EVALUATING FORECASTS OF VOLATILE ELECTRICITY PRICES MADE BY COMPETING MODELS

Thesis submitted for the degree of
Cand.Oecon.

by

Michael Bülow Pedersen

Thesis Supervisor: Timo Teräsvirta

Department of Economics
Aarhus University, Denmark

12 August 2015
(Thesis defence on 17 September 2015)

Abstract– Making accurate predictions of short-term electricity prices is of interest to many stakeholders in the electricity sector. In this thesis we investigate the forecast performance of a range of forecasting methods that include both statistical methods, machine-learning methods and a market benchmark which has previously seen little use. To emulate the situation faced by a professional forecaster, we use a range of fundamentals forecasts as explanatory variables. Empirical results are presented that analyze out-of-sample day-ahead forecast performance in the largest German/Austrian electricity market for the 12 months of 2014. These results show that an adaptive statistical model has the best individual performance, but that forecast aggregation can improve the performance significantly beyond that of a single model.

Keywords– Electricity price forecasting, electricity prices, non-linear modelling, adaptive estimation, machine learning, neural network, support vector machine, ARMAX, forecast performance evaluation, forecast aggregation, EPEX, EXAA, market benchmark, german electricity market.

Version– 1.01 (19/09/2015): Fixed typos and reworded some sentences. Added another graph of bidding curves

Contents

List of Figures

List of Tables

1 Introduction

The ability to predict short-term electricity prices accurately is of significant value to market participants engaged in bilateral trading of electricity or trading in auctions, as it enables them to make optimal bids. There has therefore been considerable focus on developing methods for predicting prices optimally in both the academic world and within the industry itself since deregulation and the introduction of electricity auctions during the 1990s and 2000s. Many papers published since deregulation have been of varying quality from a statistical method point of view, and this thesis provides an application where the proper use of statistical validation is emphasized.

In this thesis, we investigate the forecasting performance of a range of statistical and machine learning methods for estimating the hourly day-ahead electricity prices in the German/Austrian market. We test the out-of-sample forecasting performance across almost a year, using consistent statistical methods when evaluating the forecast performance. We also introduce a market benchmark that allows us to relate model performance to market performance, a coupling that has not been described in the literature until recently.

The thesis is structured as follows. Section 2 provides a brief introduction to the electricity sector and some fundamental factors that affect prices. Section 3 describes the structure of the German/Austrian electricity market. Section 4 reviews previous literature related to electricity price forecasting. Section 5 introduces the forecasting methods, model selection criteria, and how to evaluate forecast performance consistently. Section 6 explains how data were obtained and comment on features observed in the data. Section 7 deals with selecting the specific forecasting models used to forecast in the out-of-sample period. Section 8 presents the out-of-sample results. Section 9 concludes and Section 10 points to possible improvements and future research.

2 Background

In this section we provide a short introduction to the properties of electricity, the regulatory structures that have been created to address the issues of security of supply, and natural monopolies. We also see how different technologies are used in the generation of electricity and how it can be traded on exchanges.

2.1 Providing electricity

Electricity has a special importance in modern economies, because it is an essential requirement in many things that we do. At the same time it is currently non-storable in larger quantities, and supply must equal demand at all times (Aggarwal *et al.*, 2009).

2.1.1 Keeping the system balanced

The all-important goal for the electricity sector is that electricity should always be available. To achieve this there needs to be a constant balance between the production and consumption at every instant.

This requires a highly structured market, where market participants submit a schedule for all their planned electricity flows to a Transmission System Operator (TSO) before delivery and the TSO then checks if the combined schedules from all the market participants are feasible, given constraints in transmission grid

(Consentec, 2014). If this is not the case, the TSO can intervene and counter trade with generators and consumers until the schedules are feasible.

The TSO is also responsible for operating several smaller markets for reserve power - for additional inflow to/outflow of electricity from the system - that are called upon during the delivery period to completely balance the power system. When metered actual production and consumption becomes available, the market participants that caused imbalances are billed according to the TSO's costs in contracting the reserve power (Consentec, 2014).

2.1.2 A volatile product

The essential nature of electricity means that consumers have an almost unlimited willingness to pay for a minimum amount of electricity. This willingness can lead to large positive price spikes in times of high demand and/or tight supply conditions, when the price of electricity is determined in deregulated electricity markets (Amjady and Hemmati, 2006).

However, in many markets negative price spikes are also a frequent occurrence. As electricity producers have significant costs associated with starting production, there can be situations where it is optimal to produce below marginal costs for a period of time, to avoid a costly shutdown/start up cycle (Nicolosi, 2010).

Because the system always needs to balance, the demand elasticity of electricity is low, and up/downward production flexibility is limited, it is common to see both relatively frequent and large positive and negative price spikes in the hourly electricity price series.

2.1.3 The merit order curve

Electricity is produced using many different technologies that are characterized by a mix of upfront investment cost and a marginal cost of producing electricity. Some technologies such as solar, wind, or nuclear power have most of their associated costs as an up-front investment, while their marginal price is almost zero or even negative (for example, due to start-up costs or poorly designed renewable energy support schemes) (Würzburg *et al.*, 2013). These sources of electricity will almost always produce when available and crowd out sources with higher marginal costs.

At the other end of the technology mix are peak-load plants such as gas turbines and diesel generators. These generators typically have low investment costs, but high marginal costs. They are activated less frequently when demand cannot be covered by cheaper sources, and activation comes with a much higher price tag.

In the middle field are generators with more moderate fuel costs such as coal and lignite plants. These units typically run most of the time to cover their substantial investment costs, but can be throttled when prices fall below their marginal costs. Therefore, these generators frequently becomes price setting.

Together the production costs and capacities of all generators form a merit order curve as seen in Figure 1. When a certain level of consumption has to be satisfied, the low marginal cost plants are activated first and then the successively more expensive plants are activated until the electricity consumption is covered. See Nicolosi (2010) for a more detailed discussion of the merit order curve for Germany/Austria.

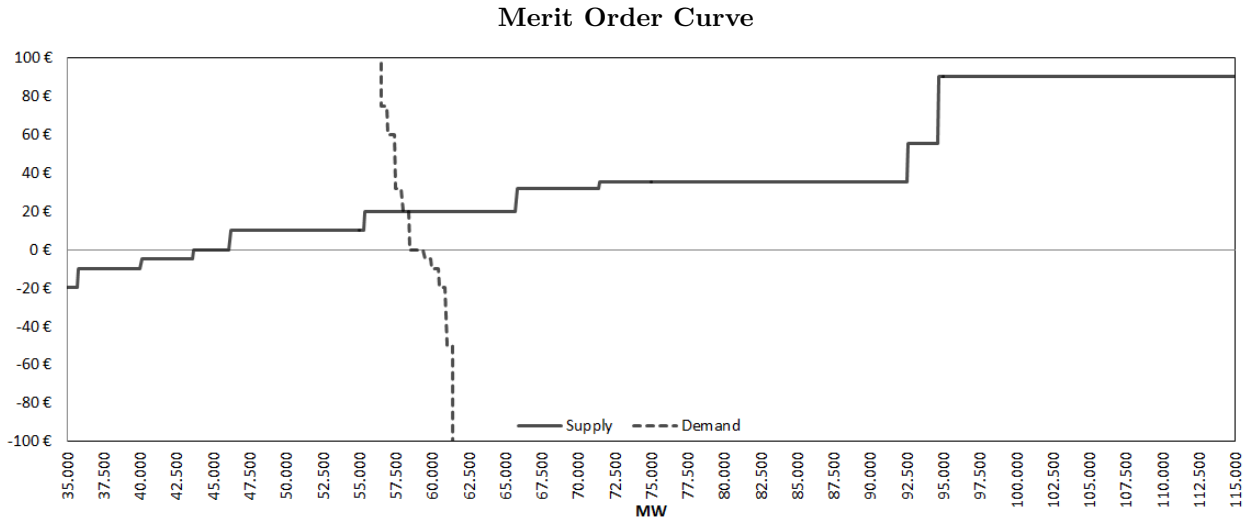

Figure 1 – An illustration of a hypothetical merit order curve. Plots the marginal production cost against the sum of production capacity of all units at or below this number. The curve shows clear jumps at the marginal costs of using a new generation type to cover demand. In real power systems these thresholds are not fixed, but vary to a certain degree between generators (depending, for example, on the efficiency of production units), leading to a smoother supply curve. Demand has also been depicted as an almost price-independent line. It is not exactly vertical because some consumers can react to prices, most notably hydro reservoirs with pumping units that can offtake significant amounts when prices are low. Source: Authors own illustration.

The merit order curve is a tool to understand price formation in electricity markets. In the perfect market, (where the maximization of total welfare is the criterion), every generator would bid their marginal cost, and the last plant to be contracted to produce at any given time, would be the plant with the highest marginal price necessary to fully cover demand.

The merit order curve also explains why price spikes are more probable when production is high (low) and consumption is low (high). During these periods the supply and demand curves risk crossing at either end of the graph, where both curves are very steep and small changes in cleared quantity lead to large changes in price.

2.1.4 Deregulation, concentration, and competition

During the last 25 years many developed countries have gone through a process of deregulating the electricity sector. Deregulation was initiated to give market incentives that improve efficiency, increase competition, shift the risk of investment from consumers to producers and in general allow market prices to better reflect the fundamental costs of providing electricity (Joskow, 2008).

Before electricity markets were deregulated, electricity was typically provided by a single vertically integrated electric utility that owned both generation units, transmission networks, distribution networks and had the monopoly on selling to electricity consumers in a certain geographical area. To contain the market power of these utilities they were mostly publicly owned or privately owned and then subject to price regulation as natural monopolies (Joskow, 2008, p. 10).

Deregulation has typically required the incumbent vertically integrated utilities to be split up, so the trans-

mission and distribution networks remained a regulated monopoly, while the other parts were exposed to competition. Deregulation has also introduced more organized power markets to facilitate economical trading opportunities between market participants (Joskow, 2008).

For these historical reasons, the electricity sector is still highly concentrated. This makes it possible, and in many situations very profitable, for larger firms to influence prices, by withholding capacity from the market, in order to receive a higher price for the remaining generation capacity.

This incentive can lead to strategic bidding that interferes with the ideal merit order curve from Figure 1, hurting consumers and reducing overall utility. To counter this market imperfection most countries have regulatory oversight in the form of a national regulatory agency (NRA) that limits strategic bidding behaviour and other market distorting practices (Joskow, 2008). Nevertheless strategic biddingremains a factor with the potential to complicate forecasting of prices in electricity markets.

2.2 Trading electricity

2.2.1 Trading schedule

The way electricity is traded between market participants differs slightly between countries. Here we focus on the setup in the market area that consists of Germany and Austria (Germany/Austria)— that is also common to many other continental European markets.

In Germany/Austria the trading of electricity is organized around both bilateral trading, trading in a public day-ahead auction, and trading in an intraday trading session where only adjustments to positions can be made. An overview of the daily trading schedule can be seen in Figure 2.

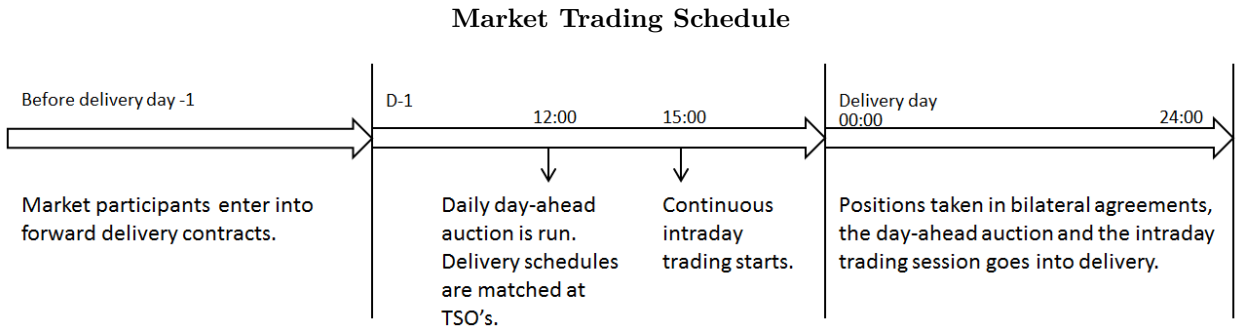

Figure 2 – The electricity trading schedule in Germany/Austria. Source: Authors own illustration.

The day-ahead auction is very important because a significant share of the Germany/Austria consumption is traded in the day-ahead auction (see Section 3 for details). Furthermore, it serves as the underlying product for many financial power derivatives traded on EEX[1]. In this thesis, we therefore focus exclusively on forecasting the day-ahead prices.

[1]see https://www.eex.com/en/products/energy/power/power-derivatives-market

2.2.2 Auction

There are two day-ahead marginal price auctions in Germany/Austria. In both, sellers and buyers can enter limit orders for delivery of power during certain contract periods until gate closure. After gate closure, the orders are combined to an aggregate buy and a sell curve, and the intersection between these curves determines the auction clearing price and volume.
All buyers with a limit price higher than the auction price pay the market clearing price, and are contracted to withdraw the contracted volume from the grid during the contract period. All sellers with a limit price lower than the auction price receive the market clearing price, and are contracted to deliver the sold volume to the grid during the contract period.

In many European countries, the day-ahead auctions are run at the same time. This allows for any connection capacity between countries to be allocated efficiently by price coupling the auctions. Under price coupling, the auction prices are determined simultaneously on all coupled exchanges, and the capacity between exchanges is used to deliver electricity from low price areas to high price areas.

2.2.3 Bidding curves

Having introduced both some fundamental concepts and how trading occurs, we take a closer look at an example of buy and sell curves from the largest day-ahead auction in Figure 3. Note that these curves represent actual auction bids from market participants that defined the final prices for this hour.

At first there seems to be an incompatibility between the flexible market buy curves in Figure 3, and our earlier graph of the merit order curve in Figure 1 that shows demand as rather inflexible and supply as more adaptive. However, this is not necessarily the case. Because market participants use bilateral forward contracts, a generator which has sold a forward supply contract can realize an economical trading opportunity by buying in the market instead of producing electricity, whenever the market clearing price falls below the units marginal cost. A generator can therefore become part of the buy side in the auction.

It is also clear from Figure 3 that there is a large amount of demand bid at the current market cap of 3000 €/MWh. This reflects price insensitivity on the part of many consumers and any export to other countries which is always bid at the market cap. Similarly, there is a significant amount of production bid at the current price floor of -500 €/MWh. This represents both must-run generation units and any import which is always bid at the price floor.

The figure also illustrates why the market sometimes shows very volatile price behaviour in response to small changes in supply/demand. If the electricity system is in a state where the bid curves cross on the steep parts at the end of either curve, small changes in volume have large price effects.

Note that these curves can shift rather dramatically between hours and days as illustrated in the changes between Hour 4 (Figure 4) and Hour 12 (Figure 3).

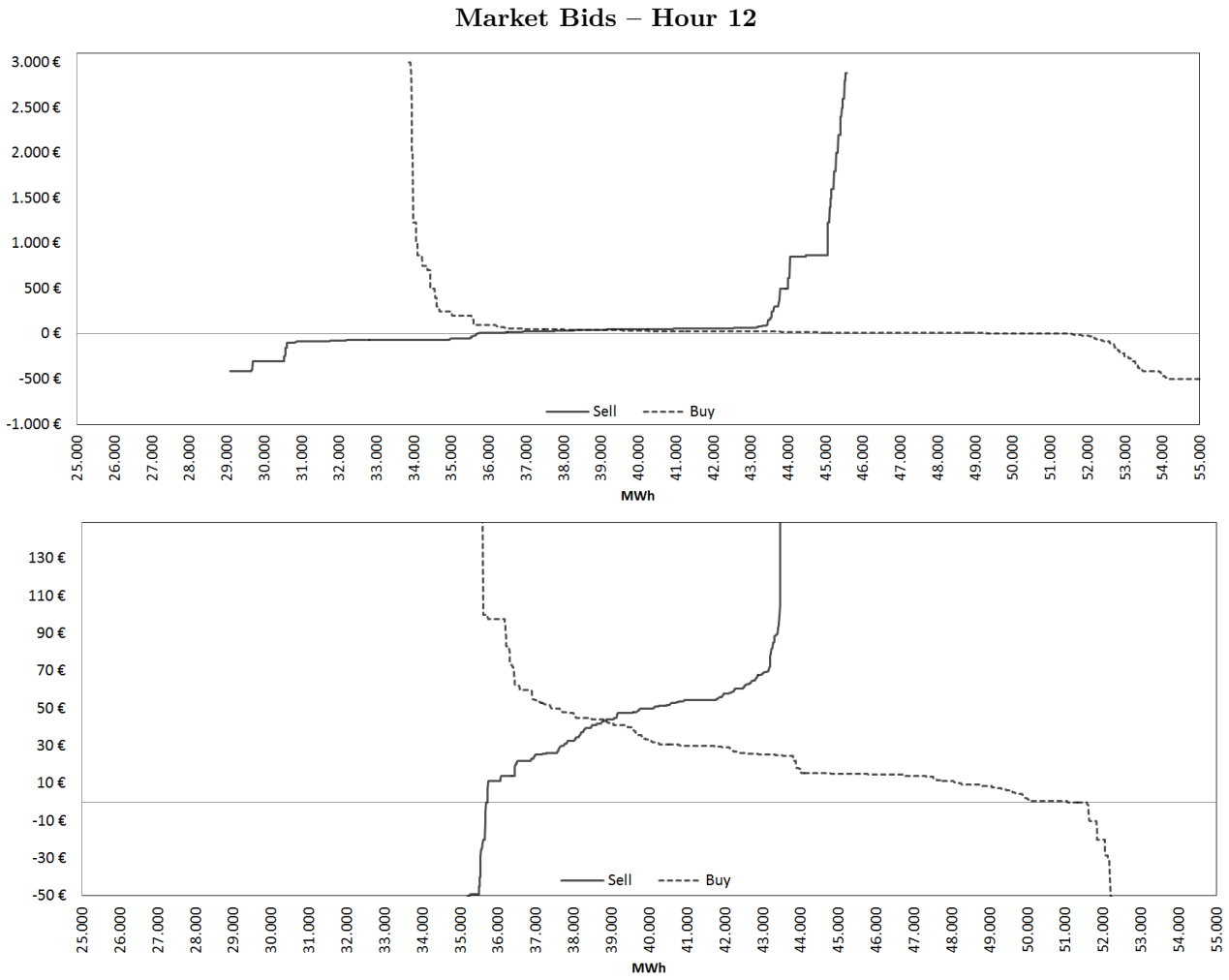

Figure 3 – Top: EPEX market bids for hour 12 on 15.12.2014. Bottom: Same as top graph but focused around more commonly occuring clearing prices. Source: EPEX Spot

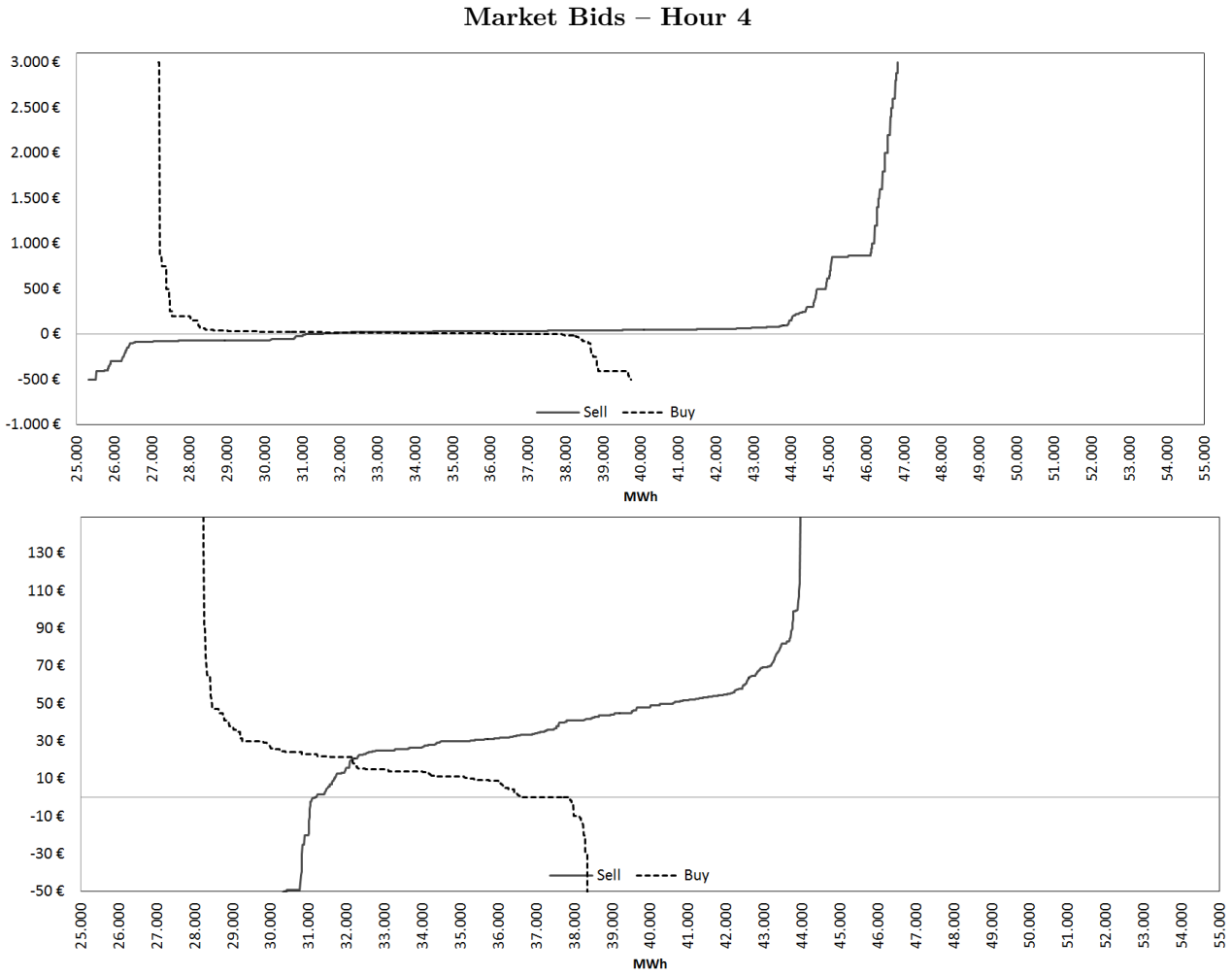

Figure 4 – Top: EPEX market bids for hour 4 on 15.12.2014. Bottom: Same as top graph but focused around more commonly occuring clearing prices. Source: EPEX Spot

3 The German Electricity Market

In this section, we take a closer look at the structure of the German/Austrian electricity sector.
First, we look at which generators produce electricity to get a better idea of the variables that are relevant for predicting prices. Then, we describe the two auctions where electricity is traded. Finally, we introduce the Market Consensus Benchmark, an easy to obtain market forecast that previously has seen little use in the literature.

3.1 Structure

The German/Austrian market area is the largest electricity market in Europe and it has five Transmission System Operators that operate the system.

Installed Production Capacity

Category\TSO	Type	Transnet	Tennet	Amprion	50Hz	APG	**Sum**
Wind	PI	682	16.469	6.669	14.347	2.140	**40.307**
Solar	PI	5.213	14.922	8.986	8.141	532	**37.794**
Gas	PD	1.041	9.303	15.600	4.293	4.901	**35.138**
Fossil Hard Coal	PD	5.127	6.938	14.166	1.479	1.173	**28.883**
Lignite / Brown Coal	PD	0	391	11.517	10.050	0	**21.958**
Hydro Storage[2]	PD	3.328	1.329	1.396	2.875	6.852	**15.780**
Nuclear	PI	2.712	5.455	4.060	0	0	**12.227**
Hydro Run-of-river	PI	808	1.719	2.047	0	5.601	**10.175**
Biomass	PI	789	2.761	1.306	1.717	254	**6.827**
Fossil Oil	PD	700	1.378	200	947	385	**3.610**
Rest[3]	-	132	580	4.620	651	1.071	**7.054**
Total		20.532	61.245	70.567	44.500	22.909	**219.753**

Table 1 – Installed production capacity in the five German/Austrian control areas at the beginning of 2015. Type is either Price Dependent (PD) or Price Independent (PI) as defined in Section 12.1.1 in the appendix. All values in installed MW. Source: transparency.entsoe.eu.

Electricity produced in Germany/Austria is generated using a range of technologies, each with different generation abilities and cost structures.
Table 1 shows the installed capacity of all generation types. At the top, we see that Renewable Energy Sources (RES) constitute the largest amount of installed capacity. Being dependent on the weather, RES are highly variable and on average they only produce a fraction of the total capacity (we will get back to this in Section 6). Due to the large capacity and volatility of renewables, the expected RES generation should be of high importance in a price forecast.

The more traditional generation types follow in Table 1 and can be characterized either as price insensitive or price sensitive (the utilization rates in Figure 33 & 34 in the appendix illustrate this).
For all generation types, any effect on auction prices arise as a result of 1) the capacity made available to the market and 2) the bidding price.

[2]Includes types: "Hydro Pumped Storage" and "Hydro Water Reservoir"
[3]Includes types: "Fossil Coal-derived gas", "Geothermal", "Other", "Other renewable" and "Waste"

For price-insensitive (PI) generators the price effect is only determined by 1), while for price-dependent (PD) generators the effect on market prices is determined by both 1) and 2).

For both PI and PD generators, it may well be relevant to include some measure of day-ahead available generation capacity in a price forecast, as this would have relevance for marginal prices due to 1). In this thesis, we have not included these kind of variables, but they could serve as an improvement.

To capture the price effect through 2), we could investigate the cost structure of PD generators, to capture any structural changes that affect their bidding prices (e.g. fuel costs, cost of CO2 permits). As changes in cost structure affect whole generation categories, they can shift large parts of the merit order curve leading to new market equilibrium prices. We have not included any cost structure variables in this thesis, but they do represent a sensible extension to the dataset.

If we look at the utilization rates in Figure 33 in the appendix, it is evident that a considerable part of the generation capacity of many generation types is not made available to the market. Possible reasons for this are discussed in Section 12.1.1 in the appendix. Therefore, we cannot directly translate Table 1 to a merit order curve like that in Figure 1 and use it for forecasting.

Exchanges with neighboring countries represent another major effect on the prices. The gross and net exchanges are shown in Figures 35 & 36 in the appendix. Here we see that Germany/Austria has almost always played a dual role as both an importer and exporter, but it was predominantly a net exporter of electricity in 2014. The external demand is bid directly at the auction as a buy or sell order, and this therefore affects German/Austrian price formation. Unfortunately, we do not have a day-ahead forecast for the net exchanges, so we will include the forecast French consumption as an imperfect proxy for external demand.

3.2 Exchanges

3.2.1 EPEX

EPEX Spot is the main power exchange for trading standardized electricity products in Germany, Austria, Switzerland and France[4].

EPEX market share in 2014

Country	Consumption	EPEX DA Volume	Share
Germany/Austria	574.156 GWh	262.920 GWh	45.8%
France	465.666 GWh	67.820 GWh	14.6%
switzerland	69.280 GWh	20.466 GWh	29.5%

Table 2 – Consumption, EPEX day-ahead volume and EPEX day-ahead share of total consumption in 2014. Sources: entsoe.eu and epexspot.com.

In 2014, EPEX Spot had a solid market share of 45.8% of the total German/Austrian consumption (Table 2). A further 1.36% of the consumption was traded on the EXAA exchange described in the next section (Table

[4]EPEX SPOT and APX Group are currently in the process of integrating their business, so in the future EPEX will also cover Belgium, Netherlands and UK.

3), while the rest was traded either bilaterally, intraday or was consumption of own production. In France and Switzerland, the market share of EPEX Spot is smaller. Here bilateral trading and own-production account for a much larger volume than the public auction.

Every day at 12:00, a number of standardized products are traded in the EPEX day-ahead auction for delivery next day. These include the hourly product (currently the smallest interval traded in the coupled EPEX DA auction), base load (delivery in each hour of the whole day), peak load (delivery between 8:00 and 20:00), off-peak (delivery between 20:00 and 8:00) and any other block interval across the day which the trade participants enter. Additionally, there are more complex order types such as linked blocks, where the execution of one block is conditional on the execution of another block, and mutually exclusive blocks where only one of several blocks can be executed[5].

Since February 4th, 2014, the EPEX auction has been a part of the North Western Europe (NWE) price coupling mechanism[6]. This mechanism ensures that the Nordic countries, the Baltics, Benelux, Germany/Austria, France and United Kingdom are price coupled. Before joining the NWE, Germany/Austria, Benelux and France formed the Central Western Europe (CWE) price coupling region, and coupling with the other parts of NWE was either with an explicit auction (UK) or volume coupling (Nordic)[7][8]. In May 2014, Portugal and Spain also joined the NWE region and, in 2015, the region has been further expanded.

From the above, it is evident that even if the generation structure in Germany/Austria had remained unchanged over the last few years (which has not been the case), there would have been structural breaks due to changes in the way the market has been operated.

As of December 2014, EPEX also runs a separate local (non-coupled) German/Austrian auction for trading quarter hours (delivery in a 15-minute period) at 15:00, one hour before intraday trading in quarters begin. It would also be interesting to forecast the prices at this auction, but this we will leave for future researchers.

3.2.2 EXAA

While the EPEX auction is linked to other auctions across the continent, the EXAA call auction is local to Germany/Austria.

It is run on weekdays at 10:12-10:15 and the results are published shortly afterwards. The Auction allows for trading in the same products as the EPEX day-ahead auction and additionally has trading in the quarter hour product. In 2014, the market share of EXAA was 1.36% of the total yearly German/Austrian consumption (Table 3). This amounts to 11.29% of the yearly Austrian consumption.

As there is declared no congestion on the border between Austria and Germany in the day-ahead trading window[9], traders on the EXAA auction can always demand delivery of the traded power directly in Germany (with any potential transmission capacity problem being solved by TSO countertrading). Participants

[5]The point of mentioning these special types of products is to show what diverse and advanced needs market participants have - needs we have to take into account when forecasting.

[6]http://static.epexspot.com/document/26088/DataNewZ_20140123_2014n15_NWE_Update.pdf

[7]Both are less efficent methods of market coupling

[8]https://www.epexspot.com/en/market-coupling

[9]http://www.apg.at/en/market/cross-border-exchange/auctions

EXAA market share in 2014

Country	Consumption	EXAA DA Turnover	Share
AT	69.294 GWh	7.825 GWh	11.29%
AT/DE	574.156 GWh		1.36%

Table 3 – Consumption, EXAA DA turnover and EXAA DA share of total consumption in 2014. Sources: entsoe.eu and exaa.at.

trading in both the EPEX and EXAA auctions can therefore speculate between the two auctions.

Note that EXAA changed the price floor from 0 € in October 2013, so the currently allowed price range is −150 € to 3000 € (Energy Exchange Austria (EXAA), 2014, p. 24) which is slightly different from EPEX.

3.3 The market consensus benchmark

Viehmann (2011) looks at the relationship between the bilateral OTC market, the EXAA auction, and the EPEX auction. He argues that the market clearing prices from the EXAA auction can be used as a snapshot of the less transparent German OTC market.

One argument for this is that the EXAA base price should coincide with the OTC base price at the time of the auction, as otherwise there will be a speculative opportunity for traders active in both marketplaces. Figure 5 confirms this view, as we see that the EXAA Base price and the Platts OTC-Index price indeed appear to be closely correlated.

EXAA Base and Platts OTC Index

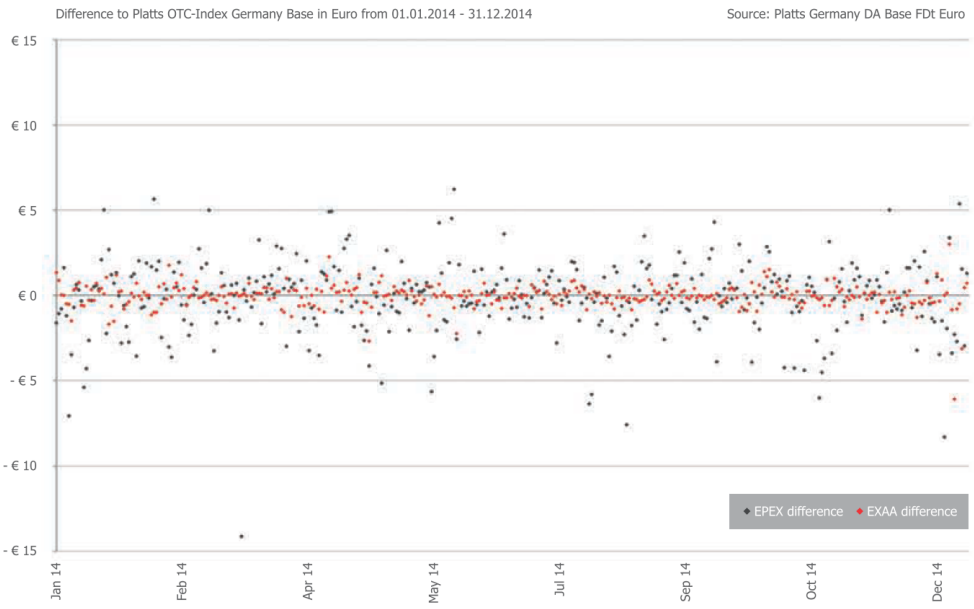

Figure 5 – Comparison of the differences of the EXAA Base price and the EPEX Base price to the Platts OTC-Index Base price. Source: Energy Exchange Austria (EXAA) (2015) (used with permission).

Having established a connection between the OTC market and the EXAA auction, Viehmann (2011) then compares the EXAA auction prices with the EPEX auction prices. He finds that EXAA "energy traders behave rationally as risk-adverse agents", that there are "negative as well as positive risk premiums that are significantly different from zero for hourly delivery contracts" and that "risk premiums are directly related to the tightness of the system" (Viehmann, 2011).

These observations all indicate that the EXAA auction prices can be seen as the OTC/EXAA market participants' 'consensus' estimate of the risk-adjusted day-ahead EPEX price for all the hourly products that also trade on EPEX. The risk adjustment appears to be related to the tightness of the electricity system. The existence of a risk adjustment seems sensible, as tight market situations are known to potentially create very large spikes in the electricity price, something a risk-averse agent would pay a premium to avoid.

We will use the EXAA auction price as a benchmark forecast in the rest of this thesis and refer to it as the **Market Consensus Benchmark (MCB)**. The only other paper found that uses EXAA in price modelling is Ziel *et al.* (2015), who uses EXAA as an explanatory variable in forecasting models of the day-ahead prices in Germany and the surrounding countries.
For reasons of brevity, we will not try to improve the MCB by accounting for the market risk adjustment, though the observations made by Viehmann (2011) indicate that this might be possible.

The existence of a market consensus benchmark, which is publicly available before the closing of the EPEX order book, is one of the great advantages of testing electricity price forecasting models on the German EPEX price.
It can be argued that the MCB puts a theoretical lower bound on the performance of any forecast model. If a model on average can forecast the EPEX price more precisely than the MCB after accounting for the expense of a round-trip trade[10], the model can be used to arbitrage between the exchanges. If this deviation becomes public knowledge traders should integrate the information and the deviation will disappear.

One caveat about using the EXAA prices as a forecast of the EPEX prices, is that EXAA is only run on weekdays at 10:12-10:15, but not on weekends and public holidays (Energy Exchange Austria (EXAA), 2014, p. 4). So, on a standard week, the Friday auction will be for delivery days Saturday, Sunday, and Monday. The information set used in the MCB for Sunday (Monday) prices will therefore be 24 (48) hours stale and we predict that this will show up as larger forecast errors on these days.

[10] A trade where a speculative position is first opened and later closed. The cost between EXAA and EPEX is, at the time of writing, between 0.095€ and 0.145€ (Details can be found in the Table 27 in the appendix).

4 Review of the Literature

In the following, we first present previous literature dealing with electricity price forecasting. After that, we briefly review literature on forecast aggregation.

4.1 Forecasting methods

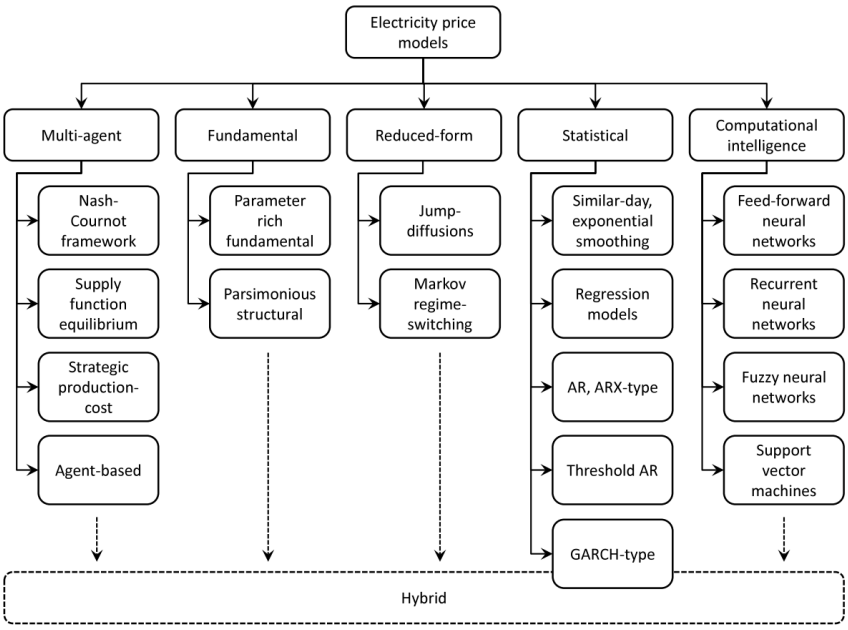

Figure 6 – A taxonomy of electricity price modelling approaches. Source: (Weron, 2014, p. 1041).

Electricitry Price Forecasting (EPF) is relatively young academic field with few publications before 2000 (Weron, 2014). Recent interest in forecasting of electricity prices has largely been fueled by the advent of electricity markets that has become central as a result of deregulation.

The main references within the electricity price forecasting field is the thorough survey article of Weron (2014), which reviews both previous literature, and the most commonly applied statistical and machine learning methods.
In Weron (2014), the author presents a taxonomy of electricity price model approaches that defines five main categories: "Multi-agent", "Fundamental", "Reduced-Form", "Statistical", and "Computational intelligence" (Figure 6).

As evident from Figure 6, there are many disparate approaches to the modelling of electricity prices. However, a lot of the methods in the figure are most frequently used to make mid and long-term forecasts (weeks–years ahead) as seen in the following quote, "Reduced-form models are generally not expected to forecast hourly prices accurately, but are expected to recover the main characteristics of electricity spot prices, typically at the daily time scale" – (Weron, 2014, p. 1049).
As our focus is on prediciting day-ahead prices, we restrict our choice of forecasting methods to the last two categories, "Statistical" and "Computational intelligence", and therefore focus the following literature

review on articles applying these methods.
Even then, the task of scrutinizing every method that has been applied in EPF articles would be a demanding exercise indeed—during the period 1989–2013 Scopus indexed 206 journal articles (Weron, 2014, p. 1034)—so we have also restricted the exposition to a smaller subset of articles that are more directly relevant to the methods we apply.

An early paper using statistical methods, is the survey of Nogales *et al.* (2002) that applies several types of ARMAX models[11] to the Spanish and Californian market in 2000[12] and evaluate model performances by comparing a special version of the mean absolute error in two single, separate weeks.
In Conejo *et al.* (2005) both an ARMAX, ARIMA, wavelet transform, and neural network model, with the forecast demand as an explanatory variable, are compared. Yet, surprisingly, no explanatory variable is used in the neural network. The authors find superior performance of the ARMAX specification with forecasts evaluated by comparing the daily percentage error—mean absolute error (MAE) divided by the average price of the day—and RMSE across four separate weeks (one in each season) on the American PJM market.
Weron and Misiorek (2008) test several types of AR models with spike preprocessing, a regime-switching model, and a semi-parametric extension and find that a spike-preprocessed ARX model, with the demand forecast as an explanatory variable, outperforms the other models. However, they compare the weekly-weighted mean absolute error—MAE divided by the average price of the week—in 10 weeks in California during 2000[13], 20 weeks in Nordic Nord Pool market during 1999, and 20 more weeks during 2004.
In Aggarwal *et al.* (2009), the authors review 47 papers on EPF that were published between 1997 and November 2006. The two main classes of models surveyed are statistical and neural network models. It is hard to use the comparison of model performances from the paper, as the reviewed papers cover price estimation on 11 different power exchanges during different time periods. However, the paper gives a very nice overview of factors influencing electricity prices, explanatory variables used in the reviewed papers, and the model design choices made in the reviewed papers.
Cruz *et al.* (2011) test a dynamic regression model, a periodic dynamic regression model (where a different dynamic regression model is used for Monday, Tuesday-Friday, Saturday, and Sunday), and a neural network on the Spanish market. The Transmission System Operator (TSO) day-ahead forecasts of wind and load are used as an explanatory variables. Out-of-sample results for 8½ months during 2007–2008 show that the linear models have a statistically significantly superior performance (using mean absolute percentage error (MAPE) and Relative MAE of (Hyndman and Koehler, 2006), and establishing that forecasts are statistically significant trough the Diebold-Mariano (DM) test).
Jónsson *et al.* (2013) fits an adaptive local approximation model with load and wind power as explanatory variables and forecasts to the Nordic DK1 market. The adaptive model is shown to outperform an ARIMAX benchmark model during a 2 year test period 2008–2009 using RMSE, the relative MAE, and relative RMSE, both from (Hyndman and Koehler, 2006).

Interesting models not used in this thesis includes both threshold AR models, mean-reverting jump-diffusion models, and semiparametric kernel models (see e.g. Weron and Misiorek (2008)). There are also numerous

[11] Here, we will refer to all ARMA models that incorporate explanatory variables as ARMAX – they also go by the names transfer function, dynamic regression, and regression with arma residuals, depending on the way the explanatory variables enter the model.
[12] A rather unfortunate choice as California suffered an energy crisis in this period due to, among other things, market manipulation by market participants (i.e. the now bankrupt Enron) (Weare, 2003)
[13] Again this period coincides with the aforementioned troublesome period

other papers, where different variations on the abovementioned statistical methods are developed, and we will refer the reader to either (Weron, 2006) or (Weron, 2014) for a more expansive review.

The other type of paper frequent in the electricity price forecasting literature, applies machine learning methods to forecast electricity prices. However, the distinction between machine learning papers and statistical papers is not always clearcut, and some of the aforementioned papers, for example, include neural networks. In Yamin *et al.* (2004), the authors use neural networks to forecast on the Californian market in 1999. Performance is evaluated in a 1 week out-of-sample period, using a special scoring function, the "MAPE using the median".
Mandal *et al.* (2006) apply a neural network to forecast half-hourly prices 1–6 hours ahead in the Australian Victoria market (this market has a somewhat different structure than that common to continental Europe) and compare performance in a single month using MAPE.
Amjady *et al.* (2010) use a neural network, which is estimated using a genetic search algorithm, to forecast Spanish day-ahead prices. The forecast period is 12 months in 1999–2000 and 4 separate weeks, coinciding with those in Conejo *et al.* (2005), and evaluation is done by comparing weekly mean errors.
Chen *et al.* (2012) test an extreme learning machine (ELM) – a type of neural network with input weights and biases randomly generated and its output weights analytically calculated – against a neural network, and a radial basis function neural network on the Australian national electricity market (NEM) market, and finds that the ELM is slightly superior, using both MAE, MAPE, and RMSE. They use a demand forecast as an explanatory variable and have separate models to each forecast period.
In Zhao *et al.* (2008) a support vector machine is used to forecast the Australian NEM regional reference price, as the first step in a procedure that generates prediction intervals, instead of point forecasts.
Chaâbane (2014a) apply a two-step procedure, where a deseasonalized price series is first forecast by an ARFIMA model, and then the residual series is forecast by a least squares support vector machine. The two-step model and a few reference models are tested on the Nordic Nordpool market on 100 hours in November 2012. Here 1-step ahead errors are evaluated using MAE and RMSE and the two-step approach is superiour. The choice of evaluating performance by comparing 1-hour ahead errors is strange, as when forecasting the day-ahead prices we would require different h-step ahead forecasts to be generated.

A few comments are in order

- The papers reviewed employ a diverse selection of scoring functions when evaluating the out-of-sample performance of different methods. MAPE, and related versions of normalized absolute errors, seem to be a popular choice. Both the mean absolute errors and root mean squared errors are popular.
- Few papers test if differences in forecast performance between models are statistically significant.
- Many papers employ a short out-of-sample validation period, leaving us uncertain about the stated results.

4.2 Forecast aggregation

Forecast aggregation has long been suggested as a technique to improve upon the forecast of a single best model (see e.g. Bates and Granger (1969); Newbold and Granger (1974)).
Within load forecasting Bunn (2000) reports that forecast aggregation can produce more accurate forecasts than a single model. In the context of electricity price forecasting Bordignon *et al.* (2013) test different

forecast aggregation schemes. Using the forecasts from four different time-series based models, they find that combining models significantly outperforms the best single model and that "the use of a simple average forecast combining method is competitive with respect to other, more sophisticated kinds of combinations" (Bordignon *et al.*, 2013, p. 102).

5 Theory

In this section we first discuss why we forecast. Then, we introduce the forecasting methods used in this thesis and the model selection criterias employed. After that, we present the theoretical foundations of how to evaluate model performance consistently. Lastly, we present the Diebold-Mariano test that allows us to provide statements on which models statistically outperform each other.

In this thesis, we will refer to a class of models as a "method", and any specific parameterization of a method is referred to as a "model".

5.1 Forecasting philosophy

As stated previously, the main goal in this thesis is to evaluate the performances of different models when predicting day-ahead prices.

We do not propose that we can find the correct model which accurately explains how and why fundamentals determine the electricity prices. This 'prediction goal' is important to have in mind, as it affects both the choice of methods (for example, choosing methods that make good predictions but does not necessarily represent the 'true' underlying data generation process) and how we select specific models (focus on measures that maximize predictive performance). See Shmueli (2010) for a detailed discussion on the philosophical and methodical differences between predicting and explaining.

Because we have the goal of prediction, we can choose freely from a wide variety of methods. At our disposal, we have both structurally simple methods, where the input variables affect prices in a relatively straightforward and easy to understand way, and we have very complex methods where an easy interpretation of the model structure is difficult to convey.

This distinction also features in Breiman (2001) who talks about two cultures in statistical modelling. The two cultures have a distinctly different view on the modelling task: The classic data modelling culture assumes a model, usually parametric, as the underlying data generation process and proceeds to estimate the parameters of this model. The machine learning modelling culture on the other hand considers the underlying data generation process as unknown and only tries to find a function that can predict the output of the underlying data generation process when given an input— usually using complex machine learning algorithms. Both these cultures are active in the field of forecasting electricity prices as stated in Weron (2014) and seen in Section 4.

The machine learning (or computational intelligence) approach has mainly been furthered by electrical engineers and the classic modelling approach has mostly been spearheaded by econometricians. As the following quote from Weron (2014) illustrates, this division is not entirely unproblematic:

> "Typically 'electric engineering' papers consider sophisticated CI [Computational Intelligence] tools and relatively simple (or not properly applied) statistical models, and when the two are compared, the former tend to perform better. On the other hand 'econometrics' or 'statistical' papers usually show that (advanced) statistical models outperform (simple) CI techniques. In addition, given that electrical engineers typically have no training or experience in a statistically

sound validation of the model performance, there is definitely room for improvement and closer cooperation between the two communities." - Weron (2014, p. 1034)

One of the major goals of this thesis is to bring these approaches closer together. We will therefore choose a diverse set of both statistical and machine learning models, and test them in a realistic information setting on the same forecasting data with consistent statistical evaluation methods.

5.2 Forecasting methods

Before we introduce the forecasting methods, let us present the forecasting problem in a general way. Assume a general model with additive errors for the spot price at time t,

$$P_t = \theta_{\mathbf{t}} \left(\mathbf{X_t} \right) + \epsilon_t \tag{1}$$

where $\theta_{\mathbf{t}}$ is an arbitrary function of a set of explanatory variables $\mathbf{X_t}$, and ϵ_t is an error term that is serially uncorrelated, have mean zero, and a finite variance.

The model in Equation 1 relates the observed prices to a set of explanatory variables— it is the underlying data generation process.

Unfortunately, we cannot implement it directly. First of all, the real set of explanatory variables affecting prices is unknown to us. Second, the exact structure of $\theta_{\mathbf{t}} \left(\bullet \right)$ which relates prices to explanatory variables is also unknown.

In an applied model we need to choose among a many potentially relevant variables. Depending on what fundamental data we can obtain and how competently we choose explanatory variables and model structure, the variance of the error term will vary.

Furthermore, most observed variables contain measurement error, so they are only an approximation of the true fundamental variables.
A prime example of this is the use wind forecasts in a model (or solar/consumption forecasts). The production of wind power on the next day is not known at the time of the day-ahead auction, so the market participants obtain forecasts from different forecast providers and use these to bid in the auction. Therefore, if we are to correctly include the effect of wind power in a model of the day-ahead prices, we would need both the forecast from all forecast providers, the available capacity from all market participants, information on the forecast provider used by each market participant, and information on any forecast aggregation scheme used internally at market participants. It is not feasible to obtain this information, so we just use a single forecast that inevitably differs from the production market participants have bid into the day-ahead auction.

So because market participants use different measures of the variables in $\mathbf{X_t}$, we face errors in the measurement of $\mathbf{X_t}$ no matter the precision of the single forecast we use.

We also have to either estimate the functional form of $\theta_{\mathbf{t}}$ in equation 1, or make assumptions on the structure of $\theta_{\mathbf{t}}$ and then estimate any parameters. This is the point where the methods presented in this section differ

the most. Broadly, the models can be split into two categories: Those that assume a parametric structure of the functional form (statistical methods) and those that estimate the functional form non-parametrically (machine learning methods).

In the following, we explain the exact methods used in this thesis.

5.2.1 Naïve

The naïve method is the very simple benchmark method where the forecast for next day is the price on a similar previous day. We choose the following model to allow for weekend differences

$$\hat{P}_t = \begin{cases} P_{t-72} & if\, Weekday = Monday \\ P_{t-24} & if\, Weekday = Tuesday \vee Wednesday \vee Thursday \vee Friday \\ P_{t-168} & if\, Weekday = Saturday \vee Sunday \end{cases} \tag{2}$$

5.2.2 ARIMA

The ARIMA method is often used as a univariate benchmark, so we include it here as a second benchmark.

The $ARIMA(p,d,q)$ model can be written as

$$\phi_j(B)\nabla^d P_{j,t} = \theta_j(B)\epsilon_{j,t}, \qquad \epsilon_{j,t} \sim N\left(0, \sigma_j^2\right) \tag{3}$$

where $\phi_j(B) = \left(1 - \phi_{j,1}B - \phi_{j,2}B^2 - \ldots - \phi_{j,p}B^p\right)$, $\theta_j(B) = \left(1 + \theta_{j,1}B + \theta_{j,2}B^2 + \ldots + \theta_{j,q}B^q\right)$, and $\nabla^d = (1-B)^d$. B is the backshift operator such that $B^p P_t = P_{t-p}$.

This specification allows for both a single model framework ($j = 1$) with all observations lined up sequentially such that $P_1, P_2 ... P_n$. It also allows for a 24 model framework ($j = 1, ..., 24$), where the hourly series are extracted and each series $P_1, P_{25,}, P_{49} \ldots$, $P_2, P_{26}, P_{50} \ldots$, $\ldots$, is estimated independently.

The parameter d in Equation 3 determines if the prices are integrated. In Equation 3, it is assumed that d is selected such that $\nabla^d P$ is a stationary process (i.e. does not contain a unit root). If the price series P does contain a unit root, we set $d = 1$ and model the differenced series. Otherwise we set $d = 0$ and model the level series.

We can also set d equal to a real number larger than 0 to obtain a fractionally integrated model, which allows the ARIMA model to exhibit long memory. While these models are interesting, they will not been considered in this thesis.

To test for a unit root in P, we employ the standard Augmented Dickey-Fuller test, which tests the null hypothesis of the series having a unit root (Said and Dickey, 1984). We also employ the KPSS test, which tests the null hypothesis that the series is stationary (Kwiatkowski *et al.*, 1992).

After having settled on d, we need to select the order of p and q. Here, we will use AIC to select between competing models and the superior model is the one with the lowest AIC (or average AIC when 24 hourly

models are used). The error terms from the model should be uncorrelated and we check the model by plotting the autocorrelation function and the partial autocorrelation function. Statistical tests for testing the null of uncorrelated errors (see e.g. Godfrey (1978) and Ljung and Box (1978)) could also be applied. However, this was not deemed necessary as the rejection of the null of uncorrelated errors is easily seen in the plots (and they offer a better diagnostic tool to explain how the errors are correlated).

When forecasting within the 24-model framework, the one-step ahead forecasts from all 24 models (one for each hour) are combined to give a day-ahead hourly price profile. The forecasting equation with $d = 0$ then becomes

$$\hat{P}_{j,t} = \hat{\phi}_{1,j} P_{j,t-1} + \hat{\phi}_{2,j} P_{j,t-2} + \ldots + \hat{\phi}_{p,j} P_{j,t-p} + \hat{\theta}_{1,j} \hat{\epsilon}_{j,t-1} + \hat{\theta}_{2,j} \hat{\epsilon}_{j,t-2} + \ldots + \hat{\theta}_{q,j} \hat{\epsilon}_{j,t-q} \tag{4}$$

In the 1-model framework we need to forecast between 1 and 24 steps ahead. We can use two approaches to forecasting several steps ahead: In direct forecasting, a separate model is estimated for each forecast horizon conditional on the known values, while in iterated forecasting, a one-step-ahead model is iterated to produce all 24 forecasts and all, at the time, unknown prices are replaced by forecasts. Marcellino *et al.* (2006, p. 1) comments, "...in theory, iterated forecasts are more efficient if the one-period ahead model is correctly specified, but direct forecasts are more robust to model misspecification". Due to time constraints, only the direct forecast approach was tested in the 1-model framework. However, it would be prudent to test both approaches, especially as the model is likely to be misspecified.
In the 1-model framework, Equation 4 changes so that we iterate the model and replace auctual prices with forecasted prices, whenever these are not yet known. Errors that are not yet observed are dropped from the equation. For estimation of ARIMA models see e.g. (Hamilton, 1994).

The $ARIMA(p, d, q)$ model replaces $\theta_{\mathbf{t}}(\bullet)$ in Equation 1 with a parametric function of past observed prices and forecast errors.

5.2.3 Linear

The linear model can be written as

$$P_{j,t} = \boldsymbol{\beta}_j^T \boldsymbol{X}_{j,t} + \epsilon_{j,t}, \qquad \epsilon_{j,t} \sim N\left(0, \sigma_j^2\right) \tag{5}$$

where $\boldsymbol{X}_{j,t}$ is a $(k \times 1)$ vector of explanatory variables, $\boldsymbol{\beta}_j$ is the corresponding $(k \times 1)$ parameter vector, and $\epsilon_{j,t}$ is a disturbance term.

The specification in Equation 5 allows for both a 1-model framework and a 24-model framework (as in the ARIMA method above), and we will test both specifications. When choosing between models with different specifications of $\boldsymbol{X}_{j,t}$ we will use AIC to choose the best model.

We estimate the model(s) using OLS and forecast using

$$\hat{P}_{j,t} = \hat{\boldsymbol{\beta}}_j^T \boldsymbol{X}_{j,t}$$

The linear model replaces $\theta_{\mathbf{t}}(\bullet)$ in equation 1 by a linear combination of the explanatory variables, $\hat{\boldsymbol{\beta}}_j^T \boldsymbol{X}_{j,t}$.

5.2.4 ARMAX

The transfer function in Equation 6 extends the ARIMA model from Equation 3 by including explanatory variables

$$\phi_j(B)\nabla^d P_{j,t} = \boldsymbol{\beta}^{\boldsymbol{T}}{}_j(B)\nabla^d \boldsymbol{X}_{j,t} + \theta_j(B)\eta_{j,t}, \qquad \eta_{j,t} \sim N\left(0, \sigma_j^2\right) \tag{6}$$

where $\boldsymbol{X}_{j,t}$ is a $(k \times 1)$ vector of explanatory variables, $\boldsymbol{\beta}_j(B) = \left(1 - \beta_{j,1}B - \beta_{j,2}B^2 - \ldots - \beta_{j,r}B^r\right)$ is a $(k \times 1)$ row vector, $\theta_j(B) = \left(1 + \theta_{j,1}B + \theta_{j,2}B^2 + \ldots + \theta_{j,q}B^q\right)$, $\phi_j(B) = \left(1 - \phi_{j,1}B - \phi_{j,2}B^2 - \ldots - \phi_{j,p}B^p\right)$, and $\nabla^d = (1-B)^d$. Where B is the backshift operator such that $B^p P_t = P_{t-p}$.

The transfer function has seen frequent use in electricity price forecasting, as it allows the lagged explanatory variables such as previously realized load to affect the future prices. However, as we only use forecasts as explanatory variables, in our case, we will assume that no complex dynamics in explanatory variables are necessary, and therefore simplify Equation 6 by setting $\boldsymbol{\beta}^{\boldsymbol{T}}{}_j(B) = \boldsymbol{\beta}_j^T \phi_j(B)$ leading to

$$\nabla^d P_{j,t} = \boldsymbol{\beta}_{\mathbf{j}}^{\mathbf{T}} \nabla^d \boldsymbol{X}_{\mathbf{j},\mathbf{t}} + \frac{\theta_j(B)}{\phi_j(B)}\eta_{j,t}, \qquad \eta_{j,t} \sim N\left(0, \sigma_j^2\right) \tag{7}$$

When we estimate the model, we use a two-step estimation approach where a linear model is estimated first, using the techniques from Section 5.2.3, and then we model the residual with an ARMA model using the techniques from Section 5.2.2. For our data, we end up selecting $d = 0$ which simplifies Equation 7 further, and leads to our definition of the ARMAX model

$$\begin{aligned} P_{j,t} &= \boldsymbol{\beta}_j^T \boldsymbol{X}_{j,t} + \epsilon_{j,t} \\ \phi_j(B)\epsilon_{j,t} &= \theta_j(B)\eta_{j,t} \quad , \qquad \eta_{j,t} \sim N\left(0, \sigma_j^2\right) \end{aligned} \tag{8}$$

Where the model is split in two equations to highlight that it is estimated in two steps. When estimating this model, we use a 1-model version for the first-step linear model and a 24-model for the second-step ARMA part— reflecting the choices in the previous sections. The number of AR and MA terms in the residual model is chosen by finding the model that minimizes the AIC (or average AIC across the 24 models).

When forecasting, we use the linear model to predict the first term and the ARMA residual model to predict the second term

$$\hat{P}_{j,t} = \hat{\boldsymbol{\beta}}_j^T \boldsymbol{X}_{j,t} + \hat{\epsilon}_{j,t} \tag{9}$$

The ARMAX model replaces $\theta_{\mathbf{t}}(\bullet)$ in Equation 1 with a parametric function of both explanatory variables, past observed prices, and past forecast errors.

5.2.5 Jonsson

The model of Jónsson *et al.* (2013) has its roots in the locally weighted regression of Cleveland and Devlin (1988) and subsequent expansions to cover recursive estimation (Nielsen *et al.*, 2000) and robust estimation (Pinson *et al.*, 2007). The Jonsson model works by creating a forecast grid from a number of local linear approximations that are conditional on explanatory variables,.

In their article, Jónsson *et al.* (2013) use the day-ahead consumption forecast issued by the Danish Transmission System Operator (TSO) and a wind forecast from a commercial forecast system to predict the day-ahead

price in the West Danish (DK1) price zone.

For the rest of this thesis, we will refer to this model as the Jonsson model.

The following description closely tracks the clear presentation found in (Jónsson *et al.*, 2013).

We will follow Jónsson *et al.* (2013) and allow for two variables in $\boldsymbol{X_t}$, and so let $\boldsymbol{X} = \begin{bmatrix} x_1 & x_2 \end{bmatrix}^T$ be a column vector representing a particular fitting point in a grid. Let $p(\boldsymbol{X}) = \begin{bmatrix} 1 & x_1 & x_2 & x_1^2 & x_1x_2 & x_2^2 \end{bmatrix}^T$ denote a column vector containing the terms of the corresponding second-order Kolmogorov-Gabor polynomial. In Jónsson *et al.* (2013) a polynomial of order two was chosen after trials with {1, 2, 3}, and here we will assume that this is sufficient to approximate the functional behaviour in the vicinity of a grid point. Now define

$$\boldsymbol{\phi}_{x,t}^T = \begin{bmatrix} \phi_{1,t} & \phi_{x_1,t} & \phi_{x_2,t} & \phi_{x_1^2,t} & \phi_{x_1x_2,t} & \phi_{x_2^2,t} \end{bmatrix} \tag{10}$$

as a column vector of coefficients such that the model

$$P_t = p^T(\boldsymbol{X}_t)\,\boldsymbol{\phi}_{\mathbf{x},\mathbf{t}} + \epsilon_t \tag{11}$$

describes the prices in the close vicinity of a grid point $\boldsymbol{X}_t$, where ϵ_t is a noise term that is mean zero and with a finite variance.
The parameters in Equation 11 are estimated using recursive and robust weighted least squares. The optimization problem becomes

$$\hat{\boldsymbol{\phi}}_{x,t} = \arg\min_{\phi_{x,t}} \sum_{s=1}^{t} \lambda^{t-s} \omega_x(X_s) \left(g(e_s, \tau)\right)^2 \tag{12}$$

where $e_t = P_t - p^T(\boldsymbol{X}_t)\,\boldsymbol{\phi}_{\mathbf{x},\mathbf{t}}$ is the forecast error using the parameters in a specific grid point, and $0 < \lambda < 1$ is a forgetting factor that exponentially discounts past observations. The tuning parameter λ is determined later on. $\omega_x(\boldsymbol{X}_t)$ is the weight assigned to the observation $\boldsymbol{X}_t$ as a function of its distance to the fitting point $\boldsymbol{X}$. $g(e_s, \tau)$ is the Huber influence function

$$g(e_t, \tau) = sgn(e_t) \times \min(e_t, \tau) \tag{13}$$

where the tuning paramter τ is a cutoff value that controls the maximum influence a single observation is allowed to have, and is determined later on. $sgn(\bullet)$ is the sign function and $\min(z, w)$ is the minimum of the variables z and w.
The weights are assigned as

$$\omega_x(X_t) = W\left(\frac{\|X - X_t\|}{h(x)}\right) \tag{14}$$

where the weighting function $W(\bullet)$ is chosen to be the tri-cube kernel in Equation 15, $\| \bullet \|$ denotes the Euclidean norm, and $h(x)$ is a fitting point dependent bandwidth. For each fitting point the bandwidth is set as the γ-quantile of the Euclidian distances between the fitting point and all observations in the training set. Where the tuning parameter γ is also determined later on.

$$W(x) = \begin{cases} \left(1 - x^3\right)^3, & if\, x \in [0;1) \\ 0 & otherwise \end{cases} \tag{15}$$

In Pinson *et al.* (2007) and Madsen (2008, Ch. 11) it is shown that the adaptive parameter estimates in Equation 12 can be found by starting from intital, guessed, values of $\hat{\phi}_{x,0}$ and $R_{u,0}$, and then updating the estimates recursively using

$$\hat{\phi}_{x,t} = \hat{\phi}_{x,t-1} + \omega_x(X_t) R_{u,t}^{-1} p(X_t) g\left(e_t, \tau\right) \tag{16}$$

and

$$R_{u,t} = \lambda R_{u,t-1} + \omega_u\left(X_t\right) \frac{\partial g\left(e_t, \tau\right)}{\partial e_t} p\left(X_t\right) p^T\left(X_t\right) \tag{17}$$

Abruptly changing parameter estimates are avoided by following Nielsen *et al.* (2000) and defining the effective forgetting factor λ_t^* as

$$\lambda_t^* = 1 - (1 - \lambda)\, \omega_u\left(X_t\right) \frac{\partial g(e_t, \tau)}{\partial e_t} \tag{18}$$

and then updating Equation 17, yielding

$$R_{u,t} = \lambda_t^* R_{u,t-1} + \omega_u\left(X_t\right) \frac{\partial g\left(e_t, \tau\right)}{\partial e_t} p\left(X_t\right) p^T\left(X_t\right) \tag{19}$$

Finally, $\theta_{\mathbf{t}}(X)$ is estimated in each fitting point $X = \begin{bmatrix} x_1 & x_2 \end{bmatrix}^T$ by

$$\hat{P} = p^T(X_t) \hat{\phi}_{x,t} \tag{20}$$

This gives a forecast surface based on the fitting points. Forecasts for specific values of X_t are found by interpolation between the grid of fitting points. Jónsson *et al.* (2013) use linear interpolation, but we will use the interpolation of Akima (1978) as implemented in the R-package Akima (Akima and Gebhardt, 2013).

The tuning parameters γ, λ, and τ are chosen as the values that minimize the RMSE of the day-ahead forecast during the training period

$$RMSE_{DA}\left(\gamma, \lambda, \tau\right) = \sqrt{\frac{1}{N} \sum_{t=1}^{N} \left(P_t - \hat{P}_t\right)^2}$$

Because of potential autocorrelation in the residuals, Jónsson *et al.* (2013) also apply a second-step model to the residuals. They test both a robustified AR method and a Double-Seasonal Holt-Winters method using 24 models— one for each hour. In their paper, they observe a modest improvement in out-of-sample performance, relative to the step-one model, when implementing either of the second step models (around a 4-5% reduction in RMSE, with AR being slightly superior).

We will also implement a second step model, but will only be using a simple ARIMA model without robustification. We model the Jonsson step-one residual series by following the method explained for the ARIMA model in Section 5.2.2. The Jonsson second-step forecast equation then becomes

$$\hat{P} = \hat{P}_{Jonsson} + \hat{E}_{ResidualModel} \tag{21}$$

where $\hat{P}_{Jonsson}$ is the forecast from the Jonsson first-step model and $\hat{E}_{ResidualModel}$ is the residual forecast from the second-step ARIMA model.

The Jonsson model operationalizes Equation 1 by approximating $\theta_{\mathbf{t}}(X_t)$ with a linear quadratic model at each point in a grid of fitting points using two explanatory variables. The second-step model also makes this approximation dependent on past forecast errors.

The R-code used to estimate the Jonsson model can be found in Section 12.2.4 in the appendix.

5.2.6 Neural network

Artificial neural networks have gained popularity in many areas of forecasting, and they have also been used extensively in the forecasting of electricity loads and prices. A neural network is a statistical method that can learn structures in training examples that is feed to it, and then use this learned structure to produce forecasts for new inputs.

A nice theoretical feature of neural networks is that even simple single-layer neural networks are universal approximators— meaning that they can approximate any linear or non-linear function arbitrary accurately (under mild regularity assumptions on the activation function) using a finite, but potentially large, number of neurons (Cybenko, 1989; Funahashi, 1989; Hornik *et al.*, 1989; White, 1990).

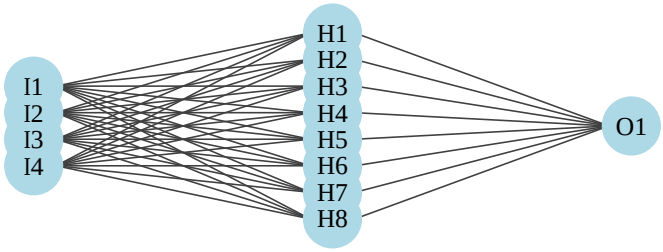

Figure 7 – Example layout of a single-layer feed-forward neural network with 4 inputs, 8 hidden neurons and 1 output. All lines represent a connection between two nodes and has a corresponding weight that has to be estimated. The figure was generated using the R package NeuralNetTools (Beck, 2015).

The neural network has an input layer, an output layer, and one or more intermediate layers of 'hidden neurons' (respectively layer I, O and H in Figure 7).
Each intermediate neuron has an activation function that converts the dot product of the nodes in the previous layer and an estimated weight vector to an output value from that neuron (typically a value between -1 and 1 or 0 and 1, depending on the activation function used). At the output layer the dot product of the previous layer of neurons and another weight vector becomes the predicted output value.

The weights in a neural network are estimated by adjusting them so that, given example input data, the model produces predicted outputs that match the observed outputs as closely as possible— this is called

training the network. A common training algorithm is the back-propagation algorithm of Rumelhart *et al.* (1986) but other optimization algorithms such as simulated annealing (Goffe *et al.*, 1994) can also be used. We will not delve further on the theory of neural networks but instead refer to (Russell and Norvig, 2010, p. 732-737) for a more in-depth introduction and (Medeiros *et al.*, 2006) for a neural network modelling strategy based on statistical inference.

The important questions we face when constructing a neural network is: How many intermediate neurons to include in each layer, how many layers of neurons to include, what activation function to use, if there should be many outputs in one model or seperate models for each output, and if the network should be fully connected or some connections should be eliminated to put structure on the neural network.

Furthermore there exist two flavors of neural networks: feed-forward networks and recurrent (feedback) networks. In a feed-forward network all links are directed from inputs through intermediate layers to outputs. In a recurrent network there can also be links backwards, so that the later layers can affect the previous layers. This complicates computation and design immensely, but also allows for more complicated behaviour.

Gupta (2000, p. 74) states, "In the function approximation or regression area where generalization capability is a major concern, three-layer perceptrons are usually preferred...". Due to this statement and the universal approximation ability, we restricts our search to fully connected feed-forward neural networks that have one hidden layer and use the common sigmoid activation function in the hidden layer (Equation 23). Furthermore we choose to construct one model for all hours. The main questions then become what input variables to choose and how many neurons to use. To choose between the different network configurations we will use cross-validation.

Below, we illustrate how the neural network is implemented using the example network from Figure 7. In Equation 22, the sum of all connections to a hidden neuron multiplied by the corresponding weights, is computed and then feed into a sigmoid activation function (Equation 23). This function attains a value between 0 and 1 and performs the function of a neuron 'firing' as a soft threshold is reached. To include a threshold in the hidden units, which h_j has to surpass to activate the neuron, we can include an input unit with a permanent value of 1.

$$h_j = \sum_{i=1.}^{4} I_i w_{1,i} \tag{22}$$

$$H_j = \frac{1}{1 + e^{-h_j}} \tag{23}$$

where I_I is the value of an explanatory variable (with 4 explanatory variables in the example neural network). $w_{1,i}$ is the corresponding parameter. h_j is then the input to hidden unit j, and H_j is the output from hidden unit j.
In Equation 24, the value at the output node is computed. Here, a linear combination of the values in the hidden layer is calculated and becomes the predicted value from the neural network. Again, we can include a bias by adding a hidden neuron without connections to the input layer with the constant value of 1.

$$O_1 = \sum_{j=1}^{8} H_j w_{2,j} \tag{24}$$

Note how in the neural network, $\theta_{\mathbf{t}}(\mathbf{X_t})$ in Equation 1 is replaced with a complex function of exogenous variables that depends both on the estimated weights and the chosen network structure, $\left(\hat{\theta}(\mathbf{X_t}, \hat{w})\right)$.

To facilitate convergence in the training algorithm, we center and scale all inputs to unit variance prior to estimation.

5.2.7 Support vector machine

The Support vector machine (SVM) framework is a relatively recent development in applied statistics, that has been made feasible by increasing computational power. Support Vector Machines used for classification tasks where first introduced by Cortes and Vapnik (1995) and extended to regression tasks in Drucker *et al.* (1997). As it is relatively easy to implement and requires little knowledge of the forecasting setting, "[the] SVM framework is currently the most popular approach for "off-the-shelf" supervised learning"–(Russell and Norvig, 2010, p. 744).

In the SVM regression framework, we fit a multidimensional 'tube' with a radius of ϵ around the data points in the training sample. Points outside of the tube are penalized according to their distance from the tube, and the tube with the lowest penalty is chosen.

The following description of the SVM regression framework follows Schölkopf and Smola (2003).

We want to estimate the linear equation

$$f(\boldsymbol{x}) = \boldsymbol{w^T} \cdot \mathbf{x} + b \tag{25}$$

where $\boldsymbol{x}$ is a $(k \times 1)$ vector of explanatory variables and $\boldsymbol{w}$ is a corresponding $(k \times 1)$ weight vector.
To estimate the $\boldsymbol{w}$ vector, allowing for the ϵ-tube, we minimize the following objective function

$$\frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{N} |y_i - f(\boldsymbol{x_i})|_\epsilon \tag{26}$$

where $|y - f(\boldsymbol{x})|_\epsilon := \max\{0, |y - f(\boldsymbol{x})| - \epsilon\}$ is the loss associated with an observation i, N is the number of observations, and C is the relative weight given to losses associated with points outside the tube compared to the regularization term $\frac{1}{2}\|\mathbf{w}\|^2$. In other words, we do penalized L1-norm regression and shrink parameters towards zero.
This can be written as a constrained optimization problem

$$\begin{aligned} minimize \quad & \tau(\mathbf{w}, \xi, \xi^*) = \tfrac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*) \\ subject\,to \quad & \left(\boldsymbol{w^T} \cdot \mathbf{x} + b\right) - y_i \leq \epsilon + \xi_i \\ & y_i - \left(\boldsymbol{w^T} \cdot \mathbf{x} + b\right) \leq \epsilon + \xi_i^* \\ & \xi_i^*, \xi_i \geq 0 \end{aligned} \tag{27}$$

so that whenever an observation lies within the ϵ-tube both ξ_i and ξ_i^* are zero, and the observation does not appear in the objective function.

We continue to follow Schölkopf and Smola (2003), and introduce Lagrange multipliers to reframe Equation 27 as the following optimization problem

$$\begin{aligned} &\textit{maximize} \quad W(\alpha, \alpha^*) = -\epsilon \sum_{i=1}^{N} (\alpha_i^* + \alpha_i) + \sum_{i=1}^{N} (\alpha_i^* - \alpha_i) y_i - \tfrac{1}{2} \sum_{i,j=1}^{N} (\alpha_i^* - \alpha_i) \left(\alpha_j^* - \alpha_j\right) \mathbf{x}_i^T \cdot \mathbf{x}_j \\ &\textit{subject to} \qquad\qquad 0 \leq \alpha_i, \alpha_i^* \leq C,\ i = 1, \ldots, N,\ \textit{and} \sum_{i=1}^{N} (\alpha_i^* - \alpha_i) = 0 \end{aligned} \tag{28}$$

where the corresponding prediction function for observations of the explanatory variables is

$$f(\mathbf{x}) = \sum_{i=1}^{N} (\alpha_i^* - \alpha_i) \mathbf{x}_i^T \cdot \mathbf{x} + b \tag{29}$$

Note that in Equation 28 & 29, the explanatory variables only enter as a dot product. In the following, this will allow us to apply what is usually termed the 'kernel trick'.
Define a mapping of the explanatory variables x into a higher dimensional 'feature space' $\mathcal{H}$.

$$\begin{aligned} \Phi: \quad \chi &\to \mathcal{H} \\ \boldsymbol{x} &\mapsto \Phi(\boldsymbol{x}) \end{aligned}$$

An example would be the mapping of a two-dimensional vector into a three-dimensional vector, $(x_1, x_2) \mapsto (x_1^2, x_2^2, x_1 x_2)$.
Now, for some mappings the dot product between the two feature space vectors can be easily calculated by a kernel that **only** operates on the original explanatory variables. Using this 'kernel trick', the dot product can be calculated without ever performing the mapping of the explanatory variables into the feature vectors.

$$k(x, y) = \left(\Phi^T(x) \cdot \Phi(y)\right)$$

If we replace $(x_i \cdot x_j)$ with $k(x_i \cdot x_j)$ in Equation 28 & 29, we obtain a kernelized version of the algorithm that finds the best fitting ϵ-tube in the feature space instead of the regular input space of the explanatory variables.

The kernelized algorithm effectively learns how to determine the best fitting tube after the explanatory variables have been mapped into the feature space, without ever mapping into the feature space. As the feature space can be very large (some kernels correspond to a feature spaces of infinitely many dimensions (Russell and Norvig, 2010, p. 748)), we might be able to find a linear ϵ-tube with a better fit, in the feature space. When the linear line in feature space is projected back into the original input space, it can show complex non-linear behaviour (Karatzoglou *et al.*, 2006).

The choice of kernel defines what feature space we map to. There are many choices of kernels, and we will test just two. The popular Radial Basis Function (RBF) kernel

$$k(x, x') = \exp\left(-\sigma \left\|x - x'\right\|^2\right)$$

and the Linear kernel, which does not map into a higher dimensional space

$$k(x, x') = x \cdot x'$$

We also test different input variable specifications. To choose between the different input specifications, we will use cross-validation.

One of the attractive features of SVM's is that the optimization problem gives a unique global solution. As with neural network, the SVM has to be trained and this requires adequate computational power. For a benchmark of different SVM implementations in R and an overview of kernels see Karatzoglou *et al.* (2006).

Note how, the SVM substitutes $\theta_{\mathbf{t}}(\mathbf{X_t})$ from equation 1 with a function of explanatory variables that might be highly non-linear if a complex kernel have been used.

5.3 Model selection

Within each method proposed above there are many choices of both the structure of the function $\hat{\theta}_{\mathbf{t}}(\bullet)$, which approximates the 'true' data generation process $\theta_{\mathbf{t}}(\bullet)$, and the input variables X_t.
To select between the different models, we present two model selection criteria in this section. Both are chosen because of their focus on selecting models for predictive ability.

5.3.1 AIC

The Akaike Information Criterion (AIC) of Akaike (1974) is a popular measure for predictive ability used for model selection in time series literature. AIC is defined as

$$\text{AIC} = -2\ln(\text{maximum likelihood}) + 2(\text{FP}) \tag{30}$$

where FP is the number of independently adjusted parameters within the model (Akaike, 1974).

We choose this measure because it reflects our focus on predictive ability, as seen in the following quote,

> "Akaike derived the AIC from a predictive viewpoint, where the model is not intended to accurately infer the "true distribution," but rather to predict future data as accurately as possible."
> – Shmueli (2010, p. 300)

When using AIC it is neither required that the models being compared are nested, (Teräsvirta and Mellin, 1986), nor that the set of models contains the true model (Hannan, 1980). However, as is seen in Equation 30, we do need to calculate the maximized likelihood to calculate the AIC. For all the statistical models surveyed we calculate the maximum likelihood, and we can therefore use AIC to select between the competing models.

5.3.2 Cross-validation

The idea behind k-fold cross-validation (CV) is that we randomly split our training data into k subsets of data. We then hold out one subset, estimate our model on the rest of the training data, and predict the values of the hold-out subset and compute a test score (e.g. RMSE). This is repeated for all k subsets, and the average test score from the k rounds of modelling is the cross-validated score (Arlot and Celisse, 2010).

By utilizing CV to select models, we avoid using the same data for estimating the model and evaluating its performance. This hopefully helps us select models that have good generalization performance.

If we set k equal to the number of observations, we get the leave-one-out cross-validation procedure. Stone (1977) shows that this procedure is asymptotically equivalent to the AIC measure presented above, so there is at least some correspondence between the two model selection methods employed (though only asymptotically and for the leave-one-out CV).
Russell and Norvig (2010) write, "Popular values for k are 5 and 10— enough to give an estimate that is statistically likely to be accurate, at a cost of 5 to 10 times longer computation time". The same range is reported in (Arlot and Celisse, 2010).

An issue when using CV with time series, and dependent observations in general, is that one of the main assumptions, independence between the validation and training samples, is not satisfied (Arlot and Celisse, 2010). When data are positively correlated, Hart and Wehrly (1986) shows that CV overfits in their setting of choosing the bandwidth of a kernel estimator. While there are customized versions of CV that take dependence into account, we will use the standard CV version when selecting models and have the potential of overfitting in mind.

For the methods employing cross-validation, we will use the cross-validated RMSE from a 7-fold CV procedure ($k = 7$) to select between models.

5.4 Comparing forecast performance

To evaluate and compare different forecasts, we need to choose a proper scoring function and then statistically test any differences in performance.

In this section, we first present the theoretical foundation that is needed when choosing a proper scoring function. Then, we present our choice of scoring function. Finally, we introduce the Diebold-Mariano test for testing and describing differences in forecast performance between models.

5.4.1 Choosing a consistent scoring function

When evaluating the performance of a deterministic forecast of a continuous variable, it is common practice to use the sample average of a scoring function to measure performance

$$\overline{L} = \frac{1}{N} \sum_{i=1}^{N} L\left(\hat{P}_i, P_i\right) \tag{31}$$

where N is the number of forecast cases with $\hat{P}_i$ and P_i respectively a forecast and the corresponding realization. Typically, the choice of scoring function L (also know as the cost function, loss function, or error measure) in Equation 31 is either the absolute error measure, $L\left(\hat{P}_i, P_i\right) = \left|\hat{P}_i - P_i\right|$, the squared error measure $L\left(\hat{P}_i, P_i\right) = \left(\hat{P}_I - P_I\right)^2$, or the absolute percentage error measure $L\left(\hat{P}_i, P_i\right) = \left|\hat{P}_i - P_i\right| / P_{:i}$.
In EPF, some authors apply more unusual scoring functions. Weron (2014, p. 1039) states that, "Probably the most common approach [to normalization of the scoring function] is to normalize the absolute error by the average price obtained in the evaluation interval (e.g. a day, a week).", which correspond to

$$L\left(\hat{P}_i, P_i\right) = \left|\hat{P}_i - P_i\right| / \bar{P}_{weekly\,mean}.$$

In the area of statistics, it has long been known that there is a direct relationship between the choice of scoring function in Equation 31 and the optimal predictor— the predictor that minimizes the expected score conditional on the information set. If one chooses the squared error measure the optimal predictor is the mean, while if one chooses the absolute error measure the optimal predictor is the median (see e.g. Granger (1993); Weiss (1996); Gneiting (2009)).

In Gneiting (2009), the observations from above are presented in a decision theoretical framework, where it is shown that there exist classes of consistent scoring functions that relate to specific 'functionals' of the forecast distribution. An example of a functional of the forecast distribution is the mean, which has, amongst other, the squared error measure as one of its corresponding consistent scoring function. Likewise a consistent scoring function for the median is the absolute error measure. The reader is referred to Gneiting (2009) for thorough treatment of this framework.

The mean absolute percentage error (MAPE), which is commonly used in electricity price forecasting in both business and academia, is only consistent for a nonstandard functional of the forecast distribution, "... which tends to support severe underforecasts as compared to the mean or median" (Gneiting, 2009, p. 758). It is unknown to this author what functional the aforementioned unusual scoring functions correspond to, and what consequences their use have.

Gneiting (2009, p. 757) comments, "If point forecasts are to be issued and evaluated, it is essential that either the scoring function be specified ex ante, or an elicitable[14] target functional be named, such as the mean or a quantile of the predictive distribution, and scoring functions be used that are consistent for the target functional".

In EPF literature there seems to have been given little thought to the issue of choosing a proper consistent scoring function. Weron (2014) reports that, "the most widely used measures of accuracy are those based on absolute errors", and "since absolute errors are hard to compare between different data sets, many authors use measures based on absolute percentage errors".

One of the invoked justifications for working with the absolute error measure, instead of the squared error measure, is that a few large forecast errors can otherwise skew the results. Another justification is that absolute errors are industry standard. One should however be aware that if absolute errors are used, models optimized to forecast the median, and not the mean, will have an advantage.
Most models surveyed are optimized to forecast the mean, so if the intention really is to perform optimally under the absolute error measure the user should instead optimize models to forecast the median (for example, using quantile regression procedures).

Some invoked justifications for working with the absolute percentage error measure, or the more unusual versions of normalization, is that it normalizes the error series and thus allows for a comparison of performance between data sets both across time and between markets.

[14] [A functional that has a consistent scoring function is called elicitable.]

There are however other score functions that are scale invariant. A candidate is the Mean Absolute (Squared) Scaled Error of Hyndman and Koehler (2006), where the absolute (or squared) error is scaled by the performance of a naïve reference method during the estimation period. This removes the scale of the underlying data, avoids many problems associated with absolute percentage errors that are especially relevant in EPF (for example, prices close to zero giving extremely large MAPE, see Hyndman and Koehler (2006) for more details), and keeps the model rankings of its non-scaled counterpart (Gneiting, 2009).
The scaled measures should therefore be preferred over the absolute percentage error measure. Scaled errors have seen some use recently, see e.g. (García-Ascanio and Maté, 2010; Cruz *et al.*, 2011; Jónsson *et al.*, 2013).

In general, it is however questionable whether a ranking of methods using the forecast performances in different data sets is even be sensible in the area of electricity price forecasting (EPF). There are a large number of differences between electricity markets. Some are highly penetrated by wind, solar, and/or hydro power, while others use mostly thermal production. Some use electric heating, while others use distributed heating systems. These differences make some markets highly volatile, while others are less so, leading to forecast performance differences that are due to inherent differences in the forecastability of the electricity price series.
Even on the same market there can be differences in the electricity system between years. One year might be a dry year with low hydro production and high prices, while another might be a wet year with high hydro production and low prices. Changes in regulatory structures, production and transmission capacity can also give year-to-year differences.

For these reasons, any comparison of EPF methods using different underlying estimation and test data is, in my opinion, at best only indicative and largely futile.

We will use one error measure that is not frequently seen in electricitry price forecasting literature but commonly used in other areas of economics. The skill score scales the error measure of one method in a period, by the error measure of a reference method during the same period. The method we use as the reference method is the market consensus benchmark.

According to Gneiting (2009), "the original score and the skill score incur the same ranking, which suggests the notion of consistency and elicitability continue to apply, at least when the test sample size n is large", which leads us to use squared errors when computing the skill score.

The skill score has the intuitive interpretation that a forecast method that performs better than the reference method will have a skill score less than 1 and one that performs worse will have a skill score larger than 1.

5.4.2 Scoring functions

For the sake of clarity, we now present the definitions of the scoring functions used in the thesis.

Mean absolute error The absolute errors are calculated as

$$AE_h = \mid P_h - \hat{P}_h \mid$$

where $\hat{P}_h$ is the forecast price in period h and P_h is the realized price in period h.

To summarize performance across a period, the errors are averaged leading to the mean absolute error (MAE) measure

$$MAE = \frac{1}{N}\sum_{h=1}^{N} \mid P_h - \hat{P}_h \mid \tag{32}$$

and different choices of N leads to different aggregation periods

$$where \begin{cases} N = 24 & \rightarrow Daily\, Mean\, Absolute\, Error\, (dMAE) \\ N = 168 & \rightarrow Weekly\, Mean\, Absolute\, Error\, (wMAE) \\ N = varies & \rightarrow Other \end{cases}$$

MAE is frequently used when evaluating forecast performance in the EPF literature. The measure is somewhat resilient to the large forecast errors that are typical when price spikes occur. But, as described in the previous section, the absolute error measure is not a consisten scoring function for the mean of the forecast distribution. As it is commonly used in the literature, we will report a few tables using this measure in the appendix.

Root mean squared error The root mean squared error (RMSE) measure is defined as

$$RMSE = \sqrt{\frac{1}{N}\sum_{h=1}^{N}\left(P_h - \hat{P}_h\right)^2} \tag{33}$$

where $\hat{P}_h$ is the forecast price in period h, P_h is the realized price in period h, and N is the number of forecast periods. We will average the RMSE over different periods as described for MAE.

The RMSE is a consistent scoring function for the mean of the predictive distribution. Because we want to model the mean of the forecast distribution, we will focus on this error measure. One problem with the measure (and the mean of a distribution with jumps in general) is that it is sensitive to extreme values. Thus the occurrences of extreme prices where the forecast errors are very large, receive a hefty penalty.

Skill score The skill score is defined by

$$Skill\, Score = \frac{\left(\frac{1}{N}\sum_{h=1}^{N}\left|\hat{P}_h - P_h\right|^k\right)}{\left(\frac{1}{N}\sum_{h=1}^{N}\left|\hat{P}_{MCB,h} - P_h\right|^k\right)} \tag{34}$$

where $\hat{P}_h$ is the forecast in period h from the model, $\hat{P}_{MCB,h}$ is the forecast from the market consensus benchmark in period h, P_h is the realized price in period h, N is the number of forecast periods and $k \in \{1, 2\}$. We mostly use the squared error measure ($k = 2$) due to the aforementioned consistency results, but in the appendix we also show tables with the absolute error measure ($k = 1$).

The skill score measures how a model performs relative to a reference model. We choose the MCB forecast as the reference model, as it puts nice theoretical bounds on how well a model should be able to perform, and it relates the model performance to the performance of professional market participants. A model having a

skill score below 1 can possibly be used to speculate between the power exchanges, so it would be surprising to find this good performance.

5.4.3 Diebold-Mariano test

The performance measures presented above allow us to measure how models perform relative to each other, by comparing their scores, but they do not tell us whether the differences are statistically significant. In this section, we present the Diebold-Mariano (DM) test, introduced by Diebold and Mariano (1995), that allows us to make statements on the statistical significance of differences in the forecast scores.

In the following description, we follow the description of the DM test in Diebold (2013).

Unconditional The standard DM test takes the forecast errors from two series and makes assumptions on these. Define the loss associated with a forecast error e_t as $L(e_t)$. For the squared errors we work with $L(e_t) = e_t^2$. Then define the loss differential between two forecasts as $d_{12t} = L(e_{1t}) - L(e_{2t})$ and make the following assumptions on the loss differential

$$Assumptions\, DM: \begin{cases} E(d_{12t}) = \mu & ,\forall t \\ cov\left(d_{12t}, d_{12(t-\tau)}\right) = \gamma(\tau) & ,\forall t \\ 0 < var(d_{12t}) = \sigma^2 < \infty \end{cases} \tag{35}$$

Then the hypothesis of equal predictive accuracy (equal expected loss) corresponds to $E(d_{12t}) = 0$ and then under the assumptions stated, the test is (Diebold and Mariano, 1995, p. 4)

$$DM_{12} = \frac{\bar{d}_{12}}{\hat{\sigma}_{\bar{d}_{12}}} \xrightarrow{d} N(0,1) \tag{36}$$

where $\bar{d} = \frac{1}{T}\sum_{t=1}^{T} d_{12t}$ is the mean loss differential of the sample and $\hat{\sigma}_{\bar{d}_{12}}$ is a consistent estimate of the standard deviation of $\bar{d}_{12}$. The loss differential might be serially correlated for many reasons, so we need to estimate the standard error of the loss differential in a robust way.
One way to do this is in a regression framework. If we regress the loss differential on an intercept and use heteroskedastic and autocorrelation robust (HAC) standard errors, the test in Equation 36 boils down to whether the intercept is statistically significant using the HAC standard errors.

Conditional Diebold and Mariano (1995, p. 11) also suggest that the regression test in Equation 36 can be made conditional by including exogenous variables, in order to test for differences in the expected loss differential across business cycles. This conditional aspect is investigated further in Giacomini and White (2006) where a single test of conditional predictive ability is developed.
Here we will just extend the unconditional DM test and use it descriptively. To implement this extension, we demean a set of exogenous variables and then regress the loss differential on an intercept, the demeaned exogenous variables, and business cycle dummies. Again using HAC standard errors (Equation 37).
We can then test the null hypothesis of equal average performance $\alpha = 0$, and the null hypothesis that the loss differential is unaffected by exogenous variables $\boldsymbol{\beta} = 0$.

$$d_{12t} = \alpha + \boldsymbol{\beta}^T \cdot Z_t + \varepsilon_t \tag{37}$$

Where $\boldsymbol{Z}_t$ is a $(k \times 1)$ vector holding all the variables we condition on, $\boldsymbol{\beta}$ is the corresponding $(k \times 1)$ parameter vector, α is the intercept, and ε_t is a potentially serially correlated error term.

The extended DM test is interesting to our case, because we can compare with the market consensus benchmark. It will allow us to test both if models on average perform equal to the market benchmark and if the performance relative to the market benchmark is unaffected by certain market conditions (for example, high wind power production or peak/off-peak).

6 Data

In this chapter we first look at how EU legislation has been a key driver behind a trend of increasing data transparency in power markets. We then describe how the data used in this thesis were obtained and cleaned. Lastly, we present the data in graphs and discuss some features of the time series.

6.1 Availability

European Commission legislation has been one of the major drivers of increased data transparency in recent years (see (The European Commission, 2003, 2009a,b, 2013) and (Viehmann, 2011, p. 1)).
The Commission has tasked the European Network of Transmission System Operators for Electricity (ENTSOE) with creating a "central information transparency platform", where European TSOs and other data owners are required to publish a

> "...minimum common set of data relating to generation, transportation and consumption of electricity to be made available to market participants." - (The European Commission, 2013, article 1).

In (The European Commission, 2013) it is specified exactly what data should be made available to the market through the new transparency platform[15] and at what point in time. This transparency data includes both the data used in this thesis and a range of other interesting data that we will not use for reasons of brevity.

While this thesis only focus on forecast performance in the German/Austrian market, the availability of a common pan-European dataset on price-driving fundamentals will allow for studies that compare model forecast performance across different markets. But as the data is not yet available at the time of writing, the interesting subject of how different fundamental structures in the European electricity markets affect model performance is left to future research.

6.2 Overview

Guided by the analysis in Section 3 of what fundamental factors drive electricity prices in Germany, we have obtained a set of explanatory variables. The dataset available consists of the following variables for the 26 months from 01.11.2012 to 31.12.2014:

- day-ahead TSO forecast of solar power production for Germany/Austria
- day-ahead TSO forecast of wind power production for Germany/Austria
- day-ahead TSO forecast of consumption for Germany/Austria
- day-ahead TSO forecast of consumption for France
- hourly auction prices from EPEX
- hourly auction prices from EXAA

[15]The ENTSOE transparency platform became operational on 5th of January 2015, but only contains data from this date onwards. Interested researchers can obtain data easily from this source, which is currently being populated with new data. Link: transparency.entsoe.eu

Additionally, we create the following calendar variables:

- month (12 dummies)
- weekday (7 dummies)
- hour of day (24 dummies)

Furthermore, we group the 3 factors above into the following 3 dummies:

- summer (1st May to 1st September)
- weekend/weekday
- peak (08:00–20:00)/off-peak (20:00–08:00)

Links to data sources can be found in Table 35 in the appendix.

The solar and wind generation forecasts for Germany/Austria were retrieved from the five TSO websites and were then supplemented with data from the transparency section of the European Energy Exchange (EEX) transparency website (where German/Austrian TSOs also report data) when TSO data was missing. As the data is on 15 minute resolution, it was averaged across the four quarters to get time series with an hourly resolution.

The current version[16] of the EEX transparency website[17] shows generation forecasts for all five TSOs, but the data is not available for downloading from the website. Though it should be possible to contact EEX to get the data, I chose to retrieve the data directly from the TSO websites instead to get as close to the source as possible. In future, data availability might improve and anyone searching for data is advised to first check with both the ENTSOE transparency website and EEX (for data prior to 2015), as this data requires the least cleaning and matching.

A few days of data from the different TSO feeds were missing (mostly related to days where clocks are changed) and the missing values were manually filled in using data from EEX. On the days when clocks changed to winter time, the second Hour 3 was deleted from the dataset to preserve a 24 hour structure. On the days when the change was to summer time, the data from Hour 2 was replicated for the missing Hour 3, again to preserve a 24 hour structure.

The data from EEX and TSO websites should always match, but this is not always the case for every TSO. Contacting one relevant TSO and EEX gave no clarification on the issue, but for most periods the discrepancies seem to be minor (on the order of 100 MW), so I will not pursue this data quality issue further.

For the very last day in the dataset (31.12.2014) the wind forecasts for Amprion is a constant 3253 MW across the day. This is obviously an erroneous forecast. It was the only occurrence of a constant wind forecast across a whole day for Amprion, so it does not seem to be a pervasive problem. The issue does not affect our results because we drop the last 3 days in order to have 52 weeks in 2014. However, it does leave

[16] as of march, 2015

[17] Note that the EEX transparency website is different from the new ENTSOE transparency website.

us with some worries about the integrity of the TSO solar and wind forecasts (whether they are sourced from TSO websites, EEX or ENTSOE).

The German load forecast was retrieved from the legacy section of the ENTSOE transparency site. For the whole period of 01.11.2012 to 31.12.2014 there were 533 values missing in the hourly load forecasts (2.81%) and 129 values missing in the load actuals (0.68%).
In five instances, the missing values overlap, but two of these are related to the phantom Hour 3 on the day of the clock change to summer time. To get a complete dataset the load forecast was taken and any missing values were replaced with actual load. For the five hours where both values were missing, the last available load forecast was used. Again, the 2nd Hour 3 on the the day of the clock change to winter time was removed to preserve a 24 hour structure.
The Austrian load forecast was retrieved from the APG website. The aforementioned clock change procedure was also applied to this data.
The French load forecast was retrieved from the RTE website. The data had already been adjusted for clock change by RTE. The method used is thus unknown. The resolution was 30 minutes, so we calculated the hourly average to get a time series on hourly resolution.

The auction prices were obtained from EPEX Spot and EXAA. Again, the aforementioned clock change adjustment procedure was applied to the price data.

One should note that while the TSO forecasts are made before the day-ahead gate closure (they are typically from about 8:00 in the morning) most are only published around 18:00 (after the day-ahead auction but before delivery), so the data cannot be used directly in operational forecasting. Nevertheless, as the TSOs use the same data sources as market participants, equivalent data should be available to the professional forecaster making a day-ahead price forecast.

Finally, we split the dataset into two periods. The training dataset runs from 01.11.2012 to 31.12.2013. The test dataset runs from 01.01.2014 to 28.12.2014[18].

When we select a particular model from one of the forecasting methods (e.g. ARIMA, SVM, etc.), we only use the observations in the training dataset.

When the final models of all forecasting methods have been selected, they will be compared by forecasting the day-ahead EPEX prices of the testing period in an online setting. This means that for each day, a model has data from all previous days available, including realized prices and forecasts of next day fundamentals and the model then produce forecasts of the 24 hourly prices on the next day.
The prediction performance in the testing period will be used to compare the different models.

6.3 Description

We now examine the different variables to get a better understanding of the data and to provide some rationalization on the choice of calendar variables.

[18]Note that we drop 29th, 30th and 31st to have 52 test weeks (though first week has 5 days only), and due to the above-mentioned Amprion data quality problem.

6.3.1 Calendar effects

In power markets the available capacity and bidding strategy of generators change over time. Some of these changes, such as unforeseen outages, are unpredictable, while others follow predictable cycles of both yearly, weekly, and daily nature.
The yearly cycle can arise from operators scheduling maintenance when prices are expected to be low. In the European markets this will usually be in the summer months, when consumption and average prices are low compared to the winter months.
The weekly and daily cycles can arise from a reluctance to operate during out-of-office hours (e.g. night or weekend), so that less capacity is made available to the market, or a higher price is requested.

All the above-mentioned effects are in addition to demand related price effects (e.g. lower prices during the weekend as a result of reduced consumption). Differences in available capacity or bidding behaviour can lead to different prices in periods requiring the same amount of generation, and which would otherwise have the same price.

To accommodate the possible effects of the above cycles we include calendar variables, and we also group them in order to limit the number of dummies.

The peak dummy captures differences that are the results of the night/day effect.
The weekend dummy captures differences caused by changes in business activity during the weekend.
The summer dummy was chosen ad hoc, (to take account of the possible yearly cycle discussed above), by peeking at the 2014 test data (see Figure 38 in the appendix— data was not available for the train period). As could be expected, it was observed that equivalent levels of demand for generation capacity seems to be associated with a higher price during the period defined as summer.

6.3.2 Consumption

When looking at the German/Austrian consumption (Figure 8) and the French consumption (Figure 9), we see that there are significant yearly variations in the time series. This variation is most pronounced in the French system where winter consumption is highest and summer consumption lowest.
We also spot large weekly variations, where weekday consumption is significantly higher than weekend consumption. Furthermore, there are clear drops in consumption during Christmas in each of the years, indicating that there could be variations in demand on other public holidays as well. For this reason, we specifically test holiday performance later on.

When we look at the German/Austrian consumption in July 2014 (Figure 10) more closely, a daily variation becomes apparent. At night the consumption frequently drops 20 GW from the day-time peak. Again, the weekly variation is apparent with lower consumption during weekends. Overall, the consumption patterns on most weekdays are similar while on Saturdays and Sundays there are different, yet distinct, patterns.

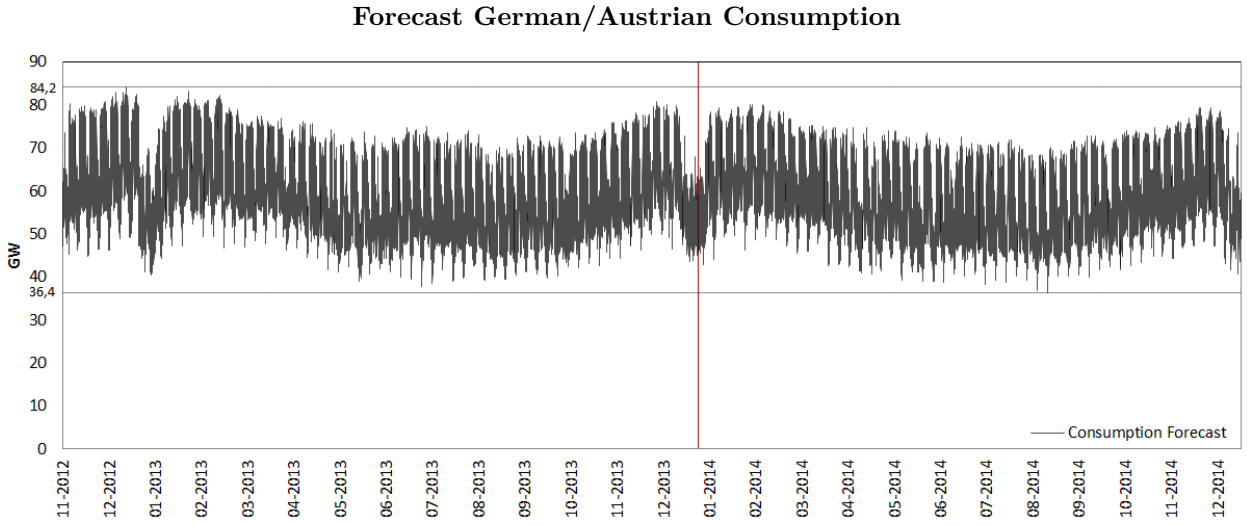

Figure 8 – Forecast German/Austrian consumption. The vertical red line separates train and test period. The horizontal black lines signify sample minimum and maximum. Source: transparency.entsoe.eu.

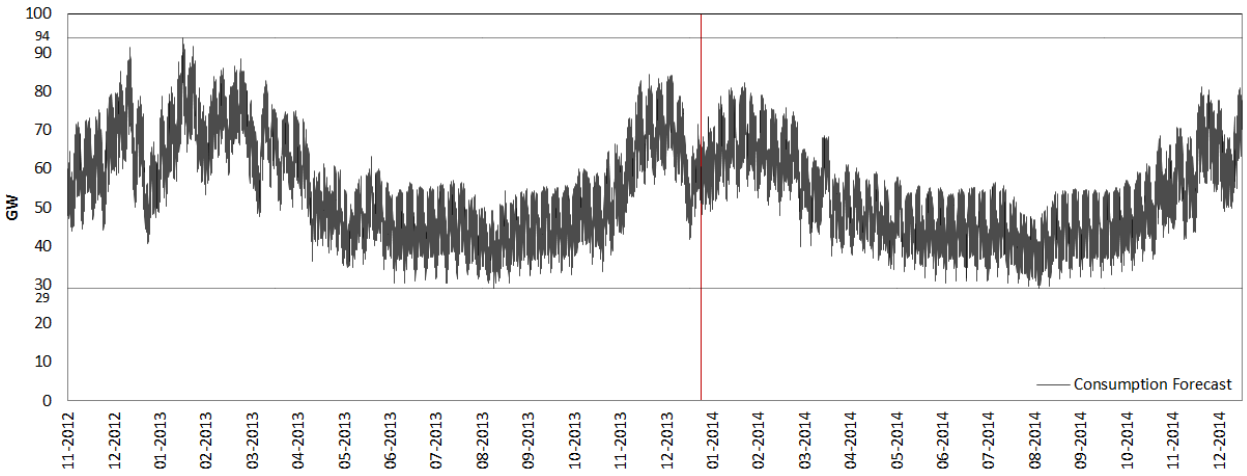

Figure 9 – Forecast French consumption. The vertical red line separates train and test period. The horizontal black lines signify sample minimum and maximum. Source: RTE.

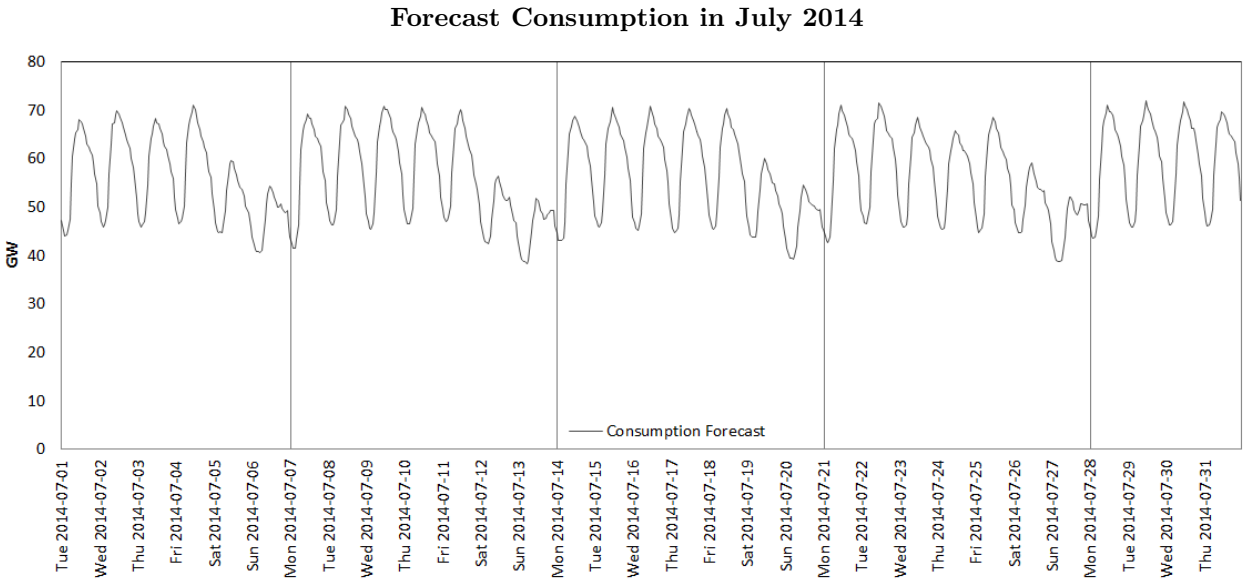

Figure 10 – Forecast German/Austrian consumption in July 2014. Source: transparency.entsoe.eu.

6.3.3 Solar

In Figure 11, we see that solar power production is highly variable. There is a clear yearly cycle with high production during summer and low production in winter.
If we look at production during the summer month of July 2014 (Figure 12), we see large daily variations with high production around noon and zero production at night. The day-to-day variation is clearly affected by cloud cover, so on a clear summer day peak production can be almost 20 GW higher than on a cloudy summer day.

The average production across the sample is around 3.4 GW produced by more than 35 GW in installed capacity (see Table 34).

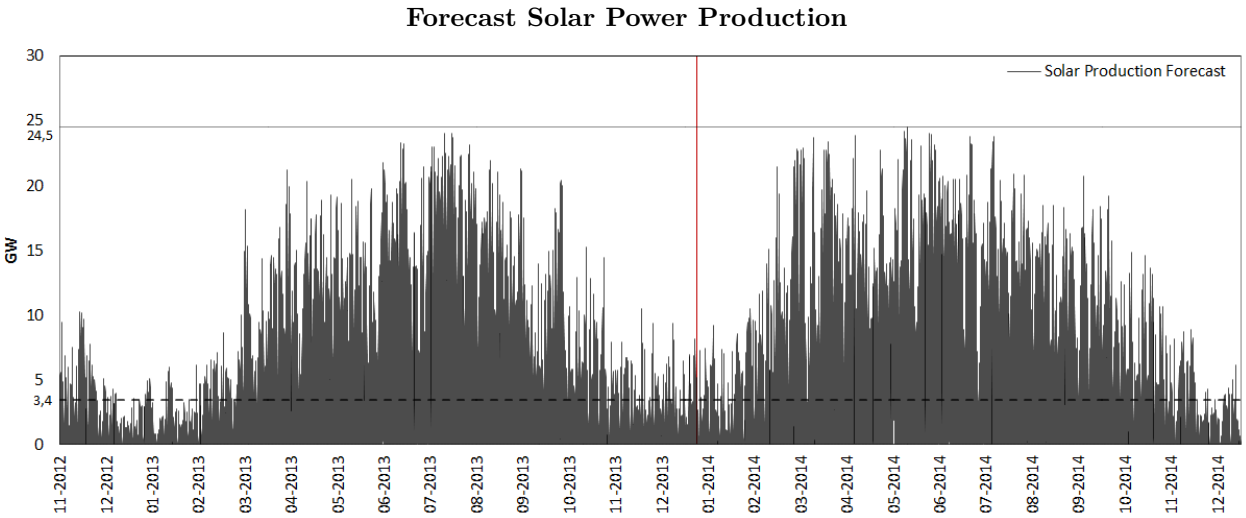

Figure 11 – Forecast hourly German/Austrian solar power production. The vertical red line separates train and test period. The horizontal black lines signify sample minimum and maximum. The horizontal dotted line represents sample average. Sources: TSO's & EEX.

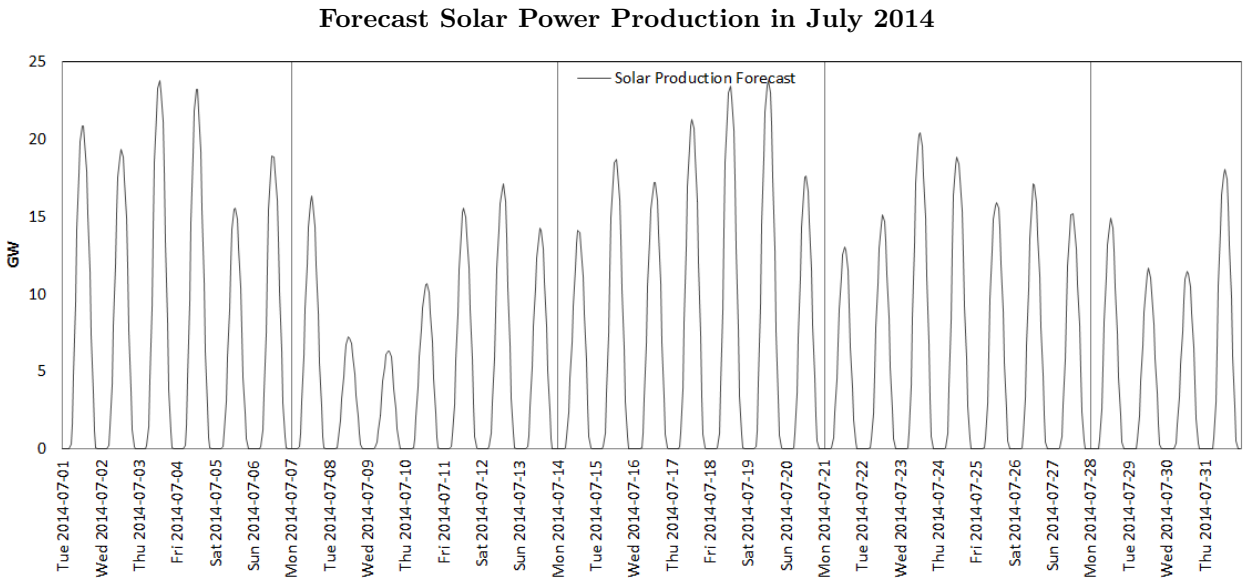

Figure 12 – Forecast German/Austrian hourly solar power production. Sources: TSO's & EEX.

6.3.4 Wind

In Figure 13, we see that wind power production also shows large variations. Although there seems to be a weak yearly cycle with lowest production in the summer, it is not nearly as pronounced as with solar or consumption. Wind power is characterized by bursts of high production across one or more days. This is seen in Figure 14 (July 2014), where wind production is more stable than solar production (though at a lower level for this particular month).

The average production across the sample was 6.3 GW from close to 40 GW in installed capacity (Table 1).

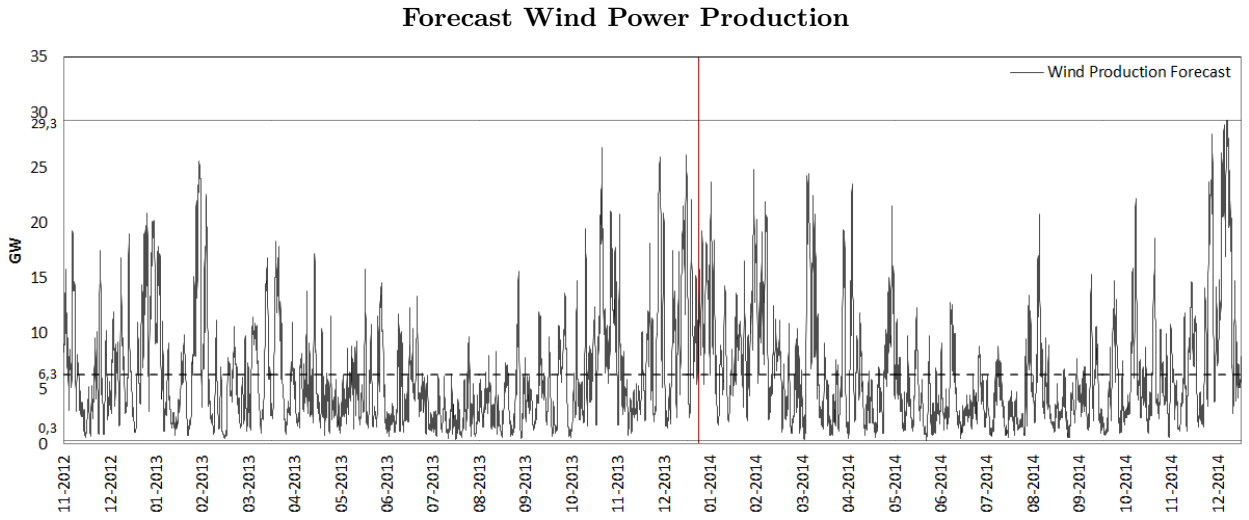

Figure 13 – Forecast hourly German/Austrian wind power production. The vertical red line separates train and test period. The horizontal black lines signify sample minimum and maximum. The horizontal dotted line represents sample average. Sources: TSO's & EEX.

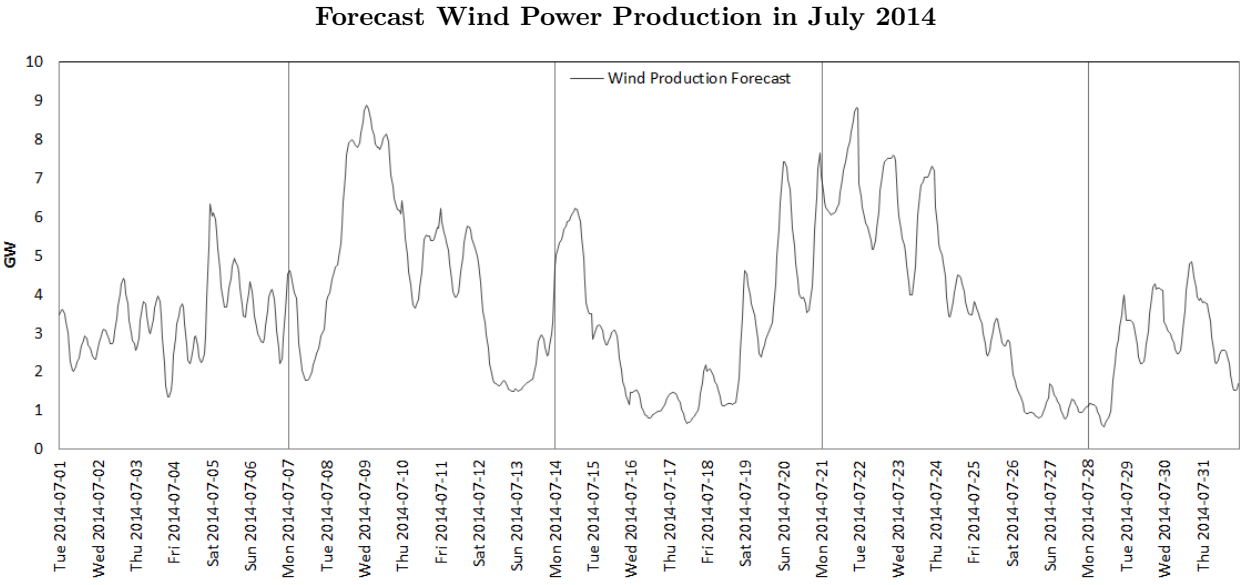

Figure 14 – Forecast German/Austrian hourly wind power production. Sources: TSO's & EEX.

6.3.5 Omitted Variables

We know that some fundamentals affect bidding in the power market and are predictable, but we simply do not have data on them. Examples of such variables are fuel prices, available capacity and planned decommissioning of power plants. These missing variables presumably cause slow and fast changes in the system that makes older data less representative of the future price-demand-production relationship that we want to model.

When facing this problem we have two options: 1) Get more data on the omitted variables or 2) give larger weights to newer and more representative data— using, for example, weighted least squares.

As seen in the Section 7, we will attempt to prevent stale data by using a 365-days rolling estimation window in all models except for the Jonsson model where adaptive equations already discount older data directly in the estimation through the λ parameter.

6.3.6 Prices

In Figure 15 we see that the EPEX auction price shows large variability across the sample. There are some notable negative spikes in both the training and the testing period. The negative spikes around Christmas 2012 are especially severe, with prices down to -222 €. There are also some occurrences of positive spikes with prices above 100 €, but these are confined to the train period. Overall, prices seem more stable in the later testing period than in the earlier training period.

When we compare the EPEX prices (Figure 16) and EXAA prices (Figure 17), we have to take into account that EXAA changed their price floor from 0 € to -150 € in October 2013. After the change in price floor, the EPEX auction seems to exhibit the most extreme prices with EXAA showing only dampened versions of the price spikes.

Focusing on July 2014 (Figure 18) we observe some interesting price dynamics. Firstly, the price at night is typically $10 - 15$ € lower than during the day. Secondly, there is a dip in prices around noon caused by the very large solar production in July (8th July saw the second lowest solar production of the month and saw no dip). Thirdly, there are generally lower prices during the weekdays.
So not surprisingly, the hourly prices follow both consumption and production patterns, and show complex dynamics.

Finally, we also note from Figure 18 that the EXAA price tracks the EPEX price rather well.

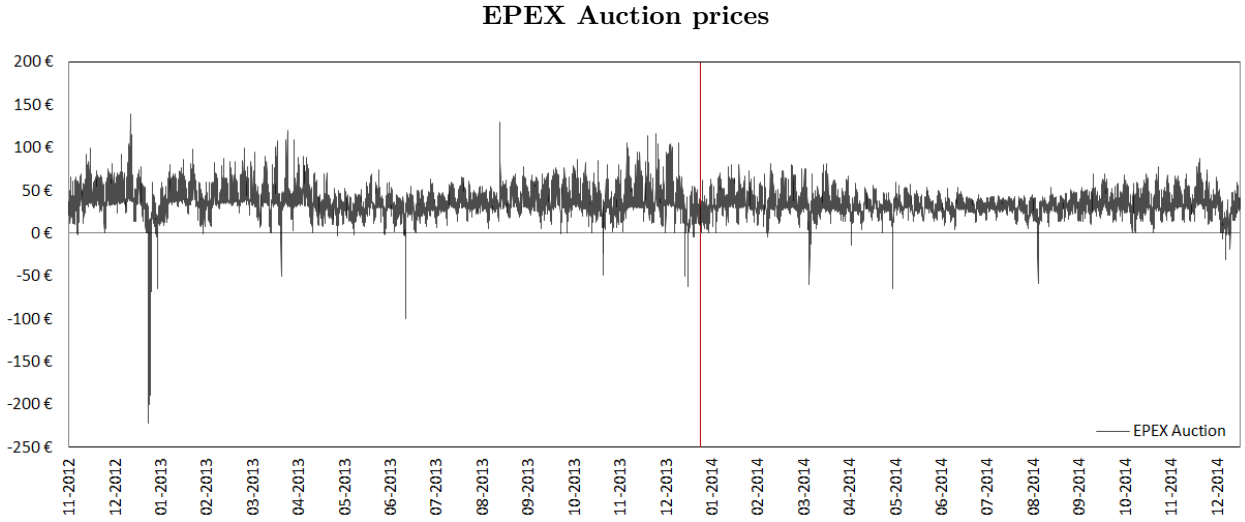

Figure 15 – EPEX auction price for Germany/Austria. The vertical red line separates the train and test period. Source: EPEX Spot.

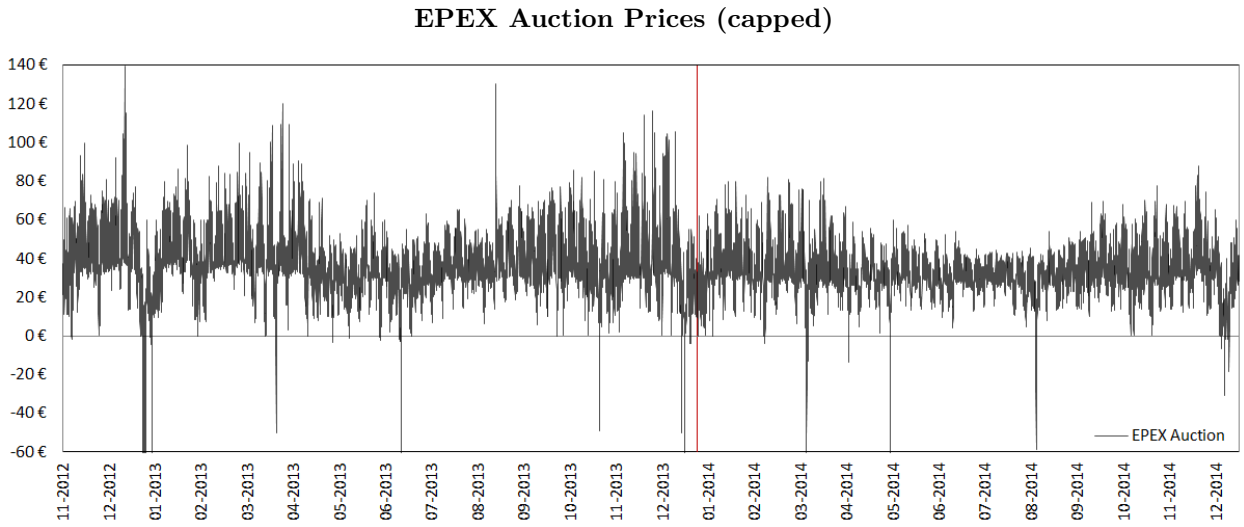

Figure 16 – EPEX auction price for Germany/Austria on the same scale as Figure 17. The vertical red line separates the train and test period. Source: EPEX Spot.

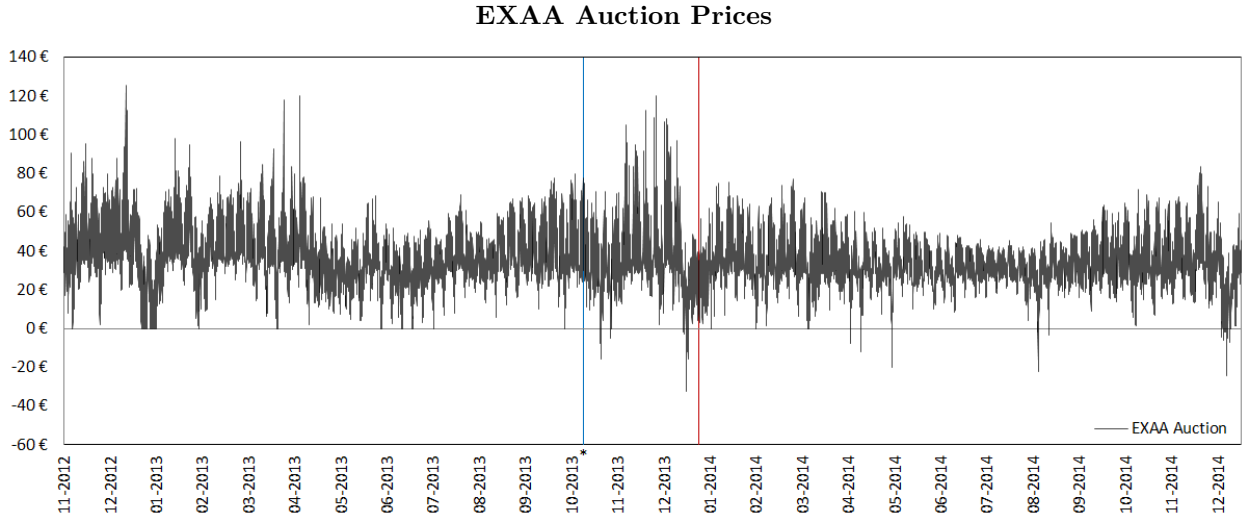

Figure 17 – EXAA auction price for Germany/Austria. *EXAA lowered the price floor from 0 €/MWh to −150 €/MWh on 16th October 2013 (EXAA, 2013, p. 5). The vertical red line separates the train and test period. Source: EXAA.

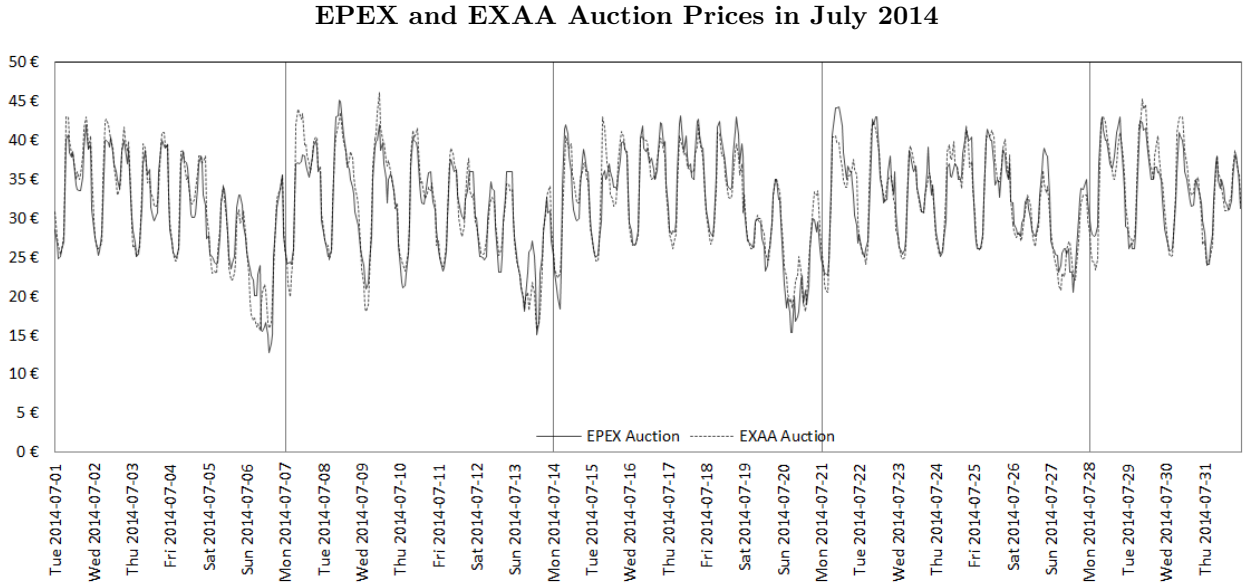

Figure 18 – Hourly EPEX and EXAA prices in July 2014. Sources: EPEX Spot and EXAA.

6.4 Descriptive statistics

Prior to modelling, we look at the distributional properties of the EPEX price series. In Table 4, we see that the informal observation described above, that prices are more stable in the test period than in the train period, is backed up by the data. The prices are slightly lower and the standard deviation of prices smaller in the test period.

While the series shows large excessive kurtosis in the train period, it is more limited in the test period. This suggests that extreme observations have become less common.

Descriptive statistics

	Train	Test	Full
Observations	10224	8688	18912
Mean	38.1	32.7	35.6
Median	36.6	31.6	34.1
SD	19	12.8	16.7
Skewness	-2.3	-0.3	-1.7
Kurtosis	32.1	6.6	29.6
Min	-222	-65	-222
Max	139.7	88	139.7
Jarque-Bera Test	370428.2**	4789.9**	567713.5**

Table 4 – ** denotes significance at a 1% level. * denotes significance at a 5% level. The null hypothesis of the Jarque-Bera test is that the series is normally distributed.

We should have these observations in mind when we compare model performance. A model that is deemed to have superior forecast performance in this test dataset, might only have so because it is good at forecasting during the more stable test period.

We also do unit root tests on the individual hourly series (H1–H24 in Table 5) and the full series (Table 6).

In the hourly series the ADF-test of a unit root is rejected except for the series H17–H20, while the KPSS test rejects the null of stationarity for the series H11–H20. There thus seems to be a paradoxical rejection where both stationarity and a unit root is rejected. For the full series we also observe the paradoxical rejection of both tests.

So why do we observe this, should it trouble us, and should we model differenced series? A probable cause can be observed in the ACF plot in Figure 19. We see strong autocorrelation at many lags, with both daily and weekly cycles. As both our tests are accounting for autocorrelation but only by including lower order lags, the tests might be mis-specified in our setting of seasonality at long lags.

Another possibility would be to deseasonalize the data, and model the deseasonlized series instead. Chaâbane (2014b) does this for the Nordpool area, and report the non-rejection of the KPSS test and rejection of the ADF-test. For reasons of brevity, we not will do this.

I have chosen not to model differenced price series because the fundamental relationships discussed in Section 2 makes it improbable that a 'regression to mean' should not be present. We simply do not expect the very volatile prices to explode indefinitely because the prices are determined by a supply/demand relationship. Therefore we expect the price series to be stationary.

As we do not attempt to find any 'correct' model but just to find a model that forecasts well, all of the abovementioned approaches could be tested empirically by fitting a model on both differenced, deseasonalized, and non-differenced data and then comparing model forecast performance. For the sake of brevity, we have chosen not to do this.

Unit root tests

Hour	ADF statistic	KPSS statistic
1	-5.70**	0.13
2	-5.69**	0.08
3	-5.60**	0.10
4	-5.23**	0.08
5	-5.50**	0.11
6	-5.47**	0.09
7	-5.32**	0.08
8	-4.37**	0.12
9	-3.80*	0.26
10	-4.33**	0.39
11	-4.51**	0.59*
12	-4.65**	0.61*
13	-4.42**	0.94**
14	-4.12**	1.03**
15	-3.80*	0.92**
16	-3.66*	0.95**
17	-3.28	1.14**
18	-2.90	1.23**
19	-3.17	1.13**
20	-3.05	0.62*
21	-3.45*	0.31
22	-3.80*	0.34
23	-4.57**	0.46
24	-5.15**	0.26

Table 5 – Unit root tests of hourly series in train period. The null hypothesis of the Augmented Dickey Fuller Test is that the series contains a unit root with the alternative that it is stationary. Lag order k=7 was used. The null hypothesis of the KPSS test is that the series is level stationary. Lag order k=4 was used.

Hours	ADF statistic	KPSS statistic
All	-6.71**	0.97**

Table 6 – Unit root tests of full series in train period. The null hypothesis of the Augmented Dickey Fuller Test is that the series contains a unit root with the alternative that it is stationary. Lag order k=21 was used. The null hypothesis of the KPSS test is that the series is level stationary. Lag order k=23 was used.

Some authors transform the data prior to modelling, to get a more stable variance, usually by taking $\log(\bullet)$ of the dependent and explanatory variables. There are, however, problems with log transforming data when negative or small numbers are present— as in German/Austrian electricity price series. Another related method of transformation, using the area hyperbolic sine function, is presented in Johnson (1949) and used on electricity prices by Schneider (2011). This transformation seems more promising in the context of electricity prices. It was tested at an early stage, but the results were not encouraging.

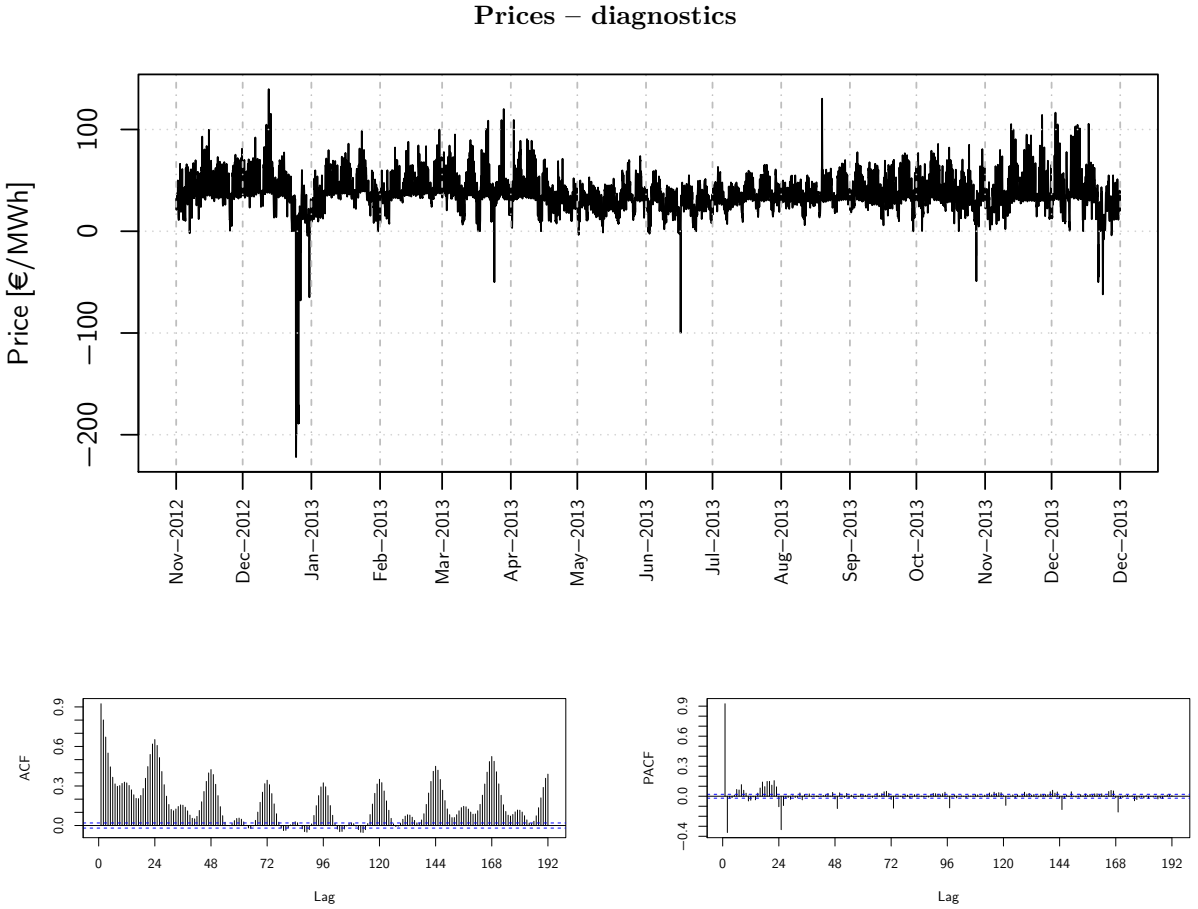

Figure 19 – Plot of price, autocorrelation function and partial autocorrelation function for the prices series in the train period. Dotted lines in the ACF/PACF plots denote significance at a 5% significance level, assuming uncorrelated and asymptotically normally distributed errors.

7 Estimation

All implementations of models were done in R (R Core Team, 2015).

7.1 ARIMA

We first test the 24-model approach to modelling the price series, by using a separate ARIMA model for each of the 24 hourly series. This 24 models ignore any relationships that might exist between the hours. After estimation, each model is used to forecast 1 step (24 hours) ahead and the 24 hourly forecasts are combined into one series of hourly forecasts.

ARIMA Specifications

No	Comment	AR	I	MA	Average AIC
(1)		24	0	0	3315.0
(2)		24	0	24	3314.7
(3)		24,48	0	24,48	3313.8
(4)		24,48...168	0	0	3251.5
(5)		24,48...168	0	24	3245.5
(6)		24,48...168	0	24,48	3226.1
(7)		24,48...168	0	24,48...168	3183.5
(8)	Using auto.Arima for model selection in each series, with 4 as the maximum lag of AR/MA terms. 10 of the selected models where integrated.		Varies		3290.4
(9)	Using auto.Arima for model selection in each series, with 7 as the maximum lag of AR/MA terms. 10 of the selected models where integrated.		Varies		3285.1

Table 7 – Overview of tested ARIMA specifications. The AR/MA columns contain the lags in the respective parts. The AIC holds the average AIC across all 24 hourly models. Specification 8 & 9 has varying model specifications, as selected by the auto.Arima() function of the R package 'forecast'. For all series in specification (9), the selected model had AR/MA lags at 5 or below.

In the ACF and PACF of the prices series (Figure 19), we observe a daily and a weekly pattern. There also seems to be a diurnal pattern around lag 12. To capture this diurnal pattern, we have to employ a model that either captures cross series correlation or treats the data as one series. The latter approach was tested in the 1-model specification for the price series[19], but this approach was found to predict worse than the naïve model and was therefore excluded. A number of different 24-model specifications were tested in Table 7.

In the tested framework, the optimal model, as selected by AIC, is model (7) that has AR and MA lags 24 through 168 for all 24 hourly models. This model come closest to removing autocorrelation in the residuals. Looking at Figure 20, we see that there is still significant residual autocorrelation around lag 24 and especially around lag 168, which is unfortunate. It is a sign that we did not extract all available information from the series.

[19]Having AR terms: 1, 2, 3, 12, 13, 23, 24, 25, 48, 72, 168 and MA terms: 1, 2, 3, 24, 168

The autocorrelation in the short-term is less problematic, as we would expect correlated errors within a day. For example, if the consumption forecast on a day overshots compared to the one used by the market, we could see correlated residuals across the whole day. As the within day error is not known when we generate the forecasts for the next day, this autocorrelation cannot be used to forecast day-ahead prices.

An automatic model selection prodedure was also tested, but the resulting average AIC was worse than that of the best model. It should be noted that 10 of the 24 series were found to contain a unit root by the procedure and thus the corresponding models were for changes. This raises the question of whether we should have differenced the some series prior to modelling. As stated previously, this is ultimately an empirical question and both models could be fitted and tested. For reason of brevity, this was not considered.

Before estimation, the train data was censored at -50 € to limit the effect that the very extreme prices in Christmas 2012 could have on model selection.

When we forecast in the test period, a rolling 365-days estimation window with daily re-estimation is employed.

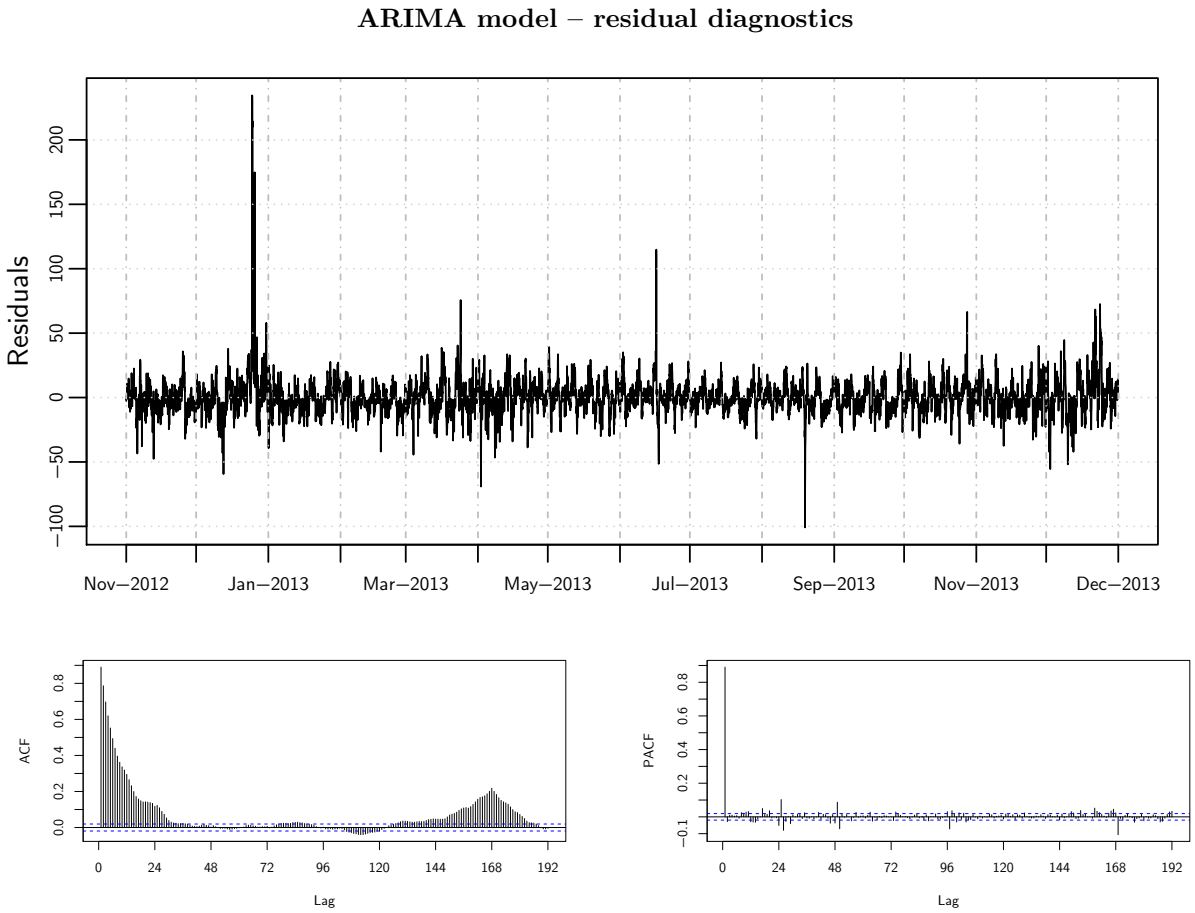

Figure 20 – Plot of residuals, autocorrelation function and partial autocorrelation function for the residuals from the ARIMA model. Dotted lines in the ACF/PACF plots denotes significance at a 5% significance level.

7.2 Linear

Several specifications for the linear model were considered.

Linear Models

No	Model	Formula	df	k	AIC
(1)	Basic	$CON_{DE} + PRO_{DE}WND + PRO_{DE}SPV + CON_{FR}$	6	-	77754
(2)	Basic with dummies in level	$CON_{DE} + PRO_{DE}WND + PRO_{DE}SPV + CON_{FR} + F[Month] + F[Hour] + F[Weekday]$	46	-	75610
(3)	Basic with dummies in level and interaction in exogenous	$(CON_{DE} + PRO_{DE}WND + PRO_{DE}SPV) \circ Weekend \circ Peak \circ Summer + CON_{FR}$ $+F[Month] + F[Hour] + F[Weekday]$	71	-	75330
(4)	Polynomial expansion with dummies in level	$poly[CON_{DE}, k] + poly[PRO_{DE}WND, k] + poly[PRO_{DE}SPV, k] + poly[CON_{FR}, k]$ $+F[Month] + F[Hour] + F[Weekday]$	46 50 54	1 2 3	75610 75312 75271
(5)	Polynomial expansion with dummies in level and interaction in exogenous	$(poly[CON_{DE}, k] + poly[PRO_{DE}WND, k] + poly[PRO_{DE}SPV, k]) \circ Weekend \circ Peak \circ Summer$ $+poly[CON_{FR}, k] + F[Month] + F[Hour] + F[Weekday]$	71 96 121	1 2 3	75330 74944 74595
(6)	Polynomial expansion with dummies in level, using RESCON	$poly[RESCON_{DE}, k] + poly[CON_{FR}, k] + F[Month] + F[Hour] + F[Weekday]$	44 46 48	1 2 3	75765 75460 74950
(7)	Polynomial expansion with dummies in level and interaction in exogenous, using RESCON	$poly[RESCON_{DE}, k] \circ Weekend \circ Peak \circ Summer + poly[CON_{FR}, k]$ $+F[Month] + F[Hour] + F[Weekday]$	55 64 73	1 2 3	75668 74918 74342

Table 8 – Overview of linear model specifications in the 1-model framework. $poly[X, k]$ denotes that an ortogonal polynomial expansion of X of degree k was performed prior to inclusion in the model (using the R-function $poly(\bullet)$). $F[X]$ means that X is a factor with a number of levels that are included as dummies. $X \circ Y$, where Y is a dummy, means that all interaction terms between all elements of X and the dummy Y are included in the regression. All models include a constant term. $RESCON = (CON_{DE} - PRO_{DE}WND - PRO_{DE}SPV)$ is the residual consumption. df is the degrees of freedom used in the model, which equals the number estimated of parameters + 1 (the estimate of the error variance).

First, we estimate the most basic specification in model (1) with 6 free parameters in the 1-model framework (Table 8).

Then, we add a range of dummies in model (2) to account for potential differences in the mean due to cycles and we observe that this model yields a significantly better AIC.

In model (3), we allow for more complex patterns of dependence in the explanatory variables by including interaction terms with the three aggregate dummies. This allows for differences in behaviour of the explanatory variables across day, week, and season, hopefully capturing some of the variations in the relationship that is due to cycles across time. This also improves the AIC.

In model (4), we introduce additional flexibility in the relationship between explanatory variables and the price by allowing for different degrees of polynomial expansions. This yields an improvement in AIC beyond that seen in model (3).

In model (5), we see that the combination of the two previous improvements beats all the previous models, but that this comes at the cost of a steep increase in the number of parameters.

In model (6), we limit the expense of degrees of freedom by putting more structure on the explanatory variables through substituting with $RESCON = (CON_{DE} - PRO_{DE}WND - PRO_{DE}SPV)$[20]. This approach beats the AIC of the unstructured model (4) and moreover lowers the degrees of freedom used in the

[20]RESCON is the amount of demand that has to be covered by generators and net import.

model.

All of the above extensions are incorporated in model (7). With a polynomial expansion of degree 3, this model yields the lowest AIC of 74342 using 73 degrees of freedom and is therefore choosen as the best linear model. A slightly lower AIC can be obtained by increasing k (up to around 9), but the increase in complexity was avoided in favor of a simpler specification, that presumably is more stable and more robust to overfitting.

We also test the 24-model framework where 24 different models are estimated, thus allowing for a different behaviour in each hour. Here we also choose model (7)— and exclude the peak and hour dummies.

To illustrate the effect of continously incorporating new data, we also test two forecasting frameworks. In the fixed framework the model parameters are only estimated once using train period data. In the rolling framework a rolling 365-days estimation window with daily re-estimation is employed. The rolling specification should allow the model to adapt better to changes in the electricity system.

Looking at the residual diagnostics of model (7) from the 1-model framework (Figure 21), we see that there is significant autocorrelation, especially around lag 24, suggesting that the series contains unexploited inter-temporal information.

7.3 ARMAX

As seen in Figure 21, the residuals from the linear model contain significant autocorrelation. To extract this information, we fit an ARMA model to the residuals from the rolling 1-model specification that was selected in Section 7.2. We choose to estimate a separate model for each of the 24 hours in the ARMA residual model.

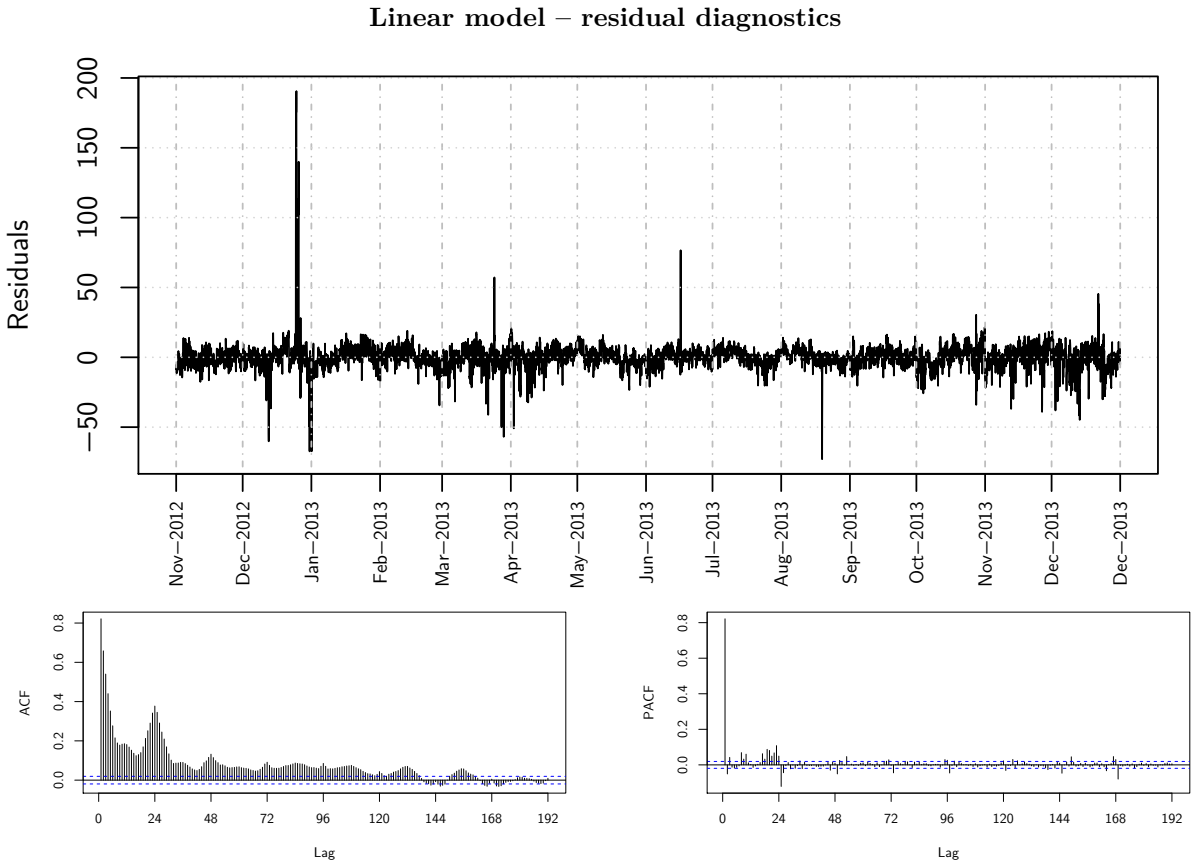

Figure 21 – Plot of residuals, autocorrelation function and partial autocorrelation function for the residuals from the linear model in the train period. Dotted lines in the ACF/PACF plots denotes significance at a 5% significance level.

ARMAX Specifications

No	Comment	AR	I	MA	Average AIC
(1)		24	0	0	2977
(2)	Selected - due to being simpler	24	0	24	2967
(3)		24,48	0	24,48	2968
(4)		24,48...168	0	0	2964
(5)		24,48...168	0	24	2965
(6)		24,48...168	0	24,48	2965
(7)		24,48...168	0	24,48...168	2963
(8)	Using auto.Arima for model selection in each series, with 4 as the maximum lag of AR/MA terms. 5 of the selected models where integrated.		Varies		2962
(9)	Using auto.Arima for model selection in each series, with 7 as the maximum lag of AR/MA terms. 5 of the selected models where integrated.		Varies		2962

Table 9 – Overview of tested ARMAX specifications. The AR/MA columns contain the number of lags in the respective parts. The AIC holds the average AIC across all 24 hourly models. Specification (8) & (9) has varying model specifications, as selected by the auto.Arima() function of the R package 'forecast'. For all series in specification (9), the selected model had AR/MA lags at 5 or below.

In Table 9, we see the tested models. While the optimal model is number (9), the performance of all models are close. We therefore opt for the simpler model (2).

After modelling the step one residuals, we observe that most autocorrelation is now removed from the step two residuals (Figure 22), though in the PACF plot we still observe significant lags at 24 and 168.

When forecasting in the test period, a rolling 365-days estimation window with daily re-estimation is employed.

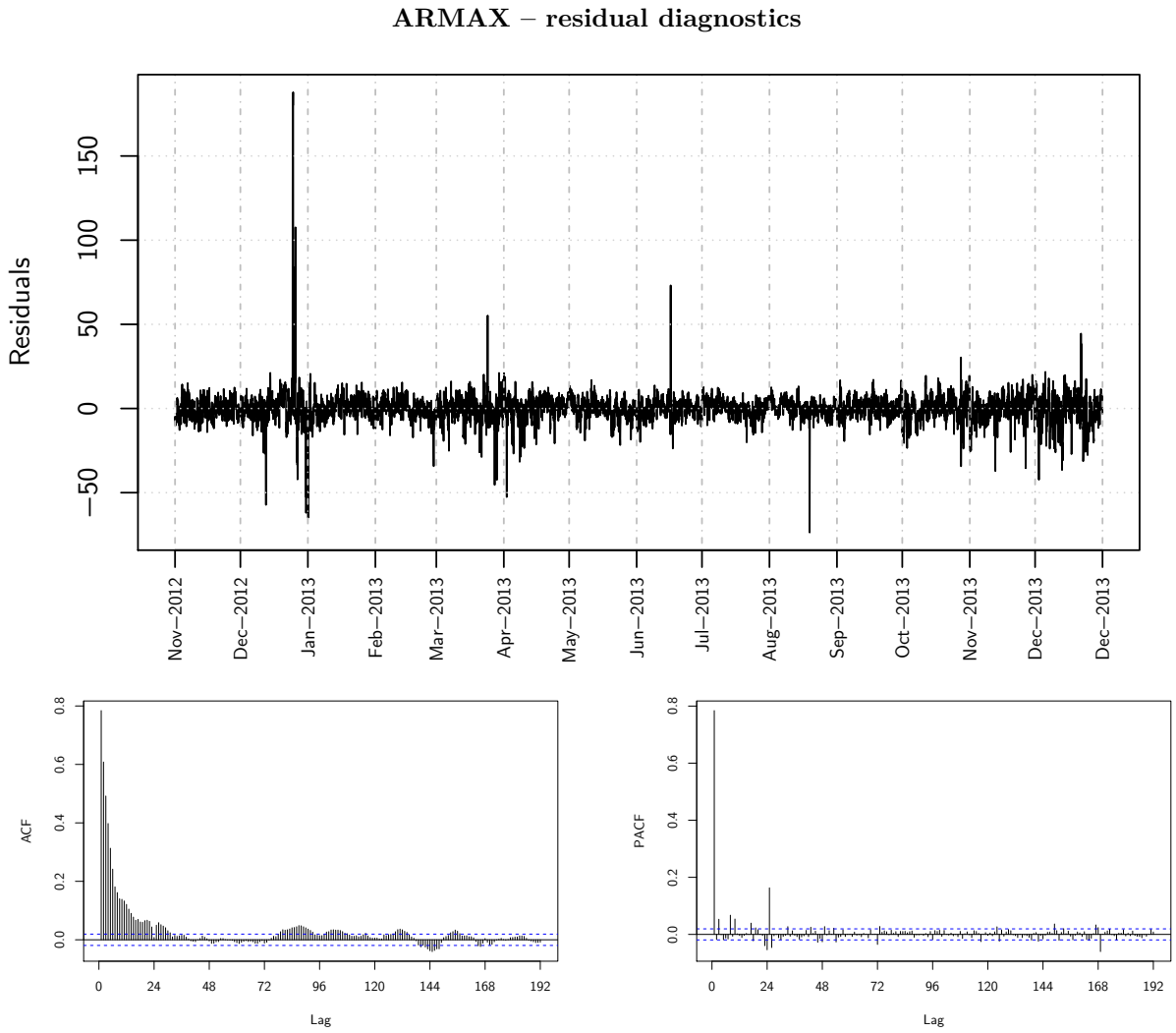

Figure 22 – Plot of residuals, autocorrelation function and partial autocorrelation function for the residuals from the ARMAX model. Dotted lines in the ACF/PACF plots denotes significance at a 5% significance level.

7.4 Jonsson step-one

To operationalize the Jonsson model we have to make several model design choices.

First and foremost, the Jonsson model suffers from a curse of dimensionality, where adding explanatory

variables will lead to a more scarce support in each fitting point, which produces a more variable forecasting surface and have higher computational requirements.
The authors in Jónsson *et al.* (2013) acknowledge the dimensionality curse and comments that, "although theoretically possible and not hard to implement, the inclusion of 1 or 2 more variables [than 2] in the model calls for a [sic] more cautiously chosen variables or merger of them in order to ensure frequent enough updates of the parameters in all fitting points".
We follow this suggestion and limit the choice of explanatory variables to the two variables $Consumption_{forc}$ and $RES_{forc} = (Solar_{forc} + Wind_{forc})$.

Before estimation, the explanatory variables are scaled to lie in the interval $[-1, 1]$**,** based on the ranges of the variables observed in the training period. We also censor the price P to lie in the interval $[0, 110]$ in all estimation equations. Both are also done in Jónsson *et al.* (2013).

Jónsson *et al.* (2013) use the BFGS algorithm to optimize the tune parameters. We initiate the BFGS optimization using the values that were found to be optimal for DK1 in Jónsson *et al.* (2013), but chose to optimize a special parameterization that restricts the tune parameters to sensible values ($\gamma, \lambda \in [0.5, 1]$ *and* $\tau \in [0, \infty]$). The optimization criterion is the RMSE of the day-ahead forecast in the test period— not including an intial 42 day burn period where the model is initialized, and the 24–25 December 2012 that had very extreme prices.

We choose 24 equidistant fitting points for each variable as in Jónsson *et al.* (2013). This leads to a grid with 576 fitting points.

The model is initialized by setting elements of the parameter vector ϕ equal to $(34, 0, 0, 0, 0, 0)$. The covariance matrix, R, is intialized as a diagonal matrix with all nonzero elements equal to 10^{-6}. During an initial burn period, we relax the updating rule of the R matrix to get more frequent updates and we disregard these forecasts when tuning parameters. Both following Jónsson *et al.* (2013).
Initially we also update the R matrix a number of times, before we start to update the parameter vector ϕ (in our case the first 100 observations ~ 4 days). Otherwise we can risk getting large and oscillating parameter estimates.

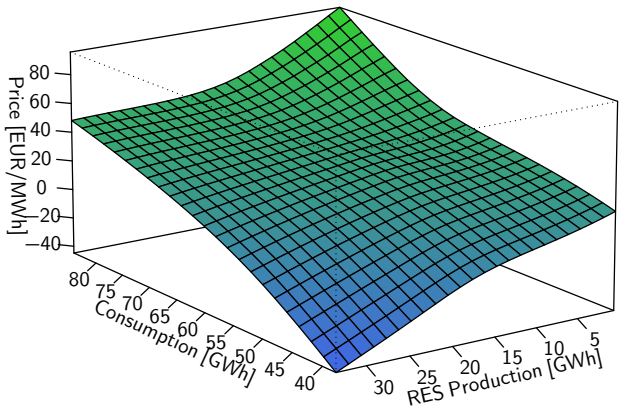

Figure 23 – Forecast surface from the Jonsson step-one model on 15.07.2014.

Figure 23 shows an actual forecast surface estimated by the model. It is evident that the model has captured a non-linear relationship between the price and the explanatory variables, where the model predicts high (low) prices when the combination of low (high) renewable energy source production and high (low) consumption occurs.

As the model is dynamically updating, the forecast surface can change in response to new data points. This gives it the ability to accommodate slow changes in the underlying relationships, such as gradual changes in the available generation capacity or changes in the bidding behaviour of market participants (for example, due to changes in fuel costs).

In Figure 24, we see that significant autocorrelation is present in the residual of the Jonsson step-one model. This issue will be adressed in the next section.

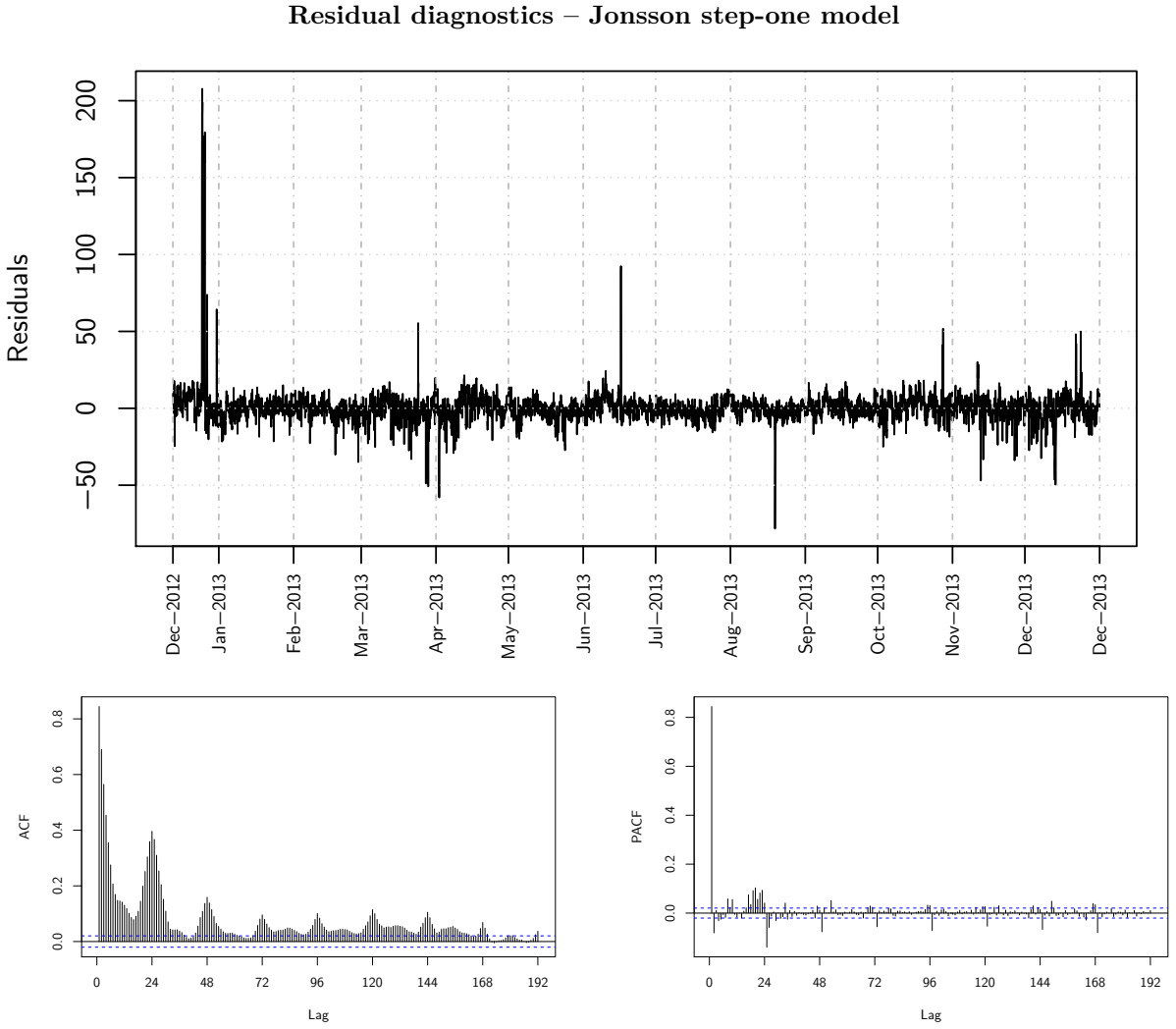

Figure 24 – Plot of residuals, autocorrelation function and partial autocorrelation function for the residuals from the Jonsson step-one model. Dotted lines in the ACF/PACF plots denotes significance at a 5% significance level.

7.5 Jonsson step-two

Observing the significant autocorrelation in the residuals from the Jonsson step-one model (Figure 24), we fit an ARIMA model to the residuals. We follow the procedure used with the ARMAX model, and estimates a separate model for each of the 24 hours.

No	Comment	AR	I	MA	Average AIC
(1)		24	0	0	2738
(2)		0	0	24	2740
(3)	Selected - due to being simpler.	24	0	24	2729
(4)		24,48	0	24,48	2730
(5)		24,48...168	0	0	2732
(6)		24,48...168	0	24	2733
(7)		24,48...168	0	24, 48	2731
(8)		24,48...168	0	24,48...168	2733
(9)	Using auto.Arima for model selection in each series, with 7 as the maximum lag of AR/MA terms. 4 of the selected models where integrated.		Varies		2727
(10)	Using auto.Arima for model selection in each series, with 4 as the maximum lag of AR/MA terms. 4 of the selected models where integrated.		Varies		2727

Table 10 – Overview of Jonsson step-two model specifications. The AR/MA columns contains the number of lags in the respective parts. The AIC holds the average AIC across all 24 hourly models. Specification 9 & 10 has varying model specifications, as selected by the auto.Arima() function of the R package 'forecast'. For all series in specification 9, the selected model had AR/MA lags at 5 or below.

In Table 10, we see that the model with automatic model selection has the best AIC. But, as the difference to the simpler ARMA model with AR lag 24 and MA lag 24 is minimal we choose this specification for the residual model.

Observing the residuals from the Jonsson step-two model, we see that a significant part of the autocorrelation has been removed. However, the errors still contain some autocorrelation around lag 24 and a significant PACF term.

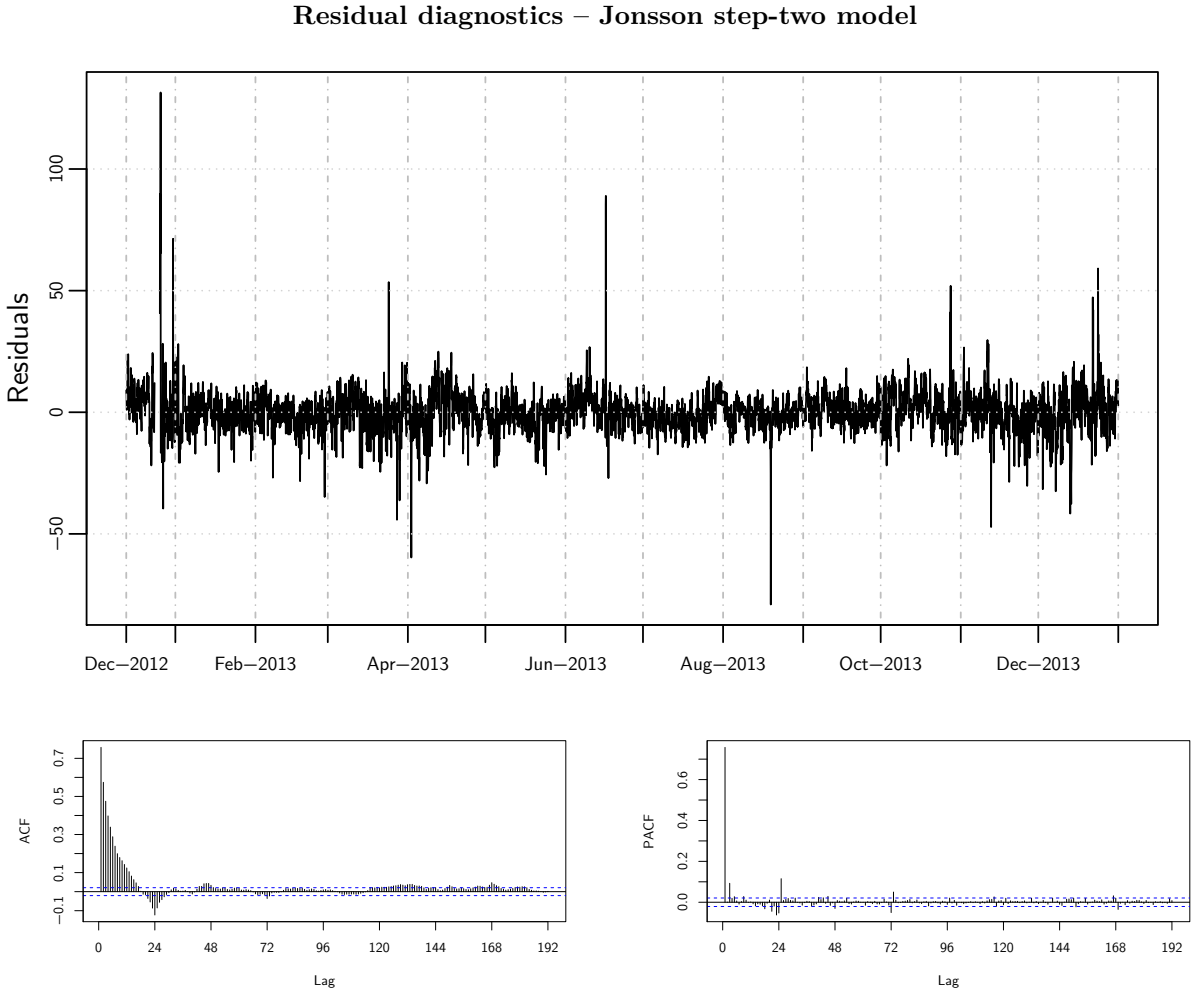

Figure 25 – Plot of residuals, autocorrelation function and partial autocorrelation function for the residuals from the Jonsson step-two model. Dotted lines in the ACF/PACF plots denotes significance at a 5% significance level.

7.6 Neural network

To implement the neural network method, we choose the model averaged neural network (avNNet) method from the R-package 'caret' (Kuhn, 2008). This method estimates a number of similar, in structure, single-layer feed-forward neural networks, that are optimized from different starting values, using the R-package 'nnet' (Venables and Ripley, 2002) and averages their predictions to obtain a final prediction. We have chosen to average the predictions from five neural networks to keep computational costs down.

When optimizing the model, we can tweak both the number of neurons in the hidden layer, a weight decay parameter, and the number of optimizing iterations.

The weight decay parameter affects the size of weight adjustments the optimization algorithm make, such that it gradually decrease during optimization. A grid search with the simplest model (1) (see Figure 37 in the appendix) showed only minor changes in performance with different values. Therefore, the weight decay parameter was fixed at 0.1.

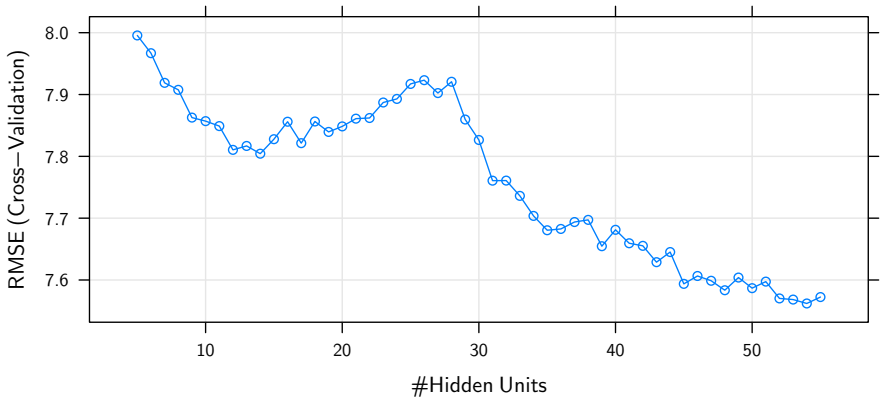

Figure 26 – Cross-validated RMSE for a different numbers of hidden units in model (1).

The maximum number of iterations was kept at 100, which is the standard of the nnet package, during selection of the best model. After the best model was selected, it was re-run with the maximum number of iterations set to 200. This yielded a slightly lower cross-validated RMSE (average of 5.117 vs. 5.254). Due to the computational costs involved, the specification with maximum number of iterations 100 was kept. This choice might result in a slight loss in performance and leaves room for further optimization of model weights.

The caret package allows for the use of bootstrap aggregation (Bagging) when constructing the model averaged neural network. Bagging works by selecting a new bootstrap sample from the data (with replacement) for each neural network. This procedure is reported to improve accuracy for unstable estimators, which include neural networks (Breiman, 1996). We therefore use bagging when estimating the neural networks.

As observed in Figure 26, the number of neurons in the hidden layer affects the performance of model (1) significantly. Observing the leveling out in RMSE around 35+ neurons, the number of neurons was fixed at 37 for all the specifications. This allows for flexibility in the neural network while still being computationally manageable.

Having selected the neural network structure, a number of different input specifications were tested in Table 11.
We see that, compared to the simple model (1), the inclusion of aggregate dummies in model (2) improves the cross-validated RMSE slightly. The inclusion of all dummies in model (3) has an even larger effect and including interaction terms in model (4) lowers RMSE slightly more, but at the cost of 60% more weights. Model (5), which introduces RESCON, shows a clear worsening in RMSE contrary to what was observed with the linear model. While in model (6), we observe that the removal of French consumption raises RMSE significantly compared to model (4) that contains this explanatory variable.
Finally in model (7), we include the price lagged by 24 hours. This model has the lowest RMSE and was therefore selected. Model (7) did not show marked changes in cross-validated RMSE when the number of neurons was varied, so the specification with 37 neurons was kept.

When forecasting in the test period, a rolling 365-days estimation window with weekly re-estimation is

Neural Network Input Specification

No	Model	Formula	#Weights	RMSE
(1)	Basic	$CON_{DE} + PRO_{DE}WND + PRO_{DE}SPV + CON_{FR}$	223	7.676
(2)	Basic with aggregate dummies	$CON_{DE} + PRO_{DE}WND + PRO_{DE}SPV + CON_{FR} + Weekend + Peak + Summer$	334	7.290
(3)	Basic with all dummies	$CON_{DE} + PRO_{DE}WND + PRO_{DE}SPV + CON_{FR} + F[Month] + F[Hour] + F[Weekday]$	1703	5.443
(4)	Basic with all dummies and interaction terms	$(CON_{DE} + PRO_{DE}WND + PRO_{DE}SPV) \circ Weekend \circ Peak \circ Summer + CON_{FR}$ $+F[Month] + F[Hour] + F[Weekday]$	2739	5.318
(5)	RESCON with all dummies and interaction terms	$RESCON_{DE}, \circ Weekend \circ Peak \circ Summer + CON_{FR}$ $+F[Month] + F[Hour] + F[Weekday]$	2147	5.972
(6)	Basic with all dummies and interaction terms minus CON_{FR}	$(CON_{DE} + PRO_{DE}WND + PRO_{DE}SPV) \circ Weekend \circ Peak \circ Summer$ $+F[Month] + F[Hour] + F[Weekday]$	2702	6.066
(7)	Basic with all dummies and interaction terms and lagged price	$(CON_{DE} + PRO_{DE}WND + PRO_{DE}SPV) \circ Weekend \circ Peak \circ Summer + CON_{FR}$ $+F[Month] + F[Hour] + F[Weekday] + PRI_{t-24}$	2776	5.254

Table 11 – Overview of specifications with different explanatory variables. $F[X]$ means that X is a factor with a number of levels that are included as dummies. $X \circ Y$, where Y is a dummy, means that all interaction terms between all elements of X and the dummy Y are included in the specification. $RESCON = (CON_{DE} - PRO_{DE}WND - PRO_{DE}SPV)$ is the residual consumption. The weights column shows the number of connections in the corresponding specification. RMSE shows the root mean square error of 7-fold cross-validation. Table 33 in the appendix shows more detailed cross-validation results.

employed.

7.7 Support vector machine

We use the R package Kernlab (Karatzoglou *et al.*, 2004) to implement the support vector machine framework, through the excellent wrapper package 'caret' (Kuhn, 2008).

When estimating a support vector machine (SVM), we are faced with a number of modelling choices. Firstly, we have to choose which kernel to use in the SVM. As stated in the Section 5.2.7, we restrict tests to the linear kernel $k(x, y) = x^T \cdot y$ and the Radial Basis Function (RBF) kernel $k(x, y) = \exp\left(-\sigma \|x - y\|^2\right)$. Note, that there are a number of other kernels that could be tested, but for reasons of brevity this is left to future research.

Secondly, we need to choose which explanatory variables to include, and how. Here we try many of the specifications used in the linear model framework.

Lastly, both the linear SVM and the radial basis function (RBF) SVM have a cost parameter that we can optimize across. We will use cross-validation to select this parameter. The RBF SVM has an additional sigma parameter, but we use a Kernlab function to automatically select a sensible value for this based on sample quantiles (more information on this can be found in the Kernlab documentation for the functions $ksvm()$ and $sigest()$).

Support Vector Machine Model Specifications

No	Model	Formula	Linear SVM	Radial SVM
(1)	Basic	$CON_{DE} + PRO_{DE}WND + PRO_{DE}SPV + CON_{FR}$	5.726	7.625
(2)	Basic with aggregate dummies	$CON_{DE} + PRO_{DE}WND + PRO_{DE}SPV + CON_{FR} + Weekend + Peak + Summer$	5.490	7.707
(3)	Basic with all dummies	$CON_{DE} + PRO_{DE}WND + PRO_{DE}SPV + CON_{FR} + F[Month] + F[Hour] + F[Weekday]$	4.984	6.205
(4)	Basic with all dummies and interaction terms	$(CON_{DE} + PRO_{DE}WND + PRO_{DE}SPV) \circ Weekend \circ Peak \circ Summer + CON_{FR}$ $+F[Month] + F[Hour] + F[Weekday]$	4.810	6.005
(5)	RESCON with all dummies and interaction terms	$RESCON_{DE,} \circ Weekend \circ Peak \circ Summer + CON_{FR}$ $+F[Month] + F[Hour] + F[Weekday]$	5.010	6.702
(6)	Basic with all dummies and interaction terms minus CON_{FR}	$(CON_{DE} + PRO_{DE}WND + PRO_{DE}SPV) \circ Weekend \circ Peak \circ Summer$ $+F[Month] + F[Hour] + F[Weekday]$	4.988	6.678
(7)	Basic with all dummies and interaction terms and lagged price	$(CON_{DE} + PRO_{DE}WND + PRO_{DE}SPV) \circ Weekend \circ Peak \circ Summer + CON_{FR}$ $+F[Month] + F[Hour] + F[Weekday] + PRI_{t-24}$	4.803	5.774

Table 12 – Overview of input specifications for the support vector machines. $F[X]$ means that X is a factor with a number of levels that are included as dummies. $X \circ Y$, where Y is a dummy, means that all interaction terms between all elements of X and the dummy Y are included in the specification. $RESCON = (CON_{DE} - PRO_{DE}WND - PRO_{DE}SPV)$ is the residual consumption. The columns "Linear SVM" and "Radial SVM" holds the average RMSE of 7-fold cross validation. More detailed distributional results can be found in Table 31 and Table 32 in the appendix.

First we tune the simplest of the models, model (1), across different values of the cost parameter. The resulting cross-validated RMSE for the linear SVM show almost no sensitivity to the cost parameter (Table 29 in appendix), so we fix the cost parameter at 1 for the linear SVM.
The RBF SVM shows more sensitivity to the cost paramter (Table 30 in appendix), but the gain in performance when selecting a higher value of the cost parameter is modest, and comes at the cost of significantly higher computational time. We therefore fix the cost parameter at 4 for the RBF SVM.
We could check the cost parameter assumption for all model specifications but, with the low sensitivity encountered, we assume that the selected parameter value is acceptable in the more complex model specifications as well.

Having established that the cost parameter is of minor importance, we move on to testing different input variable specifications.
The tested specifications are the same across both types of SVM and can be seen in Table 12. Prior to inclusion, the input variables were centered and scaled to unit variance, as recommended by Hsu *et al.* (2008).

For the linear SVM, we see gains when moving from the simple model (1) to model (2) and (3) that both contain more dummies. Contrary to the linear specification, we find only modest gains by including interaction terms in model (4) and we get slightly worse results by substituting with RESCON in model (5). This could indicate that the model has better ability to handle inputs, and therefore needs less a priori structure than the linear methods.
Comparing model (4) with model (6), we see that the inclusion of French consumption adds useful information to the model. Model (7) that contains the price lagged 24 hours is only marginally superior to model (4), which lack this term. However, we end up selecting model (7) for the linear SVM anyway, because it is clearly selected by the RBF SVM. Otherwise the simpler model (4) would be preferred.

For the RBF SVM, we see slightly worse performance when adding the aggregate dummies in model (2). However, we see significant gains by including all dummies as in model (3), and we also see gains by including interaction terms in model (4). Comparing model (4) with model (6), we observe that including the French consumption improves the performance more than for the linear SVM. Substituting with RESCON in model (5), we again get significantly worse results. For the RBF SVM, model (7) that includes the price lagged 24 hours has the lowest RMSE. We therefore end up selecting model (7).

The train sample cross-validated RMSE results clearly seem to favor the linear SVM. This could indicate that the larger feature space used by the more complex RBF SVM gives a worse generalization capability than the simpler feature space of the linear SVM. As we have selected the same model (7), this question will be answered more conclusively in Section 8, when we compare the true out-of-sample forecasting performance of the two kernels.

When forecasting in the test period, a rolling 365-days estimation window with weekly re-estimation is employed.

8 Empirical Results

In this section, we will present the out-of-sample forecast performance of all the models selected in the previous sections and compare the performance differences using the Diebold-Mariano test. We will also create aggregate forecasts that are combinations of the invididual forecasts and present their out-of-sample forecast performance. Finally, we discuss some of the observations we have made.

8.1 Individual forecasts

Having settled on a specific model for all the methods in Section 7, we simulate the daily forecasting procedure to obtain and evaluate model forecasts in the test period 01.01.2014 to 28.12.2014.

Starting with Table 15 & 16 and Figure 27, we observe that the market consensus benchmark (MCB) has superior performance compared to the other model in most weeks. Measured by RMSE, it is clearly superior across the full test period with an average weekly RMSE of 4.02.
The naïve approach, which traditionally serves as the benchmark, has a weekly RMSE of 9.38 and a relative RMSE (RRMSE) of 2.33— i.e. the weekly error of the naïve model is on average 133% higher than the MCB error.

The other commonly used reference model, ARIMA, has a RRMSE of 1.93. From the Diebold-Mariano (DM) test in Table 13, we see that this beats the naïve model but is not statistically different from the simple non-rolling linear model, which has a RRMSE of 1.84.

Comparing with the other linear methods, we observe a significant improvement when using the rolling method (linear.r, RRMSE 1.36). A probable explanation is that, as old unrepresentative data is dropped and recent data is included the estimation data becomes more representative of the future data and therefore leads to a better forecast performance.
Comparing the 1-model framework (linear, linear.r) with the 24-model framework (linear.s, linear.sr), we see that the 1-model framework shows clearly superior performance. This shows that the quantity of data lost by dropping all the other hours in the 24-model framework, is not offset by getting higher quality data.
The best linear model (linear.r) uses a rolling estimation window and has a RRMSE of 1.36 and RMSE of 5.48. The DM test in Table 13 shows that it is a statistically significant improvement compared to all the other linear models.

The ARMAX model that is closely related to the linear.r model has a RRMSE of 1.32. This is slightly better than the linear.r model, though the DM test in Table 13 shows that this is not a statistically significant difference.

Focusing on the support vector machine algorithm, we see a performance on par with linear.r and ARMAX model. SVMLinear has a slightly lower RRMSE of 1.31 against 1.38 for SVMRadial, though the difference is not statistically significant. So the question of whether the added complexity of the RBF SVM is a benefit or a disadvantage, cannot be conclusively answered.

For the neural network we observe a somewhat higher RRMSE of 1.45. This is significantly worse than

both SVMs but, interestingly, not significantly worse than the ARMAX or linear.r model! A possible explanation of why the DM test cannot reject equal performance with these models, is that the autocorrelation or variance can differ among the loss differential series and thereby give the DM test different statistical power.

Finally, we look at the Jonsson models. The Jonsson step-one model does a very decent job reaching a RRMSE of 1.21. This comfortably beats all previous models, though the difference to the ARMAX model is again, surprisingly, not statistically significant.

The Jonsson step-two model slightly improves the RRMSE to 1.17. This is significantly better than the ARMAX model at a 5% significance level. Interestingly, the Jonsson step-two model is also statistically better than the Jonsson step-one model at a 1% significance level. The explanation of why the minor difference in RMSE is statistically significant could be that the Jonsson step-one and step-two forecasts are highly correlated, due to their shared step-one part. Therefore, their loss differentials will have a low variance and thus a higher statistical power to detect minor differences in performance.

In Table 17 & 18 and Figure 28, we look at the RMSE across weekdays. Here the MCB shows a clearly worse RMSE on Mondays and Sundays, than on the other days of the week. This pattern was expected, as the MCB is set on Fridays for the next three days. The more imprecise information used on Sunday and Monday should lead to larger forecast errors— as we observe.
One surprise is that the other models **also** do worse on Sundays and Mondays. Maybe these days are inherently more difficult to forecast[21] or there could be quality issues with the forecasts issued by the TSO's during the weekend[22].
Even given the informational advantage on Sunday and Monday, the other methods are still not able to beat the MCB, except for ARMAX which is barely superior on Sundays. For all models, the performance is significantly closer though as can seen by the lower RRMSE on these days.

In Table 19, we see the performance on weekends and holidays. As expected MCB is best across weekend, weekday, and non-holiday, but it performs clearly worse on weekends and holidays. The Jonsson step-one model has the best holiday performance (though given that the train period only had 9 holidays, this difference is likely not statistically significant) and shows almost no difference to non-holidays.

Last, we look at the performance across hours of the day in Table 20 & 21 and Figure 29. In general, most models show best performance in the first 3 hours of the day and the last 4. Around hour 4, 14-16 and 19-20 most models have the highest RMSE. The models show similar performance patterns, but the ARMAX model seems to track the others to a lesser extent. This can probably explain why the DM test has more difficulties rejecting equal predictive ability when the ARMAX model is involved.

[21]For example: A) if some market participants rely on the price-information in the MCB forecast (EXAA auction) on weekdays to guide their EPEX auction strategy, but do not use it in the weekend when the information is stale and therefore bid more dispersed. B) if the market clearing point typically is lower/higher on the merit order curve, where small changes in fundamentals give higher changes in price— though this seems unable to explain the Monday performance.

[22]For example if the data received by TSO's from forecasters receive less attention/manual correction during the weekend, this could lead to more imprecise forecasts.

	mcb	naive	arima	linear	linear.r	linear.s	linear.sr	arimax	jonsson.s1	jonsson.s2	svmlinear	svmradial	avnnet
mcb		-9.11**	-11.39**	-7.47**	-6.06**	-6.38**	-7.03**	-4.93**	-6.77**	-5.31**	-7.32**	-6.32**	-6.72**
naive	9.11**		4.03**	3.43**	6.84**	1.28	6.40**	7.35**	8.04**	8.25**	7.37**	6.96**	6.58**
arima	11.39**	-4.03**		0.87	5.95**	-1.53	5.57**	7.02**	9.63**	10.03**	7.87**	7.06**	5.71**
linear	7.47**	-3.43**	-0.87		5.36**	-2.94**	3.92**	5.75**	5.77**	6.10**	5.31**	4.49**	4.06**
linear.r	6.06**	-6.84**	-5.95**	-5.36**		-4.93**	-2.77**	0.83	2.58**	3.32**	1.18	-0.27	-1.64
linear.s	6.38**	-1.28	1.53	2.94**	4.93**		5.01**	5.34**	5.62**	5.80**	5.36**	4.82**	4.47**
linear.sr	7.03**	-6.40**	-5.57**	-3.92**	2.77**	-5.01**		2.51*	4.70**	5.37**	4.27**	2.04*	0.84
arimax	4.93**	-7.35**	-7.02**	-5.75**	-0.83	-5.34**	-2.51*		1.70	2.40*	0.11	-0.83	-1.79
jonsson.s1	6.77**	-8.04**	-9.63**	-5.77**	-2.58**	-5.62**	-4.70**	-1.70		6.81**	-2.84**	-3.34**	-3.95**
jonsson.s2	5.31**	-8.25**	-10.03**	-6.10**	-3.32**	-5.80**	-5.37**	-2.40*	-6.81**		-4.01**	-4.16**	-4.65**
svmlinear	7.32**	-7.37**	-7.87**	-5.31**	-1.18	-5.36**	-4.27**	-0.11	2.84**	4.01**		-1.85	-3.19**
svmradial	6.32**	-6.96**	-7.06**	-4.49**	0.27	-4.82**	-2.04*	0.83	3.34**	4.16**	1.85		-2.41*
avnnet	6.72**	-6.58**	-5.71**	-4.06**	1.64	-4.47**	-0.84	1.79	3.95**	4.65**	3.19**	2.41*	

Table 13 – Two-way Diebold-Mariano Tests. A significant positive (negative) test statistic indicates that the top (left) model has superior forecast performance. ** denotes significance at a 1% level. * denotes significance at a 5% level.

(A)

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	3.10	1.52	2.05	0.04*
Weekend	0.62	2.69	0.23	0.82
Peak	4.75	2.21	2.15	0.03*
Summer	0.55	2.66	0.21	0.84
Cons	0.03	0.20	0.15	0.88
Wind	0.08	0.57	0.15	0.88
Solar	-0.04	0.30	-0.14	0.89

(B)

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	3.44	1.22	2.82	0.00**
Peak	4.80	2.04	2.36	0.02*

Table 14 – Conditional Diebold-Mariano test of the Jonsson step-two model against MCB. In (A) all considered regressors are included. In (B) insignificant regressors have been successively eliminated. ** denotes significance at a 1% level. * denotes significance at a 5% level.

Having found that the Jonsson step-two model performs best among all single models, we now look closer at how its forecasts differ from the MCB.

In Table 14, we apply the conditional DM test from Section 14. This shows, that the model performance differential is significantly affected by the peak variable but not by the other explanatory variables. So, the Jonsson step-two model performs significantly worse than the MCB on average, and even more so during peak hours.

The reason why we observe a significant worsening of relative performance during peak might be that we are omitting some variables relevant for forecasting peak hours (or that the TSO forecasts are of worse quality than the forecasts used by the market), that the structure of the model considered is not sufficient to accurately forecast these hours, or that the series is just more volatile and harder to forecast around peak.

Week	mcb	naive	arima	linear	linear.r	linear.s	linear.sr	arimax	jonsson.s1	jonsson.s2	svmlinear	svmradial	avnnet
1	4.33	11.77	11.97	8.38	7.98	8.39	8.06	6.75	6.22	6.24	6.60	8.17	6.67
2	5.29	11.28	11.36	6.23	5.65	5.20	5.57	5.79	5.86	5.99	5.63	5.37	5.52
3	4.31	8.84	8.03	6.31	6.31	6.71	6.05	5.72	6.12	5.81	5.08	4.61	5.33
4	5.06	7.82	9.07	4.67	4.58	5.80	4.67	4.99	6.22	5.84	4.92	5.29	5.70
5	4.26	8.65	6.39	4.41	4.33	5.74	4.67	4.41	4.97	4.90	4.32	4.04	4.38
6	4.70	8.10	8.37	5.42	5.76	6.12	6.13	5.50	5.17	5.26	6.86	5.54	6.16
7	5.25	10.70	8.56	4.94	5.24	5.48	5.37	5.09	5.17	4.97	5.39	5.05	4.72
8	4.32	9.94	7.47	5.16	5.69	5.50	4.93	5.28	4.87	4.54	5.06	4.89	5.15
9	3.76	6.78	6.38	4.79	4.95	5.33	4.72	4.77	4.76	4.42	4.51	5.10	4.79
10	4.55	9.68	7.13	6.01	5.56	6.15	5.80	5.22	4.85	4.40	5.81	5.55	6.26
11	8.97	13.56	13.67	6.72	5.68	8.17	7.16	5.88	8.04	7.82	7.85	7.67	6.89
12	4.18	18.45	9.52	5.59	4.03	6.66	3.78	3.99	4.71	4.46	6.22	4.41	4.61
13	3.44	6.84	6.39	8.20	7.20	8.61	6.69	6.54	5.76	5.16	6.74	5.27	5.15
14	3.76	5.42	5.12	4.86	4.76	6.60	5.22	5.08	4.24	3.99	4.00	5.12	5.46
15	3.49	9.46	9.51	4.84	4.27	5.80	4.18	4.24	4.74	4.31	4.53	4.02	4.74
16	4.32	9.11	6.97	7.34	5.60	8.49	5.40	4.94	4.82	4.83	5.29	4.58	4.98
17	3.41	9.05	6.67	5.17	3.57	7.23	3.39	3.45	3.20	3.13	3.17	4.02	3.67
18	3.09	9.40	8.26	4.38	4.37	5.23	4.47	4.23	5.22	4.96	4.12	3.85	4.53
19	7.79	12.41	12.71	5.97	5.75	9.92	9.29	5.53	9.55	9.40	9.30	11.42	10.29
20	2.83	16.31	7.60	4.21	3.57	4.56	3.36	3.49	3.69	4.16	3.82	4.20	4.27
21	2.85	6.51	5.49	5.69	3.82	6.20	3.41	3.02	4.08	3.48	3.33	4.26	5.63
22	3.72	6.60	5.61	9.14	6.95	9.69	6.99	4.78	4.79	4.63	5.20	4.80	5.01
23	3.52	4.82	4.23	6.02	5.72	6.36	5.79	4.57	3.95	3.85	4.11	4.69	4.73
24	3.87	6.91	6.37	6.91	4.78	7.64	6.07	3.77	3.81	3.58	4.30	3.37	4.35
25	2.50	6.45	4.66	7.68	4.77	13.00	7.09	3.83	3.68	3.51	4.97	4.90	4.85
26	2.89	7.33	4.96	6.52	4.16	7.38	4.69	3.54	3.34	3.21	3.34	3.13	3.16
27	2.19	4.05	4.40	4.63	4.11	4.16	3.72	4.64	2.51	2.46	3.11	2.94	3.28
28	2.25	3.40	4.10	7.48	6.90	8.06	7.99	5.65	3.38	3.27	5.15	5.34	5.51
29	2.17	2.84	4.28	5.41	4.32	5.38	3.99	4.02	2.61	2.50	3.90	4.39	5.75
30	2.01	4.02	4.41	5.20	3.81	4.83	3.72	3.61	2.52	2.45	2.70	2.59	2.86
31	4.22	6.20	4.99	9.16	8.38	9.66	8.57	5.71	4.65	4.19	7.57	6.74	7.13
32	2.77	4.93	5.42	9.64	8.46	10.38	9.08	4.27	2.71	2.77	7.81	8.87	10.83
33	6.73	12.21	11.95	8.76	7.69	19.03	11.81	7.36	8.29	8.12	9.19	8.63	7.93
34	3.16	12.65	5.41	6.37	4.59	9.26	5.14	5.83	3.68	3.64	3.31	4.55	3.88
35	2.10	5.09	4.29	9.80	3.13	11.18	4.59	3.41	3.50	3.30	3.03	4.77	6.11
36	2.63	4.45	4.72	10.27	8.81	10.40	9.50	4.21	3.50	3.28	8.99	10.40	11.08
37	2.23	3.22	3.39	9.40	5.57	9.44	5.82	3.58	2.97	2.51	5.59	5.47	4.87
38	3.50	5.97	5.62	9.10	4.83	9.17	4.74	6.16	4.77	4.45	4.02	3.47	3.67
39	3.22	8.28	6.18	9.91	4.06	9.83	4.05	5.14	3.84	3.71	3.37	3.33	3.84
40	3.22	8.25	7.92	6.61	4.25	6.64	4.08	6.97	4.64	4.53	4.26	5.96	6.50
41	3.93	7.52	7.10	6.45	4.48	5.83	4.87	5.13	4.69	4.44	4.27	4.43	4.36
42	2.98	7.65	6.97	8.48	5.04	8.02	4.81	3.91	5.15	4.51	4.71	6.16	6.38
43	3.85	12.87	8.28	10.41	4.78	9.99	4.85	5.77	3.94	3.67	4.28	4.10	4.45
44	4.55	9.67	8.28	10.78	6.62	10.74	6.99	5.68	5.73	5.31	6.17	5.51	5.59
45	4.07	11.58	8.59	5.91	5.45	5.84	5.12	7.09	5.62	5.62	5.01	4.20	4.07
46	2.42	6.98	5.20	4.48	3.51	4.83	3.74	4.26	5.25	5.17	2.85	4.55	5.06
47	2.37	5.76	6.03	4.65	4.32	4.60	4.27	5.05	4.50	4.63	3.21	4.07	4.94
48	3.11	7.81	6.80	4.79	4.22	5.01	3.94	4.72	3.90	3.98	4.58	5.28	4.83
49	2.88	8.71	8.74	3.88	4.95	4.51	6.20	4.75	3.85	3.84	4.70	6.58	7.61
50	3.46	11.62	10.62	5.75	4.15	6.14	5.10	4.34	4.56	4.59	4.26	6.78	6.75
51	4.10	16.28	11.10	8.10	8.33	9.31	9.39	7.07	4.74	4.60	5.25	6.19	8.23
52	6.24	17.07	15.00	19.99	6.49	23.81	7.83	12.50	6.12	5.78	5.91	6.12	6.56
	4.02	9.38	7.78	7.42	5.48	8.57	5.99	5.30	4.87	4.69	5.27	5.55	5.82

Table 15 – Root Mean Square Errors (wRMSE) by week.

Week	mcb	naive	arima	linear	linear.r	linear.s	linear.sr	arimax	jonsson.s1	jonsson.s2	svmlinear	svmradial	avnnet
1	1.00	2.72	2.76	1.93	1.84	1.94	1.86	1.56	1.43	1.44	1.52	1.89	1.54
2	1.00	2.13	2.15	1.18	1.07	0.98	1.05	1.09	1.11	1.13	1.06	1.02	1.04
3	1.00	2.05	1.87	1.46	1.47	1.56	1.40	1.33	1.42	1.35	1.18	1.07	1.24
4	1.00	1.55	1.79	0.92	0.91	1.15	0.92	0.99	1.23	1.15	0.97	1.05	1.13
5	1.00	2.03	1.50	1.03	1.02	1.35	1.10	1.04	1.17	1.15	1.02	0.95	1.03
6	1.00	1.72	1.78	1.15	1.22	1.30	1.30	1.17	1.10	1.12	1.46	1.18	1.31
7	1.00	2.04	1.63	0.94	1.00	1.04	1.02	0.97	0.98	0.95	1.03	0.96	0.90
8	1.00	2.30	1.73	1.19	1.32	1.27	1.14	1.22	1.13	1.05	1.17	1.13	1.19
9	1.00	1.80	1.70	1.28	1.32	1.42	1.26	1.27	1.27	1.18	1.20	1.36	1.28
10	1.00	2.13	1.57	1.32	1.22	1.35	1.27	1.15	1.07	0.97	1.28	1.22	1.38
11	1.00	1.51	1.52	0.75	0.63	0.91	0.80	0.66	0.90	0.87	0.88	0.86	0.77
12	1.00	4.41	2.28	1.34	0.96	1.59	0.90	0.95	1.13	1.07	1.49	1.05	1.10
13	1.00	1.99	1.86	2.39	2.09	2.50	1.94	1.90	1.67	1.50	1.96	1.53	1.50
14	1.00	1.44	1.36	1.29	1.27	1.76	1.39	1.35	1.13	1.06	1.06	1.36	1.45
15	1.00	2.71	2.72	1.38	1.22	1.66	1.20	1.21	1.36	1.23	1.30	1.15	1.36
16	1.00	2.11	1.61	1.70	1.30	1.97	1.25	1.14	1.12	1.12	1.22	1.06	1.15
17	1.00	2.65	1.96	1.52	1.05	2.12	1.00	1.01	0.94	0.92	0.93	1.18	1.07
18	1.00	3.04	2.67	1.42	1.42	1.69	1.45	1.37	1.69	1.61	1.33	1.25	1.47
19	1.00	1.59	1.63	0.77	0.74	1.27	1.19	0.71	1.23	1.21	1.19	1.47	1.32
20	1.00	5.77	2.69	1.49	1.26	1.61	1.19	1.23	1.31	1.47	1.35	1.48	1.51
21	1.00	2.28	1.93	2.00	1.34	2.18	1.20	1.06	1.43	1.22	1.17	1.49	1.98
22	1.00	1.77	1.51	2.46	1.87	2.61	1.88	1.28	1.29	1.24	1.40	1.29	1.35
23	1.00	1.37	1.20	1.71	1.63	1.81	1.65	1.30	1.12	1.09	1.17	1.33	1.35
24	1.00	1.78	1.65	1.79	1.24	1.97	1.57	0.97	0.98	0.93	1.11	0.87	1.12
25	1.00	2.58	1.86	3.07	1.91	5.20	2.84	1.53	1.47	1.41	1.99	1.96	1.94
26	1.00	2.54	1.72	2.26	1.44	2.56	1.62	1.23	1.16	1.11	1.16	1.09	1.10
27	1.00	1.85	2.01	2.11	1.88	1.90	1.70	2.12	1.15	1.13	1.42	1.34	1.50
28	1.00	1.51	1.82	3.32	3.06	3.58	3.55	2.51	1.50	1.45	2.29	2.37	2.45
29	1.00	1.31	1.97	2.49	1.99	2.48	1.83	1.85	1.20	1.15	1.79	2.02	2.65
30	1.00	2.00	2.19	2.59	1.90	2.40	1.85	1.80	1.25	1.22	1.34	1.29	1.42
31	1.00	1.47	1.18	2.17	1.98	2.29	2.03	1.35	1.10	0.99	1.79	1.60	1.69
32	1.00	1.78	1.95	3.48	3.05	3.74	3.27	1.54	0.98	1.00	2.82	3.20	3.91
33	1.00	1.81	1.78	1.30	1.14	2.83	1.76	1.09	1.23	1.21	1.37	1.28	1.18
34	1.00	4.00	1.71	2.02	1.45	2.93	1.63	1.85	1.16	1.15	1.05	1.44	1.23
35	1.00	2.43	2.05	4.68	1.49	5.33	2.19	1.63	1.67	1.58	1.44	2.27	2.92
36	1.00	1.69	1.79	3.90	3.35	3.95	3.61	1.60	1.33	1.25	3.42	3.95	4.21
37	1.00	1.44	1.52	4.21	2.50	4.23	2.61	1.61	1.33	1.13	2.50	2.45	2.18
38	1.00	1.70	1.61	2.60	1.38	2.62	1.35	1.76	1.36	1.27	1.15	0.99	1.05
39	1.00	2.57	1.92	3.08	1.26	3.06	1.26	1.60	1.20	1.15	1.05	1.03	1.20
40	1.00	2.56	2.46	2.05	1.32	2.06	1.27	2.16	1.44	1.41	1.32	1.85	2.02
41	1.00	1.91	1.81	1.64	1.14	1.48	1.24	1.31	1.19	1.13	1.09	1.13	1.11
42	1.00	2.57	2.34	2.84	1.69	2.69	1.61	1.31	1.73	1.51	1.58	2.07	2.14
43	1.00	3.34	2.15	2.70	1.24	2.59	1.26	1.50	1.02	0.95	1.11	1.06	1.16
44	1.00	2.12	1.82	2.37	1.45	2.36	1.53	1.25	1.26	1.17	1.35	1.21	1.23
45	1.00	2.84	2.11	1.45	1.34	1.43	1.26	1.74	1.38	1.38	1.23	1.03	1.00
46	1.00	2.89	2.15	1.85	1.45	2.00	1.55	1.76	2.17	2.14	1.18	1.88	2.09
47	1.00	2.43	2.54	1.96	1.82	1.94	1.80	2.13	1.90	1.95	1.36	1.72	2.09
48	1.00	2.51	2.19	1.54	1.36	1.61	1.27	1.52	1.26	1.28	1.48	1.70	1.56
49	1.00	3.02	3.04	1.35	1.72	1.57	2.16	1.65	1.34	1.33	1.63	2.28	2.64
50	1.00	3.36	3.07	1.66	1.20	1.78	1.48	1.26	1.32	1.33	1.23	1.96	1.95
51	1.00	3.97	2.71	1.98	2.03	2.27	2.29	1.72	1.16	1.12	1.28	1.51	2.01
52	1.00	2.73	2.40	3.20	1.04	3.81	1.26	2.00	0.98	0.93	0.95	0.98	1.05
	1.00	2.33	1.93	1.84	1.36	2.13	1.49	1.32	1.21	1.17	1.31	1.38	1.45

Table 16 – Relative Root Mean Squared Errors (wRRMSE) by week.

Weekday	mcb	naive	arima	linear	linear.r	linear.s	linear.sr	arimax	jonsson.s1	jonsson.s2	svmlinear	svmradial	avnnet
Mon	4.93	9.81	8.13	8.37	5.67	9.05	6.06	5.55	5.32	5.26	5.83	6.17	6.42
Tue	3.20	8.63	7.44	6.85	4.91	7.47	5.15	4.74	4.16	3.90	4.75	5.13	5.84
Wed	3.15	7.39	7.33	7.67	5.34	8.56	5.42	4.85	4.32	4.05	4.83	5.36	5.49
Thu	3.31	7.71	7.66	8.57	5.55	10.01	6.10	5.04	4.45	4.15	4.85	5.04	5.32
Fri	2.78	8.26	6.79	7.17	5.12	7.36	4.99	5.44	4.09	3.91	4.52	4.93	5.12
Sat	3.26	8.61	6.85	5.65	5.10	6.17	5.45	5.18	4.46	4.41	4.74	4.54	4.76
Sun	6.29	13.71	9.87	7.29	6.51	10.52	8.17	6.16	6.75	6.56	6.96	7.21	7.38
	4.02	9.38	7.78	7.42	5.48	8.57	5.99	5.30	4.87	4.69	5.27	5.55	5.82

Table 17 – RMSE by weekday.

Weekday	mcb	naive	arima	linear	linear.r	linear.s	linear.sr	arimax	jonsson.s1	jonsson.s2	svmlinear	svmradial	avnnet
Mon	1.00	1.99	1.65	1.70	1.15	1.84	1.23	1.13	1.08	1.07	1.18	1.25	1.30
Tue	1.00	2.69	2.32	2.14	1.53	2.33	1.61	1.48	1.30	1.22	1.48	1.60	1.82
Wed	1.00	2.34	2.32	2.43	1.69	2.72	1.72	1.54	1.37	1.29	1.53	1.70	1.74
Thu	1.00	2.33	2.32	2.59	1.68	3.03	1.84	1.53	1.35	1.25	1.47	1.52	1.61
Fri	1.00	2.97	2.44	2.58	1.84	2.65	1.80	1.96	1.47	1.41	1.63	1.77	1.84
Sat	1.00	2.64	2.10	1.73	1.56	1.89	1.67	1.59	1.37	1.35	1.45	1.39	1.46
Sun	1.00	2.18	1.57	1.16	1.03	1.67	1.30	0.98	1.07	1.04	1.11	1.15	1.17
	1.00	2.33	1.93	1.84	1.36	2.13	1.49	1.32	1.21	1.17	1.31	1.38	1.45

Table 18 – RRMSE by weekday.

	mcb	naive	arima	linear	linear.r	linear.s	linear.sr	arimax	jonsson.s1	jonsson.s2	svmlinear	svmradial	avnnet
Workday	3.55	8.40	7.48	7.76	5.33	8.55	5.56	5.14	4.49	4.28	4.97	5.34	5.65
Weekend	5.01	11.45	8.49	6.52	5.85	8.62	6.95	5.69	5.72	5.59	5.95	6.02	6.21
Holiday	5.33	13.03	13.82	12.77	5.69	14.92	6.85	10.11	4.89	4.99	6.20	8.73	7.54
Non Holiday	3.99	9.26	7.57	7.23	5.48	8.35	5.97	5.12	4.87	4.69	5.25	5.44	5.77

Table 19 – RMSE by workday, weekend, holiday, and non-holiday. A list of German holidays in 2014 can be found in Table 28 in the appendix.

Hour	mcb	naive	arima	linear	linear.r	linear.s	linear.sr	arimax	jonsson.s1	jonsson.s2	svmlinear	svmradial	avnnet
1	3.10	8.18	6.20	6.37	4.21	6.04	3.78	4.36	3.50	3.44	3.88	4.13	4.26
2	3.22	9.01	6.54	7.41	4.55	8.81	4.24	4.89	3.72	3.66	3.96	3.99	4.52
3	3.47	9.26	7.08	8.19	4.97	10.15	5.53	5.29	3.92	3.94	4.20	4.24	5.21
4	4.65	10.29	8.06	8.85	5.52	11.12	6.00	5.73	4.98	4.85	5.17	5.20	5.80
5	4.22	9.55	7.12	8.71	5.50	8.76	5.47	5.75	4.71	4.40	4.83	4.79	5.50
6	3.95	8.85	6.55	7.92	4.99	8.17	4.86	5.39	4.66	4.25	4.47	4.46	4.93
7	4.05	8.18	7.10	6.58	4.86	8.34	5.32	5.29	4.65	4.47	4.89	4.92	5.32
8	4.47	9.17	8.47	6.74	5.46	8.52	5.69	5.59	5.04	5.01	5.72	5.62	5.97
9	3.88	9.09	8.16	7.62	5.98	7.34	6.01	5.59	4.91	4.69	5.49	6.06	6.18
10	4.26	9.92	8.71	7.71	5.85	7.87	6.27	5.50	4.77	4.58	5.33	6.23	6.54
11	3.82	9.88	8.64	7.54	5.81	7.54	6.15	5.60	4.51	4.39	5.16	6.04	6.18
12	3.80	9.97	8.81	7.93	6.25	8.20	6.59	5.96	4.74	4.50	5.42	6.15	6.43
13	3.98	9.98	8.71	7.74	6.11	8.58	6.57	5.61	4.75	4.65	5.69	6.18	6.28
14	4.90	11.96	9.90	7.34	5.91	8.05	7.43	5.34	6.02	6.04	6.66	7.17	7.07
15	4.68	11.85	9.76	7.41	5.86	12.49	7.79	5.31	6.13	6.10	6.60	7.03	6.98
16	4.83	11.24	9.25	6.99	5.54	11.87	7.27	5.03	6.04	5.93	6.28	6.78	6.76
17	3.97	9.27	7.74	7.00	5.65	6.85	7.01	5.29	5.14	4.78	5.22	5.59	5.68
18	3.81	8.63	7.38	7.64	6.06	8.46	6.41	5.43	5.04	4.91	5.63	5.72	6.09
19	4.40	9.76	8.57	8.32	6.57	9.75	6.99	6.26	5.62	5.37	6.27	6.44	6.99
20	4.44	9.10	8.06	8.15	6.52	9.19	6.53	5.90	5.66	5.17	6.12	6.22	6.30
21	3.54	7.77	6.31	6.80	4.99	7.99	6.02	4.89	4.21	4.14	4.55	5.06	5.35
22	3.51	6.91	5.68	6.42	4.59	6.47	5.22	4.50	4.28	4.09	4.44	4.70	4.94
23	3.53	7.04	5.54	6.18	4.58	5.73	4.52	4.25	4.41	4.12	4.65	4.41	4.77
24	3.37	8.08	6.09	5.51	4.21	5.02	3.83	3.68	4.40	3.98	4.53	3.91	4.10
	4.02	9.38	7.78	7.42	5.48	8.57	5.99	5.30	4.87	4.69	5.27	5.55	5.82

Table 20 – RMSE by hour of day.

Hour	mcb	naive	arima	linear	linear.r	linear.s	linear.sr	arimax	jonsson.s1	jonsson.s2	svmlinear	svmradial	avnnet
1	1.00	2.64	2.00	2.06	1.36	1.95	1.22	1.41	1.13	1.11	1.25	1.33	1.37
2	1.00	2.80	2.03	2.30	1.41	2.74	1.32	1.52	1.16	1.14	1.23	1.24	1.40
3	1.00	2.67	2.04	2.36	1.43	2.92	1.59	1.52	1.13	1.13	1.21	1.22	1.50
4	1.00	2.21	1.73	1.90	1.19	2.39	1.29	1.23	1.07	1.04	1.11	1.12	1.25
5	1.00	2.26	1.69	2.06	1.30	2.07	1.30	1.36	1.12	1.04	1.14	1.13	1.30
6	1.00	2.24	1.66	2.00	1.26	2.07	1.23	1.36	1.18	1.08	1.13	1.13	1.25
7	1.00	2.02	1.75	1.62	1.20	2.06	1.31	1.31	1.15	1.10	1.21	1.22	1.32
8	1.00	2.05	1.89	1.51	1.22	1.91	1.27	1.25	1.13	1.12	1.28	1.26	1.34
9	1.00	2.34	2.10	1.96	1.54	1.89	1.55	1.44	1.26	1.21	1.41	1.56	1.59
10	1.00	2.33	2.05	1.81	1.37	1.85	1.47	1.29	1.12	1.07	1.25	1.46	1.54
11	1.00	2.59	2.26	1.98	1.52	1.98	1.61	1.47	1.18	1.15	1.35	1.58	1.62
12	1.00	2.62	2.32	2.08	1.64	2.15	1.73	1.57	1.25	1.18	1.43	1.62	1.69
13	1.00	2.51	2.19	1.95	1.54	2.16	1.65	1.41	1.19	1.17	1.43	1.56	1.58
14	1.00	2.44	2.02	1.50	1.21	1.64	1.52	1.09	1.23	1.23	1.36	1.46	1.44
15	1.00	2.53	2.08	1.58	1.25	2.67	1.66	1.13	1.31	1.30	1.41	1.50	1.49
16	1.00	2.33	1.92	1.45	1.15	2.46	1.51	1.04	1.25	1.23	1.30	1.40	1.40
17	1.00	2.33	1.95	1.76	1.42	1.73	1.76	1.33	1.29	1.20	1.31	1.41	1.43
18	1.00	2.27	1.94	2.01	1.59	2.22	1.68	1.43	1.32	1.29	1.48	1.50	1.60
19	1.00	2.22	1.95	1.89	1.49	2.21	1.59	1.42	1.28	1.22	1.42	1.46	1.59
20	1.00	2.05	1.82	1.84	1.47	2.07	1.47	1.33	1.28	1.17	1.38	1.40	1.42
21	1.00	2.20	1.78	1.92	1.41	2.26	1.70	1.38	1.19	1.17	1.29	1.43	1.51
22	1.00	1.97	1.62	1.83	1.31	1.85	1.49	1.28	1.22	1.17	1.27	1.34	1.41
23	1.00	1.99	1.57	1.75	1.30	1.62	1.28	1.20	1.25	1.17	1.32	1.25	1.35
24	1.00	2.40	1.81	1.64	1.25	1.49	1.14	1.09	1.31	1.18	1.34	1.16	1.22
	1.00	2.33	1.93	1.84	1.36	2.13	1.49	1.32	1.21	1.17	1.31	1.38	1.45

Table 21 – RRMSE by hour of day.

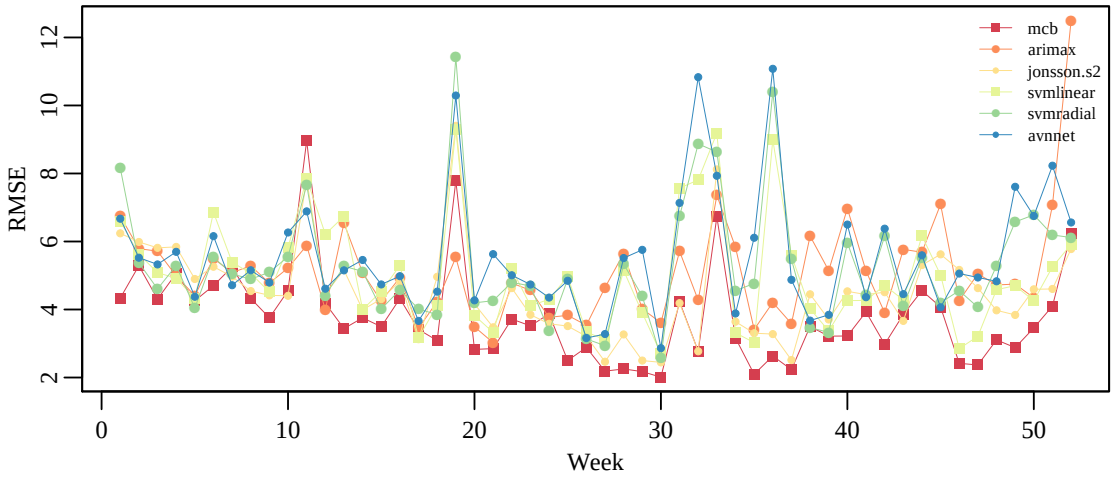

Figure 27 – RMSE by week.

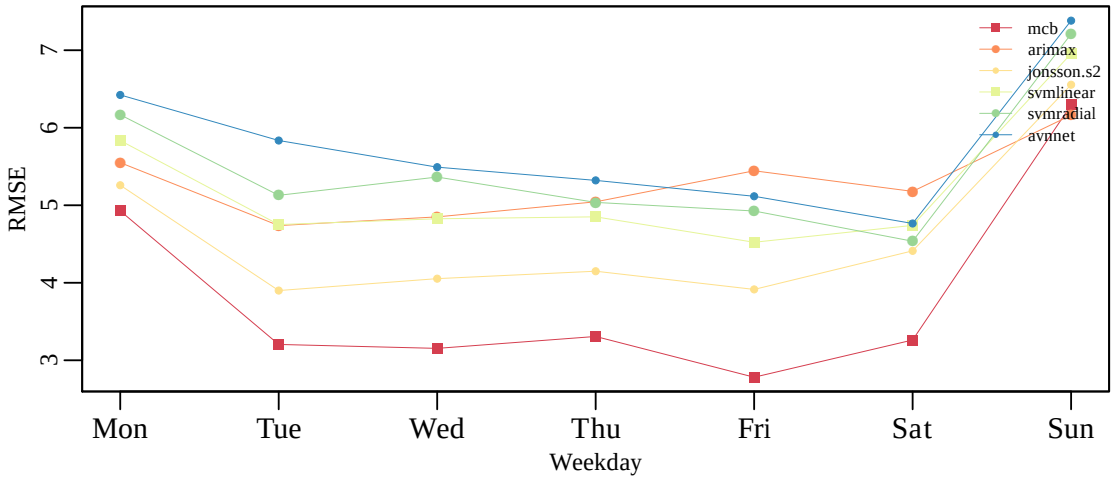

Figure 28 – RMSE by weekday.

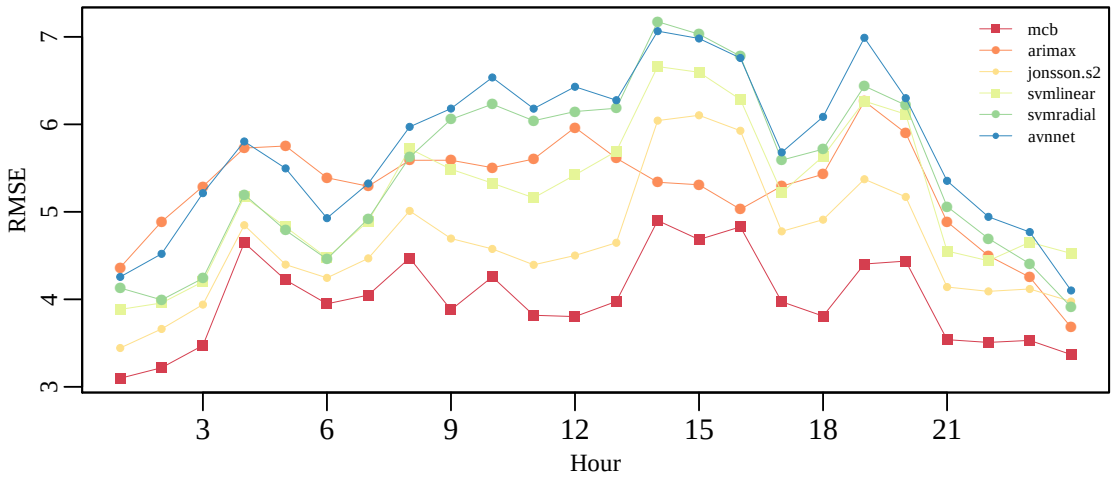

Figure 29 – RMSE by hour of day.

8.2 Combining forecasts

What is the value of an inferior forecast?

Inspired by the literature on forecast aggregation, which was reviewed in Section 4, we implement a simple forecast aggregation scheme where four of the models that show acceptable, but inferior, performance are combined into the $CF1$ forecast[23] [24] (Equation 38). Next we combine the $CF1$ forecast with the Jonsson step-two forecast to obtain the $CF2$ forecast (Equation 13).

$$CF1 = \frac{1}{4}\left(\hat{P}_{arimax} + \hat{P}_{svmlinear} + \hat{P}_{svmradial} + \hat{P}_{avnnet}\right) \tag{38}$$

$$CF2 = \frac{1}{2}\left(\hat{P}_{CF1} + \hat{P}_{jonsson.s2}\right) \tag{39}$$

In Table 24 and Figure 30, we see that the $CF1$ forecast achieves an average weekly RMSE of 4.70, which is on par with the Jonsson step-two model (RMSE of 4.69) that has the best performance of all models. In Table 22, we observe that this is a statistically significant improvement to every single constituent model.

For the $CF2$ forecast, we see a further improvement in forecast performance and it attains a weekly RMSE of 4.31. The $CF2$ forecast performs significantly better than all the other forecasts and, while still significantly worse at a 5% significance level, it loses by just 7% to the MCB.

Looking at the performance across weekdays in Figure 31, we see that the MCB is still superior on Tuesday-Saturday, but is now beaten by the $CF2$ forecast on Sunday and Monday, where the MCB uses stale information.

In Figure 32, we see that the performances of the MCB and $CF2$ forecast are almost the same in off-peak hours (H1–H8 & H21–H24), while in peak hours (H9–H20) the MCB still looks superior.

	MCB	arimax	jonsson.s2	svmlinear	svmradial	avnnet	CF1	CF2
MCB		-4.93**	-5.31**	-7.32**	-6.32**	-6.72**	-3.78**	-2.28*
arimax	4.93**		2.40*	0.11	-0.83	-1.79	2.67**	4.22**
jonsson.s2	5.31**	-2.40*		-4.01**	-4.16**	-4.65**	-0.02	4.94**
svmlinear	7.32**	-0.11	4.01**		-1.85	-3.19**	5.02**	7.79**
svmradial	6.32**	0.83	4.16**	1.85		-2.41*	5.87**	6.69**
avnnet	6.72**	1.79	4.65**	3.19**	2.41*		7.73**	7.09**
CF1	3.78**	-2.67**	0.02	-5.02**	-5.87**	-7.73**		3.62**
CF2	2.28*	-4.22**	-4.94**	-7.79**	-6.69**	-7.09**	-3.62**	

Table 22 – Two-way Diebold-Mariano tests including combination forecasts. A significant positive (negative) test statistic indicates that the top (left) model has superior forecast performance. ** denotes significance at a 1% level. * denotes significance at a 5% level.

[23] The main objectives here is to encourage diversity in the models we select, which is why we do not include the step one models, and include models that show decent forecast ability.

[24] Note that when we use the ex post performance in the test period to choose which models to average, we in effect peeked at the test data which strictly speaking invalidates it as an independent validation set. The aggregation scheme should be based on train period performance. Having this fallacity in mind, it is still believed by the author that the following results are valid due to the small freedom allowed compared with the significant performance gains.

(A)

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.05	1.54	0.68	0.50
Weekend	-2.41	2.47	-0.98	0.33
Peak	3.21	2.10	1.53	0.13
Summer	1.29	2.10	0.62	0.54
Cons	0.12	0.17	0.68	0.50
Wind	-0.52	0.53	-0.97	0.33
Solar	-0.11	0.23	-0.47	0.64

(B)

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.45	1.55	0.29	0.77
Peak	3.89	1.80	2.17	0.03*

Table 23 – Conditional Diebold-Mariano test of CF2 against MCB including combination forecasts. In (A) all considered regressors are included. In (B) insignificant regressors have been successively eliminated. ** denotes significance at a 1% level. * denotes significance at a 5% level.

Applying the conditional DM test to the loss differential between the MCB and $CF2$ forecast, we observe that the peak dummy is still significant but now the intercept is no longer significant.
This shows, that the model performance differential is significantly affected by the peak variable, but neither by the other explanatory variables nor the constant. So, equal performance of the $CF2$ forecast in off-peak hours cannot be rejected, but in peak hours the MCB still shows statistically superior forecast performance.

Week	mcb	arimax	jonsson.s2	svmlinear	svmradial	avnnet	CF1	CF2
1	4.33	6.75	6.24	6.60	8.17	6.67	5.81	5.80
2	5.29	5.79	5.99	5.63	5.37	5.52	5.06	5.20
3	4.31	5.72	5.81	5.08	4.61	5.33	4.83	5.05
4	5.06	4.99	5.84	4.92	5.29	5.70	4.91	5.12
5	4.26	4.41	4.90	4.32	4.04	4.38	3.84	4.17
6	4.70	5.50	5.26	6.86	5.54	6.16	5.63	4.89
7	5.25	5.09	4.97	5.39	5.05	4.72	4.37	4.56
8	4.32	5.28	4.54	5.06	4.89	5.15	4.72	4.39
9	3.76	4.77	4.42	4.51	5.10	4.79	4.32	3.97
10	4.55	5.22	4.40	5.81	5.55	6.26	5.13	4.18
11	8.97	5.88	7.82	7.85	7.67	6.89	6.49	6.90
12	4.18	3.99	4.46	6.22	4.41	4.61	4.07	4.13
13	3.44	6.54	5.16	6.74	5.27	5.15	5.50	5.13
14	3.76	5.08	3.99	4.00	5.12	5.46	4.26	3.71
15	3.49	4.24	4.31	4.53	4.02	4.74	3.84	3.55
16	4.32	4.94	4.83	5.29	4.58	4.98	4.39	4.27
17	3.41	3.45	3.13	3.17	4.02	3.67	3.02	2.96
18	3.09	4.23	4.96	4.12	3.85	4.53	3.61	4.11
19	7.79	5.53	9.40	9.30	11.42	10.29	8.49	8.78
20	2.83	3.49	4.16	3.82	4.20	4.27	3.38	3.45
21	2.85	3.02	3.48	3.33	4.26	5.63	3.60	3.36
22	3.72	4.78	4.63	5.20	4.80	5.01	4.44	4.16
23	3.52	4.57	3.85	4.11	4.69	4.73	4.07	3.32
24	3.87	3.77	3.58	4.30	3.37	4.35	3.13	3.02
25	2.50	3.83	3.51	4.97	4.90	4.85	3.72	2.86
26	2.89	3.54	3.21	3.34	3.13	3.16	2.43	2.59
27	2.19	4.64	2.46	3.11	2.94	3.28	3.17	2.64
28	2.25	5.65	3.27	5.15	5.34	5.51	4.98	3.81
29	2.17	4.02	2.50	3.90	4.39	5.75	4.15	2.61
30	2.01	3.61	2.45	2.70	2.59	2.86	2.29	2.25
31	4.22	5.71	4.19	7.57	6.74	7.13	6.60	5.18
32	2.77	4.27	2.77	7.81	8.87	10.83	7.58	4.33
33	6.73	7.36	8.12	9.19	8.63	7.93	6.54	6.97
34	3.16	5.83	3.64	3.31	4.55	3.88	3.34	3.26
35	2.10	3.41	3.30	3.03	4.77	6.11	3.09	2.63
36	2.63	4.21	3.28	8.99	10.40	11.08	8.16	4.58
37	2.23	3.58	2.51	5.59	5.47	4.87	3.99	2.72
38	3.50	6.16	4.45	4.02	3.47	3.67	3.68	3.90
39	3.22	5.14	3.71	3.37	3.33	3.84	2.97	2.89
40	3.22	6.97	4.53	4.26	5.96	6.50	4.57	4.09
41	3.93	5.13	4.44	4.27	4.43	4.36	3.84	3.95
42	2.98	3.91	4.51	4.71	6.16	6.38	4.59	4.33
43	3.85	5.77	3.67	4.28	4.10	4.45	3.54	3.34
44	4.55	5.68	5.31	6.17	5.51	5.59	4.86	4.98
45	4.07	7.09	5.62	5.01	4.20	4.07	4.25	4.67
46	2.42	4.26	5.17	2.85	4.55	5.06	3.62	4.19
47	2.37	5.05	4.63	3.21	4.07	4.94	3.60	3.81
48	3.11	4.72	3.98	4.58	5.28	4.83	4.51	3.88
49	2.88	4.75	3.84	4.70	6.58	7.61	4.57	3.49
50	3.46	4.34	4.59	4.26	6.78	6.75	4.27	4.14
51	4.10	7.07	4.60	5.25	6.19	8.23	5.75	4.79
52	6.24	12.50	5.78	5.91	6.12	6.56	5.48	5.02
	4.02	5.30	4.69	5.27	5.55	5.82	4.70	4.31

Table 24 – Root Mean Squared Error (wRMSE) by week, including combination forecasts. CF1 and CF2 are combined forecasts defined in Equation 38 & 39.

Week	mcb	arimax	jonsson.s2	svmlinear	svmradial	avnnet	CF1	CF2
1	1.00	1.56	1.44	1.52	1.89	1.54	1.34	1.34
2	1.00	1.09	1.13	1.06	1.02	1.04	0.96	0.98
3	1.00	1.33	1.35	1.18	1.07	1.24	1.12	1.17
4	1.00	0.99	1.15	0.97	1.05	1.13	0.97	1.01
5	1.00	1.04	1.15	1.02	0.95	1.03	0.90	0.98
6	1.00	1.17	1.12	1.46	1.18	1.31	1.20	1.04
7	1.00	0.97	0.95	1.03	0.96	0.90	0.83	0.87
8	1.00	1.22	1.05	1.17	1.13	1.19	1.09	1.02
9	1.00	1.27	1.18	1.20	1.36	1.28	1.15	1.06
10	1.00	1.15	0.97	1.28	1.22	1.38	1.13	0.92
11	1.00	0.66	0.87	0.88	0.86	0.77	0.72	0.77
12	1.00	0.95	1.07	1.49	1.05	1.10	0.97	0.99
13	1.00	1.90	1.50	1.96	1.53	1.50	1.60	1.49
14	1.00	1.35	1.06	1.06	1.36	1.45	1.13	0.99
15	1.00	1.21	1.23	1.30	1.15	1.36	1.10	1.02
16	1.00	1.14	1.12	1.22	1.06	1.15	1.02	0.99
17	1.00	1.01	0.92	0.93	1.18	1.07	0.89	0.87
18	1.00	1.37	1.61	1.33	1.25	1.47	1.17	1.33
19	1.00	0.71	1.21	1.19	1.47	1.32	1.09	1.13
20	1.00	1.23	1.47	1.35	1.48	1.51	1.20	1.22
21	1.00	1.06	1.22	1.17	1.49	1.98	1.26	1.18
22	1.00	1.28	1.24	1.40	1.29	1.35	1.19	1.12
23	1.00	1.30	1.09	1.17	1.33	1.35	1.16	0.94
24	1.00	0.97	0.93	1.11	0.87	1.12	0.81	0.78
25	1.00	1.53	1.41	1.99	1.96	1.94	1.49	1.14
26	1.00	1.23	1.11	1.16	1.09	1.10	0.84	0.90
27	1.00	2.12	1.13	1.42	1.34	1.50	1.45	1.21
28	1.00	2.51	1.45	2.29	2.37	2.45	2.21	1.69
29	1.00	1.85	1.15	1.79	2.02	2.65	1.91	1.20
30	1.00	1.80	1.22	1.34	1.29	1.42	1.14	1.12
31	1.00	1.35	0.99	1.79	1.60	1.69	1.56	1.23
32	1.00	1.54	1.00	2.82	3.20	3.91	2.73	1.56
33	1.00	1.09	1.21	1.37	1.28	1.18	0.97	1.04
34	1.00	1.85	1.15	1.05	1.44	1.23	1.06	1.03
35	1.00	1.63	1.58	1.44	2.27	2.92	1.48	1.25
36	1.00	1.60	1.25	3.42	3.95	4.21	3.10	1.74
37	1.00	1.61	1.13	2.50	2.45	2.18	1.79	1.22
38	1.00	1.76	1.27	1.15	0.99	1.05	1.05	1.11
39	1.00	1.60	1.15	1.05	1.03	1.20	0.92	0.90
40	1.00	2.16	1.41	1.32	1.85	2.02	1.42	1.27
41	1.00	1.31	1.13	1.09	1.13	1.11	0.98	1.00
42	1.00	1.31	1.51	1.58	2.07	2.14	1.54	1.45
43	1.00	1.50	0.95	1.11	1.06	1.16	0.92	0.87
44	1.00	1.25	1.17	1.35	1.21	1.23	1.07	1.09
45	1.00	1.74	1.38	1.23	1.03	1.00	1.04	1.15
46	1.00	1.76	2.14	1.18	1.88	2.09	1.50	1.74
47	1.00	2.13	1.95	1.36	1.72	2.09	1.52	1.61
48	1.00	1.52	1.28	1.48	1.70	1.56	1.45	1.25
49	1.00	1.65	1.33	1.63	2.28	2.64	1.59	1.21
50	1.00	1.26	1.33	1.23	1.96	1.95	1.24	1.20
51	1.00	1.72	1.12	1.28	1.51	2.01	1.40	1.17
52	1.00	2.00	0.93	0.95	0.98	1.05	0.88	0.81
	1.00	1.32	1.17	1.31	1.38	1.45	1.17	1.07

Table 25 – Relative Root Mean Squared Error (wRRMSE) by week, including combination forecasts. CF1 and CF2 are combined forecasts defined in Equation 38 & 39.

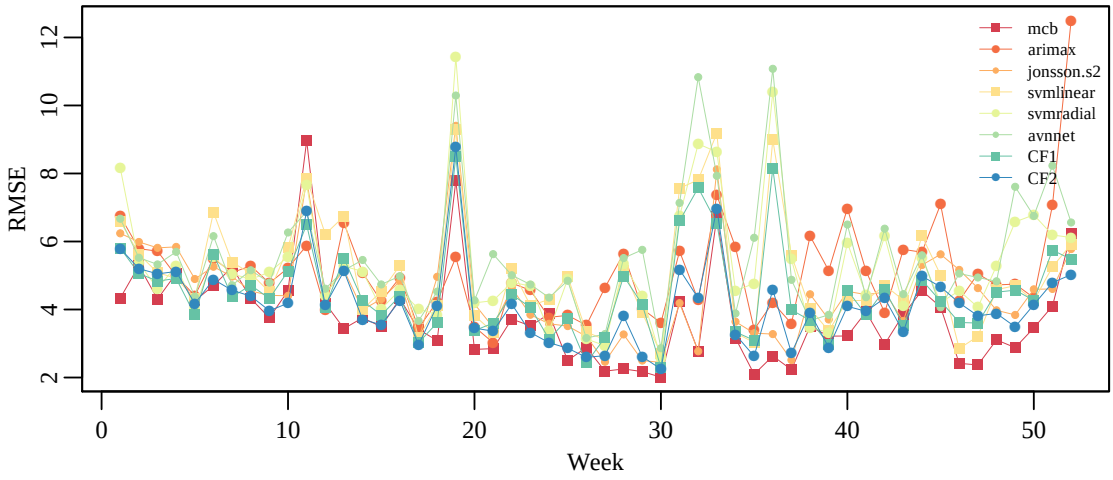

Figure 30 – RMSE by week, including combination forecasts.

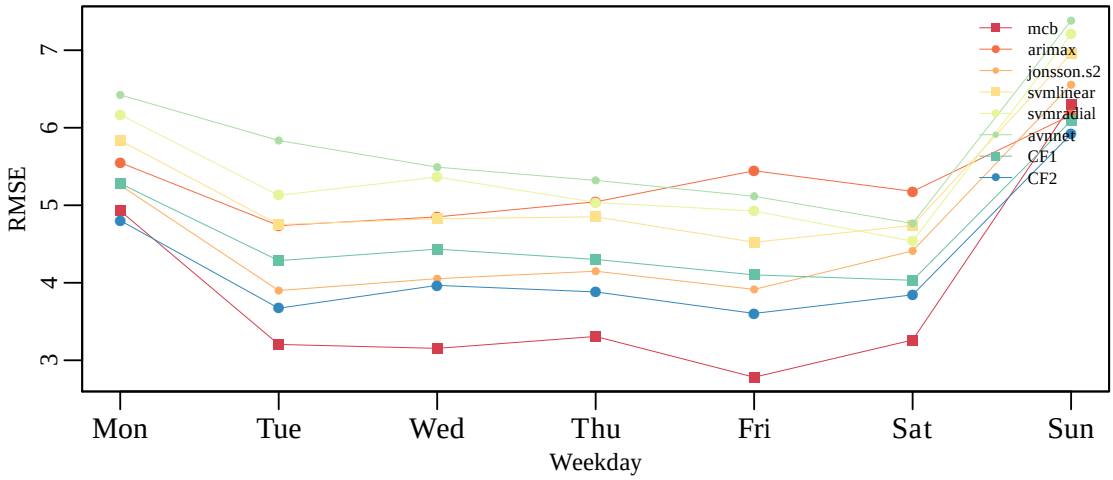

Figure 31 – RMSE by weekday, including combination forecasts.

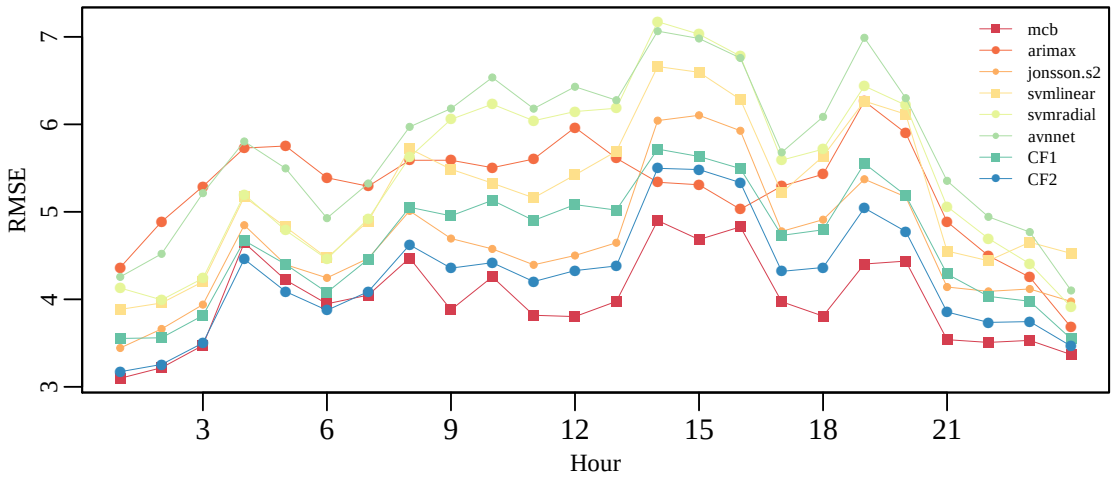

Figure 32 – RMSE by hour, including combination forecasts.

8.3 Discussion

So, how does the choice of scoring function affect the ranking of the models? In table 26, we have listed the ranking of all models by both MAE and RMSE. While the relative measures are slightly different, the rankings are mostly the same— only the 8th and 9th best model change places. So from the perspective of selecting the optimal model, there does not seem to be much difference between using RMSE or MAE as the scoring function.

Model	RMSE	RRMSE	RMSE-Rank	MAE	RMAE	MAE-Rank
mcb	4.02	1.00	1	2.71	1.00	1
CF2	4.31	1.07	2	3.15	1.16	2
jonsson.s2	4.69	1.17	3	3.38	1.25	3
CF1	4.70	1.17	4	3.49	1.29	4
jonsson.s1	4.87	1.21	5	3.52	1.30	5
svmlinear	5.27	1.31	6	3.90	1.44	6
arimax	5.30	1.32	7	3.94	1.46	7
linear.r	5.48	1.36	8	4.22	1.56	9
svmradial	5.55	1.38	9	4.02	1.48	8
avnnet	5.82	1.45	10	4.24	1.57	10
linear.sr	5.99	1.49	11	4.33	1.60	11
linear	7.42	1.84	12	5.55	2.05	12
arima	7.78	1.93	13	5.62	2.08	13
linear.s	8.57	2.13	14	5.99	2.21	14
naïve	9.38	2.33	15	6.43	2.38	15

Table 26 – Ranking of all models according to both their average RMSE and MAE in the out-of-sample period. The column RMSE is the average root mean squared error of a model. RRMSE is the RMSE relative to the market consensus benchmark. RMSE-Rank is the rank of a model according to its RMSE performance. MAE is the average absolute error of a model. RMAE is the MAE relative to the market consensus benchmark. MAE-Rank is the rank of a model according to its MAE performance. Data is from Table 15, 16, 24 & 25 in Section 8, and from Table 36, 37, 38 & 39 in the appendix.

The ARIMAX and Jonsson step-two method, which use a two-step methodology that corrects autocorrelation in the residuals of a step-one model, both perform better than their step-one counterpart. This is not really a surprise, as the autocorrelation in the step-one residuals indicate that not all information was extracted. However, it does suggest that models which show residual autocorrelation might be able to benefit, at least to a certain extent, from having an inter-temporal model fitted to the residuals. As we did not do this with the two machine learning methods, there might be room for further improvements in these if their residuals contain significant autocorrelation.

The Jonsson model displays impressive performance compared to the other statistical and machine learning models. This might be due to the fact that adaptivity is build into to model, or it might be that the conditioning on explanatory variables drives the nice performance.

However, it was shown that forecast aggregation can improve even on the performance of best individual model. This result demonstrates that inferior models can still contain useful information. A forecasting strategy might therefore benefit from fitting several disparate models and combine their forecasts.

Given the performances observed, it does not seem inconceivable that a statistical model, a machine learning model, or a combination could beat the market benchmark if the reviewed models are improved (for example, by incorporating more explanatory variables) and thoroughly calibrated (for example, by adjusting the estimation window). This result is interesting, and somewhat surprising, as we might expect that markets participants would use very capable and complex models of the electricity system to forecast prices.

9 Conclusion

In this thesis, we have investigated the forecast performance of a range of statistical and machine learning methods used to predict the day-ahead electricity prices German/Austrian.
First, we analyzed the structure of the German/Austrian electricity market, and identified key variables that have explanatory power for the German/Austrian electricity prices.
We also established that the market might show high volatility due to the fundamental characteristics of the electricity system.
Then, we introduced the EXAA auction as a powerful market benchmark that can be seen as a snapshot of the market participants expectations of day-ahead prices on the main German/Austrian power exchange, EPEX Spot.

To forecast the day-ahead electricity prices on EPEX Spot, we presented a number of different forecasting methods of both statistical and machine learning origin.
After having calibrated models, with a focus on predictive ability, day-ahead forecasts were generated for a 362 days out-of-sample testing period in 2014. The empirical results showed that the adaptive model of Jónsson *et al.* (2013) has the best forecasting performance of the surveyed models, and is statistically different to all but one contender. The good performance is probably due to either the adaptive nature of the model, which makes it able to adapt to gradual changes in underlying model relationship, or the way that the model conditions on explanatory variables.

The best individual model was, however, beaten by an aggregate forecast, where the forecasts from the five best models are combined. This forecast showed statistically superiour performance to all individual models. This result confirms that even inferior forecasts can contain information of value in forecasting. The final forecast performance was relatively close to that of the market benchmark, which is encouraging for future research on this subject.

During model evaluation, it was stressed that a consistent scoring function should be used. As we were interested in forecasting the mean prices, we chose to use root mean squared errors. However, when ranking the models according to both root mean squared errors and mean absolute errors there were only marginal differences in the ranking. This result indicates that when choosing between these two scoring functions, the wrong choice of scoring function (in relation to the forecast users intentended use) can still lead to the optimal model.

It was also noted, that differences in forecast performances have to be assessed statistically. We did this by applying the Diebold-Mariano test. This showed that for some pairs of models with almost the same average score, equal performance could be clearly rejected. However, for other pairs with more disparate

average score, equal performance could not be rejected. This highlights the need for using statistical tests when comparing model forecasting performances.

10 Future research

The field of electricity price forecasting still holds many unanswered questions. Below is a non-exhaustive list of subjects that were encountered, but not fully examined, in the thesis, either for reasons of brevity or because data was not yet available.

- All methods surveyed in this thesis could likely benefit from the inclusion of more explanatory variables. It will be interesting to see which methods can use additional variables most effectively.
- There are good arguments for using adaptive methods in electricity price forecasting. Incorporating adaptiveness in the surveyed methods, or using other methods with adaptive features, would be an interesting expansion.
- The methods surveyed might benefit from better preprocessing of the data series (for example, to remove estimation influence of outliers, or remove seasonality). A comparison of different preprocessing schemes might be interesting.
- We did not touch upon different methods for handling seasonalities, but assumed that the inclusion of calendar dummies was sufficient to account for seasonality changes. There are, however, other methods for handling seasonality, and these might lead to better forecasting performance of the models (see e.g. (Weron, 2006) or (Chaâbane, 2014b)).
- We restricted the support vector machine (SVM) analysis to only two kernels. An analysis of the forecasting performance of SVM with other kernels would be an interesting expansion to the literature.
- There are many different types of neural networks that go beyond the simple network that we considered. Furthermore, other strategies for selecting the number of hidden units and estimating weights exist (see e.g. (White, 2006)). In view of this, a more comprehensive study comparing different ways to apply neural networks to the data would be interesting.
- We used simple cross-validation (CV) to select models in the machine-learning methods. This is not appropriate in a setting where there is dependence between observations, such as time series. It would therefore be interesting to use a proper version of CV that takes account of the autocorrelation, and examine what effects it has on model selection (see e.g. (Opsomer *et al.*, 2001)).
- In the future, as more European data on price driving fundamentals become available, it will be fascinating to see the results of a cross country comparison of forecasting methods.

11 Bibliography

References

Aggarwal, S. K., Saini, L. M. and Kumar, A. 2009. Electricity price forecasting in deregulated markets: A review and evaluation. *International Journal of Electrical Power & Energy Systems* 31(1), pp. 13–22. Available at: http://dx.doi.org/10.1016/j.ijepes.2008.09.003.

Akaike, H. 1974. A new look at the statistical model identification. *IEEE Transactionson Automatic Control* 19(6), pp. 716–723.

Akima, H. 1978. A Method of Bivariate Interpolation and Smooth Surface Fitting for Irregularly Distributed Data Points. *ACM Transactions on Mathematical Software* 4(2), pp. 148–159.

Akima, H. and Gebhardt, A. 2013. *akima: Interpolation of irregularly spaced data*, [Online].

Amjady, N., Daraeepour, A. and Keynia, F. 2010. Day-ahead electricity price forecasting by modified relief algorithm and hybrid neural network. *IET generation, transmission & distribution* 4(3), pp. 432–444.

Amjady, N. and Hemmati, M. 2006. Energy price forecasting: Problems and proposals for such predictions. *IEEE Power and Energy Magazine* 4(2), pp. 20–29.

Arlot, S. and Celisse, A. 2010. A survey of cross-validation procedures for model selection. *Statistics Surveys* 4, pp. 40–79. Available at: http://eprints.pascal-network.org/archive/00006812/.

Bates, J. M. and Granger, C. W. J. 1969. The Combination of Forecasts. *Journal of the Operational Research Society* 20(4), pp. 451–468.

Beck, M. 2015. *NeuralNetTools: Visualization and Analysis Tools for Neural Networks*, Available at: http://cran.r-project.org/package=NeuralNetTools.

Bordignon, S., Bunn, D. W., Lisi, F. and Nan, F. 2013. Combining day-ahead forecasts for British electricity prices. *Energy Economics* 35, pp. 88–103. Available at: http://dx.doi.org/10.1016/j.eneco.2011.12.001.

Breiman, L. 1996. Bagging predictors. *Machine Learning* 24(2), pp. 123–140.

Breiman, L. 2001. Statistical modeling: The two cultures. *Statistical Science* 16(3), pp. 199–215.

Bunn, D. W. 2000. Forecasting loads and prices in competitive power markets. *Proceedings of the IEEE* 88(2), pp. 163–169.

Chaâbane, N. 2014a. A hybrid ARFIMA and neural network model for electricity price prediction. *International Journal of Electrical Power & Energy Systems* 55, pp. 187–194.

Chaâbane, N. 2014b. A novel auto-regressive fractionally integrated moving average least-squares support vector machine model for electricity spot prices prediction. *Journal of Applied Statistics* 41(3), pp. 635–651.

Chen, X., Dong, Z. Y., Member, S., Meng, K., Xu, Y., Member, S., Wong, K. P. and Ngan, H. W. 2012. Electricity Price Forecasting With Extreme Learning Machine and Bootstrapping. *IEEE Transactions on Power Systems* 27(4), pp. 2055–2062.

Cleveland, W. S. and Devlin, S. J. 1988. Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association* 83(403), pp. 596–610.

Conejo, A. J., Contreras, J., Espínola, R. and Plazas, M. A. 2005. Forecasting electricity prices for a day-ahead pool-based electric energy market. *International Journal of Forecasting* 21(3), pp. 435–462.

Consentec. 2014. Description of load-frequency control concept and market for control reserves - Study commissioned by the German TSOs - (ordered by 50Hertz Transmission GmbH). Tech. Rep. February, Available at: https://transparency.entsoe.eu/balancing-domain/r2/rulesOnBalancing/show.

Cortes, C. and Vapnik, V. 1995. Support-vector networks. *Machine Learning* 20(3), pp. 273–297.

Cruz, A., Muñoz, A., Zamora, J. L. and Espínola, R. 2011. The effect of wind generation and weekday on Spanish electricity spot price forecasting. *Electric Power Systems Research* 81(10), pp. 1924–1935.

Cybenko, G. 1989. Approximations by Superpositions of a Sigmoidal Function. *Mathematics of Control, Signals, and Systems* 2, pp. 303–314.

Diebold, F. X. 2013. Comparing Predictive Accuracy, Twenty Years Later: A Personal Perspective on the Use and Abuse of Diebold-Mariano Tests .

Diebold, F. X. and Mariano, R. S. 1995. Comparing Predictive Accuracy. *Journal of Business & Economic Statistics* 13(3), pp. 134–144.

Drucker, H., Burges, C. J., Kaufman, L., Smola, A. and Vapnik, V. 1997. Support vector regression machines. *Advances in neural information processing systems* 9, pp. 155–161.

Energy Exchange Austria (EXAA). 2014. RULES FOR THE TRADING OF SPOT MARKET PRODUCTS FOR ELECTRIC POWER ON THE VIENNA STOCK EXCHANGE IN ITS FUNCTION AS A - TRADING RULES SPOT MARKET. Tech. Rep. September, EXAA, Available at: http://www.exaa.at/exaa/framework/2014-09-03_trading-rules.pdf.

Energy Exchange Austria (EXAA). 2015. *EXAA Company Brochure 2015*, [Online]. Available at: http://www.exaa.at/exaa/docs/exaa_brochure_2015_web_en.pdf.

EXAA. 2013. Annual and Sustainability Report 2013 .

Funahashi, K.-i. 1989. On the Approximate Realization of Continuous Mappings by Neural Networks. *Neural Networks* 2, pp. 183–192.

García-Ascanio, C. and Maté, C. 2010. Electric power demand forecasting using interval time series: A comparison between VAR and iMLP. *Energy Policy* 38(2), pp. 715–725. Available at: http://www.sciencedirect.com/science/article/pii/S0301421509007344.

Giacomini, R. and White, H. 2006. Tests of conditional predictive ability. *Econometrica* 74(6), pp. 1545–1578.

Gneiting, T. 2009. Making and Evaluating Point Forecasts. *Journal of the American Statistical Association* 106, pp. 37–41. Available at: http://arxiv.org/abs/0912.0902.

Godfrey, L. G. 1978. Testing for Higher Order Serial Correlation in Regression Equations When the Regressors Include Lagged Dependent Variables. *Econometrica* 46(6), pp. 1303–1310.

Goffe, W. L., Ferrier, G. D. and Rogers, J. 1994. Global optimization of statistical functions with simulated annealing. *Journal of Econometrics* 60, pp. 65–99.

Granger, C. W. J. 1993. On the Limitations of Comparing Mean-Square Forecast Errors - Comment. *Journal of Forecasting* 12(8), pp. 651–652. Available at: http://wrap.warwick.ac.uk/20925/.

Gupta, K. 2000. Neural Network Structures. In: *Neural Networks for RF and Microwave Design*, chap. 3, pp. 61–103.

Hamilton, J. D. 1994. *Time series analysis*. Princeton University Press.

Hannan, E. J. 1980. The Estimation of the Order of an ARMA Process. *The Annals of Statistics* 8(5), pp. 1071–1081.

Hart, D. and Wehrly, T. E. 1986. Using Repeated Estimation Regression Kernel Data Measurements. *Journal of the American Statistical Association* 81(396), pp. 1080–1088.

Hornik, K., Tinchcombe, M. and White, H. 1989. Multilayer Feedforward Networks are Universal Approximators. *Neural Networks* 2, pp. 359–366.

Hsu, C.-W., Chang, C.-C. and Lin, C.-J. 2008. A Practical Guide to Support Vector Classification. *BJU international* 101(1), pp. 1396–400. Available at: http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf.

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

Johnson, N. L. 1949. Systems of frequency curves generated by methods of translation. *Biometrika* 36(Pt. 1-2), pp. 149–176.

Jónsson, T., Pinson, P., Nielsen, H. A., Madsen, H. and Nielsen, T. S. 2013. Forecasting electricity spot prices accounting for wind power predictions. *IEEE Transactions on Sustainable Energy* 4(1), pp. 210–218.

Joskow, P. L. 2008. Lessons Learned from Electricity Market Liberalization. *The Energy Journal* 29(01), pp. 9–42.

Karatzoglou, A., Meyer, D. and Hornik, K. 2006. Support Vector Machines in R. *Journal of Statistical Software* 15(9), p. 28. Available at: http://www.jstatsoft.org/v15/i09/paper.

Karatzoglou, A., Smola, A., Hornik, K. and Zeileis, A. 2004. kernlab - An S4 Package for Kernel Methods in R. *Journal of Statistical Software* 11(9), pp. 1–20. Available at: http://www.jstatsoft.org/v11/i09/paper.

Kuhn, M. 2008. Building Predictive Models in R Using the caret Package. *Journal Of Statistical Software* 28(5), pp. 1–26. Available at: http://www.jstatsoft.org/v28/i05/.

Kwiatkowski, D., Phillips, P. C., Schmidt, P. and 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, pp. 159–178. Available at: http://eprints.whiterose.ac.uk/64933/.

Ljung, G. M. and Box, G. E. P. 1978. On a measure of lack of fit in time series models. *Biometrika* 65(2), pp. 297–303. Available at: http://biomet.oxfordjournals.org/content/65/2/297.short.

Madsen, H. 2008. *Time Series Analysis*. Chapman & Hall.

Mandal, P., Senjyu, T. and Funabashi, T. 2006. Neural networks approach to forecast several hour ahead electricity prices and loads in deregulated market. *Energy Conversion and Management* 47(15-16), pp. 2128–2142.

Marcellino, M., Stock, J. H. and Watson, M. W. 2006. A comparison of direct and iterated multistep AR methods for forecasting macroeconomic time series. *Journal of Econometrics* 135(1-2), pp. 499–526.

Medeiros, M. C., Teräsvirta, T. and Gianluigi, R. 2006. Building Neural Network Models for Time Series: A Statistical Approach. *Journal of Forecasting* 25, pp. 49–75.

Newbold, P. and Granger, C. W. J. 1974. Experience with forecasting univariate time series and the combination of forecasts. *Journal of the Royal Statistical Society. Series A ...* 137(2), pp. 131–165. Available at: http://www.jstor.org/stable/2344546.

Nicolosi, M. 2010. Wind power integration and power system flexibility-An empirical analysis of extreme events in Germany under the new negative price regime. *Energy Policy* 38(11), pp. 7257–7268. Available at: http://dx.doi.org/10.1016/j.enpol.2010.08.002.

Nielsen, H. A., Nielsen, T. S., Joensen, A. K., Madsen, H. and Holst, J. 2000. Tracking time-varying coefficient functions.pdf.

Nogales, F. J., Contreras, J., Conejo, A. J. and Espínola, R. 2002. Forecasting next-day electricity prices by time series models. *IEEE Transactions on Power Systems* 17(2), pp. 342–348.

Opsomer, J., Wang, Y. and Yang, Y. 2001. Nonparametric Regression with Correlated Errors. *Statistical Science* 16(2), pp. 134–153. Available at: http://www.jstor.org/stable/2676791.

Pinson, P., Nielsen, H. A. and Madsen, H. 2007. Robust estimation of time-varying coefficient functions - Application to the modeling of wind power production. *Department of Informatics and Mathematical Modelling, Technical University of Denmark, Tech. Rep* .

R Core Team. 2015. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, Available at: httpwww.R-project.org.

Rumelhart, D. E., Hinton, G. E. and Williams, R. J. 1986. Learning representations by back-propagating errors. *Nature* 323(6088), pp. 533–536.

Russell, S. and Norvig, P. 2010. *Artificial Intelligence: A Modern Approach*. Pearson Education Inc., third ed.

Said, S. E. and Dickey, D. A. 1984. Testing for unit roots in autoregressive moving-average models of unknown order. *Biometrika* 71(3), pp. 599–607. Available at: http://links.jstor.org/sici?sici=0006-3444(198412)71:3<599:TFURIA>2.0.CO;2-Z.

Schneider, S. 2011. Power spot price models with negative prices. *Journal of Energy Markets* 4(4), pp. 77–102. Available at: http://mpra.ub.uni-muenchen.de/29958/.

Schölkopf, B. and Smola, A. J. 2003. A short introduction to learning with kernels. *Advanced Lectures on Machine Learning* pp. 41–64. Available at: http://www.springerlink.com/index/892U8PM7W69K3BHP.pdf.

Shmueli, G. 2010. To explain or to predict? *Statistical Science* 25(3), pp. 289–310.

Stone, M. 1977. An Asymptotic Equivalence of Choice of Model by Cross-validation An Asymptotic Akaike's Criterion. *Journal of the Royal Statistical Society Series B (Methodological)* 39(1), pp. 44–47.

Teräsvirta, T. and Mellin, I. 1986. Model selection criteria and model selection tests in regression models. *Scandinavian Journal of Statistics: Theory and Application* 13(3), p. 159.

The European Commission. 2003. *Directive 2003/54/EC - concerning common rules for the internal market in electricity and repealing Directive 96/92/EC*, [Online].

The European Commission. 2009a. *Regulation (EC) No 713/2009 - Establishing an Agency for the Cooperation of Energy Regulators*, [Online].

The European Commission. 2009b. *Regulation (EC) No 714/2009 - on conditions for access to the network for cross-border exchanges in electricity and repealing Regulation (EC) No 1228/2003*, [Online].

The European Commission. 2013. *Commision Regulation (EU) No 543/2013 - on submission and publication of data in electricity markets and amending Annex I to Regulation (EC) No 714/2009 of the European Parliament and of the Council*, [Online].

Venables, W. N. and Ripley, B. D. 2002. *Modern Applied Statistics with S*. New York: Springer, fourth ed. Available at: http://www.stats.ox.ac.uk/pub/MASS4.

Viehmann, J. 2011. Risk premiums in the German day-ahead Electricity Market. *Energy policy* 39(1), pp. 386–394.

Weare, C. 2003. The California Electricity Crisis: Causes and Policy Options. Tech. rep., Public Policy Institute of California.

Weiss, A. A. 1996. Estimating time series models using the relevant cost function. *Journal of Applied Econometics* 11, pp. 539–560.

Weron, R. 2006. *Modeling and Forecasting Electricity Loads and Prices: A statistical Approach*. John Wiley & Sons Ltd.

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, pp. 1030–1081. Available at: http://dx.doi.org/10.1016/j.ijforecast.2014.08.008.

Weron, R. and Misiorek, A. 2008. Forecasting spot electricity prices: A comparison of parametric and semiparametric time series models. *International Journal of Forecasting* 24(4), pp. 744–763.

White, H. 1990. Connectionist Nonparametric Regression: Multilayer Feedforward Networks Can Learn Arbitrary Mappings. *Neural Networks* 3, pp. 535–549.

White, H. 2006. Approximate nonlinear forecasting methods. In: *Handbook of Economic Forecasting*, Elsevier, Amsterdam, pp. 460–512.

Würzburg, K., Labandeira, X. and Linares, P. 2013. Renewable generation and electricity prices: Taking stock and new evidence for Germany and Austria. *Energy Economics* 40, pp. S159–S171. Available at: http://dx.doi.org/10.1016/j.eneco.2013.09.011.

Yamin, H., Shahidehpour, S. and Li, Z. 2004. Adaptive short-term electricity price forecasting using artificial neural networks in the restructured power markets. *International Journal of Electrical Power & Energy Systems* 26(8), pp. 571–581.

Zhao, J. H., Dong, Z. Y., Xu, Z. and Wong, K. P. 2008. A statistical approach for interval forecasting of the electricity price. *IEEE Transactions on Power Systems* 23(2), pp. 267–276.

Ziel, F., Steinert, R. and Husmann, S. 2015. Forecasting day ahead electricity spot prices: The impact of the EXAA to other European electricity markets. *ArXiv e-prints* pp. 1–17. Available at: http://arxiv.org/abs/1501.00818.

12 Appendix

12.1 The German electricity market

12.1.1 Generation

We can classify the different generation types as either price independent or price dependent.

Figure 33 shows the hourly utilization rates of price insensitive (PI) generation types. From the figure, we see that nuclear power runs at a high utilization rate that is almost independent of the day-ahead price. The utilization rate of biomass generators is also stable at around 40%, completely independent of price. Hydro run-of-river production, though quite variable, is also not correlated with price.

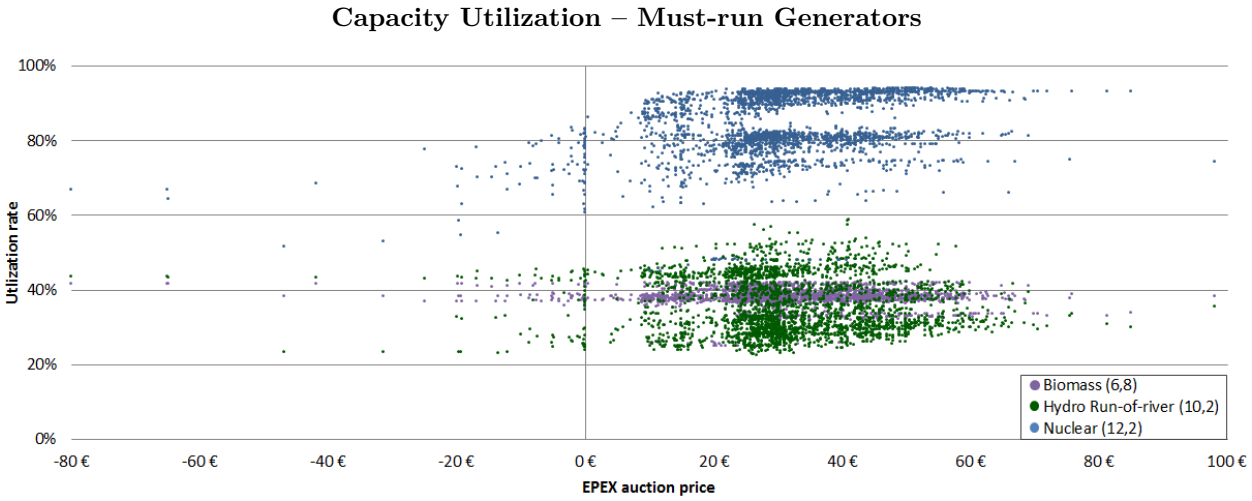

Figure 33 – Hourly capacity utilization rate (production divided by installed capacity from Table 1) plotted against the EPEX Spot price for generation types that show low sensitivity to price (Biomass, Hydro run-of-river and Nuclear). Installed capacity in GW in parenthesis. Using production data for the five German/Austrian TSOs during 1-1-2015 to 30-4-2015. Own calculations with some imputation due to missing data in single TSOs. Source transparency.entsoe.eu.

In Figure 34 we see price dependent (PD) generation types.
These units are typically marginal price setting in the market. For each generation type, we see a clear throttling of production, especially around 20-30 €, which presumably is the marginal cost of many Coal/Lignite/Hydro plants. Interestingly, we also see that hydro storage plants with pumping capacity turn to net consumers in the system, as they pump water into higher reservoirs when prices are low and use the stored water to generate electricity when prices are higher.

For all generation types, the full capacity is never completely utilized. Gas units only produce around 20% of installed capacity, even at high prices. The explanation might be that there are considerable fixed costs in keeping a power plant operational that are not covered by running for a few hours or days at high prices. So, if expected prices are low in general, the owners might not make the plant available in the day-ahead market. Or, maybe the non-operating plants are in the process of decommissioning.

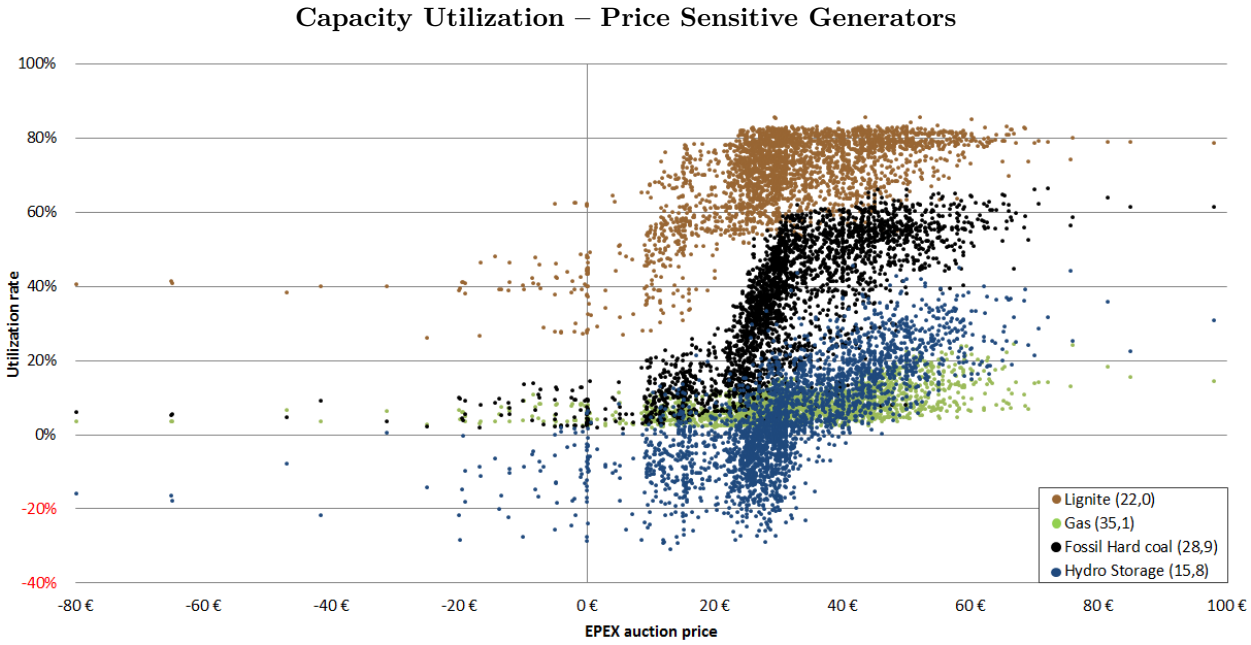

Figure 34 – Hourly capacity utilization rate (production divided by installed capacity from Table 1) plotted against the EPEX auction price for generation types that show high sensitivity to price (Lignite, Hard Coal, Hydro Storage and Gas). Installed capacity in GW in parenthesis. Using production data for the five German/Austrian TSOs from 1-1-2015 to 30-4-2015. Own calculations with some imputations due to missing data in single TSOs. Source transparency.entsoe.eu.

12.1.2 Interconnection

Interconnection capacity with neighboring countries represent another source of electricity or pool of consumers, depending on the relative needs of the importing and exporting power systems. From the gross transmission graph (Figure 35) we see that there have usually been scheduled day-ahead exchanges in both directions during 2014. This indicates that Germany/Austria probably had at least one neighbor with equal or lower prices and one with equal or higher prices.

From the net transmission graph (Figure 36) we see that Germany/Austria was predominantly an exporter of electricity in day-ahead trading in 2014, except for a monthly spell in the summer with net import.

Germany/Austria has interconnection capacity with the following countries: Denmark, Sweden, Poland, Czech Republic, Hungary, Slovenia, Italy, Switzerland, France and the Netherlands.

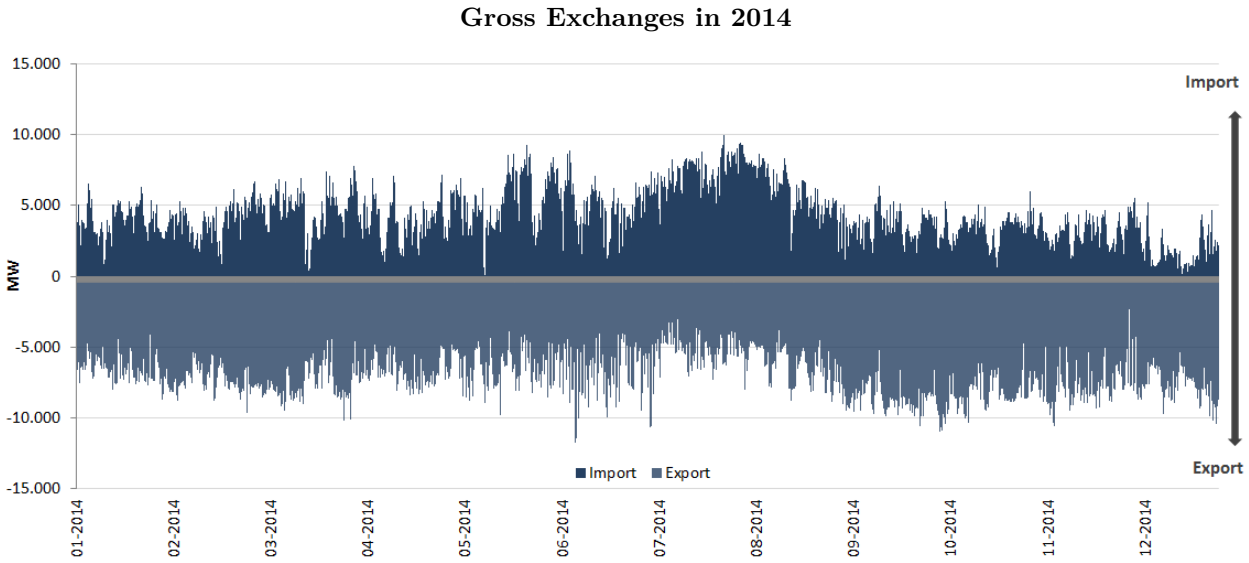

Figure 35 – Gross scheduled day-ahead exchanges between Germany/Austria and neighboring countries in 2014. Source transparency.entsoe.eu.

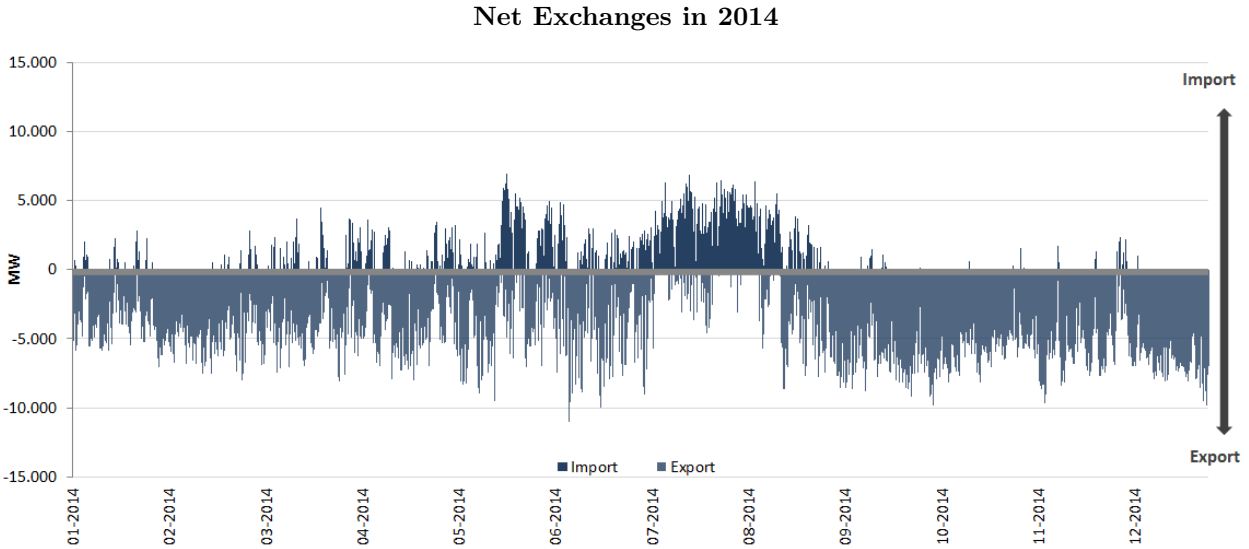

Figure 36 – Net scheduled day-ahead exchanges between Germany/Austria and neighboring countries in 2014. Source transparency.entsoe.eu.

12.1.3 Trading fees

Trade costs of a round-trip trade

Party	Fee (€/MWh)
European Commodity Clearing AG (ECC)[25]	0.015
European Power Exchange (EPEX)[26]	0.040
Energy Exchange Austria (EXAA) - Market Maker[27]	0.025
Energy Exchange Austria (EXAA) - Proprietary Trader[27]	0.075
Total Round-trip Trade Cost	0.095 - 0.145

Table 27 – Marginal trade costs of a round-trip trade between the EPEX Spot exchange and EXAA exchange (EPEX fee + EXAA fee + 2 x ECC fee).

12.2 Estimation

12.2.1 German holidays in 2014

Date	Holiday	Germany	Austria	France
01/01/2014	New Year's Day	1	1	1
18/04/2014	Maundy Friday	1	0	1
21/04/2014	Easter Monday	1	1	1
01/05/2014	Labor Day	1	1	1
29/05/2014	Ascension Day	1	1	1
09/06/2014	Whit Monday	1	1	1
03/10/2014	National Unity Day (DE)	1	0	0
25/12/2014	1st Christmas Day	1	1	1
26/12/2014	2nd Christmas Day	1	1	1

Table 28 – German holidays in 2014. Source: RTE (http://clients.rte-france.com/lang/an/visiteurs/vie/vie_jours_feries.jsp)

[25] Document Release: 11a. from http://www.ecc.de/ecc-en/about-ecc/rules/price-list
[26] Valid from 1. Jan 2015 from https://www.epexspot.com/en/extras/download-center/documentation
[27] Valid from 1. Oct 2014 from http://www.exaa.at/en/rules-docs

12.2.2 Support vector machine

C	RMSE	RMSESD
0.125	8.407	0.323
0.250	8.407	0.323
0.500	8.406	0.322
1.000	8.406	0.322
2.000	8.406	0.322
4.000	8.406	0.322
8.000	8.406	0.322
16.000	8.406	0.322
32.000	8.406	0.322

Table 29 – RMSE and R Squared of 7-fold cross validation of Linear SVM across different values of the tune parameter C, in the test period for model (1).

C	RMSE	RMSESD
0.125	8.120	0.320
0.250	8.007	0.306
0.500	7.909	0.287
1.000	7.821	0.277
2.000	7.767	0.268
4.000	7.724	0.275
8.000	7.682	0.264
16.000	7.655	0.271
32.000	7.625	0.280

Table 30 – RMSE and R Squared of 7-fold cross validation of Radial SVM across different values of the tune parameter C, in the test period for model (1).

Model	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
(1)	5.22	5.56	5.86	5.73	5.93	6.01
(2)	4.96	5.37	5.64	5.49	5.68	5.73
(3)	4.41	4.87	5.10	4.98	5.17	5.28
(4)	4.31	4.71	4.86	4.81	5.00	5.08
(5)	4.48	4.90	5.09	5.01	5.20	5.31
(6)	4.49	4.90	5.07	4.99	5.14	5.27
(7)	4.29	4.71	4.82	4.80	5.01	5.08

Table 31 – RMSE of 7-fold cross validation of Linear SVM in the test period. All models have the cost parameter set equal to 1.

Model	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
(1)	7.38	7.40	7.53	7.62	7.81	8.05
(2)	5.94	6.00	6.24	6.21	6.31	6.63
(3)	5.67	5.83	6.06	6.00	6.10	6.45
(4)	6.44	6.51	6.68	6.70	6.77	7.24
(5)	6.13	6.47	6.79	6.68	6.95	7.00
(6)	7.52	7.59	7.63	7.71	7.80	8.00

Table 32 – RMSE of 7-fold cross validation of Radial SVM in the test period. All models have the cost parameter set equal to 4.

12.2.3 Neural network

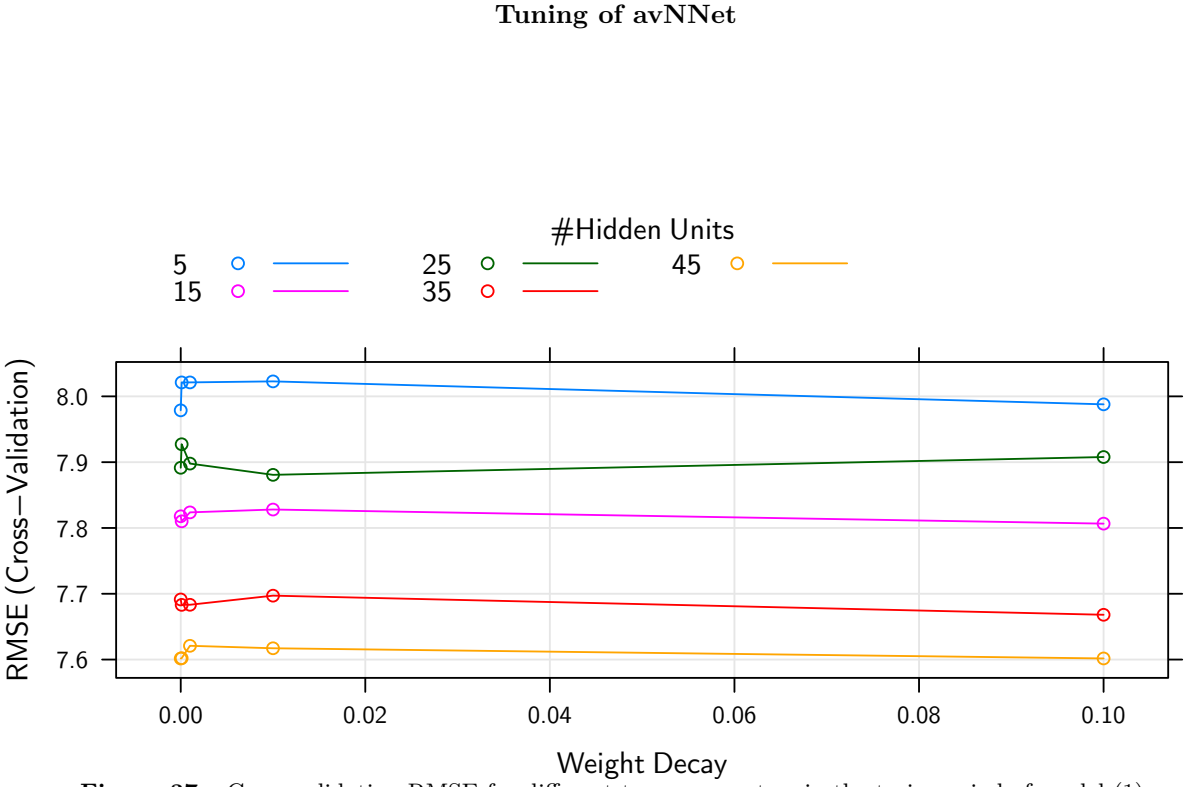

Figure 37 – Cross-validation RMSE for different tune parameters in the train period of model (1).

Model	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
(1)	7.40	7.43	7.63	7.68	7.90	8.05
(2)	6.91	7.12	7.31	7.29	7.47	7.62
(3)	5.22	5.36	5.40	5.44	5.54	5.68
(4)	5.02	5.23	5.34	5.32	5.42	5.56
(5)	5.69	5.88	5.98	5.97	6.06	6.26
(6)	5.77	5.84	6.12	6.07	6.24	6.39
(7)	5.03	5.12	5.23	5.25	5.38	5.52

Table 33 – RMSE and R Squared of 7-fold cross validation of different inputs to avNNET.

12.2.4 Jonsson model code

Not the most neat code I have produced, but it works:

```
#### Jonsson model ####
# Title:        Jonsson model
# Language:     R
# Description:  Do first step of the two-step local linear approximation model from the article
#               "Forecasting electricity spot prices accounting for wind power predictions" by Jonsson et
     al. (2013).
#               Here we fit the model to EPEX/TSO data from 2012-11-01 to 2014-12-28.
#
# Author:       Michael Bülow
# Date:         Spring, 2015

##  Initialization  ##
library(xts)
library(akima)
library(compiler)
enableJIT(2)

ptmS <- Sys.time()

setwd("/home/rstudio/") # On Amazon AWS, AMI: [http://louisaslett.com/].
Sys.setenv(TZ='UTC')

# Read data
d <- read.csv('data/tso-data.csv')
timestamp <- strptime(d[,1], format="%Y-%m-%d %H:%M")
dat <- xts(d[,2:ncol(d)], timestamp)
rm(d, timestamp)

# Set dates
burnin.start <-  as.POSIXct("2012-11-01 00:00")
burnin.end <-    as.POSIXct("2012-12-12 23:00")

train.start <-  as.POSIXct("2012-12-13 00:00")
train.end <-    as.POSIXct("2013-12-31 23:00")

test.start <-   as.POSIXct("2014-01-01 00:00")
test.end <-     as.POSIXct("2014-12-28 23:00")

# Dummy for doing train only else run complete model
isTrain <- TRUE

# Number of fitting points:
u1.fitpoints = 24 #Cons
u2.fitpoints = 24 #Res

###  PREPROCESS  ###

# Make a copy of data, convert to GWh
a <- cbind(dat[,1], dat[,4]/1000, (dat[,6] + dat[,7])/1000, dat[,1])
colnames(a) <- c("PRI_DE","CON_DE","PRO_DE_RES", "Realized Prc")

## Perform censoring of PRI_DE ##
dMinPrc <- 0    # consors 162 obs
dMaxPrc <- 110  # censors 8 obs
a[a[,1] < dMinPrc, 1] <- dMinPrc  # Replace values lower than limit
a[a[,1] > dMaxPrc, 1] <- dMaxPrc  # Replace values higher than limit
# Summary: 106 censored in train, 64 in test

## Perform [-1,1] scaling of explanatory variables using test period range ##
aT <- a[seq(burnin.start, train.end, by=60)] # Pull burn+train from dataset
aMin <- sapply(aT,FUN = min)     # Save train sample Min
aMax <- sapply(aT,FUN = max)     # Save train sample Max
# Scale explanatory variables by min/max of burn+test sample
a[, 2] <- -1 + 2*((a[, 2] - aMin[2])/(aMax[2] - aMin[2])) # Scale variables [-1:1]
a[, 3] <- -1 + 2*((a[, 3] - aMin[3])/(aMax[3] - aMin[3])) # Scale variables [-1:1]
rm(aT, dMinPrc, dMaxPrc, dat)

### END PREPROCESS ###


### INITIALIZE MODEL ###

## Define helper functions ##
triCube <- function(x){as.double(ifelse(x>=0&x<1, (1-x^3)^3, 0)) } # Tricube function
huber <- function(x,t){as.double(sign(x)*min(abs(x),t)) }     # Huber function
delta <- function(x,t){as.double(ifelse(abs(x)<t, 1, 0)) }    # Delta of huber function
euc.dist <- function(x1,x2){as.double(sqrt(sum((x1 - x2)^2))) } # Euclidian distance function
toPoly <- function(u){c(1,u[1],u[2],u[1]^2,u[2]^2,u[1]*u[2])} # Convert 2d point to 6d polynomial

# Function to calculate bandwidth in a fitting point
calc.bw <- function(u, gamma){
  distances <- apply(xt, 1, FUN=function(x){euc.dist(u,x)}) #Requires xt defined below
  as.double(quantile(distances, probs = c(gamma)))
}

## Define fitting points ##
u1 = seq(from = -1, to = 1, length.out = u1.fitpoints)
u2 = seq(from = -1, to = 1, length.out = u2.fitpoints)

## Create matrix of combinations for use in Akima interpolation ##
xy <- expand.grid(u1,u2)

## Intialize lists to hold variables ##
p = matrix(list(), nrow = u1.fitpoints, ncol=u2.fitpoints) # Parameter vector (6*1)
R = matrix(list(), nrow = u1.fitpoints, ncol=u2.fitpoints) # R matrix (6*6)
u = matrix(list(), nrow = u1.fitpoints, ncol=u2.fitpoints) # u fitpoint values (2*1)
bw = matrix(list(), nrow = u1.fitpoints, ncol=u2.fitpoints) # Bandwidth in fitpoints (1*1)
FC = matrix(list(), nrow = u1.fitpoints, ncol=u2.fitpoints) # Forecast at each U point (1*1)

## Initial values for parameters ##
```

```
pInit = c(34, 0, 0, 0, 0, 0)
RInit = diag(6)*10^(-6)

## Load data as coredata ##
if(isTrain){
  # Train
  yt <- coredata(a[seq(burnin.start, train.end, by=60),1]) #Pruned y
  xt <- coredata(a[seq(burnin.start, train.end, by=60),2:3]) #Explanatory vars
  rt <- coredata(a[seq(burnin.start, train.end, by=60),4]) #Realized y
} else {
  # Test
  yt <- coredata(a[seq(burnin.start, test.end, by=60),1]) #Pruned y
  xt <- coredata(a[seq(burnin.start, test.end, by=60),2:3]) #Explanatory vars
  rt <- coredata(a[seq(burnin.start, test.end, by=60),4]) #Realized y
}
### END INITIALIZE ###

### Jonsson et al. model ###
jonsson.model <- function(param, printInfo){
  # Set tune parameters
  gamma = param[1]
  lambda = param[2]
  tau   = param[3]

  if(printInfo){
    cat("\nGamma: ", gamma, sep="")
    cat("\nLambda :", lambda, sep="")
    cat("\nTau: ", tau, sep="")
    cat('\nStartAt: ', format.POSIXct(Sys.time(), format = "%y-%m-%d %H:%M:%S", tz="Europe/Berlin"), sep=
        "")
  }

  ## INITIALIZE VARS ##

  # dataframes to hold forecast/actual
  forc <- data.frame(forecast=matrix(NA, nrow = nrow(xt), ncol = 1))
  actual <- data.frame(actual=matrix(NA, nrow = nrow(xt), ncol = 1))

  # Fill initial values to all fitting points
  for(cVar2 in 1:u2.fitpoints){
    for(cVar1 in 1:u1.fitpoints){
      p[[cVar1,cVar2]] <- pInit
      R[[cVar1,cVar2]] <- RInit
      u[[cVar1,cVar2]] <- c(u1[cVar1], u2[cVar2])
      bw[[cVar1,cVar2]] <- calc.bw(u[[cVar1,cVar2]], gamma)
    }
  }
  ## END INITIALIZE VARS ##

  ## START Loop ##
  for(i in 0:(nrow(yt)/24 - 1)){ # Loop over each day

    # Report status
    if(printInfo){
      if(i%%100==0){
        cat("\n", formatC(i+1,width=3,format="d", flag="0"), "/", nrow(yt)/24,"..", sep="")
        ptmD <- Sys.time()
      } else if (i%%10==0){
        cat("|", sep="")
      } else {
        cat("-", sep="")
      }
    }

    # Create grid of forecasts
    for(cVar2 in 1:u2.fitpoints){
      for(cVar1 in 1:u1.fitpoints){
        FC[[cVar1,cVar2]] <- toPoly(u[[cVar1,cVar2]]) %*% p[[cVar1,cVar2]]
      }
    }

    # create vector form of forecast grid for use in Akima.
    zz <- unlist(FC)

    # Create 24 hourly forecasts by interpolation from forecast grid
    for(j in 1:24){
      #Extract explanatory variables
      point <- xt[i*24 + j, ]

      #Do interpolation using akima package.
      # To do interpolation beyond the grid we need to use extrap=TRUE, which is broken in the standard
          interp() function.
      # To get beyond this we use the old.interp() function with the deprecated ncp=2 parameter.
      singleForecast <- interp.old(xy[,1], xy[,2], zz, xo = point[1], yo= point[2], ncp=2, extrap=TRUE)

      # Save forecast
      forc[24*i + j,] <- c(singleForecast$z[1])
      # Save actual
      actual[24*i + j,] <- rt[i*24 + j, ]
    }
    ## End forecast ##

    ## Start Update of parameters ##
    for(j in 1:24){
      #Extract explanatory variables
      point <- as.vector(xt[i*24 + j, ])

      # Is observation in burn sample -> TRUE/FALSE
      isBurn <- ifelse((24*i + j)<=as.double((train.start-burnin.start)*24), TRUE, FALSE)

      #Visit each fitting point and update parameters
      for(cVar2 in 1:u2.fitpoints){
        for(cVar1 in 1:u1.fitpoints){
          # Calculate error using parameters in the fitting point
```

```
          eps <- yt[24*i + j] - toPoly(point) %*% p[[cVar1,cVar2]]

          # Calculate weight assigned to current observation in fitting point
          weight <- triCube(euc.dist(u[[cVar1,cVar2]], point) / bw[[cVar1,cVar2]])

          # Calculate lambdaEff [Eq. 12] - but allow larger updates in burn
          lambdaEff <- (1 - (1-lambda) * weight * ifelse(isBurn, 1, delta(eps, tau)))

          # Update R matrix - allow larger updates in burn
          R[[cVar1,cVar2]] <- lambdaEff * R[[cVar1,cVar2]] + weight * ifelse(isBurn, 1, delta(eps, tau))
              * toPoly(point) %*% t(toPoly(point))

          #Update p.
          #Dont update p for the first 100 observations - can give problem with oscillating p.
          if((24*i + j)>100){
            p[[cVar1,cVar2]] <- p[[cVar1,cVar2]] + weight * huber(eps, tau) * as.vector(solve(R[[cVar1,
                cVar2]]) %*% toPoly(point))
          }
        }
      }#next fitting point
    }#next hour
    ## End Update parameters ##
  }#next day

  ## Handle results ## - Not neat coding but works.
  # Write forecast and actual series
  if(isTrain){
    timestamp <- as.POSIXct(index(a[seq(burnin.start, train.end, by=60)]), format="%d/%m/%y %H:%M") #if
        train
    xtsForecast <- xts(cbind(forc,actual), timestamp)[seq(train.start, train.end, by=60)]
  } else {
    timestamp <- as.POSIXct(index(a[seq(burnin.start, test.end, by=60)]), format="%d/%m/%y %H:%M") #if
        test
    xtsForecast <- xts(cbind(forc,actual), timestamp)[seq(train.start, test.end, by=60)]
  }

  # Write above results including error measures
  summaryForecast <- cbind(xtsForecast, abs(xtsForecast[,2] - xtsForecast[,1]), (xtsForecast[,2] -
      xtsForecast[,1])^2)
  names(summaryForecast) <- c('forecast','actual','abs.error','sq.error')
  # Exclude christmas 2012.
  summaryForecast['2012-12-24/25',3:4] <- NA
  write.csv(as.data.frame(summaryForecast), file='jonsson-model-step1.csv')

  # Format output to function
  returnval <- sapply(summaryForecast, FUN = mean, na.rm=TRUE)[3:4]
  returnval[2] <- sqrt(returnval[2])
  names(returnval) <- c('MAE','RMSE')

  #return(returnval) # When running loop and MAE is of interest
  return(returnval[2]) #When running BFGS optimization
  ## END MODEL ##
}


##########################################
### Code to execute BFGS Optimization ####
##########################################

# Original values from Jonnson paper (on DK1 data)
# gamma   = 0.8529
# lambda  = 0.9877
# tau     = 55.67/7.45 = 7.47

# RMSE (train): 51.10/7.45 = 6.86
# RMSE (test): 50.89/7.45 = 6.83

gauss <- function(x){(1 + exp(-x))^(-1)}
gauss.inv <- function(x){-log((1/x)-1)}
bfgs.transform <- function(x){1/2*(1 + gauss(x))}
bfgs.transform.inv <- function(x){gauss.inv(x*2-1)}

jonsson.model.bfgs <- function(par){
  trpar <- matrix(nrow = 3, ncol = 1)
  trpar[1] <- bfgs.transform(par[1]) # Restricts parameter to be in [0.5,1]
  trpar[2] <- bfgs.transform(par[2]) # Restricts parameter to be in [0.5,1]
  trpar[3] <- exp(par[3]) # Restricts parameter to be in [0,+inf]
  jonsson.model(trpar, TRUE) #Call model
}

# The BFGS method converges when using a special parameterization such that
# par[1:2] can fall in region [0.5,1] and par[3] can fall in region [0, +inf].
# This way BFGS can only optimize across the allowed range of parameter values.
# As start values we use optimal values from jonsson paper (converted to model equivalent),
# eg: bfgs.transform.inv(0.9877)=3.680101
bfgs.optim <- optim(par=c(0.8750721,3.680101, 2.0109),
                    fn=jonsson.model.bfgs,
                    method="BFGS",
                    control=list(trace=TRUE))

# Convergence in 13 function calls, 6 gradient calls...
#           Gamma     Lambda    Tau
# Start Val 0.8750721 3.680101  2.0109
# equals... 0.8529    0.9877    7.47

# Output:   1.025308  3.247043 2.545327
# equals... 0.8680026 0.9812834 12.7474

# Final RMSE (train): 8.543875

########################################
### Code to execute grid optimization ###
########################################
## Following code loops through all combinations of tune parameters
## and produces dataframe with RMSE for all combinations.
```

```
#Final Model:
gamma = c(0.8680026)
lambda = c(0.9812834)
tau = c(12.7474)

# Expand tune parameters to matrix of all combinations
glt <- expand.grid(gamma, lambda, tau)
names(glt) <- c("gamma", "lambda", "tau")
# Create container holding results
container <- cbind(glt, matrix(nrow = nrow(glt), ncol=3))
names(container) <- c("gamma", "lambda", "tau", "MAE", "RMSE","Time Spent")

# Loop across all tune combinations
for(i in 1:nrow(glt)){
  # Write model no.
  cat("\n","Model ", i, "/", nrow(glt), " (grid: ", u1.fitpoints, "x", u1.fitpoints, ", gamma: ", glt[i
      ,1], ", lambda: ", glt[i,2], ", tau: ", glt[i,3], ")", sep="")

  # Save time
  ptmA <- Sys.time()

  # Do model
  container[i,4:5] <- jonsson.model(c(glt[i,1], glt[i,2], glt[i,3]), printInfo=TRUE)

  # Write time spent
  container[i,6] <- as.numeric(Sys.time() - ptmA, units='secs')
  cat("\n", "Run Time: ", container[i,6], '\n',sep="")
  write.csv(as.data.frame(container), file='output.csv')
}
# Print results
container
cat("\n", "Run Time: ", as.numeric(Sys.time() - ptmS, units='mins'), ' minutes',sep="")
```

12.3 Data

12.3.1 TSOs

List of TSOs

TSO	Short Name	Market	Link
Amprion GmbH	Amprion	German	www.amprion.net
TenneT TSO GmbH	Tennet	German	www.tennettso.de
50Hertz Transmission GmbH	50Hz	German	www.50hertz.com
TransnetBW GmbH	Transnet	German	www.transnetbw.com
Austrian Power Grid AG	APG	Austrian	www.apg.at
RTE EDF Transport SA	RTE	French	www.rte-france.com
Swissgrid AG	SWG	Swiss	www.swissgrid.ch

Table 34 – List of TSO names, countries, and links.

12.3.2 Data sources

Data Sources

Source	Data	Type	Link
50Hz	Solar	F	www.50hertz.com/en/Grid-Data/Photovoltaics/Archive-Photovoltaics
Amprion	Solar	F	www.amprion.net/en/photovoltaic-infeed
Tennet	Solar	F	www.tennettso.de/site/en/Transparency/publications/network-figures/actual-and-forecast-photovoltaic-energy-feed-in_land
TransnetBW	Solar	F	www.transnetbw.com/en/key-figures/renewable-energies/photovoltaic
EEX	Solar	F	www.eex-transparency.com/homepage/power/germany/production/usage/expected-solar-power-generation-
50Hz	Wind	F	www.50hertz.com/en/Grid-Data/Wind-power/Archive-Wind-power
Amprion	Wind	F	www.amprion.net/en/wind-feed-in
Tennet	Wind	F	www.tennettso.de/site/en/Transparency/publications/network-figures/actual-and-forecast-wind-energy-feed-in
TransnetBW	Wind	F	www.transnetbw.com/en/key-figures/renewable-energies/wind-infeed
APG	Wind	F	www.apg.at/en/market/generation/wind-energy-forecast
EEX	Wind	F	www.eex-transparency.com/homepage/power/germany/production/usage/expected-wind-power-generation-
APG	Load	F	www.apg.at/en/market/load/load-forecast
ENTSOE	Load	F	transparency.entsoe.eu/content/static_content/Static%20content/legacy%20data/year%20selection.html
RTE	Load	F	clients.rte-france.com/lang/an/visiteurs/vie/vie_histo_courbe_j_plus_2.jsp
ENTSOE	Load	A	www.entsoe.eu/db-query/consumption/monthly-consumption-of-a-specific-country-for-a-specific-range-of-time
EPEX Spot	Price	A	www.epexspot.com/en/market-data/dayaheadauction
EXAA	Price	A	www.exaa.at/en/marketdata/historical-data

Table 35 – Type: A=actual, F=forecast. Links were valid primo 2015.

12.3.3 Summer dummy

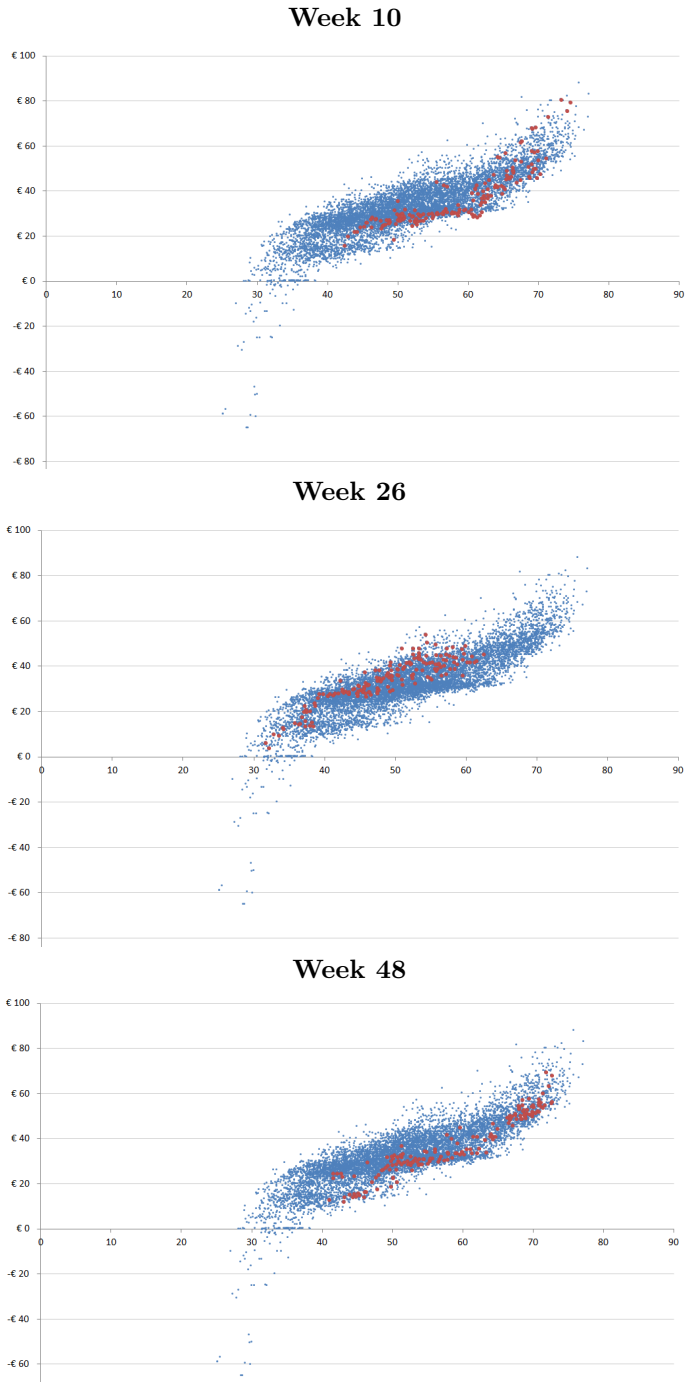

Figure 38 – EPEX day-ahead prices plotted against residual demand for a spring week (Week 10), summer week (Week 26) and winter week (Week 48) in 2014. Red dots represent the specific week. Blue dots represent all hours in 2014. Residual demand is here defined as $ResidualDemand = Consumption_{forecast} - Solar_{forecast} - Wind_{forecast} + ScheduledDayAheadNetExport$. Source: transparency.entsoe.eu and EPEX Spot.

In Figure 38, we see that for a given residual demand (defined in the figure text), the market clearing prices seem higher in the summer weeks than in the winter/spring weeks pictured. This pattern also seems present

in other weeks. It is hypothesized that this could be caused by a reduction in available generation capacity during summer months and that it is relevant to include this when modelling.

12.4 Empirical Results

12.4.1 MAE

Week	mcb	naive	arima	linear	linear.r	linear.s	linear.sr	arimax	jonsson.s1	jonsson.s2	svmlinear	svmradial	avnnet
1	3.45	9.12	9.83	6.72	6.47	7.17	6.56	5.44	5.03	4.92	5.30	5.61	5.08
2	4.22	8.96	9.69	4.90	4.34	4.17	4.48	4.56	4.71	4.83	4.32	4.50	4.60
3	2.94	6.32	6.63	5.34	5.18	5.97	5.15	4.59	4.77	4.39	4.13	3.65	4.33
4	3.57	5.60	7.09	3.67	3.51	4.70	3.47	3.88	4.78	4.44	3.87	4.14	4.36
5	3.11	6.36	4.89	3.40	3.32	4.38	3.49	3.45	4.04	3.87	3.25	3.12	3.23
6	3.49	6.02	6.41	4.31	4.71	4.99	4.94	4.60	4.24	4.35	5.68	4.02	4.66
7	4.20	8.60	7.01	3.80	4.13	4.20	4.06	3.97	4.24	3.97	4.39	4.14	3.85
8	3.00	7.37	5.80	3.88	4.28	4.46	3.99	4.10	3.81	3.56	4.05	4.00	4.33
9	2.77	5.36	5.12	3.67	3.87	4.07	3.57	3.77	3.45	3.42	3.45	3.79	3.60
10	3.19	7.43	5.37	4.87	4.35	4.86	4.47	3.92	3.63	3.43	4.80	4.47	5.02
11	3.98	7.85	8.53	4.79	4.29	5.52	4.63	4.20	4.57	4.46	4.79	4.72	4.45
12	3.05	13.05	8.02	4.20	3.13	5.25	3.02	3.08	3.34	3.40	5.04	3.29	3.71
13	2.50	5.14	4.85	6.12	5.46	6.34	5.01	5.08	4.18	4.07	5.23	3.79	3.94
14	2.85	3.82	3.98	3.85	3.75	5.41	4.09	4.05	3.07	3.03	3.13	4.14	4.51
15	2.83	7.38	7.35	3.89	3.33	4.74	3.27	3.19	3.40	3.11	3.53	3.16	3.85
16	3.16	6.43	5.62	5.73	4.54	6.90	4.44	4.08	3.83	3.89	4.30	3.73	3.90
17	2.67	7.06	5.01	4.28	2.93	6.12	2.82	2.79	2.52	2.39	2.48	3.06	2.91
18	2.33	7.13	5.59	3.16	3.21	3.92	3.32	3.22	4.03	3.81	3.10	2.97	3.56
19	3.74	7.41	6.64	4.40	4.08	5.79	4.97	3.74	3.66	3.67	4.75	5.79	5.37
20	2.18	9.99	5.11	3.38	2.87	3.51	2.60	2.70	2.63	3.10	2.96	3.11	3.15
21	2.21	4.76	4.27	4.99	3.27	5.44	2.86	2.43	3.28	2.82	2.75	3.42	4.76
22	2.98	5.35	4.27	7.60	5.36	7.94	5.38	3.46	3.93	3.91	4.26	4.07	4.10
23	2.78	3.58	3.34	5.00	4.72	5.18	4.87	3.39	3.35	3.23	3.45	3.74	3.87
24	2.68	5.48	4.86	5.37	3.78	5.41	4.22	2.77	3.00	2.80	3.23	2.59	3.24
25	1.78	4.81	3.29	6.48	3.67	8.03	4.21	2.62	2.99	2.79	4.00	3.86	3.72
26	2.15	5.42	3.71	5.63	3.41	6.01	3.73	2.86	2.49	2.29	2.46	2.36	2.50
27	1.65	2.64	3.25	3.66	3.38	3.38	2.99	4.05	2.01	1.93	2.49	2.26	2.62
28	1.70	2.53	3.26	6.34	5.70	6.80	6.31	4.91	2.83	2.62	4.14	3.96	4.21
29	1.61	2.23	3.34	4.36	3.35	4.26	3.04	3.15	2.10	2.02	3.08	3.22	3.95
30	1.47	3.21	3.47	4.27	2.90	3.97	2.80	2.84	1.95	1.87	2.11	2.10	2.37
31	2.97	4.45	3.96	8.28	7.20	8.37	7.09	4.54	3.70	3.27	6.49	5.59	5.87
32	2.16	3.94	4.03	8.59	7.70	8.91	7.98	3.64	2.15	2.16	7.24	7.39	8.47
33	3.62	7.89	7.20	6.16	5.57	8.08	6.51	4.71	4.17	4.04	5.04	5.25	4.85
34	2.38	7.83	4.26	4.90	3.66	6.05	4.05	4.28	2.95	2.86	2.44	3.52	3.02
35	1.64	3.54	3.26	8.96	2.47	9.68	3.30	2.76	2.65	2.58	2.25	3.79	4.90
36	1.88	3.52	3.72	9.08	7.92	8.98	8.19	3.49	2.77	2.64	8.20	8.86	9.32
37	1.68	2.51	2.67	8.50	5.03	8.38	4.97	2.85	2.45	2.06	4.99	4.47	3.96
38	2.75	4.80	4.41	8.00	4.12	7.90	3.96	4.84	3.70	3.32	3.25	2.75	2.88
39	2.21	6.54	4.60	8.69	3.24	8.45	3.22	4.32	3.00	2.69	2.61	2.52	2.91
40	2.47	5.91	6.19	5.65	3.25	5.55	3.17	5.51	3.70	3.64	3.32	4.21	4.96
41	2.91	5.96	6.00	5.38	3.69	4.55	3.86	4.08	3.79	3.51	3.52	3.57	3.45
42	2.27	5.41	5.39	7.33	4.34	6.89	3.92	3.26	4.31	3.76	3.97	5.29	5.28
43	2.79	10.68	6.45	9.06	3.98	8.64	3.99	4.72	3.00	2.79	3.43	3.36	3.53
44	3.49	8.05	6.83	9.16	5.58	9.32	5.69	4.44	4.70	4.38	4.98	4.37	4.39
45	3.11	8.87	6.99	4.40	4.21	4.41	3.96	5.51	4.62	4.65	3.96	3.37	3.35
46	1.80	5.65	3.91	3.44	2.63	3.83	2.83	3.53	4.26	4.18	2.22	3.51	3.75
47	1.78	3.99	4.84	3.65	3.53	3.52	3.47	4.37	3.69	3.83	2.48	3.24	3.73
48	2.31	5.94	5.65	3.69	3.49	4.02	3.29	3.90	3.10	3.08	3.71	4.11	3.83
49	2.20	6.54	7.49	2.88	3.83	3.63	4.62	3.69	2.88	2.93	3.46	5.05	5.78
50	2.68	9.05	8.95	4.51	3.26	4.78	3.73	3.31	3.57	3.63	3.23	5.41	5.45
51	3.02	13.37	8.46	5.56	5.07	6.30	5.78	4.77	3.70	3.66	4.06	4.71	4.79
52	4.71	14.45	12.81	11.19	4.93	12.84	5.60	8.04	4.60	4.20	4.05	4.41	4.52
	2.71	6.43	5.62	5.55	4.22	5.99	4.33	3.94	3.52	3.38	3.90	4.02	4.24

Table 36 – Weekly Mean Absolute Errors (wMAE).

Week	mcb	naive	arima	linear	linear.r	linear.s	linear.sr	arimax	jonsson.s1	jonsson.s2	svmlinear	svmradial	avnnet
1	1.00	2.64	2.85	1.95	1.88	2.08	1.90	1.58	1.46	1.43	1.54	1.63	1.47
2	1.00	2.12	2.30	1.16	1.03	0.99	1.06	1.08	1.12	1.14	1.02	1.07	1.09
3	1.00	2.15	2.26	1.82	1.76	2.03	1.75	1.56	1.62	1.50	1.41	1.24	1.47
4	1.00	1.57	1.99	1.03	0.98	1.32	0.97	1.09	1.34	1.24	1.09	1.16	1.22
5	1.00	2.05	1.57	1.09	1.07	1.41	1.12	1.11	1.30	1.25	1.05	1.00	1.04
6	1.00	1.72	1.83	1.23	1.35	1.43	1.41	1.32	1.21	1.24	1.62	1.15	1.33
7	1.00	2.05	1.67	0.90	0.98	1.00	0.97	0.95	1.01	0.94	1.04	0.99	0.92
8	1.00	2.45	1.93	1.29	1.42	1.48	1.33	1.37	1.27	1.18	1.35	1.33	1.44
9	1.00	1.94	1.85	1.33	1.40	1.47	1.29	1.36	1.25	1.24	1.25	1.37	1.30
10	1.00	2.33	1.69	1.53	1.37	1.53	1.40	1.23	1.14	1.08	1.51	1.40	1.58
11	1.00	1.97	2.15	1.20	1.08	1.39	1.16	1.06	1.15	1.12	1.20	1.19	1.12
12	1.00	4.28	2.63	1.38	1.03	1.72	0.99	1.01	1.10	1.12	1.66	1.08	1.22
13	1.00	2.06	1.94	2.45	2.19	2.54	2.01	2.03	1.67	1.63	2.09	1.52	1.58
14	1.00	1.34	1.39	1.35	1.31	1.90	1.43	1.42	1.08	1.06	1.10	1.45	1.58
15	1.00	2.61	2.60	1.38	1.18	1.68	1.16	1.13	1.20	1.10	1.25	1.12	1.36
16	1.00	2.04	1.78	1.81	1.44	2.19	1.41	1.29	1.21	1.23	1.36	1.18	1.24
17	1.00	2.64	1.88	1.60	1.10	2.29	1.06	1.04	0.94	0.90	0.93	1.15	1.09
18	1.00	3.06	2.40	1.36	1.38	1.68	1.43	1.38	1.73	1.64	1.33	1.27	1.53
19	1.00	1.98	1.78	1.18	1.09	1.55	1.33	1.00	0.98	0.98	1.27	1.55	1.44
20	1.00	4.58	2.34	1.55	1.32	1.61	1.19	1.24	1.21	1.42	1.36	1.43	1.45
21	1.00	2.16	1.93	2.26	1.48	2.46	1.29	1.10	1.49	1.28	1.24	1.55	2.15
22	1.00	1.79	1.43	2.55	1.80	2.66	1.80	1.16	1.32	1.31	1.43	1.36	1.37
23	1.00	1.29	1.20	1.80	1.70	1.86	1.75	1.22	1.20	1.16	1.24	1.34	1.39
24	1.00	2.05	1.82	2.00	1.41	2.02	1.58	1.03	1.12	1.05	1.21	0.97	1.21
25	1.00	2.71	1.85	3.64	2.06	4.52	2.37	1.47	1.68	1.57	2.25	2.17	2.09
26	1.00	2.53	1.73	2.62	1.59	2.80	1.74	1.33	1.16	1.07	1.15	1.10	1.16
27	1.00	1.59	1.97	2.22	2.05	2.04	1.81	2.45	1.21	1.17	1.50	1.36	1.58
28	1.00	1.49	1.92	3.73	3.35	4.01	3.72	2.89	1.66	1.54	2.44	2.33	2.48
29	1.00	1.39	2.08	2.71	2.09	2.65	1.89	1.96	1.31	1.26	1.92	2.01	2.46
30	1.00	2.18	2.36	2.90	1.97	2.70	1.90	1.93	1.33	1.27	1.43	1.43	1.61
31	1.00	1.50	1.33	2.78	2.42	2.81	2.38	1.53	1.24	1.10	2.18	1.88	1.97
32	1.00	1.83	1.87	3.98	3.57	4.13	3.70	1.69	1.00	1.00	3.36	3.43	3.93
33	1.00	2.18	1.99	1.70	1.54	2.23	1.80	1.30	1.15	1.11	1.39	1.45	1.34
34	1.00	3.28	1.78	2.05	1.53	2.54	1.70	1.80	1.24	1.20	1.03	1.48	1.26
35	1.00	2.16	1.99	5.45	1.50	5.89	2.01	1.68	1.61	1.57	1.37	2.31	2.98
36	1.00	1.87	1.98	4.83	4.21	4.78	4.36	1.86	1.47	1.41	4.36	4.71	4.96
37	1.00	1.49	1.58	5.05	2.99	4.97	2.95	1.69	1.45	1.22	2.97	2.65	2.35
38	1.00	1.75	1.61	2.91	1.50	2.88	1.44	1.76	1.35	1.21	1.18	1.00	1.05
39	1.00	2.96	2.08	3.93	1.47	3.82	1.46	1.95	1.36	1.22	1.18	1.14	1.31
40	1.00	2.39	2.50	2.28	1.31	2.24	1.28	2.23	1.50	1.47	1.34	1.70	2.00
41	1.00	2.05	2.06	1.85	1.27	1.56	1.32	1.40	1.30	1.21	1.21	1.23	1.18
42	1.00	2.38	2.37	3.23	1.91	3.03	1.73	1.43	1.90	1.65	1.75	2.33	2.32
43	1.00	3.83	2.31	3.25	1.43	3.10	1.43	1.69	1.07	1.00	1.23	1.21	1.27
44	1.00	2.31	1.96	2.62	1.60	2.67	1.63	1.27	1.35	1.25	1.43	1.25	1.26
45	1.00	2.85	2.25	1.42	1.35	1.42	1.27	1.77	1.49	1.50	1.28	1.08	1.08
46	1.00	3.14	2.17	1.91	1.46	2.13	1.57	1.96	2.37	2.32	1.24	1.95	2.09
47	1.00	2.24	2.72	2.05	1.98	1.98	1.95	2.46	2.08	2.16	1.40	1.82	2.10
48	1.00	2.57	2.45	1.60	1.51	1.74	1.43	1.69	1.34	1.33	1.61	1.78	1.66
49	1.00	2.97	3.40	1.31	1.74	1.65	2.10	1.68	1.30	1.33	1.57	2.29	2.62
50	1.00	3.38	3.34	1.68	1.22	1.78	1.39	1.24	1.33	1.36	1.21	2.02	2.03
51	1.00	4.43	2.80	1.84	1.68	2.09	1.92	1.58	1.23	1.21	1.34	1.56	1.59
52	1.00	3.07	2.72	2.38	1.05	2.72	1.19	1.71	0.98	0.89	0.86	0.94	0.96
	1.00	2.38	2.08	2.05	1.56	2.21	1.60	1.46	1.30	1.25	1.44	1.48	1.57

Table 37 – Weekly Relative Mean Absolute Errors (wRMAE).

Week	mcb	arimax	jonsson.s2	svmlinear	svmradial	avnnet	CF1	CF2
1	3.45	5.44	4.92	5.30	5.61	5.08	4.45	4.51
2	4.22	4.56	4.83	4.32	4.50	4.60	4.19	4.21
3	2.94	4.59	4.39	4.13	3.65	4.33	3.85	3.98
4	3.57	3.88	4.44	3.87	4.14	4.36	3.86	3.90
5	3.11	3.45	3.87	3.25	3.12	3.23	2.92	3.26
6	3.49	4.60	4.35	5.68	4.02	4.66	4.41	3.97
7	4.20	3.97	3.97	4.39	4.14	3.85	3.56	3.69
8	3.00	4.10	3.56	4.05	4.00	4.33	3.79	3.47
9	2.77	3.77	3.42	3.45	3.79	3.60	3.34	3.09
10	3.19	3.92	3.43	4.80	4.47	5.02	4.17	3.36
11	3.98	4.20	4.46	4.79	4.72	4.45	3.95	3.97
12	3.05	3.08	3.40	5.04	3.29	3.71	3.14	3.19
13	2.50	5.08	4.07	5.23	3.79	3.94	4.18	3.92
14	2.85	4.05	3.03	3.13	4.14	4.51	3.47	2.85
15	2.83	3.19	3.11	3.53	3.16	3.85	3.05	2.68
16	3.16	4.08	3.89	4.30	3.73	3.90	3.59	3.45
17	2.67	2.79	2.39	2.48	3.06	2.91	2.35	2.26
18	2.33	3.22	3.81	3.10	2.97	3.56	2.74	3.13
19	3.74	3.74	3.67	4.75	5.79	5.37	4.41	3.80
20	2.18	2.70	3.10	2.96	3.11	3.15	2.61	2.67
21	2.21	2.43	2.82	2.75	3.42	4.76	3.02	2.82
22	2.98	3.46	3.91	4.26	4.07	4.10	3.69	3.37
23	2.78	3.39	3.23	3.45	3.74	3.87	3.40	2.79
24	2.68	2.77	2.80	3.23	2.59	3.24	2.37	2.23
25	1.78	2.62	2.79	4.00	3.86	3.72	2.86	2.28
26	2.15	2.86	2.29	2.46	2.36	2.50	1.79	1.83
27	1.65	4.05	1.93	2.49	2.26	2.62	2.64	2.19
28	1.70	4.91	2.62	4.14	3.96	4.21	3.88	3.15
29	1.61	3.15	2.02	3.08	3.22	3.95	3.12	2.01
30	1.47	2.84	1.87	2.11	2.10	2.37	1.72	1.70
31	2.97	4.54	3.27	6.49	5.59	5.87	5.55	4.24
32	2.16	3.64	2.16	7.24	7.39	8.47	6.44	3.64
33	3.62	4.71	4.04	5.04	5.25	4.85	4.25	3.69
34	2.38	4.28	2.86	2.44	3.52	3.02	2.61	2.58
35	1.64	2.76	2.58	2.25	3.79	4.90	2.49	2.05
36	1.88	3.49	2.64	8.20	8.86	9.32	7.02	3.72
37	1.68	2.85	2.06	4.99	4.47	3.96	3.37	2.21
38	2.75	4.84	3.32	3.25	2.75	2.88	2.84	2.90
39	2.21	4.32	2.69	2.61	2.52	2.91	2.23	2.12
40	2.47	5.51	3.64	3.32	4.21	4.96	3.50	3.34
41	2.91	4.08	3.51	3.52	3.57	3.45	3.12	3.20
42	2.27	3.26	3.76	3.97	5.29	5.28	3.83	3.65
43	2.79	4.72	2.79	3.43	3.36	3.53	2.93	2.67
44	3.49	4.44	4.38	4.98	4.37	4.39	3.78	3.98
45	3.11	5.51	4.65	3.96	3.37	3.35	3.31	3.82
46	1.80	3.53	4.18	2.22	3.51	3.75	2.82	3.33
47	1.78	4.37	3.83	2.48	3.24	3.73	2.92	3.18
48	2.31	3.90	3.08	3.71	4.11	3.83	3.71	3.13
49	2.20	3.69	2.93	3.46	5.05	5.78	3.57	2.71
50	2.68	3.31	3.63	3.23	5.41	5.45	3.36	3.23
51	3.02	4.77	3.66	4.06	4.71	4.79	3.80	3.54
52	4.71	8.04	4.20	4.05	4.41	4.52	3.95	3.70
	2.71	3.94	3.38	3.90	4.02	4.24	3.49	3.15

Table 38 – Weekly Mean Absolute Errors (wMAE). CF1 and CF2 are combined forecast defined in Equation 38 & 39

Week	mcb	arimax	jonsson.s2	svmlinear	svmradial	avnnet	CF1	CF2
1	1.00	1.58	1.43	1.54	1.63	1.47	1.29	1.31
2	1.00	1.08	1.14	1.02	1.07	1.09	0.99	1.00
3	1.00	1.56	1.50	1.41	1.24	1.47	1.31	1.36
4	1.00	1.09	1.24	1.09	1.16	1.22	1.08	1.09
5	1.00	1.11	1.25	1.05	1.00	1.04	0.94	1.05
6	1.00	1.32	1.24	1.62	1.15	1.33	1.26	1.14
7	1.00	0.95	0.94	1.04	0.99	0.92	0.85	0.88
8	1.00	1.37	1.18	1.35	1.33	1.44	1.26	1.16
9	1.00	1.36	1.24	1.25	1.37	1.30	1.21	1.12
10	1.00	1.23	1.08	1.51	1.40	1.58	1.31	1.06
11	1.00	1.06	1.12	1.20	1.19	1.12	0.99	1.00
12	1.00	1.01	1.12	1.66	1.08	1.22	1.03	1.05
13	1.00	2.03	1.63	2.09	1.52	1.58	1.68	1.57
14	1.00	1.42	1.06	1.10	1.45	1.58	1.22	1.00
15	1.00	1.13	1.10	1.25	1.12	1.36	1.08	0.95
16	1.00	1.29	1.23	1.36	1.18	1.24	1.14	1.09
17	1.00	1.04	0.90	0.93	1.15	1.09	0.88	0.84
18	1.00	1.38	1.64	1.33	1.27	1.53	1.18	1.34
19	1.00	1.00	0.98	1.27	1.55	1.44	1.18	1.02
20	1.00	1.24	1.42	1.36	1.43	1.45	1.20	1.23
21	1.00	1.10	1.28	1.24	1.55	2.15	1.37	1.28
22	1.00	1.16	1.31	1.43	1.36	1.37	1.24	1.13
23	1.00	1.22	1.16	1.24	1.34	1.39	1.22	1.00
24	1.00	1.03	1.05	1.21	0.97	1.21	0.89	0.83
25	1.00	1.47	1.57	2.25	2.17	2.09	1.61	1.28
26	1.00	1.33	1.07	1.15	1.10	1.16	0.83	0.85
27	1.00	2.45	1.17	1.50	1.36	1.58	1.60	1.32
28	1.00	2.89	1.54	2.44	2.33	2.48	2.28	1.85
29	1.00	1.96	1.26	1.92	2.01	2.46	1.94	1.25
30	1.00	1.93	1.27	1.43	1.43	1.61	1.17	1.15
31	1.00	1.53	1.10	2.18	1.88	1.97	1.87	1.43
32	1.00	1.69	1.00	3.36	3.43	3.93	2.99	1.69
33	1.00	1.30	1.11	1.39	1.45	1.34	1.17	1.02
34	1.00	1.80	1.20	1.03	1.48	1.26	1.10	1.08
35	1.00	1.68	1.57	1.37	2.31	2.98	1.51	1.25
36	1.00	1.86	1.41	4.36	4.71	4.96	3.73	1.98
37	1.00	1.69	1.22	2.97	2.65	2.35	2.00	1.31
38	1.00	1.76	1.21	1.18	1.00	1.05	1.03	1.06
39	1.00	1.95	1.22	1.18	1.14	1.31	1.01	0.96
40	1.00	2.23	1.47	1.34	1.70	2.00	1.41	1.35
41	1.00	1.40	1.21	1.21	1.23	1.18	1.07	1.10
42	1.00	1.43	1.65	1.75	2.33	2.32	1.68	1.61
43	1.00	1.69	1.00	1.23	1.21	1.27	1.05	0.96
44	1.00	1.27	1.25	1.43	1.25	1.26	1.08	1.14
45	1.00	1.77	1.50	1.28	1.08	1.08	1.07	1.23
46	1.00	1.96	2.32	1.24	1.95	2.09	1.57	1.85
47	1.00	2.46	2.16	1.40	1.82	2.10	1.64	1.79
48	1.00	1.69	1.33	1.61	1.78	1.66	1.61	1.35
49	1.00	1.68	1.33	1.57	2.29	2.62	1.62	1.23
50	1.00	1.24	1.36	1.21	2.02	2.03	1.25	1.21
51	1.00	1.58	1.21	1.34	1.56	1.59	1.26	1.17
52	1.00	1.71	0.89	0.86	0.94	0.96	0.84	0.79
	1.00	1.46	1.25	1.44	1.48	1.57	1.29	1.16

Table 39 – Weekly Relative Mean Absolute Errors (wRMAE). CF1 and CF2 are combined forecast defined in Equation 38 & 39

12.5 Names and Abbreviations

50HZ - 50 Hertz TSO (German TSO)
APG - Austrian Power Grid (Austrian TSO)
AMPRION - Amprion (German TSO)
EEX - European Energy Exchange
ENBW - TransnetBW TSO (German TSO)
ENTSOE - European Network of Transmission System Operators for Electricity (Association of European TSO's)
EPEX - European Power Exchange (German/Austrian power exchange)
EXAA - Energy Exchange Austria (German/Austrian power exchange)
RTE - Réseau de transport d'électricité (French TSO)
TENNET - Tennet DE (German TSO)

ARIMA - Auto Regressive Integrated Moving Average model
CI - Computational Intelligence
EPF - Electricity Price Forecasting
DM - Diebold-Mariano [test]
MAE - Mean Absolute Error
Method - a class of models (e.g. ARIMA, linear)
Model - a specific parameterization of a method chosen by estimation
NN - Neural Network
NRA - National Regulatory Authority
PD - Price dependent [generators]
PI - Price Independent [generators]
RES - Renewable Energy Sources
RESCON - RESidual CONsumption (part of consumption that has to be covered by normal generators (and import))
RMAE - Relative Mean Absolute Error
RMSE - Root Mean Square Error
RRMSE - Relative Root Mean Square Error
SVM - Support Vector Machine
TSO - Transmission System Operator