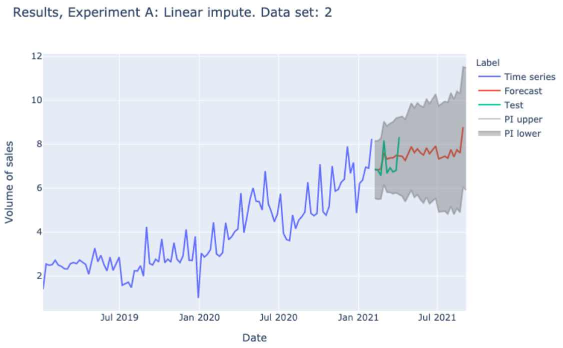

Figure 6.6: Resulting forecast of Experiment A: Linear impute, data set 2. Grey area: 95% prediction interval.

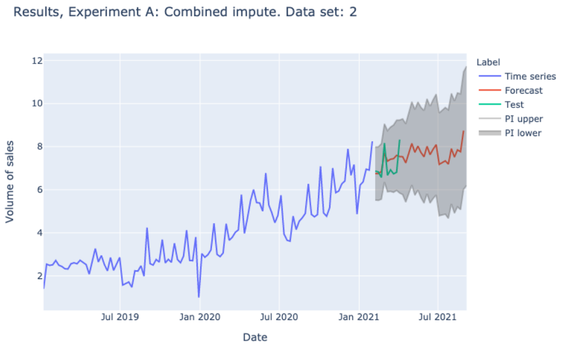

Figure 6.7: Resulting forecast of Experiment A: Combined impute, data set 2. Grey area: 95% prediction interval.

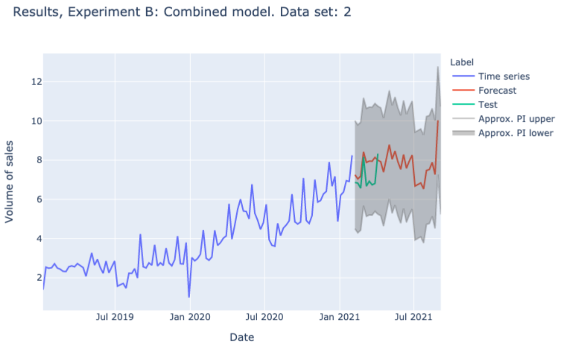

Figure 6.8: Resulting forecast of Experiment B, data set 2. Scaling parameter $a \approx 1.63$. Grey area: Approx. 95% prediction interval.

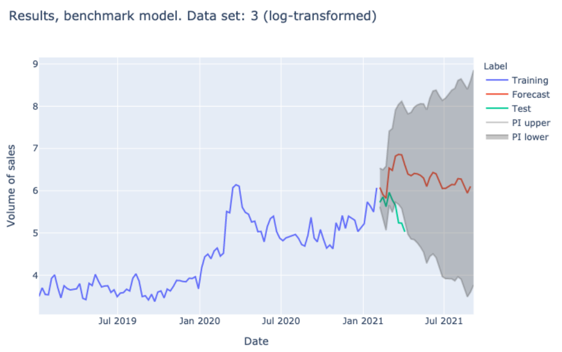

Figure 6.9: Resulting forecast of benchmark model, data set 3. Grey area: 95% prediction interval.

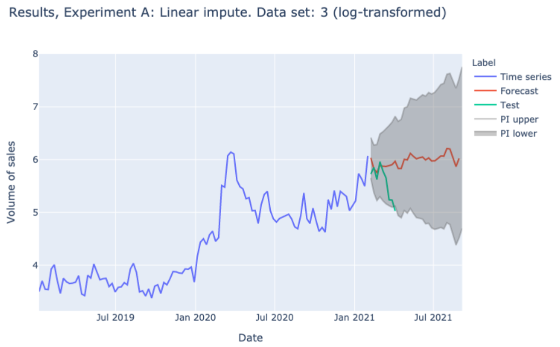

Figure 6.10: Resulting forecast of Experiment A: Linear impute, data set 3. Grey area: 95% prediction interval.

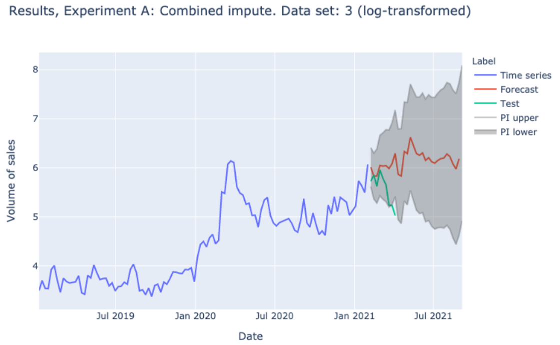

Figure 6.11: Resulting forecast of Experiment A: Combined impute, data set 3. Grey area: 95% prediction interval.

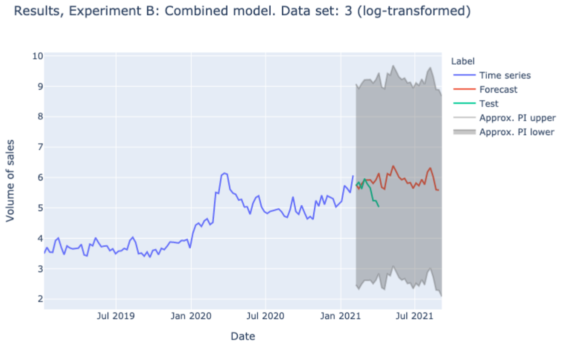

Figure 6.12: Resulting forecast of Experiment B, data set 3. Scaling parameter $a \approx 0.92$. Grey area: Approx. 95% prediction interval.

6.4 Evaluation metrics

This section presents the two evaluation metrics of the resulting forecasts. The metrics were computed in accordance to Subsection (3.3.4).

The mean absolute percentage error, MAPE, of each resulting forecast is presented in Table (6.1). The mean bias error, MBE, of each resulting forecast is presented in Table (6.2). In Tables (6.1-6.2), the metric of best performance for each series is highlighted in bold.

	Benchmark	**Exp. A: Linear**	**Exp. A: Combined**	**Exp. B**
Data set 1	214%	121%	123%	**33.4%**
Data set 2	6.97%	**6.46%**	6.78%	10.6%
Data set 3	16.2%	**6.51%**	8.62%	6.85%

Table 6.1: Primary error metric MAPE of each resulting forecast. Here, exp. is the abbreviation of experiment.

	Benchmark	**Exp. A: Linear**	**Exp. A: Combined**	**Exp. B**
Data set 1	2.37	1.28	1.31	**-0.18**
Data set 2	-0.32	**-0.14**	-0.16	-0.66
Data set 3	-0.87	-0.32	-0.45	**-0.30**

Table 6.2: Secondary error metric MBE of each resulting forecast. Here, exp. is the abbreviation of experiment.

6.5 Statistical conclusions of HLN tests

This section presents the statistical conclusions of the HLN tests. In accordance to Subsection (3.3.5), the HLN test was used for inferring significant difference between two forecasts' predictive accuracy. For each series, the predictive accuracy of the MAPE superior forecast and the benchmark forecast was compared at a significance level $\alpha = 0.05$. Here, the loss function of the forecast errors was chosen as the absolute error loss $g(e_{it}) = |e_{it}|$.

The HLN statistic, p-value and inferred conclusion of each superior forecast are presented in Table (6.3).

	HLN	p-**value**	**Inferred conclusion,** $\alpha = 0.05$
Data set 1	3.49	<0.01	Reject H_0: Significant difference
Data set 2	0.37	0.72	Accept H_0: No significant difference
Data set 3	4.20	<0.01	Reject H_0: Significant difference

Table 6.3: Inferred conclusions of the HLN test. For each series, the predictive accuracy of the MAPE superior forecast and the benchmark forecast was compared at a significance level $\alpha = 0.05$.

Chapter 7

Discussion

This chapter presents discussions on the methods and results of this thesis. Section (7.1) presents results analysis. Section (7.2) presents methods analysis.

7.1 Results analysis

This section presents an analysis of the results. Subsection (7.1.1) presents a summary of the results. Subsection (7.1.2) presents limitations of the results. Subsection (7.1.3) presents interpretations of the results. Subsection (7.1.4) presents recommendations on the results.

7.1.1 Summary

First, the results of the mean absolute percentage error presented in Table (6.1) are summarised along the results of the HLN tests in Table (6.3). Second, the results of the mean bias error presented in Table (6.2) are commented on.

As seen for the first data set in Table (6.1), the superior MAPE of 33.4% was obtained by the combined model of forecasts and the inferior MAPE of 214% by the benchmark model. The linear impute resulted in a MAPE of 121% and the combined impute in a MAPE of 123%. As seen for the first data set in Table (6.3), a significant difference in the predictive accuracy between the superior- and benchmark forecast was inferred at the significance level $\alpha = 0.05$.

As seen for the second data set in Table (6.1), the superior MAPE of 6.46% was obtained by the linear impute and the inferior MAPE of 10.6% by the combined model. The combined impute resulted in a MAPE of 6.78% and the benchmark model in a MAPE of 6.97%. As seen for the second data set in Table (6.3), no significant difference in the predictive accuracy between the superior- and benchmark forecast was inferred at the significance level $\alpha = 0.05$.

As seen for the third data set in Table (6.1), the superior MAPE of 6.51% was obtained by the linear impute and the inferior MAPE of 16.2% by the benchmark model. The combined model resulted in a MAPE of 6.85% and the combined impute in a MAPE of 8.62%. As seen for the third data set in Table (6.3), a significant difference in the predictive accuracy between the superior- and benchmark forecast was inferred at the significance level $\alpha = 0.05$.

Now, the results of the mean bias error presented in Table (6.2) are commented on. Here, as seen for the first and second data set, the method ranking in terms of MBE is unchanged to the ranking in terms of MAPE. For the third set, the method ranking of the combined model and the linear impute is of reversed order. However, as the difference of the error metrics is inconsiderably small, the reversed order is disregarded and not further commented on. Last, for each data set, the superior MBE is less than zero. That is, the superior forecasts tend to overshoot the actual observations.

7.1.2 Limitations

It is of importance to recognise the limitations of the result summary in Subsection (7.1.1). Neither the mean absolute percentage error nor the mean bias error say much about the forecasts' ability to mimic seasonal behavior. Rather, the forecasts' average deviation from the actual observations. Consequently, a superior MAPE- or MBE metric could be obtained arbitrarily by a misleading coincidence, e.g. by a naive forecast* although a naive forecast would fail miserably in mimicking seasonal behavior. Thus, upon stating interpretations of the results, it is of importance to visually analyse the forecasts along the result summary.

7.1.3 Interpretations

In prior to stating interpretations of the results for each time series, general interpretations are stated.

To begin with, the summary of the results presented in Subsection (7.1.1) highlights the importance of taking different approaches for different abnormal behaviors. Since the abnormal behavior of a series is classified by its distinct impact on the series, it could be argued as unsurprising that no method seems

* That is, a forecasted constant line: $y_{t+i|t} = C \in \mathbb{R}, \forall i = 1, \ldots, h$.

superior in all cases.

Further, it can be seen that forecasts obtained from Experiment A and Experiment B resemble similar seasonal behavior; however, of different magnitude and trend. This is clear from comparing Figures (6.2-6.4), Figures (6.6-6.8) or Figures (6.10-6.12). The similarity of the forecasts' seasonal behavior can be argued as comprehensible: forecasts of Experiment A replicate seasonal behavior in the series and forecasts of Experiment B is formed on historical (seasonal) pattern in the series. Consequently, the forecasts of Experiment A and Experiment B will resemble similar seasonal pattern.

The first time series, a level shift

As highlighted in Subsection (7.1.2), it is of importance to visually analyse the forecasts along the result summary. As seen for the first data set in Figure (6.4), the MAPE superior forecast obtained by the combined model seems to mimic the actual observations' behavior. Consequently, it could be argued that the superior metric for the combined model is not a misleading coincidence. In addition, by the statistical conclusions of the HLN test, it could be argued that the combined model statistically is superior to the benchmark model at a significance level $\alpha = 0.05$. By this, it is concluded that the combined model is superior to use for the first time series.

It is not surprising that the combined model is deemed to be superior for the first data set. As the ARIMA family aims to describe the autocorrelations in a series, it tends to replicate past events in future forecasts. For example, this can be seen from comparing Figures (6.1-6.3) and Figures (5.13-5.14): when using a SARIMA model, the forecast mimics past observations' behavior seen in the same time period of the previous year. In other words, the forecast itself mimics the behavior of the abnormal time period, regardless if the abnormal time period was imputed or not. Consequently, in the case of a level shift as in the first series, the shift will be replicated, regardless if the abnormal time period was imputed or not as the impact of a level shift is not limited within the endpoints of the abnormal time period. By this, it could be argued that the conclusions of superiority of the combined model could be generally appropriate for series exhibiting a level shift. However, large-scale conclusions like that require empirical evaluations. For this purpose, some explicit recommendations are stated in Subsection (7.1.4).

The second time series, a transient change

For the second time series, the statistical conclusions of the HLN test pointed at no significant difference in the predictive accuracy of the MAPE superior forecast and the benchmark forecast at the level $\alpha = 0.05$. Consequently, it could be argued that there is no statistical gain of performance of using a model other than the benchmark[*]. In addition, as seen in Figure (6.6), the MAPE superior forecast obtained by the linear impute does not seem to mimic the actual observations' behavior. From this, it could be argued that the superior metric for the linear impute is a misleading coincidence. Consequently, it is concluded that the linear impute not is superior to use for the second time series. Now, the benchmark model is further analysed. As seen in Figure (6.5), the benchmark forecast seems to mimic the actual observations' behavior better than the linear impute. In addition to the statistical conclusions of the HLN test and the relative low cost-complexity of the model, it could be argued that the benchmark model is more feasible than the linear impute. Consequently, it is concluded that the benchmark model is the most feasible to use for the second time series.

Although the benchmark model is concluded as the most feasible for the second data set, it could be feasible to use one other models for series exhibiting a transient change. As stated, the similarity of the forecasts' seasonal behavior in Experiment A and Experiment B is comprehensible. However, it complicates suggesting which, if any, of the experimental models that could be feasible to use for series exhibiting a transient change. Contrary to the level shift, neither the transient change nor the additive outlier transition abruptly into a new normal level. Rather, the abnormal behavior is limited within the endpoints of the abnormal time period. In other words, the global trend-change of the series pre- to post the abnormal time period is not as distinct. Consequently, at least when considering the ARIMA family of models, it could be argued that the trend-change impact of a transient change (or an additive outlier) is not as distinct in a forecast as of a level shift. Further, it could be argued that the minimal trend-change of a transient change leads to a minimal difference of the levels of the resulting forecasts. That is, regardless if the forecast is obtained by Experiment A or Experiment B, the level of the resulting forecast is similar. This further complicates suggesting which of the experimental models that could be feasible to use for other series exhibiting a transient change.

If the resulting forecasts of the second time series resembled a more

[*] That is, at least for the specific test set used in evaluation.

accurate seasonal behavior, it could be argued to suggest a combined model superior to the linear impute, as the forecast of the combined model seems to mimic the magnitude of the actual observations more accurately than the forecast of the linear impute. Although, this is not necessarily the case for all series exhibiting a transient change. An empirical evaluation of the linear impute versus the combined model is recommended for drawing general conclusions for series exhibiting a transient change.

The third time series, an additive outlier

As seen for the third data set in Figure (6.10), the MAPE superior forecast obtained by the linear impute does not seem to mimic the actual observations' behavior. Still, in Figures (6.9-6.12) it can be seen that none of the resulting forecasts mimic the series accurately as the test set expresses an unforeseen abrupt shift. This complicates stating conclusions since it is unclear whether or not the superior metric for the linear impute is a misleading coincidence. By the statistical conclusions of the HLN test, it could be argued that the linear impute is more suitable than the benchmark model. Thus, even though the forecast is not fully convincing, it is concluded that the linear impute is the most suitable to use for the third time series.

As for the transient change, it is not straightforward to suggest particular methods for series exhibiting an additive outlier. As the additive outlier yields a minimal trend-change of a series pre- to post the abnormal time period, there is a minimal difference of the levels of the resulting forecasts, regardless of forecasting approach. Even though the linear impute was concluded suitable for the third time series, it is important to recognise the ambiguousness of the superior metric. It could be argued that the most appropriate suggestion would be to perform empirical evaluations of all three experimental models for series exhibiting an additive outlier.

7.1.4 Recommendations

In order to increase the validity of the interpretations of the results presented in Subsection (7.1.3), two recommendations are stated. These recommendations are crucial for the purpose of generalizing to large-scale conclusions.

First, the sample size of the test set should be increased. That is, the models should be re-evaluated at a later time point. Specifically, this is necessary for the second time series, as the major impact of the transient change to the benchmark forecast first is present in June 2021.

Second, empirical evaluations should be carried out. The interpretations should not be seen as applicable to all time series exhibiting a certain abnormal behavior; rather, as suggestions pointing in potential directions. In order to generalize to large-scale conclusions for each abnormal behavior, empirical evaluations of the superior- and suggested methods of each series should be performed. As the interpretations are formed on a perspective of the ARIMA family, conclusions will not necessarily apply to other families of forecasting models.

7.2 Methods analysis

This section presents an analysis of the methods underlying Experiment A and Experiment B. Subsection (7.2.1) presents an analysis of the reliability of the methods. Subsection (7.2.2) presents an analysis of the validity of the methods. Subsection (7.2.3) presents an analysis of the quality of the methods.

7.2.1 Reliability

In this subsection, an analysis of the reliability of the methods is presented. First, a paragraph on the assumptions made in Experiment A and Experiment B is presented. Second, a discussion on biased models and outputs is presented. Third, a paragraph on the choice of a static scaling parameter in Experiment B is presented.

Assumptions of unquantifiable uncertainty

Several assumptions of unquantifiable uncertainty were made in Experiment A and Experiment B. First, the assumption that the abnormal time period related to the COVID-19 outbreak is reasonably detected. In other words, that a (new) normal state is transitioned to after the abnormal time period. Second, the assumption that the abnormal behavior related to the outbreak will not recur in the subsequent year(s). Third, in Experiment B, the assumption that the pre-pandemic seasonal behavior resembles the seasonal behavior in the (new) normal state. Fourth, in Experiment B, the scaling parameter a is formed on the assumption that forecasts will reflect the degree of oscillations in the (new) normal state.

It could be argued that the necessity of the assumptions is clear. As we lack prior knowledge of the new phenomena of the pandemic, we do not know how the time period will behave nor evolve. In addition, as the benchmark models tend to produce unreliable forecasts, it is argued as necessary to state assumptions that enable investigating possible solutions. For example, the small sample size post the abnormal time period disables inferring a seasonal pattern in the (new) normal state. Here, one could either assume a seasonal pattern or ignore any seasonal behavior. However, for the latter, a naive forecast could just as easily be used.

The reliability of the assumptions is either too complex or currently impossible to evaluate. However, it could be argued that the degree of reliability is dependent on data specific properties, such as the industry- and operating country of the merchant. Consequently, data anonymization complicates commenting on the degree of reliability for each data set. Regardless of the degree of reliability, it is of importance to understand that the uncertainty of the assumptions are currently unquantifiable.

Biased models and outputs

It is clear that the models of Experiment A and Experiment B are biased. As the models are not designed on an independent theoretical framework, but rather on assumptions of what should or should not happen, model outputs are not unbiased estimates. As we lack prior knowledge of the new phenomena of COVID-19, it could be argued to favour a biased model to an unbiased model.

Further, it could be argued that the most appropriate solution from a mathematical perspective is not necessarily the most appropriate solution from a domain perspective. For example, the benchmark models could be suggested as the most appropriate solutions from a mathematical perspective. However, from a domain perspective, the unbiased benchmark models are not necessarily the most appropriate, as they tend to produce unreliable forecasts.

In addition, as the sought method for tackling unreliable forecasts should be used as a complement to existing building blocks of forecasting models, it could be argued that the models of the experiments will likely be biased. Due to this and the highly uncertain phenomena currently being, it is suggested not to take the resulting long-term forecasts as reliable long-term solutions. Rather, to take the forecasts as temporary solutions, requiring more frequent re-running of models and repeated updating of predictions.

Scaling parameter in Experiment B

It can be questioned why the scaling parameter of Experiment B was chosen as a static parameter rather than a time dependent parameter.

First, in the case of the convex combination of two forecasting models, the scaling parameter is limited to $a \in [0, 1]$. An idea could be to use a time dependent parameter, $a = a(t)$, to represent a transition between the two models. In this case, the time dependency would occur during the abnormal time period. For example, a could transit from 1 to 0, representing a transition between the first and second model. However, neither the post-abnormal model (nor the pre-abnormal model) used in the combined forecast resembles a reliable forecast for the subsequent year. Thus, as neither of the models is a good representation of the actual state transited to, a time dependency as such would not yield better results.

Second, in the case of the corrected combination of two forecasting models, the scaling parameter a is not limited to $[0, 1]$*. Again, an idea could be to use a time dependent parameter, $a = a(t)$. Although, the meaning of the time dependency is not as clear as for a convex combination. Here, it could be argued that the time dependency can not be chosen as intuitively for the corrected combination of forecasting models. Thus, a static scaling parameter was chosen. A time dependent parameter could be appropriate to use, and would require further investigations. This is suggested as future improvements in Subsection (8.2.1).

7.2.2 Validity

In this subsection, an analysis of the validity of the methods is presented. For the purpose of evaluating the validity of the models in Experiment A and Experiment B, residual diagnostics of the forecast errors of each superior model were checked. Residual diagnostics is explained in Subsection (3.2.5), forecast errors are defined in accordance to Equation (3.35).

Figures (7.1-7.3) illustrate residual diagnostics of the forecast errors of each MAPE superior forecast. For all three series, the results of the residual diagnostics happen to coincide, as follows. First, the standardized residuals seem to be randomly scattered around zero without obvious outliers. As white noise behavior of the standardized residuals is desired, it could be argued that this result indicate model validity. Second, it does not seem to be any significant residual correlation of the forecast errors. As any residual correlation is undesired and should be modeled, it could be argued that this result indicate model validity. Third, for the histogram of standarized residuals, the estimated kernel density does not seem to resemble a standard Gaussian distribution. The estimated kernel density seems somewhat Gaussian

* Here, as the corrected combination of forecasting models not is a *convex* combination of models, a is not limited to $a \in [0, 1]$.

looking, which may indicate potential of improving the model. Fourth, in the normal Q-Q plot, the (non-transformed) residuals do not seem to follow the normal line accurately, which may indicate potential of improving the model.

By the residual diagnostics, it could be argued that the MAPE superior model of each series is valid. However, that each model could be improved with the purpose of increasing model validity.

7.2.3 Quality

In this subsection, an analysis of the quality of the methods is presented. For this purpose, the desired quality attributes of the sought method stated in Subsection (1.3.1) are discussed.

First, it could be argued that the models in Experiment A and Experiment B are particularly simple for the problem complexity. As the models are comprehensible and of low cost-complexity, it could be argued that the models fulfill the quality attribute of simplicity.

Second, it could be argued that the models in Experiment A and Experiment B are scalable. As the methods are automated, of low cost-complexity and do not seem overfitted to a particular time series, it could be argued as appropriate to scale the methods to larger data collections of several series. Consequently, it could be argued that the models fulfill the quality attribute of scaleability.

Third, it could be argued that the models in Experiment A and Experiment B are flexible. The models seem applicable to time series exhibiting particular abnormal behavior, regardless of the content of the series. Thus, it could be argued that the models may be adaptable to external changes such as changing the context of the problem space. Consequently, it could be argued that the models seem to fulfill the quality attribute of flexibility.

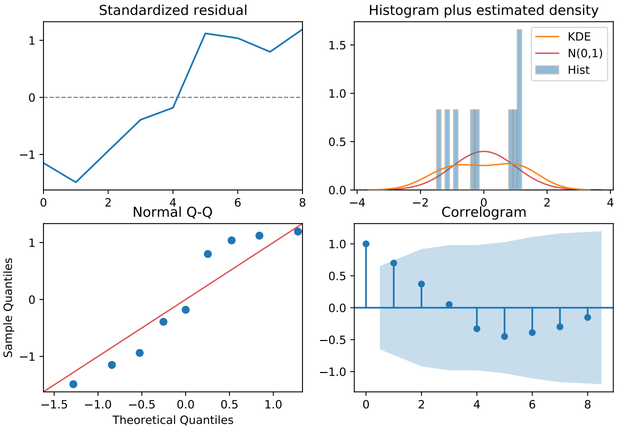

Figure 7.1: Residual diagnostics of forecast errors of MAPE superior forecast, data set 1.

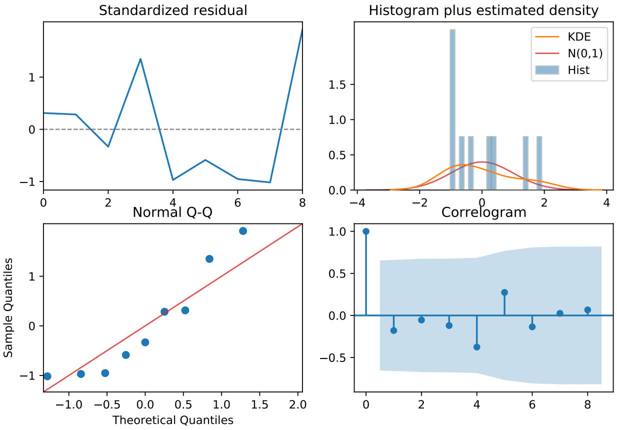

Figure 7.2: Residual diagnostics of forecast errors of MAPE superior forecast, data set 2.

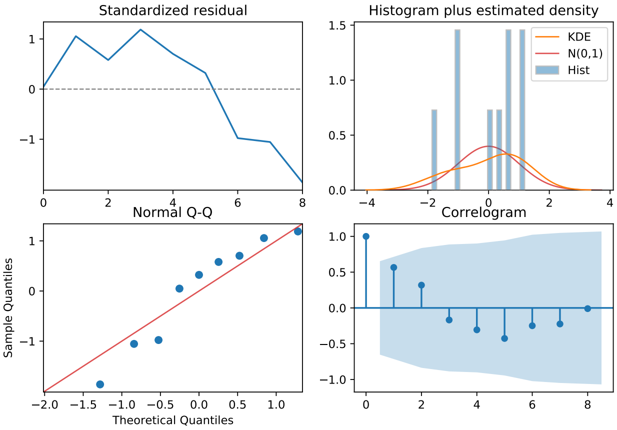

Figure 7.3: Residual diagnostics of forecast errors of MAPE superior forecast, data set 3.

Chapter 8

Conclusions and Future work

This chapter presents the conclusions of this thesis and suggestions of future work. Section (8.1) presents the major conclusions. Section (8.2) presents suggestions of future work.

8.1 Conclusions

In this thesis, an investigation was conducted on how to forecast volume of sales, to a greater extent of reliability and accuracy, during the abnormal time period of the COVID-19 pandemic. In particular, the problem of unreliable forecasts replicating certain abnormal events was investigated for three extreme cases of time series outliers. The investigation centered around two experiments, Experiment A and Experiment B, and culminated in three proposed methods: the linear impute, the combined impute and the combined model of forecasts. The methods and results of Experiment A and Experiment B were discussed.

The major conclusions of the analysis of the results highlighted the importance of taking different approaches for different abnormal behaviors. First, for the time series exhibiting a level shift, the combined model of forecasts was concluded as superior to the benchmark- and the other proposed models. The superiority of the combined model was argued as comprehensible and unsurprising. A discussion on the possibility of generalizing the conclusion of superiority to a large-scale conclusion for series exhibiting a level shift was presented. The importance of empirical evaluations for large-scale conclusions as such was highlighted. Second, for the time series exhibiting a transient change, the benchmark model was concluded as the most feasible, even though the superior error metric was obtained by the linear impute. In addition, a discussion on the potential use for the other proposed models for series exhibiting a transient change was presented. Here, suggestions of empirical evaluations of the linear impute versus the combined model were stated and it was even argued to suggest a combined model superior to the linear impute. Third, for the time series exhibiting an additive outlier, the linear impute was concluded as more suitable than the benchmark- and the other proposed models. A discussion on the potential use for the other proposed models for series exhibiting an additive outlier was presented. Here, due to ambiguousness of the superior error metric for the linear impute, suggestions of empirical evaluations of all three proposed models were stated.

In the analysis of the methods underlying Experiment A and Experiment B, the reliability, validity and quality of the models were discussed. Here, the problem complexity of handling the extreme cases of unreliable forecasts was prominent. The major conclusions of the analysis of the methods were that the proposed models fulfilled the desired attributes of simplicity, scaleability and flexibility. Further, residual diagnostics of the forecast errors of the MAPE superior forecast for each series indicated model validity. However, potential improvements of the models can be made for the purpose of increasing model validity. In regards to the reliability of the models and the highly uncertain phenomena of the COVID-19 pandemic, it was concluded to suggest not to take the resulting long-term forecasts as reliable long-term solutions. Rather, to take the forecasts as temporary solutions, requiring more frequent re-running of models and updating of predictions regardless of forecasting approach.

8.2 Future work

This section presents suggestions of future work. In addition, it is suggested to perform empirical evaluations of the proposed models for the purpose of stating large-scale conclusions for particular time series outliers. Subsection (8.2.1) presents suggestions on improvements within the study. Subsection (8.2.2) presents suggestions on future studies.

8.2.1 Improvements within study

Here, two suggestions of potential improvements within the study are stated.

Non-static scaling parameter in Experiment B

The first suggestion is to introduce a non-static scaling parameter in the combined model of forecasts.

As stated in Subsection (7.2.1), an idea is to use a time dependent scaling parameter $a = a(t)$ rather than a static scaling parameter $a \neq a(t)$. Here, the time dependency could be defined by data specific properties or by abnormal behavior specific properties. Ideally, the procedure of computing $a = a(t)$ should be automated yet cost-effective. In particular, the automation and cost-effectiveness is of importance if the method should remain scaleable to larger data collections.

Approximate prediction intervals in Experiment B

The second suggestion is to improve the computation of approximate prediction intervals of the combined model of forecasts.

The naive computation of approximate prediction intervals outlined in Subsection (5.4.2) could be made more rigorous. For example, by an explicit computation of the covariance of the two models, rather than assuming independence of the models. Or, by including an h-dependence in the computation of the uncertainty as the uncertainty should increase with the forecast horizon h.

8.2.2 Future studies

Here, two suggestions of future studies on forecasting during abnormal time periods are stated.

Feature engineering of exogenous variables

The first suggestion is to study the use of feature engineering of exogenous variables. In particular, to study it as a complement to existing building blocks of forecasting models.

As seen in related work stated in Subsection (2.3), it could be successful to introduce a new feature to the time series for the purpose of tackling complicated forecasting procedure arising from abnormal time periods. In the setting of this thesis, a suggestion would be to include exogenous variables such as dates of introducing and lifting social distancing restrictions, features representing population movement patterns or features representing the time period of the first-, second-, third-, et.c.-, pandemic waves.

Volatility modeling

The second suggestion is to study volatility modeling.

As seen in related work stated in Subsection (2.3), it could be successful to add a volatility variable to the benchmark forecasting models for the purpose of tackling abnormal time period forecasting. The volatility modeling could be performed by incorporating the idea of a hybrid ARIMA-GARCH model.

In this thesis, an initial implementation of a SARIMA-GARCH model was performed. The hybrid model was excluded in this thesis due to time limitations to further improving the implementation. The model seemed feasible for some time series, at least for some particular high parameter choices of the GARCH model. Consequently, it is suggested to further study the use of a hybrid SARIMA-GARCH model for the purpose of volatility modeling.

Bibliography

Akouemo, H. N. and R. J. Povinelli (2014). “Time series outlier detection and imputation”. In: *IEEE PES General Meeting | Conference Exposition*, pp. 1–5. ISSN: 1932-5517. DOI: 10.1109/PESGM.2014.6939802.

Bates, J. M. and C. W. J. Granger (1969). “The Combination of Forecasts”. In: *OR* 20 (4), pp. 451–468. DOI: 10.2307/3008764.

Bollerslev, T. (1986). “Generalized autoregressive conditional heteroskedasticity”. In: *Journal of Econometrics* 31 (3), pp. 307–327. ISSN: 0304-4076. DOI: 10.1016/0304-4076(86)90063-1.

Box, G. E. P et al. (2016). *Time Series Analysis: Forecasting and Control*. 5th. New Jersey, USA: John Wiley & Sons, Inc. ISBN: 978-1-118-67502-1.

Brockwell, P. J. and R. A. Davis (2002). *Introduction to Time Series and Forecasting*. 2nd. New York, USA: Springer-Verlag New York, Inc. ISBN: 0-387-95351-5.

Chatfield, C. (2003). *The Analysis of Time Series: An Introduction*. 6th. Rhode Island, USA: Chapman & Hall/CRC. ISBN: 978-1584883173.

Chen, Y., W. Yang, and B. Zhang (2020). *Using Mobility for Electrical Load Forecasting During the COVID-19 Pandemic*. arXiv: 2006.08826v1 [eess.SP].

Diebold, F. X. and R. S. Mariano (1995). “Comparing Predictive Accuracy”. In: *Journal of Business and Economic Statistics* 13 (3), pp. 253–263. DOI: 10.2307/1392185.

Ding, Z. and M. Fei (2013). “An Anomaly Detection Approach Based on Isolation Forest Algorithm for Streaming Data using Sliding Window”. In: *IFAC Proceedings Volumes* 46 (20), pp. 12–17. ISSN: 1474-6670. DOI: 10.3182/20130902-3-CN-3020.00044.

Güngör, B. O., H. M. Ertuğrul, and U. Soytaş (2021). “Impact of Covid-19 outbreak on Turkish gasoline consumption”. In: *Technological Forecasting and Social Change* 166.120637. ISSN: 0040-1625. DOI: 10.1016/j.techfore.2021.120637.

Harvey, D., S. Leybourne, and P. Newbold (1997). “Testing the equality of prediction mean squared errors”. In: *International Journal of Forecasting* 13 (2), pp. 281–291. ISSN: 0169-2070. DOI: 10.1016/S0169-2070(96)00719-4.

Hyndman, R. J. and G. Athanasopoulos (2018). *Forecasting: Principles and Practice*. 2nd. OTexts: Melbourne, Australia. URL: https://otexts.com/fpp2/ (visited on 05/15/2021).

Kieu, T. et al. (2019). “Outlier Detection for Time Series with Recurrent Autoencoder Ensembles”. In: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, Main track*, pp. 2725–2732. DOI: 10.24963/ijcai.2019/378.

Kwiatkowski, D. et al. (1992). “Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root?” In: *Journal of Econometrics* 54 (1–3), pp. 159–178. ISSN: 0304-4076. DOI: 10.1016/0304-4076(92)90104-Y.

Lin, Y. et al. (2021). “Forecasting stock index price using the CEEMDAN-LSTM model”. In: *The North American Journal of Economics and Finance* 57.101421. ISSN: 1062-9408. DOI: 10.1016/j.najef.2021.101421.

Tsay, R. S. (1988). “Outliers, Level Shifts, and Variance Changes in Time Series”. In: *Journal of Forecasting* 7 (1), pp. 1–20. ISSN: 0277-6693. DOI: 10.1002/for.3980070102.

Vishwakarma, G. K., C. Paul, and A. M. Elsawah (2020). “An algorithm for outlier detection in a time series model using backpropagation neural network”. In: *Journal of King Saud University - Science* 32 (8), pp. 3328–3336. ISSN: 1018-3647. DOI: 10.1016/j.jksus.2020.09.018.

Yan, R. et al. (2021). “Multi-hour and multi-site air quality index forecasting in Beijing using CNN, LSTM, CNN-LSTM, and spatiotemporal clustering”. In: *Expert Systems with Applications* 169.114513. ISSN: 0957-4174. DOI: 10.1016/j.eswa.2020.114513.

Yu, Y. et al. (2014). “Time Series Outlier Detection Based on Sliding Window Prediction”. In: *Mathematical Problems in Engineering* 2014.879736. DOI: 10.1155/2014/879736.

Zou, H. and Y. Yang (2004). “Combining time series models for forecasting”. In: *International Journal of Forecasting* 20 (1), pp. 69–84. ISSN: 0169-2070. DOI: 10.1016/S0169-2070(03)00004-9.

TRITA-SCI-GRU 2021:182