Norwegian University
of Life Sciences

Master's Thesis 2022 30 ECTS
School of Economics and Business

Short-term load forecasting in times of unprecedented price movements

Kevin Gulliksen Berghagen &
Vegard Hågård Hoch-Nielsen
Business Administration

Preface

This thesis marks the end of a two-year master's degree in Business Administration with specialization in Finance at the Norwegian University of Life Sciences (NMBU).

The interest in power markets has been a big part of our master’s degree and stems from the increasing importance of renewable electricity production and the electrification of our society. In this thesis we combined this interest with our field of study by researching forecasting methods, useful for the market participants. Throughout these last five months, we have learned a lot about various forecasting methods and the different software options for the building of prediction models. A special thank you to Olvar Bergland for your guidance and valuable feedback throughout this process. We also want to thank our family and friends for their continuous support throughout our years of study.

Kevin Gulliksen Berghagen & Vegard Hågård Hoch-Nielsen

Ås, May 2022

Abstract

In this thesis we aimed to find the best methods for short-term load forecasting in the Norwegian electricity market during times of unprecedented price movements. We answered three questions related to this aim. The first was which model achieved the most accurate forecast. The second was whether our proposed models outperform the official forecasts published on the Entso-E platform. The third question asked was if the price movements had any effect on the accuracy of the load forecast.

We constructed two SARIMAX models, a Gradient boosted decision tree, a Random Forest, and a Multilayer perceptron model. Our findings show the two SARIMAX models to be most accurate. These models outperformed the forecasts published on the Entso-E platform in four out of the five Norwegian bidding zones, measured in MAPE and RMSE. Finally, we have shown that forecasting load with and without price information did not result in significant differences in accuracy. Our findings did not indicate an increase in difficulty of forecasting 2021 compared to 2019, neither for the three southern bidding zones with higher price increase nor the northern two zones.

Sammendrag

I denne masteroppgaven har vi forsøkt å finne den beste metoden for kortsiktig prognostisering av elektrisitets-etterspørsel i perioder med ekstreme prisbevegelser. Vi har besvart tre spørsmål knyttet til denne problemstillingen. Det første var hvilken modell som oppnår høyest nøyaktighet. Det andre var om våre modeller presterer bedre enn de publiserte prognosene på Entso-Es offentlig tilgjengelige data-plattform. Det tredje spørsmålet var om de ekstreme prisbevegelsene hadde noen effekt på nøyaktigheten av prognosene.

Vi har laget to SARIMAX modeller, en Gradient boosting decision tree-, en Random Forest og en Multilayer perceptron-modell. Gjennom arbeidet har vi vist at de to SARIMAX-modellene presterer best. Disse modellene er mer nøyaktig enn prognosene publisert på Entso-Es plattform for fire av de fem norske strømregionene, målt i MAPE og RMSE. Til slutt har vi vist at prognoser gjort både med og uten prisinformasjon ikke gir signifikante forskjeller i nøyaktighet. Det ble heller ikke påvist en klar forskjell i vanskelighetsgraden av å prognostisere 2021 sammenlignet med 2019, verken for de sørlige prissonene med høy prisvekst eller de nordlige sonene med en lavere prisvekst.

Contents

1. Introduction

The power grid and the supply of electricity is one of the highly critical infrastructures of the modern world, and in the years to come the importance of electric power will only grow further. As we aim to reduce the share of high-emission energy sources in the total energy production globally, the share of renewable electric power is set to increase. The two largest areas of growth in the production of renewable power are, according to the IEA (2021), solar- and wind power. These are variable renewable energy (VRE) sources dependent on the weather conditions, meaning the production capacity will vary out of the producer's control. With a larger share of the global supply of energy dependent on the whims of the weather, proper planning is paramount.

We aim to contribute to the power production planning by researching how best to forecast day-ahead load in the Norwegian market. The forecasting of load plays a role in the price formation in the physical and financial power market, and thus a more accurate forecast will benefit market participants on both the buy- and sell-side. The accuracy of the forecast is also of importance to the Transmission System Operators (TSO) to ensure the electricity infrastructure security and reliability, by balancing the supply and demand of the physical market.

We will focus on times of greater than usual price movements, represented in our testing by the period of abnormally high electricity-prices which occurred during the fall of 2021. The interest for this period specifically, comes partly from the deeply rooted assumption that the price sensitivity of demand in the electricity markets is close to zero. As such, we would expect to find that the price increase will not be a significant factor in load forecasting. All else equal, the methods of load forecasting should be no less accurate during the fall of last year, as the price-factor which changed significantly should make little difference in the demand.

The aim of the thesis is to find the best model to forecast load in the Norwegian electricity market when the price is higher than usual. To answer this, the work is centered around three more detailed questions. The first question is designed to find the best forecasting model. In answering the second question, we find whether our two best models add anything to the forecasting work, by comparing them to the officially published forecasts. The third question asked is if the recent price increase had any impact on the quality of the forecast.

The first question is regarding the best method of load forecasting for the timespan selected. We will approach this question by utilizing two different, but related types of methodologies. The first of which is the more traditional way of analyzing time series data, by using statistical regression-based methods. For the second type of methods, we will use a rapidly emerging way of analysis in the finance field, a set of different machine learning models. The thesis will not discuss the inner workings of machine learning and artificial intelligence in great depth, as the basis for the work is in the financial aspect of load forecasting, not the technical programming aspect of machine learning. The aim is to compare the methods, and to analyze whether the use of machine learning techniques provides better results for short-term load forecasting, or if the statistical methods prove to be the superior forecasters.

The second question is whether it is possible to improve today's official forecasts published on the Entso-E Transparency Platform. This will be evaluated by the performance of the forecasts developed in this thesis measured by two main metrics. The first of which is the overall performance of the forecast compared to the official forecast over the time periods in question. This will be measured as the average error of the forecast. The second metric will be the size of the outlier forecasting residuals. The argument for both metrics to be used being that the average forecasting performance best describes the models fit. However, the average error should be seen in accordance with large errors, to account for outlier risk in the forecast.

The third question is whether the forecasting performance is significantly affected by the price increase in the autumn of 2021. The last half of 2021 is a period with larger than usual price increase compared to previous years, meanwhile the autumn of 2019 experienced prices at more normal levels. The first test is to compare the forecasting accuracy of models including a price variable, to the accuracy of the same model blind to the price. If the model with price performs different to the one without price, the price variable is providing either information or noise, depending on whether the accuracy is better or worse. The second test, is a comparison of the accuracy in the autumn of 2019 to the autumn of 2021, using the two best models without price information. Again, if the price increase has influenced the load, we would expect a model not accounting for price changes to perform poorer in 2021 than 2019. We would also expect to see the three southernmost bidding zones where the price increase was steepest, to be comparatively more difficult to forecast in 2021 than the two northern zones with a lower price increase. As such, the two northern zones should be closer in forecasting accuracy between 2019 and 2021 than the southern three.

The forecasts will be made as rolling 24-hours in ahead predictions. This means that the predictions made for the first hour of any given day, is made using all information observable at the latest 24 hours in advance. To evaluate the forecasting performance two benchmarks is selected. The first of which is a seasonal naïve forecast. If forecasting models are unable to outperform the seasonal naïve forecasts, assuming the load is equal to the load 24 hours in advance, we would argue it provides no value to the forecasting work. The second benchmark is the official forecasts gathered from the Entso-E Transparency Platform data bank. Using this benchmark, we will be able to see whether the models designed in this thesis provides informational value exceeding what the established forecast does.

The thesis will be organized in 9 chapters, the first of which is this introduction. Chapter two contains background information about the markets relevant for the thesis and practical aspects of load forecasting. In chapter three a review of load forecasting in previous literature can be found. Chapter four describes the theoretical framework of the thesis, including the relevant models and evaluation metrics. Chapter five contains information on the data used in building the forecasting models. Chapter six is a model description, where the decisions of relevant variables and the model construction for all models used are described. In chapter seven the forecasting results will be presented, before they are discussed in chapter eight. Chapter nine will provide a conclusion to the questions the thesis aims to answer.

2. Background

In this chapter, background information about the Norwegian power market and the Entso-E platform is provided. There will be a short description of the key concepts of the power market. The chapter concludes with some general information regarding load forecasting.

2.1. The Norwegian power markets

While most European countries have one internal bidding zone for their power market, the Norwegian power market consists of five different bidding zones. The power markets in Norway have internal bottlenecks and can thus experience different power prices between the zones, as it has in the autumn of 2021. During the latter part of the year the prices has reached record highs in the three southernmost zones driven in part by power prices in the European markets, while the two northern zones have experienced a lower price increase. This has sparked a heated debate over the export of power to other European nations through cross-border interconnectors.

One of the reasons why the population of Norway has been so appalled by the rise in the price of electricity to households, is that the Norwegian power market has traditionally had some of the cheapest electricity in Europe. While other European nations has relied on a power mix consisting of a range of energy sources, Norway gets a large part of its electricity from hydroelectric powerplants. Electricity production using hydropower plants remains one of the cheapest forms of power production. Due to the favorable weather conditions and topography of Norway for utilizing impoundment and diversion hydropower facilities, the Norwegian households has been able to rely on this renewable and cheap power for many years.

In the coming decades, the rest of Europe is in dire need of access to renewable power if we are to reach the zero emission climate goals. The Norwegian hydropower production capacity will prove to be important for balancing the peak hours of supply and demand for the neighboring countries relaying in larger parts on variable renewable energy sources. Countries such as Denmark and Germany, where a larger part of the total power production comes from wind power will need an alternative source of electric power during the off-peak hours for wind.

The reason why the Norwegian reservoir-based hydropower production will be important is the storable nature of the production method. While it is expected that wind- and solar power becomes a larger part of the total electricity production in the future, the production capacities fluctuate with the weather conditions. As electricity-storage in large quantities is difficult, the electricity demand during the off-peak production hours needs to be covered by alternative production methods. Currently, this is achieved in too large part by fossil fuels such as natural gas and coal in many countries. These sources, like impoundment hydropower, can produce electricity at the time of need, when the wind or solar production is insufficient. If we are to reach the net-zero emissions goals we do however need to phase out much of the non-renewable power, especially coal.

This is where the cross-border connections and power trading capacities between nations will be crucial. To balance the power supply and demand during all hours of the day and all days of the year, with different geographical locations being suited for different production methods, we need to be able to exchange power. The Norwegian market currently has 17 cross-border connections, according to Entso-E (s.a.), the first of which started operating in the 1960s for just this reason. In years of heavy precipitation, where water would be sent passed the hydropower plants unused, it was now possible to utilize some of the excess waterflow to produce electricity which was exchanged over the border to Sweden. In years of low precipitation, where the 1950s had seen power rationing, this opened the opportunity to purchase power from the Swedish network. The same reasoning applies for the future, where the balancing of production and demand across Europe will require a network of complimentary power production.

2.2. Entso-E

The Entso-E system was created to ease the cooperation between European nations in their grid-to-grid power exchange and claims to promote a competitive pan-European market, (Entso-E, s.a., b). It was given legal mandate in 2009 by the EU in a push for liberalization of the power markets within the EU-area. The organizations consist of a larger set of key departments and areas of work, but we will limit the scope of explanation to the two most important factors for this thesis.

One of the areas of Entso-E's work is the integration of renewable energy sources in the European power grid. To get the European power markets ready for the future of power production, they work on both system development and market design to ensure best integration of renewable power production in line with the EU's 2030 energy policy. This integration will require flexible generation, demand response and interconnections between national grids. The reliability of the power supply using flexible production and demand across Europe will be built on good forecasts, both for the supply- and demand-side.

Another area of importance for this thesis, and for the function of Entso-E as an integration-system for European nations, is the transparency platform. The data-platform is essential for the creation of an internal pan-European marketplace for electricity. The member nations are required to submit information on amongst other things; electricity production, load, and transmission to a transparent data-bank open to all market participants. This limits the potential informational inefficiencies and promotes an efficient and competitive market. In addition, the transparent load and production information allows better planning of future systems and capacities across the continent.

2.3. Power market characteristics

The supply and demand of electricity is made up of the power producers on the supplier side and the greater society on the demand side. This includes everyday home consumption, industry, and everything in between. The modern world runs on electricity, and the consumption is critical for most of the day-to-day operations of our lives and to produce goods and services. Due to the importance of electricity, Hofmann & Lindberg (2019) has found that the short-term price elasticity is close to zero in the Norwegian power market. As such, the consumption of electricity, or the load, is generally not affected in the same way as other commodities might be by changing prices. This means the markets for electricity behaves somewhat different from other commodity markets. In this section we will define some key concepts in the power market.

The suppliers of electricity are dependent on balancing the supply with the load in the market, which makes the avoidance of under- and oversupply a vital part of the electricity markets. If the power-grids are not balanced efficiently by the Transmission System Operators at all times, it causes shutdowns and incurs large costs. For this reason, the act of forecasting and planning load is of importance to both producers and TSOs.

The production methods of electricity can be divided into two categories, renewable and non-renewable production. The non-renewable production of electricity includes nuclear power and burning fossil fuels, which is generally depletable power production resources. Among the largest sources of renewable power today are wind-, solar-, and hydroelectric power. While the world is run in large parts on the power from the burning of fossil fuels, the future of power production looks to be in the renewable sources of energy. These production methods, however, are often variable. The variable renewable power production creates new challenges for the balancing of electricity, where an increased part of the production is limited by uncontrolled weather factors, which is why we argue the forecasting of load will be important in a greener future.

The electricity load is driven by a multitude of factors, some of which can be described as daily and seasonal variations in power demand. We can divide it into base- and peak load based on either the time of day or time of year. By base load we mean the lows of observed load over a period, which can be viewed as the minimum load needed to keep society running during the period. The peak load is the highs of power consumption over the same period. The peak load on a daily basis in Norway is typically during the morning and afternoon, while on a yearly basis the peak is during winter.

Load peaks during winter times is caused by one of the large power consuming activities in the Norwegian market being heating, and the time of day coincides with when a sizeable part of society is aligned before and after working hours in using electric appliances and running their water heaters. With heating being one of the biggest electricity consumptions for Norwegian households, the weather is one of the main drivers of load. Cold periods increase load, while mild winters sees the power consumption peak lowered. In warmer climates, the cooling of buildings has an effect on load we do not see much of in Norway as the temperature rarely rises above threshold levels for wider air condition use.

2.4. Load forecasting

Load forecasting is often divided into three categories based on the forecast horizon, short-, medium- and long-term forecasting, as described by (Hammad et al., 2020). Short-term forecasts predict the load from minutes to days ahead, the medium-term horizon ranges from a week to a year and long-term forecasts are any horizon further in the future than a year.

With the different forecasting spans, so comes differences in what drives load. In shorter terms, the biggest impact on load comes from factors such as seasonality and weather, while long-term forecasts can benefit from including factors as economic growth and the implementation of power saving measures. While we can assume that the economic situation of the country can change in the coming months, it has a low influence on the electricity consumption in the short term. When forecasting over multiple years, these factors increase in importance.

There are multiple of popular methods for load forecasting, both in literature and practice. In this thesis the methods are split into two groups. The first is what could be described as the traditional method of statistical methods, while the second is machine learning- or artificial intelligence methods. While the statistical methods are still very much in use, the field of data science and use of machine learning is gaining popularity in finance for time-series analysis and forecasting.

3. Literature review

In this chapter, forecasting work in previous literature will be reviewed. Provided is a review of load forecasting in general, and of previous literature findings and results using different statistical methods and machine learning to forecast electricity load. Furthermore, in continuation of this chapter, an in-depth description is included in chapter 4, where we outline the theoretical framework for the specific methods used to forecast in this thesis.

The literature on electricity load forecasting (ELF) can be split into three or four categories depending on the time horizon according to (Hammad et al., 2020). The categories are long-, medium- and short-term load forecasting referred to as LTLF, MTLF and STLF respectively. The STLF category consists of forecasting intervals from one hour to a week. This category is important for daily operations for utility managers and have implications for generation and transmission scheduling. Some researchers also include a fourth class called ultra/very short-term load forecasting (VSTLF). VSTLF is for forecasting less than an hour ahead and are used for real-time control.

Hong & Fan (2016) provide a tutorial review on probabilistic electric load forecasting. In their review they argue that some of the empirical reviews comparing different STLF technique are misleading. They state that STLF techniques can be set at a disadvantage depending on the researcher's expertise and/or the case study setup. Therefore, there is no clear answer to which techniques performs best.

Numerous statistical time series models, artificial intelligence (AI) and hybrid models have been used to develop STLF's the last decades. Nti et al. (2020) has reviewed 77 articles within ELF published over nine years (2010-2020). They found that AI-based models were most commonly used, where 9 out of the 10 most popular models being AI. The exception being Autoregressive Integrated Moving Average (ARIMA) models, which is the third most used. Among the AI-based models, artificial neural networks (ANN) are the most popular representing 28% of AI models used in the electricity load forecasting work.

A number of studies in the last decades has been applying novel approaches to improving the STLF accuracy of the conventional Box & Jenkins (1976) ARIMA approach. Lee & Ko (2011) proposed an approach, embedding a lifting scheme into the ARIMA model. Simulation results showed the proposed algorithm superior to a back-propagation network (BPN) algorithm and a traditional ARIMA model.

A study by Tarsitano & Amerise (2017) proposes using a two-stage SARIMAX model, a combination of a linear regression and ARMA models for STLF. The study did not conclude on whether the proposed model achieved an improved forecasting ability, but the residual autocorrelation is reduced, shown by a reduction of the Ljung-Box test statistic. The reduction of autocorrelation in the residuals indicate an improved model fit.

Elamin & Fukushige (2018) used a SARIMAX model to perform STLF on a region in Japan. Their goal was to compare a SARIMAX model with main effects to a SARIMAX model with interactions. The model with interactions included cross effects in addition to main effects, as proposed by Hong et al. (2010). The SARIMAX model with interactions resulted in an improvement in MAPE by 22,2% compared to the SARIMAX model with main effects.

With the rise in computational power in the early 1990s, artificial intelligence-based methods have been widely studied and used to forecast electric load. One of probably the most popular AI-based methods, the artificial neural network (ANN) has according to Weron (2006) risen in popularity, because it requires no prior modeling experience to obtain reasonable load forecast. Another set of machine learning models made popular by their ease of use are the decision tree-based regression models, including simple regression trees, gradient boosted regression trees and random forests.

The comparison between statistical models and machine learning models have been made a number of times in previous literature. Papadopoulos & Karakatsanis (2015) compared the day-ahead forecasting performance of two statistical methods, one SARIMA and one SARIMAX, and two decision tree models, a Random Forest (RF) and a Gradient Boosting Regression Tree (GBRT). With hourly data from the ISO New England Control Area (ISO-NE CA) from 2009 to 2012, they found the GBRT to produce the most accurate 24 hours ahead load predictions. Measured in Mean Absolute Percentage Error (MAPE) the GBRT had errors of 1,32% compared to the RF errors of 1,96%, SARIMAX errors of 2,54% and SARIMA errors of 2,62%. While this represents a notable outperformance, the authors attribute the larger SARIMA and SARIMAX errors to their failure to model the multiple seasonality's in the data due to software limitations. For further work the authors suggests the inclusion of additional exogenous variables such as humidity or direct solar irradiation.

Another review of short-term load forecasting methods was done by Zor et al. (2017). In their publication they compare the accuracy of an Artificial neural network (ANN), a Support vector machine (SVM), and an Adaptive neuro-fuzzy inference system. Of the three methods, the artificial neural network and the support vector machine is applied to the bidding zone of New England, USA, while the adaptive neuro-fuzzy inference system was used for the bidding zone of New South Wales, Australia. The two directly comparable methods, the ANN and the SVM performed with an accuracy measured in MAPE of 1,95% and 1,79% respectively, with the authors arguing both methods being valuable for load forecasting work.

One of the important factors for the short-term load forecasting work is the ability to capture the different seasonalities and calendar effects on the electricity demand. Bakirtzis et al. (1996) noted the improvements of including the holiday effects in their artificial neural networks model for predicting the 24 hours ahead load in the Greek market for 1993. In their proposed model including the holidays effect they found a small improvement in the forecasting performance on the holiday, and interestingly a 30% improved accuracy over the two days following holidays.

Khwaja et al. (2015) shows the effect of bagging in the use of artificial neural networks. Their work shows the single artificial neural network model achieving accuracies in the range of 1,8 to 2,8 measured in MAPE. In comparison, the bagging neural network was shown to have accuracies in the range of 1,74 to 1,8 MAPE. While a single artificial neural network has the capability to achieve good results, the authors argue that the act of creating a set of uncorrelated learners should reduce the variation range of forecasting models.

4. Theoretical framework

In this chapter the theoretical framework of the thesis is explained. The chapter starts by looking at the theory behind the statistical methods. In the following part the architecture of the machine learning methods is described. The third part of the chapter contains general theoretical elements relevant to forecasting, including evaluating metrics for model fit and forecast accuracy.

4.1. Statistical methods

In this subchapter, we will outline the theoretical framework behind ARMA models. The theoretical framework is background for the work presented in chapter 6. In chapter 6 we are constructing three statistical models, a seasonal naïve autoregressive (SAR) and two extended ARMA models.

4.1.1. ARMA models

It was Box & Jenkins (1976) who first popularized the autoregressive process of predicting a variable based on previous values of the same variable. They did so when they introduced the ARMA model and the Box-Jenkins methodology for forecasting time series. The forecasting methodology consists of three steps: model identification, estimation of parameters, and prediction and validation.

ARMA models uses previous values and errors of the dependent variable to forecast. The model consists of two parts, the Autoregressive process (AR) and the Moving Average process (MA). Stationarity is also a requirement for ARMA models and in the case of non-stationary, the data series can be differenced to achieve stationarity. This results in an Autoregressive Integrated Moving Average (ARIMA) model.

Brooks (2014) states that one of the reasons why ARMA models do well compared to other statistical approaches is due to the use of previous values of the dependent variable, also referred to as "lags". This approach is especially effective in the case of electricity load forecasting where the load for a specific hour of the day, is often similar to the load the same hour the previous day. Therefore, a simple AR model with no more than a few lags, often called a naïve model, is used as a benchmark for more complex models.

An autoregressive process is as mentioned when the current value of a variable, y, only depends on the value of previous values of y, and an error term. The process can be denoted AR(p), where the (p) expresses the lag length. The model can be expressed as

$$y_t = \mu + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + u_t \qquad \textit{Eq: 1}$$

where y_t is the estimated dependent variable in period t, μ is the constant, ϕ_p is the coefficient determining the weight of the observation p, y_{t-p} is the lagged dependent variable for period t-p and u_t is a white noise disturbance term.

The moving average process also uses lagged values, but instead of a variable it uses previous forecasting errors. The model is a linear combination of white noise processes, where the current value of y, depends on the current and previous values of the errors. The white noise process has a constant and expected zero mean, $E(u_t) = 0$, a constant variance, $var(u_t) = \sigma^2$ and zero autocovariance, except when not lagged. The process can be denoted MA(q), with q expressing the lag length. The model can be express as

$$y_t = \mu + u_t + \theta_1 u_{t-1} + \theta_2 u_{t-2} + \cdots + \theta_q u_{t-q} \qquad \textit{Eq: 2}$$

where y_t is the estimated dependent variable in period t, μ is the constant, θ_q is the coefficient determining the weight of the observation q, $u_t \; and \; u_{t-q}$ is the current and lagged dependent variable for period t and t-q.

The combination of AR(p) and MA(q) processes results in an ARMA (p, q) model, where y_t is linearly dependent of its on previous values and a combination of current and previous white noise disturbance terms. By combining and shortening the equations from AR(p) and MA(q) we get the ARMA (p, q) model express below:

$$y_t = \mu + \sum_{i=1}^{p} \phi_i \, y_{t-i} + \sum_{i=1}^{q} \theta_i \, u_{t-i} + u_t \qquad \textit{Eq: 3}$$

Identifying the relevant number of lags (p, q) can be done by interpreting the output from the autocorrelation function (ACF) and the partially autocorrelation function (PACF), or by using and comparing versions of an information criterion, (Brooks, 2014). The ACF determines whether the dependent variable and the lag(s) are autocorrelated. Autocorrelation occurs when values of a time series is correlated with previous values over time. The difference between

ACF and PACF is that PACF shows the correlation with one specific lag, while AFC also include correlation between previous lags. When determining the correlation for the second lag, the value of the ACF will be determined by the correlation between y_t and y_{t-1}, and between y_{t-1} and y_{t-2}. Therefore, assume the only correlation in the dataset is 0.9 for the first lag, which also means a 0.9 correlation between y_{t-1} and y_{t-2}. This results in an AFC value for the second lag of $0.9\ x\ 0.9\ =\ 0.81$. PACF on the other hand, control for the correlation between y_{t-1} and y_{t-2} when determining the correlation for the second lag, thus resulting in a correlation of zero and some random error for the second lag. With this being the only difference, the ACF and PACF gives the same value for the first lag.

According to Brooks (2014), researchers use the AFC and PACF to find patterns that characterizes a time series. The usual patterns for an AR (1) process are a significant spike in PACF at lag one, followed by a number of near-zero values at higher lags. While for ACF there is usually a high value at lag one and then geometrically declining for higher lags. The interpretation of this pattern is that there's only a correlation between today's value and yesterday's value in the time series. If this pattern is reversed for the PACF and ACF, it suggests that we are dealing with an MA (1) process. Furthermore, a combination of both an AR and MA process usually has a geometrically declining PACF and ACF.

A fourth pattern can occur where the ACF never decay all the way to zero or it does so very slowly. This can indicate that the times series has a trend, which would make the series non-stationary. An ARMA model requires a stationary time series, which will be discussed in the next section.

Stationarity

Determining whether a series is stationary or non-stationary is important because it can strongly influence the series behavior and properties, (Brooks, 2014). An example of this, is the pattern observed in the previous section when determining the relevant lags in an AR and MA process. A stationary time series inhibits the characteristics of a constant mean, constant variance and a constant autocovariance structure. These requirements are expressed beneath, respectively:

$$E(y_t) \ = \ \mu \qquad \textit{Eq: 4}$$

$$E(y_t - \mu)(y_t - \mu) = \sigma^2 < \infty \qquad \textit{Eq: 5}$$

$$E(y_{t_1} - \mu)(y_{t_2} - \mu) = \ y_{t_2 - t_1} \quad \forall \, t_1, t_2 \qquad \textit{Eq: 6}$$

A stationary series possesses a mean-reverting process which can be observed as a time series which frequently crosses its mean value. This characteristic can be illustrated when unexpected changes occur, often referred to as "shocks". For stationary series, a shock will gradually go away. This is because the effect of a shock occurring at time t, has a smaller effect at time $t + 2$ than at time $t + 1$. In contrast, a shock will have an infinite effect in a stochastic non-stationary series with a unit root, as the effect of the shock has an equally large effect at time $t + 1$, $t + 2 \dots$ and so on. This non-stationary effect can be observed as discussed in the previous section, with an ACF value of close to one which is slowly declining. A trend-stationary process, also known as deterministic non-stationarity, is also mean reverting, but it doesn't fulfill the requirement of a constant mean.

In Brooks (2014) it is stated that cases of non-stationary in a time series with a trend, is a negative quality which can cause spurious regressions. Regressing two unrelated non-stationary variables which are trending over time, can result in a high R-square and significant coefficient estimates. This is of course valueless since they are unrelated. In contrast, two independent stationary variables regressed on the other will be expected to have non-significant coefficients and a low R-square. A non-stationary variable in a regression model will also make the standard assumptions for asymptotic analysis invalid, resulting in t-ratios and f-statistics not following their respective distributions. As a result, it is not possible to validly undertake hypothesis tests about the regression with non-stationary variables.

In order to apply an ARMA model to a non-stationary series, the series can be integrated of order d to achieve stationarity. Stochastic non-stationary series have been found to describe financial and econometric times series best, and this type can be differenced d times, equal to the number of unit roots to become stationary. The first difference is taken by subtracting the previous from the current observation: $\Delta y_t = y_t - y_{t-1}$. Transforming a non-stationary data series to a stationary series result in an ARIMA (p, d, q) model.

Extending ARIMA models

The ARIMA model can further be extended with seasonality and exogenous variables. When performing electric load forecasting with time series over longer periods, a seasonal or periodic component should be included, (Soliman & Al-Kandari, 2010). Extending the ARIMA with seasonality, results in a SARIMA (p, d, q) (P, D, Q)$_S$ model. Whereas P is the seasonal AR process, denoting lags from previous seasons, D denotes seasonal integration, Q denotes the number of MA processes from previous seasons and S denotes the number of observations in the seasonal pattern. This approach is useful as yearly, weekly, and daily seasonality is common in electricity load time series, (Weron, 2006).

Researchers using AR methods have usually dealt with these patterns in demand by using dummy variables. Seasonality can occur in many ways and for variables like hourly electricity load there is a daily, weekly, and yearly seasonality. Accounting for the seasonality helps the model adjust for the patterns and can increase accuracy.

Seasonality can also be dealt with by using a similar-day approach, or similar hour when dealing with hourly data. Weron & Misiorek (2005) divided all 24 hours into separate models, which was generally favored over the multi-model specification for STLF. Another irregular seasonality is holidays, which are often idiosyncratic and have caused significant forecasting errors, (Myung, 2013). Holiday seasonality is often dealt with using dummy variables for all public holidays or divided into multiple dummies based on the holiday's characteristics.

In a review by Soliman & Al-Kandari (2010), they argue that the lack of exogenous variables affecting load in SARIMA models as temperature, wind speed, humidity, and illumination in time series models, limits their forecasting ability. SARIMA models should be extended to include exogenous input variables also known as transfer functions in (Weron, 2014). By including exogenous input variables into the model, we now have an SARIMAX model. In this model the current value of the dependent variable is expressed linearly in terms of its previous values, past values of the noise, and in terms of the current and previous values of the exogenous variables.

4.2. Machine learning

Machine learning is defined as the use of computer algorithms which are able to improve performance in a task by gaining experience through the input of data on which to train. As a subset of artificial intelligence, it is a way of mimicking the natural intelligence of living creatures in their ability to implement prior experiences in their decision-making process. The goal of machine learning in the analysis and predictions of time-series data, is to create algorithms with the ability to find relationships in large sets of data, which would prove too time consuming to do manually.

Machine learning can be divided into three categories based on the type of feedback given to the algorithm. There are supervised-, unsupervised and reinforcement learning, where supervised learning is what is used for time series forecasting. Supervised learning is defined by labeled data with a set of inputs and output values, as described by Zhao & Liu (2007). The algorithm has a targeted designated output value for each set of inputs in the training data and works by learning the relationships between inputs and the target value, assigning weights of importance to the individual input values provided.

For times series analysis there are many different models of machine learning techniques available. In the following sub-chapters, the theoretical framework of relevant models for this thesis will be presented.

4.2.1. Artificial Neural Networks

Artificial neural networks are inspired by the neural network of the human brain with its many biological neurons and connections between them. It is first proposed by McCulloch & Pitts (1943) who developed the computational model for neural networks decisions, based on threshold logic algorithms. Artificial neural networks were constructed of a certain number of neurons which are either activated or not activated, just like biological neurons. Today there exists many versions of neural networks logic models, built on signaling through a set of nodes.

Each node in the network implements all input from the previous layer using a weight for each connection. The weight represents a single inputs importance, relative to all inputs. In addition to weighted inputs, all nodes contain a bias, representing a constant term. The internal value of each node can thus be expressed as

$$z = \sum_{n=i}^{N} X_i * W_i + B * 1 \qquad \textit{Eq: 7}$$

where X represents input i from the previous layer, W is the weighting of the respective input i, and B is the bias of the node. This value is put into what is called an activation function, which determine the output of each node. There are several different activation functions, some of the most popular in the literature being the Sigmoid, Tanh and Rectified Linear unit expressed as:

$$Sigmoid\colon \quad f(z) = \frac{1}{1 + e^{-z}} \qquad \textit{Eq: 8}$$

$$Tanh\colon \quad f(z) = \frac{e^{z} - e^{-z}}{e^{z} + e^{-z}} \qquad \textit{Eq: 9}$$

$$ReLU\colon \quad f(z) = \max{(0, z)} \qquad \textit{Eq: 10}$$

The Sigmoid and Tanh activation functions are characterized by s-shaped output ranges, ranging from 0 to 1 for the sigmoid function and -1 to 1 for the Tanh function. The Rectified Linear Unit has a linear output for values above zero, but a given node will not be activated for negative values. The non-activation of negative nodes makes the Rectified Linear Unit activation function computationally more efficient than the sigmoid and Tanh functions.

A frequently utilized way of learning for a neural network is called backpropagation, as was first proposed by Rumelhart et al. (1986). In the backpropagation process, the algorithm calculates the gradient of the errors for each weighting. The goal of the backpropagation process is to update the weights for both inputs and the bias, to find the local minimum of the loss function. The loss function will be specifically defined in chapter 4.3.1, here we limit the explanation to that it is a function representing the output errors of the predicting model.

A multi-layer perceptron network is a type of feed forward neural networks model. What makes it a feed forward network, is that the layers are connected unidirectionally from the input to output layer. This is different to some neural networks such as the recurrent neural networks, which is using bidirectional connections or loops linking the output back to the inputs of layers and nodes in the network.

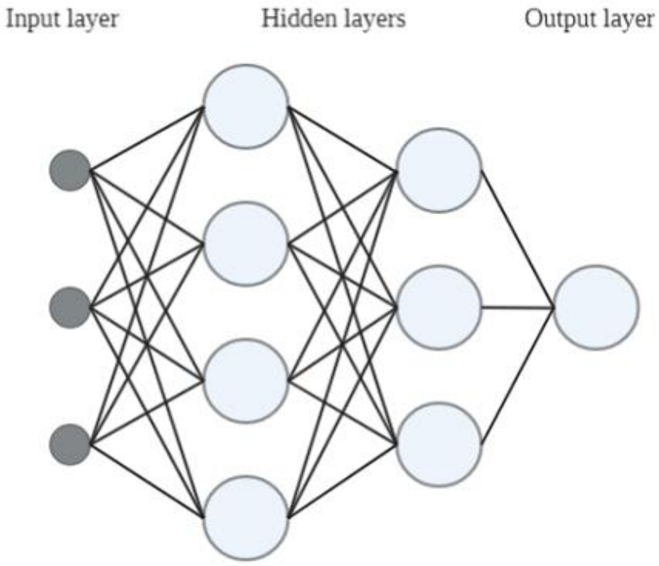

Figure 1: Neural network consisting of two hidden layers of four and three nodes

The most basic variant of the neural networks, the feed forward neural network consists of minimum three layers of nodes. One input layer, which is often called the visible layer. The reason for this is that the input layer is simply the variables fed into the model. The hidden layer or layers, consists of the nodes which takes the input variables and process them as explained in the previous section. The final layer, the output layer, is where the model's output is calculated. For a regression problem there is usually one output node, as we are looking for one output value, while classification problems often have multiple nodes in the output layer representing different output targets.

4.2.2. Decision Trees and Random Forests

A decision tree model is another way of analyzing regression type problems using machine learning. The model operates by creating a set of questions regarding the input variables, before running each observation in the dataset though the model, resembling a tree of decision points. It starts at what is often called the root node, the singular node at the start of the model. In a regression type problem, the nodes contain a question regarding a numerical value, or a Boolean represented by 1 or 0 for true or false. At every node the model poses a true or false question, in the style of:

$$Value \geq Threshold\ value \qquad \textit{Eq: 11}$$

When answering this question the model creates two branches, one for observations where the answer is true and one where it is false. This is the method of work from the root node until the model reaches what is often called the leaf node, the point at which splitting the data further gives no more accurate output value. When reaching the leaf node, the model's output is the mean observed target values of the data points in the training set, which fits into this final node of the decision tree. For training the decision tree, the gradient of the loss function is calculated when splitting the data differently at each node, optimizing for a local minimum.

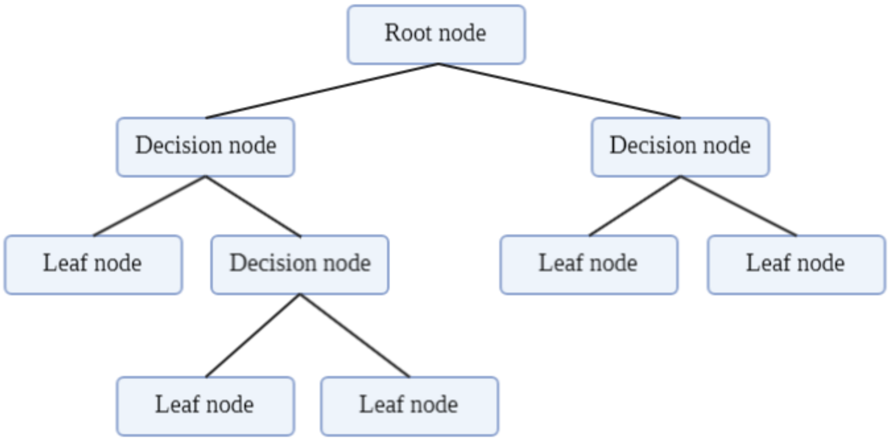

Figure 2: Decision tree model

An improved version of the simple decision tree is found in the Gradient Boosted Decision Tree model. What differentiates a gradient boosted tree from the regular version is that the boosted tree model constructs a series of sequential trees, each new tree aiming to account for the residuals of the former. This method of stagewise prediction of the former learners' residuals was discussed by Friedman (2001) and should improve model accuracy, especially in cases needing more complex data mining.

As another method building on the decision tree learning algorithms, Ho (1995) introduced Random Decision Forests. Decision tree models have a known tendency to overfit, defined as having very low bias, meaning errors in the training data, but high variance, meaning large errors in the validation data. To combat the tendency of overfitting, Random forests were built on the random subspace method popularly called "feature bagging", to reduce the feature correlation effects on the final prediction. The act of bagging is to rather than create one learner, several learners are created and assigned random parts of the dataset. In the case of feature bagging, they are assigned different features of the dataset. By using the average predictions of the set of uncorrelated learners, the random forest models aim to reduce the effects of noise in the training data and thus the model's variance.

4.3. Forecasting

Time series forecasts can be generated to predict both in-sample and out-of-sample. In-sample forecasts makes a prediction on the same dataset used to estimate the parameters in the model. Meanwhile, out-of-sample forecasts is estimated on one part of the dataset or time horizon and then used to forecast another part of the dataset or another time horizon. In-sample forecasts are expected to perform better, as the estimated model is fit to the exact dataset which it predicts, (Brooks, 2014). In this thesis we will forecast the future load, which means we are unable to train on the same data we are forecasting. Therefore, we are performing out-of-sample forecasts in this thesis.

Furthermore, there is two methods, Dynamic and Static forecasting. The dynamic method forecasts multiple steps ahead starting from the first period in the forecasting sample. This model does not add new information to the model for each forecasted step. Furthermore, depending on the number of steps and model design, a dynamic forecast uses forecasted values to forecast further than one step ahead. The Static method forecasts one-step-ahead, while rolling the actual data sample forwards. In this case, new information is added to the model for each step it forecasts, as it uses actual data to forecast further. In this thesis the goal is to forecast the day-ahead load in which the static method is the best fit to our purpose.

4.3.1. In-sample model fit

In this subchapter we outline the numerous criteria for evaluating to what degree a model fit the training data.

Specifying the correct number of lags and variables in ARMA models can be done in a number of ways. As previously outlined, we can use the ACF and PACF and use the patterns to specify the number of lags, but this isn't easy, as real-world datasets rarely exhibits the patterns described. A more popular technique is what is known as an information criterion (IC). One of the benefits with IC techniques is the removal of some of the subjectivity of interpreting patterns. There are multiple variants of the IC, but the general factors are the logarithm of the likelihood function and a penalty for adding extra parameters, also referred to variables. By adding an extra parameter, the IC increases if the parameter fails to increase the log likelihood function more than the penalty. The goal is to reduce the IC as much as possible. The different ICs presented below vary by how strict the penalty term is. The three IC used in EViews follows the conventions of Akaike's (1987), Schwarz's (1978), and Hannan-Quinn (HQIC), which are expressed respectively as:

$$AIC = -2(\frac{log(L)}{T}) + \frac{2k}{T} \qquad \textit{Eq: 12}$$

$$SIC = -2(\frac{log(L)}{T}) + \frac{klog(T)}{T} \qquad \textit{Eq: 13}$$

$$HQIC = -2(\frac{log(L)}{T}) + \frac{2klog(log(T))}{T} \qquad \textit{Eq: 14}$$

Where log(L) is the log of the likelihood function divided by the number of observations, k is equal to the total number of parameters estimated $k = p + q + 1$ and T is the sample size. Of the three IC, SIC is the stricter one in term of the penalty term, then HQIC and at last AIC. There is no clear answer to which model is the superior, but according to Brooks (2014) SIC will more often deliver the correct model, while AIC tent to deliver a too large a model.

For machine learning models, the in-sample model fitting is performed by adjusting the model to minimize a given loss function. While a loss function could, in theory, be any function which measures the errors of the fit between the target value and the in-sample predictions, there are two functions most commonly used. These are the L1 and L2 loss functions. The L1 loss function measures the absolute errors, while the L2 loss measures the squared errors.

$$L1 = \sum_{i=1}^{n} |Actual\ value - predicted\ value| \qquad \textit{Eq: 15}$$

$$L2 = \sum_{i=1}^{n} (Actual\ value - predicted\ value)^2 \qquad \textit{Eq: 16}$$

The main difference to note when deciding between which of the two functions to select is for when the data containing large outliers. The L2 loss function, being the sum of squaring the errors, will punish larger errors more severely than the L1 would. One has to consider whether the outlier errors should be of more significance to the model, or if the model is best served by using the absolute errors measure in the L1 loss function.

4.3.2. Out of sample forecast performance

Forecast evaluation is an important part of forecasting. It allows forecasters to test and select what models perform better, and it allows stakeholders to understand the performance of the forecasts, (Hong & Fan, 2016). Evaluating a model's forecasting performance, is often done by comparing error metrics with a baseline and to other models. To determine the forecasts accuracy, the whole out of sample forecast period are compared to actual value, and the difference is aggregated in an error metric. The model with the lowest measured error is argued to be the most accurate model. There are multiple error metrics that can be used to evaluate forecasts; Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE). MAE is the simplest metrics, measuring the mean absolute forecast error. MAE can be express as

$$MAE = \frac{1}{n}\sum_{t=1}^{n} |y_t - f_t| \qquad \textit{Eq: 17}$$

where n is total forecasting steps, y_t is the actual value at time t and f_t is the forecast value at time t. The MSE metric squares the difference between the forecast and the actual value in time t, and then takes the average over the period. This metric values large errors disproportionally more serious than small errors. Forecasts with large errors will be put at a disadvantage using MSE, which is a useful property if large errors are more serious than small errors. Transforming MSE back to the original scale while keeping the properties of MSE, is achieved with RMSE. MSE and RMSE can be expressed as

$$MSE = \frac{1}{n}\sum_{t=1}^{n} (y_t - f_t)^2 \quad , \quad RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n} (y_t - f_t)^2} \qquad \textit{Eq: 18, 19}$$

Lastly, MAPE measures the absolute error like MAE, but presents the absolute error in percentage of the actual value. MAPE is a good metric to compare forecasts of different scales, or when a forecasted value change scales over the forecast horizon. It also has the attractive property that it can be interpreted as a percentage error. MAPE can be expressed as:

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\frac{|y_t - f_{t|}}{|y_t|} \qquad \textit{Eq: 20}$$

According to a study by Nti et al. (2020) the most common error metrics in load forecasting are RMSE and MAPE used in 38% and 35% of studies, respectively. Hyndman & Koehler (2006) states that MAPE is a widely used metric in load forecasting, because of its simplicity and transparency. The MAPE's weaknesses are data of very different scales and data values close to zero or negative. The weaknesses of MAPE are not very relevant for load forecasting, as few level load series are close to zero and a negative load is not possible. Following the norms of previous STLF research, forecasting errors in this thesis are presented in MAPE and RMSE.

5. Data description

This chapter consists of a description of the time series data collected and how the datasets are pre-processed if that is the case. There is a description of the characteristics of historical load demand, temperature, humidity, and the price development used it the models designed.

5.1. Electricity load

The electricity load data are collected from Entso-E's transparency platform. The dataset consists of hourly observations for all five bidding zones in Norway from 01.01.2015 to 31.12.2021. The dataset used consists of an insignificant number of missing values and are not adjusted for any extreme values. The load data are also adjusted for changes between summer and wintertime.

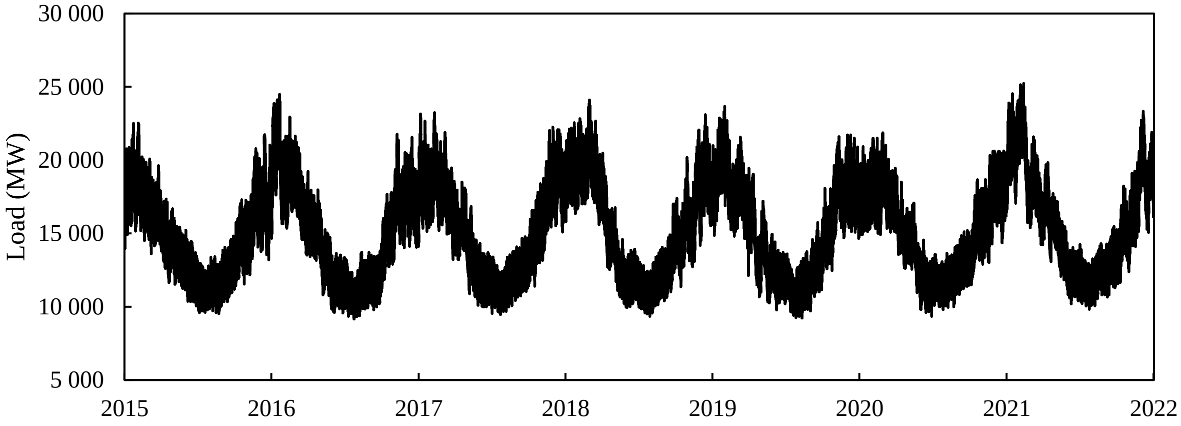

Figure 3: Hourly aggregated electricity load in Norway from 01.01.2015 to 31.12.2021

Norway's hourly aggregated electricity load used in this thesis is shown in figure 3. The annual seasonality in load can be observed, as the electricity load is higher in the winter months and higher during summer. The load does not appear to be trending over the period. The data is fairly regular, having very few extreme outlier values, except for shorter periods of very high demand during the winter of 2016 and 2021. To observe the weekly and daily seasonality in the load demand, we can reduce the resolution to the average load for all hours the week, displayed in an hourly frequency in figure 4.

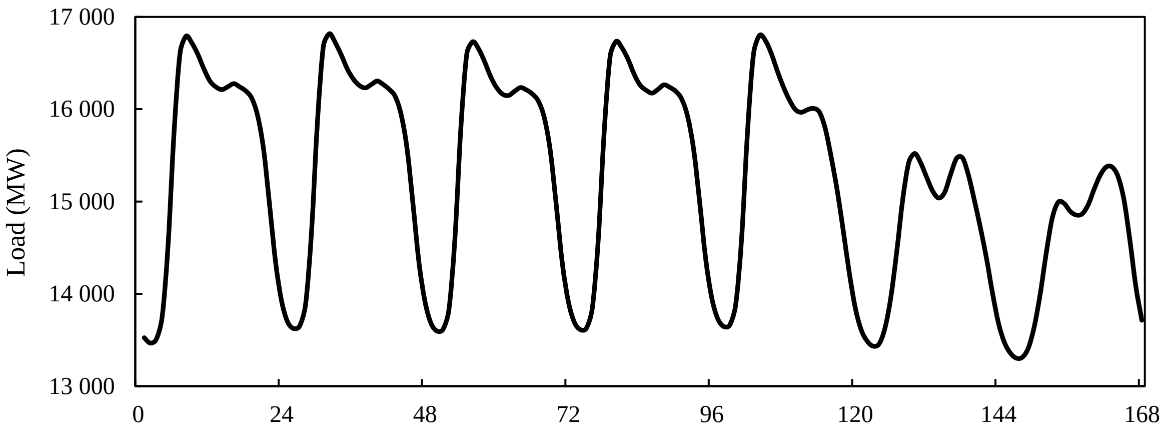

Figure 4: Hourly average load for all hours of the week in Norway from 01.01.2015 to 31.12.2021

The load curve indicates that on average Monday to Thursday have the same load characteristics, while Friday, Saturday and Sunday are unique. For the workdays the load curve increases rapidly from 05:00 and reaches its peak at 08:00. Meanwhile, in the weekend the load increase begins later in the morning. Friday resamples the other workdays up to midday but have a lower load in the afternoon and evening. Saturday and Sunday have a lower consumption throughout the day, and a later morning peak load at around 10:00, and then another peak at 18:00.

The load consumption varies throughout the day depending largely on human activity. As the normal work hours approaches the consumption increases rapidly. This can largely be explained by increased heating of air and water, and start-up in production facilities. This continues throughout workhours and as they end there is a slight increase around 16:00. After 16:00 consumption decreases to a daily low around 02:00. The pattern of hours with high and low demand is often referred to as a peak hours and off-peak hours. Peak hours being the hours of high demand and off-peak hours of low demand.

5.2. Weather data

Weather data is collected through the Norwegian meteorological institute, using one weather station for each bidding zone. We have collected hourly data on air temperature and relative humidity which have the highest influence on load apart from time factors according to Weron (2006). There are few missing values in the dataset for temperature and humidity over the period, and thus not compromising the dataset in any significant way. To smooth the data inputs at the points of missing values, a strategy of mean replacing has been adopted. The weather stations used for each bidding zone is shown in table 1.

Table 1: Weather station overview

Bidding zone	Weather station	Location
NO1	SN18700	Oslo
NO2	SN44300	Sandnes
NO3	SN69150	Stjørdal
NO4	SN90400	Tromsø
NO5	SN50540	Bergen

The weather stations are selected based on the quality of their data and by location, being in close proximity to the most inhabited place in the region. Using weather stations closest to the most inhabited areas are more likely to fit load demand better, as this is place account for the highest consumption. Hong et al. (2015) states that choosing weather station is important for load forecasting and can have an impact on forecasting accuracy. They also propose an algorithm from selecting the best weather station selection, which would be of interest as further work.

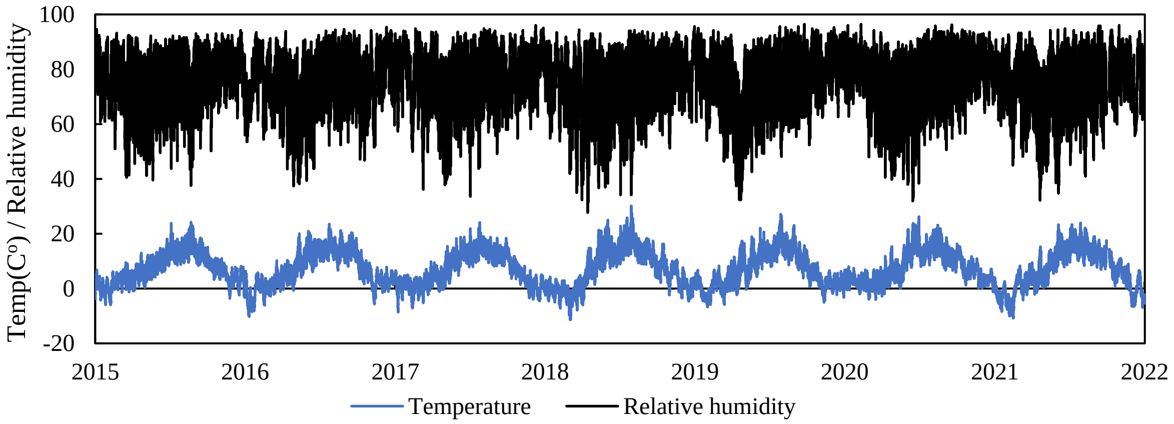

Figure 5: Average hourly temperature and relative humidity in Norway from 01.01.2015 to 31.12.2021

Figure 5 shows the hourly average temperature and relative humidity for the five weather stations chosen to represent the five bidding zones in Norway. As expected, there is a seasonality in the temperature. Relative humidity is the relationship between the absolute humidity and the maximum humidity. Maximum humidity is when the air no longer can hold more humidity without creating clouds or cause rainfall, given the temperature. Warm air can hold more humidity than cold air. As the relative humidity varies with temperature, we can also observe that the relative humidity tends to be lower more frequently in summer than winter.

5.3. Price history

The price data are gathered from Nord Pool and are only adjusted for the change between summer and winter time. Throughout 2021, the Norwegian power prices have increased tremendously. For most of the period considered in this thesis, the power prices have been low compared to the current Norwegian power prices. Especially in the three most southern bidding zones of Norway where the prices have increased more than for NO3 and NO4, which is the middle and northern part of Norway. From the beginning of 2015 to the end of 2021, prices in the southern parts of Norway NO1, NO2 and NO5 have been closely connected. The bidding zones of the middle and northern part of Norway are also closely connected, but the three southern and two northern bidding zones are not as closely connected as seen in table 2. This is due to a bottle neck in the power transmission system between the southern and middle part of Norway, which at times result in a price difference.

Table 2: Correlogram for all Norwegian bidding zones in the period 01.01.15 to 31.12.21

	NO1	NO2	NO5	NO3	NO4
NO2	0,99	1			
NO5	0,99	0,99	1		
NO3	0,65	0,64	0,64	1	
NO4	0,64	0,62	0,63	0,94	1

In figure 6 a graph of the hourly price development from 2015 to the end of 2021 for the bidding zones NO1 and NO4. For the greater part of this period, all prices were low and close to equal. During the spring, summer and autumn of 2020 Norway experienced very low prices compared to the previous years. Some price peaks in both NO1 and NO4 are observable over the period, whereas the other zone's price did not follow. A few months into 2021 the prices in both areas increase dramatically and the prices disconnect. The disconnection between the prices is due to the bottleneck which unable the zones from sustaining equal prices.

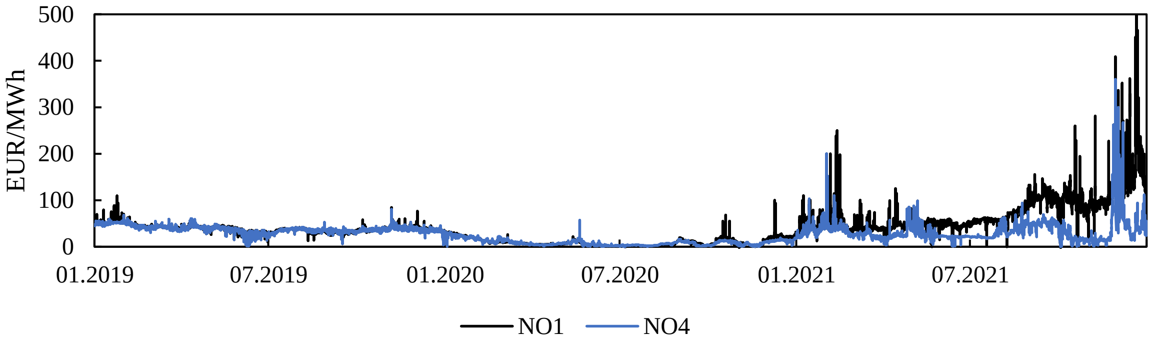

Figure 6: Price history for bidding zone NO1 and NO4 in the period 01.01.15 to 21.12.21

6. Model description

In this chapter of the thesis, the steps of model selection are outlined. Six models based on four methods will be created. Three of the models are based on a statistical approach: SAR and SARIMAX, and the three remaining are based on a machine learning approach: Decision Tree, Random Forest, and Artificial Neural Network.

6.1. Naïve model

To create a baseline for comparing the more complex models we introduce a seasonal autoregressive (SAR) model. This model consists of one SAR part, a 24-hour lag. This model can be written as a SARIMA(0,0,0)(1,0,0)$_{24}$. This model will be able to capture the daily seasonality as it uses the same hour from the previous day, but it will be unable to capture any of the weekly and annual seasonality. The model will predict the hour 24 hours ahead within the same season, and its weakness is that it will for example be unable to know whether the current season is autumn or spring. The naïve model can be expressed as:

$$\hat{y} \ = y_{t-24} \qquad \textit{Eq: 21}$$

6.2. SARIMAX model selection

In this section we outline every step of deriving the SARIMAX model. The model selection steps displayed in this chapter is done for NO1 in the estimation period from 01.01.2015 to 30.06.2019 with the aim to forecast the last half of 2019. This process is also done for all bidding zones and forecasting horizons resulting in the same model design. The forecasts for the last half of 2021 are also trained on the 4,5 prior and not from the start of 2015, as a shorter training period reduced computational runtime with no reduction in the accuracy. This process is based on Box & Jenkins's (1976) approach which consists of model identification, estimation of parameters, validation, and predictions. The first two parts of the process are outlined in this chapter and the predictions are displayed in chapter 7.

6.2.1. Model identification

The first step involves determining the order of the model required to fit the features of the time series. The first step is to investigate what AR and MA order fits the load series. As we outlined in chapter 5, the load demand for Norway has an annual, weekly, and daily seasonality. The same is observed for all five bidding zones, as we can observe for NO1 in figure 7.

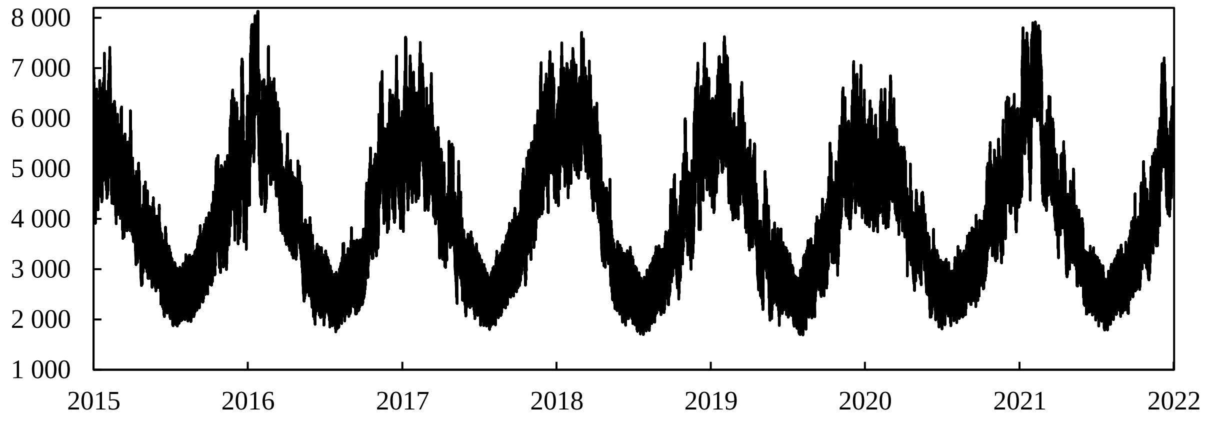

Figure 7: Hourly load demand for NO1 from 01.01.2015 to 31.12.2021

To achieve a meaningful interpretation from plotting the ACF and PACF, the series has to be stationary. Seeing the seasonal patterns of the level series we assume a seasonally stationarity series, as it has a slow mean-reverting effect and no apparent trend. The first and/or seasonal difference can be taken in the attempt to achieve stationary. The first difference is taken and graphed in figure 8. The series is centered around a mean, but it seems to not have a constant variance as there is a higher variation in load in the winter than in the summer.

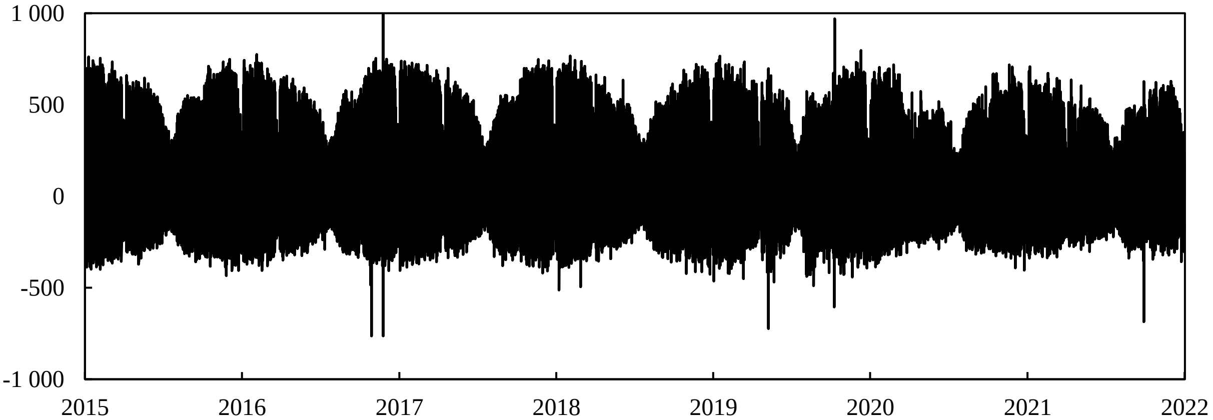

Figure 8: First difference of the hourly load demand in NO1 from 01.01.2015 to 31.12.2021

To reduce the impact of heteroscedasticity in the series, the logarithmic values of the series is used. Figure 9 displays the log first differenced and figure 10 displays the series after taking the first and seasonal difference (s = 24). By taking the logarithmic values of the series the change in variation between summer and winter is reduced, and by visual inspection the series looks stationary after both taking the first and seasonal difference.

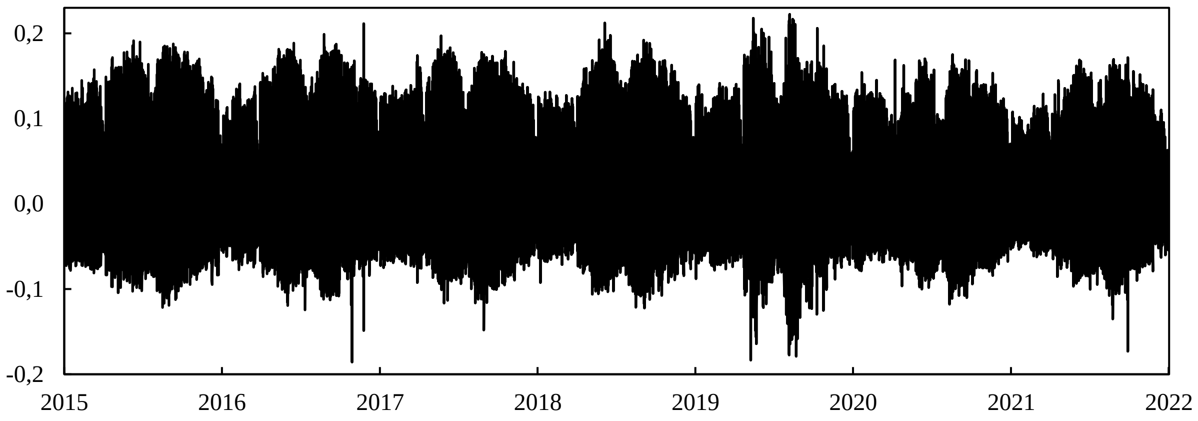

Figure 9: Log first difference of the hourly load demand in NO1 from 01.01.2015 to 31.12.2021

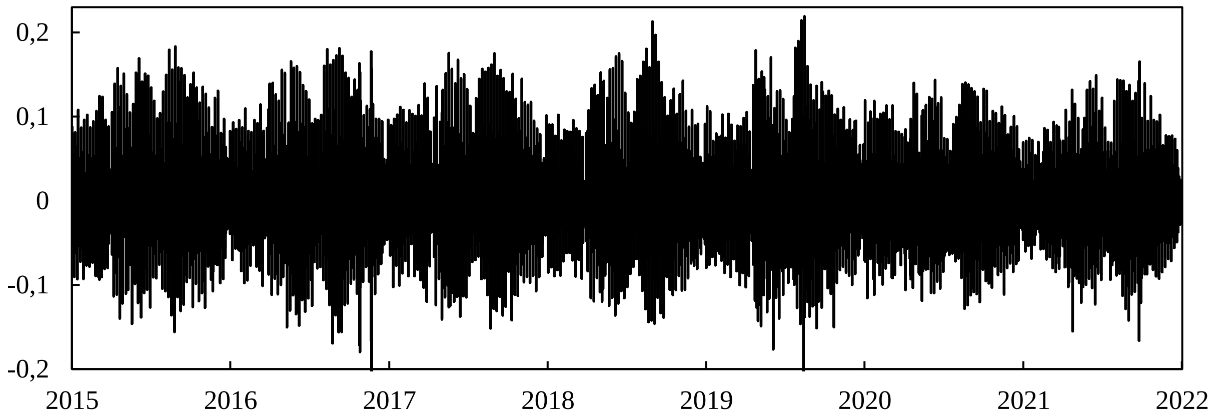

Figure 10: First and seasonal difference of the log hourly load demand in NO1 from 01.01.2015 to 31.12.2021

To formally verify that the series in stationary we apply a unit root tests. Both the Augmented Dickey-Fuller (ADF) (Dickey & Fuller, 1979) and Kwiatkowski, Phillips, Schmidt, and Shin (KPSS) (Kwiatkowski et al. 1992) are tested on the series. The ADF test's null hypothesis is non-stationary, and the alternative is stationarity. The 5% critical value for ADF test is -2,86. The KPSS test's null hypothesis is a trend-stationary series, and the alternative is non-stationarity. If the KPSS test indicate trend-stationarity in the level series, the correct approach is to remove the trend instead of differencing. The KPSS critical value at 5% is 0,46. Testing is performed on the training dataset which ranges from 01.01.15 to 01.07.19 and 01.01.17 to 01.07.21. The lag length in the two test is specified by AIC and Schwert (1989) for the ADF and KPSS test respectively. Both criteria result in a lag length of 54.

Table 3: ADF and KPSS test statistic on NO1 load from 01.01.2015 to 31.06.2019

Test	Level	First diff.	First and Seasonal diff.	Log First diff.	Log First and Seasonal diff.
ADF	-4,7*	-30,8*	-44,6*	-31,4*	-44,9*
KPSS	0,864*	0,005	0,001	0,005	0,000

**Significant at the 1% level*

The level series is stationary according to the ADF test, and trend-stationary according to KPSS at the 1% level. The suggested method to deal with a trend-stationary series is detrending. Since there is no trend in the series, but seasonality, we also tested the differenced series. Testing the both the first and first and seasonally differenced series shows a stationary series with no trend as the ADF test is significant and the KPSS test shows non-significant results. To visually display the seasonal autocorrelation, we plot the ACF in figure 11.

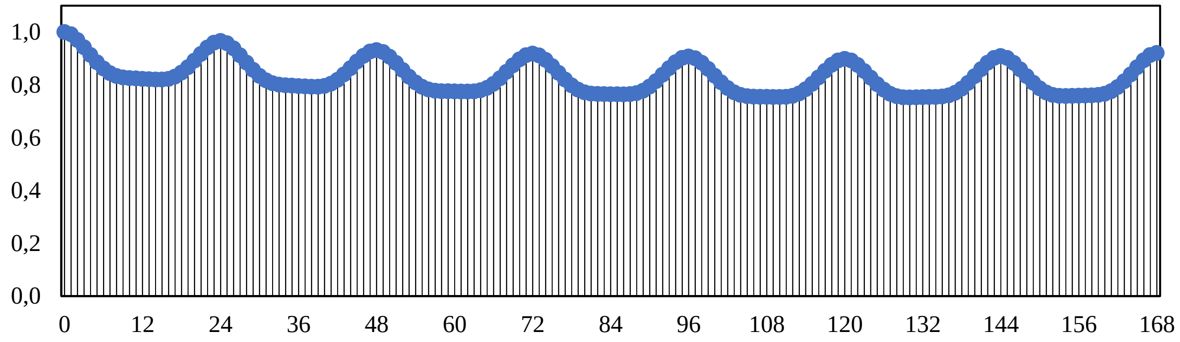

Figure 11: ACF plot of the level NO1 series for all hours of the week

According to Brooks (2014) a trend can display itself on a ACF plot by never going to zero, or by doing so slowly. In this instance the ACF goes up and down in 24 hours intervals in line with the daily seasonality. In conformity with previous test, it is not indicative of a trend, but seasonality in the dataset. Therefore, the first and seasonal difference (s=24) of the series is the correct approach, instead of detrending. After obtaining a stationary series, the next step is to plot the ACF and PACF and determine the ARMA order.

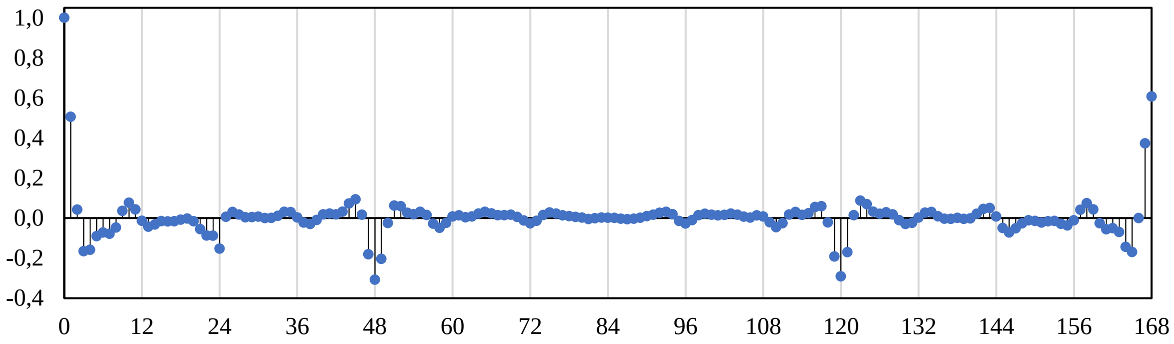

Figure 12: ACF plot of the first and seasonal differenced NO1 series for all hours of the week

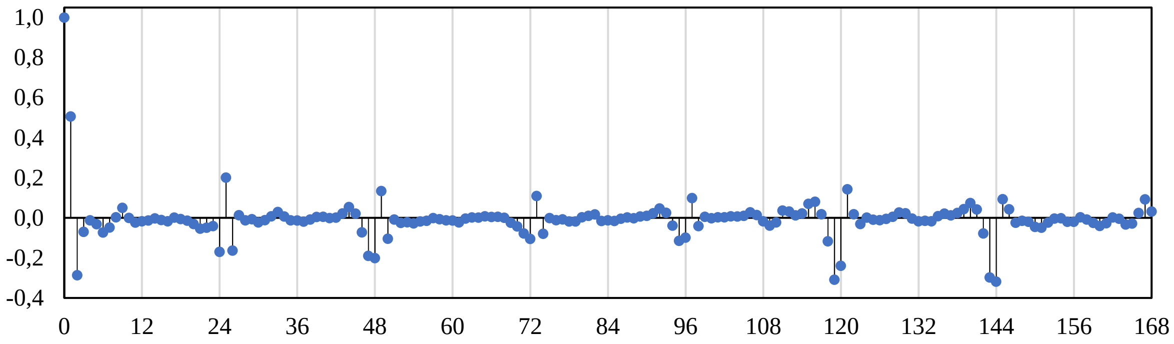

Figure 13: PACF plot of the first and seasonal differenced NO1 series for all hours of the week

The ACF plot of the first and seasonally differenced series suggests that we have multiple relevant MA and SMA parts. Since we are forecasting 24 hours ahead, we cannot use more recent lags than t-24. In the ACF plot there is significant spikes around the seasonal MA lags 24, 48, 120 and 168. The same goes for the PACF plot where we have significant seasonal AR spikes around the lags 24, 48, 72, 96, 120, 144 and 168. To determine the number of ARMA parts that best fit the load, we perform further tests in the next step of the Box-Jenkins approach.

Feature selection

The next step is to determine the best exogenous variable that fit load demand best. We divide the exogenous variables into two categories: main effects and cross effects. Main effects consist of weather variables and dummy variables for calendar effects. Cross effects are both dummy and weather variables created by two variables interacting.

Weather variables: Temperature and humidity are the most used weather variables used as load predictors according to Arora & Taylor (2013). As discussed in chapter 5, weather data from a reliable source in the respective bidding zone is used. Although humidity is not as extensively studied in the literature, Elamin & Fukushige (2018) has shown in to be useful for load forecasting. In the SARIMAX model we include hourly temperature in Celsius degrees and relative humidity as exogenous variables. We are using actual data at time t, and t-1 instead of forecasted weather for the day ahead. This is to ease the data gathering process and the impact is low due to the assumption of unsystematic weather forecasting errors.

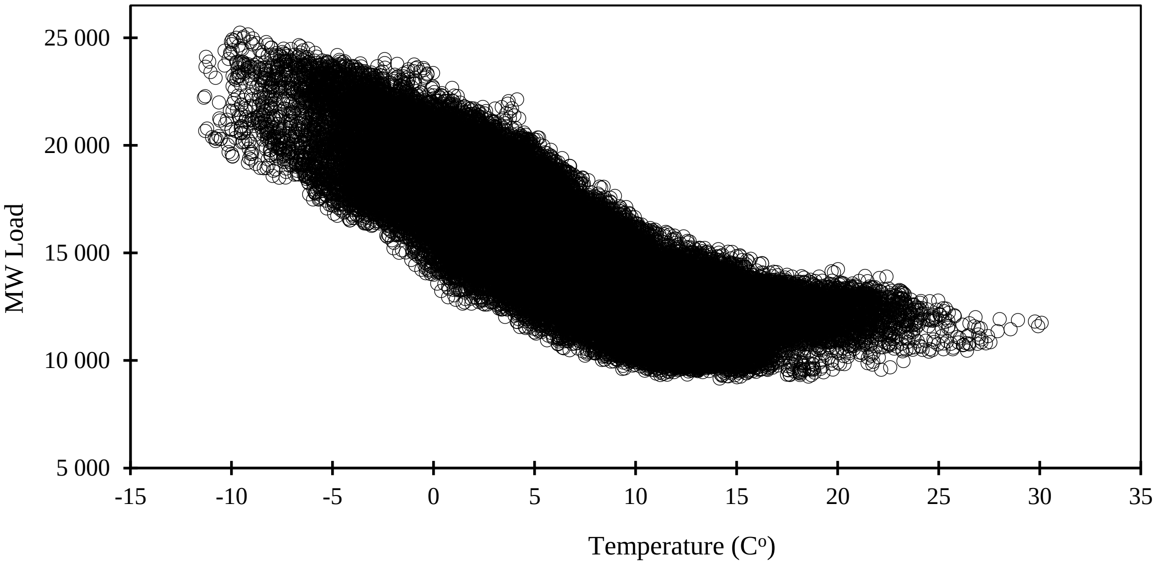

Figure 14: Average hourly temperature and load from 01.01.2015 to 31.12.2021

Temperature is closely related to power consumption in Norway, as electricity is the main utility for heating. Plotting temperature against load in figure 14 shows a non-linear relationship with a negative correlation. As the air temperature rises, the demand for electricity declines, except for temperatures above about 16 degrees. Beyond this point the load peak flattens and the minimum load starts to increase. This can be explained by the increased demand for electricity to air-condition instead. The latter effect is often observed clearly in countries with a warmer climate, where the load increases when the temperature rises beyond about 30 degrees. To capture this nonlinear relationship between load and temperature we introduce the variables heating degrees and temperature squared. Through testing we find that temperature squared performed best of the two. To capture the influence of temperature and humidity has on load, we introduce the weather variables: $temp_t$, $temp_{t-1}$, $temp_{t-24}$, $temp^2_t$, $temp^2_{t-1}$, $temp^2_{t-24}$, $humi_t$ $humi_{t-1.}$

Annual seasonality: To capture the annual seasonality, we tested the inclusion of monthly and weekly dummies. In theory the 52 weeks in a year should be able to capture the gradually change in seasonality better than monthly dummies, but at the same time, weekly dummies are more sensitive to variations. To avoid overparameterizing we decided to opt for monthly dummy variables. Eleven dummy variables are used: Month(k), k=1, 2...,11 from February to December and January are left as the reference.

Furthermore, we want to take public holidays into consideration. One approach would be to create one dummy variable for each holiday, or by grouping the different types of holidays. We decide to take a simple approach and make one dummy variable for all holidays, where Holiday is set to 1 for public holidays and 0 otherwise.

Weekly seasonality: The load demand also varies by the day of the week. Workdays differs from the weekend and the occasional holiday. To capture this seasonality, we have multiple approaches. We can group the workdays, except from Friday as they have a similar load curve on average, (figure 15). Saturday and Sunday have different load curves and thus grouping them can reduce the model fit. Through testing we decide on creating one separate dummy variable for all days, introducing dummy variable Days(j), 1,2…,6 from Tuesday to Sunday leaving Monday as the reference. Dummy Day(1) = Tuesday takes value 1 for Tuesday and 0 otherwise.

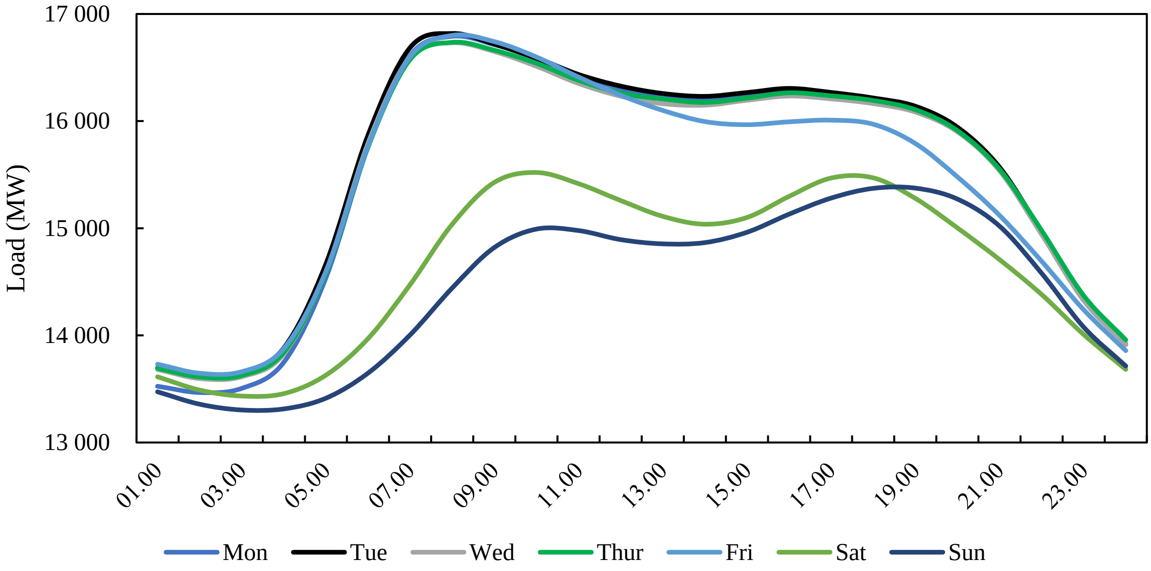

Figure 15: Hourly average load for each weekday in Norway from 01.01.2015 to 31.12.2021

Daily seasonality: Figure 15 also shows the daily seasonality, as the load varies depending on the hour of the day. To determine how to capture the daily seasonality best, we experimented with grouping the hours and with one dummy for each hour. We grouped the hours into peak hours and off-peak hours and compared it to having one dummy for all hours. One dummy for each hour resulted in a better fit, thus we introduce 23 dummies, Hour(i) 1,2...23 where the reference hour is from 23:00 to 00:00. For example, the dummy Hour(1) takes value 1 for hour 00:00 to 01:00 and 0 otherwise.

Price*: Lastly, the hourly actual price at time t is added in some models. The actual price is used as a proxy for ahead price. This will not affect the performance, assuming the forecasting errors of the day-ahead price is unsystematic.

Cross effects

In our model we also include interaction variables between weather variables and calendar variables as proposed in Hong et al. (2010). The interaction variables are created by simply multiplying two existing variables. The interaction between for example temperature and the hour of day lets the model variate by how much the temperature affects the load demand at a specific hour. This is beneficial as the demand for heating based on temperature can differ during the night versus in the morning. The same principle applies to the other cross effects.

We firstly introduce the interaction variable day(j) * hour(i). In this variable we add every hour of the week except Sunday from 23:00 to 00:00. This allows for every hour of the week to be considered independently.

Furthermore, we add the interaction variable holiday * hour(j) to allow for every hour of public holidays to be considered independently. Holidays primarily affect the load demand during workhours and can resemble the weekend. This indicate that the off-peak hours are less affected by holidays.

We also add three temperature and calendar interaction variables; $temp_t$ * hour(j), $temp_{t-1}$ * hour(j), $temp_t$ * month(k) and two interaction variables for relative humidity and calendar variables; hour(j), $humi_t$ * hour(j), $humi_t$ * month(k). These interaction variables allow for temperature and humidity to have a different effect on load, based on the current hour and month. Lastly, we introduce an interaction between temperature and humidity; $temp_t$ * $humi_t$**,** as the humidity varies between seasons based on air temperature.

6.2.2. Parameter estimation and diagnostic testing

This subchapter consists of estimating and testing the SARIMAX model. The most common methods to determine ARMA parts are by interpreting the ACF and PACF plots, minimizing IC as AIC and SIC, the maximum likelihood and choosing variables resulting in the lowest forecasting error. We decided to select the best AR and MA parts based on what results in the lowest AIC. Through extensive testing we found that the model SARIMA (0,0,0) $(7,0,7)_{24}$ achieved the lowest score on all IC measurements. As denoted in the model, the seasonal difference of the load series for forecasting is not taken, as differencing resulted in a higher forecasting error. According to Hyndman & Khandakar (2008) it is best to make as few differences as possible because over-differencing harms forecasts.

Table 4: Estimation output for SARIMA(0,0,0)(7,0,7)$_{24}$ for NO1 from 01.01.15 to 30.06.19

Variable	Coefficient	Std, Error	t-Statistic
C	4123,563	75,463	54,6
SAR(24)	-0,0153	0,00008	-193,1
SAR(48)	-0,0153	0,00005	-260,6
SAR(72)	-0,0158	0,00008	-189,6
SAR(96)	-0,0151	0,00007	-209,9
SAR(120)	-0,0156	0,00006	-244,0
SAR(144)	-0,0156	0,00007	-217,7
SAR(168)	0,9844	0,00000	136339,5
SMA(24)	0,9479	0,00673	140,7
SMA(48)	0,9434	0,00490	192,2
SMA(72)	0,9447	0,00661	142,9
SMA(96)	0,9460	0,01381	68,5
SMA(120)	0,9470	0,01535	61,7
SMA(144)	0,9462	0,00984	96,1
SMA(168)	-0,0499	0,00392	-12,8
R-squared	0,967	SIC	13,871
AIC	13,867	HQIC	13,868

All the SAR and SMA parts of the equation are significant at the 1% level. To check if the model captures all the autocorrelation, we plot a correlogram of the residuals. Of the ACF and PACF plot, we can observe that the residual autocorrelation is reduced, but still present. Although determining if there is any residual autocorrelation do not need any formal testing, the Ljung-Box statistic at lag 168 is 270 768 for comparison with later models.

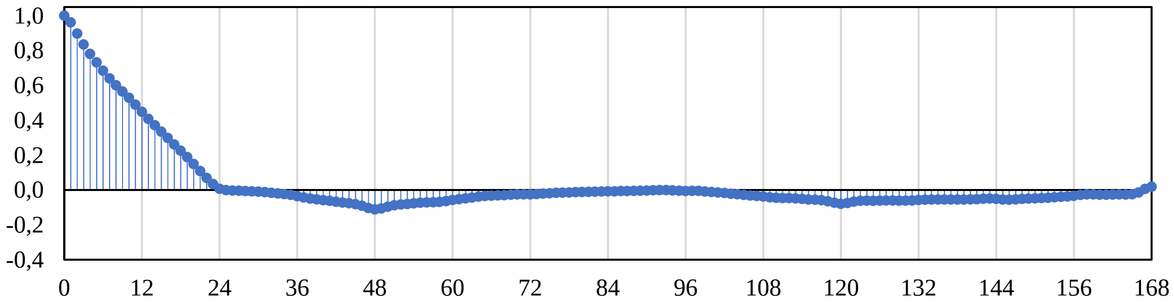

Figure 16: ACF plot of SARIMA(0,0,0)(7,0,7)$_{24}$ estimation residuals for all hours of the week

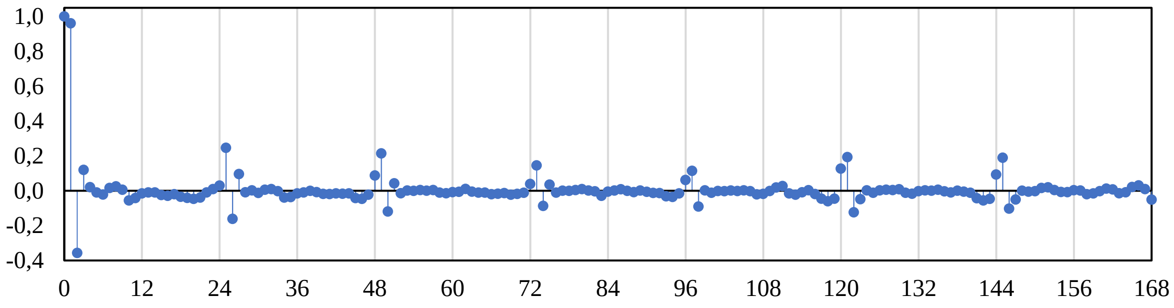

Figure 17: PACF plot of SARIMA(0,0,0)(7,0,7)$_{24}$ estimation residuals for all hours of the week

Of the ACF and PACF plot there is clearly substantial part of autocorrelations left in the residuals. The residual autocorrelation after lag 24 in the PACF plot is in the few lags following the seasonal lag. Since we are unable to take the first difference and use more recent lags prior to t-24 and chose not to take the seasonal difference, we are unable to capture more of the autocorrelation with a simple SARMA model. If we could manage to remove all the autocorrelation, the residuals would be unsystematic, also referred to as white noise. From figure 18, there seems to be a yearly seasonality, with larger residuals during the winter months.

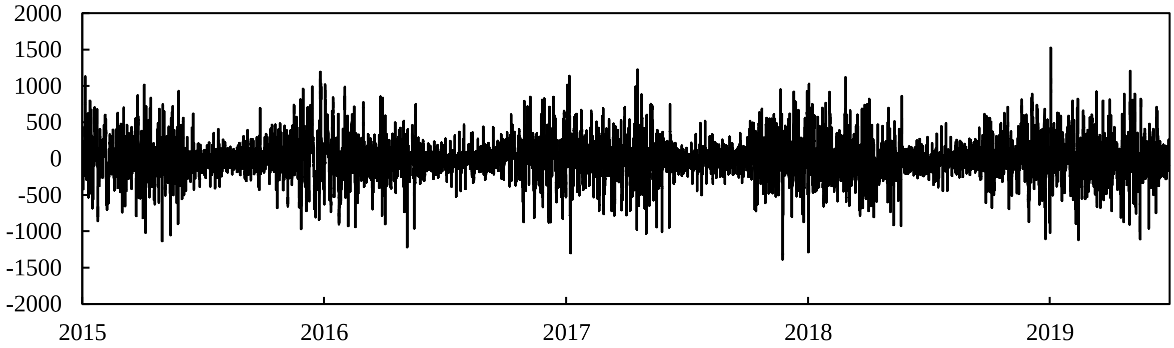

Figure 18: Graph of SARIMA(0,0,0)(7,0,7)$_{24}$ estimation residuals

Higher residuals in the winter months indicates that the model is unable to capture the yearly seasonality. Another explanation is that the residuals are in absolute terms which causes higher errors as the load increases in the winter. Turning the residuals into percentages reduce the seasonal variation and shows that the model misses equally much in the summer and winter. In attempt to capture the yearly seasonality, we introduce monthly dummy variables in later models.

The next step is to fit the model with the exogenous variables. The main effects consist of the weather variables; $temp_t$, $temp_{t-1}$, $temp_{t-24}$, $temp^2_t$, $temp^2_{t-1}$, $temp^2_{t-24}$, $humi_t$ $humi_{t-1}$ and dummy variables for daily hours, weekdays, months, and holidays.

The first model, herby referred to as “SARIMAX main” is estimated with only the main effects. The estimation period is 4,5 years prior to the second half of 2019 and 2021. Table 5 shows the model estimated for NO1 prior to the second half of 2019. Insignificant variables are omitted; thus, the table only shows the significant parameters at the 1% level.

Table 5: Estimation output for SARIMAX(0,0,0)(7,0,7)$_{24}$ main for NO1 from 01.01.15 to 30.06.19

Variable	Coefficient	t-Statistic	Variable	Coefficient	t-Statistic	Variable	Coefficient	t-Statistic
C	4178,749 (257,343)	16,24	Hour8	949,217 (313,165)	3,03	Mar	99,894 (15,342)	6,51
SAR(24)	-0,010 (0,001)	-10,95	Hour9	1019,386 (329,803)	3,09	May	-76,476 (17,660)	-4,33
SAR(48)	-0,007 (0,001)	-7,90	Hour10	1051,285 (338,405)	3,11	Jun	-119,786 (22,689)	-5,28
SAR(72)	-0,009 (0,001)	-9,25	Hour11	1065,477 (338,558)	3,15	Jul	-145,045 (30,296)	-4,79
SAR(96)	-0,008 (0,001)	-8,73	Hour12	1067,197 (338,183)	3,16	Aug	-143,197 (29,911)	-4,79
SAR(120)	-0,008 (0,001)	-8,70	Hour13	1060,390 (335,792)	3,16	Sep	-150,663 (24,816)	-6,07
SAR(144)	-0,010 (0,001)	-10,89	Hour14	1053,618 (333,946)	3,16	Oct	-110,353 (19,270)	-5,73
SAR(168)	0,986 (0,001)	1126,96	Hour15	1075,264 (330,017)	3,26	Nov	-134,108 (12,843)	-10,44
SMA(24)	0,639 (0,004)	164,04	Hour16	1110,720 (329,976)	3,37	Dec	-90,405 (6,526)	-13,85
SMA(48)	0,583 (0,004)	143,94	Hour17	1096,523 (336,990)	3,25	Temp	-37,240 (1,264)	-29,47
SMA(72)	0,578 (0,004)	132,86	Hour18	1048,310 (345,494)	3,03	Temp$_{t-1}$	-43,489 (1,260)	-34,52
SMA(96)	0,564 (0,004)	126,91	Hour19	971,209 (352,476)	2,76	Temp$_{t-24}$	-27,137 (0,277)	-98,07
SMA(120)	0,528 (0,004)	120,32	Sat	-374,613 (30,288)	-12,37	Temp^2	0,742 (0,050)	14,81
SMA(144)	0,565 (0,004)	129,56	Sun	-423,332 (19,402)	-21,82	Temp$^2_{t-1}$	0,807 (0,050)	16,12
SMA(168)	-0,341 (0,004)	-83,81	Holiday	-253,527 (2,344)	-108,17	Temp$^2_{t-24}$	0,569 (0,017)	33,85
Hour7	812,310 (308,063)	2,64	Feb	74,280 (10,856)	6,84	Hum$_t$	1,728 (0,189)	9,14
						Hum$_{t-1}$	-1,249 (0,191)	-6,54
R-square	**0,988**		**SIC**	**12,842**				
AIC	**12,828**		**HQIC**	**12,832**				

Estimating the SARIMAX with main effects results in an AIC of 12,828, compared to the SARIMA’s model AIC of 13,867. This indicates that the model has improved its fit more than the punishment for introducing 49 new variables. This is also supported by an improvement in the adjusted R-square from 0,967 to 0,988, showing that SARIMAX is able to explain more of the observed variation in the load demand. This is expected as we introduce dummy variables to capture the seasonality and the effect of temperature and relative humidity on load.

Furthermore, SARIMAX main’s residuals are observably lower in the estimation period than the residuals of the SARIMA. The yearly seasonality is still evident but significantly decreased. The model seems unable to capture the higher variation in load during the winter than summer. The Ljung-Box test statistic for the SARIMAX main model is 145 874, dramatically lower than for SARMA, indicating that the model has improved.

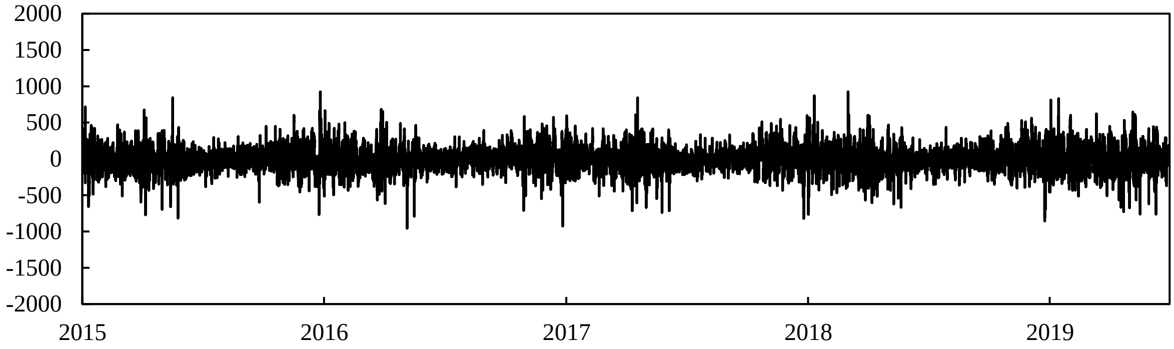

Figure 19: Graph of SARIMAX(0,0,0)(7,0,7)$_{24}$ main estimation residuals

Extending the SARIMAX main model further we add the cross effects in a new model herby referred to as “SARIMAX interaction”. The cross effects consist of the interaction terms: day(j) * hour(i), holiday * hour(i), temp$_t$ * hour(i), temp$_{t-1}$ * hour(i), hum$_t$ * hour(i), temp$_t$ * month(k), hum$_t$ * month(k) and temp$_t$ * hum$_t$.

In addition to introducing all the interaction variables and letting variables as holiday, temperature and humidity vary by hour and month, we remove the dummy series day(i) and hour(j) to avoid multicollinearity with day(i) * hour(j). This change opens up to not only letting the hour of day and weekdays to be modeled independently, but every hour of the week.

Table 6 display the SARIMAX interaction model’s R-squared, and ICs for NO1 in the estimation prior to the second half of 2019. The estimation output is shown in appendix 1, where insignificant variables are omitted, and the table only shows the significant parameters at the 1% level.

Table 6: SARIMAX(0,0,0)(7,0,7)$_{24}$ interaction IC and R-square for NO1 from 01.01.15 to 30.06.19

Test	Value
R-squared	0,9905
AIC	12,61
SIC	12,68
HQIC	12,64

Compared to the SARIMAX main model the number of variables is increased by 254. The addition of a large number of variables is punished by all the IC, but all IC for the SARIMAX interaction model is about 0,2 lower compared to SARIMAX main. The AIC of SARIMAX main was 12,828 and is improved to 12,68 by adding the cross effects. Graphing the residuals in figure 20, there still seems to be a seasonal change in the residuals, but significantly reduced compared to the SARIMA model.

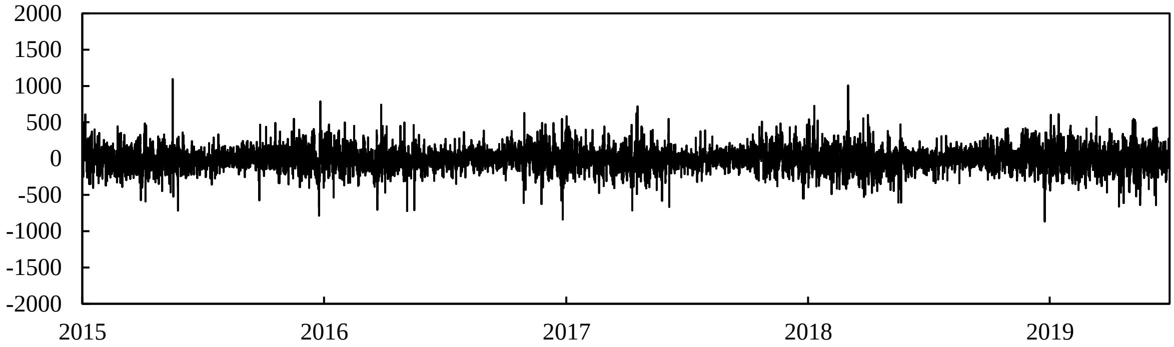

Figure 20: Graph of SARIMAX(0,0,0)(7,0,7)$_{24}$ interaction estimation residuals

Testing the residual autocorrelating formally with a the Ljung-Box test, the SARIMAX interaction model's test statistic is 143 667, improving slightly compared to the SARIMAX main model.

6.3. Multilayer perceptron regressor

To create a feed forward neural network model we use the Multi-Layer Perceptron Regressor (MLPRegressor) from the library of scikit-learn, (Pedregosa et al. 2011). This is a supervised learning version of a multilayer perceptron neural network. The model is built in Python.

The first step of building the model is importing and adapting the data. For this thesis, predictions for two separate timespans are made. One set of predictions are made using a model trained on data from 01.01.2015 to 30.06.2019, and the second set of predictions are made using training data from 01.01.2015 to 30.06.2021. The validation period for both models are 01.07. to 31.12, in their respective end-of-training years. Apart from the differing length of training data timespan, each model is built using the same steps as follows.

First, the datasets are scanned for missing- and non-numerical values. Each variable is adapted as to be a float value. For true false variables the value is set as a Boolean 0 or 1 representation. The datasets contain a small number of missing values in some exogenous variables which is replaced by the mean value of the variable. For larger sets of missing values, the replacement strategy should be reassessed. However, because of how few missing values the dataset contained the mean replacement strategy proved not to compromise the final results and is thus selected.

The multilayer perceptron model often benefits from normalizing or standardizing the data. This is done by using scaling functions from the scikit-learn (Pedregosa et al. 2011) library. To properly scale data, the training and validation data is separately scaled, not allowing information leakage between the two sets of data. The scaler functions are fit on the training data before the same scale is applied to the validation data. Fitting the scaler being the process of the function learning how to scale the data to the given range. The predictions are made on the scaled data before the results are inverted to actual values. This process of scaling data shows no improved results and is consequently not used for the final predicating model of the thesis.

The structure and parameters of the MLPRegressor are selected by extensive testing. The best results are found using three hidden layers with 64, 32 and 16 nodes. The solver used is ‘adam’, the activation function ‘ReLU’ and maximum iterations of 1000 although the model mostly converge well before this.

The selection of exogenous variables is made on the basis of previous literature, as well as the statistical analysis of the load time series shown in previous chapters.

From the literature we know that weather variables are common inputs in load forecasting models. For the machine learning models, temperature and humidity are used. In addition to these, heating degrees are computed. While the SARIMAX models benefited more from squared temperature, the machine learning models marginally prefer heating degrees. Heating degrees are defined here as the number of degrees Celsius below 15,5. The final weather variables used are a set of minimum and maximum temperature over the past 24- and 48 hours values.

From the analysis in previous chapters, we have shown a yearly, weekly, and daily seasonality in the load data. This is represented in the input values by dummy variables for months, weekdays and for the hour of the day. Additionally, there is a dummy variable for the year, and one for public holidays. The load-series are also found to have autoregressive properties. To account for the predictive value of previous timesteps, lagged load variables is added to the model. After testing, the effects are best captured using the load 24, 48, and 168 hours in advance. The maximum and minimum load of the past day is also used, by this we mean the period starting 48 ending 24 hours before the predicted timepoint. This gives us the final inputs of the model, shown in the table 7 below.

Table 7: Overview of exogenous variables used in ML models

Variable	Input type
Year	Dummy
Month	Dummy
Day	Dummy
Hour	Dummy
Holiday	Dummy
Temperature	Value
Humidity	Value
Price*	Value
Heating Degrees	Value
Max temp - 24 hours	Value
Max temp - 48 hours	Value
Min temp - 24 hours	Value
Min temp - 48 hours	Value
Average temp - 24 hours	Value
Max load - day before	Value
Min load - day before	Value
Load - t-24	Value
Load - t-48	Value
Load - t-168	Value

** Price only used for a selection of forecasts*

6.4. Decision Tree- and Random Forest regressor

The decision tree and random forest models both use the same set of input variables as described in the multilayer perceptron regression method, and which are found in table 7. The data processing is also the same, to ensure comparable results. The only difference being that the decision tree and random forest models does not benefit from scaling the data, meaning the scaling step is skipped. This is one of the arguments for these types of models, the ease of use, where the model can take most input data as is.

For the decision tree model, scikit-learns (Pedregosa et al. 2011) ‘LBGMRegressor’ is used. This is a gradient boosting decision tree model, which has been known to outperform the bagging method of optimizing of random forests that was described in chapter 4.2.2. The boosting type used is ‘gbdt’, a gradient boosted decision tree. For the number of gradient-boosted trees parameter, the ‘n_estimators’, the model uses 100.

The random forest model utilizes the ‘RandomForestRegressor’ from scikit-learn (Pedregosa et al. 2011). This model is constructed using the number of estimators, or number of trees in the decision forest, of 1000.

7. Results

In this chapter the results of the forecasting results will be presented. The chapter is divided in three main parts.

In the first part the results of the six models developed in this thesis will be displayed, with one subpart for each of the five Norwegian bidding zones. In the second part, two overall best performing forecasts, one statistical and one machine learning method will be compared to the official forecast of Entso-E. The comparison is done for all bidding zones and areas. The third part shows the results from the analysis of the power price’s influence on forecasting performance. Two analyses are performed, one where the day-ahead price is used as a variable in two models, compared to the same models without price as a variable. In the second analysis we compare the forecasting performance of the two best models in the second half of 2019 with 2021.

7.1. Comparison of forecasting methods

In this subchapter the forecasting results for the autumn of 2019 and 2021 for all models developed in this thesis are displayed. The models are compared and evaluated in terms of average accuracy and an evaluation of the model’s errors. The metrics shown are the average performance of each model measured in both MAPE and RMSE, and the error range of each model measured in the extreme points of error in both directions. The best statistical and the best machine learning models measured in average performance for the autumn of 2021 period will be used to present a graph of the forecast residuals. Residuals for models not shown in this chapter can be found in appendix 2. The last element is a comparison of the best two models to the actual load, over two selected weeks. The weeks selected are the first full week of July, Monday through Sunday 05.07.21 – 11.07.21, and the first full week of December, Monday through Sunday 06.12.21 – 12.12.21. The weeks are selected to showcase the model performance under different seasonal circumstances.

7.1.1. Forecasting results for NO1

The forecasting results for NO1 is displayed in table 8 which shows the average error metrics for all forecasts both in MAPE and RMSE. All models perform better than the naïve model indicating that the more complex models are able capture more information. Both in terms of MAPE and RMSE the SARIMAX model with interaction performed the best of the models designed in this thesis, while SARIMAX with only main effects is second. Of the machine learning methods, the Decision tree model achieves the best accuracy, closely behind

SARIMAX main. The ANN model MLPRegressor performed by far the poorest for 2021, but not as far behind the rest of the methods for 2019. A noteworthy point for the MLPRegressor is that it is the only method for which the forecast accuracy was worse in 2021 compared to the 2019 results.

Table 8: Forecasting accuracy for NO1, 01.07. – 31.12. of 2019 and 2021.

	Autumn 2019		Autumn 2021	
Model	**MAPE**	**RMSE**	**MAPE**	**RMSE**
Naive	6,15 %	345,51	5,61 %	303,18
SARIMAX main	2,93 %	147,11	2,56 %	128,05
SARIMAX interaction	**2,60 %**	**131,37**	**2,40 %**	**125,27**
Decision tree	2,97%	158,50	2,67%	145,34
Random forest	3,31%	179,34	2,83%	151,38
MLPRegressor	3,24%	163,21	3,92%	212,71

In table 9 the extreme errors of the models are presented. The Decision tree model is the best performing model measured by the smallest range of errors in 2021. The largest negative error was 14,63% below the actualized load and the largest positive error of 13,95% above the actualized load, resulting in a range of 28,58 percentage points (pp). The ranges of the other four complex models are at similar levels in the low- to mid-thirties, while the naïve model has the widest error range by some margin. The most interesting point to take note of is that the SARIMAX with interactions which was the best model measured by average forecasting accuracy over the period was not the model with the smallest error range.

Table 9: Extreme forecasting errors NO1 for 2021

Model	**Low**	**High**	**Range**
Naive	-33,13 %	35,91 %	69,03 pp
SARIMAX main	-21,13 %	12,94 %	34,08 pp
SARIMAX interaction	-18,49 %	14,76 %	33,24 pp
Decision tree	-14,63 %	13,95 %	**28,58 pp**
Random forest	-16,19 %	19,65 %	35,85 pp
MLPRegressor	-13,93 %	21,76 %	35,69 pp

The extreme errors are however, only one part of the picture. To provide a more complete view of the error distribution the residuals of the two best models, the SARIMAX with interactions- and the Decision tree model, are graphed in figure 21 and 22. What is evident looking at the residuals graph is that the SARIMAX interactions suffers in the extreme errors measure from a small number of larger outlier errors. The overall distribution of the errors however is better than the Decision tree model, as previously shown by the lower root mean squared error.

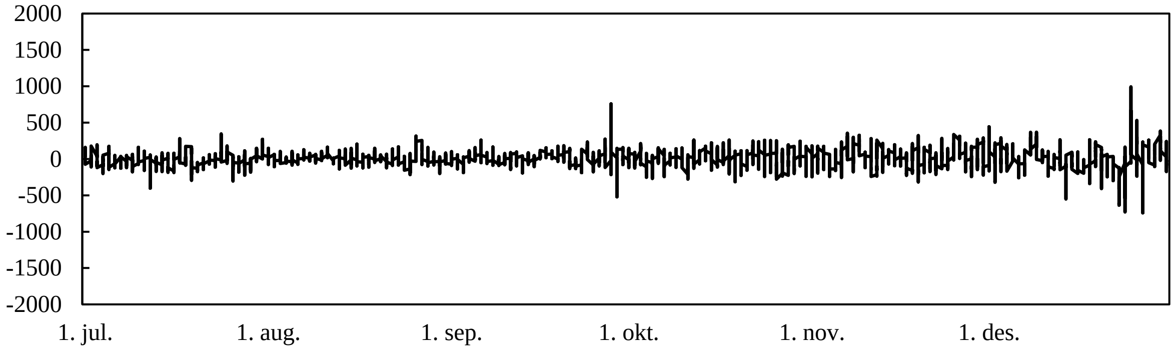

Figure 21: Residuals from SARIMAX interactions forecast of load in NO1 01.07.21 - 31.12.21

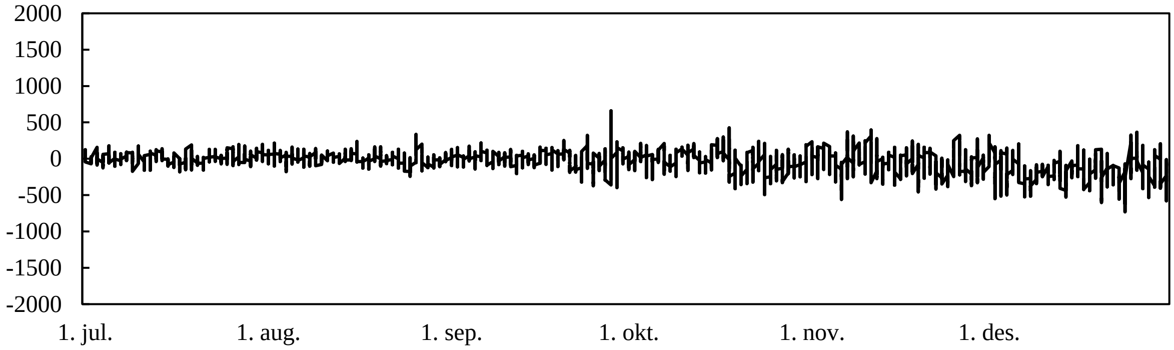

Figure 22: Residuals from Decision tree forecast of load in NO1 01.07.21 - 31.12.21

From the graphing of both models, it is shown there is an increase in the residuals towards the end of the period. This is due to the effect of higher absolute levels of load during the winter months of the end of the period compared to the summer months at the beginning. As figure 23 shows, the residuals measured in percentage of actual load sees a more consistent distribution throughout the complete period.

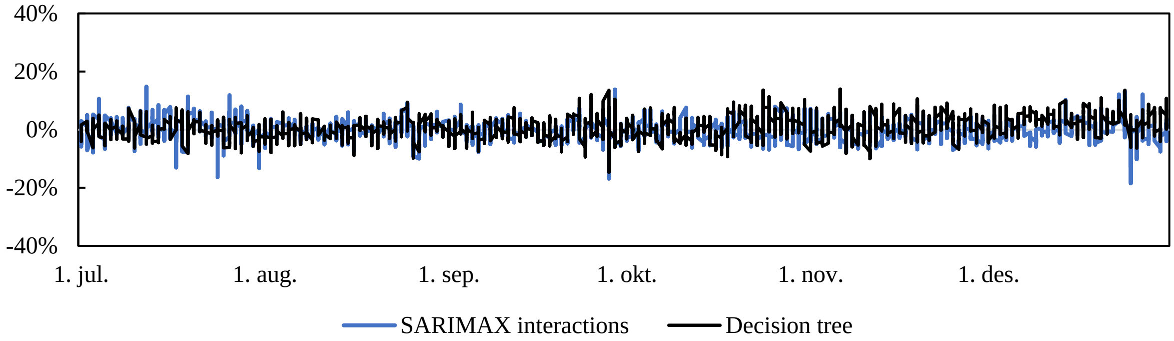

Figure 23: Percentage residuals for Decision tree and SARIMAX interaction 01.07.21 - 31.12.21

Figures 24 and 25 shows the load prediction for SARIMAX interactions and Decision tree, the first full weeks of July and December of 2021. For the July week in figure 24 the visualization shows that both the SARIMAX interactions- and Decision tree models manage to capture the trends well throughout the working days, while the Decision tree model miss the peak hours of the weekend in a larger degree than the SARIMAX interactions model.

For the selected December week in figure 25 both the SARIMAX interactions- and the Decision tree model have a larger error term during Tuesday through Thursday, while improving during the weekend. Interestingly, the SARIMAX interactions model undershoots the observed load more, while the Decision tree model overshoots it. However, both models seem to capture the general directions and trends of the week.

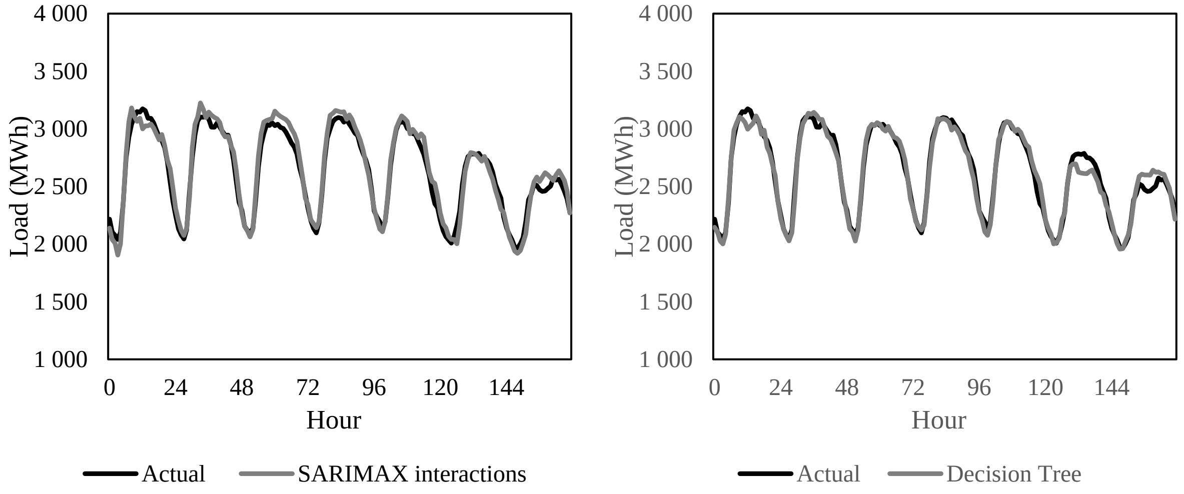

Figure 24: Actual load and forecasted load for one week in July 2021 (05.07.21 - 11.07.21)

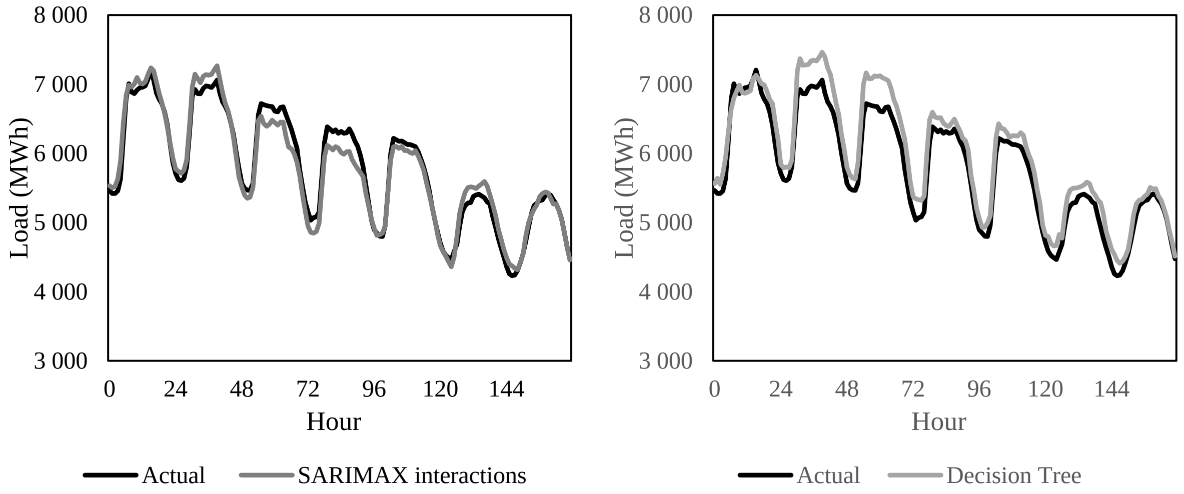

Figure 25: Actual load and forecasted load for one week in December 2021 (06.12.21 - 12.12.21)

7.1.2. Forecasting results for NO2

The forecasting results for NO2 is somewhat similar to the results for NO1. The SARIMAX models are still performing among the best, but the difference between the two models is smaller than for NO1. In this case the cross effects added in SARIMAX interactions seems to have a small positive effect in 2019 and a negative influence on the performance in 2021. Among the machine learning models, The Decision tree model also performs best in this bidding zone. In terms of both MAPE and RMSE the Decision tree model performs as well as the worst performing SARIMAX in both periods. The MLPRegressor also perform better in comparison to the other models for NO2, than it did for NO1. The interpretation of these results can also be seen in relation to the performance of the naïve model. Compared to NO1 the naïve model has a 2 percentage points lower MAPE, indicating that the average change in load from one day to another is lower for NO2.

Table 10: Forecasting accuracy for NO2

	Autumn 2019		Autumn 2021	
Model	**MAPE**	**RMSE**	**MAPE**	**RMSE**
Naive	4,01 %	227,52	3,89 %	222,94
SARIMAX main	2,38 %	130,75	**2,43 %**	**132,90**
SARIMAX interaction	**2,31 %**	**125,27**	2,55 %	144,74
Decision tree	2,40%	129,14	2,54%	139,16
Random forest	2,66%	143,43	2,74%	149,73
MLPRegressor	2,66%	143,32	2,67%	146,07

The extreme errors in the predicted load of each model for NO2 shows that the smallest range belongs to SARIMAX main. Unlike the results for NO1, it is the model with the best results on average over the period, which also performs best measured in extreme errors. The difference between the ranges of the complex methods, however, are small and arguably negligible in NO2. The lower average error of the naïve model also clearly shows in the error range, with the model being far closer to the error ranges of the complex models in NO2 than it was for NO1. As the naïve model is built by assuming the load is equal to the load 24 hours before, the error range of the naïve model is a measure of the largest changes of load of the same hour from day to day during the period.

Table 11: Extreme forecasting errors NO2 for 2021

Model	Low	High	Range
Naive	-20,32 %	28,29 %	48,61 pp
SARIMAX main	-19,91 %	12,87 %	**32,78 pp**
SARIMAX interaction	-20,19 %	14,37 %	34,56 pp
Decision tree	-21,08 %	12,65 %	33,73 pp
Random forest	-20,11 %	14,00 %	34,12 pp
MLPRegressor	-20,19 %	14,53 %	34,72 pp

From the graphing of the SARIMAX main- and Decision tree models in figures 26 and 27, it looks as though the prediction series for NO2 is less defined by a small number of extreme errors. Both models do show a larger than usual point of error at the start of the prediction series.

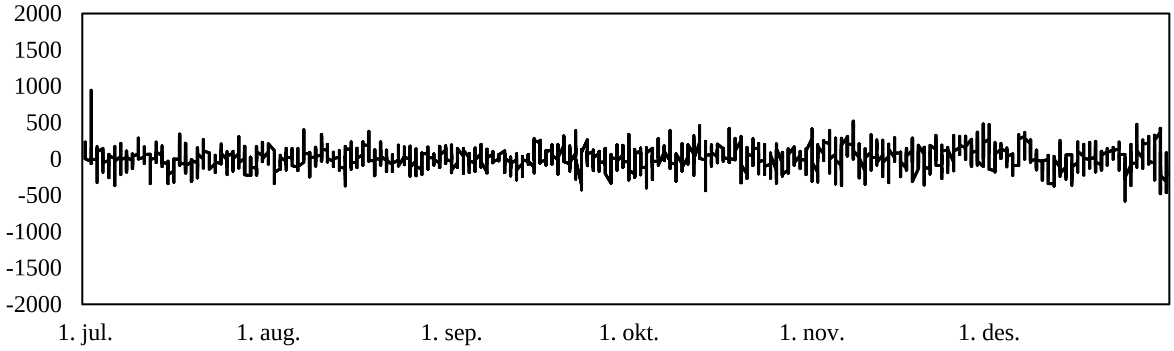

Figure 26: Residuals from SARIMAX main forecast of load in NO2 01.07.21 - 31.12.21

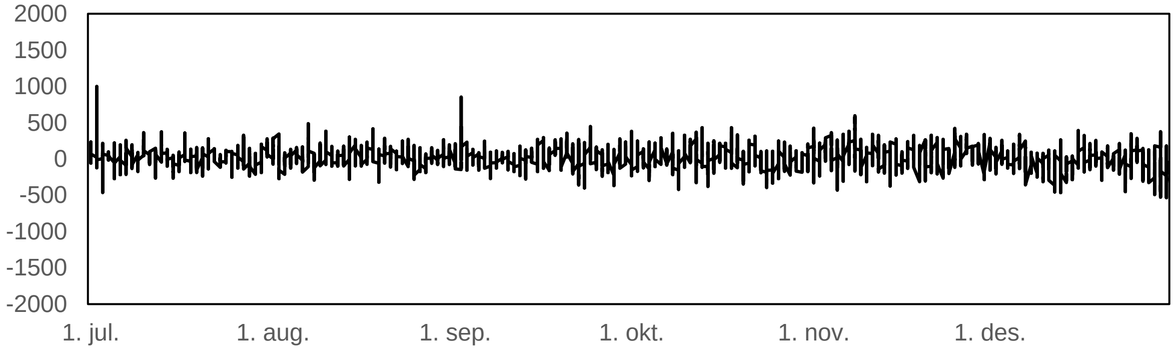

Figure 27: Residuals from Decision tree forecast of load in NO2 01.07.21 - 31.12.21

Comparing the predictions to the actual load for the first full week of July in figure 28, it is clear that both the SARIMAX main and Decision tree are able to predict the general trend of the load. They do however fail to accurately predict the finer variations, especially at peak hours. This might be a sign of factors not included in the models, which affect the load during the daily peaks. Both models also predict lower loads the Saturday in question. For the December week in figure 29 the models cannot explain the peak of Monday, and the night-time load of Saturday to Sunday. For the remaining days of the week both models capture the general trends and shape of the load curve quite closely.

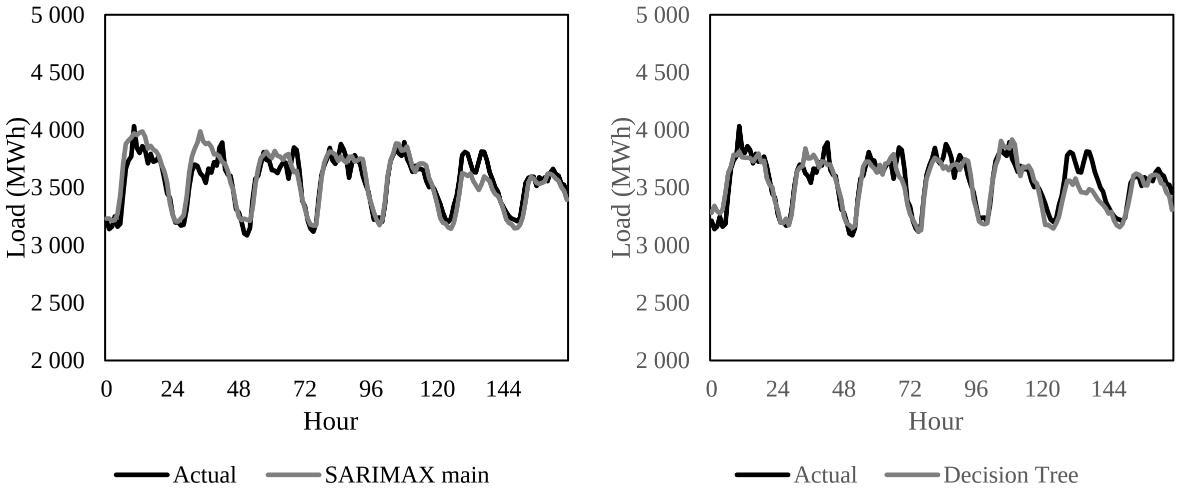

Figure 28: Actual load and forecasted load for one week in July 2021 (05.07.21 - 11.07.21)

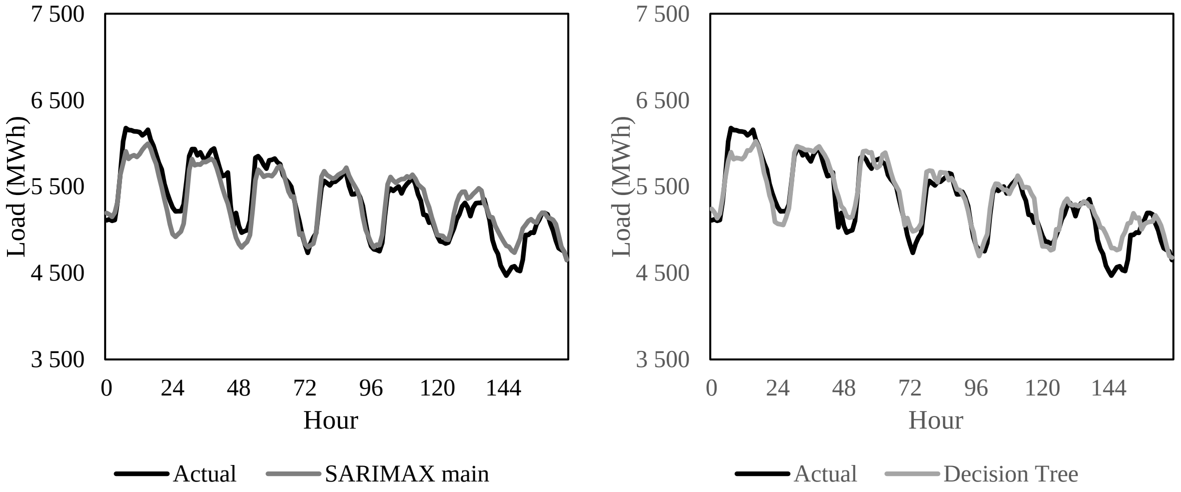

Figure 29: Actual load and forecasted load for one week in December 2021 (06.12.21 - 12.12.21)

7.1.3. Forecasting results for NO3

The forecasting results for NO3 resembles the results for NO1 and 2 in terms of ranking the model's performance. Again, the SARIMAX models are on top with both the lowest MAPE and RMSE, but this time the SARIMAX with interaction performs best in both periods. Among the machine learning models the Decision tree and MLPRegressor perform the best, closely behind the SARIMAX main model. Furthermore, The Random Forest model is the least accurate in both periods, in terms of MAPE and RMSE. For all models, the consistent relationship between MAPE and RMSE in both periods is an indication that no model has abnormally large outlier errors. The naïve model's forecast performs poorer than the complex models, with a MAPE of about one percentage point higher than the complex models.

Table 12: Forecasting accuracy metrics for NO3

	Autumn 2019		Autumn 2021	
Model	**MAPE**	**RMSE**	**MAPE**	**RMSE**
Naive	3,95 %	153,58	4,26 %	171,33
SARIMAX main	2,76 %	103,98	3,07 %	119,39
SARIMAX interaction	**2,75 %**	**103,90**	**2,96 %**	**116,38**
Decision tree	2,82 %	106,68	3,10 %	122,74
Random forest	2,99 %	114,54	3,43 %	136,11
MLPRegressor	2,77 %	105,30	3,12 %	124,02

For the extreme errors, the SARIMAX main has the lowest range despite performing worse on average than the SARIMAX interaction model. Once again, it is shown that the error ranges of the five complex models are comparable in size, from 34 to 38 percentage points. The naïve model has a range not much higher, showing that, just as for NO2, the load of the same hour the day before was at all times during the period in somewhat close approximation of the load.

Table 13: Extreme forecasting errors NO3 for 2021

Model	**Low**	**High**	**Range**
Naive	-19,73 %	23,69 %	43,42 pp
SARIMAX main	-12,76 %	21,18 %	**33,95 pp**
SARIMAX interaction	-13,49 %	21,72 %	35,21 pp
Decision tree	-13,68 %	23,47 %	37,15 pp
Random forest	-16,50 %	21,09 %	37,58 pp
MLPRegressor	-13,41 %	21,08 %	34,48 pp

Figures 30 and 31 shows that there was for neither series a small number of abnormally large errors. It rather shows that the extreme error range is a more representative measure of the range of errors throughout the full period, than it was for the NO1 and NO2 predictions. As for the other prediction series, the residuals are larger during the latter part of the forecasting period, which is due to the higher levels of absolute load as shown in the NO1 section (figure 23).

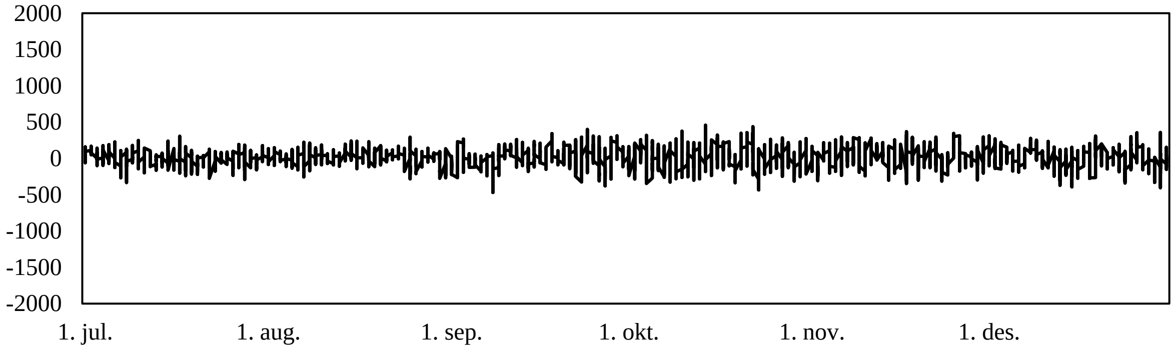

Figure 30: Residuals from SARIMAX interaction forecast of load in NO3 01.07.21 - 31.12.21

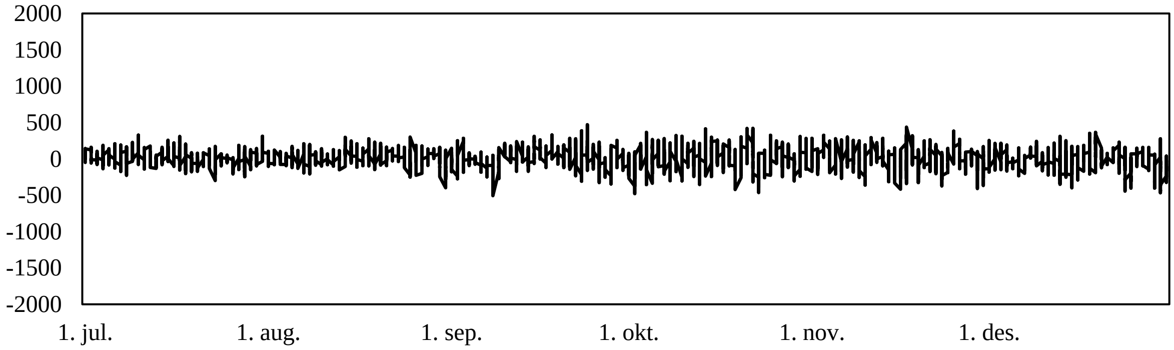

Figure 31: Residuals from Decision tree forecast of load in NO3 01.07.21 - 31.12.21

The July week (figure 32) shows the same tendencies for NO3 as it did for NO2. Both the SARIMAX interactions- and the Decision tree model capture the daily variations, while having some issues with predicting the variations at peak times of day. The SARIMAX interactions model outperforms the Decision tree model for the prediction of the Saturday, where the Decision tree err on the low side. For the December week in figure 33, the performance is reversed, where the Decision tree model seems to be a good forecast, while the SARIMAX interactions model has some more pronounced errors for Wednesday and Thursday.

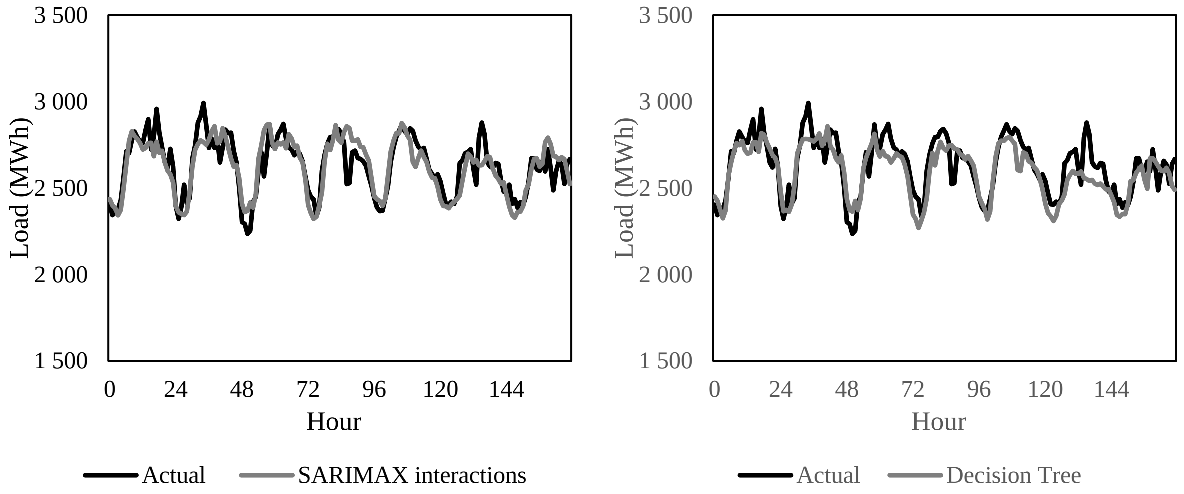

Figure 32: Actual load and forecasted load for one week in July 2021 (05.07.21 - 11.07.21)

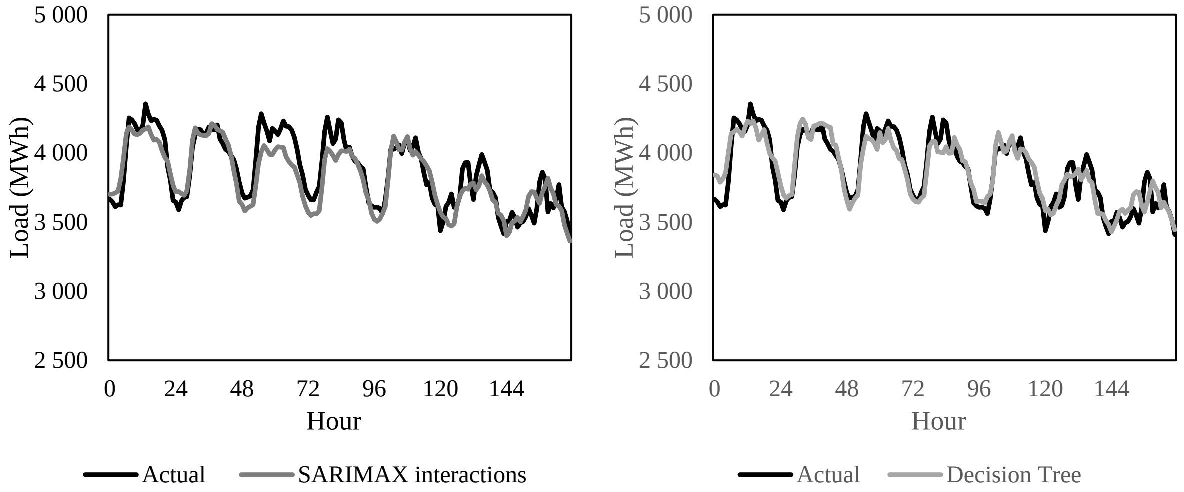

Figure 33: Actual load and forecasted load for one week in December 2021 (06.12.21 - 12.12.21)

7.1.4. Forecasting results for NO4

The forecasting performance in NO4 do somewhat deviate from the previously shown bidding zones. The similarities are the ranking of forecast, which is in line with previous bidding zones, where the SARIMAX models and Decision tree are the top performing models. In this bidding zone, the SARIMAX with main effects is more accurate than the SARIMAX with interactions in both periods.

In contrast to the previous bidding zones, there is a large increase in the error metrics across all models from 2019 to 2021. The 2021 forecasts for NO4 have the worst average performance out of all bidding zones forecasted. This indicates that the training period prior to 2019 resembles the autumn of 2019 closer than what the respective training period does for the autumn of 2021. This results in models that is fitted less accurate to predict the load in the autumn of 2021 than for 2019. Furthermore, the relationship between MAPE and RMSE looks to be inherently the same for both periods. This indicates that the cause of the increase is not likely to result from an increase in a few large deviations from the actual load, but higher errors on average across the period.

Table 14: Forecasting accuracy metrics for NO4

	Autumn 2019		Autumn 2021	
Model	**MAPE**	**RMSE**	**MAPE**	**RMSE**
Naive	3,91 %	106,79	4,92 %	132,53
SARIMAX main	2,79 %	74,36	**3,61 %**	**94,73**
SARIMAX interaction	2,83 %	74,82	3,69 %	97,75
Decision tree	**2,75 %**	**74,05**	3,74 %	97,18
Random forest	2,87 %	77,63	3,95 %	101,26
MLPRegressor	2,95 %	77,95	3,81 %	97,71

With the average accuracy of the models in 2021 suffering, there is also a markedly higher extreme errors range as well. The range of errors in NO4 is higher than for any other bidding zone. Interestingly, the most accurate forecasting models on average was not the ones with the lower error ranges. The MLPRegressor while being only fourth best on average, is the model with the smallest range.

Table 15: Extreme forecasting errors NO4 for 2021

Model	Low	High	Range
Naive	-37,83 %	53,02 %	90,85 pp
SARIMAX main	-28,01 %	31,40 %	59,41 pp
SARIMAX interaction	-28,01 %	28,97 %	56,97 pp
Decision tree	-19,47 %	35,68 %	55,14 pp
Random forest	-19,17 %	38,42 %	57,59 pp
MLPRegressor	-17,34 %	34,38 %	**51,72 pp**

The graphing of the residuals for the SARIMAX main- and Decision tree models in figure 34 and 35 shows no marked period of higher errors. The accuracy is stable, but lower than for the other zones throughout. The errors are also in line with the relationship between MAPE and RMSE showing very few abnormally large errors. Keep in mind that NO4, the zone covering northern Norway, has the lowest average load of all five zones. As such the residuals for NO4, while lower, are when measured in percentage higher than for other zones.

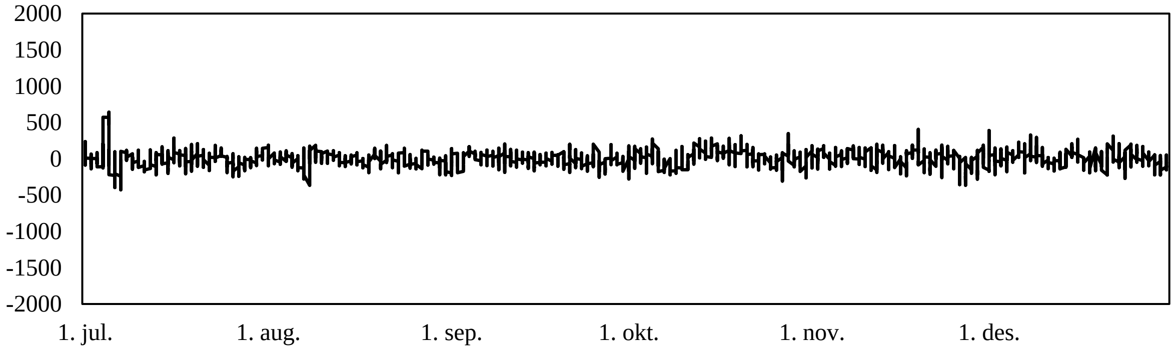

Figure 34: Residuals from SARIMAX main forecast of load in NO4 01.07.21 - 31.12.21

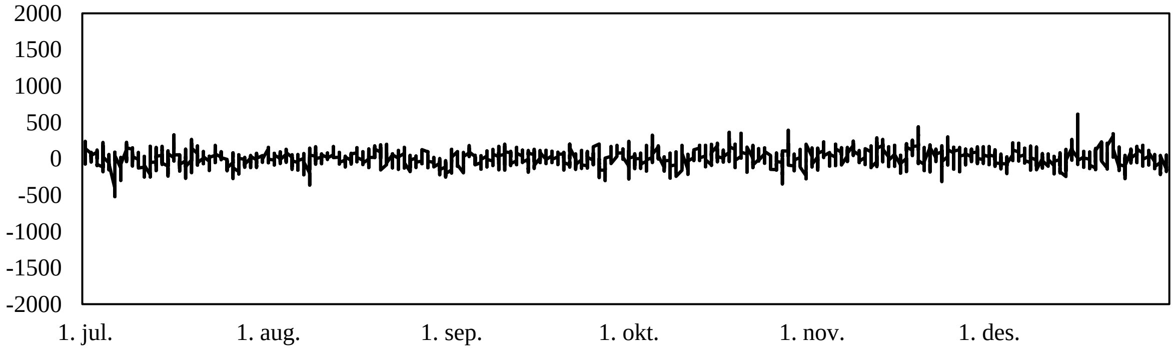

Figure 35: Residuals from Decision tree forecast of load in NO4 01.07.21 - 31.12.21

Figure 36 shows the forecasted and actual load for the first week of July and it is clear the model has some issues. The night between Monday and Tuesday both models predict a high peak in load, while the actual load was falling, as is usual at nighttime. Nothing in the preceding days load or the weather data seem to explain this deviation. For the rest of the week the predicted movements are not as pronounced in the opposite direction, but the models seem unable to capture the changes in load still. For the December week shown in figure 37, both models are markedly better. While there still are some errors, the predicted load follows the actual load in direction throughout the week, providing reason to believe that the models are working albeit worse for some weeks of the period.

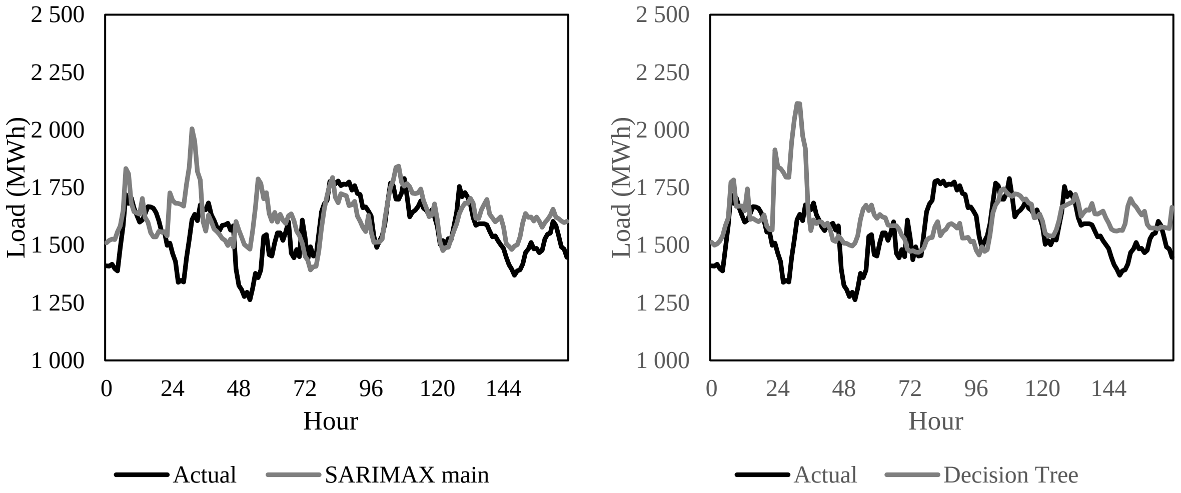

Figure 36: Actual load and forecasted load for one week in July 2021 (05.07.21 - 11.07.21)

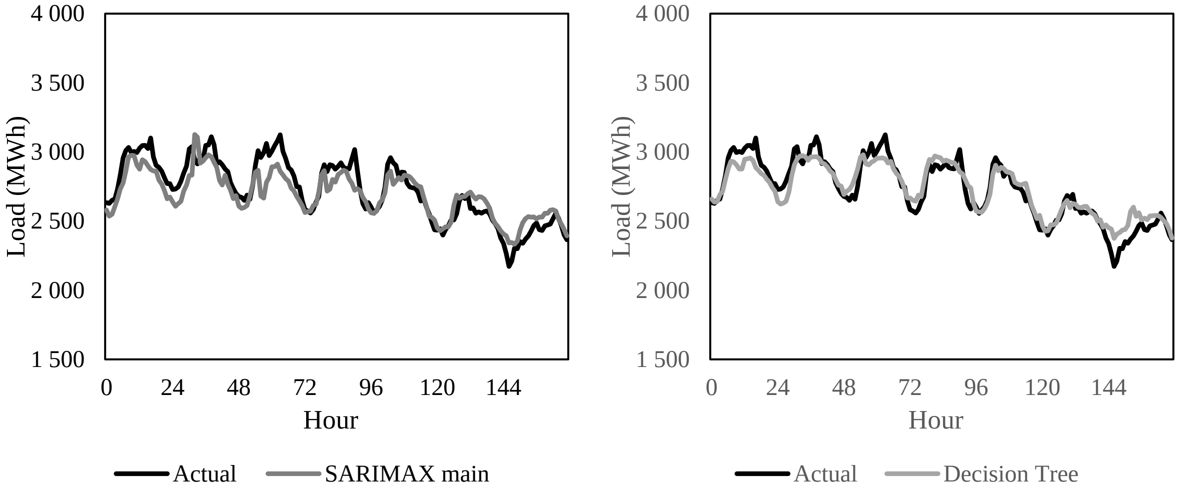

Figure 37: Actual load and forecasted load for one week in December 2021 (06.12.21 - 12.12.21)

7.1.5. Forecasting results for NO5

Once again, the rankings of the models are consistent with previous bidding zones, with the SARIMAX models performing best, followed by MLPRegressor and Decision tree. The Random Forest model is again among the worst performing models. Furthermore, interestingly all models have about 1 percentage point higher MAPE in 2019 compared to 2021. This is the opposite of the results in NO4.

Table 16: Forecasting accuracy metrics for NO5

	Autumn 2019		Autumn 2021	
Model	**MAPE**	**RMSE**	**MAPE**	**RMSE**
Naive	5,11 %	114,48	4,07 %	104,12
SARIMAX main	3,85 %	81,51	2,84 %	**68,69**
SARIMAX interaction	**3,78 %**	**80,66**	**2,75 %**	69,21
Decision tree	4,11 %	85,81	3,06 %	74,03
Random forest	4,11 %	87,01	3,27 %	79,93
MLPRegressor	3,99 %	83,94	3,01 %	73,92

For the extreme errors measure of NO5 table 17 indicates that the fluctuations have been large. The ranges are not as large as they were for NO4, while still being high compared to the first three zones. Again, we see that the best performing model on average, the SARIMAX interactions was beaten in the extreme errors measure. For NO5 the top three is comprised of the machine learning models.

Table 17: Extreme forecasting errors NO5 for 2021

Model	**Low**	**High**	**Range**
Naive	-26,78 %	34,20 %	60,98 pp
SARIMAX main	-17,93 %	37,38 %	55,31 pp
SARIMAX interaction	-15,32 %	36,21 %	51,52 pp
Decision tree	-15,80 %	30,37 %	**46,18 pp**
Random forest	-19,69 %	29,38 %	49,06 pp
MLPRegressor	-16,44 %	30,01 %	46,45 pp

The residuals in figure 38 and 39 indicates no clear period of larger errors, except the increase of residuals with higher levels of load at the latter part of the series. There does however seem to be some error spikes at the start of August and mid-September, as well as during the December month for both models.

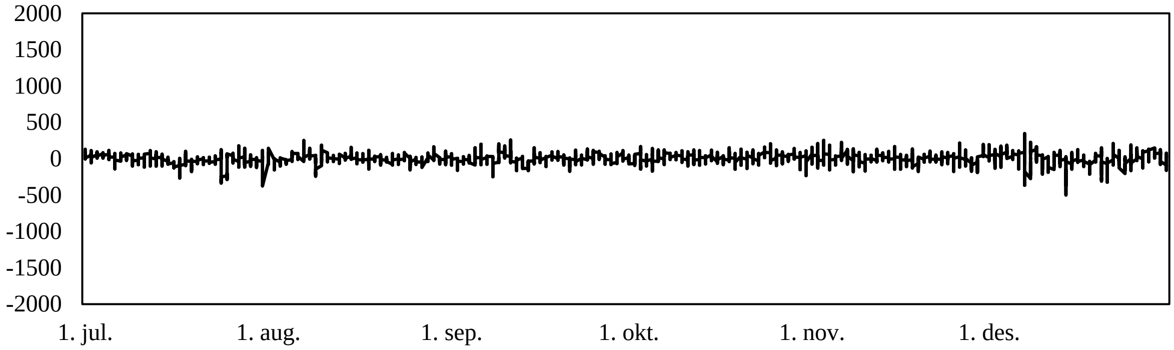

Figure 38: Residuals from SARIMAX interaction forecast of load in NO5 01.07.21 - 31.12.21

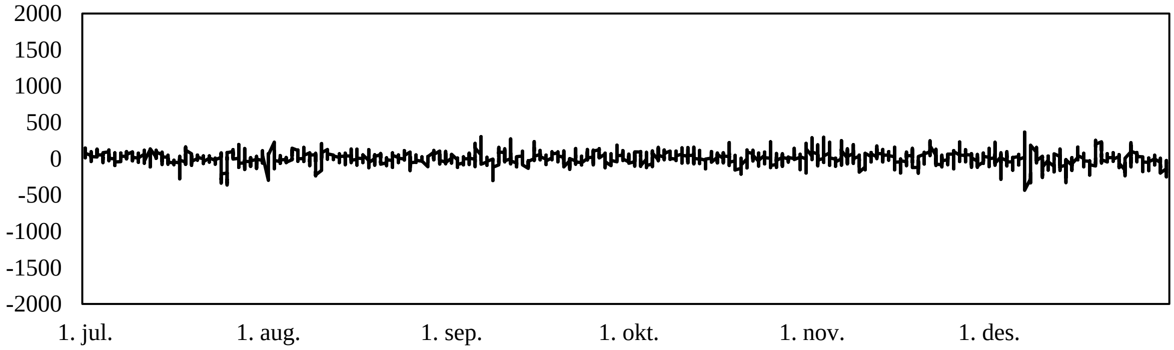

Figure 39: Residuals from MLPRegressor forecast of load in NO 01.07.21 - 31.12.21

Figures 40 and 41 shows that the models are closely fit to the actual load for the most part of both the July and December week. There is however an unusual load spike the Tuesday of the December week and a larger nighttime drop, in addition to a dip in the peak hours of Friday, which both models are unable to account for.

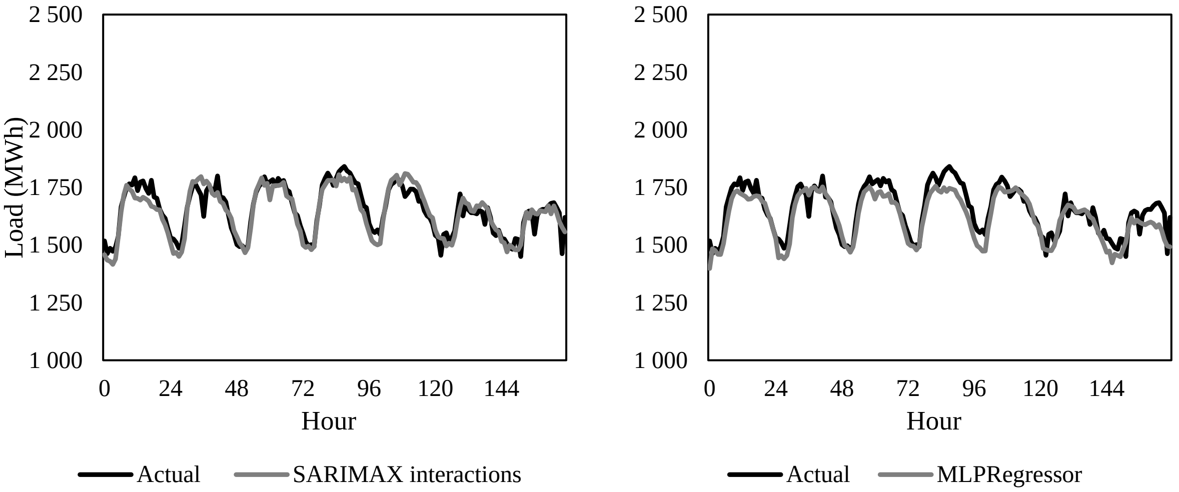

Figure 40: Actual load and forecasted load for one week in July 2021 (05.07.21 - 11.07.21)

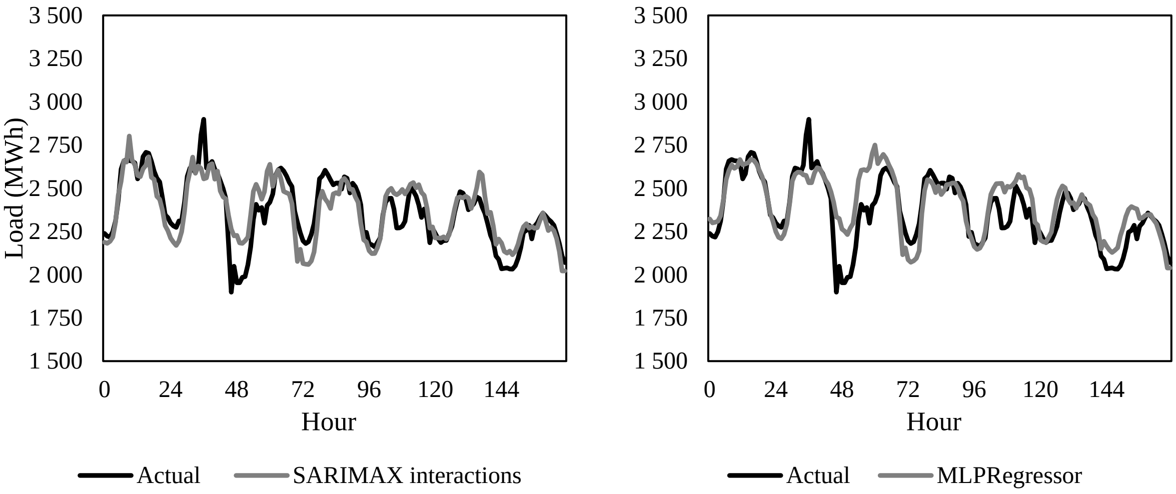

Figure 41: Actual load and forecasted load for one week in December 2021 (06.12.21 - 12.12.21)

7.2. Forecast compared to Entso-E

This subchapter presents the forecasting results of the overall best statistical and machine learning model in comparison to the Entso-E's Day-ahead forecast. Our findings are surprising as the official forecast Entso-E performs consistently poor in NO2-NO5 for both periods. Comparing results for the other biding zones is of little value as Entso-E perform poorer than the naïve benchmark. The reason for this is unknow, and any attempt at unfolding the reason this would be speculative. In table 18 the results are displayed for all bidding zones, time periods and models.

Table 18: Entso-E, SARIMAX interaction and Decision tree forecast performance

Bidding zone	Model	Autumn 2019		Autumn 2021	
		MAPE	**RMSE**	**MAPE**	**RMSE**
NO1	SARIMAX interaction	2,60 %	**131,37**	2,40 %	125,27
	Decision tree	2,97 %	158,5	2,67 %	145,34
	ENTSO-E	**1,94 %**	164,32	**1,79 %**	**105,61**
NO2	SARIMAX interaction	**2,31 %**	**125,27**	2,55 %	144,74
	Decision tree	2,40 %	129,14	**2,54 %**	**139,16**
	ENTSO-E	7,44 %	326,17	14,25 %	641,47
NO3	SARIMAX interaction	**2,75 %**	**103,9**	**2,96 %**	**116,38**
	Decision tree	2,82 %	106,68	3,10 %	122,74
	ENTSO-E	6,44 %	201,5	8,77 %	304,63
NO4	SARIMAX interaction	2,83 %	74,82	**3,69 %**	97,75
	Decision tree	**2,75 %**	**74,05**	3,74 %	**97,18**
	ENTSO-E	7,18 %	194,8	6,38 %	148,08
NO5	SARIMAX interaction	**3,78 %**	**80,66**	**2,75 %**	**69,21**
	Decision tree	4,11 %	85,81	3,06 %	74,03
	ENTSO-E	7,92 %	193,92	14,22 %	293,93

For NO1 the Entso-E forecast perform significantly better in terms of MAPE than the best models designed in this thesis. For the last half of 2019, Entso-E's forecast has a lower MAPE of 0,66pp and 1,03pp compared to the SARIMAX interactions and Decision tree, respectively. The only exception to this is accuracy measured in RMSE, where both the SARIMAX and Decision tree perform better. This indicates that Entso-E does have some large errors and are being punished by RMSE. In the last half of 2021 Entso-E's forecast perform better in both metrics measured, with a lower absolute MAPE of 0,61% and 0,88% compared SARIMAX interactions and Decision tree.

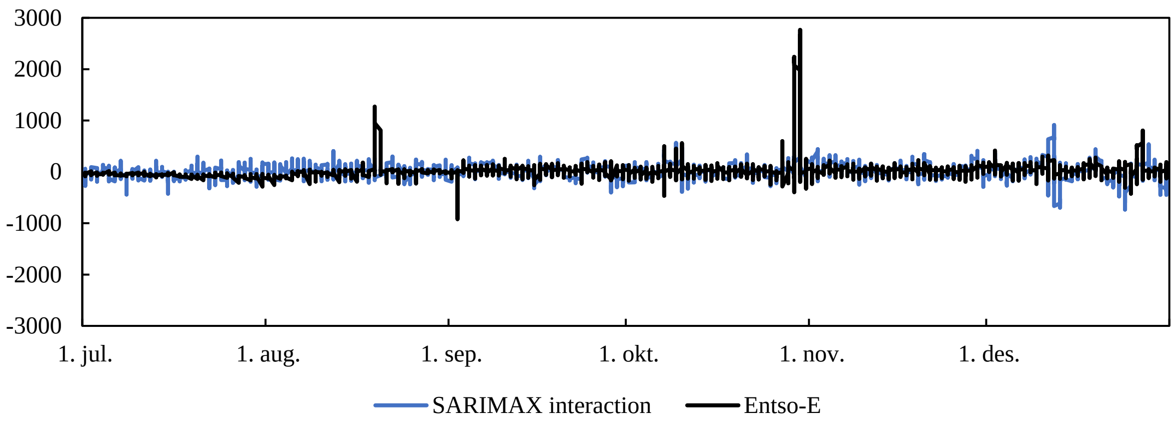

Figure 42: Forecasting errors for SARIMAX interactions and Entso-E, NO1 – 2019

Inspecting the forecasting errors of SARIMAX and Entso-E in NO1 for the last half of 2019 enlightens the reason behind our results. In figure 42 we can observe Entso-Es forecast having large residuals on few occasions, resulting in a high RMSE. The largest errors in the end of October 2019 occurs on Tuesday evening the 29th, as Entso-E severely underpredicts the load for multiple hours, whereas the load for the same period follows the usual pattern. The errors of SARIMAX interaction are one average somewhat higher than that of Entso-E, but its biggest errors are smaller than those of Entso-E. Interestingly, the SARIMAX has consistently sized errors over the whole period, while Entso-E's errors increase over time.

In the five figures below, we show the residuals of the forecast published on Entso-E for the autumn of 2021, compared to our SARIMAX interactions model. Similarly, to what was shown for the autumn of 2019 in NO1, the Entso-E published forecast for NO1 is accurate on average while containing some points of extreme errors. For the bidding zones of NO2 to NO5 the graphing of residuals tells a different story. These four zones have no such pronounced outlier errors. However, the forecast is far inferior to the SARIMAX interactions predictions for the complete period.

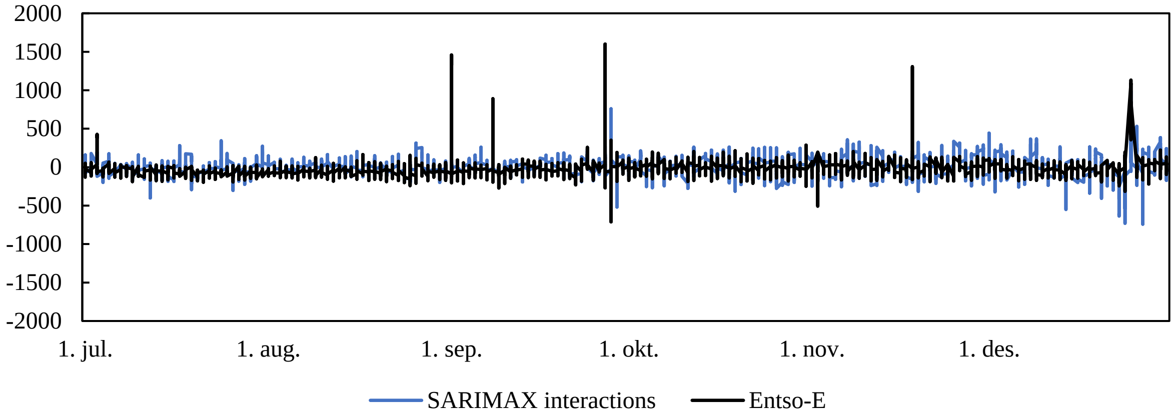

Figure 43: Forecasting errors for SARIMAX interactions and Entso-E, NO1 – 2021

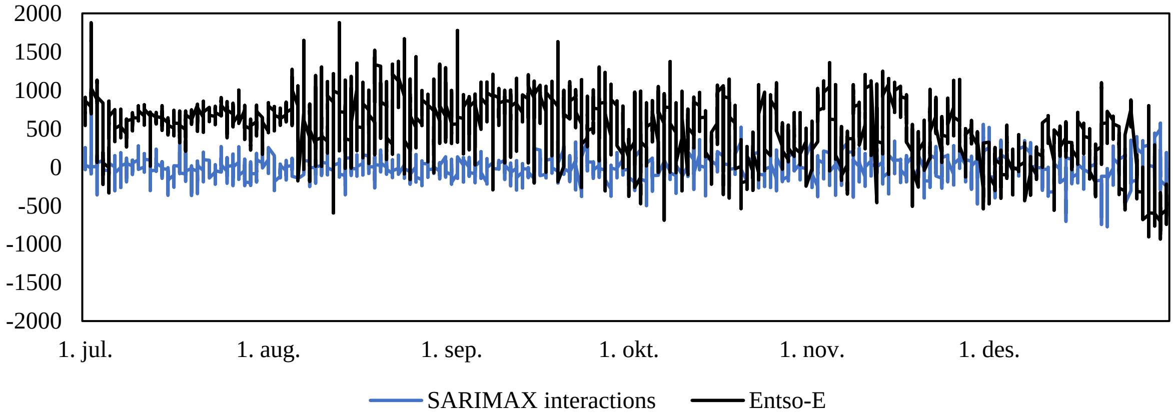

Figure 44: Forecasting errors for SARIMAX interactions and Entso-E, NO2 - 2021

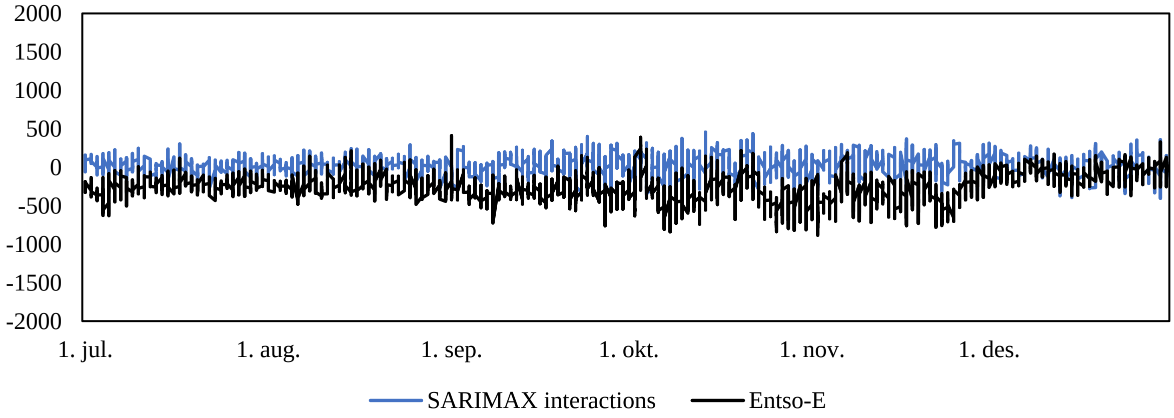

Figure 45: Forecasting errors for SARIMAX interactions and Entso-E, NO3 – 2021

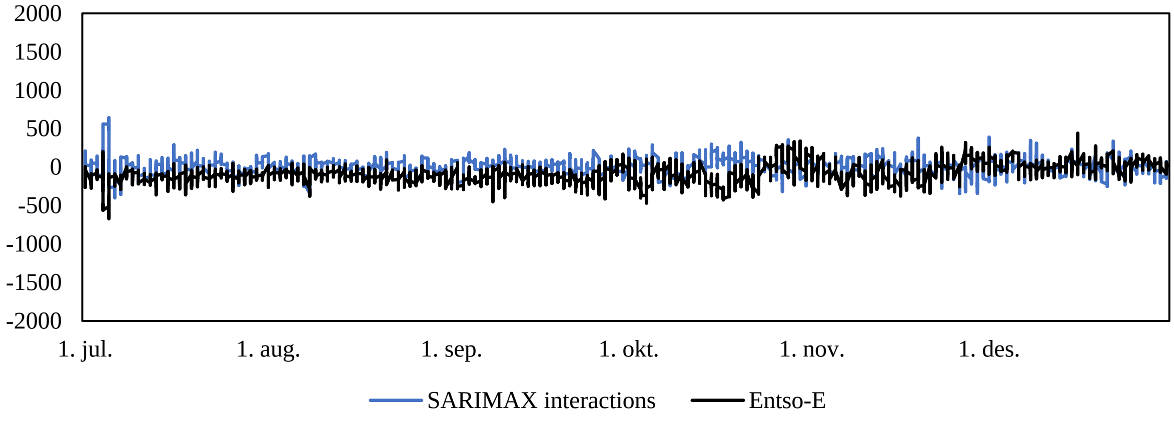

Figure 46: Forecasting errors for SARIMAX interactions and Entso-E, NO4 – 2021

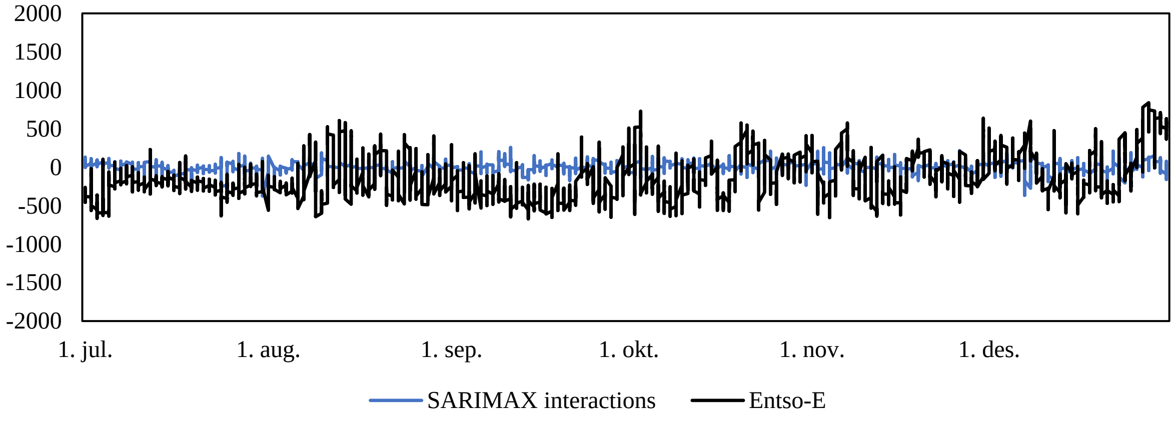

Figure 47: Forecasting errors for SARIMAX interactions and Entso-E, NO5 – 2021

7.3. The effect of price change

This subchapter presents the results regarding our assessment of whether the price increase in 2021 has had an influence on forecasting performance. In this thesis there is conducted two analyses to examine this relationship. The first analysis is to determine if adding the day-ahead price to the model improves the out of sample forecast performance. The result from this analysis performed with the SARIMAX main and Decision tree model is displayed in table 19. The "original" model is without price as a variable, and the "Price incl." model does have price as a variable.

Table 19: Forecasting comparison with and without a price-variable

Bidding Zone	**Model**	Autumn 2019		Autumn 2021	
		Original	**Price incl.**	**Original**	**Price incl.**
NO1	SARIMAX interaction	2,93%	2,93%	2,56%	2,58%
	Decision tree	2,97%	2,96%	2,67%	2,73%
NO2	SARIMAX interaction	2,38 %	2,36 %	2,43 %	2,43 %
	Decision tree	2,40%	2,41%	2,54%	2,56%
NO3	SARIMAX interaction	2,76 %	2,76 %	3,07 %	3,07 %
	Decision tree	2,82 %	2,82%	3,10 %	3,10%
NO4	SARIMAX interaction	2,79 %	2,76 %	3,61 %	3,67 %
	Decision tree	2,75 %	2,73%	3,74 %	3,72%
NO5	SARIMAX interaction	3,85%	3,83%	2,84%	2,94%
	Decision tree	4,11 %	4,16%	3,06 %	2,90%

Adding price as a variable to the both the SARIMAX main and Decision tree model resulted for the most part in close to equally accurate forecasts. This indicate that tomorrow's forecasted price does not provide any additional information. It can also be argued that the forecasted day-ahead price does not affect demand in a way that our original models are unable to pick up.

The second analysis is to test whether our forecast systematically performed better or worse in the last half of 2021 than for 2019. For this analysis we compare our forecasting results for both periods using the best statistical and machine learning models for all bidding zones. The results are presented in table 20, as the difference between the last half of 2019 and 2021. For example, a negative value means that the forecast for 2019 had a lower MAPE than the forecast for 2021.

Table 20: Difference between forecast performance in the fall of 2019 and 2021

Bidding Zone	Model	MAPE	RMSE
NO1	SARIMAX interaction	0,20 pp	6,1
	Decision tree	0,30 pp	13,2
NO2	SARIMAX interaction	-0,24 pp	-19,5
	Decision tree	-0,14 pp	-10,0
NO3	SARIMAX interaction	-0,21 pp	-12,5
	Decision tree	-0,28 pp	-16,1
NO4	SARIMAX interaction	-0,86 pp	-22,9
	Decision tree	-0,99 pp	-23,1
NO5	SARIMAX interaction	1,03 pp	11,5
	Decision tree	1,05 pp	11,8

The results presented in table 20 shows no systematic difference in forecasting performance between the fall of 2019 with no price increase and 2021 with a large price increase. The bidding zones with the largest price increase; NO1, NO2 and NO5 shows no clearly trending difference in performance for 2019 and 2021. The forecasting performance in NO1 and NO5 performs better in 2021, while the opposite is the case for NO2. As for NO3 and NO4 who experienced a lower price increase, the results show that the forecasts did perform better in 2019. The spurious results of this analyze provides no indication that one year is more difficult to forecast. The difference appears to be random and is likely caused by other factors than price.

8. Discussion

In this chapter the results are discussed in relation to the research questions asked in this thesis and compared to earlier findings. The topics are presented in the same order as they were presented in chapter 7. Firstly, a discussion of what model design perform the best, then whether the model design in this thesis is able to outperform Entso-E's official forecast, and lastly, whether or not the recent price increase affect forecast performance.

8.1. Model performance

In the quest to identify the method which best forecasts the electricity demand in the prize zones of Norway, the SARIMAX models have consistently performed among the very best on average accuracy. Whether adding cross effects to the SARIMAX model in addition to the main effects as proposed by Hong et al. (2010) resulted in a better forecast, is ambiguous. For all bidding zones the cross effects either improved the forecasting performance from the SARIMAX main or performed close to equal. The results however do not conclusively show improvement from adding the cross effects. Compared to the results of Elamin & Fukushige (2018) which found a clear improvement in MAPE by adding cross effects, our results are spurious.

Among the machine learning models the Decision tree method consistently performs the best or equal to the other machine learning models. At the other end of the scale the Random Forest model consistently performs the worst. Following logically from the design of Random Forest models, consisting of a thousand decision trees, one could expect it would be able to capture more information and tune its predictions more accurately. However, our results are in line with earlier findings as the Decision tree model uses a gradient booster which has been known to outperform a standard Random Forest model, as noted in Caruana & Niculescu-Mizil (2006).

Although the best average performance is achieved using the SARIMAX models, there is a trade-off to consider. The results of the Decision tree model are close to equal accuracy in all five zones and do inherent some desired qualities other than accuracy. The SARIMAX models require extensive preprocessing of the input data and analysis to fit the model. Meanwhile, the main advantage of a Decision tree model is that it handles data well as is, without differencing, normalizing, or standardizing. In addition, the Decision tree model proved to be the most computationally efficient of the complex models, requiring both the least computational power and having the lowest runtimes. The choice of model is thus heavily influenced by what constraints one operates under. With limited time or computing power, one might prefer the

Decision trees easy implementation and quicker runtimes. Without these constraints the SARIMAX models deliver the best average forecast and would be preferred.

Another consideration is between the average accuracy and the size of the extreme errors. While the average accuracy is the most telling metric for the overall quality of the forecasting model, some users of forecasting might be more sensitive to large outlier errors. In the extreme range of the errors the Decision tree performs best in two, the MLPRegressor in one, while the SARIMAX main has the lowest range in the remaining two zones. However, as the residual graphs in chapter 7.1 shows, the extreme errors are not necessarily indicative of a model with many large errors. While possible, the results do not show any model with large errors and near zero errors averaging to the best accuracy for the zone. The extreme range for the models is mostly at comparable sizes in each zone, and the errors of each model seems to have similar distributions within their respective ranges. With this in mind, we argue the results of the extreme errors metric do not show a model clearly outperforming the rest.

Some differences between the statistical models and the machine learning models can also be found in the input features. While the base input features are the same for both sets, some differences developed through testing and adapting each model. All models were built using the base variables capturing seasonal- and weather-effects. Where they differ is in the adaptation of these variables to capture the effects most efficiently for each model type to produce the most accurate forecast. This we would argue does not hurt the comparability of the models, apart from the differing preprocessing needs trade-off which was discussed earlier.

Finally, attempting to conclude on what forecasting model performs the best, can at best be applied to those who have the same experience with statistics and machine learning as ourselves. As noted by Hong & Fan (2016) empirical reviews on different STLF techniques can be misleading, depending on the researcher’s expertise and/or case study setup. In this thesis we argue that both techniques are treated equally fair in terms of case study setup, but our own expertise is in favor of the statistical approach. Although this is the first time we design a statistical forecast of this complexity, neither of us had prior knowledge or experience with machine learning techniques.

With this in mind, the machine learning techniques performed close to the SARIMAX model and is easier to implement and run. In addition, running the SARIMAX model in Eviews is demanding in terms of computational power and the model had a significantly longer runtime.

However, the computational speed of EViews might be outperformed in a prediction setting by other software such as R or Python.

8.2. Model performance compared to Entso-E

Comparing the forecast accuracy for NO1 the last half of 2019, Entso-E’s official forecast perform significantly better in terms of MAPE, but poorer measured by RMSE. When inspecting the forecast residuals, it shows that Entso-Es forecast has lower average errors, but a few very large errors. Low forecasting errors is very beneficial for market participants and TSO’s as it provides more reliable insights, which can improve market efficiency, production planning and network balancing. Although a lower average error is very beneficial, so is minimizing few but large errors. Very large forecasting errors can reduce efficiency of the day-ahead market and balancing of the transmission network. For the last half of 2021 Entso-E’s forecast performs best in NO1 measured in both MAPE and RMSE. Even with the few large errors in mind the forecast published on the Entso-E platform for NO1 is a very good approximation of the day-ahead load. The models developed in this thesis has not been able to match the average accuracy the forecast published on Entso-E achieves for this bidding zone.

The accuracy of the forecast published on the Entso-E platform for NO1 raises the question of why zones NO2 through NO5 is predicted with much lower quality. This thesis’ models are shown to be applicable to all five zones with comparably accurate results, all of which clears the bar of being better than the naïve forecast. The day-ahead forecast published on the Entso-E platform, however, consistently performs worse than the naïve forecast for all zones but NO1. Following from the fact that our models perform in all five zones, and the fact that the forecast published on the Entso-E platform for NO1 is accurate, one would expect it to be entirely possible to generalize the forecasts used to predict NO2-5 more accurately as well. Why this is not the case is unbeknown to us and our inquiries into the matter has not resulted in anything.

The inaccuracy of the published forecasts for all zones excluding NO1 makes the comparison between the models developed in this thesis and the official forecasts difficult to interpret. While it is clear if one were to take the published forecasts as best effort models for the four zones, we have in this thesis been able to develop better forecasts outperforming the official forecast in four out of five zones. However, as the results seem to indicate that something other than lack of ability plays a part, we would exercise caution in the interpretation of zones NO2 to 5.

Another consideration to make is that the forecasts developed are performed as rolling 24-hour predictions. This makes the forecasts not directly comparable with the forecasts gathered from the Entso-E Transparency Platform, as those forecasts are made for the complete following day at a set time the day ahead. The forecasts made in this thesis will thus have some informational advantage over the official forecasts for the last part of the day, while having some disadvantage for the first part. On balance we argue the results can be compared when keeping this in mind.

For the models designed in this thesis to compete with Entso-E in NO1, we believe some improvements can be made. Firstly, a more in-depth analysis of the model errors could be conducted, to identify under which conditions the models can be improved. This can be useful in search of further optimizing feature and variable selection. By looking at the errors split into their seasonal components as discussed throughout the thesis, annually, weekly, and daily sections, it would be possible to identify whether there are seasonally dependent effects on the load for which our models do not account. Building on the understanding of what seasonal conditions the models are less optimized for, it would be possible to search for variables which could be affecting load under these specific conditions.

Another improvement to be considered, of the weather variable quality, can be done by performing the weather station selection suggested by Hong et al. (2015). The suggestion made by Hong et al. is a framework design to determine how many and which weather stations to use for the load forecasting of a geographic area. As demonstrated by their research, this can increase the quality of the weather variables and lead to better forecast accuracy. No publications of such analysis pertaining to the forecasting of load in the Norwegian markets has been found. As such, it is possible improvements could be made by using a weather station selection more adapted to capture effects of for instance population density in combination with weather conditions.

8.3. The power price’s influence on load forecasting

The price increase seen in Norway the last year is unprecedented and might have resulted in a short-term demand effect. A short-term change in demand based on the power price can result in less accurate forecasts, because they are trained on historical data and are therefore unable to account for new demand patterns. In addition, the day-ahead price is traditionally not used in load forecasts for two reasons. Firstly, the predicted day-ahead load is used in the day-ahead market price formation, and secondly, the short-term price elasticity is found to be close to zero, (Hofmann & Lindberg, 2019). In other words, forecasts assume inelastic price sensitivity and are therefore unable to react to potential short-term price responses. This is the reason for the work in this thesis, to find whether the price changes of the last half of 2021 have a significant effect on the forecasting of load in the Norwegian market.

Thus, examining the relationship and causality between price and load is made difficult by the way price formation works in the electricity markets. The day-ahead prices are determined using amongst other factors the day-ahead load forecasts, a higher predicted load means the day-ahead price is increased. As such we observe a market where, counterintuitively, models including price as an exogenous variable will show a positive relation indicating higher prices leads to higher demand. This is where the age-old problem of regression analysis is apparent, correlation does not necessarily indicate causation.

In light of this, and the fact that a pure price elasticity analysis was outside the scope of the thesis, the solution is to be a mapping of whether adding the price-information would significantly change the prediction quality of the models. This approach has some obvious weaknesses. The most important of which is the interpretation of the results. Should the resulting accuracy of the models change significantly when adding the price information, the interpretation would be limited to the fact that there is some information about day-ahead load to be captured in the price variable. Whether we have found changed price elasticity compared to what previously has been assumed, or if the difference comes from other effects will have to be left to further work. The strength, however, is in its practical approach. We are able to shed light on whether the price increase has had any implication for load which can be captured by the forecasting models. This would lay the groundwork for further analysis of the exact market impacts of the price increase.

In our first analysis, adding price as a variable shows no improvement in forecasting performance for either 2019 or 2021. The accuracy, measured in MAPE rarely deviated by more than 0,03 percentage points when adding the price information compared to the predictions without price. Neither did the results change in one uniform direction. This indicates quite clearly only small random changes from adding a new variable. Nothing indicates the models was able to learn any significant new information from the price. This is in line with previously discussed assumptions for the electricity markets, in that the price elasticity is likely to still be close to zero.

The second analysis of the price effect on load forecasting, the changes in model performance between 2019 and 2021, gives no reason to believe the price changes significantly changed forecasting neither. Of the five zones, three shows slightly worse accuracy, while two show improvements from 2019 to 2021. Of the three southern zones with the largest price increase in 2021, one improved while two worsened. Once again, the testing has failed to prove a clear and uniform change, which leads us to argue the forecasting of load was not made more difficult by the large price increases.

9. Conclusion

We started the thesis attempting to answer how best to forecast load in periods of extreme price movements. To answer this, we divided the question into three sub-questions. The first was what is the best type of model for load forecasting in the period in question. The second was whether the best models can compete with the official forecasts published on the Entso-E transparency platform. The third and final question was whether the price movements changed the conditions for forecasting load in the Norwegian market. We offer the conclusions in the same order.

In the attempt to find the best model to accurately forecast the day-ahead load in the Norwegian markets during the last half of 2021's extreme price movements for southern Norway we have found two models to perform close to equal. While the SARIMAX model including interactions has the most accurate forecasts on average, the Decision tree model is not far behind. For zones NO1 through NO5 the SARIMAX with interactions showed accuracy measured by MAPE of 2,40%, 2,55%, 2,96%, 3,69% and 2,75% respectively. The Decision tree models accuracy was 2,67%, 2,54%, 3,10%, 3,74%, and 3,06% respectively. While the SARIMAX model is more accurate, the Decision tree model is both faster to compute and easier to set up. The best model is thus dependent on the need of the forecaster. For absolute performance, we have shown the SARIMAX with interactions to be the best model. When in need for speed and ease of implementation, the Decision tree model delivers adequate results.

When comparing the models developed in this thesis to the forecasts published on the Entso-E platform we have shown our models to be the most accurate for four out of five zones. This comes with the caveat that the predictions published for all zones except NO1 are far behind both our models and a naïve forecast, for which we can offer no explanation. For NO1, the forecast published on the Entso-E platform has a MAPE of 1,79%, clearly outperforming our best model, the SARIMAX interactions with its 2,40%. All models developed in this thesis, however, are able to remain comparably accurate for the remaining four zones unlike the forecast published on the Entso-E platform.

Lastly, our results show that the price changes seen during the autumn of 2021 appear to not have any influence on the accuracy in the forecasting of load. Forecasting with and without the price variable showed no significant difference in results, indicating that there was no information to be gained by the models from the price changes. We have also shown that the difficulty of load-forecasting in the three zones where we have seen the largest price increase in 2021, NO1, NO2 and NO5, did not see a clearly larger change than in the two zones with lower price increases. This leads us to conclude that the extreme price movements likely have not been an important factor in the determination of the load in the five Norwegian bidding zones.

Reference list

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika,* 52(3): 317-332.

Arora, S. & Taylor, J. W. (2013). Short-Term Forecasting of Anomalous Load Using Rule-Based Triple Seasonal Methods. *IEEE Transactions on Power Systems*, 28 (3): 3235-3242.

Bakirtzis G., Petridis V., Kiartzis S. J., Alexiadis M. C., and Maissis A. H. (1996). A neural network short term load forecasting model for the Greek power system. *IEEE Transactions on Power Systems*, 11(2): 858-863.

Box, G. E. P. & Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control*. 2nd ed. San Francisco: Holden-Day.

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

Caruana, R. & Niculescu-Mizil, A. (2006). *An Empirical Comparison of Supervised Learning Algorithms. Proceedings of the 23rd international conference on Machine learning,* Pittsburgh, USA.

Dickey, D. A. & Fuller, W. A. (1979). Distribution of Estimators for Time Series Regressions with a Unit Root. *Journal of the American Statistical Association*, 74 (366): 427-431.

Elamin, N., & Fukushige, M. (2018). Modeling and Forecasting Hourly Electricity Demand by SARIMAX with Interactions. *Energy*, 165 (PB), 257-268.

Entso-E (s.a.) *Grid map.* Available from: https://www.entsoe.eu/data/map/ (read. 13.05.22)

Entso-E. (s.a., b) *ENTSO-E Mission Statement.* Available from: https://www.entsoe.eu/about/inside-entsoe/objectives/ (read. 27.04.22)

Friedman, J.H. (2001) Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29 (5): 1189 – 1232

Hammad, M., Jereb, B., Rosi, B., & Dragan, D. (2020). Methods and Models for Electric Load Forecasting: A Comprehensive Review. *Logistics & Sustainable Transport,* 11(1): 51-76.

Ho, T.K. (1995) *Random Decision Forest. Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, 14-16 August 1995.*

Hofmann, M., & Lindberg, K. B. (2019). *Price elasticity of electricity demand in metropolitan areas - Case of Oslo, 16th International Conference on the European Energy Market, September 2019.*

Hong T. & Fan S. (2016). Probabilistic electric load forecasting: A tutorial review. *International Journal of Forecasting*, 32(3): 914-938.

Hong, T., Gui, M., Baran, M., & Willis, H. (2010). *Modeling and forecasting hourly electric load by multiple linear regression with interactions. IEEE PES General Meeting. Minneapolis, USA. July 25-29, 2010*. IEEE.

Hong, T., Wang, P., & White, L. (2015). Weather station selection for electric load forecasting. *International Journal of Forecasting*, 31 (2): 286-295.

Hyndman, R.J. & Khandakar, Y. (2008) Automatic Time Series Forecasting: The forecast Package for R. *Journal of Statistical Software*, 27(3): 1–22.

Hyndman, R. J. & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22 (4): 679-688.

IEA (2021), *Net Zero by 2050*, Report by IEA 05/21, available from: https://www.iea.org/reports/net-zero-by-2050

Khwaja A.S., Naeem M., Anpalagan A., Venetsanopoulos A., Venkatesh B. (2015) Improved short-term load forecasting using bagged neural networks. *Electric Power Systems Research*, 125: 109-115

Kwiatkowski, D., Phillips, P.C.B., Schmidt, P., & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root?. *Journal of Econometrics*, 54(1-3): 159-178.

Lee, C-M., & Ko, C-N. (2011) Short-term load forecasting using lifting scheme and ARIMA models. *Expert Systems with Applications*, 38 (5): 5902-5911.

McCulloch, W.S. & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics* 5(4): 115–133.

Myung, S. K. (2013). Modeling special-day effects for forecasting intraday electricity demand. *European Journal of Operational Research*, 230 (1): 170-180.

Nti, I. K., Teimeh, M., Nyarko-Boateng, O., & Adekoya, A. F. (2020). Electricity load forecasting: a systematic review. *Journal of Electrical Systems and Information Technology*, 7(1).

Papadopoulos, S. & Karakatsanis, I. (2015) *Short-term electricity load forecasting using time series and ensemble learning methods. IEEE Power and Energy Conference. Illinois, USA. February 20-21, 2015*. IEE

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12 (): 2825-2830.

Rumelhart, D., Hinton, G. & Williams, R. (1986) Learning representations by back-propagating errors. *Nature,* 323: 533–536.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2): 461-464.

Schwert, G. W. (1989). Tests for Unit Roots: A Monte Carlo Investigation. *Journal of Business & Economic Statistics,* 7(2): 147-159.

Soliman S., & Al-Kandari A. (2010). *Electrical Load Forecasting: Modeling and Model Construction,* 1st ed. Butterworth–Heineman.

Tarsitano, A. & Amerise, I. L. (2017). Short-term load forecasting using a two-stage SARIMAX model. *Energy,* 133(C): 108-114

Weron, R. (2006). *Modeling and Forecasting Electricity Loads and Prices: A Statistical Approach.* HSC Books, Hugo Steinhaus Center, Wroclaw University of Technology.

Weron, R. (2014). Electricity price forecasting: A review of the state-of-the-art with a look into the future. *International Journal of Forecasting,* 30(4): 1030-1081.

Weron, R., & Misiorek, A. (2005, May 10-12, 2005). *Forecasting spot electricity prices with time series models Proceedings of the European electricity market EEM-05 conference,* Lodz, Poland.

Zhao, Z &. Liu, H. (2007). Spectral feature selection for supervised and unsupervised learning. *Proceedings of the 24th international conference on Machine learning*, New York, USA.

Zor K., Timur O., & Teke A. (2017). A state-of-the-art review of artificial intelligence techniques for short-term electric load forecasting. *2017 6th International Youth Conference on Energy (IYCE)*, Budapest, Hungary.

Appendix 1: Estimation output

Table 21: Estimation output for SARIMAX(0,0,0)(7,0,7)$_{24}$ interaction for NO1 from 01.01.15 to 30.06.19

Variable	Coefficient	t-Statistic	Variable	Coefficient	t-Statistic	Variable	Coefficient	t-Statistic
C	4293,8 (266,8)	16,09	Wed* Hour1	70,8 (20,6)	3,43	Hour17* Holiday	-260,6 (20,5)	-12,70
SAR(24)	1,505 (0,056)	26,68	Wed* Hour6	1018,1 (337,3)	3,02	Hour18* Holiday	-231,0 (21,2)	-10,92
SAR(48)	-0,980 (0,071)	-13,73	Wed* Hour7	1361,7 (335,8)	4,06	Hour19* Holiday	-212,6 (21,5)	-9,90
SAR(72)	0,332 (0,080)	4,16	Wed* Hour8	1391,9 (341,4)	4,08	Hour20* Holiday	-177,3 (21,5)	-8,24
SAR(96)	0,292 (0,077)	3,79	Wed* Hour9	1351,2 (356,1)	3,79	Hour21* Holiday	-147,5 (22,2)	-6,65
SAR(120)	-0,780 (0,062)	-12,54	Wed* Hour10	1285,9 (362,7)	3,55	Hour22* Holiday	-107,5 (23,4)	-4,60
SAR(144)	1,144 (0,044)	26,01	Wed* Hour11	1305,1 (361,0)	3,62	Hour23* Holiday	-66,3 (24,2)	-2,73
SAR(168)	-0,520 (0,036)	-14,28	Wed* Hour12	1301,3 (359,9)	3,62	Temp$_t$* Hour4	-25,0 (8,0)	-3,12
SMA(24)	-0,849 (0,057)	-14,90	Wed* Hour13	1291,3 (358,9)	3,60	Temp$_t$* Hour5	-26,7 (6,7)	-4,00
SMA(48)	0,518 (0,043)	12,10	Wed* Hour14	1282,3 (357,7)	3,58	Temp$_t$* Hour6	-32,2 (6,4)	-5,04
SMA(96)	-0,311 (0,043)	-7,21	Wed* Hour15	1296,1 (352,1)	3,68	Temp$_t$* Hour7	-39,4 (6,1)	-6,42
SMA(120)	0,622 (0,032)	19,31	Wed* Hour16	1335,4 (353,6)	3,78	Temp$_t$* Hour8	-36,2 (6,4)	-5,65
SMA(144)	-0,669 (0,031)	-21,82	Wed* Hour17	1312,0 (359,8)	3,65	Temp$_t$* Hour9	-39,7 (6,7)	-5,94
SMA(168)	0,142 (0,02)	8,45	Wed* Hour18	1245,3 (365,0)	3,41	Temp$_t$* Hour10	-40,9 (6,9)	-5,89
Feb	-157,1 (19,0)	-8,27	Wed* Hour19	1162,5 (370,8)	3,13	Temp$_t$* Hour11	-44,5 (6,6)	-6,73
Mar	-341,9 (18,4)	-18,54	Wed* Hour20	1072,4 (372,9)	2,88	Temp$_t$* Hour12	-40,1 (6,7)	-5,97
Apr	-391,4 (20,1)	-19,44	Thu* Hour1	74,0 (21,9)	3,37	Temp$_t$* Hour13	-38,2 (6,9)	-5,57
May	-789,8 (25,6)	-30,90	Thu* Hour6	1022,6 (337,8)	3,03	Temp$_t$* Hour14	-35,2 (6,8)	-5,20
Jun	-1148,0 (34,3)	-33,47	Thu* Hour7	1368,7 (336,6)	4,07	Temp$_t$* Hour15	-21,5 (6,6)	-3,27
Jul	-1315,4 (55,3)	-23,80	Thu* Hour8	1405,8 (342,4)	4,11	Temp$_t$* Hour16	-30,2 (6,5)	-4,62
Aug	-1222,3 (55,5)	-22,02	Thu* Hour9	1367,2 (357,0)	3,83	Temp$_{t-1}$* Hour4	27,1 (8,3)	3,27
Sep	-935,4 (42,5)	-22,02	Thu* Hour10	1309,2 (363,5)	3,60	Temp$_{t-1}$* Hour5	27,1 (6,9)	3,94
Oct	-600,6 (28,5)	-21,10	Thu* Hour11	1323,6 (361,5)	3,66	Temp$_{t-1}$* Hour6	28,8 (6,6)	4,38
Nov	-221,5 (24,0)	-9,23	Thu* Hour12	1318,3 (360,4)	3,66	Temp$_{t-1}$* Hour7	32,1 (6,3)	5,07
Dec	-165,3 (18,8)	-8,78	Thu* Hour13	1312,1 (359,0)	3,65	Temp$_{t-1}$* Hour8	26,7 (6,6)	4,04
Temp$_{t-1}$	-54,5 (6,0)	-9,07	Thu* Hour14	1297,4 (357,9)	3,62	Temp$_{t-1}$* Hour9	28,4 (6,9)	4,13
Temp$_{t-24}$	-28,4	-109,01	Thu*	1308,0	3,72		28,8	4,08

	(0,26)		Hour15	(352,1)		Temp$_{t-1}$* Hour10	(7,1)	
Temp^{2}	0,26 (0,06)	4,55	Thu* Hour16	1348,7 (353,8)	3,81	Temp$_{t-1}$* Hour11	28,6 (6,8)	4,23
Temp$^{2}_{t-1}$	0,77 (0,05)	15,52	Thu* Hour17	1321,7 (360,0)	3,67	Temp$_{t-1}$* Hour12	22,7 (6,9)	3,29
Temp$^{2}_{t-24}$	0,59 (0,02)	38,31	Thu* Hour18	1253,6 (365,5)	3,43	Temp$_{t-1}$* Hour13	18,8 (7,0)	2,69
Hum	-2,99 (0,36)	-8,37	Thu* Hour19	1157,6 (371,4)	3,12	Hum$_{t}$* Hour9	0,93 (0,36)	2,60
Hum$_{t-1}$	-1,41 (0,17)	-8,12	Thu* Hour20	1065,1 (373,2)	2,85	Hum$_{t}$* Hour10	1,72 (0,36)	4,76
Mon* Hour6	995,8 (336,3)	2,96	Fri* Hour1	83,1 (21,4)	3,88	Hum$_{t}$* Hour11	1,99 (0,36)	5,56
Mon* Hour7	1342,3 (334,5)	4,01	Fri* Hour6	995,5 (338,7)	2,94	Hum$_{t}$* Hour12	2,37 (0,35)	6,76
Mon* Hour8	1386,8 (340,1)	4,08	Fri* Hour7	1344,1 (337,3)	3,98	Hum$_{t}$* Hour13	2,76 (0,35)	7,79
Mon* Hour9	1361,7 (355,2)	3,83	Fri* Hour8	1388,1 (342,8)	4,05	Hum$_{t}$* Hour14	3,08 (0,35)	8,74
Mon* Hour10	1314,6 (362,5)	3,63	Fri* Hour9	1359,1 (357,3)	3,80	Hum$_{t}$* Hour15	3,24 (0,35)	9,35
Mon* Hour11	1342,4 (360,8)	3,72	Fri* Hour10	1305,7 (363,7)	3,59	Hum$_{t}$* Hour16	3,05 (0,35)	8,80
Mon* Hour12	1334,2 (359,7)	3,71	Fri* Hour11	1314,0 (361,8)	3,63	Hum$_{t}$* Hour17	2,70 (0,35)	7,66
Mon* Hour13	1324,4 (358,7)	3,69	Fri* Hour12	1288,9 (360,5)	3,58	Hum$_{t}$* Hour18	2,25 (0,36)	6,25
Mon* Hour14	1312,9 (358,0)	3,67	Fri* Hour13	1261,1 (359,0)	3,51	Hum$_{t}$* Hour19	1,95 (0,36)	5,40
Mon* Hour15	1324,7 (352,4)	3,76	Fri* Hour14	1225,9 (357,9)	3,43	Hum$_{t}$* Hour20	1,34 (0,37)	3,62
Mon* Hour16	1359,6 (353,6)	3,85	Fri* Hour15	1215,0 (352,0)	3,45	Hum$_{t}$* Hour21	1,05 (0,37)	2,85
Mon* Hour17	1330,2 (359,8)	3,70	Fri* Hour16	1239,4 (354,0)	3,50	Temp$_{t}$* Apr	-7,37 (1,09)	-6,77
Mon* Hour18	1263,4 (364,8)	3,46	Fri* Hour17	1221,7 (360,6)	3,39	Temp$_{t}$* May	7,53 (1,19)	6,33
Mon* Hour19	1175,6 (370,7)	3,17	Fri* Hour18	1154,1 (365,8)	3,16	Temp$_{t}$* Jun	25,89 (1,49)	17,32
Mon* Hour20	1080,5 (372,7)	2,90	Fri* Hour19	1020,4 (372,0)	2,74	Temp$_{t}$* Jul	32,43 (2,13)	15,25
Tue* Hour1	65,6 (15,3)	4,30	Sat* Hour11	933,5 (362,4)	2,58	Temp$_{t}$* Aug	29,49 (2,31)	12,75
Tue* Hour6	1044,9 (336,8)	3,10	Sat* Hour16	917,2 (354,3)	2,59	Temp$_{t}$* Sep	16,72 (1,96)	8,55
Tue* Hour7	1384,9 (335,0)	4,13	Sat* Hour17	957,1 (360,8)	2,65	Temp$_{t}$* Oct	-3,55 (1,19)	-2,99
Tue* Hour8	1418,7 (340,8)	4,16	Hour4* Holiday	-103,0 (22,1)	-4,66	Temp$_{t}$* Nov	-10,90 (0,94)	-11,55
Tue* Hour9	1377,9 (356,0)	3,87	Hour5* Holiday	-267,6 (19,5)	-13,72	Temp$_{t}$* Dec	-12,47 (0,79)	-15,79
Tue* Hour10	1317,4 (362,9)	3,63	Hour6* Holiday	-463,6 (18,4)	-25,25	Hum$_{t}$* Feb	2,55 (0,20)	12,49
Tue* Hour11	1339,6 (361,0)	3,71	Hour7* Holiday	-492,1 (18,2)	-27,00	Hum$_{t}$* Mar	4,92 (0,18)	27,88
Tue* Hour12	1331,1 (360,1)	3,70	Hour8* Holiday	-431,7 (18,6)	-23,20	Hum$_{t}$* Apr	4,59 (0,18)	25,36
Tue* Hour13	1325,7 (359,0)	3,69	Hour9* Holiday	-381,4 (19,1)	-19,96	Hum$_{t}$* May	7,28 (0,21)	34,10

Tue* Hour14	1311,3 (358,1)	3,66	Hour10* Holiday	-338,1 (19,3)	-17,51	Hum_t* Jun	7,68 (0,27)	28,51
Tue* Hour15	1325,6 (352,5)	3,76	Hour11* Holiday	-324,3 (19,4)	-16,74	Hum_t* Jul	7,50 (0,38)	19,76
Tue* Hour16	1364,8 (353,9)	3,86	Hour12* Holiday	-313,4 (19,8)	-15,85	Hum_t* Aug	7,42 (0,38)	19,45
Tue* Hour17	1341,6 (360,4)	3,72	Hour13* Holiday	-314,8 (19,9)	-15,82	Hum_t* Sep	6,19 (0,34)	17,96
Tue* Hour18	1277,2 (365,4)	3,49	Hour14* Holiday	-313,8 (20,0)	-15,68	Hum_t* Oct	5,18 (0,23)	22,19
Tue* Hour19	1186,6 (371,3)	3,20	Hour15* Holiday	-306,8 (19,9)	-15,38	Hum_t* Nov	1,29 (0,24)	5,32
Tue* Hour20	1098,3 (373,1)	2,94	Hour16* Holiday	-293,6 (20,0)	-14,65	Hum_t* Dec	1,07 (0,20)	5,28
						$Temp_t$* Hum_t	-0,23 (0,01)	-19,10
R-squared	**0,9905**		**SIC**	**12,68**				
AIC	**12,61**		**HQIC**	**12,64**				

Appendix 2: Residual graphs

Residual graphs NO1

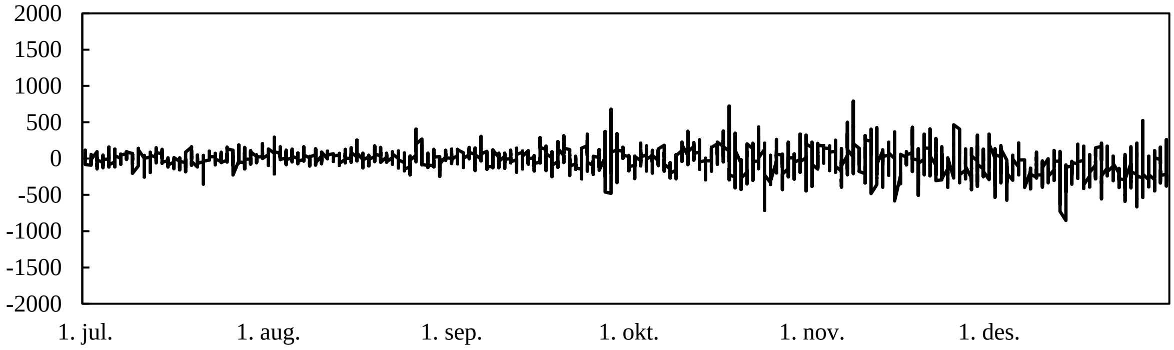

Figure 48: Residuals from RandomForestRegressor forecast of load in NO1 01.07.21 - 31.12.21

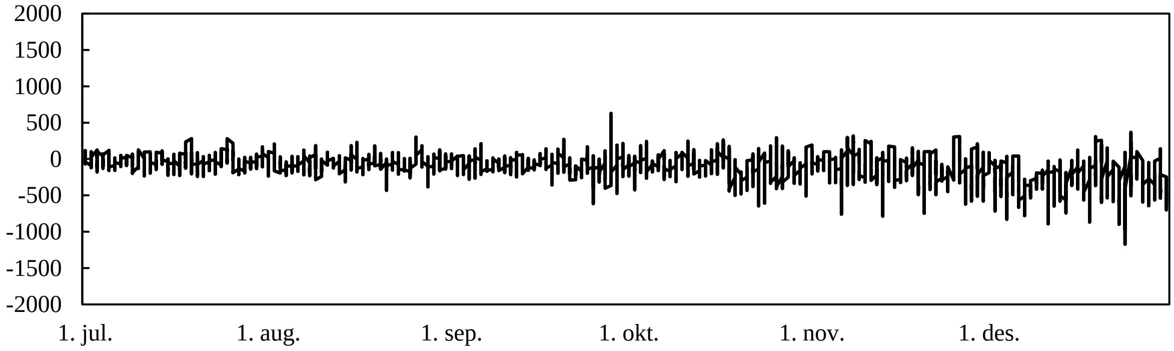

Figure 49: Residuals from MLPRegressor forecast of load in NO1 01.07.21 - 31.12.21

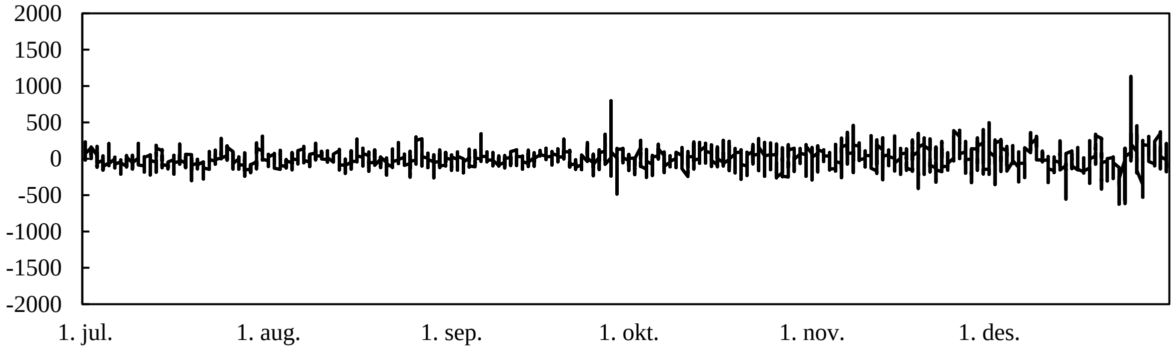

Figure 50: Residuals from SARIMAX main forecast of load in NO1 01.07.21 - 31.12.21

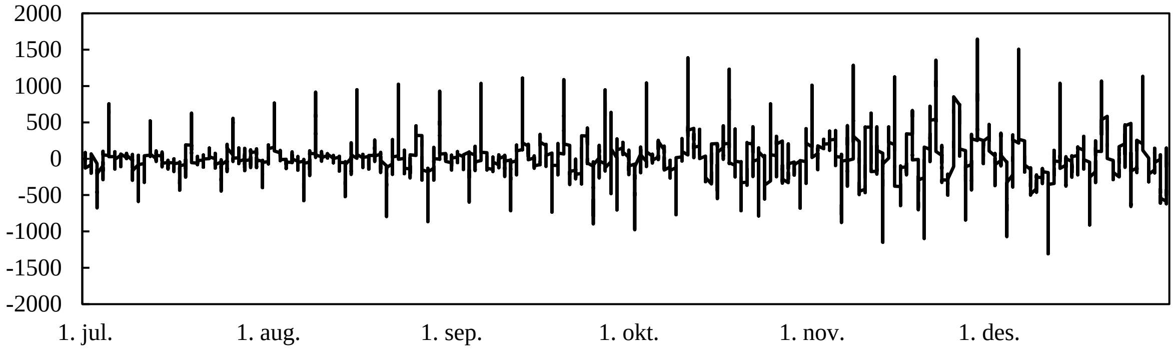

Figure 51: Residuals from Naïve forecast of load in NO1 01.07.21 - 31.12.21

Residual graphs NO2

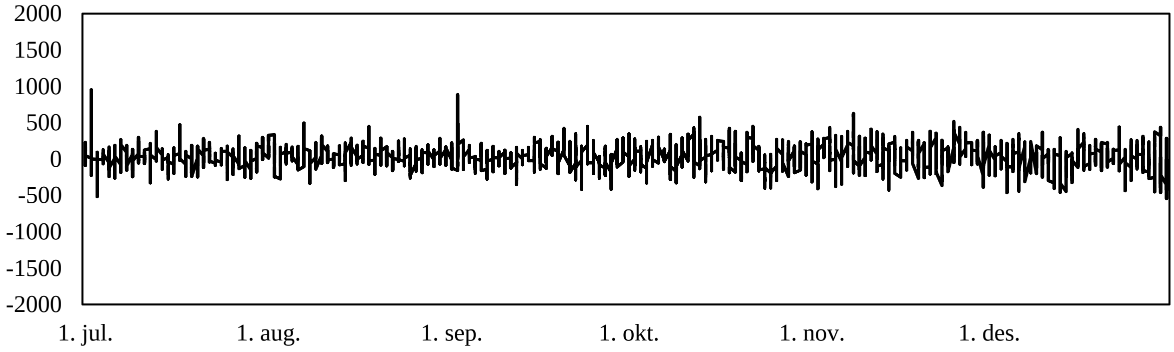

Figure 52: Residuals from RandomForestRegressor forecast of load in NO2 01.07.21 - 31.12.21

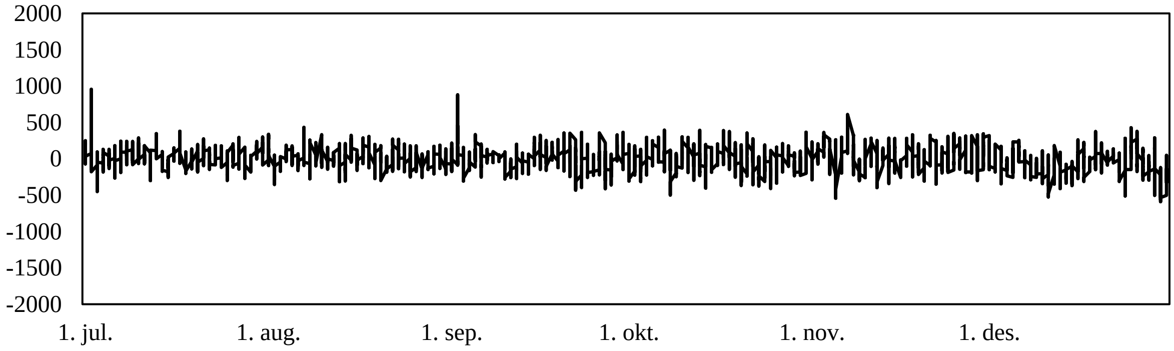

Figure 53: Residuals from MLPRegressor forecast of load in NO2 01.07.21 - 31.12.21

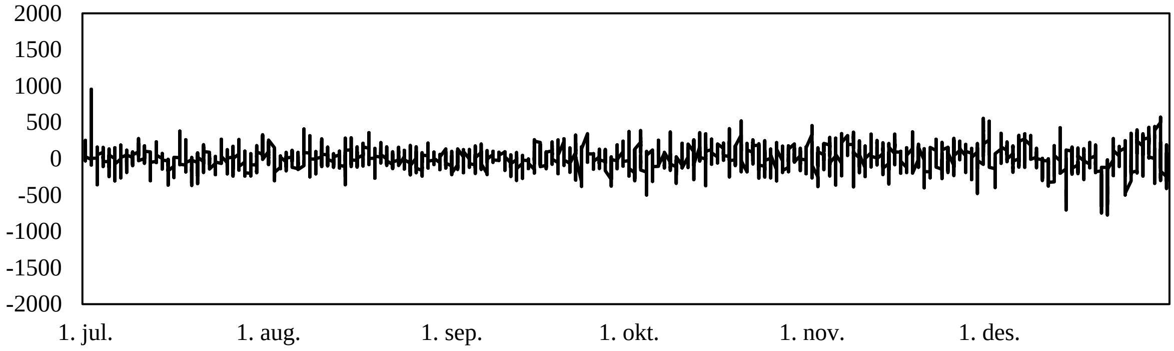

Figure 54: Residuals from SARIMAX interactions forecast of load in NO2 01.07.21 - 31.12.21

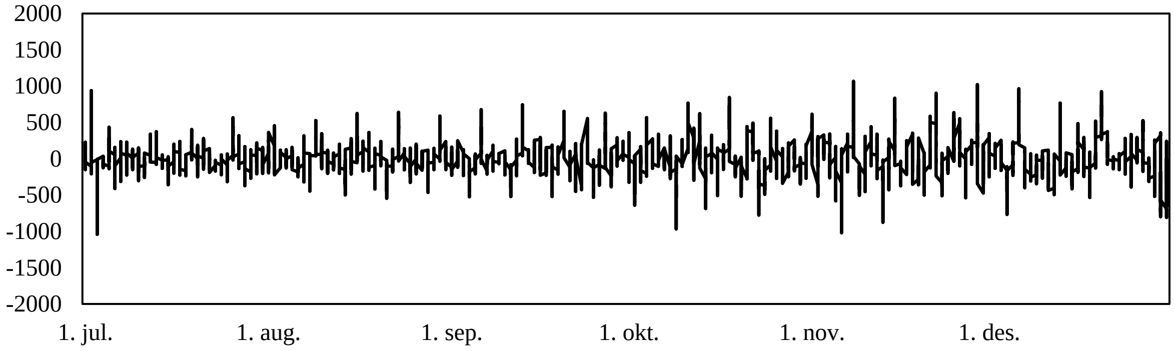

Figure 55: Residuals from Naïve forecast of load in NO2 01.07.21 - 31.12.21

Residual graphs NO3

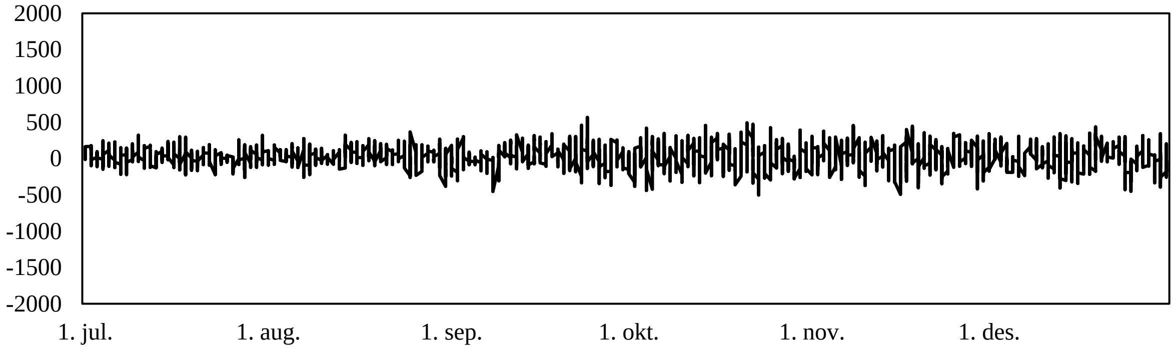

Figure 56: Residuals from Random Forest forecast of load in NO3 01.07.21 - 31.12.21

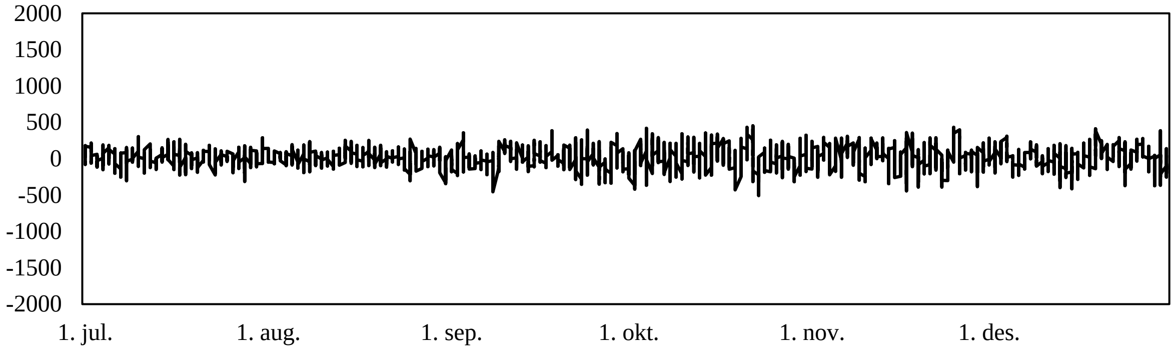

Figure 57: Residuals from MLPRegressor forecast of load in NO3 01.07.21 - 31.12.21

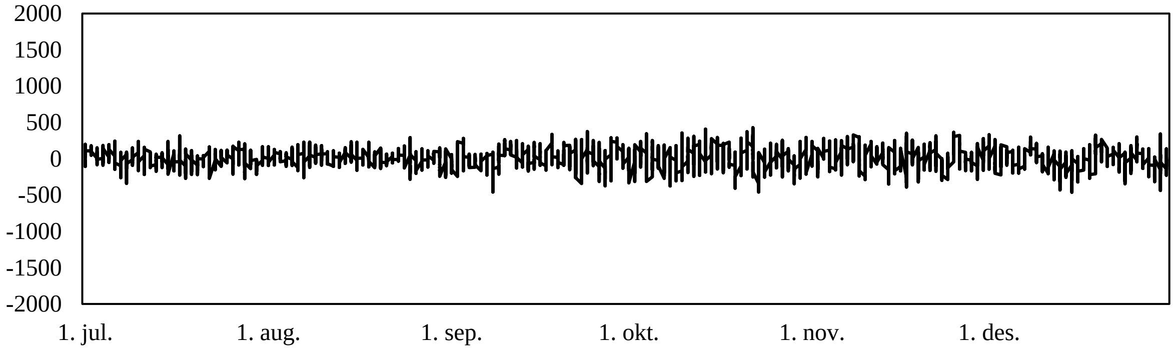

Figure 58: Residuals from SARIMAX main forecast of load in NO3 01.07.21 - 31.12.21

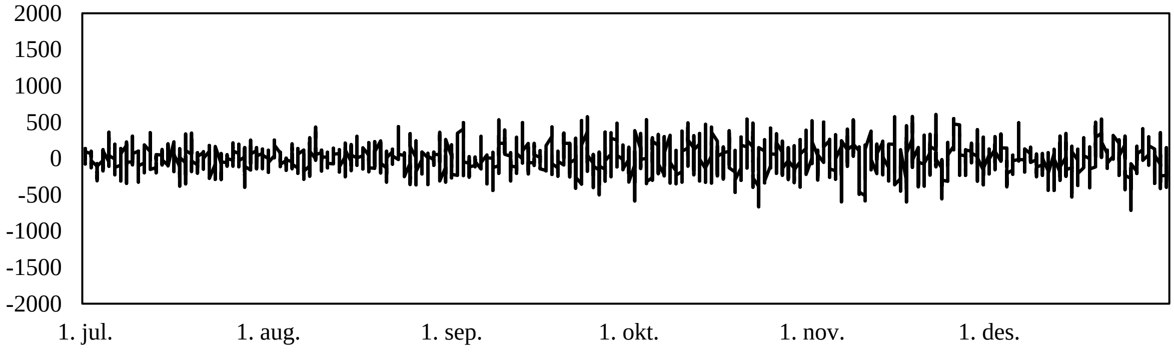

Figure 59: Residuals from Naïve forecast of load in NO3 01.07.21 - 31.12.21

Residual graphs NO4

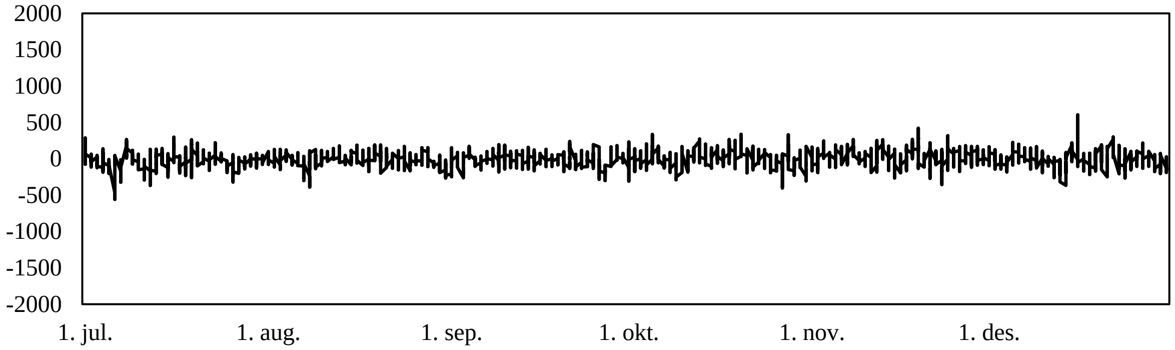

Figure 60: Residuals from Random Forest forecast of load in NO4 01.07.21 - 31.12.21

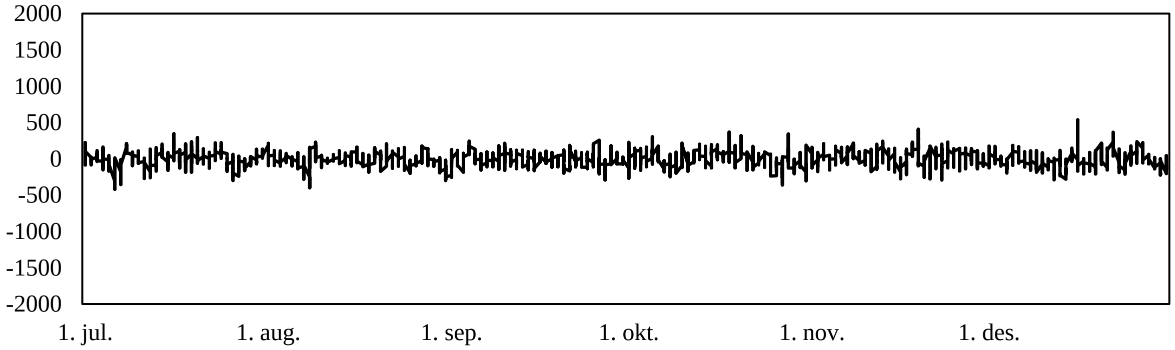

Figure 61: Residuals from MLPRegressor forecast of load in NO4 01.07.21 - 31.12.21

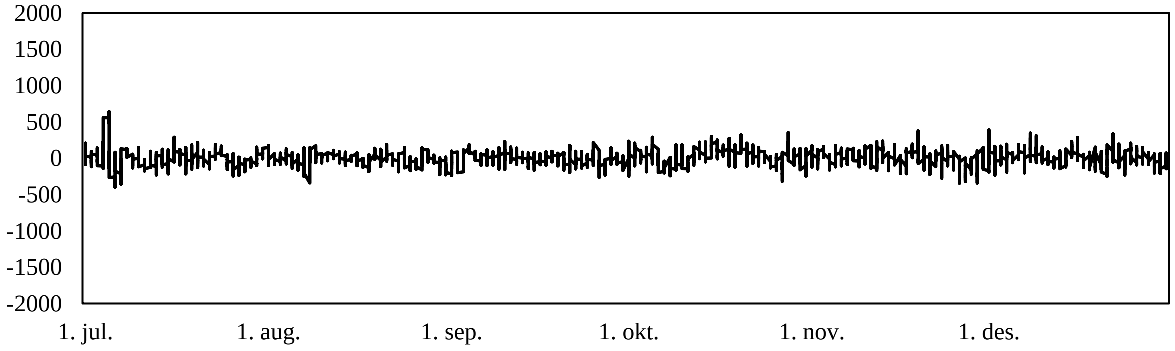

Figure 62: Residuals from SARIMAX interaction forecast of load in NO4 01.07.21 - 31.12.21

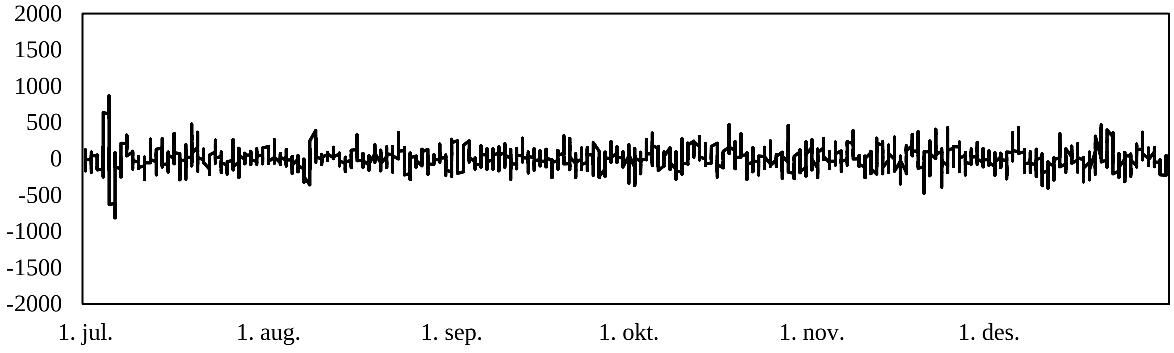

Figure 63: Residuals from Naïve forecast of load in NO4 01.07.21 - 31.12.21

Residual graphs NO5

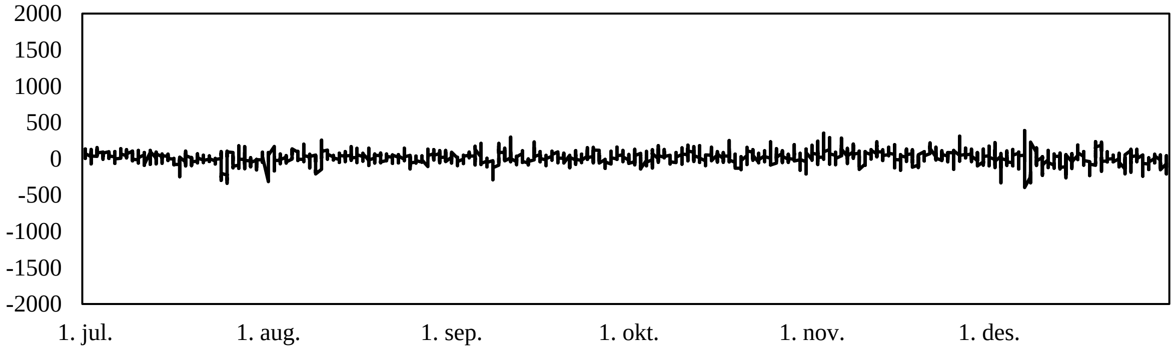

Figure 64: Residuals from Decision tree forecast of load in NO5 01.07.21 - 31.12.21

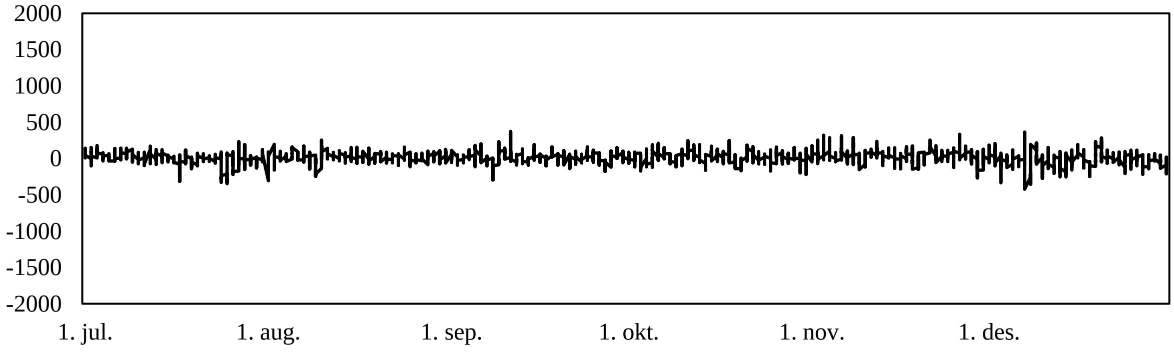

Figure 65: Residuals from Random Forest forecast of load in NO5 01.07.21 - 31.12.21

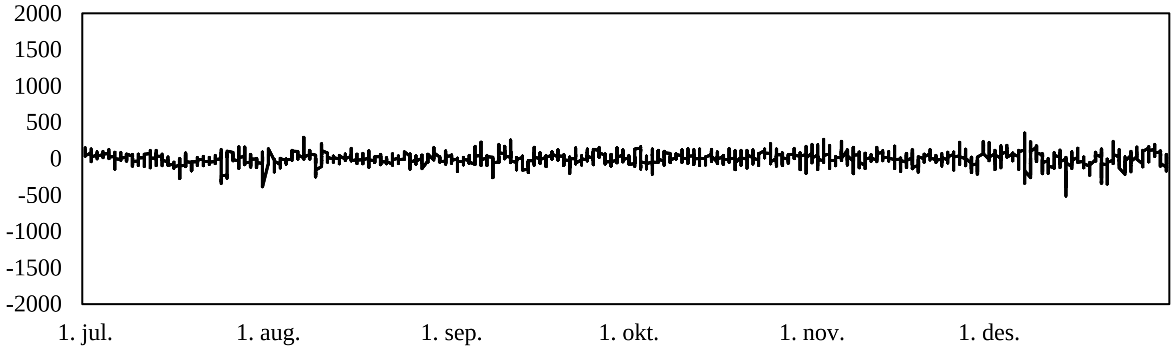

Figure 66: Residuals from SARIMAX main forecast of load in NO5 01.07.21 - 31.12.21

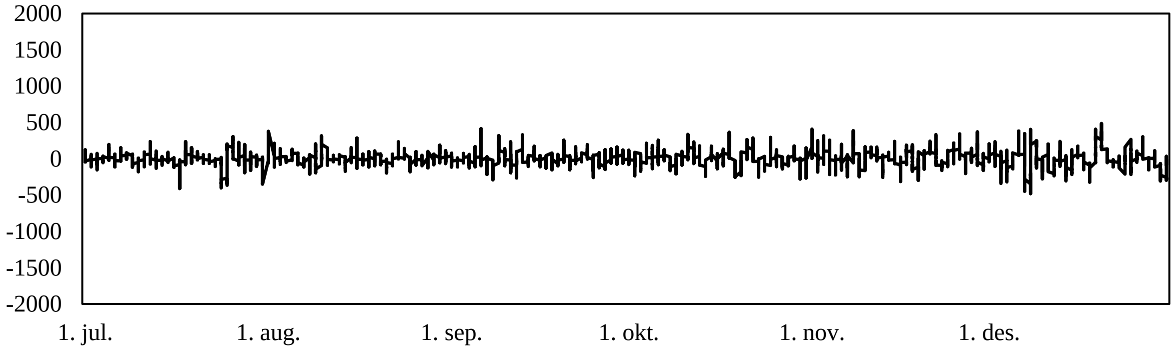

Figure 67: Residuals from Naïve forecast of load in NO5 01.07.21 - 31.12.21

Appendix 3: List of abbreviations

ACF – Autocorrelation Function

ADF – Augmented Dickey Fuller

AI - Artificial intelligence

AIC – Akaike’s Information Criteria

ANN – Artificial Neural Network

AR - Autoregressive

ARMA – Autoregressive Moving Average

ARIMA – Autoregressive Integrated Moving Average

BPN – Back-Propagation Network

ELF – Electricity Load Forecasting

ENTSO-E – the European Network of Transmission System Operators

GBRT – Gradient Boosting Regression Tree

HQIC – Hannan-Quinn Information Criteria

IC – Information Criteria

LTLF – Long-Term Load Forecast

KPSS – Kwiatkowski, Phillips, Schmidt, Shin

MA – Moving Average

MAE – Mean Absolute Error

MAPE – Mean Absolute Percentage Error

ML – Machine Learning

MLP – Muti-Layer Perceptron

MSE – Mean Squared Error

MTLF – Medium-Term Load Forecast

MW - Megawatt

PACF – Partial Autocorrelation Function

pp – Percentage points

RF – Random Forest

RMSE – Root Mean Squared Error

SAR – Seasonal Autoregressive

SARMA – Seasonal Autoregressive Moving Average

SARIMA – Seasonal Autoregressive Integrated Moving Average

SARIMAX – Seasonal Autoregressive Integrated Moving Average with Exogenous variables

SIC – Schwarz’s Information Criteria

SMA – Seasonal Moving Average

STLF – Short-Term Load Forecast

SVM – Support Vector Machine

TSO – Transmission System Operator

VRE – Variable Renewable Energy

VSTELF – Very Short-Term Load Forecasting

Appendix 4: List of Tables

Appendix 5: List of figures

Norges miljø- og biovitenskapelige universitet
Noregs miljø- og biovitskapelege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway