h_da
HOCHSCHULE DARMSTADT
UNIVERSITY OF APPLIED SCIENCES

Hochschule Darmstadt

Fachbereiche Mathematik
und Naturwissenschaften
& Informatik

Leading Indicator Search for Time Series Forecasting

Abschlussarbeit zur Erlangung des akademischen Grades

Master of Science (M.Sc.)

vorgelegt von

Lucas Möller

Matrikelnummer: 766755

Referent	:	Prof. Dr. Markus Döhring
Korreferent	:	Prof. Dr. Christoph Becker
Ausgabedatum	:	01.10.2021
Abgabedatum	:	16.05.2022

ERKLÄRUNG

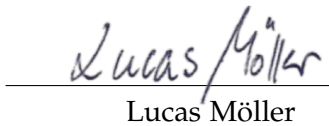
Ich versichere hiermit, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die im Literaturverzeichnis angegebenen Quellen benutzt habe.

Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten oder noch nicht veröffentlichten Quellen entnommen sind, sind als solche kenntlich gemacht.

Die Zeichnungen oder Abbildungen in dieser Arbeit sind von mir selbst erstellt worden oder mit einem entsprechenden Quellennachweis versehen.

Diese Arbeit ist in gleicher oder ähnlicher Form noch bei keiner anderen Prüfungsbehörde eingereicht worden.

Karlsruhe, 16. Mai 2022



Lucas Möller

ABSTRACT

Accurate forecasts of demand play an important role in many businesses and industries. Especially in the retail sector, these forecasts build the basis for planning various supply-chain activities such as stock management or the allocation of scarce resources and personnel. In this context, the underlying data are often time series. What makes time series data special is that successive observations are usually not independent. In order to make predictions of the future course of time series, established time series models therefore attempt to model the inherent structures and patterns based on past observations. Potentials that external data sources may offer in the form of leading indicators are often neglected.

Therefore, this work addresses the research question whether and under which conditions the integration of external data contributes to improving the accuracy of forecasts. Based on a similarity measure, leading indicators from an external online open data source are determined for the time series of a German retail company. Within an experiment, these leading indicators are incorporated individually as external regressors into a linear time series model. The comparison of the forecasting performance between the univariate and bivariate models is intended to provide information on which factors are responsible for the successful inclusion of leading indicators.

The results of the experiment reveal that a certain time series pattern has a statistically significant influence on the outcome of the inclusion. When both the time series of the retail company and the time series of their corresponding leading indicators exhibit this pattern, the integration of the external regressors can be particularly beneficial for the forecast accuracy. Moreover, the similarity of the time series pairs plays a subordinate role as the results additionally show.

Keywords— Forecasting; Leading Indicator; Similarity Matching; Google Trends; ARIMAX model.

ZUSAMMENFASSUNG

Möglichst genaue Nachfrageprognosen spielen in vielen Unternehmen und Branchen eine wichtige Rolle. Insbesondere im Einzelhandel bilden robuste Prognosen von Abverkäufen die Grundlage für die Planung verschiedener Aktivitäten innerhalb einer Lieferkette, wie z. B. die Lagerverwaltung oder die Personal- und Ressourcenzuteilung. Häufig handelt es sich bei den zugrundeliegenden Daten um Zeitreihen. Das Besondere an Zeitreihendaten ist, dass aufeinanderfolgende Beobachtungen in der Regel nicht unabhängig voneinander sind. Etablierte Zeitreihenmodelle versuchen daher die der Zeitreihen inhärenten Strukturen und Muster auf Basis historischer Werte zu modellieren, um Vorhersagen für den zukünftigen Verlauf zu treffen. Mögliche Potenziale, die externe Datenquellen in Form von Frühindikatoren bieten können, werden dabei oftmals vernachlässigt.

Die vorliegende Arbeit beschäftigt sich daher mit der Forschungsfrage, ob und unter welchen Bedingungen die Integration von externen Daten zur Verbesserung der Prognosegenauigkeit beitragen kann. Hierfür werden auf Basis eines Ähnlichkeitsmaßes Frühindikatoren aus einer Online-Datenquelle für die Zeitreihen eines deutsches Einzelhandelsunternehmens bestimmt. Im Rahmen eines Experiments werden diese Frühindikatoren anschließend einzeln als externe Regressoren in ein lineares Zeitreihenmodell aufgenommen. Der Vergleich der Vorhersagegenauigkeit zwischen den univariaten und bivariaten Modellen soll Aufschluss darüber geben, welche Faktoren für die Einbeziehung von Frühindikatoren erfolgskritisch sind.

Die Ergebnisse des Experiments zeigen, dass ein bestimmtes Zeitreihenmuster einen statistisch signifikanten Einfluss auf den Ausgang der Einbeziehung hat. Wenn sowohl die Zeitreihen des Einzelhandelsunternehmens als auch die Zeitreihen ihrer entsprechenden Frühindikatoren dieses Muster aufweisen, kann die Integration der externen Regressoren besonders vorteilhaft für die Prognosegenauigkeit sein. Wie die Ergebnisse ebenfalls veranschaulichen, spielt die Ähnlichkeit der Zeitreihenpaare dabei eine untergeordnete Rolle.

Schlagworte— Forecasting; Leading Indicator; Similarity Matching; Google Trends; ARIMAX model.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ANOVA Analysis of variance

API Application Programming Interface

AR Autoregressive

ARIMA Auto-Regressive Integrated Moving Average

ARIMAX Auto-Regressive Integrated Moving Average Exogenous Variable

ARMA Mixed Autoregressive Moving average

ART Align Rank Transform

ART-C Align Rank Transform Contrasts

CV Cross-validation

FMCG Fast Moving Consumer Goods

i.i.d. independent and identically distributed

MA Moving average

MAE Mean Absolute Error

MSTL Multiple Seasonal-Trend decomposition using Loess

OOS out-of-sample

SARIMA Seasonal Auto-Regressive Integrated Moving Average

SS Sum of Squares

STL Seasonal and Trend decomposition using Loess

Part I

THESIS

1 INTRODUCTION

1.1 MOTIVATION

An accurate and robust forecasting of data plays an important role in many businesses and industries as well as macroeconomic issues. Forecasting the supply and demand within an industrial sector, the demand for goods and labour within a production environment or forecasting the weather are just a few examples. For the retail sector in particular, the accurate prediction of future sales and demand has a sustainable impact, both economically and environmentally. This impact can relate to several levels of the complete supply-chain of a retail store. With improving the forecast accuracy, products receive a higher availability which prevents out-of-stock situations. In addition, an accurate demand planning can provide the ability to minimize waste due to overstocking. Moreover, precisely predicting the future amount of products sold enables an effective stock management and an efficient allocation of scarce resources and personnel (Fredén and Larsson [42], Taylor and Letham [90]).

In this context, the underlying data are often time series. What makes time series unique compared to other data structures is the dimension of *time*, which increases the complexity of data analysis (Tavakoli et al. [89]). Despite its importance, there are serious challenges associated with producing reliable and high accurate forecasts for time series data. An intuitive approach to forecasting time series is to analyze historical data. The focus is often on trends and seasonal patterns which will be extrapolated into the future. It is assumed that the temporal structure of past observations provides information about the future course of the time series (Sagaert et al. [81]).

Here, one important factor, namely information from external data sources, is often neglected. The integration of external data sources can be discussed intensively since it can be associated with various obstacles. The search for publicly accessible data sources is tedious. External data sources are often associated with costs, if providers charge for the use of their data. Then, this data must first be purchased and subsequently preprocessed. In this case, a reliable data quality is not always guaranteed. There is even less guarantee that the use of external data will add any value to the forecast accuracy. Therefore, they have to be evaluated by specialized departments or experts in a complex process in which valuable potentials for modeling may be lost. However, the inclusion of external data sources and the identification of indicators, that have a leading effect on the time series being forecast, can improve the sales forecast accuracy. If these indicators exhibit similarities in

terms of patterns and temporal structures, they may contain leading context information that explain some of the historical variation (Currie and Rowley [30], Stock and Watson [87]). Selecting appropriate leading indicators and their respective lead order is not trivial. For this reason, it is crucial to examine approaches, in which external time series can be merged and processed so that they potentially lead to an improving forecast performance.

1.2 RESEARCH OBJECTIVES

The purpose of this work is to explore an approach to merge time series from two datasets. One of these datasets shall be a freely accessible dataset from an external data source. Based on a similarity measure, leading indicators will then be determined and incorporated as external regressors in the forecasting process of univariate time series. The motivation and purpose lead to the core research question of this work.

Does the integration of external data sources and leading indicators contribute to improving the accuracy of forecasts?

From this central question, further sub-questions can be derived.

- *How can time series be merged?*
- *When is it beneficial to add external regressors and which conditions are critical for success?*
- *Do time series patterns and similarity have a significant influence?*

These research questions are decisive for the content structure and approach within this work. Eventually, they lead to the two fundamental hypotheses.

- *The inclusion of external time series as leading indicators can improve forecast accuracy.*
- *The higher the similarity between two time series, the more beneficial the integration will be.*

The main objective is to examine these hypotheses with scientifically sound methods and to answer the core research question including its sub-questions. Here, it is essential to note that in this work no novel forecasting methods are developed. Furthermore, it is not the goal to achieve the best possible forecasts for the individual datasets. The question of *whether* and *when* the addition of a leading indicator can be useful should rather be answered based on established time series models. The critical factors can then be taken into account in further forecasting tasks.

1.3 METHODOLOGY

The hypotheses of this work are examined within the framework of an experiment. This experiment should include and implement scientific methods. It

must provide an approach that can be used to address the formulated hypotheses. For this reason, the underlying methodology must be well defined. The main features of the approach followed in this work will now be described.

The experiment is based on two datasets. One dataset is provided by a German retail company which is active in the discounter and construction market as well as in the consumer electronics business. It contains retail sales figures, spanning over two years and including roughly 800 thousand different products with a daily sampling rate.[1] This dataset forms the basis of the experiment since predictions and forecast errors are obtained for its time series. The second dataset will be retrieved from *Google Trends* as an online open data source. Google Trends is a tool developed by Google that provides fine-grained data on the popularity of customer queries on certain search terms in the Google search engine (Cebrián and Domenech [23]). The service is freely available for various research activities. The search terms used to create the Google Trends dataset will be collected based on meta-information of the retail sales products. Both datasets include time series data.

After the two datasets have been gathered and preprocessed for the experiment, selected time series models will first be fitted for every univariate retail sales time series. These models include rather simple baseline models up to a stochastic, linear time series model. Then, forecast errors, which can be used for a performance measure and comparative metric for model accuracy, are calculated. Subsequently, one leading indicator from the Google Trends dataset is determined for each individual retail sales time series. For this purpose, the retail sales time series and external time series are merged based on a similarity measure. There are several approaches to define similarity between two time series. In particular, the similarity measure should be capable of representing similar patterns and quantifying the strength and direction of the relationship. The chosen leading indicators are then incorporated individually as external regressors into the linear time series model converting a univariate model to a bivariate model. Forecast errors are calculated analogously. Ultimately, the performances of the models are compared with each other. Especially the performance difference between the univariate and bivariate linear time series models is central for answering the research questions. Finally, the obtained results are validated with three multivariate analysis methods.

Further information on the individual components of the experimental pipeline and detailed descriptions of the datasets are provided in chapter 3.

1 The name of the company cannot be unveiled due to a disclosure agreement.

1.4 RELATED WORK

This section provides a brief overview of previous work related to the research questions formulated in this work. The objective is to understand the current depth in this field of research and to find out if there are some possible gaps in the literature.

Time series are often affected by external factors that can influence the future course. Such factors may include legislative activities, policy changes or environmental regulations (Durka and Pastorekova [37]). For this reason, a great research area and research interest have emerged from the isolated consideration of historical values to the integration of external leading indicators in prediction models. As versatile the resources of external influences are, so diverse are the use cases for leading indicator search. There are multiple cases in which macroeconomic indicators and natural phenomena like weather are examined and integrated into time series models. The search term *leading indicator time series* resulted in approximately 2,320,000 hits at *Google Scholar*, indicating that much research has been done in this area.

For instance, Durka and Pastorekova [37] modeled and predicted the Gross Domestic Product (GDP) per capita in Slovakia while examining the impact of unemployment rate as an external regressor. It was demonstrated that the model including the external regressor was able to explain much of the variance in the target variable GDP. Moreover, the variable unemployment rate contributed to a superior forecast performance. De Felice, Alessandri, and Ruti [34] have shown that the use of weather data leads to a clear improvement of forecasting accuracy for electricity demand in Italy. The inclusion of external variables has also been studied in public health care. Wangdi et al. [93] have shown that certain weather conditions have a leading effect on the number of cases of malaria in endemic areas of Bhutan. It turned out that the mean maximum temperature lagged at one month was a strong positive predictor of increased malaria cases.

There are many comparable examples in academic literature. For the retail sector, it especially important to incorporate variables that may explain customer shopping behavior and thus correlate with the number of products sold. Murray et al. [68] provide empirical evidence to explain in detail the psychological mechanism of how different aspects of weather affect consumer spending. In addition, Bertrand, Brusset, and Fortin [14] have shown that unseasonal weather has a significant impact on predicting retail sales. Siwerz and Dahlén [85], Žliobaitė, Bakker, and Pechenizkiy [101] and Pavlyshenko [76] suggest that calendar events and public holidays such as Christmas and Easter and even specific weekdays can correlate with product sales and improve forecast accuracy. Furthermore, Huang, Fildes, and Soopramanien [48] have demonstrated that incorporating price and promotional data can lead to substantially more accurate forecasts across a range

of product categories.

With the recent advancements in digital technologies and widespread use of social media as a category of online open data sources, an enormous amount of user-generated content is grown (Asur and Huberman [7]). The relevance of such online open data sources has already been verified in many research tasks. Elshendy et al. [40] analyzed in their study the relationship between the West Texas Intermediate daily crude oil price and multiple predictors extracted from *Twitter*, *Google Trends*, *Wikipedia*, and the Global Data on Events, Language, and Tone database (*GDELT*). Their results have shown that the combined analysis of the four media platforms carries valuable information in making financial forecasting. Particularly Google Trends has proven to be a convenient open data source. In 2012, Choi, and Varian [25] presented in their report short-term forecasts of multiple economic indicators (including unemployment rates, automobile demand and vacation destinations). It turned out that the inclusion of Google Trends in the forecasting process could improve model outcomes by 5% to 20%. The examples revealed a positive association of the volume of search queries with the financial and economic indicators. Here, the areas of application are again very heterogeneous and researchers have shown that Google Trends data can be successfully used to predict social and economic trends (Boone, Ganeshan, and Hicks [16]).

However, despite the increasing use of Google Trends, comparatively little amount of research has been made to incorporate its customer queries data to enhance retail sales forecasts. Boone, Ganeshan, and Hicks [16] and Boone et al. [17] made first attempts to examine if search volumes for certain search terms can improve the sales forecasts of specific products. They have been working with an online retailer specialized in food and cookware. Google Trends time series were retrieved for selected search terms that may lead the customer to the retailer according to the business owner. The premise is that if a customer searches for a certain term, it may shows an intent of the customer to explore and potentially buy the product. Then, the external time series were included in forecast models resulting in a decrease of forecast errors.

The idea and structure of the presented case studies regarding the use of Google Trends are leading the way for the selected approach taken in this work. External time series are extracted from Google Trends and integrated into forecast models in order to improve the forecasting performance. Boone et al. [17] and Elshendy et al. [40] even use the same linear time series models that will be applied later. However, there are various aspects in which this work differs greatly from the related work mentioned so far. All examples have in common that the search terms have already been determined a priori. Boone et al. [17] have fixed the search terms after consultation with the retailer either according to high-level product categories or based on unique

selling propositions. This way, external regressors refer to the population and not to specifications of individual products. In addition, in all examples the whole set of external time series was integrated into the forecasting models and subsequently evaluated together. Accordingly, no similarity measure has been computed to merge time series and to determine individual leading indicators. Therefore, the similarity is not considered in the forecasting process as the indicators are already set. Elshendy et al. [40] do informally report similarity measures, but they were not used in a preceding step to define leading indicators. Ultimately, only the results about whether there have been improvements or deteriorations due to the integration of the external data source are communicated. No factors or conditions are mentioned that are essential for the successful inclusion of leading indicators. This work aims to connect precisely at these points in order to close the apparent gap. The procedure in which leading indicators are determined and the unveiling of critical success factors distinguish this work from the previous work.

1.5 OUTLINE

Following the introduction, chapter 2 presents the theoretical foundations of this work. First, the statistical properties and features of time series data are introduced in section 2.1. This introduction is essential to understand how time series models use the inherent structures of time series to make forecasts. The time series models presented in this section consist of simple baseline models and a more complex linear time series model with its potential extensions. This section ends with the presentation of a forecast evaluation procedure that can be utilized to assess and compare the forecasting performance of the models. Subsequently, the decomposition of time series and one well known decomposition method will be explained in section 2.2. This section furthermore defines a parameter to properly quantify the strength of possible seasonal time series patterns. Section 2.3 describes the similarity matching for the leading indicator search. Here, a statistic will be introduced to measure the similarity between two time series. The chapter ends with the presentation of multivariate analysis methods, which are used to evaluate the results of this work.

Chapter 3 specifies the design of the experiment. This includes a detailed description of the datasets in section 3.2 and a comprehensive explanation of the experimental pipeline in section 3.3. Within the experimental pipeline, the underlying methodology will be illustrated step by step.

The results obtained in the experiment are presented in chapter 4. This covers a performance comparison of the time series models in section 4.1 as well as the analysis of selected time series pairs in section 4.2. Based on these two sections, the cross-correlation and the seasonal strength are analyzed as key indicators in section 4.3. Up to this point, qualitative statements and hypotheses will be formulated, which are then will be evaluated using

multivariate analysis methods. Their outcomes will be explained in section 4.4. The key findings and special aspects that should be considered when conducting the experiment are discussed in section 4.5.

The final chapter 5 summarizes the main findings and answers the research questions of this work.

2 THEORETICAL FOUNDATIONS

This chapter provides the theoretical background of this work. This includes the definition of concepts and the presentation of selected models for time series forecasting in section 2.1. This section particularly introduces the statistical properties and features of time series that are fundamental for time series analysis. In addition to the presented time series models, a procedure for evaluating the corresponding forecasting accuracies will be addressed. With the results of this procedure, the models can be compared and potential key factors for the experiment can be extracted. Section 2.1 is followed by the decomposition of time series data which enables an analysis of time series patterns and components. This decomposition is essential for defining indicators that can be used to measure the strength of time series patterns. Subsequently, the leading indicator search with its similarity matching will be explained in section 2.3. Here, the idea of leading indicator search is highlighted, and a method is presented with which two time series can be tested for similarity. Last, three multivariate analysis methods are proposed for the validation of the experimental results.

2.1 TIME SERIES FORECASTING

2.1.1 *Terminology*

This subsection introduces the fundamental terms in form of basic definitions. These definitional delimitations are necessary to understand the core concepts of the time series models that follow.

2.1.1.1 *Time Series*

A *time series* is an ordered collection of observations $(X_t)_{t=1,\dots,T}$ measured sequentially through time, where T denotes the number of observations obtained until the current point in time. A time series can be continuous or discrete, depending on whether the measurements are made continuously through time or are taken at a discrete set of time points. For a discrete time series, the data are typically recorded at equal intervals of time (Bourier [18, p. 155], Chatfield [24, p. 11]).[1] This may be the hourly development of crude oil prices, the minute-by-minute change of a company share value or the weekly sales of a product in a certain retail store. The frequency, in which the observations are made, is called the *sampling rate*. What makes time series data special is that successive observations are usually not independent and time series analysis must take account of the order in which

1 All the time series examined in this work are discrete. For this reason, all subsequent descriptions refer to discrete time series data.

the observations are collected. In addition, time series analysis is different from other statistical problems in that the observed time series is mostly the only realization that will ever be observed. Therefore, describing the data using summary statistics, finding suitable statistical models to describe the data generating process and predict future values of the series constitute the three main objectives of time series analysis (Chatfield [24, p. 12 f.]).

2.1.1.2 *Backward shift and backward difference operator*

Throughout this work, the *backward shift* operator $\mathcal{B}$ and *backward difference* operator ∇ are employed extensively. The backward shift operator $\mathcal{B}$ is defined by $\mathcal{B}X_t = X_{t-1}$ and has the effect of shifting the data back one period. Hence, multiple applications of $\mathcal{B}$ to X_t shifts the data back multiple periods: $\mathcal{B}^j X_t = X_{t-j}$. The backward difference operator ∇ is defined by $\nabla X_t = X_t - X_{t-1}$. Since $\mathcal{B}$ is also a convenient operator for describing the process of differencing, the previous formula can alternatively be written as $\nabla X_t = X_t - X_{t-1} = (1 - \mathcal{B})X_t$. A second-order differencing can be written as $(1 - \mathcal{B})^2 X_t$. Accordingly, the dth-order difference of X_t is provided by $(1 - \mathcal{B})^d X_t$. In general, the backshift notation is particularly useful when combining differences, as the operators can be treated using ordinary algebraic rules (Hyndman and Athanasopoulos [51], Box et al. [20, p. 7]).

2.1.1.3 *Stationarity*

A common approach in the analysis of time series data is to consider the observed time series as part of a realization of a process. This process can either be of deterministic or stochastic nature (Box et al. [20, p. 6 f.]). When a series is said to be deterministic, its future values can be predicted exactly from past values. On the other hand, a time series is stochastic in that the future is only partly determined by past values. Since most series are stochastic, the following explanations restrict attention on a model for a stochastic time series, often called a stochastic process[2]. The latter can be described as a family of random variables indexed by time and will be denoted by $\{X_t, t \in T\}$, where T denotes the index set of times on which the process is defined. In this context, the observed value at time t, namely x_t, will be regarded as an observation on an underlying random variable, X_t. Eventually, the observed time series is called a sample realization of a stochastic process that describes its probability structure (Chatfield [24, p. 24 f.], Brockwell and Davis [22, p. 7]).

An important class of stochastic processes for describing time series are *stationary processes*. Stationary processes assume that the probabilistic properties of the process do not change over time, in particular varying about a fixed constant mean level and with constant variance (Box et al. [20, p. 7]). At this point, a distinction must be made between strict and weak stationarity.

2 Most authors use the terms *model* and *process* interchangeably.

The following definitions will help to clarify the difference between these two forms (Chatfield [24, p. 25], Brockwell and Davis [22, p. 15]).

Definition 1 (Strict stationarity) A stochastic process is said to be strictly stationary, if $(X_{t_1}, \ldots, X_{t_n})$ and $(X_{t_1+k}, \ldots, X_{t_n+k})$ have the same joint distributions for all integers k and $n > 0$. Thus, the joint distribution of any set of observations must be unaffected by shifting all the times of observations forward or backward by any integer amount k (Box et al. [20, p. 24]):

$$(X_{t_1}, \ldots, X_{t_n}) \sim (X_{t_1+k}, \ldots, X_{t_n+k}), \quad \forall t_1, \ldots, t_n, k \tag{2.1}$$

Definition 2 (Weak stationarity) A stochastic process is said to be weakly stationary, if its first- and second-order moments are finite and do not change through time. Conversely, this means that the mean function as the first-order moment

$$\mu_X(t) = \mathrm{E}(X_t), \quad \forall t \in \mathbb{Z}$$

and the covariance between X_t and X_{t+k} as the second-order moment for different values of t and k

$$\gamma_X(t, t+k) = \mathrm{Cov}(X_t, X_{t+k}) = \mathrm{E}[(X_t - \mu)(X_{t+k} - \mu)], \quad \forall t, k \in \mathbb{Z}$$

are independent of t,

$$\begin{aligned} \mu_X(t) &\equiv \mu, & \forall t \in \mathbb{Z}, \\ \gamma_X(t, t+k) &\equiv \gamma_X(k), & \forall t, k \in \mathbb{Z}. \end{aligned} \tag{2.2}$$

$\gamma(k)$ is called *autocovariance function* and $\gamma(0)$ equals the variance σ^2 when the lag k is zero (Brockwell and Davis [22, p. 15]). The most fundamental example of a stationary process is a sequence of independent and identically distributed (i.i.d.) random variables with zero mean $\mathrm{E}(X_t) = 0$ and constant variance $\mathrm{Var}(X_t) = \sigma^2$. This process is weakly stationary and is referred to as a *white noise process*. The autocovariance function is given by:

$$\gamma_X(k) = \begin{cases} 0, & k \neq 0 \\ \sigma^2, & k = 0. \end{cases}$$

By convention, this work uses Z_t (variously called series of *innovations*, *shocks* or *errors*) and $Z \sim \mathcal{WN}(0, \sigma^2)$ to denote a white noise process. Although the white noise process has very basic properties and is rarely used to describe data directly, it is often used to model the random disturbances in more complicated processes (Box et al. [20, p. 28 f.], Chatfield [24, p. 28]).

2.1.1.4 *Linear processes*

The class of linear processes provides a general framework for studying stationary processes (Brockwell and Davis [22, p. 51]). If a time series X_t has the representation

$$X_t = c + \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j} \tag{2.3}$$

for all t, where $Z \sim \mathcal{WN}(0, \sigma^2)$ and c is some origin (e.g., its mean μ), the series is a linear process. By using the backward shift operator $\mathcal{B}$, equation 2.3 can be written more compactly as

$$X_t = \psi(\mathcal{B}) Z_t, \tag{2.4}$$

where $\psi(\mathcal{B}) = \sum_{j=-\infty}^{\infty} \psi_j \mathcal{B}^j$. In this equation the operator $\psi(\mathcal{B})$ can be thought of as a linear filter. When applied to a white noise input series Z_t, it produces the output X_t. Equation 2.4 allows the linear process to represent X_t as a weighted sum of present and past values of the white noise process Z_t (Box et al. [20, p. 47], Brockwell and Davis [22, p. 51]).

However, the representation in 2.3 of the general linear process would not be very useful in practice if it contained an infinite number of parameters ψ_j. *Autoregressive (AR)*, *Moving average (MA)* and *Mixed Autoregressive Moving average (ARMA)* processes introduce parsimony and are representationally useful in modeling time series data (Box et al. [20, p. 52]).

AUTOREGRESSIVE PROCESSES The first special case of general linear processes are the AR processes, in which only the first p of the weights are nonzero. The term *autoregressive* indicates that the process is a regression of the variable X_t against itself using a linear combination of its past values (Hyndman and Athanasopoulos [51]). The process can be written as

$$X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + Z_t, \tag{2.5}$$

where the symbols $\phi_1, \phi_2, \ldots, \phi_p$ represent a finite set of weight parameters (Box et al. [20, p. 8 f.]). The AR process is referred as an AR(p) model, an autoregressive model of order p. AR models are remarkably flexible at handling a wide range of different time series patterns (Box et al. [20, p. 52]).

MOVING AVERAGE PROCESSES Compared to the AR process, the MA process uses past white noise shocks or errors in a regression-like model (Hyndman and Athanasopoulos [51]):

$$X_t = c + \theta_1 Z_{t-1} + \theta_2 Z_{t-2} + \cdots + \theta_q Z_{t-q} + Z_t. \tag{2.6}$$

The symbols $\theta_1, \theta_2, \ldots, \theta_q$ represent analogously the finite set of weight parameters and X_t can be thought of as a weighted linear sum of the last q

random shocks Z_t (Box et al. [20, p. 9 f.]). The MA process is referred as a MA(q) model, a moving average model of order q (Box et al. [20, p. 53]).

Changing the AR(p) parameters $\phi_1, \ldots, \phi_p$ and MA(q) parameters $\theta_1, \ldots, \theta_q$ results in different time series patterns. The variance of the random shocks Z_t will only change the scale of the series.

MIXED AUTOREGRESSIVE MOVING AVERAGE PROCESSES If certain constraints on the values of the parameters are met, it is possible to write any stationary AR(p) model as an MA(∞) model, and vice versa (Box et al. [20, p. 51 f.], Chatfield [24, p. 36 ff.]). However, this may involve an infinite number of parameters which are impossible to estimate from a finite set of data. Many real-world time series data can be approximated in a more parsimonious way (meaning fewer parameters are needed) (Chatfield [24, p. 38 f.]). Therefore, it is often useful to include both AR and MA terms in a class of mixed ARMA models, which aim to use as few parameters as possible (Box et al. [20, p. 10]). The resulting process

$$\begin{gathered} X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = c + Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q} \\ \text{or} \\ \phi(\mathcal{B}) X_t = c + \theta(\mathcal{B}) Z_t \end{gathered} \tag{2.7}$$

with constant c, is said to be an ARMA model of order (p, q) (Box et al. [20, p. 10]). Figure 2.1 shows a simulated time series with weekly sampling rate based on an ARMA$(1, 1)$ process with $\phi_1 = 0.8$ and $\theta_1 = 0.4$ scaled on a range of $[0, 1]$.

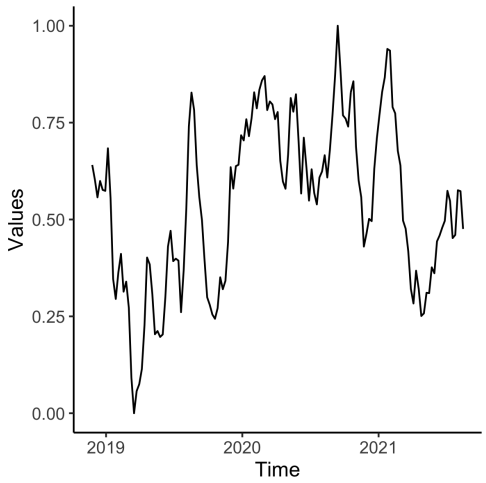


Figure 2.1: Simulated time series with weekly sampling rate based on an ARMA$(1, 1)$ process with $\phi_1 = 0.8$ and $\theta_1 = 0.4$ scaled on a range of $[0, 1]$.

2.1.1.5 *(Partial) Autocorrelation function*

The previous descriptions introduced a time series as a sample realization of a stochastic process, either as a linear combination of its past values, as

a weighted linear sum of the last q random shocks or as an inclusion of both. Along with the autocovariance function, the *autocorrelation function* can be used at this point to assess the degree of dependence in the time series data. Both functions play a crucial role in the problem of constructing an appropriate model for the data (Box et al. [20, p. 43]). Just as the correlation measures the extent of a linear relationship between two variables, the autocorrelation function measures for stationary processes the linear relationship between X_t and X_{t+k} at lag k. An additional useful function, which is complementary to the autocorrelation function, is the *partial autocorrelation function*, which measures the excess correlation at lag k which has not already been accounted for by autocorrelations at lower lags (Hyndman and Athanasopoulos [51]). The autocorrelation function of a time series X_t can be written as

$$\begin{aligned} p_X(k) &= \mathrm{Cor}(X_t, X_{t+k}) \\ p_X(k) &= \frac{\mathrm{E}[(X_t - \mu)(X_{t+k} - \mu)]}{\sqrt{\mathrm{E}[(X_t - \mu)^2]\mathrm{E}[(X_{t+k} - \mu)^2]}} \\ p_X(k) &= \frac{\mathrm{E}[(X_t - \mu)(X_{t+k} - \mu)]}{\sigma_X^2} \\ p_X(k) &= \frac{\gamma_X(k)}{\gamma_X(0)}. \end{aligned} \tag{2.8}$$

Following the definition of a stationary process, the variance $\sigma_X^2 = \gamma_X(0)$ is the same at time $t + k$ as at time t. An autocorrelation function $p_X(k)$ has all the properties of an autocovariance function (e.g., it is an even function, since $p_X(k) = p_X(-k)$). Furthermore, it satisfies the additional condition $p_X(0) = 1$ and has the usual property of correlation that $|p_X(k)| \leq 1$ (Box et al. [20, p. 25]).

However, in practical problems the theoretical autocorrelation function is unknown and there is a finite time series $x_1, x_2, \ldots, x_T$ of T observations, from which only estimates of the autocorrelations can be obtained (Box et al. [20, p. 30 f.]). The most satisfactory estimate of the autocorrelation function is called *sample autocorrelation function* or *correlogram* and can be written as

$$r_X(k) = \hat{p}_X(k) = \frac{c_X(k)}{c_X(0)}, \tag{2.9}$$

where

$$c_X(k) = \hat{\gamma}_X(k) = \frac{1}{T} \sum_{t=1}^{T-|k|} (X_t - \bar{X})(X_{t+|k|} - \bar{X}), \quad -T < k < T \tag{2.10}$$

is the estimate of the autocovariance function $\gamma_X(k)$ and $\bar{X}$ is the sample mean of the time series. The sample (partial) autocorrelation function is one of the most useful tools in assessing the behavior and properties of a time series. It provides a general procedure for identifying which of the many

possible stationary time series models is a suitable candidate for representing the dependence in the data (in the same way that plotting a histogram helps to indicate which family of distributions may be appropriate) (Brockwell and Davis [22, p. 18], Chatfield [24, p. 30 f.]). One common pattern for stationary time series is to see short-term correlation where perhaps the first three or four values of $r_X(k)$ are significantly different from zero. If the sample autocorrelation coefficients seem to decrease in an approximately exponential way, then an AR(1) model is indicated. A higher-order AR model may be appropriate, if the coefficients behave in a more complicated way. A MA(1) is indicated, if the only significant autocorrelation is at lag one. When a time series exhibits a trend, the coefficients for small lags tend to be large and positive because observations nearby in time are also nearby in value. Therefore, the sample autocorrelation function of a trended time series tends to have positive values that slowly decrease as the lags increase. When a time series exhibits a seasonal pattern, the autocorrelations will be larger for the seasonal lags than for other lags. Hence, the sample autocorrelation function can help to determine the order of a stationary process and is often used to see if seasonality is present (Hyndman and Athanasopoulos [51], Chatfield [24, p. 32]). The terms *trend* and *seasonality* will be further explained in subsection 2.2.1.

2.1.1.6 *Forecasting*

Forecasting a time series can be defined by finding a linear combination of observed values $x_1, x_2, \ldots, x_T$ that result in minimum forecast errors for future values $x_{T+h}, h > 0$. The forecasts of x_{T+h} made at time T for h steps ahead will be denoted by $\hat{x}_T(h)$, where the integer h is called the *lead time* or *forecasting horizon* (Chatfield [24, p. 3], Box et al. [20, p. 2]). Forecast errors can be described as the deviation between the prediction and the actual observed value. To understand the concept, one can think of daily sales of a product in a certain retail store last month ($T = 30$). Based on the historical values $x_1, \ldots, x_{30}$, the task is to predict the sales of this product for the next week, i.e., $\hat{x}_{30+1}, \ldots, \hat{x}_{30+7}$ for a daily sampling rate with lead time $h = 7$. Since most series are not deterministic, it is crucial to recognize the structure, patterns and influencing factors of a time series. However, those factors are typically random so that time series forecasting is rather finding an appropriate model to accurately represent this random behavior. If this random behavior is of stationary nature, valid techniques can be developed for time series forecasting (Chatfield [24, p. 24]).

2.1.2 *Baseline models*

The following descriptions are highly condensed and merely give an overview of selected models that intend to serve as a benchmark for the experiment. The forecast errors achieved by these models form the baseline since they follow rather simple rule-based approaches. Therefore, only the basic concepts

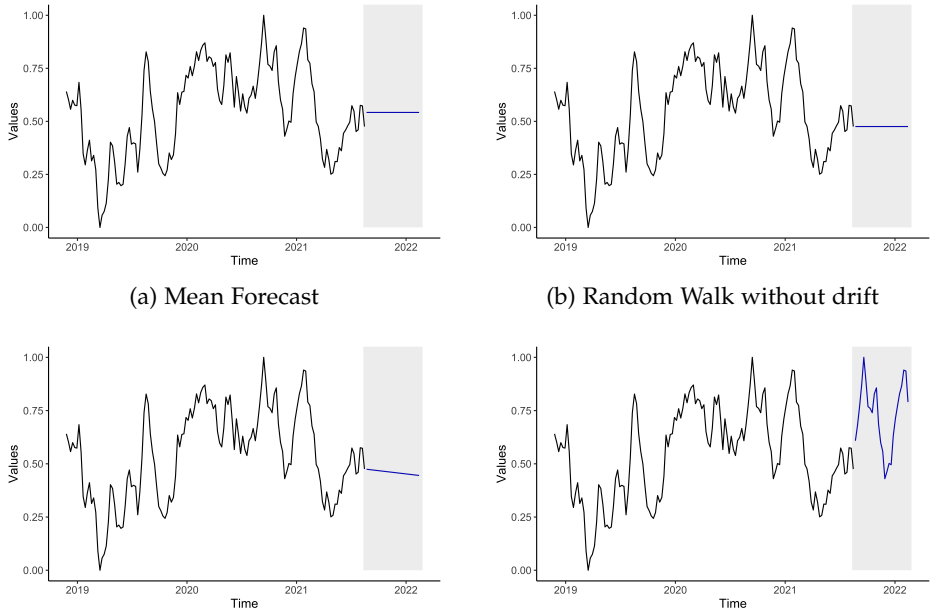
of these models are briefly explained. They all have in common that they are univariate forecasting models that work on isolated time series.

MEAN FORECAST The Mean Forecast model returns forecasts and prediction intervals for an i.i.d. model applied to time series Y_t based on the historical mean. The underlying model can be described with $Y_t = \mu + Z_t$. The resulting forecasts for a specified horizon h are given by $\hat{Y}_{t+h} = \mu$, where μ is estimated by the sample mean (Hyndman and Khandakar [53], Hyndman et al. [52]).

RANDOM WALK To start with, a *random walk process* X_t is defined by $X_t = X_{t-1} + Z_t$ (Chatfield [24, p. 29]). If the random walk starts at time $t = 0$ the process can also be written as $X_t = X_0 + \sum_{j=1}^{t} Z_j$ so that X_t is the accumulation of all past innovations (Mills [67, p. 283]). According to the hypothesis of the random walk process, future steps or directions cannot be predicted on the basis of past history (Malkiel [65, p. 26]). Nonetheless, the corresponding Random Walk model, firstly employed by Pearson [77] in 1905, is one of the simplest and yet most important models in time series forecasting (Nau [69]). The model can be divided into a model with and without *drift*. Both variants assume that in each period, X_t takes a random step away from its previous value, and the steps are i.i.d.. If the average value of the step size is zero, a Random Walk model without drift is present. The corresponding h-step-ahead forecast for a time series Y_t is given by $\hat{Y}_{t+h} = Y_t$. This formula denotes that all future values will equal the last observed value. If the average step size is a nonzero value c, the model is said to be a Random Walk model with drift. It follows the process $X_t = c + X_{t-1} + Z_t$ where c is the nonzero drift parameter. This parameter can be defined as the average increase from one period to the next (Nau [69]). The h-step-ahead forecast is provided by $\hat{Y}_{t+h} = ch + Y_t$ (Hyndman and Khandakar [53], Hyndman et al. [52]). The resulting long-term forecasts are looking like a trend line with slope c, but it is always re-anchored on the last observed value (Nau [69]).

SEASONAL NAÏVE If the data follow a random walk process, a Naïve model is optimal, since all forecasts for time series Y_t are set to be the value of the last observation $\hat{Y}_{t+h} = Y_t$. This is equivalent to a Random Walk model without drift ($c = 0$). If the underlying time series is highly seasonal, a similar model, called Seasonal Naïve, can be applied. In this case, each forecast is set to be equal to the last observed value from the same season (e.g., the same week of the previous year). The model can be written as $Y_t = Y_{t-s} + Z_t$ where Z_t is a normal i.i.d. error and s is the seasonal period. The h-step-ahead forecast is given by $\hat{Y}_{t+h} = Y_{t+h-s(k+1)}$ where k is the integer part of $(h-1)/s$ (i.e., the number of complete years in the forecast period prior to time $t+h$) (Hyndman and Athanasopoulos [51]).

Figure 2.2 shows forecasts of the described baseline models for the simulated ARMA$(1,1)$ time series.



(a) Mean Forecast

(b) Random Walk without drift

(c) Random Walk with drift

(d) Seasonal Naïve

Figure 2.2: Forecasts of baseline models for a simulated time series based on an ARMA$(1,1)$ process with weekly sampling rate and a forecast horizon of $h = 26$.

2.1.3 *Linear Time Series models*

With the recent advancement in computational power of computers and more importantly the development of cutting-edge machine learning approaches such as deep learning, new algorithms are developed to analyze and forecast time series data (Siami-Namini, Tavakoli, and Namin [84], Abdoli, MehrAra, and Ardalani [1]). However, for more than half a century linear time series models have dominated many areas of time series forecasting (Xie and Goh [98]). Especially one linear time series model, namely Auto-Regressive Integrated Moving Average (ARIMA), has demonstrated its outperformance in precision and accuracy in many real-world applications (Siami-Namini, Tavakoli, and Namin [84]). Due to its statistical properties as well as the underlying well known methodology it is on the most important and widely used time series models (Zhang [99]). For this reason, this model will constitute the core time series model for this work and the corresponding experiment. The following subsection will introduce this model and two of its potential extensions.

2.1.3.1 *Auto-Regressive Integrated Moving Average model*

In practice, many empirical time series are non-stationary and stationary AR, MA or ARMA processes cannot be applied directly. One possible way of handling non-stationary series is to apply differencing assuming that some suitable difference of the process is eventually stationary, on which an ARMA(p,q)

model can then be fitted in the usual way. As described in subsection 2.1.1, differencing can be applied multiple times $(1-\mathcal{B})^d X_t$, where d is the order of differencing. There are an unlimited number of ways in which a time series can reveal non-stationary behavior. However, if the original time series is differenced d times before fitting an ARMA(p,q) process, then the model for the original undifferenced series that can handle non-stationary behavior is said to be an ARIMA process of order p, d and q: ARIMA(p,d,q).[3] The letter I in the acronym ARIMA stands for integrated and d denotes the number of differences taken (Box et al. [20, p. 80 ff.], Chatfield [24, p. 41 f.]). Therefore, the ARIMA model is a modified form of the ARMA model and can be written as (Hyndman and Athanasopoulos [51])

$$\begin{gathered}\underbrace{(1-\phi_1\mathcal{B}-\cdots-\phi_p\mathcal{B}^p)}_{\text{AR(p)}}\ \underbrace{(1-\mathcal{B})^d X_t}_{d\text{ differences}} = c + \underbrace{(1+\theta_1\mathcal{B}+\cdots+\theta_q\mathcal{B}^q)Z_t}_{\text{MA(q)}} \\ \phi(\mathcal{B})(1-\mathcal{B})^d X_t = c + \theta(\mathcal{B})Z_t.\end{gathered} \tag{2.11}$$

All the models already presented in this work are special cases of the ARIMA model, as shown in table 2.1.

White noise	ARIMA(0, 0, 0) with no constant
Random walk	ARIMA(0, 1, 0) with no constant
Random walk with drift	ARIMA(0, 1, 0) with a constant
Autoregression	ARIMA(p, 0, 0)
Moving average	ARIMA(0, 0, q)

Table 2.1: Special cases of ARIMA models (Hyndman and Athanasopoulos [51]).

Analogously to the baseline models presented in 2.1.2, an ARIMA model can also be used to make forecasts. Point forecasts of this model are obtained based on the following three steps (Hyndman and Athanasopoulos [51]):

1. Expand the ARIMA equation 2.11 so that X_t is on the left hand side and all other terms are on the right,
2. Rewrite the equation by replacing t with $T+h$,
3. On the right hand side of the equation, replace future observations with their forecasts, future errors with zero, and past errors with the corresponding residuals.

This procedure starts with a forecast horizon of $h=1$ and is repeated continuously for $h=2,3,\ldots$ until all forecasts have been calculated.

2.1.3.2 *Seasonal Auto-Regressive Integrated Moving Average model*

Up to this point, all previous remarks to stationary linear processes were restricted to non-seasonal data. However, ARIMA models are also capable of

3 ARIMA models are sometimes called *Box-Jenkins models* based on the original key reference Box and Jenkins [19] from 1970.

modeling a wide range of seasonal data (Hyndman and Athanasopoulos [51]). In general, a time series is said to exhibit periodic behavior with s time periods per year, when similarities in the series occur after s basic time intervals. This means that observations that are s intervals apart are similar. Let $\mathcal{B}^s$ denote the operator such that $\mathcal{B}^s X_t = X_{t-s}$ and since a non-stationary series is still assumed let the simplifying operation $\nabla_s X_t = (1 - \mathcal{B}^s) X_t = X_t - X_{t-s}$ denote seasonal differencing (Box et al. [20, p. 306 ff.], Chatfield [24, p. 42 f.]). Based on these operators a general Seasonal Auto-Regressive Integrated Moving Average (SARIMA) model with non-seasonal terms of order (p, d, q) and seasonal terms of order (P, D, Q) may be written as

$$\phi(\mathcal{B})\Phi(\mathcal{B}^s)(1 - \mathcal{B})^d(1 - \mathcal{B}^s)^D X_t = \theta(\mathcal{B})\Theta(\mathcal{B}^s)Z_t, \tag{2.12}$$

where Φ, Θ denote polynomials in $\mathcal{B}^s$ of order P, Q respectively (Chatfield [24, p. 42]). One can see from equation 2.12 that a SARIMA model is formed by including additional seasonal terms in the ARIMA model. Ultimately, a SARIMA model can be abbreviated as

$$\text{ARIMA} \quad \underbrace{(p, d, q)}_{\substack{\text{Non-seasonal part} \\ \text{of the model}}} \quad \times \quad \underbrace{(P, D, Q)_s}_{\substack{\text{Seasonal part} \\ \text{of the model}}},$$

where s is the seasonal period (e.g., number of observations per year) (Hyndman and Athanasopoulos [51]).

2.1.3.3 *Auto-Regressive Integrated Moving Average Exogenous Variable model*

ARIMA processes can be used to model non-stationary time series. Furthermore, the SARIMA model was introduced as an extension of the general ARIMA model by including the ability to handle seasonal data with seasonal terms. Hence, ARIMA models are capable of modeling a wide range of various time series patterns. However, in the current form as denoted in 2.11 respectively 2.12 ARIMA models are only applicable for univariate time series data. Therefore, the model must meet the requirement to be able to also integrate external regressors in the modeling and forecasting process in order to be used as the primary model within experiment. Fortunately, there exist an additional extension of the ARIMA model for this requirement. This extension is called an Auto-Regressive Integrated Moving Average Exogenous Variable (ARIMAX) model. It enlarges the ARIMA model through the inclusion of external variables and at the same time give evidence of the contribution of each of them (Elshendy et al. [40]). As an extension of ARIMA, ARIMAX inherits the capacity to identify the underlying patterns in time series data and to quantify the impact of each external regressor (Victor-Edema and Essi [91]). The following equation 2.13 will describe an ARIMAX process for a time series Y_t with non-seasonal data and the inclusion of one external regressor

X_t. The idea can be easily extended to include seasonal terms and multiple external variables. Using backward shift operators, the model is given by

$$\phi(\mathcal{B})(1-\mathcal{B})^d Y_t = c + \beta X_t + \theta(\mathcal{B})Z_t, \tag{2.13}$$

where $\phi(\mathcal{B}) = 1 - \phi_1\mathcal{B} - \cdots - \phi_p\mathcal{B}^p$, $\theta(\mathcal{B}) = 1 + \theta_1\mathcal{B} + \cdots + \theta_q\mathcal{B}^q$, X_t is an external variable at time t and β is its coefficient (Fredén and Larsson [42, p. 10 f.], Victor-Edema and Essi [91]).

While this extension looks straightforward, one disadvantage is that the coefficient β of the external variable is hard to interpret. In a linear regression setting, the value of β is the effect on Y_t when X_t is increased by one unit. This is not the case for an ARIMAX model. The presence of lagged values of the time series Y_t in $\phi(\mathcal{B})(1-\mathcal{B})^d Y_t$ means that β can only be interpreted conditional on the value of previous values of Y_t, which is hardly intuitive (Hyndman [49]).

2.1.3.4 *Model selection and diagnostic checking*

The linear time series models presented so far reveal numerous parameters that need to be estimated from a finite set of data. Hence, time series analysis and forecasting usually involves finding a suitable model for this set of data. However, it should be mentioned that this suitable model might be only an approximation to the data. Depending on the complexity of the time series patterns being modelled and the complexity and accuracy of the model, there will be departures from the model to a greater or lesser extent. Statistical model building is generally an iterative, interactive process and usually has three main stages (Chatfield [24, p. 71 f.]):

1. Model specification (or model identification),
2. Model fitting (or model estimation),
3. Model checking (or model verification or model criticism).

In terms of ARIMA models and its extensions, these three stages can be further refined using the following general procedure provided by Hyndman and Athanasopoulos [51]:

1. Plot the data and identify any unusual observations.
2. If necessary, transform the data to stabilize the variance.
3. If the data are non-stationary, take first differences of the data until the data are stationary.
4. Examine the (partial) autocorrelation function: Is an ARIMA$(p, d, 0)$ or ARIMA$(0, d, q)$ model appropriate?
5. Try the chosen model(s), and use a model selection criterion to search for a better model.

6. Perform diagnostic checks by doing appropriate statistical tests.
7. Once the diagnostic checks are passed, calculate forecasts.

This procedure is necessary since determining an appropriate ARIMA model to represent an observed time series involves a number of interrelated problems. These include the choice of p and q (*order selection*) and the estimation of the mean μ, the parameters $\{\phi_i, i = 1, \ldots, p\}$, $\{\theta_i, i = 1, \ldots, q\}$, the white noise variance σ^2 and especially the order of differencing d respectively D for SARIMA models. In particular, the order selection is a crucial issue since in practice the true order of the model generating the data is unknown. For this reason, techniques are required to find an order that represents the data optimally in some sense (Brockwell and Davis [22, p. 137 ff.]).

Analyzing the sample (partial) autocorrelation function is one technique that can be helpful in determining the value of p or q. However, if p and q are both positive, the sample autocorrelation function and partial autocorrelation function are difficult to recognize and do not help in finding suitable values (Brockwell and Davis [22, p. 155]). In the case of $p > 0$ and $q > 0$ there need to be a more systematic approach using a sophisticated model selection criteria that gives a numerical-valued ranking of all value combinations. One of the most commonly used model selection criteria is the bias-corrected version of the Akaike's Information Criterion (AIC), denoted by AIC_C (Chatfield [24, p. 76]). AIC_C uses a maximum likelihood estimation technique that finds the values of the parameters which maximize the probability of obtaining the observed data. For ARIMA models, the AIC_C statistic can be written as

$$AIC_C = AIC + \frac{2(p+q+k+1)(p+q+k+2)}{T-p-q-k-2}, \tag{2.14}$$

where $AIC = -2log(\mathcal{L}) + 2(p+q+k+1)$, $k = 1$ if $c \neq 0$, $k = 0$ if $c = 0$ and $\mathcal{L}$ is the likelihood of the data (Hyndman and Athanasopoulos [51]). Measured by the likelihood function, this criterion essentially chooses the parameters and thus the model which minimize the AIC_C statistic, subject to a penalty term that increases with the number of parameters fitted in the model.[4] However, AIC_C tend not to be a good guide for selecting the appropriate order of differencing d respectively D of a model, but only for selecting the values of p and q. AIC_C values between models with different orders of differencing are not comparable, since differencing changes the data on which the likelihood is computed. Therefore, there need to be another approach to choose the order of differencing. One approach is to simply difference the non-stationary and potentially seasonal series d respectively D times until the series appears to be stationary and (most of) the seasonality is removed. An additional approach for further evaluation is the use of formal procedures, including testing for stationarity (Chatfield [24, p. 43], Hyndman and

4 Readers are referred to Brockwell and Davis [22, p. 171 ff.] for detailed information about the AIC_C model selection criterion.

Athanasopoulos [51]).[5]

After choosing appropriate values for p, q, d and D, the final selection of a model depends on a variety of diagnostic checks (also called *goodness of fit tests*[6]). These diagnostic checks can be systematized to a large degree by the use of model selection criteria such as the minimization of the AIC_C statistic. A common diagnostic check involves some sort of residual analysis. Residuals are in time series analysis generally the one-step-ahead forecasting errors written as

$$e_t = x_t - \hat{x}_{t-1},$$

where x_t denotes the observed value of X_t at time t and $\hat{x}_{t-1}$ is the forecast made for time t at time $t-1$. If the model is satisfactory, then the residuals should form a random series and resemble white noise. Residuals can be examined in several ways to make sure that they are consistent with their expected behavior under the model.[7] In order to assess the overall fit of the model, the residuals can be plotted against time over the whole period of fit and treated as a time series in its own right (Brockwell and Davis [22, p. 179 f.]). If the diagnostic checks suggest that the fitted model is inadequate, then the forecasting method based on it will not be optimal (Chatfield [24, p. 77 f.]).

2.1.4 *Evaluation of forecasts*

The performance of a time series model is usually measured by how well the model is able to predict the future. Intuitively, this is reflected in the deviation between the actual value Y_{T+h} and its corresponding forecast $\hat{Y}_{T+h}$. Based on forecast errors, several error metrics can be defined to assess the forecast accuracy and to enable an overarching model comparison. The question arises, whether there is a standard procedure to evaluate the performance and forecasting ability of time series models, especially on data the models have not yet seen.

To assess the generalizability of algorithms in classification and regression, Cross-validation (CV) is one of the most widely used methods (Hastie, Tibshirani, and Friedman [45, p. 241]). Because of its simplicity and universality, many results on model selection performances exist on CV procedures (Arlot and Celisse [6]). Larson [60] already noticed in 1931 that training an

5 A common statistical test for testing the presence of a unit root against the alternative hypothesis of stationarity is the *Dickey–Fuller test* (Dickey and Fuller [36]). An additional statistical test for testing the null hypothesis that a time series is stationary is the *Kwiatkowski–Phillips–Schmidt–Shin* (KPSS) test. The resulting p-value provides an indication whether differencing is required (Kwiatkowski et al. [59]).

6 Box et al. [20, Ch. 8, p. 284 ff.] provide full details on several diagnostic checks.

7 A well known test for residual analysis is the *Box-Ljung portmanteau lack-of-fit test*. This test tests if the residuals form a stationary series and resemble white noise (Box et al. [20, p. 289 ff.]).

algorithm and evaluating its statistical performance on the same data yields an overoptimistic result. In fact, the output of the algorithm should rather be tested on new, unseen data. Against this background, CV was raised to fix this issue and consists in its most primitive but nevertheless useful form in the controlled or uncontrolled splitting of the data into two samples (Stone [88]). One sample (often called training sample or training set) is used for fitting the algorithm. The remaining data (often called validation sample or validation set) are used for evaluation and can play the role of new data as long as data are i.i.d.. This becomes particularly useful when only a limited amount of data is available and a simple split in training and test set is insufficient (Arlot and Celisse [6]). At this point, it is important to mention that there are various splitting strategies that lead to various CV methods. Arlot and Celisse [6] provide a detailed overview of each method and what factors should be taken into account.

One of the most widely applied procedures is the well known K-fold CV. This procedure uses part of the available data to fit the model, and a different part to test it. The data is divided into K parts of approximately equal cardinality n/K. After the preliminary partitioning of data, the model is fitted to the $K-1$ parts of the data and the performance of the model is evaluated for the kth part. This operation is performed for the parts $k = 1, 2, \ldots, K$ ensuring that each part successively plays the role of the validation sample. The evaluation results of the K operations can then be combined (Hastie, Tibshirani, and Friedman [45, p. 241 f.], Arlot and Celisse [6]).

When it comes to time series forecasting, however, practitioners are often unsure of the best way to evaluate their models. There is often a feeling that one should not be using future data to predict the past (Bergmeir, Hyndman, and Koo [13]). A standard evaluation procedure in the traditional forecasting literature is the out-of-sample (OOS) procedure, where a section from the end of the series is withheld for evaluation. By using OOS, only one evaluation on a test set is accordingly considered, whereas with the use of CV, various such evaluations are performed. Therefore, the benefits of CV, especially for small datasets, cannot be exploited (Bergmeir, Hyndman, and Koo [13]).

For this reason, a more sophisticated version of the basic training and test split, called *time series CV*, has been developed. What distinguishes this procedure is that there are a series of test sets, each consisting of a single observation. Only the observations that occurred prior to the observation that forms the test set are part of the corresponding training set. Each iteration yield to new data in the training set and with each new data point a h-step-ahead forecast is made for a selected forecast horizon h. Hence, no future observations can be used in the forecasting process. According to this procedure, several forecast errors can then be calculated. Figure 2.3 illustrates the series of training and test sets, where the blue observations denote training

data, the red cells denote test data and the grey cells denote data that is not used in the specific iteration (Hyndman and Athanasopoulos [51]).

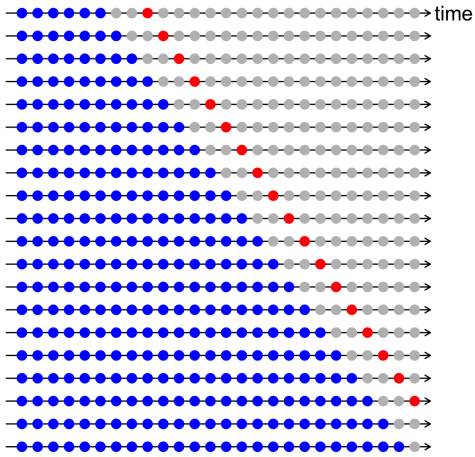


Figure 2.3: Time series CV with a multi-step-ahead forecast and a forecast horizon of $h = 3$ (Hyndman and Athanasopoulos [51]).

The forecast accuracy in this procedure is computed by averaging over the generated forecast errors made on the test sets. This technique is also known as *evaluation on a rolling forecasting origin* since the *origin* at which the forecast is based rolls forward in time (Hyndman and Athanasopoulos [51]). The forecast errors can then be utilized to compute error metrics in order to assess the achieved forecast accuracy of the underlying model.

2.2 DECOMPOSITION OF TIME SERIES

Many real-world time series data exhibit a variety of complex patterns. In order to make further investigations into the behavior of time series, it is often helpful to split a time series into several components, each representing an underlying pattern category. Decomposing a time series accurately into these components can reveal insights of the time series data from different perspectives and facilitate further analysis and time series tasks (Wen et al. [94], Hyndman and Athanasopoulos [51]). This section contains the definitional delimitation of time series components, the presentation of a common method for extracting these components and the definition of seasonal strength.

2.2.1 *Time Series components*

There are commonly three types of patterns a time series can reveal: *Trend*, *Seasonality* and *Cycles*. When a time series is decomposed into components, the trend and cycle are usually combined into a single trend-cycle component (sometimes called trend for simplicity). Ultimately, a time series comprises of a trend-cycle component, a seasonal component and a remainder

component. In describing time series data, these patterns and components need to be defined carefully (Hyndman and Athanasopoulos [51], Chatfield [24, p. 13 f.], Wen et al. [94]):

Definition 3 (Trend) A trend is defined as a long-term increase or decrease in the data. Hence, it describes the long-term direction of a time series. The values of a time series scatter around the trend over time.

Definition 4 (Seasonality) The seasonal component generally refers to repeated seasonal factors which affect the data. Those seasonal factors can be the time of the year (e.g., special events such as Christmas or Easter), the day of the week or other repeating patterns within any fixed period.

Definition 5 (Cycles) A cycle occurs when the time series exhibit rises and falls that are not of a fixed frequency. Economic conditions are usually the cause of these fluctuations and they are often related to the business cycle.

Definition 6 (Remainder) In addition to the components described so far, other variables can have an effect on the time series data. These can be factors which have a one-time effect or variables that are mostly unknown and repeatedly but irregularly affect the time series in their intensity and direction. These effects are summarized under the remainder component.

The decomposition of a time series can usually be of an additive or multiplicative form. If an additive form is assumed, the decomposition can be written as

$$Y_t = S_t + T_t + R_t, \tag{2.15}$$

where Y_t is the time series data, S_t the seasonal component, T_t the trend-cycle component and R_t is the remainder component, all at period t. A multiplicative decomposition would alternatively be written as:

$$Y_t = S_t \times T_t \times R_t. \tag{2.16}$$

The choice of decomposition form depends on the patterns the time series shows (Hyndman and Athanasopoulos [51], Brockwell and Davis [22, p. 23 f.]).

2.2.2 *Decomposition methods*

When choosing a forecasting model for a time series that includes trend, cycles and seasonality, it is crucial to apply a method that is able to capture its patterns properly (Hyndman and Athanasopoulos [51]). This subsection will briefly describe Seasonal and Trend decomposition using Loess (STL) as a method of time series decomposition.

STL, developed by Cleveland et al. [27], is one of the most classical and widely used decomposition methods (Wen et al. [94]). It is based on a sequence of smoothing and filtering procedures using a locally weighted regression, commonly known as Loess (*LOcal regrESSion*) (Cleveland et al. [27], Rojo et al. [80]). For a given time of observation, the Loess smoother is based on fitting a weighted polynomial regression, where weights decrease with increasing distance from the nearest neighbor (Dagum and Luati [33]). In a multi-step process in which moving averages alternate with Loess smoothing, the time series is fitted iteratively until trend and seasonality stabilize. The degree of smoothing in the trend and seasonal components are determined by six parameters. The components are extracted from the data series at the end of the STL process (Rojo et al. [80]). A detailed description of the alternating algorithm can be found in Cleveland et al. [27]. Zhou et al. [100] provide a graphical representation of the internal circulation process of STL. Figure 2.4 shows the simulated ARMA(1, 1) time series and its three components estimated by STL.

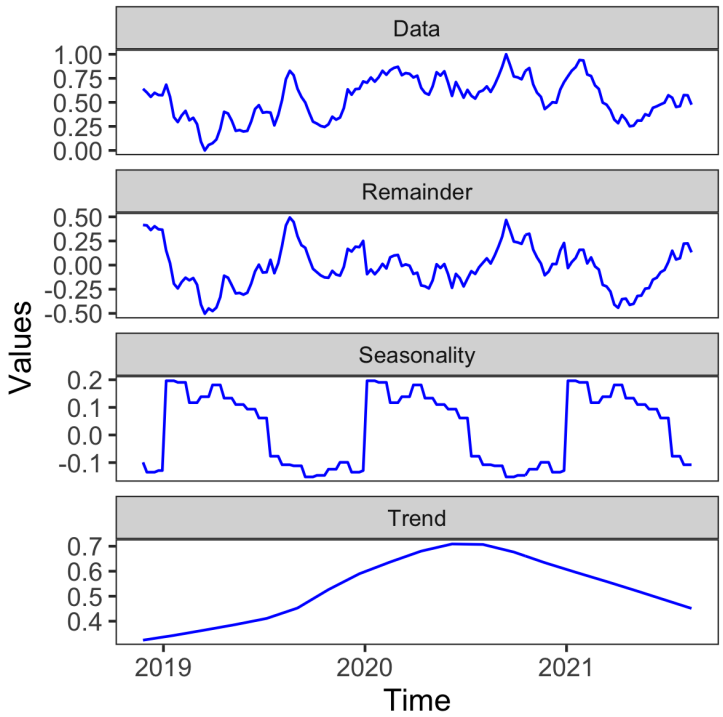


Figure 2.4: STL decomposition of the simulated ARMA(1, 1) time series.

The components are shown separately so that their relative behavior can be visualized. To reconstruct the data in the top panel, one can add the three components together. When the seasonal and trend-cycle components have been subtracted from the data, the remainder component in the second panel is left (Hyndman and Athanasopoulos [51]). However, many time series data from real-world applications are affected by several repeated activities or schedules, which introduce multiple seasonality. STL suffers from the inability to handle abrupt trend changes and less flexibility in the presence of

multiple seasonality. Therefore, an extension of STL called Multiple Seasonal-Trend decomposition using Loess (MSTL) has been proposed allowing the decomposition of time series with multiple seasonal patterns (Bandara, Hyndman, and Bergmeir [11]). The difference between these two methods lies in how to estimate multiple seasonal components (Wen et al. [94]).

2.2.3 *Definition of seasonal strength*

Retail sales, in particular, often exhibit strong seasonal variations, consequently making an effective modeling of retail sales time series a challenging task (Chu and Zhang [26]). Certain products may be bought frequently during the Christmas or Easter period. Outside of these events, however, these products may be less in demand. For this reason, it is even more important to define an indicator that can be used to estimate the seasonal strength of a time series. The decomposition method and the resulting time series components described in the previous subsections can also be used to handle this task (Wang, Smith, and Hyndman [92]). If a component is removed from the original data, the resulting values are the *component adjusted* data (e.g., the seasonally adjusted data for an additive decomposition is given by $Y_t - S_t$). If a time series shows strong seasonal patterns, the detrended data should have much more variation than the remainder component. Hence, the seasonal strength of a time series can be defined by

$$F_S = max\left(0, 1 - \frac{Var(R_t)}{Var(S_t + R_t)}\right), \tag{2.17}$$

where $Var(\cdot)$ is the variance of the corresponding component. A value of F_S close to 0 indicates almost no seasonality within the time series. The two variances should be approximately the same. Contrarily, a series with strong seasonality will have F_S close to 1 because $Var(R_t)$ will be much smaller than $Var(S_t + R_t)$ (Hyndman and Athanasopoulos [51]).

2.3 LEADING INDICATOR SEARCH

One of the main objectives of this work is to examine to what extent external data sources and thus external time series can be utilized to improve the forecast accuracy of a retail sales time series. Finding external time series that benefits the forecasting process is an approach, which is referred to in the micro- and macroeconomic literature as a *leading indicator search* (Bloom, Buckeridge, and Cheng [15]). The following explanations will clarify what leading indicator search is about and what methodology can be used to implement it.

2.3.1 *Definitions*

The term *leading indicator* can be defined as a variable whose significant fluctuations anticipate significant fluctuations in a target variable. In the time series domain, it shows similar patterns and thus exhibit a similar course as the target variable but delayed by a specific lag. If a time series is shown to provide an early indication of structural changes consistently, then the series is called a leading indicator (Bloom, Buckeridge, and Cheng [15]). Regarding the objectives proclaimed in this work, a retail sales time series is the target variable and will be denoted as *response* time series. The external time series is a potential leading indicator. This means that external time series whose structural changes provide an early indication of the development of the response time series may eventually have a beneficial impact on predicting its future values. Figure 2.5 illustrates highly simplified the concept of a lead-lag relationship between two time series.

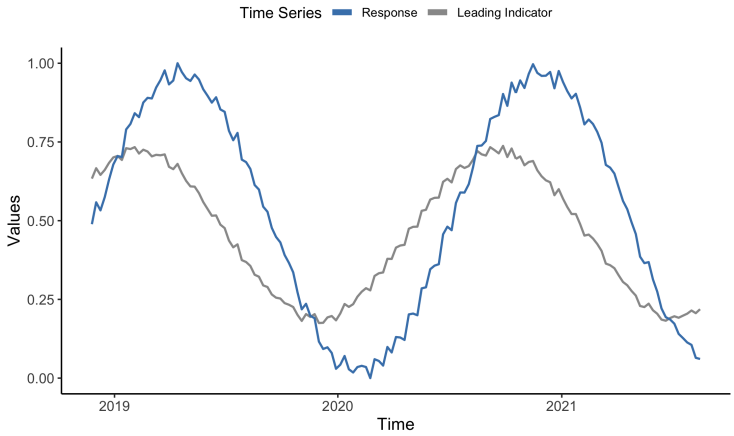


Figure 2.5: Illustration of a sample lead-lag relationship between two time series. Both time series have a weekly sampling rate. It can be seen that the leading indicator may be used to anticipate future fluctuations of the response time series.

The problem arises, how to describe and model the similarity and interrelationship existing between two time series. There are several methodological choices and approaches on identifying lead-lag relationships. A well known and commonly used statistical tool in evaluating and quantifying the strength and direction of the lead–lag relationship between two time series, is the *cross-correlation* analysis. By computing the correlation, expressed as the cross-correlation coefficient, between two time series at a large number of lags, the similarity between these two can be measured (Olden and Neff [75]). For this reason, the cross-correlation is a suitable metric to address the research questions and thus will excessively be exploited as a similarity measure in this work.

2.3.2 *Similarity matching with cross-correlation*

As presented in section 2.1, a stationary stochastic time series Y_t can be described by its mean μ, autocovariance function $\gamma_Y(k)$ and autocorrelation function $p_Y(k)$. For the measurement of similarity between two time series X_t and Y_t, the *cross-correlation function* can be employed as an analysis tool. It is a natural metric for measuring the similarity between segments of time series and helps to identify lead times between time series by lagging X_t to maximize the cross-correlation function (Bloom, Buckeridge, and Cheng [15]). For this purpose, it is useful to regard the pair of time series as realizations of a hypothetical population of pairs of time series, called a *bivariate stochastic process* (X_t, Y_t). In this case, X_t is the external time series that got classified as a leading indicator and Y_t is the response time series. In this work, the assumption is made that the bivariate stochastic process (X_t, Y_t) is stationary. This implies that the two time series X_t and Y_t have constant means (μ_X, μ_Y) and constant variances (σ_X^2, σ_Y^2) (Box et al. [20, p. 429]). Following this assumption, the corresponding *cross-covariance function* between X_t and Y_t at lag k can be written as

$$\begin{aligned} \gamma_{XY}(k) &= \mathrm{E}[(X_{t+k} - \mu_X)(Y_t - \mu_Y)] \\ \gamma_{XY}(k) &= \mathrm{Cov}(X_{t+k}, Y_t), \quad k = 0, \pm 1, \pm 2, \ldots. \end{aligned} \tag{2.18}$$

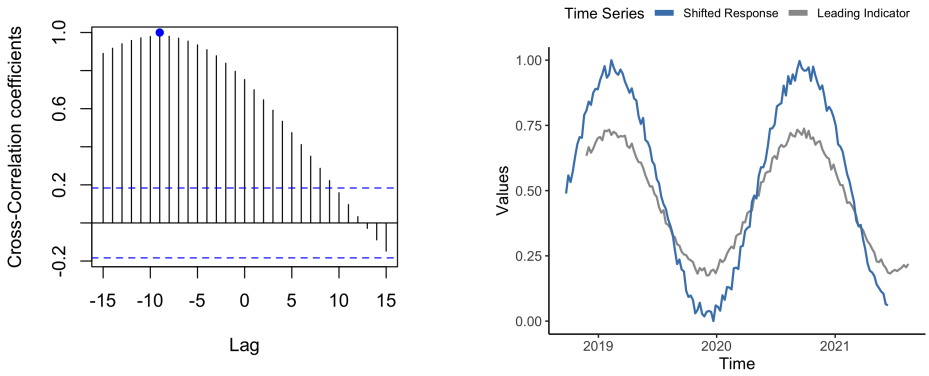
The cross-correlation coefficient at lag k between X_{t+k} and Y_t is similarly given by

$$p_{XY}(k) = \frac{\gamma_{XY}(k)}{\sigma_X \sigma_Y}. \tag{2.19}$$

p_{XY} is called the cross-correlation function of the stationary bivariate process. In contrast to the autocorrelation function, the cross-correlation function is not symmetric about $k = 0$, since $p_{XY}(k)$ is not in general equal to $p_{XY}(-k)$ (Chatfield [24, p. 27]). Similar to the autocorrelation function, the theoretical cross-correlation function remains unknown. The cross-correlation coefficients have to be estimated from the data, constituting the *sample cross-correlation function*. The sample cross-correlation function is provided by the estimate $r_{XY}(k)$ of the cross-correlation coefficient at lag k

$$r_{XY}(k) = \frac{c_{XY}(k)}{S_X S_Y}, \quad k = 0, \pm 1, \pm 2, \ldots, \tag{2.20}$$

where S_X and S_Y are the estimates of $\sqrt{c_{XX}(0)} = \sigma_X$ respectively $\sqrt{c_{YY}(0)} = \sigma_Y$ and $c_{XY}(k)$ is the estimate of the cross-covariance coefficient at lag k (Box et al. [20, p. 431 f.]). Figure 2.6a shows the estimated cross-correlation function of the two time series plotted in figure 2.5.



(a) Estimated cross-correlation function.

(b) Shifted response time series.

Figure 2.6: Estimated cross-correlation function and shifted response time series. Highest cross-correlation coefficient at lag $k = -9$. The blue dashed lines represent the approximated confidence interval for a 5% significance level $(\pm 1.96/\sqrt{n})$. Cross-correlation coefficients exceeding these limits can be considered significantly different from zero. Response time series shifted by lag $k = -9$.

In the context of a lead-lag relationship, only negative values of k are important, since the goal is to identify which external time series has a leading effect on the response time series Y_t. A negative value for k corresponds to a correlation between the external series X_t at a time before t and the time series Y_t at time t. In the example illustrated in figure 2.6a, the highest estimated cross-correlation coefficient is at lag $k = -9$. Since a similar course is expected, this means that knowing the value $X_{t=-9}$ of the external time series at time $t = -1$ may be advantageous for predicting $\hat{Y}_{t=0}$ (in case the forecast horizon is $h = 1$). To take advantage of this potential benefit, either the external time series or the response time series need to be shifted by lag k in an appropriate way. Figure 2.6b shows the sample response time series shifted by lag $k = -9$.

However, calculating the sample cross-correlation is not enough. Including leading indicators in a forecasting model introduces two modeling stages, which have to be taken into account in the further course of this work (Sagaert et al. [81]):

1. Selection of the appropriate leading indicators from the complete set of potential ones.

2. Evaluate their impact on the forecasting accuracy.

This work concentrates on an application, in which a bivariate time series forecast scheme is developed using the leading indicator in concert with the response time series. The ARIMAX model described in 2.1.3.3 provides such a scheme, in which the leading indicator is incorporated as an external regressor.

2.4 MULTIVARIATE ANALYSIS METHODS

With the previous sections, the similarity matching and the time series decomposition were introduced. They result in two factors that can have a significant impact on the leading indicator search. The cross-correlation, which quantifies the similarity between two time series and the seasonal strength, which can be used to estimate the effect of seasonal patterns. To assess the impact of these factors, statistical analysis methods are required. For this reason, some of the basic characteristics of selected multivariate analysis methods are presented in the following sections. Two methods, namely *multiple linear regression* and *Analysis of variance (ANOVA)*, are considered parametric approaches and one method, namely *Align Rank Transform Contrasts (ART-C)*, is a nonparametric alternative. The variable set that will be used later on in these methods consists of one variable that will be regarded as the dependent variable (referred to as *response* variable) and multiple independent variables (referred to as *predictors* or *covariates*). More detailed information about the variable set will be provided in chapter 3. The parametric methods were selected due to the presumed linear relationship between the predictor variables and the response variable. All three multivariate analysis methods will be used to analyze the experimental results obtained throughout this work. It is important to mention that these methods are not explained in their entirety as the focus is primarily on the application of these methods and not on their derivations. The aim is to provide enough information to be able to interpret the analysis results. References to comprehensive information from the relevant literature are provided at selected explanations.

Before the analysis methods are introduced, however, the term *interaction effect* will first be defined. An interaction between two independent variables X_1 and X_2 is said to occur when the effect of variable X_1 on the response variable Y changes whenever there is change in the value of X_2. This represents the idea that the value of the dependent variable may relate in some nonadditive way to the values of both predictor variables (Navarro [72], Norman and Streiner [74, p. 91]). The three presented analysis methods can integrate and analyze interaction effects.

2.4.1 *Multiple linear regression*

In statistical learning, there is usually an assumption that there is some relationship between the quantitative response variable Y and p different predictor variables $X_1, X_2, \ldots, X_p$. This relationship can be written in the general form

$$Y = f(X) + \epsilon,$$

where f is some fixed but unknown function representing the systematic information that $X_1, \ldots, X_p$ provide about Y and ϵ is a random error term representing the stochastic component (James et al. [54, p. 16]). Multiple

linear regression assumes that this relationship is of linear form and that f can be approximately estimated by a linear function given by

$$Y = X^T\beta + \epsilon = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon, \tag{2.21}$$

where X_j represents the jth predictor, β_0 is the intercept and β_j with $j \in \{1, \ldots, p\}$ is the regression coefficient that quantifies the association between predictor variable X_j and the response variable. In this form, multiple linear regression is one of the most widely used statistical techniques in various fields of science (James et al. [54, p. 59 ff.], Czado and Schmidt [32, p. 191 ff.]).

To include potential interaction effects between predictor variables, equation 2.21 can easily be extended. Considering a standard multiple linear regression model with two predictor variables

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon,$$

where the association between X_1 and Y is a linear constant represented by the regression coefficient β_1. An interaction effect between X_1 and X_2 can now be included by introducing a third predictor, which is constructed by computing the product of X_1 and X_2:

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon \\ &= \beta_0 + (\beta_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + \epsilon \\ &= \beta_0 + \tilde{\beta}_1 X_1 + \beta_2 X_2 + \epsilon. \end{aligned}$$

The association between X_1 and Y is no longer constant since $\tilde{\beta}_1$ is now a function of X_2. This means that a change in the value of X_2 will change the association between X_1 and Y (James et al. [54, p. 88], Balli and Sørensen [10]).

In general, there are two major use cases of multiple linear regression: *Prediction* and *inference*. In a prediction setting, the goal is to estimate f that it yields in accurate predictions for Y, based on the observed values of the predictor variables. Within the inference analysis, the goal is to understand the association between Y and $X_1, \ldots, X_p$. The following questions may arise:

- Which predictors are associated with the response?
- What is the relationship between the response and each predictor?
- How strong is the relationship?
- Are there any interaction effects between the predictor variables?

It should be mentioned that these two uses of multiple linear regression are not mutually exclusive (James et al. [54, p. 17 ff.], Allison [4, p. 1 f.]). However, in the context of this work, the focus is primarily on the inference

analysis.

In order to answer the questions above, the regression coefficients β_j need to be analyzed. In practice, these coefficients are unknown and need to be estimated such that the linear model in 2.21 fits the available data well. The most common approach is the *ordinary least squares* (OLS) method, in which the estimates $\hat{\beta}_j$ are chosen that minimize the *Residual Sum of Squares* (RSS) (James et al. [54, p. 72 f.]).[8]

With the help of the estimated covariance matrix $\widehat{\sum}(\hat{\beta}) = \hat{\sigma}^2(\mathbf{X}^T\mathbf{X})^1$, where $\mathbf{X}$ is the design matrix and $\hat{\sigma}^2$ is the estimate of the residual variance, it is possible to perform hypothesis tests on the coefficients. In terms of inference analysis, particularly two tests, namely F-test and t-test, are of special importance for this work. The F-test tests the null hypothesis whether all the regression coefficients are zero (no relationship between response and predictor variables)

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

versus the alternative hypothesis that at least one coefficient is nonzero

$$H_1 : \exists j = 1, \ldots, p, \beta_j \neq 0.$$

The t-test tests the null hypothesis whether one selected regression coefficient is zero (no relationship between response and selected predictor variable)

$$H_0 : \beta_j = 0$$

versus the alternative hypothesis that the coefficient is nonzero (a relationship exists)

$$H_1 : \beta_j \neq 0.$$

Based on the resulting p-values, it can be determined whether to reject the null hypotheses (James et al. [54, p. 75 ff.], Czado and Schmidt [32, p. 213 ff.]).

However, these hypothesis tests are subject to a number of assumptions. Most of these assumptions relate to the error term ϵ in equation 2.21. The error terms are considered random variables $\epsilon_1, \ldots, \epsilon_n$ that represent all the unmeasured causes of the dependent variable Y. The assumptions, formally known as *Gauss-Markov assumptions*, are (James et al. [54, p. 66], Kutner et al. [58, p. 18]):

1. $\mathrm{E}(\epsilon_j) = 0 \quad \forall j,$

8 Detailed information about the OLS method can be found in Czado and Schmidt [32, p. 197 ff.].

2. $\text{Var}(\epsilon_j) = \sigma^2 \quad \forall j$,
3. $\text{Cov}(\epsilon_i, \epsilon_j) = 0 \quad \forall i \neq j$,
4. $\epsilon \sim \mathcal{N}(0, \sigma^2)$.

These assumptions imply that the error terms are mean independent, variance independent (also known as *homoscedasticity*), uncorrelated and follow a normal distribution.[9] Their distribution parameters do not depend on the realizations of X_i. The first assumption guarantees that the least squares estimates $\hat{\beta}_j$ are unbiased estimates. The second and third assumption guarantee that least squares coefficients are efficient. Their standard errors are at least as small as those produced by any other unbiased, linear estimation method. The normality assumption implies that the results of statistical tests including p-values will be correct. These assumptions need to be satisfied for the results and inferential conclusions to be strictly valid (Allison [4, p. 122 ff.]). In addition, multiple linear regression assumes that the causal mechanism with which the values of Y are generated is of linear form. However, if the true relationship is far from linear, then virtually all the conclusions and results from the regression are suspect. A useful graphical tool for identifying non-linearity is the residual plot, in which the residuals are plotted against the predicted values $\hat{y}_i$. If the relationship is linear, the residual plot will ideally show no discernible pattern. On the other hand, a pattern present in the plot may indicate a problem with some aspect of the linear model (James et al. [54, p. 93 f.]).

2.4.2 *Analysis of variance*

The following subsection presents ANOVA as an additional parametric method. The ANOVA is a procedure based on a factorial design that examines the effect of one (or more) independent variables (referred to as *factors* or *grouping variables*) on one dependent variable (analogously called response variable). It also assumes a linear relationship and a corresponding ANOVA model can be assigned to the class of linear models. The factors $X_1, X_2, \ldots, X_p$ need to be nominally scaled with k factor levels (referred to as *groups*). For the response variable Y, a metrical scaling is required. The different types of ANOVA can be differentiated according to the number of factors used. While in the setting of a one-way ANOVA, only one factor with k different groups will be analyzed, the two-way ANOVA examine two factors with k respectively j different groups. This can be continued to a multi-factor ANOVA with more than two grouping variables. The effects of individual factors are called *main effects*. This delimitation is necessary regarding potential interaction effects between factors that can also be examined with a multi-factor ANOVA. Furthermore, a distinction is made between a *balanced* and *unbalanced* design

9 A statistical test for testing the null hypothesis of homoscedasticity is the *Breusch-Pagan* test (Breusch and Pagan [21]). A well known statistical test for testing the null hypothesis of normality is the *Shapiro-Wilk normality* test (Shapiro and Wilk [83]).

(Backhaus et al. [9, p. 174 f.], Fromm [43, p. 13 f.]).[10]

The overall objective for all ANOVA types is to test differences in the response variable across different factors or different groups for significance. The question arises whether the mean response values of the individual factors or groups are equal or significantly different. Hence, ANOVA analyzes the differences between mean values. However, the variances of the observed values around these mean values also play a decisive role. The principle is heavily based on a decomposition of variance into a systematic part, which can be explained by the factors and interactions, and a stochastic part represented by a random error term ϵ, which cannot be explained (Herrmann and Seilheimer [46, p. 267], Navarro [71]). If the elements are squared and summed up over the observations (SS), the decomposition of variance can be written as

$$SS_{total} = SS_{between} + SS_{within}, \tag{2.22}$$

where $SS_{between}$ is the variance between the factor groups and interactions and SS_{within} is the remaining variance that is neither due to the factors nor to interaction effects, i.e., random effects on the response variable. Based on this decomposition, the following hypotheses can be formulated:

$H_0 : \mu_1 = \mu_2 = \mu_3 = \ldots$
versus
H_1 : not all mean values are equal and
at least one main effect or interaction effect is $\neq 0$.

These hypotheses can be tested using an F-test. The idea of the test is that the variances within the groups are small and the variances between the groups are large, when the groups differ. If the result turns out to be statistically significant (p-value < significance level α), further questions concerning the isolated analysis of individual factors or their interactions can be investigated. In these cases, the null hypothesis is that the analyzed factor has no effect or that there are no interactions present. Table 2.2 illustrates a sample two-way ANOVA table with two factors k and j and the interaction $k \times j$.

10 In a balanced design all sample sizes within the different groups are equal. Contrarily, in an unbalanced design the sample sizes are unequal. Unbalanced designs need to be treated with a lot more care since they have a major impact on how ANOVAs are performed and how they are interpreted. In fact, it turns out that there are three fundamentally different ways in how to run an ANOVA in an unbalanced design and they are not all equally appropriate to every situation. The three different ways are conventionally referred to as *Type I, Type II and Type III Sum of Squares (SS)*. All three types lead to different hypothesis testing strategies and thus to different SS values. Most statistical computer programs have these three types implemented. Type III is the most conservative type and is usually the default option since it does not give greater weight to groups with larger sample sizes (Bender and Lange [12]). The main differences between these three types are out of scope. Readers are referred to Navarro [73] and Norman and Streiner [74, p. 98 f.] for detailed information.

Source of variance	df	Sum of Squares	Mean Squares	F
Factor k	$K-1$	SS_k	$MS_k = \frac{SS_k}{K-1}$	$F_k = \frac{MS_k}{MS_{within}}$
Factor j	$J-1$	SS_j	$MS_j = \frac{SS_j}{J-1}$	$F_j = \frac{MS_j}{MS_{within}}$
Interaction $k \times j$	$(K-1)(J-1)$	$SS_{k\times j}$	$MS_{k\times j} = \frac{SS_{k\times j}}{(K-1)(J-1)}$	$F_{k\times j} = \frac{MS_{k\times j}}{MS_{within}}$
Error	$K*J(I-1)$	SS_{within}	$MS_{within} = \frac{SS_{within}}{K*J(I-1)}$	-
Total	$K*J*I-1$	SS_{total}	-	-

Table 2.2: Sample two-way ANOVA with interaction effect. This table shows the different sources of variance, their degrees of freedom, their SS and Mean Squares and the corresponding formulas for the F-statistic (Bender and Lange [12]).

A significant $k \times j$ interaction would indicate that the effect k has on response variable Y is significantly different for different groups of j, and vice-versa.

The underlying F-test is a so-called *omnibus test*. It tests whether there are differences between groups, but not whether all groups are different from each other. It is not possible to determine which groups of one or more factors exert a significant influence on the response variable and how large these effects are. Thus, if the F-test shows that a factor has a significant influence on the response variable, it cannot be concluded that all group means are different. To analyze such differences, so-called *post-hoc contrast tests* can be performed. Post-hoc contrast tests allow pairwise comparisons by testing all possible combinations of groups against each other. This way they will reveal which groups of each factor cause these influencing effects. They are only performed if the F-test of an ANOVA led to a significant result and the user subsequently wants to know (ex post; a posteriori) which factor groups account for differences in the means (Backhaus et al. [9, p. 196 f.], Bachman [8, p. 250 f.]).

However, since multiple comparisons with the same null hypothesis (no differences in group means) lead to multiple tests, there is an accumulation of the α-error. The α-error is the *Type I error* and represents the probability of rejecting the null hypothesis even though the null hypothesis is correct. The more statistical tests with a significance level of $\alpha = 5\%$ are run, the greater the probability of finding at least one test result that is significant by chance (problem of multiplicity). Therefore, the α-error must be corrected in such a way that the desired significance level is retained in case of multiple comparisons. A well known correction is the *Bonferroni correction*, in which α

is divided by the number of test repetitions (Norman and Streiner [74, p. 80 f.], Edgington [38, p. 80 ff.]). There are various post-hoc contrast tests available that differ in their assumptions and procedures. Werner [95, p. 322 ff.] and Norman and Streiner [74, p. 81 ff.] provide an overview of alternative post-hoc contrast tests and their main differences.

Analogously to multiple linear regression, the testing procedures of ANOVA rely on specific assumptions about the random error term ϵ. Here, the error terms are assumed to be normally i.i.d. with zero mean: $\epsilon \sim \mathcal{N}(0, \sigma^2)$. This can be assessed by looking at Quantil-Quantil (Q-Q) plots or running a Shapiro-Wilk normality test. However, the F-test is quite robust to violations of the normal distribution assumption. An additional assumption is that the variances caused by the error terms should be approximately equal across all groups (referred to as the *homogeneity of variance* or *homoscedasticity*) (Fromm [43, p. 13 f.], Navarro [70]).[11] Again, ANOVA is fairly robust to deviations from the homoscedasticity assumption, especially if there is a balanced design (Norman and Streiner [74, p. 80]). Nevertheless, these assumptions should be examined in any case to ensure the validity of the test results.

2.4.3 *Align Rank Transform Contrasts*

The multivariate analysis methods presented so far are of parametric nature. This means they use parametric tests such a t-tests or F-tests for testing hypotheses. These tests are based on certain assumptions, in particular regarding the distribution of some error terms. If these assumptions are violated (i.e., *nonconforming data*), the test results and the associated conclusions may lose validity. For this reason, nonparametric methods become very important in case of violations.

A nonparametric alternative for a multifactor ANOVA is called ART-C. ART-C is a new procedure for nonparametric multifactor analysis proposed by Elkin et al. [39], with which multifactor post-hoc contrast tests can be conducted. ANOVA assumes normal distribution of the error terms ϵ. This alternative circumvents this limitation by using an aligning and ranking procedure. ART-C is assigned to the Align Rank Transform (ART) paradigm. Its procedure is similar to the ART algorithm, with specifications for post-hoc contrasts tests. For this reason, ART will be roughly outlined in order to present ART-C.

ART is an algorithm for factorial data analysis that can handle nonconforming data in a factorial design. With ART one can examine interaction effects even if multiple factors are involved. The algorithm relies on a two-fold preprocessing step that (1) aligns the response variable Y for each effect (main or interaction) before (2) assigning ranks, averaged in the case of ties. While

11 One statistical test for comparing the variances of two or more groups is the *Levene's test*. This test tests the null hypothesis that all group variances are equal. The alternative hypothesis is that at least two of the compared groups differ (Fromm [43, p. 24 f.]).

aligning the response variable, effects are estimated as marginal means[12] and then stripped from the response variable so that all effects but one are removed. After this alignment and ranking step, common ANOVA procedures can be used, making the ART algorithm accessible to anyone familiar with the F-test (Wobbrock et al. [97]). Wobbrock et al. [97] provide a detailed description of the ART procedure for N factors. However, it was shown by Kay [56] that subsequent post-hoc contrast tests involving combinations of groups across multiple factors cannot be conducted on ART's aligned and ranked data without exploding Type I errors.

Against this background, ART-C can be seen as an extension of ART that enables correct post-hoc contrast tests. In contrast to ART, ART-C offers an alignment process specific to post-hoc contrast tests involving one or more factors. The response variable is first aligned and then ranked with ascending midranks. Within the alignment step, the response variable is aligned not for main effects and interactions, but for intended post-hoc contrast tests. Therefore, by using ART-C Y must be aligned and ranked for each set of factors whose groups will be compared. If there is an interaction effect of two factors k and j, then Y need to be aligned and ranked separately to compare combinations of groups of k and j. Elkin et al. [39] provide a comprehensive example with three factors in the presentation of their extension.

12 Considering two factors with multiple groups and a corresponding contingency table, then the marginal means of one factor are the means for that factor averaged across every group of the other factor.

3 EXPERIMENTAL DESIGN

This chapter provides a step-by-step description of the experimental pipeline and detailed information about the datasets used within the experiment. In addition, this chapter contains information about the applied software and hardware.

3.1 SETUP

The experiment was conducted on a Windows computer with an AMD Ryzen 5 3600 6-Core 3.59 GHz AM4 35MB Cache processor, 32 GB 3200 MHz DDR4 RAM and a NVIDIA GeForce GTX 1660 Super 6 GB GPU. The associated pipeline was written in *Python* and in *R*. Python is an interpreted, object-oriented, high-level programming language[1] and was primarily used to retrieve and preprocess time series data from the external online data source. R is a language and environment for statistical computing that provides a wide variety of graphical techniques.[2] It was mainly used to write the program with which the experiment was carried out. In addition, R was used to visualize and analyze the achieved results. Information on key Python libraries, R packages and corresponding functions are provided at selected steps throughout the experimental pipeline.

3.2 DATASETS

The next two sections present the two datasets on which the experiment is based. The description of the external dataset contains additional information about the applied methodology to create the dataset.

3.2.1 *Retail sales figures*

This dataset was provided by a German retail company which is one of the leading trading companies in Germany. The data is stored in several comma-separated values (*CSV*) files. Almost 800 thousand unique sales time series of different products are available. These products are assigned to roughly five thousand product groups, which in turn are assigned to four superordinate retail areas (Fast Moving Consumer Goods (FMCG), household, kitchenware, electronics). The time series data is spread over a total of 59 distinct stores. The dataset covers a time period of more than two years, starting in November 2018 with a daily sampling rate. The last record was noted in August 2021. In addition to the sales figures, there are also various meta information

1 https://www.python.org/doc/essays/blurb/
2 https://www.r-project.org/about.html

of the individual products, which enable a unique identification of the products. The most important meta information is the product description, which plays an important role especially for the second dataset. However, due to local computing capacities, a limitation of the amount of data must be made in order to process the data efficiently. For this purpose, after consultation with the company, only products of the superordinate retail area FMCG were considered in the experiment. FMCG are relatively low-priced products with a high turnover. They satisfy immediate wants and needs and are products of daily use. This includes food and beverages, cleaning products but also products for personal hygiene (Kaiser [55]). In addition, a single store was selected that had the most FMCG products sold over the period of two years. The forecast horizon of FMCG products is set to be three weeks ($h = 3$). Ultimately, there are 3335 unique products that lead to 3335 sales time series in total.

3.2.2 *Google Trends*

Given the research questions, the second dataset must be obtained from an external data source. Since external data sources often charge money for the use of their datasets, the search was limited to freely available data sources. Furthermore, it should be possible to query the data automatically via an appropriate Application Programming Interface (API).

For this reason, Google Trends was selected as the external data source. Google Trends is a free online service of the company *Google LLC*, which provides information about which and how often search terms (referred to as *keywords*) were entered by users of the Google search engine. Hence, Google Trends offers one of the largest real time datasets. In order to speed up the process time, Google Trends provides a sample of Google's search database each time the service is used. Its data is an unbiased, anonymized, categorized and aggregated representative of the Google search data. Google Trends makes it possible to measure the interest in a particular topic, from around the globe right down to city-level geography. More importantly, Google Trends offers a free data explorer[3] as a tool with which users can analyze the search interest in a keyword over time, where it is most searched, or what else people search for in connection with it. Once the keyword and the interested time period and region are entered, Google Trends returns the associated search interest over time as a univariate time series. This provides a unique perspective on what people search for, what they are currently interested in and curious about (Rogers [79]).

This analysis option can be particularly useful for FMCG products. When customers use the Google search engine to search for certain products, it may be an indication that they have an immediate need and may purchase these products in the near future. This is why the product descriptions from

3 https://trends.google.com

the meta information of the retail sales dataset are very valuable. They can be used as keywords that need to be passed to Google Trends. By doing so, a dataset with keyword related time series data can be generated. This idea is especially interesting for the goal of leading indicator search. However, the product descriptions from the retail sales dataset reveal some serious differences in terms of quality. For this reason, the descriptions must go through some preprocessing steps before they can be passed to Google Trends. The following preprocessing steps have proven to be useful:

- Removal of special characters (e.g., *?*, *!*, *&*, +, -, # etc.),
- Filtering of English and German stopwords[4] (e.g., *the*, *are* or *is*),
- Removal of quantity information (e.g., *500g* or *250ml*),
- Only keep words with length > 3,
- Apply lowercase notation.

Google Trends works best with concise and specific keywords. A product description, however, can be composed of a lot of product-specific information resulting in long descriptions. The possibility of not obtaining adequate time series data from Google Trends becomes more likely as the number of words increases. For this reason, *n-grams* are extracted for every description. A *n-gram* is a contiguous sequence of n items from a given sample of text or speech (Soffer [86]). In this work n is set to be 1. When $n = 1$, n-grams are called *unigrams*. If a product description consists of four different words after preprocessing, then four individual unigrams can be extracted. Table 3.1 shows three exemplary product descriptions before and after preprocessing and the resulting unigrams.

Product description before preprocessing	**Product description after preprocessing**	**Unigrams**
Knorr Rahmsauce Braten & Schmoren 250ml	knorr rahmsauce braten schmoren	[knorr, rahmsauce, braten, schmoren]
Bratkartoffeln 400g	bratkartoffeln	[bratkartoffeln]
Bio Lacroix Paste 40g, Gemüse	lacroix paste gemüse	[lacroix, paste, gemüse]

Table 3.1: Product descriptions before and after preprocessing and corresponding unigrams.

There are 3335 unique products with corresponding product descriptions in the retail sales dataset. After applying the preprocessing steps, 9274 unigrams in total could be extracted. These unigrams served as keywords to generate the Google Trends dataset. However, these keywords need to be passed individually to Google Trends and the time series data must be exported as a CSV file manually. Given the number of keywords, this causes a

4 Stopwords are high-frequency words that usually have little lexical content.

very time-consuming manual process. To overcome this problem, the Python library pytrends offers an API with which the Google Trends service can be utilized within a Python program. This API is not officially provided and maintained by Google Trends, but with its help thousands of keywords and keyword related time series can be processed and retrieved in no time. Further information on the use of pytrends will be provided in section 3.3.

3.3 PIPELINE

This section contains the step-by-step description of the experimental pipeline which is required to implement the methodology presented in section 1.3. The sequence of the individual steps is shown in figure 3.1. Each step will be described individually.

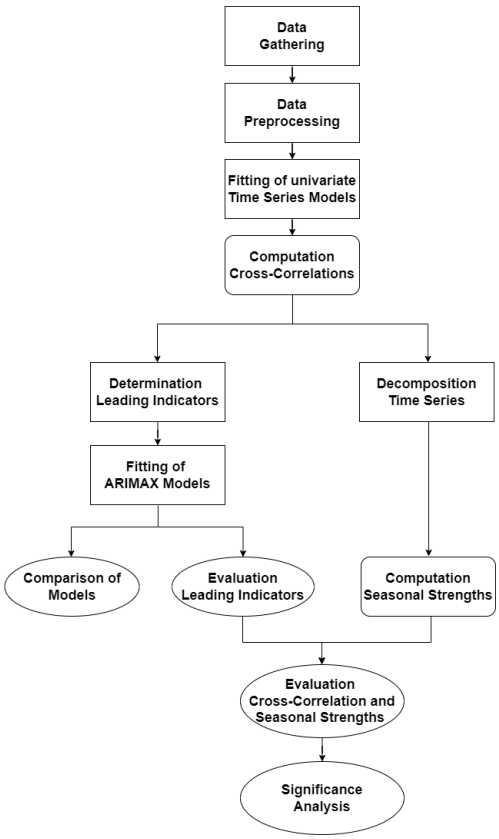


Figure 3.1: Pipeline of experiment.

DATA GATHERING In this step, the Google Trends dataset will be generated based on the extracted unigrams from the product descriptions. These unigrams serve as keywords in the further process. The intention is to create a keyword time series database. In order to retrieve the time series data for each keyword, the Python library pytrends[5] was utilized. This library offers an API with which thousands of keyword related time series can be requested

5 https://github.com/GeneralMills/pytrends

automatically within a Python program. For this purpose, the library offers some useful functions. Before requesting time series data from the Google Trends service, a connection to Google must be established. This connection can be built with the function *TrendReq*() from the `request` module. Several meta connection information such as the time zone offset (*tz*), the host language (*hl*) and the country abbreviation (*geo*) must be passed to the function in order to narrow down the geography. Since the descriptions (and ultimately unigrams) are of products that are sold in a German retail store, the parameters were set to the values $tz = 60$, $hl =$ "de" and $geo =$ "DE". After connecting to Google, a payload must be built to specify the relevant request information. Here, a list of keywords to get data for is passed to the function *build_payload*(). In addition, the timeframe of interest must be set. Since the first sales figures were noted in November 2018 and the last sales figures in August 2021, the timeframe was set to be five years. Once the payload is built, historical time series data of the searched keywords according to the specified timeframe can be retrieved with the function *interest_over_time*() (Aganjuomo [2]). This procedure was applied to all 9274 keywords. However, not every keyword leads to search queries in Google and thus to corresponding time series data. From 9274 unigrams in total only 1848 keywords have search queries in Google. Therefore, the Google Trends dataset eventually consists of 1848 keyword related time series.

DATA PREPROCESSING The time series returned by Google Trends have a weekly sampling rate and are normalized. Each data point is divided by the total searches of the location and timeframe specified in the request. Subsequently, these data points are scaled from 0 to 100 based on the keyword's proportion to all searches on all keywords, where 100 is the maximum search interest for the selected request parameters. Therefore, each keyword time series has a range of $[0, 100]$ (Rogers [79], Medeiros and Pires [66]). On the other hand, the retail sales time series have a daily sampling rate with an undefined range of values from theoretically $[0, +\infty]$. Hence, in order to match the two datasets in the similarity matching, the retail sales data must first be summed up on a weekly basis. Since not every product is sold on every day, days with no sales figures are filled with 0. The sum serves as a representative figure for the week. In total, each retail sales time series consists of 143 weeks and thus 143 data points. The Google Trends time series must also be limited to this period (week 47 of 2018 to week 32 of 2021). In addition, the time series of the two datasets have different scales. For this reason, *Min-Max normalization* is applied to all time series. This way, the minimum value of each time series gets transformed into 0, the maximum value gets transformed into 1, and every other value gets transformed into a decimal between 0 and 1. Unfortunately, the retail sales dataset also includes time series of products that were sold only irregularly (the proportion of the value 0 is high). To ensure that such time series do not distort the results of the experiment, they will be excluded from the dataset beforehand. Therefore, retail sales time series, in which the value 0 has a relative proportion of

$\geq$ 50% over the entire period and $\geq$ 25% in year 2021, will be eliminated. The remaining time series are considered efficient. Ultimately, there are 1633 retail sales time series for which results will be obtained in the experiment.

FITTING OF UNIVARIATE TIME SERIES MODELS In this step, univariate time series models will be fitted for each retail sales time series. The package that is mostly used in this step is the forecast package (Hyndman et al. [52], Hyndman and Khandakar [53]). This package provides functions and tools for displaying, analyzing and forecasting univariate time series. Furthermore, it offers functions for automatic ARIMA modeling. Before the time series models can be fitted, the time series must first be initialized as time series objects. This initialization can be done with the $ts()$ function of the core package stats. Here, it is important to determine the *frequency* which refers to the number of observations before a seasonal pattern repeats (Hyndman and Athanasopoulos [50]). After preprocessing, both the retail sales and keyword time series have weekly sampling rates. Therefore, the frequency is set to be 52. The univariate models that are fitted in this step are the three baseline models *Mean Forecast*, *Random Walk without drift* and *Seasonal Naïve* presented in subsection 2.1.2 and the ARIMA model as a linear time series model. The baseline models will be initialized with the functions $meanf()$, $rwf()$ respectively $snaive()$ (all with default settings). The ARIMA model will be initialized with the function $auto.arima()$. This model is different from the other models and must be treated separately. The function conducts a search over possible models by testing multiple order constraints and returns the best ARIMA model according to a model selection criterion (such as AIC_C). By doing so, this function uses a variation of the *Hyndman-Khandakar algorithm* (Hyndman and Athanasopoulos [50]). This algorithm covers the steps three to five of the ARIMA modeling procedure presented in section 2.1.3. Various arguments can be passed to the function allowing multiple variations of the algorithm. The most important argument is the *seasonal* parameter. If set to true, the search will be extended to seasonal SARIMA models. Against the background that retail sales time series often exhibit strong seasonal patterns, this option is of particular importance for the experiment. The default behavior of the algorithm can be found in appendix A.2. All four models are passed bundled with the $forecast()$ function to the $tsCV()$ function. This function then computes forecast errors using a rolling forecast origin until a forecast error is computed for every test observation. With the help of this function, the evaluation presented in subsection 2.1.4 can be performed. The forecast horizon is set to be three weeks ($h = 3$). This means that sales are predicted in three weeks time. The parameter *initial* is passed a split point which represents an initial period of the time series where no cross-validation is performed. In this step, this split point is set to be 80% of the time series length (114 weeks). The remaining 29 weeks constitute the test set. Eventually, $tsCV()$ returns a numerical time series object containing the forecast errors as a matrix where the time index corresponds to the last pe-

riod of the training set and the columns correspond to the forecast horizon (Hyndman and Khandakar [53]).

COMPUTATION CROSS-CORRELATIONS After the keyword time series database is created in step one, the process of the similarity matching begins. For each retail sales time series, the cross-correlation between every keyword time series is computed for a number of lags. The number of lags is set to be 4, which corresponds to four weeks. This limit goes in hand with the nature of FMCG products that have a high turnover. The assumption is made that if a customer has an urgent need for a product, the customer will presumably not search for the product on Google more than four weeks prior to the purchase. Ultimately, there are $12,071,136$ $(1633 * 1848 * 4)$ cross-correlation computations. However, these computations are only performed on the training set of each time series (first 80% respectively 114 weeks of time series values). Future values of the test set which may affect the estimates of the coefficients should not be included in the computation. The cross-correlations in this step are computed with the $ccf()$ function of the `stats` package. Here, only positive cross-correlations are considered, because it is assumed that higher values lead to higher similarity. For each keyword time series, the highest of the four cross-correlation values with its corresponding lag will be stored temporarily.

DETERMINATION LEADING INDICATORS The next step is to determine the leading indicator for each retail sales time series. For this purpose, the keywords will be ranked in descending order using the cross-correlation values stored in the previous step. The keyword that has the highest cross-correlation with the respective retail sales time series is set to be the leading indicator.

FITTING OF ARIMAX MODELS In the previous step, a leading indicator was determined for each retail sales time series. These leading indicators will now be incorporated as external regressors into the linear time series model, converting the univariate ARIMA models to bivariate ARIMAX models. For this operation, however, it is important to appropriately shift the two time series by the lag at which the highest cross-correlation was obtained. This pre-step is crucial in order to take full advantage of the leading effect the indicator may possess. Either the retail sales time series need to be shifted back in time or the leading indicator time series need to be shifted forth in time by the corresponding lag k (as indicated in section 2.3.2). Subsequently, the leading indicator can be included to the model via the parameter *xreg* of the function $auto.arima()$. If this parameter is set, a regression model with ARIMA errors is fitted as an ARIMAX model. In this form, the errors $\epsilon_1, \ldots, \epsilon_t$ from a regression (e.g., as shown in subsection 2.4.1) are allowed to contain autocorrelation. The model can be written as

$$Y_t = \beta_0 + \beta_1 X_{1,t} + \cdots + \beta_p X_{p,t} + \eta_t,$$

where η_t replaces ϵ_t and is assumed to follow an ARIMA model. This model has two error terms, the error from the regression model denoted by η_t and the error from the ARIMA model denoted by Z_t (which assumed to be white noise) (Hyndman and Athanasopoulos [50]). This representation of an ARIMAX model differs from the one presented in section 2.1.3 and is only intended to show how the concept of this model can be implemented with the *auto.arima*() function. An advantage of this approach is that, analogously to the other models, forecast errors can be calculated with the *tsCV*() function.

COMPARISON OF MODELS Up to this point, there are multiple forecast errors for each time series based on the initial split within the *tsCV*() function. For each observation of the test set exist a forecast and thus a forecast error. These forecast errors are available for all 1633 retail sales time series and for each of the five time series models. The next step is to compare the forecasting performance of the five models. However, the models cannot be assessed on the basis of pure forecast errors alone. For this reason, an error metric is required that enables an overall model comparison. In this work, the error metric Mean Absolute Error (MAE) will be selected. This decision is rather arbitrary at this point and other error metrics can be used for comparison as well. After the previous steps, n forecast errors are available for a single retail sales time series. These errors represent both positive and negative deviations. The MAE then computes the mean error of the n absolute deviations: $\frac{\sum_{i=1}^{n} |\varepsilon_i|}{n}$, where $|\varepsilon_i|$ is the absolute deviation between the forecast of observation i and its actual value. Performing this computation for every retail sales time series, there are ultimately 1633 MAEs for each time series model. It then can be analyzed for each time series which model performed best based on the achieved MAE. The focus in this step is on exploratory data analysis. With the help of the package ggplot2 (Wickham [96]) numerous plots will be created to analyze the performances.

EVALUATION LEADING INDICATORS This step constitutes on the core steps of the experiment. To answer the research questions whether and under which conditions the inclusion of an external regressor is beneficial, the performances of the ARIMA[6] and ARIMAX models will be compared with each other. First, it will be examined in how many cases there was an improvement ($MAE_{ARIMA} > MAE_{ARIMAX}$) and in how many cases there was a deterioration ($MAE_{ARIMA} \leq MAE_{ARIMAX}$). This comparison provides the basis for defining the dependent variable

$$Impact = MAE_{ARIMA} - MAE_{ARIMAX},$$

which will be examined in the multivariate analysis methods. If the value is > 0, the external regressor is said to have a *positive* impact on the ARIMA model. The inclusion has resulted in an improvement of forecast accuracy

6 Due to the seasonality of the retail sales time series these ARIMA models may also take the form of SARIMA models. For the comparison with ARIMAX models, both variants will be summarized under the term *ARIMA models*.

since the MAE of the ARIMAX model is lower than the MAE of the counterpart ARIMA model. Contrarily, if the value is ≤ 0, the external regressor is said to have a *negative* impact on the ARIMA model. These terms will be used numerous times throughout the analyzes. After the comparison is completed for all retail sales time series, individual time series will selectively be analyzed where improvements and deteriorations occurred.

DECOMPOSITION TIME SERIES As already indicated in the previous sections, especially retail sales time series can exhibit strong seasonal patterns, making an effective modeling of these series a challenging task. For this reason, it is assumed that the seasonality of a time series may affect the forecasting performance of ARIMA and ARIMAX models and thus the comparison of them. Therefore, it is crucial to apply methods that can capture seasonal patterns. Since it is not known beforehand how many seasonal patterns and periods the time series of the two datasets may have, the decomposition method MSTL is applied for all time series instead of the STL method. This method can be implemented with the function *mstl*() of the `forecast` package. Function *mstl*() allows multiple seasonal periods and decomposes a time series into seasonal, trend and remainder components, where seasonal components are estimated iteratively using STL (Hyndman and Khandakar [53]). The function will be called with its default parameter settings.

COMPUTATION SEASONAL STRENGTHS In order to quantify the potential influence of seasonality on the results of the experiment, an indicator is required that can measure the seasonality of a time series. The seasonal strength defined in subsection 2.2.3 can be used for this purpose. The formula in equation 2.2.3 will be applied to compute the seasonal strength of all time series. In contrast to the cross-correlation, the seasonal strength will be calculated based on the complete length of the time series. The time series are rather short with just 143 data points. This approach ensures that all possible seasonal periods will be considered.

EVALUATION CROSS-CORRELATION AND SEASONAL STRENGTHS In this step, qualitative statements will be made about the influence of the two variables *cross-correlation* and *seasonal strength* on the dependent variable *impact*. These statements are primarily based on exploratory data analysis.

SIGNIFICANCE ANALYSIS The results of the model comparison, the evaluation of leading indicators and the evaluation of the variables cross-correlation and seasonal strength constitute the results of the experiment. In the previous steps, only qualitative statements are made about these results. For this reason, the last step of the experimental pipeline is to evaluate these statements quantitatively using multivariate analysis methods. These consist of two parametric methods and one nonparametric method. They will also be used to investigate potential interaction effects between the variables. For both parametric methods, the underlying assumptions are examined with

various diagnostic plots and statistical tests. The functions and packages used for each method are provided at the corresponding evaluation (starting at subsection 4.4.1). All three analysis methods use a variable set composed of the following variables (including value ranges):

- Impact $[-\infty, +\infty]$,
- Cross-correlation $[0, 1]$,
- Seasonal strength retail sales time series $[0, 1]$,
- Seasonal strength leading indicator time series $[0, 1]$.

Both seasonal strength variables and the cross-correlation forms the set of independent variables. The variable impact is included in the modeling process as the dependent variable. However, before conducting the analysis methods, data points that are considered outliers according to the boxplot method are removed from the results dataset. This means that values of the dependent variable above Quantile $3 + 1.5 * \text{IQR}$[7] or below Quantile $1 - 1.5 *$ IQR will be considered outliers and removed.

7 Interquartile range

4 RESULTS

This chapter presents the results of the experiment. The first three sections build on exploratory data analysis and make qualitative statements about the findings. Section 4.1 provide the results of the model comparison. Subsequently, two time series are examined, where the inclusion of the leading indicator led to both, a noticeable improvement and a significant deterioration. Based on this, the influences of seasonal strength and cross-correlation on the response variable *impact* are analyzed in section 4.3. Within the first three sections, additional hypotheses are formulated that are examined within the multivariate analysis methods. Their results are evaluated in section 4.4. Last, the main findings and essential aspects to consider when conducting the experiment are discussed in section 4.5.

4.1 COMPARISON OF MODELS

Within the experiment, the $tsCV()$ function was applied to five different time series models with 1633 time series each to capture forecast errors on a rolling forecasting origin with a forecast horizon of $h = 3$. Based on these forecast errors, the error metric MAE was calculated. Ultimately, there are 1633 MAEs for each time series model. Since every MAE is computed using the same procedure, an overall model comparison is possible. This, in turn, makes it possible to determine which model has achieved the best forecast accuracy and thus the lowest MAE for each time series. In figure 4.1 a simple count statistic is illustrated of how often which model was chosen as the best time series model.

It turns out that the ARIMAX model is the best time series model for approximately a third of all time series. The ARIMA and Mean Forecast model almost share the same count statistic. The Random Walk model without drift was the best model for 264 time series. The lowest count statistic is achieved by the Seasonal Naive model with only 34 occurrences. Based on this simple count statistic, the ARIMAX model performs best in a relative comparison.

However, the distribution of the model related MAEs reveal that the first four models apparently have similar measures of central tendency and dispersion (see figure 4.2). The medians, represented by a line across the boxes within the boxplots, are nearly on the same level. It seems that the MAEs follow the same distribution with minor differences in skewness. Conversely, this means that the forecast accuracy of the four models does not differ greatly and that the choice of the best model is very close for the majority of time series. Here, the Seasonal Naive model is an exception. The estimated

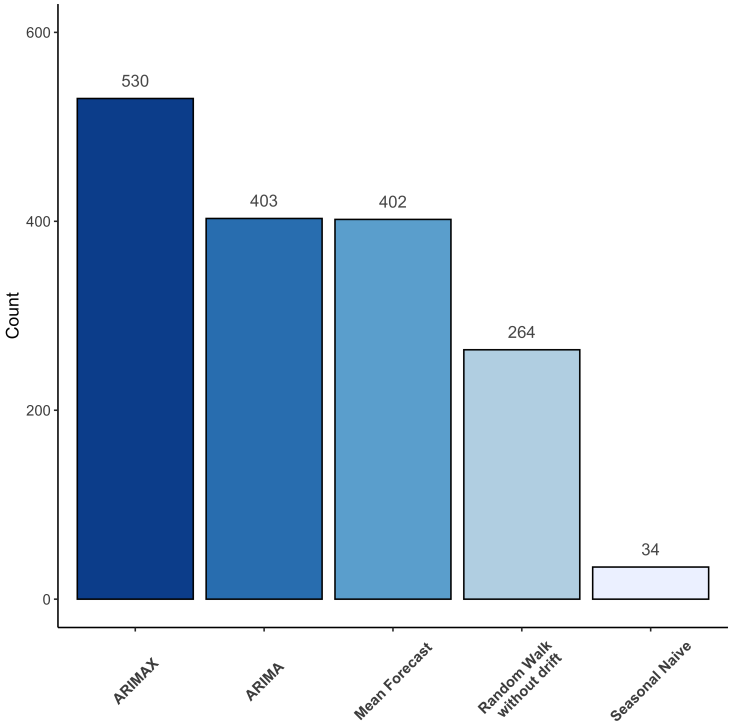


Figure 4.1: Count statistic of how often which model was considered the best time series model.

distribution indicates that the forecast accuracy is comparatively poor. The median of the MAE values is at 0.156 and therefore the highest among the models. This also explains why the Seasonal Naive model is the best time series model for only 34 time series. All models have in common that several data points are declared as outliers.

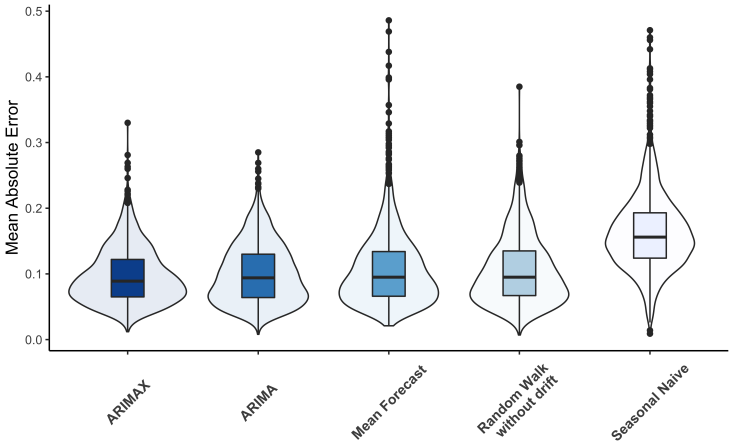


Figure 4.2: Boxplots and violin plots of MAEs for each model.

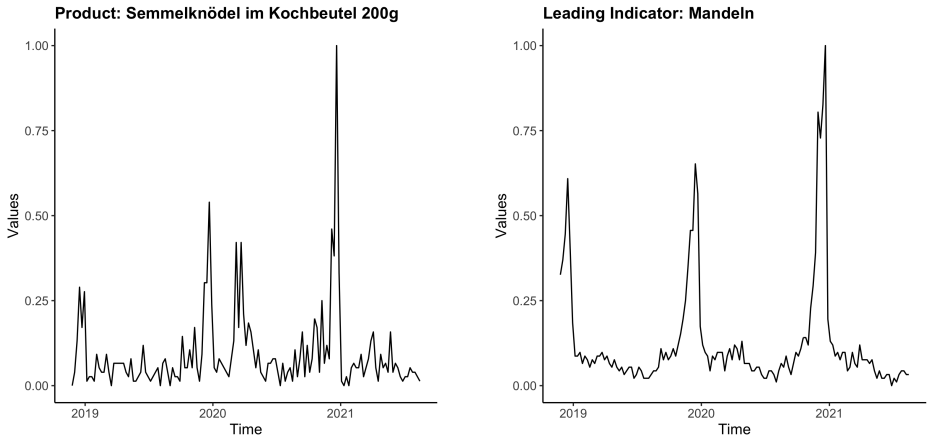
For the further analysis, it is particularly important to find out which factors are responsible for the performance of the ARIMAX models being better than the counterpart ARIMA models. The dependent variable *impact* is ulti-

mately defined from this comparison. These factors can provide information on when it is especially worthwhile to incorporate external regressors and when univariate ARIMA models are preferable. For this reason, the overall model comparison is broken down to these two time series models. It turns out that in total, there was an improvement ($MAE_{ARIMA} > MAE_{ARIMAX}$) in 51% and a deterioration ($MAE_{ARIMA} \leq MAE_{ARIMAX}$) in 49% of the time series. Restricting this comparison to time series where the two aforementioned models performed best, then there was an improvement in 57% and a deterioration in 43% of the cases.

4.2 UNIVARIATE RESULTS

An improvement in 51% and a deterioration in 49% of the cases resembles the result of a coin toss. There is no clear tendency that the inclusion of an external leading indicator has a positive effect on the forecast accuracy of ARIMA models. However, there were time series with a noticeable improvement but also cases with a significant deterioration. For this reason, several retail sales time series from both extremes were reviewed individually. The results of two randomly selected time series will be presented in this section.

For the first time series, the ARIMA model achieved an MAE of 0.087 and the ARIMAX model achieved an MAE of 0.028. There is a difference of 0.059 resulting in an improvement of roughly 68%. The retail sales time series and its corresponding leading indicator series are shown in figure 4.3.



(a) Retail sales time series.

(b) Leading indicator time series.

Figure 4.3: Retail sales time series and its corresponding leading indicator series. Improvement of roughly 68%. Highest cross-correlation is 0.71 at lag $k = -1$.

There seems to be no trend in both time series since there is no long-term increase or decrease in the data. However, both series reveal strong yearly seasonal patterns around Christmas time. These seasonal patterns are also considered in the univariate modeling process of both time series. For

the retail sales time series an ARIMA$(0,1,1) \times (0,1,0)_{52}$ model is fitted. An ARIMA$(0,0,3) \times (0,1,0)_{52}$ model is fitted for the leading indicator time series. Therefore, a seasonal differencing for the seasonal period $s = 52$ takes place with an order of $D = 1$. The associated ARIMAX model is implemented by a regression with ARIMA$(2,0,0)$ errors. Table 4.1 shows the *summary*()-output of the model.

Data: Retail sales time series with leading indicator.

Coefficients:

	AR1	AR2	Intercept	Leading Indicator
	0.0716	0.2011	0.0231	0.5426
Standard Error	0.0828	0.0861	0.0118	0.0524

Estimated σ^2: 0.00722

Table 4.1: *summary*()-output of regression with ARIMA$(2,0,0)$ errors.

The model residuals illustrated in figure 4.4 seem to follow a white noise process.

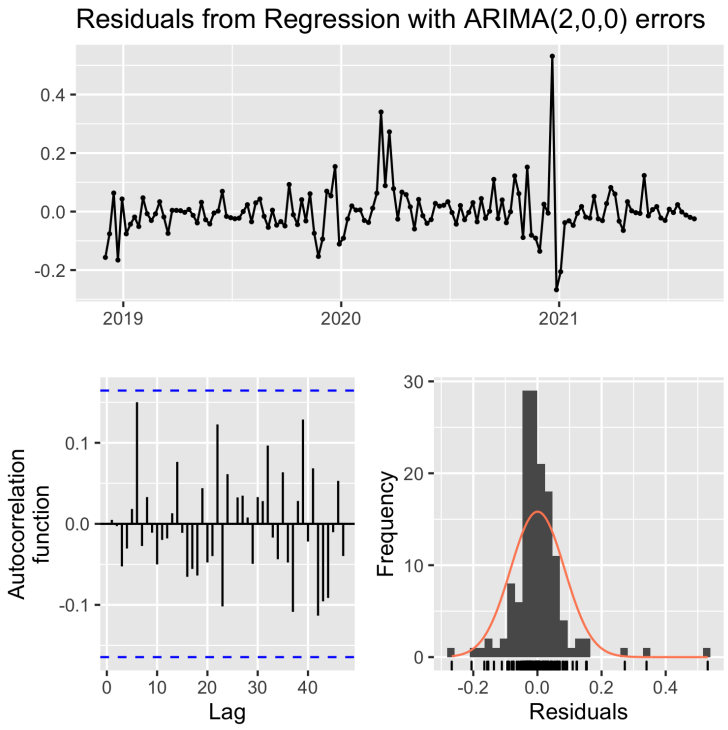


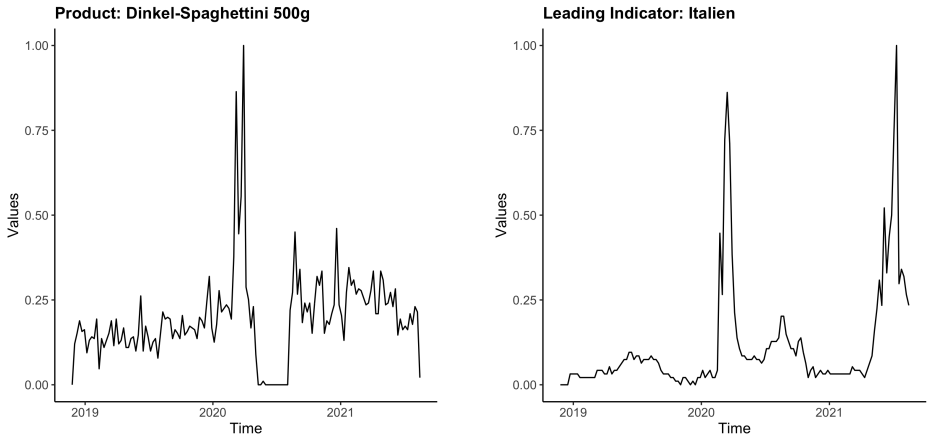
Figure 4.4: Various diagnostic plots of the residuals from a regression with ARIMA$(2,0,0)$ errors. The plots are generated with the *checkresiduals*() function of the forecast package.

This is confirmed by the high p-value (> 0.05) of the Box-Ljung portmanteau lack-of-fit (abbreviated with *Box-Ljung*) test in table 4.2. The residuals form a stationary series and resemble white noise. The diagnostic checks suggest that the fitted model is adequate.

Data	Test	p-value
Residuals from Regression with ARIMA$(2,0,0)$ errors	Box-Ljung test	0.953

Table 4.2: Result of Box-Ljung test with residuals from regression with ARIMA$(2,0,0)$ errors. The p-value indicates that residuals resemble white noise.

For the second time series, the ARIMA model achieved an MAE of 0.044 and the ARIMAX model achieved an MAE of 0.174. There is a difference of -0.13 resulting in a significant deterioration of nearly 300%. The retail sales time series and its corresponding leading indicator series are shown in figure 4.5.



(a) Retail sales time series.

(b) Leading indicator time series.

Figure 4.5: Retail sales time series and its corresponding leading indicator series. Deterioration of nearly 300%. Highest cross-correlation is 0.62 at lag $k = -2$.

For the retail sales time series, the mean level seems to increase slightly in time indicating a trend. The same does not apply to the leading indicator time series. Both time series apparently exhibit no strong seasonal patterns. The retail sales time series and leading indicator series have one peak at the beginning of 2020. In addition, the leading indicator reveals a peak in the middle of 2021. However, these peaks cannot be considered seasonality since there are no repeated seasonal factors. They may have been caused by random external factors. This can be confirmed by the fitted univariate models. For the retail sales time series an ARIMA$(0,1,1)$ model is fitted. On the other hand, an ARIMA$(0,1,0)$ model is fitted for the leading indicator time series. Both time series are not seasonal differenced and the univariate ARIMA models have no seasonal terms P and Q.

The associated ARIMAX model is implemented by a regression with ARIMA$(1,0,1)$ errors. Table 4.3 shows the *summary*()-output of the model.

Data: Retail sales time series with leading indicator.

Coefficients:

	AR1	MA1	Intercept	Leading Indicator
	0.8757	-0.4571	0.1463	0.3903
Standard Error	0.0577	0.0986	0.0347	0.0777

Estimated σ^2: 0.009046

Table 4.3: *summary*()-output of regression with ARIMA$(1,0,1)$ errors.

Despite the comparatively poor forecast accuracy, the residuals illustrated in figure 4.6 also seem to form a stationary series and resemble white noise.

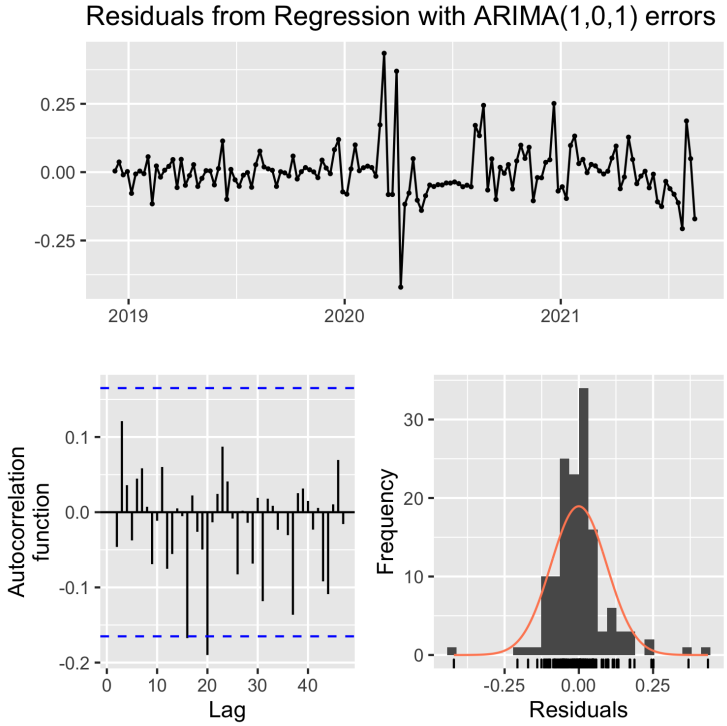


Figure 4.6: Various diagnostic plots of the residuals from a regression with ARIMA$(1,0,1)$ errors. The plots are generated with the *checkresiduals*() function of the forecast package.

Again, this is verified by the high p-value (> 0.05) of the Box-Ljung test in table 4.4. The ARIMAX model appears to adequately represent the data.

Data	Test	p-value
Residuals from Regression with ARIMA$(1,0,1)$ errors	Box-Ljung test	0.990

Table 4.4: Result of Box-Ljung test with residuals from regression with ARIMA$(1,0,1)$ errors. The p-value indicates that residuals resemble white noise.

The ARIMAX models of both examples pass the diagnostic checks and adequately represent the data. The main difference between these two examples is that in the example with improvement, both time series reveal strong seasonal patterns. In the example with deterioration, figure 4.5 and the univariate ARIMA models indicate that the retail sales time series and its leading indicator series show rather weak seasonality. This phenomenon could be observed for the majority of examined time series from the small sample of both extremes. For this reason, the seasonal strength as a measure of seasonality was determined for these examples. For this purpose, the definition from subsection 2.2.3 was applied. It turns out that in the example of improvement the retail sales time series has a seasonal strength of $F_S = 0.72$ and the leading indicator series of $F_S = 0.92$. Both values are close to 1 indicating strong seasonality within both series. Contrarily, the retail sales time series of the deterioration example has a seasonal strength of $F_S = 0.46$ and its leading indicator series of $F_S = 0.31$. Both values are rather close to 0 which supports the assumption of weak seasonality. This suggests that the improvement may be influenced by how strong the seasonality of the retail sales time series and its leading indicator series is. What stands out in these comparisons is that the cross-correlations in all examples are moderate (here 0.71 and 0.62) and do not differ greatly from each other. It may be possible that the cross-correlation exerts less influence on the outcome of the inclusion.

4.3 EVALUATION OF CROSS-CORRELATION AND SEASONAL STRENGTH

The results of the previous section suggest that there may be a relationship between the level of seasonal strength and the impact on ARIMA models. In addition, it was apparent from the sample that the cross-correlation values do not appear to differ greatly between the two extremes mentioned. For this reason, the analysis was extended to all 1633 time series. This section presents the exploratory results of the evaluation of cross-correlation and seasonal strength as independent variables.

The influence of the cross-correlation on the dependent variable *impact* is analyzed first. The distribution of the cross-correlation values is shown in figure 4.7. Moderate cross-correlations of around 0.6 are obtained for most of the time series. The average cross-correlation is 0.59.

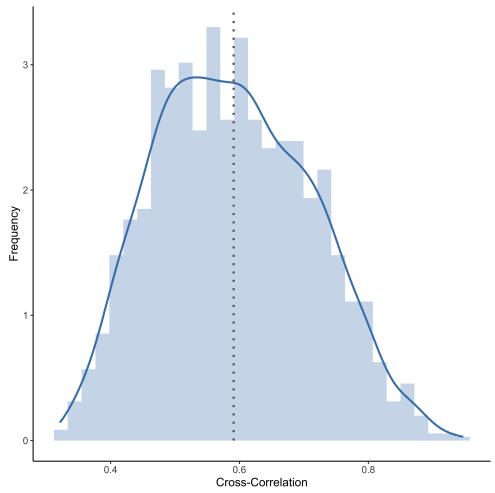


Figure 4.7: Distribution of cross-correlation. Average cross-correlation of 0.59.

To examine the influence of cross-correlation on the impact on ARIMA models, a different view on the distribution of the cross-correlation is illustrated in figure 4.8. Here, the distribution is shown discretized and separated by positive ($MAE_{ARIMA} > MAE_{ARIMAX}$) and negative ($MAE_{ARIMA} \leq MAE_{ARIMAX}$) impact with a simple count statistic.

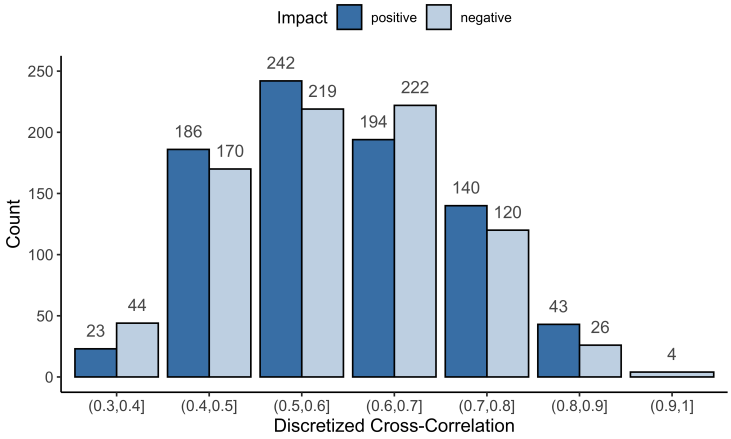
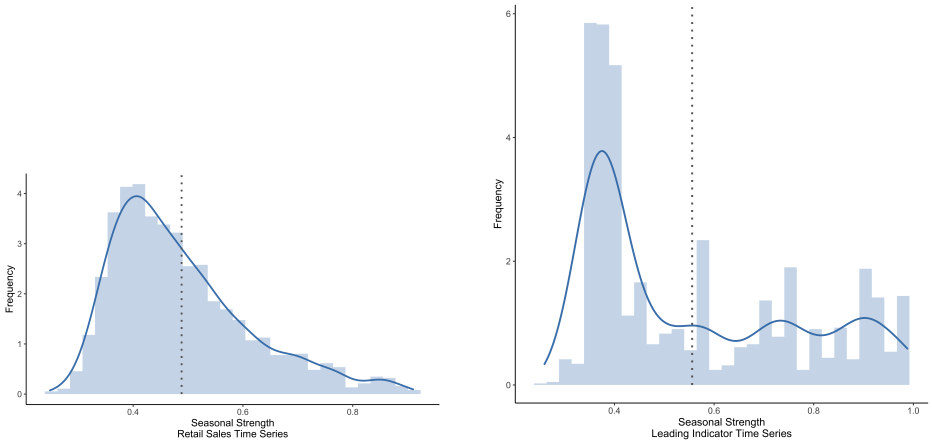


Figure 4.8: Discretized cross-correlation separated by impact on ARIMA model.

Most of the cross-correlation values are in a range around $(0.4, 0.8]$. Furthermore, it can be seen that there are no cross-correlation ranges where the proportion of the positive impact excels and vice versa. In both, high and low cross-correlation areas, are cases of improvement as well as deterioration. The distribution is balanced. Based on this distribution, it cannot be assumed that a higher cross-correlation, and thus a higher similarity, automatically leads to an improvement and to a positive impact on ARIMA models.

The analysis of seasonal strength, on the other hand, reveals a different picture. The distribution for the retail sales time series and the leading indi-

cator series is shown in figure 4.9. While the seasonal strength for the retail sales time series follows a right-skewed distribution, there is no clear pattern in the distribution for the leading indicator time series. The average seasonal strength is 0.49 respectively 0.56.



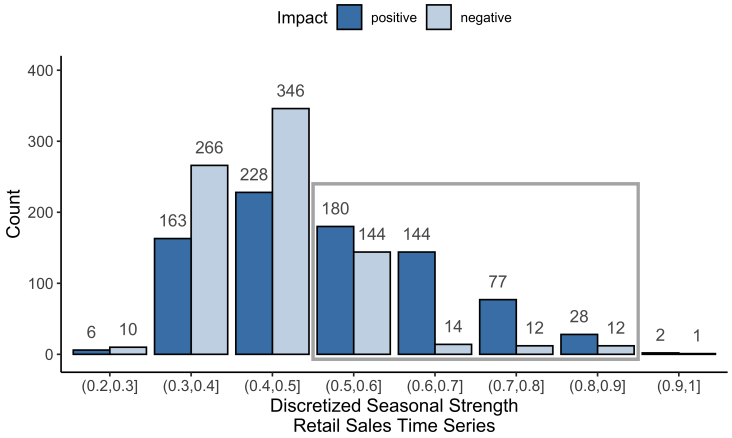
(a) Distribution of seasonal strength from retail sales time series. Average seasonal strength of 0.49.

(b) Distribution of seasonal strength from leading indicator time series. Average seasonal strength of 0.56.

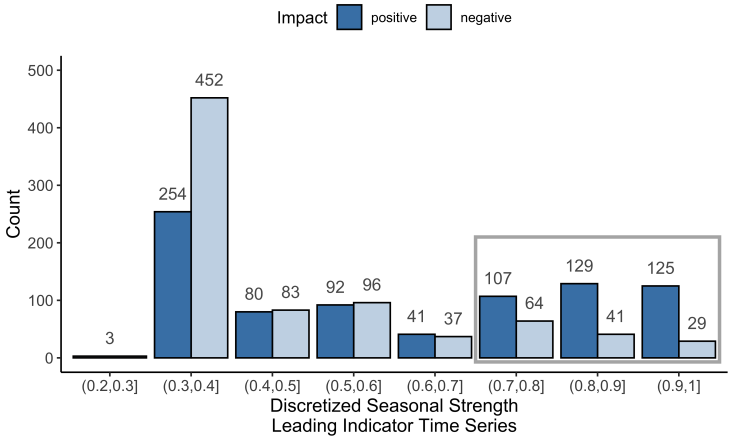
Figure 4.9: Distribution of seasonal strengths.

Figure 4.10 illustrates analogously the distribution discretized and separated by impact. Here, interesting differences can be observed compared to the cross-correlation. For the retail sales time series, the proportion of the negative impact is greater in the range of $(0.2, 0.5]$. From a range of 0.5, this ratio changes. In particular, the proportion of the positive impact is notably greater in the range of $(0.5, 0.8]$. A similar picture emerges for the leading indicator time series. Here, the proportion of the positive impact is greater from a seasonal strength of 0.7. This pattern suggests that the inclusion of an external regressor can especially be worthwhile as the seasonal strength for the retail sales and its leading indicator time series increases.

Last, the question arises how the other time series models perform as seasonal strength increases. For this reason, the count statistic of the best time series model is illustrated again, this time by the discretized seasonal strength of the retail sales time series (see figure 4.11a). It turns out that the ARIMA model has its highest absolute proportions in the range of $(0.3, 0.6]$. However, as seasonal strength increases, this proportion decreases in absolute and relative terms. On the other hand, the proportion of the ARIMAX model is becoming relatively large compared to the other models. The relative proportion of the Random Walk model without drift increases with seasonal strength. As expected, the proportions of the Seasonal Naive model are low.



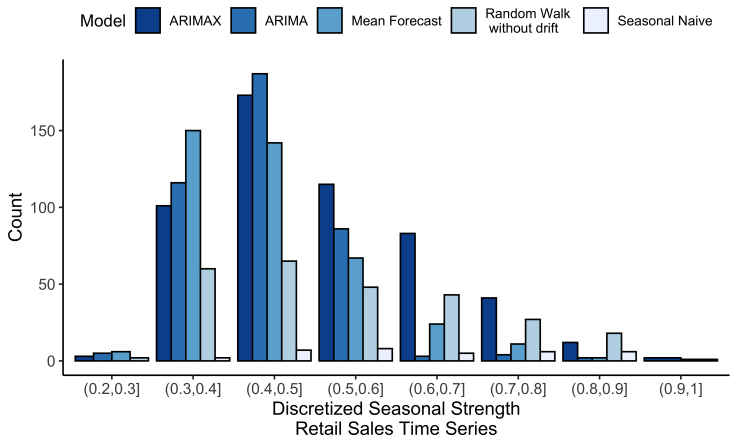
(a) Discretized seasonal strength from retail sales time series separated by impact on ARIMA model.



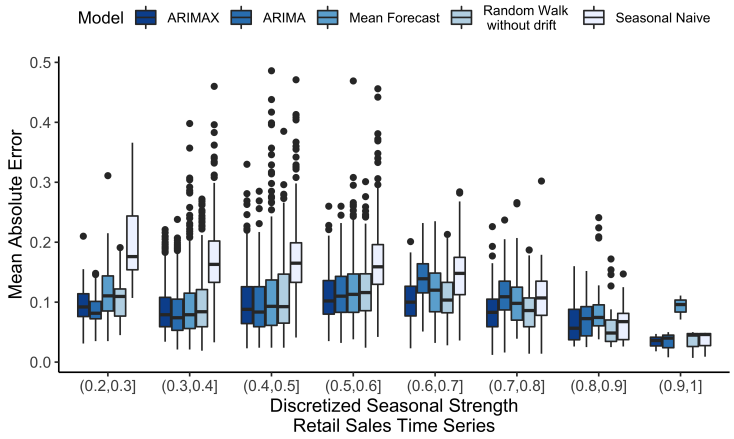
(b) Discretized seasonal strength from leading indicator time series separated by impact on ARIMA model.

Figure 4.10: Discretized seasonal strengths separated by impact on ARIMA model.

Again, these statements must be put into relation with the model-related distribution of the MAEs illustrated in figure 4.11b. Up to a seasonal strength of 0.6, the distributions of the first four models suggest similar performances. The choice of the best model is very close in these ranges of seasonal strength. However, a shift can be seen from a seasonal strength of > 0.6. Especially the boxplots for the ARIMA and ARIMAX model indicate that there are differences between these two models. In comparison, the forecast accuracy of the ARIMA model tends to deteriorate with increasing seasonality. What stands out is that the Random Walk model without drift performs comparatively well even at high seasonality ranges. In addition, as seasonal strength increases, the performance of the Seasonal Naive model improves. This may be where the ability of the Seasonal Naive model comes into play to model seasonal periods.



(a) Count statistic of best model by discretized seasonal strength of retail sales time series.



(b) Distribution of MAEs by model and discretized seasonal strength of retail sales time series.

Figure 4.11: Count statistic of best model and distribution of MAEs by model and discretized seasonal strength of retail sales time series.

Based on these figures, it can be assumed that ARIMA models are unable to model highly seasonal time series adequately. The superior performance of the ARIMAX models could be solely due to the fact that the external regressors only represent the unmodeled seasonality and do not add any value to the forecast accuracy.

This assumption is verified by examining the univariate fitted ARIMA models of the retail sales time series. If a time series is seasonal differenced or the seasonal terms P and Q of the model are $\neq 0$, then there is evidence that the seasonality is correctly captured. This test was performed for every retail sales time series. The result is shown in figure 4.12. At low to moderate seasonal strength values, it turns out that seasonality is not modeled. The result also demonstrates, however, that the seasonality of retails sales time series with high seasonal strength is correctly captured by the univariate

ARIMA models (or to be exact SARIMA models). Thus, the assumption that external regressors only cancel out the inability of ARIMA models to properly model seasonality and add no value to the forecast accuracy cannot be confirmed. The improvements by ARIMAX models can be attributed to the external regressors.

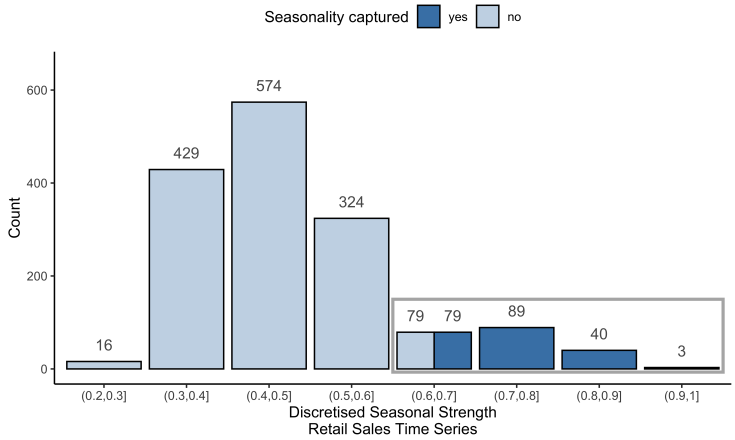


Figure 4.12: Test if seasonality is correctly captured by univariate ARIMA models.

Based on these exploratory results, the following two additional hypotheses are proposed:

1. There is no relationship between cross-correlation and the impact on ARIMA models. Improvements and deteriorations are both obtained equally at low and high cross-correlation values.

2. If both the retail sales and leading indicator time series exhibit strong seasonal patterns, then the probability of a positive impact increases. In such cases, the inclusion of the external regressor can be particularly beneficial.

4.4 SIGNIFICANCE ANALYSIS

In addition to the hypotheses formulated in section 1.2, two further hypotheses have been proposed based on the results obtained so far. However, these results are of qualitative nature and were obtained through exploratory data analysis. For this reason, multivariate analysis methods are required that can be used to quantitatively validate the results and to test the formulated hypotheses. The analysis methods utilized in this work are multiple linear regression, ANOVA and ART-C. All three methods offer hypothesis tests for this purpose.

The variable set consists of the dependent variable *Impact* and the three independent variables *Seasonal Strength Retail Sales Time Series*, *Seasonal Strength Leading Indicator Time Series* and *Cross-Correlation*. In all three analysis methods, a three-way interaction between the three independent variables is ini-

tially modeled by default. Based on the model results, it will be decided if and how to examine the interaction effect further. For the purpose of clarity and readability, the following abbreviations are introduced for the three independent variables:

- Cross-Correlation - *CC*,
- Seasonal Strength Retail Sales Time Series - *SSRS*,
- Seasonal Strength Leading Indicator Time Series - *SSLI*.

The distributions of the three independent variables have already been discussed in the previous sections. This is why the distribution of the dependent variable will be presented here (see figure 4.13). The average impact is 0.004, which is > 0. This means that on average, an improvement takes place.

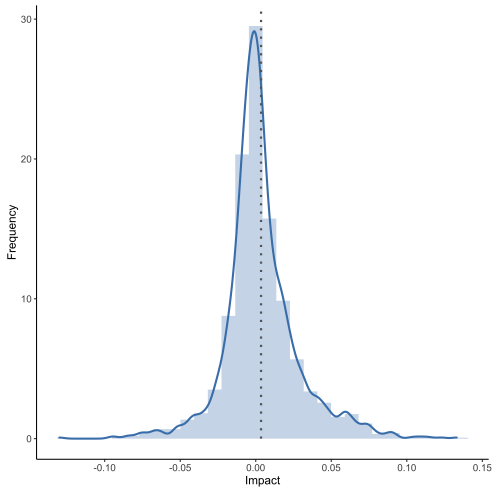


Figure 4.13: Distribution of dependent variable *Impact*. Average impact of 0.004.

Moreover, data points, that were considered outliers (in total 157) according to the boxplot method presented in section 3.3, are removed from the results dataset. Eventually, there are 1476 time series with experimental results.

4.4.1 *Evaluation of multiple linear regression*

This subsection presents the results of the multiple linear regression. While applying this method of analysis, the predictor variables have been centered (by subtracting the variable means) and scaled (by dividing the (centered) variables by their standard deviations) beforehand. Centering and scaling improve the interpretability of regression coefficients and main effects even when involved in interactions (Schielzeth [82], Gelman [44]). The multiple linear regression model was initialized and fitted with the function $lm()$ of

the stats package: $lm(Impact \sim SSRS * SSLI * CC)$. The following linear relationship was assumed:

$$Impact = \beta_0 + \beta_1 SSRS + \beta_2 SSLI + \beta_3 CC + \beta_4 SSRS * SSLI + \beta_5 SSRS * CC + \beta_6 SSLI * CC + \beta_7 SSRS * SSLI * CC.$$

With the intercept β_0, a total of eight regression coefficients must be estimated from the data. The regression coefficients $\beta_1, \ldots, \beta_3$ quantify the association between the predictor variables and the response variable and $\beta_4, \ldots, \beta_7$ represent the possible interaction terms. The F-test tests if there is any statistically significant relationship between response variable and predictor variables and the t-test tests each predictor and interaction individually. The results of the multiple linear regression model are shown in table 4.5.

Coefficients	Estimate	Std. Error	t-value	Pr(>|t|)	
(Intercept)	0.0013	0.0005	2.58	0.0101	*
SSRS	0.0038	0.0006	6.58	0.0000	***
SSLI	0.0032	0.0006	5.61	0.0000	***
CC	0.0004	0.0005	0.73	0.4657	
SSRS*SSLI	0.0014	0.0005	2.75	0.0061	**
SSRS*CC	-0.0011	0.0005	-2.09	0.0372	*
SSLI*CC	-0.0002	0.0006	-0.30	0.7605	
SSRS*SSLI*CC	-0.0010	0.0004	-2.76	0.0058	**
Residual standard error: 0.01418 on 1468 degrees of freedom.					
Multiple R^2: 0.1579, Adjusted R^2: 0.1539.					
F-statistic: 39.33 on 7 and 1468 degrees of freedom.					
p-value: $< 2.2e-16$.					

Note: $^{***}p < 0.001; {}^{**}p < 0.01; {}^{*}p < 0.05$

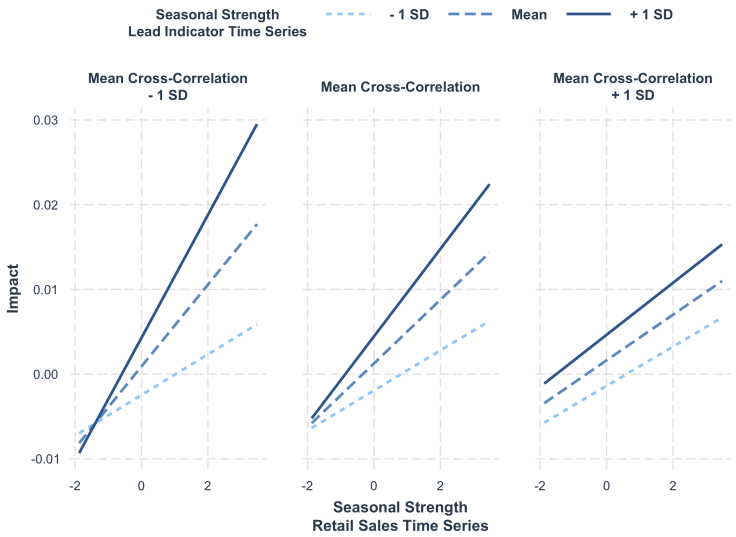
Table 4.5: $summary()$-output of multiple linear regression model.

The F-test has a p-value $\ll 0.05$ indicating a highly significant result. This means that there is a high probability that at least one variable or interaction exerts a significant influence on the dependent variable ($\beta_j \neq 0$). The p-values of the individual t-tests reveal which variables and/or interactions exert a significant influence. It turns out that the influence of the two seasonal strength variables are highly significant. Their two-way interaction is also statistically significant. On the other hand, the cross-correlation does not exert a statistically significant influence. Only within a two-way interaction with the seasonal strength of the retail sales time series and within a three-way interaction with both seasonal strength variables, the influence of the cross-correlation is significant (although the two-way interaction is only slightly significant).

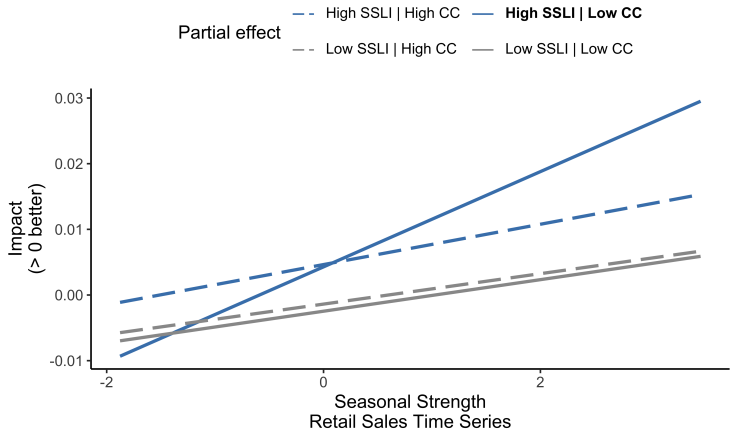
An interesting aspect of these results is that the interaction between all three predictor variables is significant. Thus, the regression coefficients $\hat{\beta}_1, \ldots, \hat{\beta}_6$ can only be interpreted to a limited extent. The association of the corresponding predictor variables and two-way interactions to the response variable is affected by the values of the three-way interaction. For this reason, the three-way interaction effect is examined primarily in the further course of this subsection. The underlying approach is adapted from Houslay [47] and Long [63]. Both refer to a method proposed by Cohen et al. [29] and popularized by Aiken and West [3]. Within this approach, two of the three predictor variables are set to be so-called *moderator* variables. Subsequently, the slopes of the response variable on the remaining predictor variable is computed while the moderator variables are held constant at different combinations of high and low values (Houslay [47]). High values and low values are defined as $+1$ *standard deviation (SD)* respectively -1 *standard deviation (SD)* from the moderator means (Long [63]). Here, the seasonal strength of the leading indicator time series and the cross-correlation are set to be the moderator variables. The slopes will be computed on the seasonal strength of the retail sales time series. This procedure can be performed automatically with the function *interact_plot()* of the interactions package (Long [62]) and its results are shown in figure 4.14.

If the lines run parallel at different combinations of high and low values of the moderator variables, this may be an indication that there is no interaction effect. Figure 4.14a reveals that, especially at the $mean - 1$ *SD* level of cross-correlation, there is an interaction effect that particularly affects high values of *SSLI*. When the seasonal strength variables increase at a low level of cross-correlation, higher values are obtained for the response variable. When the three levels of moderator variables are combined (figure 4.14b), the interaction effect becomes even clearer. The relationship between *SSRS* and the response variable changes noticeably when there are high values of *SSLI* in combination with low values of *CC*. This combination may occur if there are strong non-seasonal fluctuations between seasonal periods of seasonal time series. These intraseasonal fluctuations can have a mitigating impact on the corresponding cross-correlation. Two time series pairs with such a combination of values are shown in appendix A.1. In particular the seasonal structures of the retail sales time series are strongly obscured by the intraseasonal fluctuations. This is also reflected in the seasonal strength values $(0.54, 0.57)$. In these constellations, the leading indicators may provide information through their seasonal patterns ($F_s = (0.81, 0.89)$) that smooth out these influences.

In summary, both seasonal strength variables have a positive, statistically significant influence on the response variable. This influence is further amplified when the cross-correlation between the retail sales time series and its leading indicator series is low. If the cross-correlation is examined on its own, then it can be concluded that this variable does not exert a statistically sig-



(a) Three-way interaction effect at different combinations of high and low values of *CC* and *SSLI*.



(b) Summarized illustration of three-way interaction effect.

Figure 4.14: Three-way interaction effect of multiple linear regression model.

nificant influence. There is no relationship between the cross-correlation and the impact on ARIMA models. Only in interaction with the seasonal strength variables a significant influence does emerge for the cross-correlation variable. Accordingly, the hypotheses stated in section 4.3 can be confirmed to a large extent.

For the obtained results to be strictly valid, the assumptions of the multiple linear regression regarding the model residuals need to be met. They have to be variance independent (homoscedasticity) and normally distributed. In addition, the assumption of linearity should not be violated. However, as the following results show, these assumptions are partially violated.

The assumption of linearity can be tested by plotting the model residuals against the fitted values. This is illustrated in figure 4.15. The residual plot shows no discernible pattern. There appears to be a linear relationship in the data. This assumption is apparently not violated.

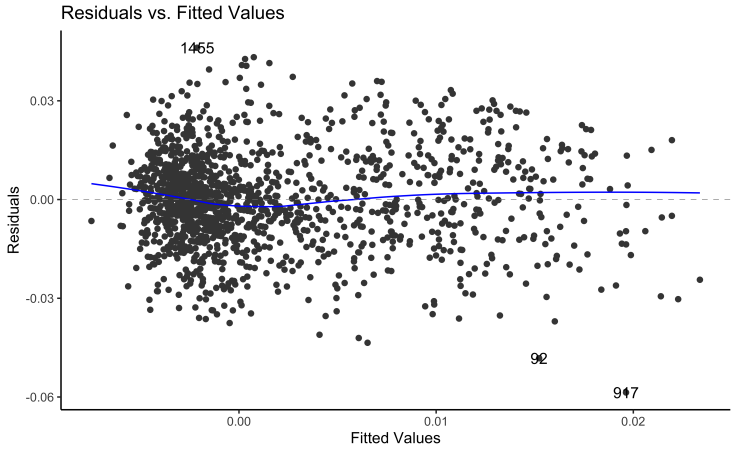


Figure 4.15: Residuals plotted against fitted values.

The normal distribution assumption can be examined analyzing a Q-Q plot. If the residuals are normally distributed, then they should form an approximately straight line. This is clearly not the case as illustrated in figure 4.16. In addition, the distribution has *heavy tails* indicating that the residuals have more extreme values than would be expected if they truly followed a normal distribution.

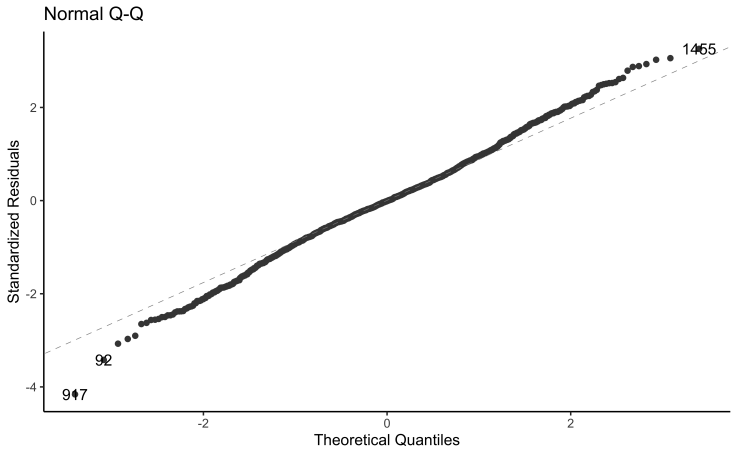


Figure 4.16: Normal Q-Q plot of residuals.

The p-value of the Shapiro-Wilk normality test (< 0.05) in table 4.6 confirms that the residuals are not normally distributed. Therefore, the normal distribution assumption can be considered violated.

Data	Test	p-value
Residuals from multiple linear regression model	Shapiro-Wilk normality test	0.001699

Table 4.6: Result of Shapiro-Wilk normality test with residuals from multiple linear regression model. The p-value < 0.05 indicates that the residuals are not normally distributed.

Moreover, the p-value of the studentized Breusch-Pagan test in table 4.7 implies that the model residuals are not variance independent. Therefore, two important assumptions of the Gauss-Markov assumptions are violated, which negatively affects the validity of the test results and the estimation and thus interpretability of the regression coefficients.

Data	Test	p-value
Residuals from multiple linear regression model	Studentized Breusch-Pagan test	1.162e-12

Table 4.7: Result of studentized Breusch-Pagan test with residuals from multiple linear regression model. The p-value $\ll 0.05$ indicates heteroscedasticity. The residuals are not variance independent.

Due to the partial violations of the assumptions, the multiple linear regression model may not be an appropriate model to represent the data. This is also partially confirmed by the low adjusted R^2 value of 0.1539. Most of the variance within the data is not explained by the model. Overall, the results obtained must therefore be handled with great caution.

4.4.2 *Evaluation of analysis of variance*

The multiple linear regression results revealed that the two seasonal strength variables and the three-way interaction with cross-correlation have a significant influence of on the response variable. However, the corresponding assumptions were partially violated, negatively effecting the validity of the results. Therefore, the purpose of this subsection is to verify whether the results of the ANOVA can confirm the results already obtained from a different statistical point of view.

ANOVA tests if the mean response values of the individual factors or groups are equal or significantly different. The analysis is based on factors that need to be nominally scaled with k groups. Since the predictor variables in the variable set are all metrically scaled, they must be nominalized beforehand. Here, a dichotomization strategy was selected, using the median in combination with the mean values of the predictors as cutpoints. The cutpoint for the cross-correlation was set at 0.6 and for the seasonal strength variables at 0.5. Values exceeding these cutpoints are considered *high*. On the other hand, values below these cutpoints are considered *low*. This way the analysis problem

was artificially transformed in order to conduct an ANOVA with a factorial design. Eventually, there is one response variable and three independent factors with $k = 2$ groups (high and low). This leads to a $2 \times 2 \times 2$ design with eight distinct group combinations. Table 4.8 provides some descriptive statistics for each group combination.

CC	SSRS	SSLI	Response	N	Mean	Standard deviation
low	low	low	impact	306	-0.00	0.01
high	low	low	impact	391	-0.00	0.01
low	low	high	impact	199	0.00	0.01
high	low	high	impact	72	-0.00	0.01
low	high	low	impact	87	0.00	0.02
high	high	low	impact	35	-0.00	0.02
low	high	high	impact	212	0.01	0.02
high	high	high	impact	174	0.01	0.01

Table 4.8: Descriptive statistics for three-way ANOVA with $2 \times 2 \times 2$ design.

Column N in table 4.8 shows the number of observations in each group combination. It turns out that the ANOVA needs to be performed with an unbalanced design. The sample sizes are all unequal. The group combinations, in which the seasonal strength variables share the same level (low & low, high & high), have the most observations. The linear model of the ANOVA is initialized and fitted analogously to multiple linear regression with the function $lm()$. The actual ANOVA is conducted with the $Anova()$ function of the car package. This function can be used to calculate *Type II* and *Type III* ANOVA tables for model objects produced by the $lm()$ function (Fox and Weisberg [41]). Here, *Type III* was selected since it does not give greater weight to group combinations with larger sample sizes. The results of the three-way ANOVA are shown in table 4.9.

Both seasonal strength factors exert a statistically significant influence on the response variable. This means that for these factors the mean response values of the two groups are significantly different. Also in the context of ANOVA, the cross-correlation is not significant (not even in interaction with the other factors). Therefore, it can be considered to completely exclude this main effect in the modeling process. Compared to the multiple linear regression, the three-way interaction is not statistically significant. It seems that the dichotomization eliminated the minimal effect the cross-correlation had (although the p-value is close to the 5% significance level). However, the two-way interaction between the seasonal strength main effects $SSRS * SSLI$ is again significant. This implies that the effect *SSRS* has on the response variable is significantly different for the two different groups of *SSLI*, and vice-versa.

Coefficients	Sum of Squares	Degrees of freedom	F-value	Pr(>F)	
(Intercept)	0.0034	1	16.64	0.0000	***
SSRS	0.0017	1	8.44	0.0037	**
SSLI	0.0016	1	7.82	0.0052	**
CC	0.0001	1	0.56	0.4556	
SSRS*SSLI	0.0013	1	6.18	0.0130	*
SSRS*CC	0.0005	1	2.27	0.1322	
SSLI*CC	0.0003	1	1.45	0.2293	
SSRS*SSLI*CC	0.0006	1	3.11	0.0778	
Residuals	0.3000	1468			

Note: $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

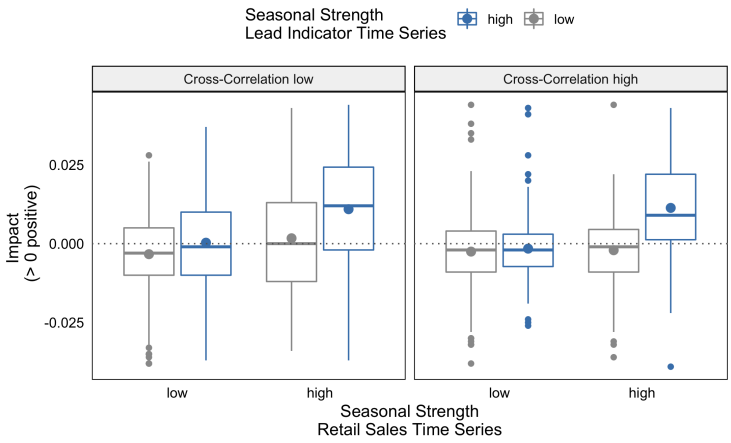
Table 4.9: ANOVA table (Type III tests).

The probability table 4.10 already shows in which direction this interaction effect goes and between which groups the significant differences exist.

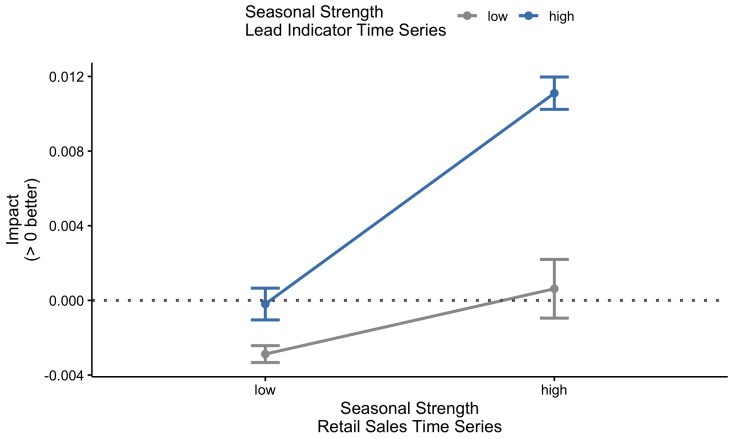
			Impact	positive	negative
SSRS	SSLI	CC			
low	low	low		**0.39**	**0.61**
		high		**0.39**	**0.61**
	high	low		0.48	0.52
		high		0.38	0.62
high	low	low		0.48	0.52
		high		0.40	0.60
	high	low		**0.68**	**0.32**
		high		**0.77**	**0.23**

Table 4.10: Conditional observed probabilities of dichotomized response variable given three independent factors.

The probabilities indicate that it is more likely to achieve a positive impact when the seasonal strengths of the retail sales and leading indicator time series are high. In contrast, when seasonal strengths are low, it is more likely to have a negative impact. If both time series have low seasonal strength, the inclusion of the leading indicator is probably counterproductive. The ratio between positive and negative impact is almost identical for the two cross-correlation groups in these seasonal strength constellations (low & low versus high & high). These relationships can also be seen in the corresponding interaction plots in figure 4.17.



(a) Three-way interaction.



(b) Two-way interaction.

Figure 4.17: Exploratory analysis of interaction effects.

The mean response values of the low and high groups of the seasonal strength factors suggest that there are significant differences and a significant interaction. Higher mean response values (> 0), and thus a positive impact, are achieved when the seasonal strengths are high, regardless of the cross-correlation group (see figure 4.17a). The two-way interaction plot in figure 4.17b supports this assumption. The difference between the two mean values of the groups low and high of the factor *SSLI* is greater in group *high* than in group *low* of factor *SSRS*.

Since the two-way interaction effect is statistically significant, the main effects *SSLI* and *SSRS* are being compromised by the interaction. For this reason, post-hoc contrast tests are conducted on the two-way interaction effect only. Multiple pairwise comparisons were made to determine which factor groups account for differences in the means. All separate group combinations have been tested against each other. However, since the ANOVA is conducted on an unbalanced design, the *estimated marginal means* were compared rather than the observed mean values. These marginal means are

based on the underlying statistical model, not on the observed data. They represent what the mean values would have been, had there been a balanced design with equal sample sizes. With the function *emmeans*() from the emmeans package, the estimated marginal means for specified factors or factor combinations in a linear model can be computed and pairwise comparisons can be carried out automatically (Lenth [61]). The results of the post-hoc contrast tests are shown in table 4.11.

SSRS	SSLI	Estimated marginal mean	Standard error	Degrees of freedom	Grouping
low	low	-0.0029	0.0005	1468	a
low	high	-0.0006	0.0010	1468	a
high	low	-0.0002	0.0014	1468	a
high	**high**	**0.0111**	**0.0007**	**1468**	**b**

Results are averaged over the levels of: Cross-correlation.
P-value adjustment: Bonferroni correction.
Significance level used: $\alpha = 0.05$.

Table 4.11: Results of post-hoc contrast tests on estimated marginal means. Column *Grouping* indicates which means and thus which combinations are similar and which are significantly different.

It turns out that the group combination (high, high) is significantly different from all other combinations. This combination has the only positive mean value of 0.0111. This result confirms what the interaction plots have already illustrated. If both time series have a high seasonal strength, the probability of achieving an improved forecast accuracy with the inclusion of the leading indicator ($MAE_{ARIMA} > MAE_{ARIMAX}$) increases. This finding is almost congruent with the results of the multiple linear regression, except that with ANOVA the cross-correlation loses further statistical significance.

Analogously to multiple linear regression, certain assumptions regarding the ANOVA residuals need to be met. They are assumed to be normally i.i.d. with zero mean. Furthermore, the variances caused by the residuals should be approximately equal across all groups (homogeneity of variance). Again, the residuals should form an approximately straight line, if they are normally distributed. The Q-Q plot in figure 4.18 shows that the residuals do not form a straight line. They do not seem to follow a normal distribution and normality cannot be assumed. This conclusion is supported by the Shapiro-Wilk normality test. The *p*-value is significant (0.00158). Therefore, the normal distribution assumption can be considered violated.

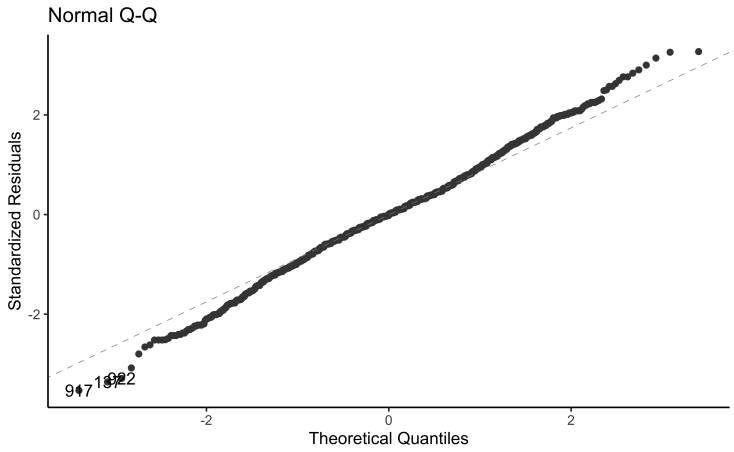


Figure 4.18: Normal Q-Q plot of residuals.

To examine the homogeneity of variance, the residuals can be plotted against the fitted values, which is illustrated in figure 4.19. There seems to be no evident relationships between residuals and fitted values (the mean of each group). However, there are differences in variance noticeable. The Levene's test for homogeneity of variance confirms (p-value $\ll 0.05$) that there are at least two statistically significant different groups in terms of residual variance. Therefore, the assumption of homogeneity of variances can also be considered violated.

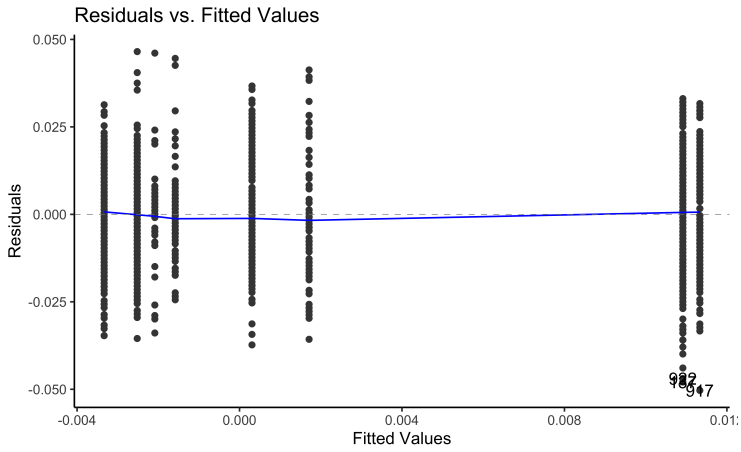


Figure 4.19: Residuals plotted against fitted values.

Ultimately, the two fundamental assumptions of ANOVA are violated. Although ANOVA is said to be robust to violations, they still have a negative impact on the validity of the results and conclusions. However, the exploratory analyses provide first evidence that the seasonal strength plays an important role in answering the research questions of this work.

4.4.3 *Evaluation of Align Rank Transform Contrasts*

In both parametric analysis methods, the key assumptions were violated. The underlying data is *nonconforming* and does not hold the assumptions. All the conclusions and results from these methods can therefore be considered suspect. For this reason, the nonparametric multifactor analysis method ART-C was selected as an alternative, especially for ANOVA, to evaluate the results of the experiment. ART-C circumvents the normal distribution assumption by using an aligning and ranking procedure, in which the response variable is first aligned for intended post-hoc contrast tests and then ranked with ascending midranks. After this step, common ANOVA procedures can be applied to the aligned and ranked data. The ART-C method can be initialized with the $art()$ function of the ARTool package (Kay et al. [57]): $art(Impact \sim SSRS * SSLI * CC)$. Subsequently, a common three-way ANOVA (Type III) can be applied on the returned object with the $anova()$ function of the stats package. The results of the three-way ANOVA on the aligned and ranked data are shown in table 4.12.

Term	Degrees of freedom	Sum of Squares	F-value	Pr(>F)	
SSRS	1	18465073.94	112.64	0.0000	***
SSLI	1	13306694.06	79.41	0.0000	***
CC	1	82148.09	0.45	0.5019	
SSRS*SSLI	1	4430106.24	25.54	0.0000	***
SSRS*CC	1	128632.69	0.71	0.4009	
SSLI*CC	1	55452.80	0.30	0.5809	
SSRS*SSLI*CC	1	445089.61	2.46	0.1170	

Note: $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 4.12: ANOVA table (Type III tests) for ART-C method.

In terms of significance, the results are identical to those of the common ANOVA presented in table 4.9. Even after the data has been aligned and ranked, the significant influences of the seasonal strength factors and their two-way interaction remain the same (as the statements about the individual main and interaction effects). The result of the post-hoc contrast test for the significant two-way interaction is shown in table 4.13.

SSRS	SSLI	Value	F-value	Pr(>F)	
low-high	low-high	288.73	25.54	0.0000	***

Note: $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 4.13: Results of post-hoc contrast test on two-way interaction effect.

The difference *low-high low-high* can be interpreted as the difference between $(low - high|low)$ and $(low - high|high)$, which is estimated as 288.73.

The interpretation of this value is hardly intuitive, but the comparison states that there is a significant positive difference between *SSRS low* and *SSRS high* in group *SSLI low* compared to *SSRS low* and *SSRS high* in group *SSLI high*. This indicates that higher values of the response variable are achieved when the seasonal strengths of the two time series are high. This supports and emphasises the results of the post-hoc contrast tests from the two-way ANOVA interaction effect in table 4.11.

4.5 DISCUSSION

The results presented in the previous sections illustrate that four out of the five tested time series models perform at a comparable level. The forecast accuracies of the Random Walk without drift, Mean Forecast, ARIMA and ARIMAX model are close to each other and the MAEs seem to follow a similar distribution. The baseline models, that follow rather simple rule-based approaches, can achieve comparatively good results. An exception is the Seasonal Naive model. One of the key objectives of this work is to identify factors that reveal when the inclusion of an external regressor may be worthwhile and when it is not. For this reason, the comparison of the ARIMA and ARIMAX models is particularly important. Restricting the analysis to this comparison, there is an improvement in 51% and a deterioration in 49% of the time series. The examination of individual examples with noticeable improvements but also with significant deteriorations have shown that the impact on the ARIMA model possibly depends on the seasonal strengths of the retail sales and its corresponding leading indicator time series. Moreover, the cross-correlation achieved in the similarity matching seems to play a subordinate role. This analysis was extended to all time series. It was found that positive and negative results were obtained in all cross-correlation ranges. There are no ranges where the proportion of the positive impact excels and vice versa. The ratio between positive and negative impact is balanced. For the seasonal strength variable, however, the proportion of the positive impact is notably greater in higher ranges. Based on these results, two additional hypotheses were formulated, which were examined and tested within the framework of multiple analysis methods (two parametric and one nonparametric). All three analysis methods conclude that when the seasonal strength of the retail sales time series and its leading indicator series is high, the probability of achieving an improving forecast accuracy increases. On the other hand, the cross-correlation exerts no statistically significant influence as an independent variable. The violations of the model assumptions of the parametric methods can be compensated by the explorative findings and the results of the nonparametric analysis method ART-C. Ultimately, the hypotheses stated in section 4.3 can be confirmed. There is no relationship between cross-correlation and the impact on ARIMA models. A higher similarity does not guarantee a higher forecast accuracy with ARIMAX models. Contrarily, the interaction between the seasonal strength variables is statistically significant in a way that higher values lead to positive effects. One possible

interpretation is that the historical variation of highly seasonal retail sales time series can be explained with information provided by the strong seasonal components of their leading indicators. The short-term development of these leading indicators may anticipate upcoming seasonal patterns of the retail sales time series. As the results show, these information can be particularly valuable for forecast accuracy.

However, there are four aspects that need to be considered while conducting the experiment. All four aspects affect important steps in the experimental pipeline, especially the steps data gathering, similarity matching, leading indicator determination and analysis with ANOVA. The four aspects that will be discussed in this section are:

1. Challenges utilizing Google Trends,
2. Spurious correlations,
3. Ranking and selection,
4. Dichotomization of continuous variables.

1. CHALLENGES UTILIZING GOOGLE TRENDS Google Trends offers access to the relative popularity of actual search requests made to Google and provides the frequency in which a particular term is searched for from various regions around the globe down to city-level geography. However, Google Trends only provides access to a small sample of Google's search database since the entire dataset would be too large to process efficiently (Google process billions of search requests per day). From a performance point of view, this sampling strategy may make sense, but it comes with a major challenge. The small sample provided is not always the same and in fact, it is constantly changing. The time series data requested for a keyword today may already look completely different tomorrow even if the same filters as time and location are applied (Medeiros and Pires [66], Cebrián and Domenech [23]). In addition, Medeiros and Pires [66] have shown that the differences between various samples of the same keyword will be higher the less often the keyword is searched. For the keywords requested and processed in this work, it was not tested whether they were high volume search terms. Moreover, only one sample per keyword was considered in the experiment. This in in turn could lead to the challenge that other keyword samples could have led to different leading indicator time series and thus to different conclusions. Medeiros and Pires [66] offer one solution to overcome this possible problem. Instead of processing one sample per keyword, many different samples should be gathered over a longer period. Subsequently, the time series data could be averaged across multiple samples in order to get a more reliable time series of that keyword. This approach should be considered in further experiments in order to substantiate the findings.

2. SPURIOUS CORRELATIONS The similarity matching between the two datasets within the experiment was performed using the cross-correlation

function. The cross-correlation is a natural metric to evaluate and quantify the strength and direction of the lead–lag relationship between two time series X_t and Y_t. Y_t was set to be the retail sales time series and X_t one of the corresponding keyword time series from the Google Trends dataset.. In this work, the assumption was made that the underlying bivariate stochastic process (X_t, Y_t) is stationary. As a result, it was assumed that the two time series X_t and Y_t have constant means (μ_X, μ_Y) and constant variances (σ_X^2, σ_Y^2). Based on this assumption, the sample cross-correlation has been calculated for a number of lags according to the formula in equation 2.20. With the help of the estimated standard deviations of both series (S_X and S_Y), 5% significance levels $(\pm 1.96/\sqrt{n})$ for confidence intervals could be approximated. Cross-correlation coefficients exceeding these limits have been considered significantly different from zero. They are suggesting significant associations between the two examined time series. However, these significance limits and thus the estimates of the cross-correlation coefficients are only correct, if the two time series X_t and Y_t are independent and stationary in a sense that they are serially uncorrelated (no autocorrelation). In practice, many empirical time series exhibit autocorrelated structures since they do not comprise independent values. These autocorrelated structures can lead to non-stationarity. When the time series themselves are non-stationary, their standard deviations can be higher or lower since the variances of both series are no longer constant. This in turn has a direct influence on the approximation of the confidence intervals and thus on the decision whether a cross-correlation coefficient is significantly different from zero (Dean and Dunsmuir [35]). Therefore, cross-correlations that have been calculated without taking autocorrelations into account may provide misleading statistical evidence of a linear relationship between independent non-stationary time series. In fact, pairs of autocorrelated time series that are completely independent of each other can show significant cross-correlations, even when neither has a causal effect on the other. The two sample time series illustrated in figure 2.5 may exhibit no significant lead-lag relationship once the autoregressive structure of both series is removed. Hence, temporal autocorrelations can inflate estimates of cross-correlations coefficients and cause high rates of incorrectly concluding linear relationships (i.e., Type I error) (Olden and Neff [75]). These misleading cross-correlations are called *spurious correlations* since they indicate spurious relationships. The cross-correlations in this work were calculated with the aforementioned assumptions. There is a risk that many such spurious correlations, that can be interpreted in no meaningful way, exist in the results. The time series were not tested individually for autocorrelations beforehand. To overcome this problem, one common solution is to remove the autocorrelation from at least one of the pair of series under study. This process is known as *prewhitening*. Prewhitening can be seen as a linear filtering operation and consists of the following three steps (Bloom, Buckeridge, and Cheng [15], Razavi and Vogel [78]):

1. Fitting a time series model (such as ARIMA) to the original time series X_t (external keyword) and store the residuals from this model.

2. Filter the response time series Y_t (retail sales) using the estimated coefficients from the model of step one.

3. Compute the cross-correlation between the residuals of step one and the filtered values of Y_t of step two. Both, the residuals and filtered values, form stand-alone time series.

The resulting cross-correlations relating the two (previously autocorrelated) time series can then be assessed and interpreted more reliably (Dean and Dunsmuir [35]). In order to evaluate the impact of prewhitening, this process was performed as a preceding step before computing the cross-correlations between the retail sales time series and each keyword series. It turned out that this process had a negative impact on the overall comparison between the ARIMA and ARIMAX models. The proportion of improvement decreased from 51% to 39% and the proportion of deterioration increased from 49% to 61%. This means that for more retail sales time series, leading indicators were selected as external regressors, which had a negative impact on the forecast accuracy. Consequently, the more robust interpretability of the cross-correlations was accompanied by a deterioration in the forecast accuracy. In addition, for 171 prewhitened retail sales time series no adequate ARIMA models could be found with the *auto.arima()* function. As the exploratory analysis of the results makes clear (see A.4), there may also be no contrary statements regarding the statistical significance of the cross-correlation. However, in the original computation of cross-correlation, there were ranges in which the proportion of the positive impact was greater. For the prewhitened cross-correlation, the proportion of the negative impact is greater in almost all ranges (in some cases even clearly). The hypothesis that the higher the similarity, the more beneficial the integration, still cannot be confirmed. There are also changes in the seasonal strength analysis of the retail sales time series. The proportion of the positive impact is still greater in higher ranges of seasonal strength, but the difference to the proportion of the negative impact seems to be no longer noticeable. This is presumably caused by the overall lower proportion of improvements and the reduced results dataset. Ultimately, this leads to the question of what is more important from a forecasting perspective. Since this work is primarily concerned with the question of which factors are critical for the success of the integration of leading indicators, the interpretability of the cross-correlations is of secondary importance. Nevertheless, this aspect should be critically questioned and considered when analyzing individual cases.

3. RANKING AND SELECTION For the purpose of finding appropriate leading indicators, cross-correlation coefficients for numerous lags were computed between each retail sales time series and keyword series from the Google Trends dataset in a 1:N fashion, where N represents the number of keyword time series. The highest cross-correlation and its corresponding lag for each keyword were stored temporarily. Subsequently, these keywords have been ranked by their achieved cross-correlation. From these N keywords, the keyword with the highest cross-correlation was then chosen to

be the leading indicator. This operation was performed for each time series of the retail sales dataset. Choosing the highest out of N cross-correlations, however, exhibit randomness. This randomness refers to two steps within this operation and multiplies from step to step. First, the cross-correlation values of the individual lags were compared with each other and the highest cross-correlation was selected and stored temporarily. Second, the cross-correlation values of the individual keywords were ranked in descending order and the keyword with the highest cross-correlation was considered leading indicator. In both steps, the selection was limited to absolute values. But what if the cross-correlation values of the individual lags or keywords are close to each other? It is possible that the corresponding second or third highest keyword makes a better leading indicator in terms of forecast accuracy. The lead-lag relationship may be more beneficial at other lags with similar cross-correlation values. The main problem is that absolute cross-correlation coefficients do not indicate whether they are significantly different from zero. Therefore, instead of choosing the highest cross-correlation per lag and keyword, the smallest p-value of a correlation-test per lag and per keyword should be chosen to ensure that the cross-correlations are statistically significant. Conducting multiple correlation-tests, however, increases the risk of a Type I error, i.e., to erroneously conclude the presence of a significant cross-correlation. For this reason, a Bonferroni correction should be applied in both steps to adjust the level of significance (Curtin and Schulz [31]). This alternative approach was implemented in the similarity matching with the functions $cor.test()$ and $p.adjust()$ from the `stats` package. It turned out that in only 383 cases ($< 25\%$ of all time series) different keywords and thus different leading indicators were selected based on p-values. The proportions of improvement and deterioration also remained unchanged. The seasonal strength analysis did not produce significantly different results either. The proportion of the positive impact remains notably greater in the range of $(0.7, 1.0]$ for the leading indicator time series (see A.3). Ultimately, this alternative approach did not lead to any major changes.

4. DICHOTOMIZATION OF CONTINUOUS VARIABLES The independent variables in the variable set are all metrically scaled. In order to conduct an ANOVA, these variables have been categorized using a simple dichotomization strategy. The cutpoints were set near the mean and the median values. Values that exceeded these cutpoints were classified as *high*. On the other hand, values below these cutpoints were classified as *low*. The values of the independent variables have been converted into two groups. This dichotomization simplifies the statistical analysis and leads to a straightforward interpretation and presentation of results. A binary split enables a comparison of groups of observations with high or low values of the independent variables (Altman and Royston [5]). This comparison can be realized with an ANOVA and a corresponding F-test, leading to an estimate of the difference between factors and groups. However, the dichotomization of continuous variables also comes with some drawbacks that can lead to several

problems. First, there is no good reason in general to suppose that an underlying dichotomy exits in the data, and if one exists, there is no reason why it should be at the median or mean value (MacCallum et al. [64]). The question arises where the cutpoints should be set. This decision is crucial since individual observations close to but on opposite sides of the cutpoints are characterized as being very different rather than very similar. Second, the statistical power to detect a relation between the factors and the response variable is reduced since dichotomization causes a high loss of information. The strategy applied in the ANOVA of this work reduces power by the same amount as discarding a third of the data would do (MacCallum et al. [64], Cohen [28]). Accordingly, not only the assumptions of the ANOVA were violated, but also the dichotomization of continuous variables led to a reduced statistical power of the results obtained.

For future research questions regarding leading indicator search, these four aspects should be an integral part in the planning and design of experiments. When calculating cross-correlations within the similarity matching, it should first be clarified which research objective should be weighted higher. Is it primarily about the interpretability of the achieved cross-correlations or are potential spurious correlations accepted for an improvement of forecast accuracy? If the latter is chosen, causal relationships should be assessed by experts in a subsequent step. In addition, multiple samples of keyword time series should be retrieved from Google Trends in order to avoid high variances within individual time series. If metric variables constitute the results of the experiment, the evaluation by ANOVA should be avoided due to the disadvantages of the required dichotomization, especially if the underlying assumptions are violated. The latter can be at least partially compensated by nonparametric alternatives such as ART-C. The alternative ranking and selection approach did not produce any significantly different results. The cross-correlation computation can be utilized as a methodology for similarity matching. Nevertheless, correlation-tests can be performed as evidence of statistical significance.

5 CONCLUSION

The objective of this work was to answer the core research question, whether the integration of external data sources and leading indicators can contribute to an improvement of forecast accuracy. To answer this question, leading indicators from an external online open data source were determined for each time series of a retail sales dataset. The approach chosen to address the question of how time series can be merged was based on a similarity matching with the cross-correlation as a similarity measure. Time series with a high similarity to those of the retail sales time series were included individually as external regressors in the linear time series model ARIMA, converting it into an ARIMAX model. The tool Google Trends was selected as a freely available online data source. The comparison of the forecasting performance between the bivariate ARIMAX models and their counterpart univariate ARIMA models reveals that there is an improvement in 51% and a deterioration in 49% of the cases. Therefore, the hypothesis that the inclusion of external time series as leading indicators can improve the forecast accuracy can be confirmed. However, there is no clear tendency and the improvements can be solely by chance.

In order to exclude this chance, individual time series with noticeable improvements and significant deteriorations have been examined. The analysis results suggest that there might be factors that are responsible for both extremes. Following the analysis of individual time series, it was then demonstrated, that the successful addition of external regressors only depends on the seasonal strengths of the merged time series. Moreover, the cross-correlation seems to have no significant influence. Improvements and deteriorations were both obtained equally at low and high cross-correlation values. The results of the multivariate analysis methods confirm these findings. For this reason, the hypothesis that the higher the similarity between two time series, the more beneficial the integration will be, cannot be confirmed. The seasonal pattern, on the other hand, turns out to be a critical success factor for the experiment. When both time series have high seasonal strengths, the probability of the leading indicator improving the forecast accuracy increases. This interaction between the two seasonal strength variables proves to be statistically significant. These results make it possible to answer the question, when it is useful to include external regressors and which conditions are critical for success.

Ultimately, it can be concluded that high and low cross-correlations can contribute to an improving forecast accuracy, provided the merged time series have strong seasonal patterns. Time series with high seasonal strengths

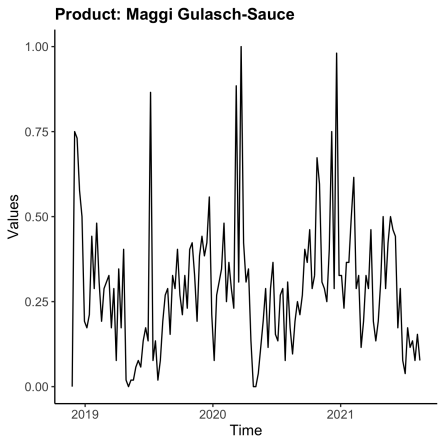
are therefore suitable candidates for leading indicator search. For this reason, it is important to consider the seasonal strength of time series within the matching process in further research questions. In addition, further similarity measures that may enable a robust interpretability in connection with an improvement of forecast performance should be tested.

Part II

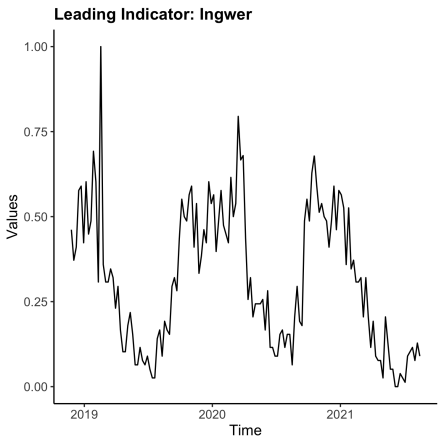
APPENDIX

APPENDIX

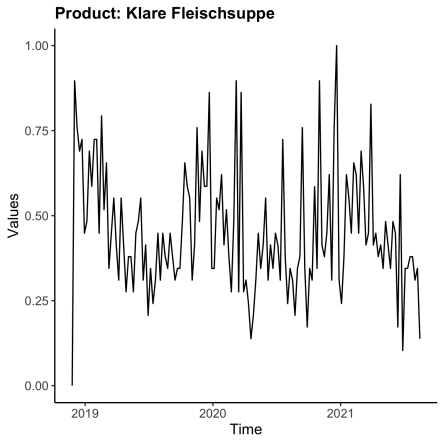
A.1 SELECTED TIME SERIES PAIRS



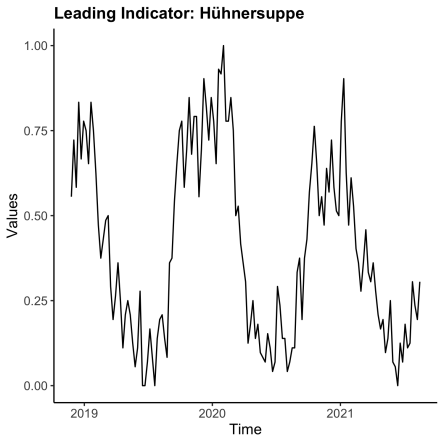
(a) Retail sales time series *first* time series pair.



(b) Leading indicator time series *first* time series pair.



(c) Retail sales time series *second* time series pair.



(d) Leading indicator time series *second* time series pair.

Figure A.1: Time series pairs from evaluation of multiple linear regression. Cross-correlation values of 0.47 and 0.45.

A.2 HYNDMAN-KHANDAKAR ALGORITHM

Hyndman-Khandakar algorithm for automatic ARIMA modelling

1. The number of differences $0 \leq d \leq 2$ is determined using repeated KPSS tests.
2. The values of p and q are then chosen by minimising the AICc after differencing the data d times. Rather than considering every possible combination of p and q, the algorithm uses a stepwise search to traverse the model space.
   a. Four initial models are fitted:
      - ARIMA$(0, d, 0)$,
      - ARIMA$(2, d, 2)$,
      - ARIMA$(1, d, 0)$,
      - ARIMA$(0, d, 1)$.

      A constant is included unless $d = 2$. If $d \leq 1$, an additional model is also fitted:
      - ARIMA$(0, d, 0)$ without a constant.

   b. The best model (with the smallest AICc value) fitted in step (a) is set to be the "current model".

   c. Variations on the current model are considered:
      - vary p and/or q from the current model by ± 1;
      - include/exclude c from the current model.

      The best model considered so far (either the current model or one of these variations) becomes the new current model.

   d. Repeat Step 2(c) until no lower AICc can be found.

Figure A.2: Default behavior of the Hyndman-Khandakar algorithm used in the function *auto.arima*() (Hyndman and Khandakar [53]).

A.3 ALTERNATIVE RANKING AND SELECTION PROCEDURE

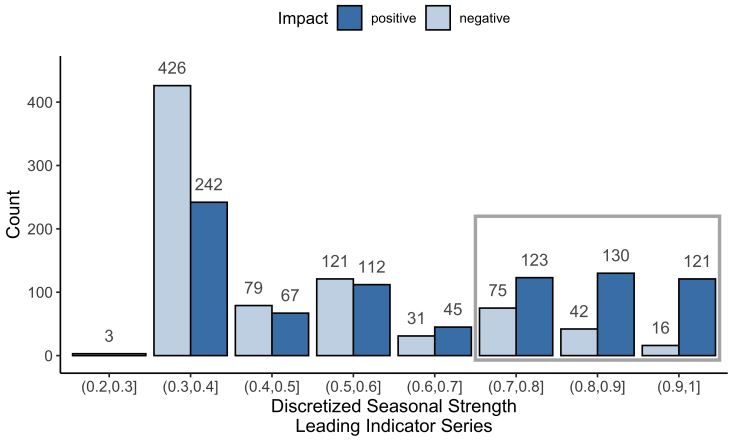
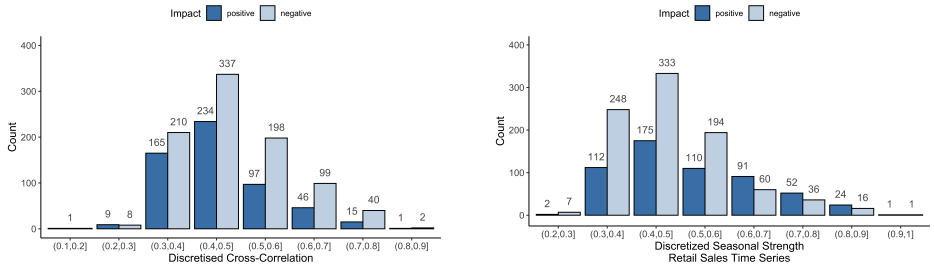


Figure A.3: Discretized seasonal strength from leading indicator time series separated by impact on ARIMA model. Results are obtained from the alternative ranking and selection procedure.

A.4 RESULTS OF THE PREWHITENING PROCESS



(a) Discretized cross-correlation separated by impact.

(b) Discretized seasonal strength from retail sales time series separated by impact.

Figure A.4: Results of the prewhitening process.

BIBLIOGRAPHY

[1] Ghahreman Abdoli, Mohsen MehrAra, and Mohammad Ardalani. "COMPARING THE PREDICTION ACCURACY OF LSTM AND ARIMA MODELS FOR TIME-SERIES WITH PERMANENT FLUCTUATION." In: *Gênero & Direito* 9 (2020). DOI: 10.22478/ufpb.2179-7137.2020v9n2.50782.

[2] Olu Aganjuomo. *How To Pull Data From Google Trends API*. https://medium.com/@oluaganju/how-to-pull-data-from-google-trends-api-a9ce975dd80f. [Online; accessed 2022-05-01]. 2021.

[3] Leona S. Aiken and Stephen G. West. *Multiple Regression: Testing and Interpreting Interactions*. SAGE Publications Inc., 1991. ISBN: 9780803936058.

[4] Paul D. Allison. *Multiple Regression: A Primer*. Pine Forge Press series in research methods and statistics. SAGE Publications Inc., 1999. ISBN: 9780761985334.

[5] Douglas G. Altman and Patrick Royston. "The cost of dichotomising continuous variables." In: *BMJ (Clinical research ed.)* 332.7549 (2006), pp. 1080–1080. DOI: 10.1136/bmj.332.7549.1080. URL: https://doi.org/10.1136/bmj.332.7549.1080.

[6] Sylvain Arlot and Alain Celisse. "A survey of cross-validation procedures for model selection." In: *Statistics Surveys* 4 (2010), pp. 40–79. DOI: 10.1214/09-SS054. URL: https://doi.org/10.1214/09-SS054.

[7] Sitaram Asur and Bernardo A. Huberman. "Predicting the Future with Social Media." In: *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*. Vol. 1. IEEE, 2010, pp. 492–499. DOI: 10.1109/WI-IAT.2010.63.

[8] Lyle F. Bachman. *Statistical Analyses for Language Assessment Book*. Cambridge Language Assessment. Cambridge University Press, 2004. ISBN: 9780521802772.

[9] Klaus Backhaus, Bernd Erichson, Wulff Plinke, and Rolf Weiber. *Multivariate Analysemethoden. Eine anwendungsorientierte Einführung*. 14th ed. Springer, Berlin, Heidelberg, 2016. ISBN: 9783662460757. DOI: https://doi.org/10.1007/978-3-662-46076-4.

[10] Hatice Ozer Balli and Bent E. Sørensen. "Interaction effects in econometrics." In: *Empirical Economics* 45.1 (2013), pp. 583–603. DOI: https://doi.org/10.1007/s00181-012-0604-2.

[11] Kasun Bandara, Rob J. Hyndman, and Christoph Bergmeir. *MSTL: A Seasonal-Trend Decomposition Algorithm for Time Series with Multiple Seasonal Patterns*. 2021. DOI: 10.48550/ARXIV.2107.13462. URL: https://arxiv.org/abs/2107.13462.

[12] Ralf Bender and Stefan Lange. "Varianzanalyse." In: *Dtsch Med Wochenschr* 132.21 (2007), pp. 57–60. DOI: 10.1055/s-2007-959044.

[13] Christoph Bergmeir, Rob J. Hyndman, and Bonsoo Koo. "A note on the validity of cross-validation for evaluating autoregressive time series prediction." In: *Computational Statistics & Data Analysis* 120 (2018), pp. 70–83. DOI: https://doi.org/10.1016/j.csda.2017.11.003.

[14] Jean-Louis Bertrand, Xavier Brusset, and Maxime Fortin. "Assessing and hedging the cost of unseasonal weather: Case of the apparel sector." In: *European Journal of Operational Research* 244.1 (2015), pp. 261–276. DOI: https://doi.org/10.1016/j.ejor.2015.01.012. URL: https://www.sciencedirect.com/science/article/pii/S0377221715000326.

[15] Ronald M. Bloom, David L. Buckeridge, and Karen E. Cheng. "Finding Leading Indicators for Disease Outbreaks: Filtering, Cross-correlation, and Caveats." In: *Journal of the American Medical Informatics Association* 14.1 (2007), pp. 76–85. DOI: 10.1197/jamia.M2178.

[16] Tonya Boone, Ram Ganeshan, and Robert L. Hicks. "Incorporating Google Trends Data Into Sales Forecasting." In: *Foresight: The International Journal of Applied Forecasting* 38 (2015), pp. 9–14.

[17] Tonya Boone, Ram Ganeshan, Robert L. Hicks, and Nada R. Sanders. "Can Google Trends Improve Your Sales Forecast?" In: *Production and Operations Management* 27.10 (2018), pp. 1770–1774. DOI: https://doi.org/10.1111/poms.12839. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/poms.12839.

[18] Günther Bourier. "Zeitreihenanalyse." In: *Beschreibende Statistik*. Gabler Verlag, 2012. Chap. 6, pp. 155–194.

[19] George E. P. Box and Gwilym M. Jenkins. *Time Series Analysis: Forecasting and Control*. Holden-Day series in time series analysis and digital processing. Holden-Day, 1970. ISBN: 9780816210947.

[20] George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung. *Time Series Analysis: Forecasting and Control*. 5th ed. Wiley Series in Probability and Statistics. Wiley, 2015. ISBN: 9781118674925.

[21] Trevor S. Breusch and Adrian R. Pagan. "A Simple Test for Heteroscedasticity and Random Coefficient Variation." In: *Econometrica* 47.5 (1979), pp. 1287–1294.

[22] Peter J. Brockwell and Richard A. Davis. *Introduction to Time Series and Forecasting*. 2nd ed. Springer Texts in Statistics. Springer, New York, NY, 2002. ISBN: 9780387953519. DOI: https://doi.org/10.1007/b97391.

[23] Eduardo Cebrián and Josep Domenech. "Is Google Trends a quality data source?" In: *Applied Economics Letters* (2022), pp. 1–5. DOI: 10.1080/13504851.2021.2023088. URL: https://doi.org/10.1080/13504851.2021.2023088.

[24] Chris Chatfield. *Time-Series Forecasting*. Chapman & Hall, CRC Press, 2000. ISBN: 9781584880639.

[25] Hyunyoung Choi, and Hal Varian. "Predicting the Present with Google Trends." In: *Economic Record* 88.s1 (2012), pp. 2–9. DOI: https://doi.org/10.1111/j.1475-4932.2012.00809.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1475-4932.2012.00809.x.

[26] Ching-Wu Chu and Guoqiang Peter Zhang. "A comparative study of linear and nonlinear models for aggregate retail sales forecasting." In: *International Journal of Production Economics* 86.3 (2003), pp. 217–231. DOI: https://doi.org/10.1016/S0925-5273(03)00068-9.

[27] Robert B. Cleveland, William S. Cleveland, Jean E. McRae, and Irma Terpenning. "STL: A Seasonal-Trend Decomposition Procedure Based on Loess." In: *Journal of Official Statistics* 6.1 (1990), pp. 3–33.

[28] Jacob Cohen. "The Cost of Dichotomization." In: *Applied Psychological Measurement* 7.3 (1983), pp. 249–253. DOI: 10.1177/014662168300700301. URL: https://doi.org/10.1177/014662168300700301.

[29] Jacob Cohen, Patricia Cohen, Stephen G. West, and Leona S. Aiken. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. 3rd ed. Mahwah, New Jersey, USA: Lawrence Erlbaum Associates, Inc., 2003.

[30] Christine S. M. Currie and Ian T. Rowley. "Consumer behaviour and sales forecast accuracy: What's going on and how should revenue managers respond?" In: *Journal of Revenue and Pricing Management* 9.4 (2010), pp. 374–376. DOI: 10.1057/rpm.2010.22.

[31] François Curtin and Pierre Schulz. "Multiple correlations and Bonferroni's correction." In: *Biological Psychiatry* 44.8 (1998), pp. 775–777. DOI: 10.1016/S0006-3223(98)00043-2. URL: https://doi.org/10.1016/S0006-3223(98)00043-2.

[32] Claudia Czado and Thorsten Schmidt. *Mathematische Statistik*. Statistik und ihre Anwendungen. Springer, Berlin, Heidelberg, 2011. ISBN: 9783642172601. DOI: https://doi.org/10.1007/978-3-642-17261-8.

[33] Estela Dagum and Alessandra Luati. "Global and local statistical properties of fixed-length nonparametric smoothers." In: *Statistical Methods and Applications* 11 (2002), pp. 313–333. DOI: 10.1007/BF02509830.

[34] Matteo De Felice, Andrea Alessandri, and Paolo M. Ruti. "Electricity demand forecasting over Italy: Potential benefits using numerical weather prediction models." In: *Electric Power Systems Research* 104 (2013), pp. 71–79. DOI: https://doi.org/10.1016/j.epsr.2013.06.004.

[35] Roger T. Dean and William T. M. Dunsmuir. "Dangers and uses of cross-correlation in analyzing time series in perception, performance, movement, and neuroscience: The importance of constructing transfer function autoregressive models." In: *Behavior Research Methods* 48.2 (2016), pp. 783–802. DOI: 10.3758/s13428-015-0611-2. URL: https://doi.org/10.3758/s13428-015-0611-2.

[36] David A. Dickey and Wayne A. Fuller. "Distribution of the Estimators for Autoregressive Time Series With a Unit Root." In: *Journal of the American Statistical Association* 74.366 (1979), pp. 427–431.

[37] Peter Durka and Silvia Pastorekova. "ARIMA vs. ARIMAX - which approach is better to analyze and forecast macroeconomic time series?" In: *Proceedings of 30th international conference on mathematical methods in economics* (2012), pp. 136–140.

[38] Eugene Edgington. *Randomization Tests, Fourth Edition*. Statistics: A Series of Textbooks and Monographs. Taylor & Francis, 1995. ISBN: 9780824796693.

[39] Lisa A. Elkin, Matthew Kay, James J. Higgins, and Jacob O. Wobbrock. "An Aligned Rank Transform Procedure for Multifactor Contrast Tests." In: *The 34th Annual ACM Symposium on User Interface Software and Technology*. Association for Computing Machinery, 2021.

[40] Mohammed Elshendy, Andrea Fronzetti Colladon, Elisa Battistoni, and Peter A. Gloor. "Using four different online media sources to forecast the crude oil price." In: *Journal of Information Science* 44.3 (2018), pp. 408–421. DOI: 10.1177/0165551517698298. URL: https://doi.org/10.1177/0165551517698298.

[41] John Fox and Sanford Weisberg. *An R Companion to Applied Regression*. Third. Thousand Oaks, California, USA: Sage, 2019. URL: https://socialsciences.mcmaster.ca/jfox/Books/Companion/.

[42] Daniel Fredén and Hampus Larsson. "Forecasting Daily Supermarkets Sales with Machine Learning." PhD thesis. Stockholm, Sweden: KTH Royal Institute of Technology School of Engineering Science, 2020.

[43] Sabine Fromm. *Datenanalyse mit SPSS für Fortgeschrittene 2: Multivariate Verfahren für Querschnittsdaten*. VS Verlag für Sozialwissenschaften, 2010. ISBN: 9783531147925. DOI: https://doi.org/10.1007/978-3-531-92026-9.

[44] Andrew Gelman. "Scaling Regression Inputs by Dividing by Two Standard Deviations." In: *Statistics in medicine* 27 (July 2008), pp. 2865–73. DOI: 10.1002/sim.3107.

[45] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference and prediction*. 2nd ed. Springer series in statistics. Springer, 2009. ISBN: 9780387848846. URL: http://www-stat.stanford.edu/~tibs/ElemStatLearn/.

[46] Andreas Herrmann and Christian Seilheimer. "Varianz- und Kovarianzanalyse." In: *Marktforschung: Methoden, Anwendungen & Praxisbeispiele*. Ed. by Andreas Herrmann and Christian Homburg. 2nd ed. Wiesbaden, Germany: Gabler, 2000, pp. 267–294.

[47] Tom Houslay. *UNDERSTANDING 3-WAY INTERACTIONS BETWEEN CONTINUOUS VARIABLES*. https://tomhouslay.com/2014/03/21/understanding-3-way-interactions-between-continuous-variables/. [Online; accessed 2022-04-14]. 2014.

[48] Tao Huang, Robert Fildes, and Didier Soopramanien. "The value of competitive information in forecasting FMCG retail product sales and the variable selection problem." In: *European Journal of Operational Research* 237.2 (2014), pp. 738–748. DOI: https://doi.org/10.1016/j.ejor.2014.02.022. URL: https://www.sciencedirect.com/science/article/pii/S0377221714001374.

[49] Rob J. Hyndman. *The ARIMAX model muddle*. https://robjhyndman.com/hyndsight/arimax/. [Online; accessed 2022-03-14]. 2010.

[50] Rob J. Hyndman and George Athanasopoulos. *Forecasting: principles and practice*. 2nd ed. [Online; accessed 08-March-2022]. Melbourne, Australia: OTexts, 2018.

[51] Rob J. Hyndman and George Athanasopoulos. *Forecasting: principles and practice*. 3rd ed. [Online; accessed 21-March-2022]. Melbourne, Australia: OTexts, 2021.

[52] Rob J. Hyndman, George Athanasopoulos, Christoph Bergmeir, Gabriel Caceres, Leanne Chhay, Mitchell O'Hara-Wild, Fotios Petropoulos, Slava Razbash, Earo Wang, and Farah Yasmeen. *forecast: Forecasting functions for time series and linear models*. R package version 8.16. 2022. URL: https://pkg.robjhyndman.com/forecast/.

[53] Rob J. Hyndman and Yeasmin Khandakar. "Automatic time series forecasting: the forecast package for R." In: *Journal of Statistical Software* 26.3 (2008), pp. 1–22. DOI: 10.18637/jss.v027.i03.

[54] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. 2nd ed. Springer, 2021. ISBN: 9781071614174.

[55] Werner Kaiser. "Fast Moving Consumer Goods." In: *Qualitative Marktforschung in Theorie und Praxis: Grundlagen, Methoden und Anwendungen*. Ed. by Gabriele Naderer and Eva Balzer. Wiesbaden, Germany: Gabler, 2011, pp. 605–615. ISBN: 9783834967909. DOI: 10.1007/978-3-8349-6790-9_31. URL: https://doi.org/10.1007/978-3-8349-6790-9_31.

[56] Matthew Kay. *Contrast tests with ART*. https://cran.r-project.org/web/packages/ARTool/vignettes/art-contrasts.html. [Online; accessed 2022-04-20]. 2020.

[57] Matthew Kay, Lisa A. Elkin, James J. Higgins, and Jacob O. Wobbrock. *ARTool: Aligned Rank Transform for Nonparametric Factorial ANOVAs*. R package version 0.11.1. 2021. DOI: 10.5281/zenodo.594511. URL: https://github.com/mjskay/ARTool.

[58] Michael H. Kutner, Christopher J. Nachtsheim, William Wasserman, and John Neter. *Applied Linear Statistical Models*. Chicago: Irwin, 1996.

[59] Denis Kwiatkowski, Peter C.B. Phillips, Peter Schmidt, and Yongcheol Shin. "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 (1992), pp. 159–178. DOI: https://doi.org/10.1016/0304-4076(92)90104-Y. URL: https://www.sciencedirect.com/science/article/pii/030440769290104Y.

[60] Sara C. Larson. "The shrinkage of the coefficient of multiple correlation." In: *Journal of Educational Psychology* 22 (1931), pp. 45–55.

[61] Russell V. Lenth. *emmeans: Estimated Marginal Means, aka Least-Squares Means*. R package version 1.7.2. 2022. URL: https://CRAN.R-project.org/package=emmeans.

[62] Jacob A. Long. *interactions: Comprehensive, User-Friendly Toolkit for Probing Interactions*. R package version 1.1.0. 2019. URL: https://cran.r-project.org/package=interactions.

[63] Jacob A. Long. *Exploring interactions with continuous predictors in regression models*. https://cran.r-project.org/web/packages/interactions/vignettes/interactions.html. [Online; accessed 2022-04-14]. 2021.

[64] Robert C. MacCallum, Shaobo Zhang, Kristopher J. Preacher, and Derek D. Rucker. "On the practice of dichotomization of quantitative variables." In: *Psychological methods* 7(1) (2002), pp. 19–40. URL: https://doi.org/10.1037/1082-989X.7.1.19.

[65] Burton G. Malkiel. *A Random Walk Down Wall Street: The Time-tested Strategy for Successful Investing*. W. W. Norton & Company, 2020. ISBN: 9780393358384.

[66] Marcelo C. Medeiros and Henrique Pires. *The proper use of Google Trends in forecasting models*. Tech. rep. 683. Rio de Janeiro, Brazil, 2021.

[67] Terence C. Mills. *The Foundations of Modern Time Series Analysis*. Palgrave Advanced Texts in Econometrics. Palgrave Macmillan UK, 2011. ISBN: 978-0230290181.

[68] Kyle B. Murray, Fabrizio Di Muro, Adam Finn, and Peter Popkowski Leszczyc. "The effect of weather on consumer spending." In: *Journal of Retailing and Consumer Services* 17.6 (2010), pp. 512–520. DOI: https://doi.org/10.1016/j.jretconser.2010.08.006.

[69] Robert Nau. *Notes on the random walk model*. https://people.duke.edu/~rnau/Notes_on_the_random_walk_model--Robert_Nau.pdf. [Online; accessed 02-03-2022]. 2014.

[70] Danielle Navarro. *Assumptions of One-way ANOVA*. https://stats.libretexts.org/@go/page/4033. [Online; accessed 2022-04-20]. 2020.

[71] Danielle Navarro. *Comparing Several Means (One-way ANOVA)*. https://stats.libretexts.org/@go/page/4034. [Online; accessed 2022-04-20]. 2020.

[72] Danielle Navarro. *Factorial ANOVA 2 - Balanced Designs, Interactions Allowed*. https://stats.libretexts.org/@go/page/4043. [Online; accessed 2022-04-22]. 2020.

[73] Danielle Navarro. *Factorial ANOVA 3 Unbalanced Designs*. https://stats.libretexts.org/@go/page/8303. [Online; accessed 2022-04-20]. 2020.

[74] Geoffrey R. Norman and David L. Streiner. *Biostatistics: The Bare Essentials*. B.C. Decker, 2008. ISBN: 9781550094008.

[75] Julian Olden and Bryan Neff. "Cross-correlation bias in lag analysis of aquatic time series." In: *Marine Biology* 138 (2001), pp. 1063–1070. DOI: 10.1007/s002270000517.

[76] Bohdan M. Pavlyshenko. "Machine-Learning Models for Sales Time Series Forecasting." In: *Data* 4.1 (2019). DOI: 10.3390/data4010015. URL: https://www.mdpi.com/2306-5729/4/1/15.

[77] Karl Pearson. "The Problem of the Random Walk." In: *Nature* 72 (1905), p. 318. ISSN: 1476-4687. DOI: 10.1038/072318a0.

[78] Saman Razavi and Richard Vogel. "Prewhitening of hydroclimatic time series? Implications for inferred change and variability across time scales." In: *Journal of Hydrology* 557 (2018), pp. 109–115. DOI: https://doi.org/10.1016/j.jhydrol.2017.11.053. URL: https://www.sciencedirect.com/science/article/pii/S0022169417308181.

[79] Simon Rogers. *What is Google Trends data - and what does it mean?* https://medium.com/google-news-lab/what-is-google-trends-data-and-what-does-it-mean-b48f07342ee8. [Online; accessed 2022-05-01]. 2016.

[80] Jesús Rojo, Rosario Rivero, Jorge R. Morte, Federico Fernández-González, and Rosa Pérez-Badia. "Modeling pollen time series using seasonal-trend decomposition procedure based on LOESS smoothing." In: *International Journal of Biometeorology* 61 (2017), pp. 335–348. DOI: 10.1007/s00484-016-1215-y.

[81] Yves R. Sagaert, El-Houssaine Aghezzaf, Nikolaos Kourentzes, and Bram Desmet. "Tactical sales forecasting using a very large set of macroeconomic indicators." In: *European Journal of Operational Research* 264.2 (2018), pp. 558–569. DOI: https://doi.org/10.1016/j.ejor.2017.06.054.

[82] Holger Schielzeth. "Simple means to improve the interpretability of regression coefficients." In: *Methods in Ecology and Evolution* 1.2 (2010), pp. 103–113. DOI: https://doi.org/10.1111/j.2041-210X.2010.00012.x.

[83] Samuel S. Shapiro and Martin B. Wilk. "An Analysis of Variance Test for Normality (Complete Samples)." In: *Biometrika* 52.3/4 (1965), pp. 591–611.

[84] Sima Siami-Namini, Neda Tavakoli, and Akbar Siami Namin. "A Comparison of ARIMA and LSTM in Forecasting Time Series." In: *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. 2018, pp. 1394–1401. DOI: 10.1109/ICMLA.2018.00227.

[85] Robert Siwerz and Christopher Dahlén. "Predicting sales in a food store department using machine learning." PhD thesis. Stockholm, Sweden: KTH Royal Institute of Technology School of Computer Science and Communication, 2017. URL: http://www.diva-portal.org/smash/get/diva2:1108597/FULLTEXT01.pdf.

[86] Aya Soffer. "Image categorization using texture features." In: *Proceedings of the Fourth International Conference on Document Analysis and Recognition*. Vol. 1. IEEE, 1997, pp. 233–237. DOI: 10.1109/ICDAR.1997.619847.

[87] James H. Stock and Mark W. Watson. *Forecasting Inflation*. Working Paper 7023. National Bureau of Economic Research, 1999. DOI: 10.3386/w7023.

[88] Mervyn Stone. "Cross-Validatory Choice and Assessment of Statistical Predictions." In: *Journal of the royal statistical society series b-methodological* 36 (1974), pp. 111–147.

[89] Neda Tavakoli, Sima Siami Namini, Mahdi Khanghah, Fahimeh Soltani, and Akbar Siami Namin. "An autoencoder-based deep learning approach for clustering time series data." In: *SN Applied Sciences* 2 (2020). DOI: 10.1007/s42452-020-2584-8.

[90] Sean J. Taylor and Benjamin Letham. "Forecasting at Scale." In: *The American Statistician* 72.1 (2018), pp. 37–45. DOI: 10.1080/00031305.2017.1380080.

[91] Uyodhu A. Victor-Edema and Isaac D. Essi. "Autoregressive Integrated Moving Average with Exogenous Variable (ARIMAX) Model for Nigerian Non Oil Export." In: *European Journal of Business and Management* 8 (2016), pp. 29–34.

[92] Xiaozhe Wang, Kate Smith, and Rob J. Hyndman. "Characteristic-Based Clustering for Time Series Data." In: *Data Mining and Knowledge Discovery* 13.3 (2006), pp. 335–364. DOI: 10.1007/s10618-005-0039-x. URL: https://doi.org/10.1007/s10618-005-0039-x.

[93] Kinley Wangdi, Pratap Singhasivanon, Tassanee Silawan, Saranath Lawpoolsri, Nicholas J. White, and Jaranit Kaewkungwal. "Development of temporal modelling for forecasting and prediction of malaria infections using time-series and ARIMAX analyses: A case study in endemic districts of Bhutan." In: *Malaria Journal* 9.251 (2010). DOI: 10.1186/1475-2875-9-251. URL: https://doi.org/10.1186/1475-2875-9-251.

[94] Qingsong Wen, Zhe Zhang, Yan Li, and Liang Sun. "Fast RobustSTL: Efficient and Robust Seasonal-Trend Decomposition for Time Series with Complex Patterns." In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (2020), pp. 2203–2213. DOI: 10.1145/3394486.3403271.

[95] Joachim Werner. *Lineare Statistik*. BeltzPVU, 1997. ISBN: 9783621275095.

[96] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN: 9783319242774. URL: https://ggplot2.tidyverse.org.

[97] Jacob O. Wobbrock, Leah Findlater, Darren Gergle, and James J. Higgins. "The Aligned Rank Transform for Nonparametric Factorial Analyses Using Only ANOVA Procedures." In: vol. 2011. 2011, pp. 143–146. DOI: 10.1145/1978942.1978963.

[98] Min Xie and Thong Ngee Goh. "A comparative study of neural network and Box-Jenkins ARIMA modeling in time series prediction." In: *Computers & Industrial Engineering* 42.2 (2002), pp. 371–375. DOI: https://doi.org/10.1016/S0360-8352(02)00036-0.

[99] Guoqiang Peter Zhang. "Time series forecasting using a hybrid ARIMA and neural network model." In: *Neurocomputing* 50 (2003), pp. 159–175. DOI: https://doi.org/10.1016/S0925-2312(01)00702-0.

[100] Jing Zhou, Zhongyao Liang, Yong Liu, Huaicheng Guo, Dan He, and Lei Zhao. "Six-decade temporal change and seasonal decomposition of climate variables in Lake Dianchi watershed (China): stable trend or abrupt shift?" In: *Theoretical and Applied Climatology* 119.1 (2015), pp. 181–191.

[101] Indrė Žliobaitė, Jorn Bakker, and Mykola Pechenizkiy. "Beating the baseline prediction in food sales: How intelligent an intelligent predictor is?" In: *Expert Systems with Applications* 39.1 (2012), pp. 806–815. DOI: https://doi.org/10.1016/j.eswa.2011.07.078.