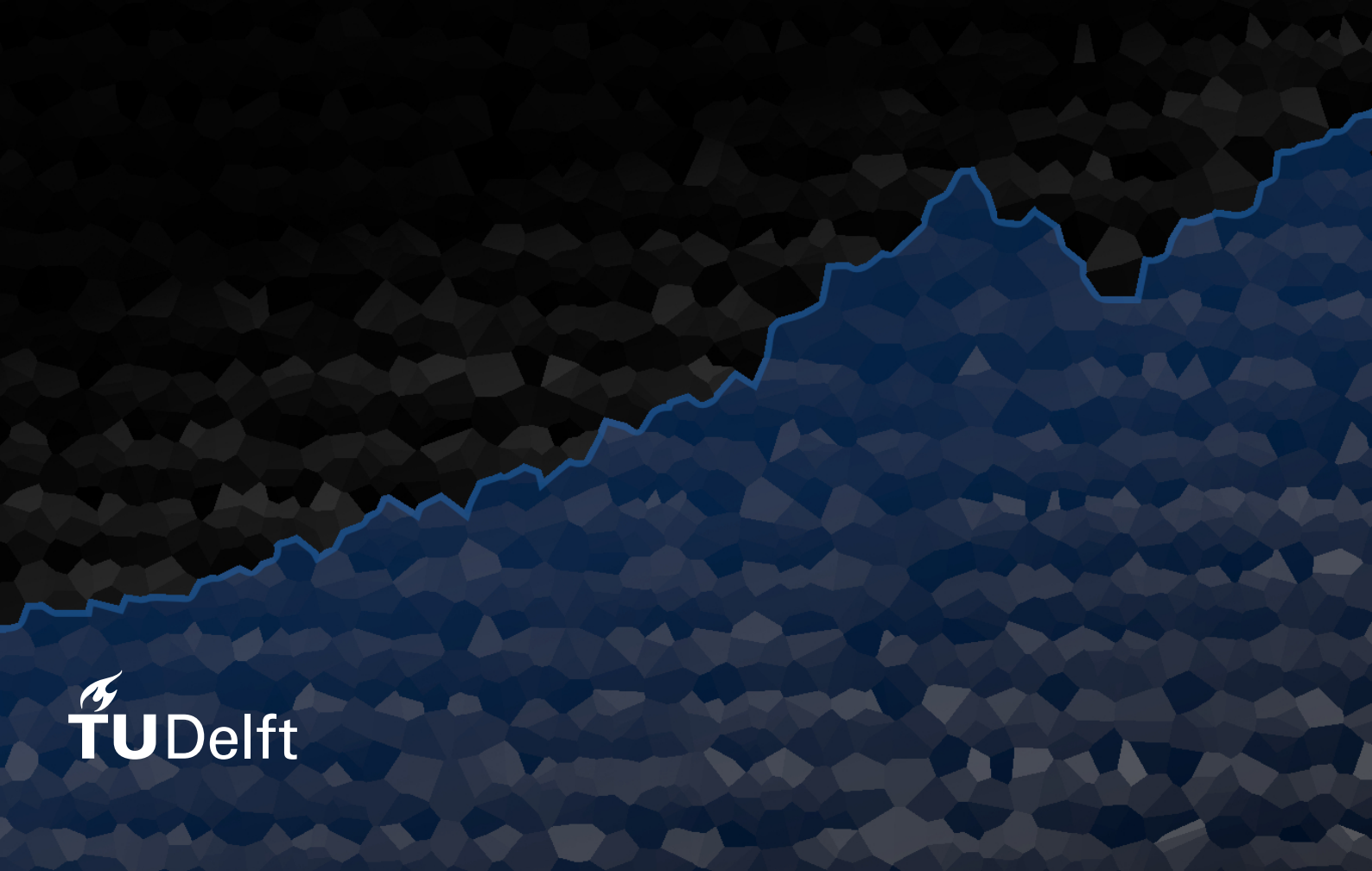
Predicting financial time series with incomplete information due to late publications of financial reports

Master Thesis

Jeroen Esseveld

Delft University of Technology



TU Delft

Predicting financial time series with incomplete information due to late publications of financial reports

Master Thesis

by

Jeroen Esseveld

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Friday July 10, 2020 at 14:00.

Student number:	4481119		
Supervisors:	Prof. dr. M. Loog		
	Ir. E. Rambier		
Thesis committee:	Dr. J. van Gemert,	Committee chair	TU Delft
	Prof. dr. M. Loog,	Supervisor	TU Delft
	Dr. C. Lofi,	Committee member	TU Delft
	Ir. E. Rambier,	Supervisor	Exact

TU Delft

= exact

A complex system that works is invariably found to have evolved from a simple system that worked.
— JOHN GALL (1975)

Preface

This thesis report is the result of nine months work. First of all I want to thank Marco Loog, Estelle Rambier and Judith Redi for giving me the opportunity to do this great project. Thank you for supervising me throughout the process. I have had a great time during my internship at Exact. The Data Science team of Exact was very welcoming to me and it has always been a pleasure working with you. I want to thank you, the Data Science team, for the good memories and delicious cappuccinos we have shared.

During the last three months of my thesis project, I e-mailed Marco a lot about my thesis. I worked from home due to COVID-19 measures, so video calls and e-mails was our discussion platform. Marco's teamwork is characterized with fast replies to my emails. In times when Marco was busy, he would make time to reply my email late in the evenings or during the weekends. My sincere thanks for your amazing supervision.

Estelle and I have shared countless of coffees and have played very competitive table tennis games. When I had questions, I could always turn to you Estelle and discuss anything with you. I am grateful for that. Thank you for being such a great supervisor!

Almost five years ago I started as a Bachelor student in Computer Science. it keeps amazing me how wonderful the world of computers and programming is. This motivated me to continue studying for the Master's degree. Looking back at those five years, I never regretted enrolling for Computer Science at this university. I hereby thank Ivo Swartjes, who supervised me during my internship in 2013, and inspired me to go to the university.

I want to thank Rebecca Kemeling for supporting me throughout my thesis project and for all your advice and guidance. I thank my friends and family for their help and support.

Finally, I thank my parents, Rob and Nelly, for always supporting me and believing in me. With their support, I was able to climb the ladder of education.

The Appendix of this report contains further notes and is not a part of the core study. Materials of the Appendix are inessential to the understanding of this report, but gives additional insight in the topic.

Jeroen Esseveld
Delft, July 2020

Disclaimer

The information made available by Exact for this research is provided for use of this research only and under strict confidentiality. Tables and figures displaying time series use example data because real world data should remain non-disclosed.

Abstract

Records from ledgers of Dutch companies all across the Netherlands are used in this study. Records can be submitted in the ledgers with various lags, because the data of many different bookkeepers is involved with different workflows. Bookkeepers can be punctual or late, therefore records can be submitted with various lags in the ledgers. This causes missing data, which results in a deformation of a time series that is constructed from these records. Using a technique called *now-casting*, a prediction can be made of how these series with no missing data would have looked like.

This study sheds light on how information of an incomplete time series from ledgers of Dutch companies can be used to *now-cast* on that series, without the use of external indicators. To better utilize the information available from the series, an addition to the Seasonal Auto Regressive Integrated Moving Average with eXogenous regressors (SARIMAX) model is proposed. The addition to the SARIMAX model is presented in two forms: the additive- and multiplicative relation between indicator- and target series. These are modeled with the goal to improve the information utilization and therefore improve the *now-casting* accuracy. Experiments have shown that this addition to the model does not give a direct improvement in accuracy compared to an ordinary SARIMAX model. Thereafter an iterative *now-cast* procedure is proposed to utilize information from highly lagged records. It has been shown that this gives a slight increase in accuracy for the overall *now-cast*.

Contents

List of Figures

List of Tables

Nomenclature

Abbreviations and Acronyms

AR	Auto Regressive (component)
ARMAX	Auto Regressive Moving Average with eXogenous regressors
CBS	Dutch national institute of statistics
DFM	Dynamic Factor Model
ECB	European Central Bank
GDP	Gross Domestic Product
I	Integrated (component)
KPI	Key Performance Indicator
MA	Moving Average (component)
MASE	Mean-Absolute-Scaled-Error
MF-VAR	Mixed Frequency Vector Auto Regressive
MIDAS	Mixed Data Sampling
nRMSE	normalised Root-Mean-Squared-Error
PCA	Principal Component Analysis
RCSFI	Reference Classification System of Financial Information
S	Seasonal (component)
SARIMA	Seasonal Auto Regressive Integrated Moving Average
SARIMAX	Seasonal Auto Regressive Integrated Moving Average with eXogenous regressors
SME	Small to Medium Enterprise
X	Exogenous (component)

1

Introduction and Related work

Exact is a Dutch software company that offers a service for online accounting[1]. One of the goals of Exact is to give financial insights into the Dutch economy by means of Key Performance Indicators (KPIs) to its customers[2]. Timely and reliable KPIs play a key role for important decisions of companies [3]. Bookkeeping records from ledgers of Dutch Small- to Medium Enterprises (SMEs) are used to produce these KPIs. To generate KPIs which give a representative view on the current state of the Dutch economy, records from the bleeding edge are used. These records are submitted in ledgers of the online accounting software. They are submitted by many bookkeepers from different companies and sectors all across the Netherlands. Bookkeeping can be done in many ways. A common bookkeeping workflow is to submit records of transactions after the end of each quarter. This way, records are submitted in time to publish quarterly financial reports. Various workflows can differ in punctuality, which influences the delay in which records are submitted. As an example, a sales transaction that happened in January could be submitted in the ledger by a bookkeeper as a record after the first quarter, in April. Bookkeepers which are more punctual, would for example submit records of all transactions before the end of the week. These reporting lags can vary from a day to over a year.

The following pattern from aggregated records is observed: a lot of transactions are not yet recorded in the recent past. Further back in the past, fewer records are still missing. Figure 1.1 displays the average distribution of records which are still not reported after one or more months have passed since the transaction date. From the figure, one could observe that when three months have passed, less than 10% of records are still expected to be submitted.

1.1. Problem statement

At Exact, KPI time series are produced by summing a selection of records, partitioned in a monthly interval. A problem arises: KPI time series constructed with records (of which some are still missing due to submission delays) can give a wrong representation of the real world. A sudden drop is observed at the end of time series constructed from the presently known records. The missing records cause a downward bias in the time series. Figure 1.2 shows an example financial time series, observed at the end of January 2016 (Figure 1.2a) and January 2018 (Figure 1.2b). The bias is especially clear in the last three points in the two plots of Figure 1.2 because the time series drops to zero. This behavior renders the correctness of possible interpretations on the data debatable.

The objective of this thesis is to make a projection of the series in such a way that useful interpretations can be made (e.g. calculate the year-over-year growth rate per month). The desired output is a time series with no downward bias. This means that the data points in 2016 in Figure 1.2a would be projected in such a way that they approach the data points of 2016 observed two years later, depicted in Figure 1.2b. This projection can be produced by financial *now-casting*. This term is a contraction of

[1] Exact offers more products, for the complete list visit `https://www.exact.com/products`.

[2] Some of the KPIs are publicly displayed and can be seen at `https://www.exact.com/nl/over-ons/mkb-monitor` (At the time of writing).

[3] Exact published a (Dutch) white paper with clarifications of the motivation and displayed KPIs: `https://files.exact.com/static/web/downloads/NL-OTH-Whitepaper-Exact-MKB-Monitor.pdf`

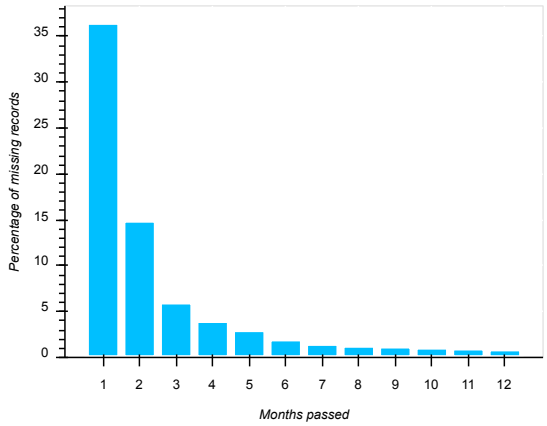
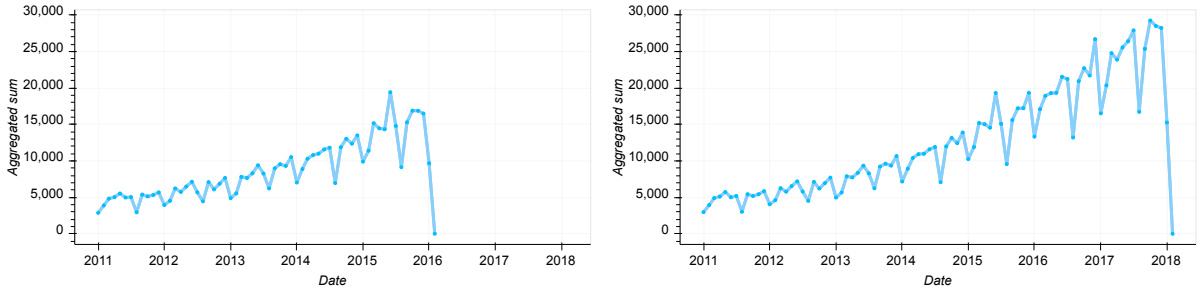


Figure 1.1: Average distribution of missing records as a function of time. The distribution can differ for records from other sectors or dates. This figure serves for the purpose of giving a general intuition.

now and *forecasting*. *Now-casting* is the prediction of the present, the very near future and the very recent past in economics [2]. In Section 1.2 the difference between *forecasting* and *now-casting* is explained in more detail. The main question of this thesis is how to *now-cast* in such a way that the downward bias is removed. Early reported records might give a premature glance of the time series. Many other background indicators could be used for *now-casting*. However, this study focuses on how much information a financial time series provides for its own *now-cast*. In this research it is studied how this premature view could be used as auxiliary information for the *now-cast*.



(a) Time series as seen from Januari 2016

(b) Time series as seen from Januari 2018

Figure 1.2: Aggregated sum of a selection of records

1.2. Related Work

The growth rate of Quarterly Gross Domestic Product (GDP) is a key indicator for the state of the economy. The GDP is of great importance to decision-makers in governments, central banks, financial markets and non-financial firms [6]. Banks estimate GDP prematurely because GDP is subject to substantial lags of financial publications. A timely and reliable evaluation of economic conditions is a key element in the assessment of the monetary policy stance [9]. Financial statements are published with high delays, therefore the European Central Bank (ECB) publishes preliminary estimates approximately 30 days after the end of the reference quarter [9]. To acquire a real-time estimate of the real GDP, banks use indicators with higher publication frequencies to produce short term *now-casts*.

Generally, *now-casting* is the act of predicting for the pending or just finished period (e.g. quarter). A *now-cast* can be seen as an intra-period forecast, with the information about that period already available, which is used for the forecast [24]. *Now-casting* is different than forecasting in the sense that a *now-cast* uses information available from the pending period which is being *now-casted*, whereas forecasts use the information available to predict subsequent periods from which no information is available yet.

For the estimation of current quarter GDP, usually a mix of indicators is used such as industrial production, unemployment, consumer confidence, stock markets and prices of goods and services [6]. Richardson et al. used Factor Models and machine learning approaches to *now-cast* the GDP growth of New Zealand [20]. The Factor Model combines different indicators with a linear combination. The Factor Model is described by:

$$y_t = \alpha_0 + \sum_{i=1}^{k} \alpha_i f_i + \varepsilon_t \tag{1.1}$$

Where α_0 to α_n are parameters and ε_t is the residual term. f_i are factors obtained using Principal Component Analysis (PCA) on the indicator time series.

Giannone et al. used about 200 macroeconomic indicators to *now-cast* the GPD. A combination of indicator time series from different sources usually increases the forecasting accuracy but comes at the cost of using different types of data being released in a non-synchronous manner and with different degrees of lag and frequency. This results in datasets with a so-called jagged edge [12]. *Now-casting* with a combination of mixed-frequency indicators can be challenging because of the unstructured nature. Bridge equations are used to translate the linear combination of factors to GDP. The information contained in various short-term indicators gets transferred, or bridged, to the coherent structure implied by the National Accounts[4] [13]. In other words, high-frequency time series gets converted to quarterly time series using a dynamic linear equation:

$$y_t = \alpha + \sum_{s=0}^{q} \beta_{i,s} x_{i,t-s} + \varepsilon_{i,t} \tag{1.2}$$

α is a constant and $\beta_{i,s}$ are regression coefficients, q is the number of high-frequency periods that fit in the low-frequency period (e.g. 3 months per quarter). $x_{i,t}$ is the i^{th} indicator at time t.

Mixed Frequency Vector Auto Regressive models (MF-VAR) are more recent approaches which are capable of using many data sources with different reporting- frequencies and lags. Ouwehand used MF-VAR models to *now-cast* Quarterly GDP with Monthly time series. The quarterly time series is modeled as a monthly time series with missing data in the first two months of the quarter [18]. Kalman filters are used to estimate regressors in a model and is capable of working with missing data in time series.

Mixed Data Sampling Regression Models (MIDAS) is yet another model, introduced by Ghysels et al. [11]. MIDAS includes indicators in the regression at their original observation frequency. This approach has the advantage of preserved timing information in the indicators. Different types of MIDAS implementations have successfully been applied to data produced by Portuguese automated teller machines and points-of-sale [8].

The Dynamic Factor Model (DFM) is a technique that models the motions of unobserved factors in a time series. Doz et al. introduced a two-step estimator that combines DFM with a Kalman filter to perform *now-casting* [7, 21]. DFM is generally written as two equations:

$$y_t = \Lambda f_t + \beta x_t + \varepsilon_t, \quad \varepsilon_t \sim N(0, \Sigma) \tag{1.3}$$

$$f_t = \sum_{i=1}^{p} \alpha_i f_{t-i} + \xi_t, \quad \xi_t \sim N(0, I) \tag{1.4}$$

Where f_t are the latent factors. Λ is a matrix of factor loadings[5]. x_t are optional indicators. Σ is a covariance matrix. I is the identity matrix. p is the number of autoregressive factors. Λ, α_i and β are parameters. Equation 1.4 describes the motion of the unobserved factors. The DFM described in Equations 1.3 and 1.4 can be cast into state space form to estimate the parameters with a Kalman Filter. Bańbura et al. and Schiavoni et al. adopted and extended the model from Doz et al. in their studies [1, 21].

Xie et al. *now-casted* electricity prices in Sweden using a SARIMAX model. SARIMAX is short for Seasonal Auto regressive Integrated Moving Average with eXogenous regressors. Different power production sources are used as indicator data for the *now-cast* of electricity prices. This model has a satisfactory performance on the domain [26].

[4] A National Account implements a technique for measuring economic activity of a nation.

[5] A factor loading matrix is a matrix of size $p \times k$ with p observable random variables and k unobserved random variables. The matrix is used to indicate the relationship between each observable- and unobservable random variable.

1.3. Report Outline

In Chapter 2, the data used during this study is described and two new ideas are proposed. This is followed by the experiments in Chapter 3 where the experimental setup is defined and results are shown. The results are interpreted in Chapter 4 with a conclusion, discussion and recommendations for future work. The Appendix contains work that is not a part of the core study but gives additional insight in the topic.

2

Data and Methodology

In this chapter the problem is identified with more details. First, it is described with what kind of data has been worked and how the data is pre-processed. Thereafter a model for *now-casting* is described and explained. Furthermore, a change to the model is proposed.

2.1. Dataset

Exact provides data from their online accounting software for this study. This data comprises bilions of transactions. These transactions are structured as records in a ledger. Together they represent the accounting of over 300,000 Dutch SMEs. These records are submitted in the ledgers with a delay. This delay (or lag) marks the difference between the execution date of a transaction and the date this transaction is submitted as record in the ledger. The month a transaction is executed in, will be referred to as transaction month (or transaction date) later in this report. Different ledgers are used per account to distinguish the purpose of the transaction. Revenue transactions represent sold products which are recorded in the revenue ledger, salary payments are recorded in the salaries ledger, and so on.

Time series can be acquired from these different ledgers. As an example, the revenue ledger can be shown as function of time by using the revenue transactions from a company as time series. To take a step further, a time series can be acquired with the revenue transactions of thousands of companies from the same sector. This represents the revenue of a whole sector. Series like these are used to construct KPIs, such as the year-over-year growth rate[1]. From the provided data, time series can be acquired from different sectors and different ledger types. Later in this report, hypothesis are defined and validated. These tests are explained in Chapter 3 using time series with a variety of characteristics. For this reason datasets are generated from the data provided by Exact.

2.1.1. Data acquisition and Pre-processing

The transactions are subdivided in different ledgers for different companies. Every ledger is identified by a Reference Classification System (RCSFI) code[2] and company. The timestamps of all transactions in the ledgers are partitioned to monthly intervals. All transactions from companies in the same sector are aggregated together by summing the values per month and per sector. Time series are produced on sector level, not for individual companies. Sector time series represent the total sum of the transactions from companies in the same sector. This way, so much transactions are used to construct the time series that the series is much smoother than series constructed from individual companies. Therefore the fluctuations caused by noise in the series are reduced and yearly patterns become more emphasized. From the data, 15 sectors and 7 RCSFI codes are identified. From every sector/RCFI-code combination is a time series constructed. In total, $15 \times 7 = 105$ datasets are obtained.

[1] Year-over-year growth rate compares a statistic for a period with the same period from one year ago.

[2] Reference Classification System of Financial Information, or Referentie GrootboekSchema in Dutch, is a scheme introduced to use standardized codes in bookkeeping, general ledger, profit and loss accounts and balance sheets. Example accounts for these codes are: Salary payment, Revenue, Taxes. More information about RCSFI: `https://www.referentiegrootboekschema.nl/`.

Publicly available data from the Dutch National Institute of Statistics (CBS) is used as a reference in this study to check whether the produced time series are representative. For example, the revenue growth of the construction sector obtained from CBS[3] is compared with the revenue growth time series constructed from the transactions of the construction sector. The time series are compared by means of the Pearson correlation coefficient. Pearson correlation is widely used as a measure of the linear relationship between two variables and can also be used to measure the noise between two signals [3, 4]. The Pearson correlation coefficient of the publicly available data and the constructed time series is about 0.75.

The records that are used to construct time series might contain mistakes. Records of transaction which are accidentally an order of magnitude bigger than they should have been (e.g. an extra 0 at the end), can disrupt the time series. Therefore, outlier removal methods have been explored to increase the correlation. Removing transactions with values outside the Lower- and Upper-limit range[4] shows an improvement in correlation. Removing transactions with even stricter bounds, values outside the lower Quartile and upper quartile, will improve the correlation still. After outlier removal, the correlation coefficient between the two time series has improved to over 0.8. Such a correlation coefficient indicates that the the available data is relatively representative, especially when considering that the data from Exact represents about 20% of all SMEs in the Netherlands. Outlier transactions are ignored during the construction of the 105 datasets to obtain more realistic time series.

2.1.2. Two-dimensional dataset

Each dataset contains all records for a particular sector/RCSFI-code combination. In the dataset, the record submission dates are preserved. Therefore it is possible to consider older versions from the time series. Records submitted after a particular date could be ignored when a time series is constructed. This is effectively changing the present time to some point in the past. The date that the present time is set to, will from now on be referred to as the observation date. Records submitted after the observation date are not used in the time series. Figures 1.2a and 1.2b are examples where the observation date is changed to January 2016 and January 2018 respectively. All the datasets that are produced as described in Section 2.1.1, are two-dimensional: they can be used to construct a time series of transactions for different observation dates. The observation date will be used for a *now-casting* procedure, described in Section 2.3.3.

2.2. Forecasting methods

In this study we consider two general methods to produce a forecast for this projection: time series forecasting using historical data and *now-casting* using indicator data.

2.2.1. Forecasting on historical data

A lot of financial time series from individual companies in our dataset do not always show clear patterns such as seasonality or trend. As an example, monthly salary payments of a company might show a trend, while business purchases might be much more irregular and unpredictable. During preprocessing of the data as described in Section 2.1.1, the fluctuations caused by noise are reduced and the patterns in the time series are emphasized. The noise and fluctuation is reduced by aggregating financial data of companies from the same sector together, resulting in a dataset with time series which are likely to contain a trend-, seasonal- or cyclical component. These components can be modeled with time series models.

In order to get a representative forecast, a model should be fitted on unbiased data. Less recent data contains almost no bias. As explained in Section 1.1 the bias is caused by the absence of lagged records. When fitting a model from the start of a time series until a point where the bias increases over a certain threshold ω, then there is a risk of generalizing on data that lost relevance due to economic change. Economic change can be caused by many factors. Therefore the economy as a whole is treated as a hidden context[5], which is subject to gradual Concept Drift [23]. Informally, gradual eco-

[3] Publicly available database with year-over-year revenue growth of the Dutch construction sector, provided by CBS: `https://opendata.cbs.nl/statline/#/CBS/nl/dataset/83837ned/table?ts=1575465775727`

[4] Lower-limit and Upper-limit are defined as $Q1 - 1.5IQR$ and $Q3 + 1.5IQR$ respectively, according to Box plot terminology.

[5] A system is considered a Hidden Context if the rule set that would express the outputs of that system is unknown or difficult to uncover [25].

nomic Concept Drift refers to a slow shift in the structure of an economic system [10]. These shifts could for example be caused by governments which are embarking subsidies to shift a market. A model fitted on data from too far in the past, could suffer from a wrong generalization for the present due to possible Concept Drift. A model fitted on data too close to the present suffers from a bias caused by missing records. The optimum is somewhere in the middle.

To be able to train models on unbiased data, it is important to quantify the bias and find a point where the bias becomes insignificant. To do this, the bias threshold ω is defined as a change of 0.1% compared to the observation of the prior month. Records are submitted with delays according to an exponential distribution. In the recent past of a transaction month a lot of records are submitted. whereas further in the past, fewer records are submitted for that transaction month. The amount of records submitted with high delays becomes small enough after 12 to 24 months that the impact of these delayed records on the total observed value less than 0.1%. Over 95% of the time series in our dataset reach below threshold ω between 12 and 24 months. In other words, 24 months after a transaction date, the aggregated sum of records for the transaction date does not change more than 0.1% and is assumed to not change anymore. This is used as a motivation to assume every time series in the dataset is saturated with records, and therefore has a negligible bias, after 24 months. We define the number of months passed since the transaction month e as Δt. $\Delta t = \tau - e$, where τ is the month of the present date. In Section 2.3.2 is further elaborated on the motivation for the choice of 24 months. Throughout this study, it is assumed that for every transaction month e in the data, e is considered unbiased if $\Delta t \geq 24$. In Section 2.3.2 is this assumption used to train models on unbiased data.

2.2.2. *Now-casting* with indicator data

External indicators can be incorporated in the model to improve the accuracy. Indicators are referred to as exogenous data[6]. The time series that is being modeled is referred to as endogenous data. Different external data sources can be used as exogenous data. Ouwehand and Gianonne et al. have observed that the use of relevant indicators play an important role in the stability and accuracy of a *now-cast* [12, 18]. The use of many indicator data-sources for *now-casting* has to be done with caution; too many data-sources used together can lead to over-fitting due to the Curse of Dimensionality. MIDAS and MF-VAR are two techniques which are especially sensitive for this problem, as their design encourages the use of a broad range of data-sources [6, 16].

In this study, the option of multiple external data-sources to improve the stability and accuracy is disregarded. As explained in Section 1.1, this study focuses on how much information early records can give when used as indicator series. Instead, the information from early submitted records are used as indicator data. As explained earlier records are submitted in the ledgers with a delay. These records in the ledgers are used to construct a time series. They also provide some extra information about the time series in two ways: (1) they reflect the punctuality of bookkeepers which can be observed with a delay distribution of the records. This delay distribution is similar to the missing records distribution displayed in Figure 1.1. It is assumed that the punctuality of bookkeepers stays somewhat the same. (2) They also reflect the state of the economy at the transaction date. The latter is interesting because it can serve as an indicator for a *now-cast*. Records with low delays, e.g. records submitted within a month after the transaction date, give a premature glance of the current state of the economy. These low delay records can be used as an indicator time series. In the following sections, this concept is used in the model design.

2.3. Proposed Methods

In the sections below, models addressing the problem of time series with a downward bias are proposed.

2.3.1. Time series Model

A time series model is created and configured for our data. Most of the financial time series in the dataset show a strong yearly auto correlation, this is displayed for one of the datasets with an Auto Correlation Function shown in Figure 2.1.

Some time series also show an Auto Correlation at lags three, six and nine months. This indicates that they have quarterly auto correlations. The yearly auto correlation is the most significant in many

[6] An Exogenous variable is one whose value determined externally and independently of the endogenous variable.

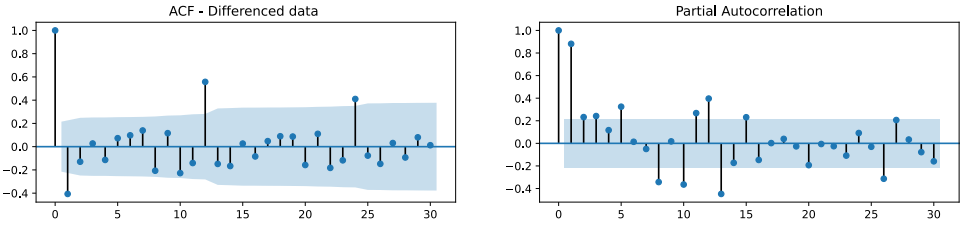


Figure 2.1: Auto Correlation and Partial Auto Correlation of one of the financial time series in the dataset

time series, therefore it is chosen to model two Auto Regressive (AR) parameters and a seasonal component. The seasonal component will be discussed later. An AR(p) process is described as follows:

$$Y_t = \phi_0 + \sum_{i=1}^{p} \phi_i Y_{t-i} + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma_\varepsilon^2) \tag{2.1}$$

ϕ_i are the parameters and p is the AR order. ε_t is a noise term, drawn from an independent and identically distributed normal distribution. Subsequently, Moving Average (MA) terms are incorporated for the lagged prediction errors. An MA(q) process is described as follows:

$$Y_t = \xi_t + \sum_{i=1}^{q} \theta_i \xi_{t-i}, \quad \xi_t \sim N(0, \sigma_\xi^2) \tag{2.2}$$

θ_i are the parameters and q is the MA order.

In time series, seasonality is modeled to capture the variations that occur in every period. This component contributes by modeling the seasonal adjustment [15]. It is a way of modeling in a desea-sonalized fashion, which Pijpers applied for *now-casting* unemployment payments[19]. In this study, a Seasonal component with a period of 12 lags with 2 Seasonal AR parameters is used.

The constructed time series in this study usually have an upward trend which means that the time series is not stationary. A non-stationary time series can be transformed into a stationary one, by taking the difference of the series [27]. For this reason an Integrated component is modeled. The integration part is realized by taking the d^{th} difference of Y_t. For $d = 1$, first order differencing is applied: $Y'_t = Y_t - Y_{t-1}$. Similarly, a seasonal difference $D = 1$ and seasonal period of s lags is specified with: $Y'_t = Y_t - Y_{t-s}$.

Early submitted records are used as indicator data, which form a premature observation of the current state of economy. The indicator time series is treated as exogenous data. From now on we refer to exogenous time series with X_t. The indicator time series X_t contains aggregated sums with only transactions from records that were reported in the same month as they were executed. In other words, X_t is in essence a time series of early submitted records. X_t is used to steer the *now-cast* according to the current economy. The exogenous data contains the initial transactions that happened in the recent past and reveal the presence of a potential economic- depression or growth. X_t can be seen as a time series that contains a subset of all aggregated transactions as a function of time t. It is important to note that X_t is discretized per month, just like Y_t. Therefore adjusting the frequency of X_t to match Y_t's frequency is not needed. This renders Bridge Equations unnecessary.

To illustrate the operation of X_t, an ARMAX(p, q) process is shown:

$$Y_t = \psi X_t + \sum_{i=1}^{p} \phi_i Y_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma^2) \tag{2.3}$$

ψ, ϕ_i, θ_i and σ^2 are the parameters in this model. p is the AR order. q is the MA order.

The use of premature observations of aggregated transactions is only effective if the policies driving bookkeeping behavior stay somewhat the same. If this would not be the case, then the relation between X_t and Y_t would suddenly change. This might lead to a wrong generalization, resulting in erroneous estimations. As an example, if bookkeepers would stop reporting records throughout the year and report almost all records at the end of the year instead, then the modeled behavior is not applicable anymore and wrong estimations will be made for Y_t with respect to X_t. Therefore the bookkeeping behavior is assumed to stay somewhat the same or changes gradually so generalizations can be made.

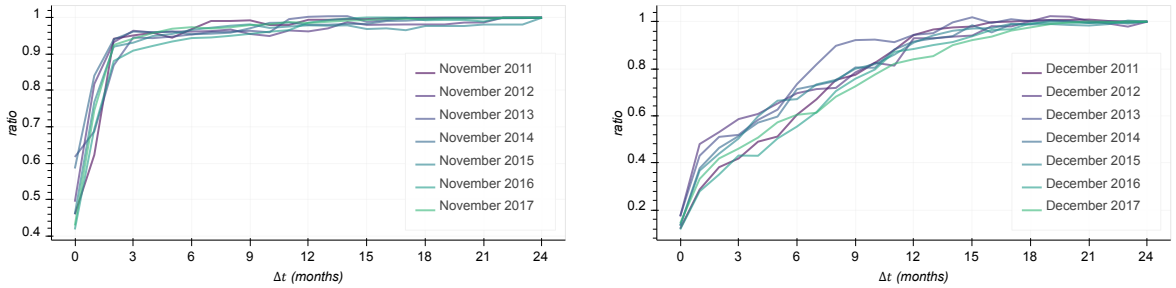
The cash flow of a month depends on the number of day per month and the number of working days per month. For simplicity reasons, the number of days per month are assumed to be constant across time.

With a combination of the components S, AR, I, MA and X described earlier, a SARIMAX model is created. This model is suited for time series from the datasets and can anticipate its predictions on economic turbulence due to the influence of X_t. A SARIMAX model is specified with: $(p,d,q) \times (P,D,Q)_s$. The orders of AR, I and MA are specified with p, d and q respectively. The seasonal- AR, I and MA orders are specified with P, D and Q respectively.

This model forms the basis of this study. New additions that are proposed in the following sections will extend on this model.

2.3.2. Extended model

Indicator time series used in Equation 2.3 can be very reliable if X_t is a process that shows a somewhat consistent relation with respect to Y_t. The relation of X_t to Y_t however, is time-varying in many datasets. Figure 2.2 depicts two charts of the A_t/Y_t ratio as a function of the time passed Δt. Where A_t is the aggregated sum of records for transaction month e submitted between e and $e+\Delta t$. With $e \in E$ where E is the set of transaction dates: all Novembers from 2011 to 2017 in Figure 2.2a and all Decembers 2011 to 2017 in Figure 2.2b. As the observation time progresses (Δt goes up), more records have become available and at $\Delta t = 24$: $Y_t = A_t$. The distribution of lags from the submitted records are somewhat consistent throughout the years. These two charts are plotted from the same dataset, but they show that the lag distribution differs per month of the year. As an example: In November 2017 (Figure 2.2a) at $\Delta t = 0$, the sum of available records is ~ 0.42 times the sum for November 2017 $\Delta t = 24$. While one month later, December 2017 (Figure 2.2b) at $\Delta t = 0$, the sum of available records is only ~ 0.12 times the sum for December 2017 $\Delta t = 24$.



(a) Ratio of all November months from 2011 till 2017

(b) Ratio of all December months from 2011 till 2017

Figure 2.2: The ratio A_t/Y_t as a function of Δt for November and December. A_t is the sum of records available after Δt.

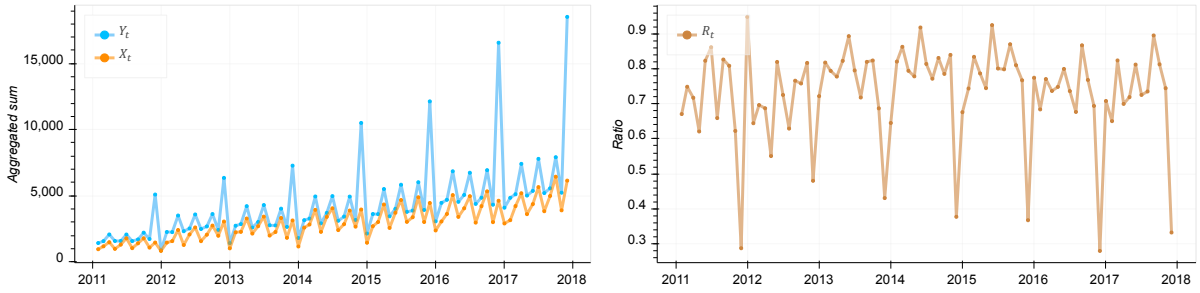
Figure 2.2 shows how the punctuality of the records can vastly differ. Let us now change the perspective and consider what implication this has on the information carried by the exogenous data. In Section 2.3.1 is described that X_t is a time series obtained from aggregated records reported between transaction month e and $e + \Delta t$. The behavior of X_t, constructed with records from e to $e + 1$, as a function of all transaction months e, is illustrated in Figure 2.3a. The graph shows that the December months of Y_t are distinct compared to the other months of the year. These distinct data points are present in Figure 2.3b as a dip. The data points of the December months in Figure 2.3b are the same as the points at $\Delta t = 1$ in Figure 2.2b.

These patterns observed in the data can be used to better utilize the information. Information utilization is defined as: how much information is contributing to the prediction of the *now-cast*. If a model is better capable of utilizing the available indicator information, then the accuracy of the *now-cast* should increase.

It is inviting to incorporate this time-varying relation as an extension on the earlier discussed SARIMAX model. The metric used to measure utilized information is explained in Section 3.1.1. To test whether these additions increase the information utilization, a hypothesis is defined:

Hypothesis 1. *Modeling a time-varying relation between exogenous- and endogenous time series for now-casting, improves the information utilization. With exogenous data being a time series constructed with early submitted records.*

To validate Hypothesis 1, two processes are proposed which model a relation. They are described below.



(a) Y_t and X_t as a function of t.

(b) The relation $R_t = X_t/Y_t$ as function of time t.

Figure 2.3: The relation between Endogenous- (Y_t) and Exogenous (X_t) time series. X_t is the sum of records available at $\Delta t = 1$

The relation between X_t and Y_t can be modeled as a process. Inspired by Factor Models, a linear combination of the time series X_t can be used to model the time-varying relation between X_t and Y_t. Let us define the relation R_t as $R_t = X_t/Y_t$. R_t is modeled as an AR process and is defined in Equation 2.4. ψ_i and σ^2 are the parameters and p' is the AR order. $p' = 12$ could be a suitable order to capture the seasonality, because the time-varying behavior is yearly.

The process from Equation 2.4 with estimated parameters is used for the process that models Y_t in Equation 2.5. This equation is an ARMAX model with parameters Ψ, ϕ_i, θ_i and σ_ξ^2. R_t is used as exogenous data. The ratio R_t is used to remove the time variance of X_t. This is done through: X_t/R_t and can be seen in Equation 2.5.

$$R_t = \sum_{i=1}^{p'} \psi_i R_{t-i} + \varepsilon_t, \quad \varepsilon_t, \sim N(0, \sigma_\varepsilon^2) \tag{2.4}$$

$$Y_t = \Psi \frac{X_t}{R_t} + \sum_{i=1}^{p} \phi_i Y_{t-i} + \sum_{i=1}^{q} \theta_i \xi_{t-i} + \xi_t, \quad \xi_t \sim N(0, \sigma_\xi^2) \tag{2.5}$$

The formulas described with Equation 2.4 and 2.5 are capable of modeling financial time series in the dataset with a multiplicative time varying-relation between the endogenous and exogenous data. Instead of the ratio R_t, also the difference can be used to model additive exogenous time variance. This would be useful if the time variant relation can better be described by a time series representing the difference between Y_t and X_t. Therefore $D_t = Y_t - X_t$ is introduced as an alternative to R_t. This would give two new but similar equations:

$$D_t = \sum_{i=1}^{p'} \psi_i D_{t-i} + \varepsilon_t, \quad \varepsilon_t, \sim N(0, \sigma_\varepsilon^2) \tag{2.6}$$

$$Y_t = \Psi(X_t + D_t) + \sum_{i=1}^{p} \phi_i Y_{t-i} + \sum_{i=1}^{q} \theta_i \xi_{t-i} + \xi_t, \quad \xi_t \sim N(0, \sigma_\xi^2) \tag{2.7}$$

The equations 2.4 and 2.6 describe an AR(p') process and Equations 2.5 and 2.7 describe an ARMAX(p, q) process. These processes can extended with the Seasonal and Integrated component to better fit on the data used in this study.

These models are used in Section 3.1 to validate Hypothesis 1.

2.3.3. Iterative Method

The model described in the previous section can be extended even further. The exogenous data X_t gives a premature view of Y_t. Records of various delays can be incorporated as indicator time series. The use of information available from records submitted with higher lags could increase the information utilization during *now-casting*. The reasoning for this is explained in this section.

The data-source used for the exogenous time series has an interesting property: it can be constructed with more delayed records. For every time step (Δt) of new information, an exogenous time series can be obtained. For higher Δt values, more records are available. Therefore exogenous time series X_t that incorporates records with higher lags are assumed to contain more information about Y_t.

Giannone et al. observed from forecasting errors on their data that new information has a monotonic and negative effect on the forecasting uncertainty [12]. The data used in this study might also show a negative effect on the uncertainty with new information from higher lagged records. A hypothesis is defined to test this:

Δt	Jan 2018	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan 2019
12	×												
11	×	×											
10	×	×	×										
9	×	×	×	×									
8	×	×	×	×	×								
7	×	×	×	×	×	×							
6	×	×	×	×	×	×	×						
5	×	×	×	×	×	×	×	×					
4	×	×	×	×	×	×	×	×	×				
3	×	×	×	×	×	×	×	×	×	×			
2	×	×	×	×	×	×	×	×	×	×	×		
1	×	×	×	×	×	×	×	×	×	×	×	×	
0	×	×	×	×	×	×	×	×	×	×	×	×	×

Table 2.1: Records submitted over time as seen from January 2019, crosses indicate submitted records.

Hypothesis 2. *The use of delayed records as indicator data for now-casting increases the information utilization.*

To validate Hypothesis 2, a method is designed that uses records submitted with higher delays. This is described below.

Table 2.1 illustrates how much months of records are available after Δt months passed. As an example: in January 2019, 12 months have passed since January 2018 and therefore records for January 2018 with a lag of up to 12 months have been recorded. Hence, predictions for January 2018 with 12 months of data will be much more accurate than predictions for January 2019 with only one month of data.

To exploit this property, *now-casts* can be made with exogenous data that contains more information. This is achieved by constructing X_t using records submitted with higher lags. This approach is limited by the fact that records with high lags have not been submitted for the months in the recent past. Older months contain more information, but might be less representative for the economy at present date. Many factors influence the economy and is therefore treated as a hidden context. Let us consider this ballpark example: the economy of January 2018 is likely to be similar to the economy of February 2018, but the economy might gradually change after one or more years, which means that the economy of January 2018 might not be similar to the economy of January 2019.

To summarize: older months used for the exogenous time series give a less biased and therefore more representative view of the economy at that time, because many records are known now. Older dates give less insight in present day economy compared to the months in the recent past. Months from the recent past have more missing records and therefore have a downward bias. For this reason a procedure is designed that uses both a recent- and unbiased view of the economy. This is done by exploiting these properties by means of iteratively *now-casting* for each step Δt with $\Delta t \in \{0, 1, \cdots, 23\}$ [7]. The procedure starts at $\Delta t = 23$ and goes down with each step. Each step uses less information and therefore *now-casts* become less accurate. From now on, this procedure is referred to as the iterative *now-cast*.

The iterative *now-cast* procedure uses a matrix M that consists of cumulative aggregated sums of records. M is structured with the same intuition as Table 2.1. M is a lower triangular matrix of size $k \times k$, with k the number of time steps. Every column represents the date in which a transaction from a record happened. Every row represents the total sum of aggregated records which are submitted within i months, starting with $i = 0$ at the bottom of the matrix. A cell at row i and column j in M is referred to as $m_{i,j}$. Cells in M above the diagonal contain a Not-a-Number value, indicated with a dash ($-$). Cells are cumulative, meaning that $m_{i,j}$ equals the aggregated sum of the new records plus $m_{i-1,j}$. Accessing a complete row i of M is done using the colon symbol (:), $m_{i,0:k}$. Accessing a range of cells from the p^{th} column to q^{th} column on row i is done with: $m_{i,p:q}$. The matrix M is shown in Equation 2.8.

[7] Take note that $X_t = Y_t$ if X_t contains records all with a delay $\Delta t \leq 24$.

$$M = \begin{bmatrix} m_{k,1} & - & \cdots & - & - \\ m_{k-1,1} & m_{k-1,2} & \cdots & - & - \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ m_{2,1} & m_{2,2} & \cdots & m_{2,k-1} & - \\ m_{1,1} & m_{1,2} & \cdots & m_{1,k-1} & m_{1,k} \end{bmatrix} \tag{2.8}$$

The iterative *now-cast* procedure is described in Algorithm 1. The procedure makes a one-point prediction in every iteration. n is the number of observations (the length of the time series to train the model on). r is the total number of iterative steps. As assumed in Section 2.2.1, the bias becomes negligible after two years. Therefore, the total number of iterative steps is 24 (months). At line 6, f is a function that takes the endogenous- and exogenous time series and returns an instance of a time series model, for which the parameters are fitted to the data with the `fit()`-function. Various time series models can be used, this will be discussed in Section 3.2.1. The `model.predict()`-function at line 7, predicts $n\_predict$ steps ahead and requires exogenous data to make predictions. Line 8 writes the prediction to the matrix M, the prediction will be used in later iterations as endogenous data. At line 9, n is incremented to increase the time series length for fitting the model with every iteration.

Algorithm 1 Iterative *Now-cast*

1: **procedure** iterativeNowcast(M, n, r)
2: **for** $(i = 0; i < r; ++i)$ **do**
3: $Y_{0:n} \leftarrow m_{r,0:n}$
4: $X_{0:n} \leftarrow m_{r-i-1,0:n}$
5: $X_{n+1} \leftarrow m_{r-i-1,n+1}$
6: model $\leftarrow f(Y_{0:n}, X_{0:n})$.fit()
7: $\hat{Y}_{n+1} \leftarrow$ model.predict($n\_predict \leftarrow 1$, $exog \leftarrow X_{n+1}$)
8: $m_{r,n+1} \leftarrow \hat{Y}_{n+1}$
9: $n \leftarrow n + 1$
10: **return** $m_{r,0:n}$
11:
12: $n \leftarrow 60$ *// Number of observations, example initialization*
13: $r \leftarrow 24$ *// Total number of iterations, Two years*
14: $\hat{Y}_{0:60+24} \leftarrow$ iterativeNowcast(M, n, r)

The procedure in Algorithm 1 is used to *now-cast* from n to $n + r$, where $n + r$ is the present time. The output of the iterative *now-cast* procedure is a time series $\hat{Y}$ with the intention to have removed the bias due to missing records. An example *now-cast* from the procedure is shown in Figure 2.4. In the figure, Y_t and *now-cast* $\hat{Y}_t$ are slowly deviating halfway the prediction. At that point $\hat{Y}_t$ starts making a difference because the portion of missing records becomes significant enough to see. Whether the

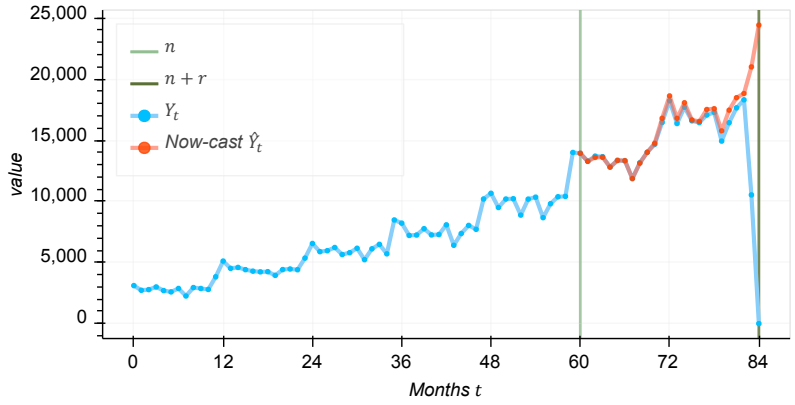


Figure 2.4: An example *now-cast* from the iterative *now-cast* procedure

iterative procedure is better able to utilize information can not be judged from Figure 2.4, this will be analyzed in Section 3.2.

3

Experiments

Two experiments are performed to analyze how the proposed *now-casting* models perform on financial data. The first experiment tests if the information can be utilized in a more effective way. Modeling the relation between exogenous- and endogenous time series might improve the efficiency of information utilization. Two relations are considered in this experiment which are described in Section 2.3.2: the additive relation and the multiplicative relation. The information utilization is assessed by measuring the accuracy of the model. This will be explained in more detail in Section 3.1.2.

In the second experiment is tested if highly lagged records can be incorporated in the *now-cast* and provide some extra information. The iterative procedure provides means for utilizing records with higher delays. The contribution of using records with higher delays is assessed by means of the prediction accuracy.

3.1. Time-varying relation experiment

This experiment aims to verify the validity of Hypothesis 1, defined in Section 2.3.2.

3.1.1. Setup

The goal of the experiment is to analyze the contribution of modeling time variant exogenous behavior. To do so, different models are tested on various financial datasets. The first model is a SARIMAX $(2,1,2) \times (1,1,2)_{12}$ without time varying X_t. This model is is shown as #1 in Table 3.1. As second and third model, the Equations 2.5 and 2.7 with additional Seasonal and Integrated components are used. These models are explicitly capturing the time variant behavior of the exogenous time series in either R_t or D_t. All three SARIMAX models use 9 free parameters with the earlier specified configuration and have the free parameters assigned as follows: 2 for AR, 2 for MA, 1 for seasonal AR, 2 for seasonal MA, 1 for exogenous data and 1 parameter for the state covariance[1]. The two models are shown as #2.1 and #3.1 in Table 3.1. Model #2.1 and #3.1 both use the relation modeled with R_t and D_t respectively. These two processes are modeled by a SARIMA $(2,1,2) \times (1,1,2)_{12}$, without exogenous data. R_t and D_t are estimated with 8 free parameters. The models are used to simulate the process R_t and D_t, this yields time series $\hat{R}_t$ and $\hat{D}_t$. These series are used as exogenous data in #2.1 and #3.1 respectively. Table 3.1 shows these additional processes as #2.2 and #3.2.

#	Model(endog. data, exog. data)	Explicitly model time variance X_t	N^o parameters
1	SARIMAX (Y_t, X_t)	No	9
2.1	SARIMAX (Y_t, $X_t/\hat{R}_t$), with $\hat{R}_t$ from #2.2	Yes, multiplicative	9
2.2	SARIMA (R_t), with $R_t = X_t/Y_t$		8
3.1	SARIMAX (Y_t, $X_t + \hat{D}_t$), with $\hat{D}_t$ from #3.2	Yes, additive	9
3.2	SARIMA (D_t), with $D_t = Y_t - X_t$		8

Table 3.1: Models used in experiments.

[1]The state covariance is a covariance matrix for the current state and next state of a system estimated by a Kalman Filter.

The data provided by Exact is used to compare the models. The financial data ranges from 2011 to the end of 2018 and is separated into many time series. With the pre-processing method explained in Section 2.1.1, time series for different sectors and different RCSFI-codes are obtained. A total of 105 datasets are acquired. From these datasets the endogenous data Y_t is selected and three different exogenous time series: X_t containing all records for transaction months $e \in E$ with a report date from e to $e + \Delta t$, with $\Delta t \in \{1, 2, 6\}$ and $E = \{$January 2011, February 2011, $\cdots$, November 2018, December 2018$\}$. 315 datasets are harvested, of which 3 datasets were not usable. These 3 datasets have a low record density which resulted in time series with empty partitions. In total 312 problem instances (datasets) are used in this experiment. From every dataset the time series Y_t and X_t are obtained. These are used as training data for the models with $n = 60$ observations. Every model has access to: $Y_0, \cdots, Y_{n-1}$ and $X_0, \cdots, X_\tau$ as training data. The models have to *now-cast* 24 data points: $Y_n, \cdots, Y_\tau$, with τ as the observation date.

3.1.2. Error Analysis

The models are assessed on the *now-casting* accuracy. The accuracy is measured with the Mean-Absolute-Scaled-Error (MASE) and normalized Root-Mean-Squared-Error (nRMSE). In the field of time series forecasting, MASE is one of the metrics that is widely used for time series with different scales, such as for method comparison in the M4-competition [17].

MASE scales the error down by the magnitude of fluctuation in Y_t, this makes comparisons between *now-casts* of time series with different scales and fluctuations possible [14]. MASE is shown in Equation 3.1, with Y_t the time series and $\hat{Y}_t$ the predicted time series. A model is assessed from its first prediction, at time n, to the last prediction, at time τ, with $\tau = r + n$ and $r = 24$.

$$\epsilon_{MASE} = \frac{\frac{1}{r}\sum_{t=n}^{\tau} |Y_t - \hat{Y}_t|}{\frac{1}{\tau-1}\sum_{t=2}^{\tau} |Y_t - Y_{t-1}|} \tag{3.1}$$

The nRMSE measure is used to normalise for different time series scales. nRMSE measure scales the error down with the mean value $\bar{Y}$ of Y_t. nRMSE is in contrast with MASE disregarding the fluctuations in the time series [22]. It is shown in Equation 3.2.

$$\epsilon_{nRMSE} = \frac{1}{\bar{Y}}\sqrt{\frac{1}{\tau - n}\sum_{t=n}^{\tau}(Y_t - \hat{Y}_t)^2} \tag{3.2}$$

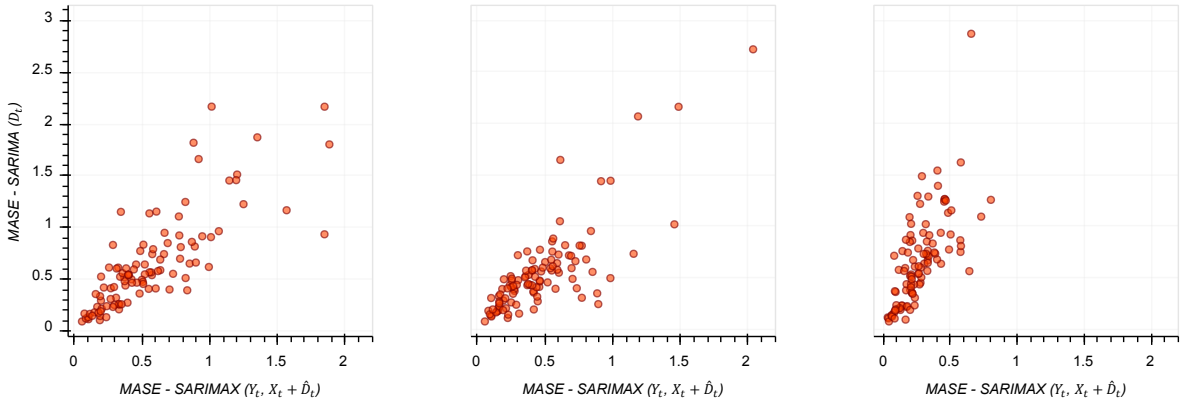
3.1.3. Results

The accuracy results are too extensive to show all in one table, therefore the results are presented in a summarized fashion. Table 3.2 shows the average values of the *now-casting* accuracy for the three models. The results are averaged per different exogenous Δt values. All measurements are shown with the standard deviation. The standard deviation of the accuracy measures MASE and nRMSE is relatively high for all models. The table also shows that no model outperforms the other two models. Models #1 and #3.1 have a competing performance while model #2.1 scores somewhat worse. From Table 3.2 can be seen that a more saturated exogenous dataset positively contributes to lower error measures for #1, #2.1 and #3.1. In other words: as time passes (Δt goes up), the uncertainty goes down.

The models #2.1 and #3.1 are both dependent on the quality of the modeled processes R_t and D_t respectively. In other words, if by accident #2.2 or #3.2 does not manage to fit its parameters properly and produces an $\hat{R}_t$ or $\hat{D}_t$ which can be considered worthless, then model #2.1 or #3.1 will suffer from it. Figure 3.1 shows the MASE of process #3.1 versus the MASE of process #3.2 for the three different dataset groups: $\Delta t \in \{1, 2, 6\}$. Every dot in the figure resembles a sample from the dataset and is located according to the accuracy of #3.1 and #3.2. The dependence of #3.1 on #3.2 can be seen by the correlation between the errors from #3.1 and #3.2 in the three sub-figures. The average accuracy increases for both #3.1 and #3.2 with higher Δt datasets. Similar patterns are observed for processes #2.1 and #2.2. The accuracy of #2.2 displayed in Table 3.2 is relatively bad, this negatively impacts the accuracy of #2.1. A paired T-Test is performed to determine if the accuracy of model #3.1 is improved relative to model #1. The following hypotheses are defined:

Dataset	#	Model	MASE		nRMSE	
$\Delta t = 1$	1	SARIMAX (Y_t, X_t)	0.589	$\pm$0.47	0.117	$\pm$0.078
	2.1	SARIMAX $(Y_t, X_t/\hat{R}_t)$	0.730	$\pm$0.55	0.172	$\pm$0.213
	2.2	SARIMA (R_t)	1.170	$\pm$1.89	0.276	$\pm$0.728
	3.1	SARIMAX $(Y_t, X_t + \hat{D}_t)$	0.585	$\pm$0.45	0.116	$\pm$0.080
	3.2	SARIMA (D_t)	0.672	$\pm$0.53	0.133	$\pm$0.093
$\Delta t = 2$	1	SARIMAX (Y_t, X_t)	0.484	$\pm$0.31	0.100	$\pm$0.061
	2.1	SARIMAX $(Y_t, X_t/\hat{R}_t)$	0.584	$\pm$0.46	0.144	$\pm$0.178
	2.2	SARIMA (R_t)	0.878	$\pm$1.17	0.281	$\pm$0.332
	3.1	SARIMAX $(Y_t, X_t + \hat{D}_t)$	0.471	$\pm$0.32	0.099	$\pm$0.064
	3.2	SARIMA (D_t)	0.543	$\pm$0.41	0.386	$\pm$0.478
$\Delta t = 6$	1	SARIMAX (Y_t, X_t)	0.288	$\pm$0.18	0.061	$\pm$0.041
	2.1	SARIMAX $(Y_t, X_t/\hat{R}_t)$	0.351	$\pm$0.41	0.105	$\pm$0.351
	2.2	SARIMA (R_t)	1.197	$\pm$1.27	0.249	$\pm$1.498
	3.1	SARIMAX $(Y_t, X_t + \hat{D}_t)$	0.296	$\pm$0.20	0.065	$\pm$0.047
	3.2	SARIMA (D_t)	0.758	$\pm$0.70	0.291	$\pm$0.324

Table 3.2: Averages of MASE and nRMSE for all $\Delta t \in \{1, 2, 6\}$.



(a) Dataset: $\Delta t = 1$ (b) Dataset: $\Delta t = 2$ (c) Dataset: $\Delta t = 6$

Figure 3.1: Three scatter plots of accuracy for SARIMAX $(Y_t, X_t + \hat{D}_t)$ versus SARIMA (D_t). Every dot is a sample from one of the datasets.

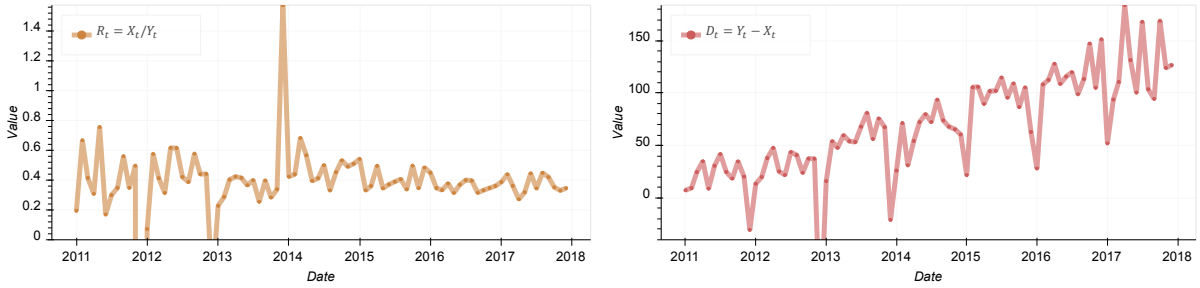
H_0: $\mu_{\#3.1} = \mu_{\#1}$

H_1: $\mu_{\#3.1} < \mu_{\#1}$

$\mu_{\#i}$ is the average MASE error of model $\#i$. **H_0** is not rejected. There is not enough evidence in the data to prove whether modeling a time variance improves the accuracy. Considering the fact that model #1 also requires less parameters to produce similar results, makes model #1 is the preferred model in this experiment.

This experiment is performed to validate Hypothesis 1, defined in Section 2.3.2. Modeling a time-varying relation between exogenous and endogenous data have not been shown to improve the information utilization. With the models and datasets used in this experiment, there is not enough evidence to accept Hypothesis 1, therefore it is rejected.

From the experiments, various *now-cast* accuracies are observed for different problem instances among the models #1, #2.1 and #3.1. This variation in accuracy is caused by the different ways in which the information of X_t is used in the models. For some problem instances, X_t is more informative for a *now-cast* than R_t or D_t. For other problem instances R_t or D_t can be more informative. R_t and D_t are shown in Figure 3.2 from an example problem instance. R_t (in Figure 3.2a) is less informative for a *now-cast* because the relation does not reveal patterns (which makes it hard to model the series). D_t (in Figure 3.2b), on the other hand, shows a yearly seasonality and a trend. For this problem instance, D_t is more useful than R_t when used as auxiliary data in a *now-cast*.



(a) The relation R_t of the problem instance.

(b) D_t of the problem instance.

Figure 3.2: Relations R_t and D_t of a problem instance from the dataset.

3.2. Lagged records experiment

This experiment aims to verify the validity of Hypothesis 2, defined in Section 2.3.3.

3.2.1. Setup

In the second experiment the performance of the Iterative procedure, introduced in section 2.3.3, is tested. f was left unspecified in Algorithm 1. The results in Section 3.1.3 have shown that a SARIMAX (Y_t, X_t) model would be the preferred implementation. For this reason a SARIMAX (Y_t, X_t) model with the configuration as described in Section 3.1.1 is chosen as implementation for f in the second experiment.

The goal of this experiment is to analyze the contribution of incorporating lagged records in a *now-cast*, by means of the iterative procedure. The contribution of lagged records is measured in terms of the *now-cast* accuracy.

In this experiment, 24 time steps are being *now-casted*. For some months of the year it is harder to *now-cast* than other months of the year. Therefore a twelve-iteration forward validation (or walk-forward validation) method is used to weight every month of the year equally. Forward validation is used to measure how well a model generalizes for time series [5]. Every iteration in the forward validation has a training set of 48 or more data point and a test set of 24 data points. The first three forward validation iterations are shown in Table 3.3. The table shows how the Y_τ shifts with every iteration. Referring to the table: the training set is indicated with orange, the test set is indicated with red. The iterative procedure will be tested with 105 datasets, which are obtained as described in Section 2.1.1. With the use of forward validation, in total with 105 datasets and 12 iterations $105 \cdot 12 = 1260$ tests are performed.

Iteration	2011 Jan	⋯	2014 Nov	2014 Dec	2015 Jan	2015 Feb	⋯	2017 Jan	2017 Feb	2017 Mar
1	Y_0	⋯		$Y_{\tau-24}$			⋯	Y_τ		
2	Y_0	⋯			$Y_{\tau-24}$		⋯		Y_τ	
⋮	Y_0	⋯				$Y_{\tau-24}$	⋯			Y_τ

Table 3.3: First three iterations of the forward validation. Orange is the training set and red is the test set.

The accuracy of the *now-cast* is measured by the mean error of the 24 predicted points: from $Y_{\tau-24}$ to Y_τ, with Y_τ the last- and most recent point. Every predicted point has equal weight in the accuracy measure. The prediction of the most recent point Y_τ, usually is the most uncertain because only one month of information from the exogenous data is available. Therefore the accuracy of the prediction of the most recent point Y_τ is of interest and will also be measured as a separate score metric.

Two additional models are compared in this experiment. As second process, a SARIMAX (Y_t, X_t) model predicting 24 steps, with X_t containing records with lags e to $e + 1$ ($\Delta t = 1$). Informally, the second model will use a time series of early submitted records as exogenous data. As third process a SARIMA (Y_t) model is used, which does not utilize the information that X_t provides. The third model in this experiment disregards X_t whereas the iterative procedure and second model do not. With this difference, the contribution of exogenous data is measured.

The accuracy is measured with MASE and nRMSE (described with Equations 3.1 and 3.2).

3.2.2. Results

In total, 1260 time series are tested and the results are summarized in Table 3.4. This table shows the average errors accompanied with the standard deviation. The first MASE- and nRMSE columns show the average error of 24 *now-casted* time steps. The second MASE- and nRMSE columns only considers the accuracy of the most recent point on the same *now-casts*.

		Accuracy of *now-casted* $Y_{\tau-24}$ to Y_τ				Accuracy of *now-casted* Y_τ			
#	Method	MASE		nRMSE		MASE		nRMSE	
1	Iterative Procedure	0.2540	$\pm$0.4543	0.0874	$\pm$0.0935	0.6420	$\pm$1.2663	0.1246	$\pm$0.2106
2	SARIMAX (Y_t, X_t)	0.5894	$\pm$0.4604	0.1887	$\pm$0.1516	0.8458	$\pm$0.9134	0.1645	$\pm$0.2098
3	SARIMA (Y_t)	0.6818	$\pm$0.6745	0.1906	$\pm$0.1569	1.1097	$\pm$1.2181	0.1655	$\pm$0.1314

Table 3.4: Average *now-casting* accuracies, expressed in MASE and nRMSE.

In Table 3.4 can be seen that the Y_τ errors are much higher than the error of $Y_{\tau-24}$ to Y_τ. This aligns with the other experiments and agrees with the assumption that recent points are *now-casted* with more uncertainty. It is interesting to notice that the standard deviation of the most recent prediction errors from the Iterative Procedure is the highest compared to the other models. This is caused by the way the iterative procedure is constructed. In every iteration, it uses its prior predictions. This is done with the idea to utilize the information in X_t as much as possible. The downside of this approach is that prediction errors can be compounded. Therefore this procedure is prone to over- or under predictions. An example of a *now-cast* with an over prediction is depicted in Figure 3.3. The dashed blue line indicates a hypothetical correct *now-cast*.

From the accuracy results of models #2 and #3 in Table 3.4 can be concluded that the use of information available in X_t somewhat positively contributes to the *now-cast*. Even more information in X_t can be utilized if more lagged records are incorporated by means of the iterative procedure. This also increases the accuracy as can be seen with the accuracy measures of #1 and #2 in Table 3.4.

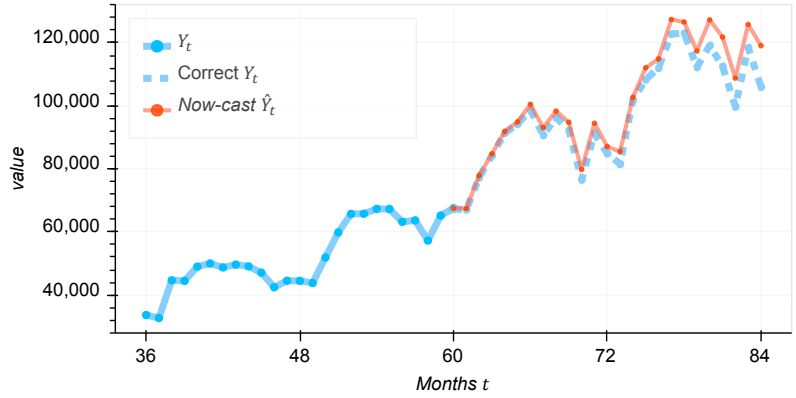


Figure 3.3: An example of an over-*now-cast*

A second Paired T-Test is performed to see if the iterative procedure is able to better utilize records with higher delays. The following hypothesis are defined for the models used in this experiment:

H_0: $\mu_{\#1} = \mu_{\#2}$

H_1: $\mu_{\#1} < \mu_{\#2}$

$\mu_{\#i}$ is the average MASE of the prediction for $Y_{\tau-24}$ to Y_τ, from model $\#i$. From the Paired T-Test a p-value of 7.69^{-110} is obtained, therefore **H_0** is rejected.

This experiment is performed to validate Hypothesis 2, defined in Section 2.3.3. The results show that the use of delayed records as indicator data for *now-casting* increases the accuracy. With this result can be concluded that the information utilization is increased if records are used with higher delays. Hypothesis 2 is accepted.

4

Conclusion, discussion and Recommendations

This chapter presents the conclusions from this thesis research, followed by a discussion about the design choices and the assumptions that are made. The last section gives recommendations for future work.

4.1. Conclusion

In this thesis the problem of incomplete time series due to missing records is addressed. The main problem was the downward bias in the time series as a consequence of lagged records. The bias can be removed by the practice of *now-casting*. Different studies applied financial *now-casting* to estimate the quarterly GDP. An important aspect in these studies was the acquisition of data. Various data-sources are used together as indicator time series to serve as exogenous data for the *now-casts*.

In this study is analyzed how records in a ledger can be used as auxiliary data for a *now-cast*. The lag distribution of the submitted records is generated by the workflows of bookkeepers. This information is used for *now-casting*. Furthermore, it is studied if more information of records can be utilized by modeling the relation between the exogenous and endogenous series. Experiments have been performed to uncover whether modeling the time-varying behavior can improve the utility of the available information. Utility is measured by comparing *now-cast* accuracies of models which use different auxiliary data. In the first experiment three different models are tested, of which one does not use the time-varying relation. The results of the experiments do not show an improved accuracy when modeling the time-varying relation between indicator- and target time series. There is not enough evidence to prove that modeling a time-varying relation increases the information utility. The models that use a time-varying relation for the *now-cast* depend on two processes, which together requires more parameters. It is preferred to not model the relation between endogenous and exogenous, because this addition does not show an improvement. For some problem instances however, the use of an additive- or multiplicative relation results in a better accuracy. This means that some properties, extracted from the indicator data can be used to better utilize the available information in the data. Although modeled relations do not improve the overall accuracy, they can provide some extra information for a *now-cast*.

Records which are submitted recently after the transaction date, can be used to give a premature view of the economy. Records with higher delays can give an even better view of the economy at that time. This motivation is used to analyze whether incorporating lagged records can improve the utilization of the available data to *now-cast*. An iterative fashion is used to *now-cast* with different levels of delayed records. This approach makes more use of the available information than a model that only uses early reported records. Results from the experiments show there is a slight improvement when using more information by means of an iterative *now-cast*. This comes at a cost of possible compounding errors, which makes the method more prone for over- or under predictions. The most recent data point is *now-casted* with the most uncertainty, because for this point the least information is available.

4.2. Discussion

This section will elaborate on some of the design choices made during this project.

4.2.1. Model design

There is not enough evidence to show that modeling a time variant relation between X_t and Y_t increases the amount of utilized information during *now-casting*. It could be possible that the datasets used in this experiment have a less significant time variant relation than initially thought.

In the Section 2.3.2 is described how the time variant relation between X_t and Y_t can be modeled with two processes each. One downside is that two processes are used to model the time variance while only one was needed. A linear combination of X_t and Y_t would reduce the model to only one process. Such a process is described with Equation 4.1.

$$Y_t = \sum_{i=0}^{P} \psi_i X_{t-i} + \sum_{i=1}^{p} \phi_i Y_{t-i} + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma_\varepsilon^2) \tag{4.1}$$

p is the AR order, and P is the exogenous-AR order. ψ_i, ϕ_i, σ_ε^2 are the parameters. This model can be extended by also incorporating the additive or multiplicative relation between X_t and Y_t, as shown in Equations 4.2 and 4.3. Even a mix of additive and multiplicative relation can be used.

$$Y_t = \psi_0 X_t + \sum_{i=1}^{P} \psi_i \frac{X_{t-i}}{Y_{t-i}} + \sum_{i=1}^{p} \phi_i Y_{t-i} + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma_\varepsilon^2) \tag{4.2}$$

$$Y_t = \psi_0 X_t + \sum_{i=1}^{P} \psi_i (Y_{t-i} - X_{t-i}) + \sum_{i=1}^{p} \phi_i Y_{t-i} + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma_\varepsilon^2) \tag{4.3}$$

During the project, the models described in the equations above are implemented in State Space form, which is convenient for the Kalman Filter to fit the parameters of the model. The State Space form of Equation 4.1 was somewhat challenging to implement, but it was realised. This was challenging because the library (Statsmodels[1]) that was used to build the models was not very transparent. It is even more complicated when the Integrated and Seasonal components are modeled in State Space form. These components are difficult to build because they require a big change in the matrices used inside the State Space model. The Seasonal and Integrated components are needed to be able to test on real data, such as the dataset provided by Exact. Something went wrong in the design of the models during the developments, because the parameters of the State Space equivalent of a SARIMA with Auto Regressive X were not able to converge properly. It took too much time to correct the mistakes, therefore an alternative solution was tried. The alternatives are the processes described with Equations 2.4, 2.5, 2.6 and 2.7.

One of the important questions which is still left unanswered is whether $\sum_{i=0}^{P} \psi_i X_{t-i}$ in Equation 4.1 is able to capture a time-varying relation such that the Equations 4.2 and 4.3 are not needed. Section A.1 provides an artificial scenario in which Equation 4.1 would not be capable of properly modeling without the relation. The scenario is crafted with artificial data, and therefore does not prove the idea. It is unfortunate that this question can not be answered at the end of this project.

In this study, statistical models similar to SARIMAX are analyzed and tested. To gain more insight, a comparison with other models from the *now-casting* literature should be performed. The use of different techniques would allow to extract other information from the indicator data. As an example: Bridge Equations can be used discretize the indicator data with smaller time intervals. This means that the date property of the records have higher resolution and therefore less information is lost.

4.2.2. Assumptions

A few assumptions had to be made for the *now-casts* on the data. This section briefly goes through them and discusses the consequences of these assumptions.

As mentioned in Chapter 1, bookkeepers have different record submission punctualities. The punctuality of bookkeeping influences the value of indicator X_t used for the *now-casts*. As an example, if bookkeepers would suddenly be less punctual, then X_t observed at $\Delta t = 1$ would change. During model fitting it is assumed that bookkeepers will remain about as punctual as they were previously. Of course bookkeepers can differ with their punctuality from time to time, but because the time series involve so much different companies, the punctuality distribution will smooth out. It is also assumed

[1] Statsmodels is a Python library for statistical modeling. Documentation about State Space of Statsmodels: `https://www.statsmodels.org/stable/statespace.html`

that the overall punctuality might gradually change. The punctuality can be impacted in many ways, such as system down-time for a long period that causes bookkeepers to submit records at a later time. In this study, these domain specific factors are not taken into account. It is also assumed that economic turbulence will hardly impact the punctuality.

Monthly cash flow is influenced by the number of days per month and the number of working days per month. All months are assumed to have the same number of days and the same number of working days for the sake of simplicity. In reality, on average every month has about 21 working days. The number of working days per month can vary from 18 days (in April 2017) to 23 days (in March 2017) [2].

During the pre-processing described in Section 2.1.1 it is assumed that some bookkeeping mistakes can disrupt a *now-cast*. Records of transaction which are accidentally an order of magnitude bigger than they should have been because of an extra 0 at the end, can have a major impact on the time series. It is not investigated how these mistakes can be detected and therefore it is chosen to strip the outliers from the dataset. Outliers which are benign are also stripped from the dataset, this is the price that is payed to reduce the overall noise in the dataset.

4.3. Recommendations

There are still some directions left to explore due to limited time and scope of this thesis. In this section, possible future steps will be recommended. Some of these recommendations are follow-ups from the research done in this project.

It is observed in the data of this study that after 24 months the downward bias is vanished, because almost all the records are submitted by that time. Therefore, for all the time series in the dataset the *now-casting* starts 24 months before the present. Beginning a *now-cast* 24 months before the present is not necessary for all datasets. How long it takes before enough records have been submitted for the bias to disappear can differ per time series and per month of the year. If this would have been measured, then variable starting times could have been used for *now-casting*. It would be preferred to start a *now-cast* closer to the present if the downward bias vanishes early. This gives less overhead and a *now-cast* starting point which might better reflect the present economy.

The use of more indicators is an interesting direction to explore. More relevant indicators for *now-casting* can be obtained either from the dataset provided by Exact or external economic data. Various data sources might improve the accuracy, but more importantly, opens up a more techniques which can be explored. For example, the Factor Model or Dynamic Factor Model described in the Related Work, could be extended to additionally model the relation between endogenous data and exogenous indicators.

The data that is used for this study is discretized per month. This means that the submission date and transaction date of each record is partitioned in intervals of months. This is convenient because it is the same frequency as the KPIs that need to be *now-casted*. This results in some loss in information due to the discretization. The records could be partitioned in higher frequency periods, such as weeks or days. It is also possible to not discretize at all. To be able to still *now-cast* on a monthly frequency Bridge Equations could be used. The use of models designed for mixed frequency data will become more inviting to use.

In the data that is used were also records submitted with future transactions. The records with future transactions are expected to occur in the future, usually because they happen on a regular basis, like fixed costs. These future transactions give an insight in the foresighted cash flow from the bookkeeper's perspective. The foresights are disregarded during this study but might be able to contribute in a *now-cast*.

[2] Calendar of working days for the Netherlands: `http://www.vakantiespreiding.eu/aantal-werkdagen/`.

A

Further Notes

A.1. Experiments on artificial data

This section illustrates the possible contribution of modeling a time varying relation. Artificial time series are used to show how the principle works in simple circumstances. In this section is shown how a model, that would be able to capture a time varying relation between exogenous and endogenous data, would be designed to (almost) perfectly fit on the data.

In Figure A.1 is a time series Y_t displayed with an auto regressive lag of 3. In this artificial setup, Y_t has no noise and follows this signal: $\{4, 3, 7, 4, 3, 7, \cdots\}$. An AR(3) process for which its parameters are fitted on time series Y_t is displayed with the green line in Figure A.1. The AR(p)-process is described in Equation A.1, with ϕ_i free parameters. The model is trained on Y_t from $t = 0$ to $t = 29$. For $t \geq 30$ the model forecasts subsequent points. The dashed line starting from $t = 30$ for Y_t in Figure A.1 indicates that this is not shown to the model. The AR(3) model is able to (almost) perfectly reproduce and forecast Y_t.

$$Y_t = \sum_{i=1}^{p} \phi_i Y_{t-i} \tag{A.1}$$

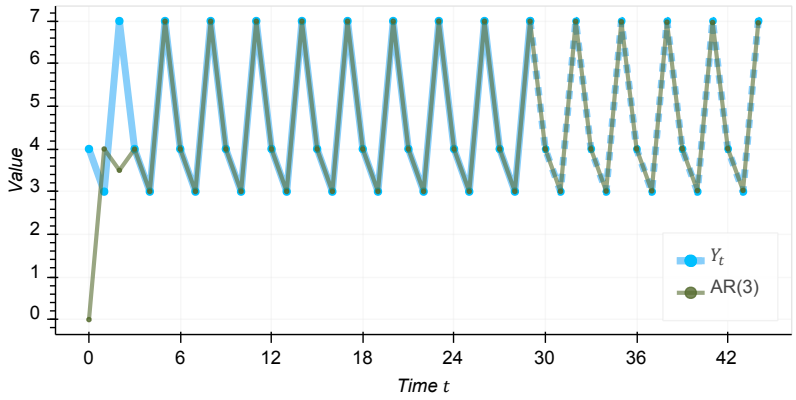


Figure A.1: Time series Y_t and a fitted AR(3)-process

An external factor X_t can be observed. X_t has 4 auto regressive lags and is described by: $\{1, 0.5, 1.5, 0.75, 1, 0.5, 1.5, 0.75, \cdots\}$. A new time series Y'_t is introduced. Y'_t is influenced by X_t. X_t has a time varying influence on Y'_t. This time varying influence is described with relation R_t. The relation R_t is described by an artificial time series which has 4 auto regressive lags and follows this signal: $\{1, 3, 2, 4, 1, 3, 2, 4, \cdots\}$. In figure A.2 is the new time series Y'_t depicted which is influenced by the external factor X_t according to: $Y'_t = Y_t \cdot R_t \cdot X_t$. The different auto regressive lags of R_t and Y_t cause a more

complex structure in Y_t'. The external influence causes Y_t' to obtain an auto regressive lag of 12. An AR(3)-process would not be able to fit on Y_t'. An AR(12)-process however would be able to (almost) perfectly fit on the training data. An AR(12)-process requires 12 parameters. We can do better in terms of parameters and generalization. A new model is introduced which models an AR- and exogenous-AR (ARXAR) process. an ARXAR(p, P) is described with Equation A.2. An ARXAR(3, 4) is trained on Y_t' with exogenous time series: $X_t \cdot R_t$. 7 parameters are used for this model. It is able to (almost) perfectly fit on the new data Y_t'

$$Y_t = \sum_{i=1}^{p} \phi_i Y_{t-i} \sum_{i=0}^{P} \psi_i X_{t-i} \tag{A.2}$$

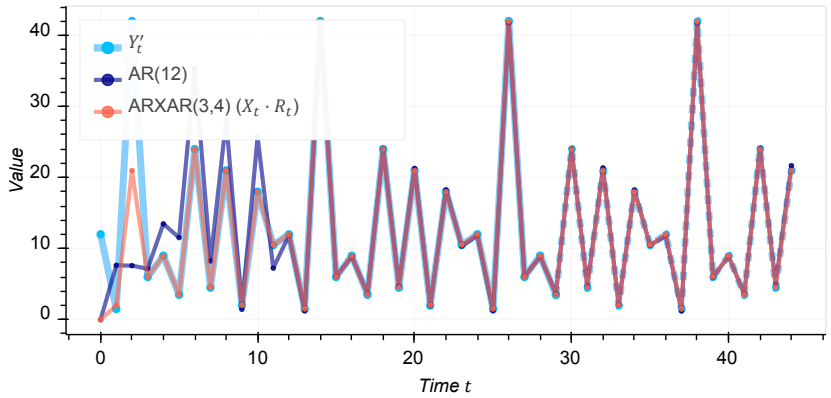
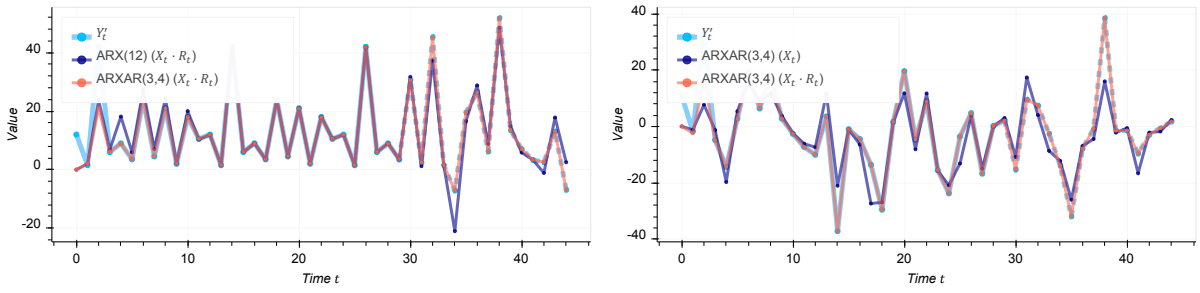


Figure A.2: Time series Y_t' and two processes

The AR(12)-process does not correctly generalize because it has not learned to model the external influence. This can be tested by changing the values of X_t in with $t \geq 30$. To allow the process to anticipate on external changes, an ARX model is introduced. ARX is desribed by Equation A.3. In Figure A.3a is shown that the ARX(12) process is not able to fit properly on the data. An ARXAR(3, 4) requires $X_t \cdot R_t$ as exogenous time series to be able fit on Y_t'.

Let us change X_t to an IID random variable: $X_t \sim N(0, 1)$. Y_t' is still defined as: $Y_t' = Y_t \cdot R_t \cdot X_t$. The ARXAR is still able to obtain a good fit on the data if the relation is as exogenous data. This can be seen in Figure A.3b. An ARXAR(3, 4) with X_t is provided as exogenous data. The model is not able to properly fit.

$$Y_t = \psi X_t \sum_{i=1}^{p} \phi_i Y_{t-i} \tag{A.3}$$



(a) (b)

Figure A.3: Time series Y_t' influenced by external variable X_t

We have seen that Y_t' is influenced by an observable external variable X_t. X_t has a time dependent influence on Y_t'. This time dependent influence is described by R_t which is not directly observable, but

can be uncovered: $R_t = Y'_t/(X_t \cdot Y_t)$. To be able to use relation R_t, it needs to be modeled because it is derived as the relation between X_t and Y'_t, but Y'_t is unknown after $t \geq 30$. The modeled relation R_t and external variable X_t are used together in a model to properly fit on the data.

Bibliography

[1] Marta Bańbura and Gerhard Rünstler. A look into the factor model black box: Publication lags and the role of hard and soft data in forecasting GDP. *International Journal of Forecasting*, 27(2): 333–346, 2011. ISSN 01692070. doi: 10.1016/j.ijforecast.2010.01.011.

[2] Marta Bańbura, Domenico Giannone, Michele Modugno, and Lucrezia Reichlin. Now-casting and the real-time data flow. In *Handbook of Economic Forecasting*, volume 2, pages 195–237. Elsevier B.V., 2013. doi: 10.1016/B978-0-444-53683-9.00004-9.

[3] Jacob Benesty, Jingdong Chen, and Yiteng Huang. On the importance of the pearson correlation coefficient in noise reduction. *IEEE Transactions on Audio, Speech and Language Processing*, 16(4):757–765, 2008. ISSN 15587916. doi: 10.1109/TASL.2008.919072.

[4] Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Noise Reduction in Speech Processing. *Noise reduction in speech …*, 2:229, 2009. doi: 10.1007/978-3-642-00296-0.

[5] Christoph Bergmeir and José M. Benítez. On the use of cross-validation for time series predictor evaluation. *Information Sciences*, 191:192–213, 2012. ISSN 00200255. doi: 10.1016/j.ins.2011.12.028.

[6] Jasper M De Winter. *Nowcasting GDP Growth: statistical models versus professional analysts*. 2016. ISBN 9789491602733.

[7] Catherine Doz, Domenico Giannone, and Lucrezia Reichlin. A two-step estimator for large approximate dynamic factor models based on Kalman filtering. *Journal of Econometrics*, 164(1): 188–205, 2011. ISSN 03044076. doi: 10.1016/j.jeconom.2011.02.012.

[8] Cláudia Duarte, Paulo M.M. Rodrigues, and António Rua. A mixed frequency approach to the forecasting of private consumption with ATM/POS data. *International Journal of Forecasting*, 33 (1):61–75, jan 2017. ISSN 01692070. doi: 10.1016/j.ijforecast.2016.08.003.

[9] ECB. Economic Bulletin, Issue 2 / 2020. *Economic Bulletin*, (2):146, 2020.

[10] Ryan Elwell, Robi Polikar, and Senior Member. Incremental Learning of Concept Drift in Nonstationary Environments. *IEEE TRANSACTIONS ON NEURAL NETWORKS*, 22(10), 2011. doi: 10.1109/TNN.2011.2160459.

[11] Eric Ghysels, Pedro Santa-Clara, and Rossen Valkanov. The MIDAS Touch: Mixed Data Sampling Regression Mod. *Finance*, 2004.

[12] Domenico Giannone, Lucrezia Reichlin, and David Small. Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics*, 55(4):665–676, may 2008. ISSN 03043932. doi: 10.1016/j.jmoneco.2008.05.010.

[13] Thomas B. Götz and Thomas A. Knetsch. Google data in bridge equation models for German GDP. *International Journal of Forecasting*, 35(1):45–66, jan 2019. ISSN 01692070. doi: 10.1016/j.ijforecast.2018.08.001.

[14] Rob J. Hyndman and Anne B. Koehler. Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679–688, oct 2006. ISSN 01692070. doi: 10.1016/j.ijforecast.2006.03.001.

[15] Prajakta S Kalekar. Time series Forecasting using Holt-Winters Exponential Smoothing. Technical report, 2004.

[16] Vladimir Kuzin, Massimiliano Marcellino, Christian Schumacher, and Deutsche Bundesbank. *MIDAS versus mixed-frequency VAR: nowcasting GDP in the euro area*. 2009. ISBN 9783865585080.

[17] Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. The M4 Competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, 36(1):54–74, jan 2020. ISSN 01692070. doi: 10.1016/j.ijforecast.2019.04.014.

[18] Pim Ouwehand. Rapport Nowcasting voor de horeca. 2016.

[19] Frank P Pijpers. Nowcasting using linear time series filters. Technical report, 2018.

[20] Adam Richardson, Thomas Van, Florenstein Mulder, Thomas Van Florenstein Mulder, and Tuğrul Vehbi. Crawford School of Public Policy CAMA Centre for Applied Macroeconomic Analysis Nowcasting New Zealand GDP Using Machine Learning Algorithms Nowcasting New Zealand GDP Using Machine Learning Algorithms *. 2018. ISSN 2206-0332.

[21] Caterina Schiavoni, Franz Palm, Stephan Smeekes, and Jan van den Brakel. A dynamic factor model approach to incorporate Big Data in state space models for official statistics. 2019.

[22] Maxim Vladimirovich Shcherbakov, Adriaan Brebels, Nataliya Lvovna Shcherbakova, Anton Pavlovich Tyukov, Timur Alexandrovich Janovsky, and Valeriy Anatol evich Kamaev. A survey of forecast error measures. *World Applied Sciences Journal*, 24(24):171–176, 2013. ISSN 18184952. doi: 10.5829/idosi.wasj.2013.24.itmies.80032.

[23] Alexey Tsymbal. The problem of concept drift: definitions and related work. Technical report, 2004.

[24] Roy Verbaan and Carin Bolt, Wilko van der Cruijsen. Using Debit Card Payments Data for Nowcasting Dutch Household Consumption. *SSRN Electronic Journal*, apr 2018. doi: 10.2139/ssrn.3047122.

[25] Gerhard Widmer. Learning in the presence of concept drift and hidden contexts. *Machine Learning*, 23(1):69–101, 1996. ISSN 08856125. doi: 10.1007/bf00116900.

[26] Mengchen Xie, Claes Sandels, Kun Zhu, and Lars Nordstrom. A seasonal ARIMA model with exogenous variables for elspot electricity prices in Sweden. In *International Conference on the European Energy Market, EEM*, 2013. ISBN 9781479920082. doi: 10.1109/EEM.2013.6607293.

[27] Eric R Ziegel. Analysis of Financial Time Series. *Technometrics*, 44(4):408–408, 2002. ISSN 0040-1706. doi: 10.1198/tech.2002.s96.