THESIS FOR THE DEGREE OF MASTER OF SCIENCE IN ENGINEERING PHYSICS

Stochastic modeling of electricity prices using ARMA and ARMAX models

Robin Axelsson

Department of Mathematical Sciences
Division of Mathematical Statistics
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2012

Acknowledgements

I'd like to begin by thanking Patrik Albin for being my supervisor and instructor for this thesis. He's easy to work with, his expertise and support has been very useful. I'd also like to thank John Ekwall as SMHI for supplying the weather data that I used in my computations. Lastly, I would like to thank Alexander Herbertsson and Claes Breitholtz for helping me to better understand the Kalman filtration process that I used for the model estimations.

Abstract

In this thesis we examine the Swedish electricity market and the effects from the deregulation of it. The fact that the deregulation has given rise to highly volatile prices has created a strong need for more sophisticated models to yield better electricity price forecasts.
We examine and evaluate the ARMA models as forecasting models for electricity prices. We also examine how outdoor temperatures influence electricity consumption and electricity prices. Hourly and daily temperature data are then used as exogeneous input to ARMAX models which we also evaluate for use as forecasting models.
Both ARMA models and ARMAX models provide good fits during shorter time–periods and using temperature data has turned out to provide a significant improvement to the models. However, they all fail to properly capture the extreme behaviour with price spikes that occur during the winter seasons. They are also not suitable for long run predictions without further sophistications e.g. by using time–varying parameters such as time–varying volatilities and mean values.
The conclusions are that the need for improved forecasting methods indeed are imminent and that the ARMA and ARMAX models in some circumstances has the potential to provide good predictions.

Table of Contents

1. Theory and background

Since the deregulation of the electricity market, electricity prices have become extremely volatile putting a potentially serious strain on consumers' budgets, especially during the winter months. This development has created a need for financial instruments that protect against the fluctuating nature of the electricity prices and more sophisticated forecasting methods that predict the electricity prices movements in order to be able to make fair estimations on such financial instruments. The ARMA and ARMAX models have capabilities of capturing some of the characteristics inherent in electricity price movements that traditional random walk models fail to capture. This thesis intends to evaluate using ARMA models and ARMAX models with outdoor temperatures as exogeneous data as forecasting models for electricity price movements.

This thesis is divided as follows; Chapter 1 gives a background of the electricity market and introduces the models that are intended be used for estimations on this market. Chapter 2 provides the methodology behind model estimations and the evaluation of these estimations. Chapter 3 presents the results and we provide concluding remarks in Chapter 4.

Chapter 1 begins by analyzing the Swedish electricity market discussing the actors on the market, their roles and their market power in Section 1.1. Section 1.2 discusses the producers vs. the consumers and various factors that affect supply and demand of electricity. Section 1.3 discusses the electricity price movements and what distinguishes these movements from e.g. the movements of stocks. Section 1.4 conducts a statistical analysis of hourly observed electricity price data for 2010. Section 1.5 introduces the ARMA and ARMAX models and discusses their theoretical properties.

1.1 The Swedish electricity market

The electricity market has grown considerably over the past 20 years and consists today of many different actors [Svenska Kraftnät 2007]. On the supply side we have the electricity producers and on the demand side, we have the consumers. The connection between consumers and producers is mediated through brokers and other intermediaries that interact with the commodities market that is maintained by Nord Pool Spot A/S[1]. The electricity is delivered through the national power grid and regional networks. The national grid is maintained by Svenska Kraftnät whom also have the role to ensure that the balance between supply and demand is in harmony and that the supply of electricity is stable throughout the country. The roles of the different actors are illustrated in **Figure 1.1**.

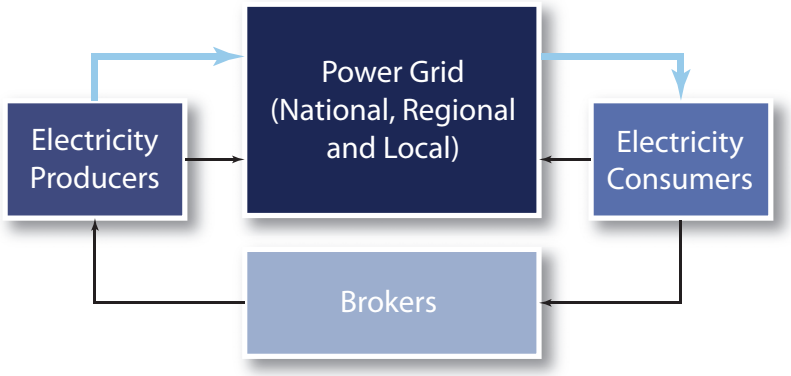

Figure 1.1: An illustration of the relationship between the different roles on the electricity market.

Source: Svenska Kraftnät

Electricity markets are also characterized by capacity constraints in the distribution networks. Due to these constraints, Svenska Kraftnät have decided to divide Sweden into four separate regions with different tariffs. This division is illustrated in **Figure 1.2**.

[1] Nord Pool Spot A/S is owned by the Nasdaq OMX Group Inc which is an American financial services corporation that runs the Scandinavian stock exchange.

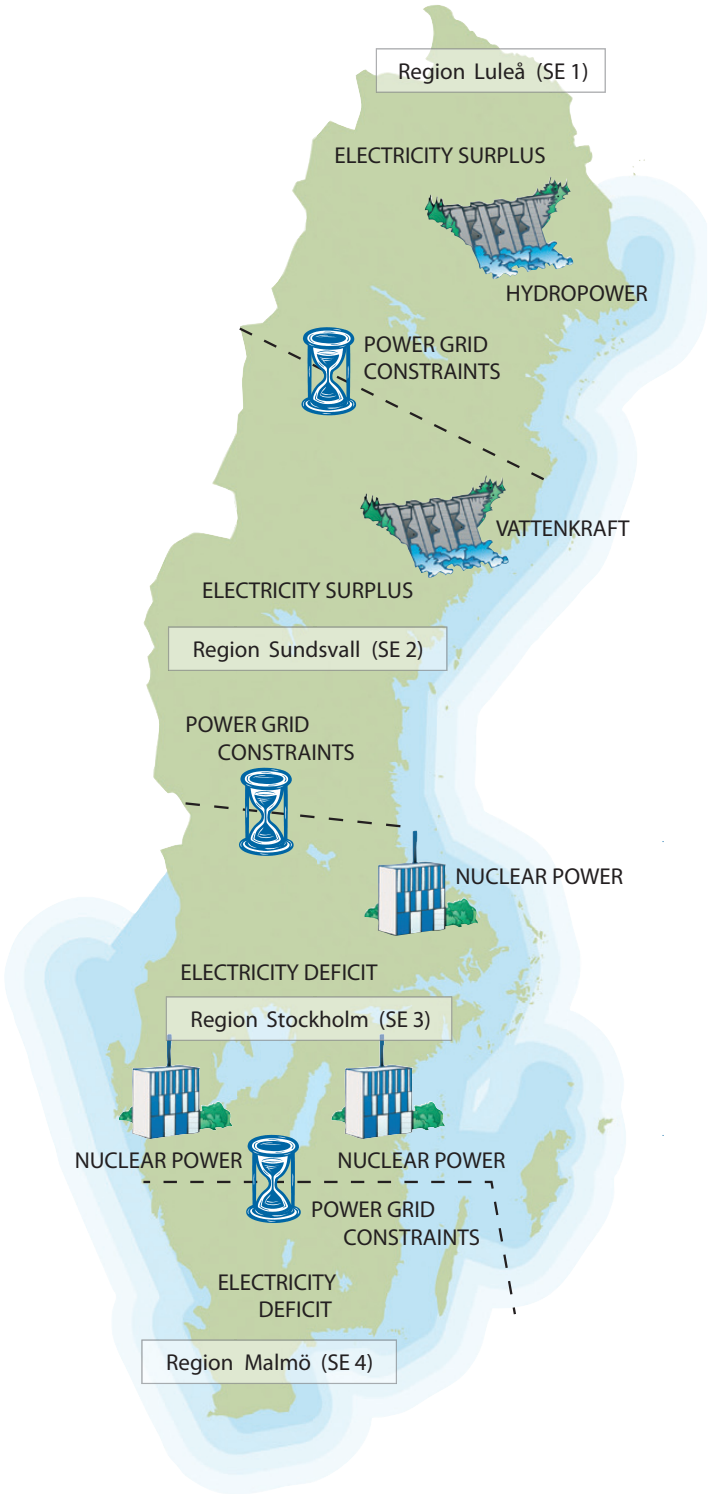

Figure 1.2: Division of Sweden into four price regions (SE 1 – SE 4) as decided in December 1 2011 by Svenska Kraftnät.

Source: Svenska Kraftnät

The distribution between consumers and producers is uneven which tends to create a surplus in the north where most of the electricity is produced and a deficit in the more densely populated regions in the south where most of the electricity is consumed. Each line within the power grid has a maximum amount of electricity that it can carry at a given moment which tends to put a serious strain on the marginal costs of transmission. This has a considerable potential to isolate regions of the electricity market from the rest of the market. In such situations, the producers and market makers have the ability to exercise a greater level of market power and influence over prices as has been demonstrated in e.g. the Californian electricity market after the restructuring and deregulation of electricity prices in that market [Borenstein et al 1999]. In general, suppliers have a considerable potential to exercise market power. Some studies have indicated this by deducing large Lerner indices (even though a large part of the Lerner index can be explained by the inelastic nature of electricity demand) in e.g. the British electricity markets [Wolfram 1999]. Also political agendas and energy policies such as the trade with carbon emissions, the political aftermath after the Fukushima incident, and mandates requiring the production of electricity to only come from renewable sources have shown to give rise to increasing electricity prices [Bryce 2012] and [Svenska Kraftnät 2007].

During the early years of the reform and deregulation of the electricity market, there was a rule of thumb that set the electricity costs to consist of three equally large parts:

- The cost of the electrical energy itself
- Network fees and charges for the transmission of the electricity
- Energy taxes and VAT

This rule of thumb has been blurred out as the cost of electrical energy by 2007 constituted 35% of the final tariff, 20% network fees and 45% taxes [Svenska Kraftnät 2007]. Today, the cost of energy can account for as much as 50 % followed by a 40% tax and a transmission fee down to 10% of the total cost of electricity consumption.

1.2 Producers, consumers and factors affecting supply and demand

The electricity is mainly produced by nuclear power reactors and hydropower plants. In Sweden there are a total of 10 nuclear reactors, each of them located in Ringhals, Forsmark, and Oskarshamn. Together they constitute about half of the electricity production, the other half comes from hydropower. By 2010, about 144.9 TWh of electricity was produced [SCB 2012]. The sources of electricity production are summarized in **Figure 1.3**.

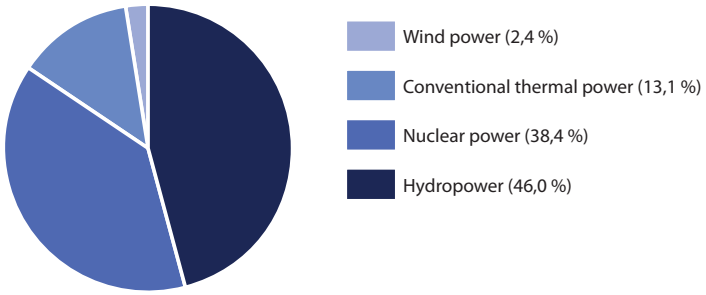

Figure 1.3: The Swedish electricity production as of 2010 by energy source as a percentage of the total electricity production.

Source: SCB

Factors that affect the production capacity are the number of reactors that are online and the water levels in the dams. A period of low precipitation reduces the production capacity of the hydropower plants. Also, the need to take down a nuclear reactor for maintenance also reduces the production capacity.

The majority of electricity consumption comes from heating of real estate and residential consumption which constitutes about half of the total consumption in Sweden. The other half of the consumption comes from the industry such as the paper industry. The total domestic usage of electricity at 2010 was at 136.3 TWh. The electricity consumption is summarized in **Figure 1.4**.

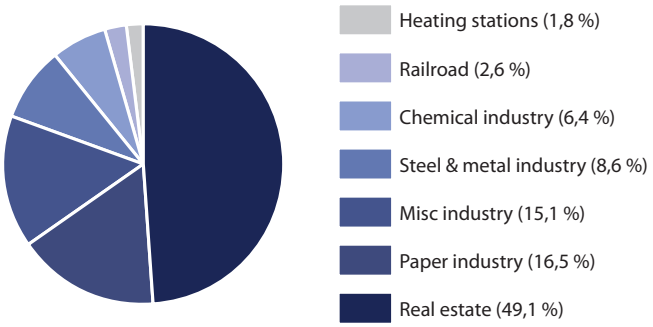

Figure 1.4: Major applications for electricity consumption as a percentage of total Swedish consumption by 2010.

Source: SCB

The major factor that affects the electricity consumption in Sweden is the temperature. The lower the outdoor temperature, the more electricity is consumed as houses need more heating. The correlation between the electricity consumption and outdoor temperature is illustrated in Figure 1.5.

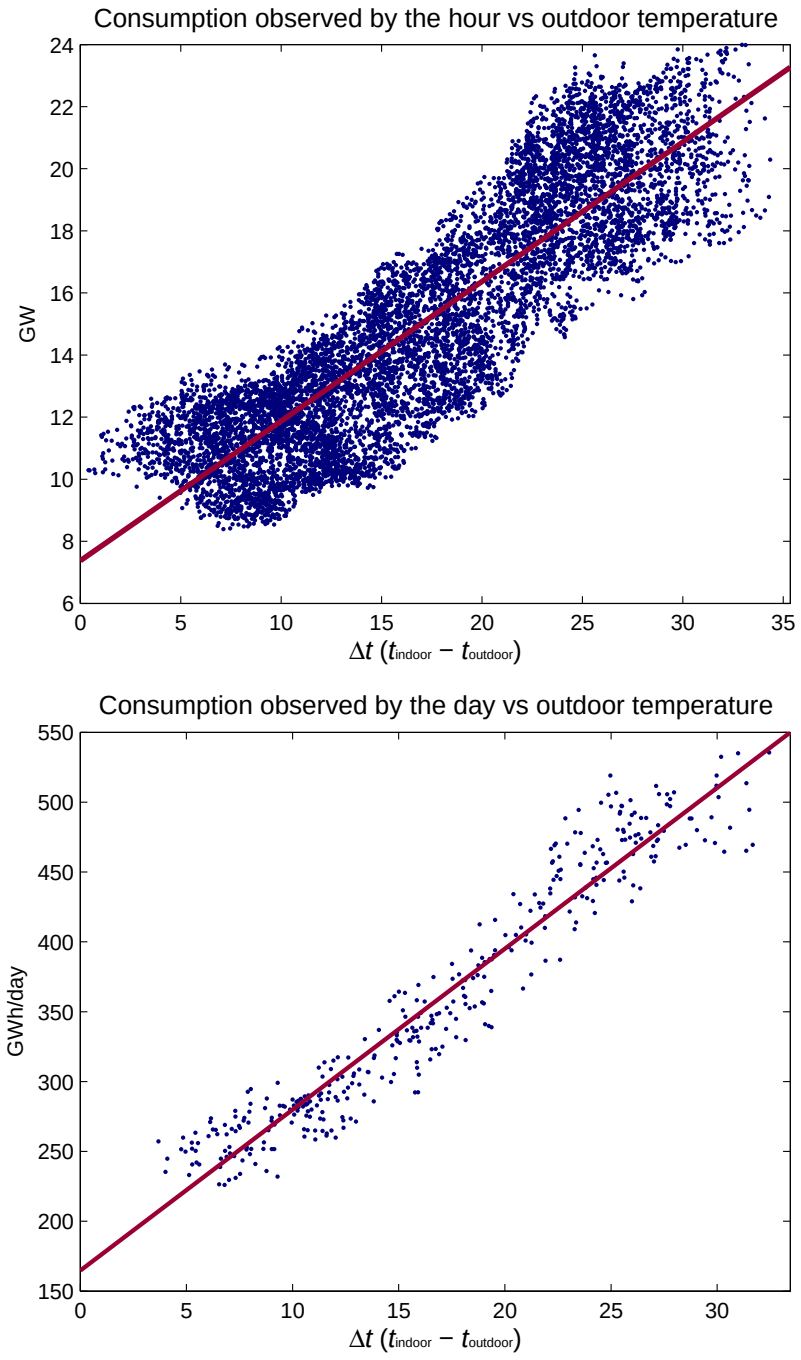

Figure 1.5: The hourly electricity consumption (upper figure) and daily electricity consumption (lower figure) as a function of temperature difference between weighted average nationwide outdoor temperature and average indoor temperature (20°C).

Source: SMHI & Nord Pool Spot A/S consumption data

Another factor that affects consumption is the business cycle. At the peak of the economy, more electricity is consumed for production whereas less is consumed during a recession. The Swedish power grid is connected to Norway, Finland, Germany and Poland where some electricity is exported during a surplus production and imported whenever there is a national deficit.

1.3 Properties of electricity price movements

In 1992 the electricity market became deregulated and restructured which gave more freedom for electricity producers' pricing strategies. Since 2002 electricity has become available for public trade in the spot market through Nord Pool Spot. The deregulation and public trade has made the electricity prices considerably more volatile than in the past. Since the means to store electricity are limited, electricity prices are very sensitive to sudden shortages which make prices even more volatile, especially during the winter seasons. This has created a stronger need for protection against such fluctuations in the past few years and the volatile nature of the electricity prices have put a stronger demand for sharper instruments that give proper protection at the right price–premium.

The most commonly used instrument to protect or hedge against such price fluctuations is the swap agreement which basically means that a user can enter into a bound agreement to either pay a fixed price for the electricity or a floating price that depends on time to maturity of the agreement [Hull 2007]. Pricing of financial derivatives such as swap agreements basically hinges on the assumption that the underlying asset follows some sort of stochastic process. The most basic derivatives are the European put and call options that are mostly used for stocks and other equities. Such financial instruments build upon statistical models that are specifically designed for equities. The problem with such statistical models is that they fail to properly capture the movements in electricity prices and therefore more sophisticated models are needed for pricing electricity based derivatives. As we have mentioned before, electricity prices have the following three characteristics that stock prices don't have [Knittel Roberts 2005]

- Short term stationarity – they are stationary in the short term and have no trend when looking at shorter time intervals. There is also a mean–reversion property inherent in the electricity prices which means that the electricity prices tend to converge toward a long–term mean over time.
- Seasonalities – they have different types of seasonal cycles that come from the fact that people tend to take showers in the mornings and therefore consume more electricity during such times. Also more electricity is consumed during weekdays than during weekends and during winters than during summers.

- Sudden price spikes – sudden shortages and demand that saturates the production tend to give rise to price spikes which comes from the difficulty to store electricity.

A mathematical model must take these factors into account in order to achieve a fair accuracy. There is also a positive skew larger during periods of high demand variability and smaller during periods of low demand variability [Bessembinder Lemmon 2002]. The volatility of spot prices is considerably higher in periods of high demand than during periods of low demand.

There is also an "inverse leverage effect" inherent in the spot price movements. The traditional leverage effect is experienced when e.g. stock prices drop which tends to panic the market and raise the implied volatilities. The higher implied volatility is a result of the inherent risk–aversion among traders and investors for markets that are going down. Another way too look at this is through financial instruments that are used to hedge against such developments, such as put options. The increased demand for out of the money puts gives rise to higher implied volatility for lower strikes[2] [Hull 2007]. With many commodities such as the electricity, this effect is reversed which means that the implied volatility increases when the commodities prices increase. So there is an aversion in the market against rises in commodities prices. Also, price increases also increases demand for derivative instruments such as forward agreements, futures contracts, swap agreements or call options which are instruments that are used to hedge against rising commodity prices.

Another feature of electricity prices that is quite distinct from equities is that they sometimes can be negative [Brandstätt et al 2011]. This happens when supply outstrips demand and the producers literally don't know where to put it. Measures have been taken to prevent such events from happening. In Scandinavia, Danish wind power is used to pump water into Norwegian and Swedish reservoirs which are later used to drive hydropower plants whenever the demand requires it. Also, at Nord Pool there is a minimum price policy which most likely explains why we don't see any negative prices in our electricity price data.

[2] The strike price of a put–option, or put which is the price at which a holder of it has the right but not the obligation to sell the underlying asset. A put is out of the money when the strike price of the put is lower than the price of the underlying asset and it is at–the–money or in–the–money when the strike price of the put is higher than the price of the underlying asset.

1.4 A statistical analysis of observed electricity price data

Since 2010, trading data for electricity spot prices have become publicly available via Nord Pool Spot A/S and in this Section we analyze the Swedish electricity prices for the year of 2010 as noted on the Nord Pool Spot commodities exchange using the framework as provided by Knittel, Roberts [Knittel Roberts 2005]. Electricity price data show distinct variations within one day. In **Figure 1.6** we see the average hourly electricity prices measured in Swedish Krona per megawatt hour (SEK/MWh).

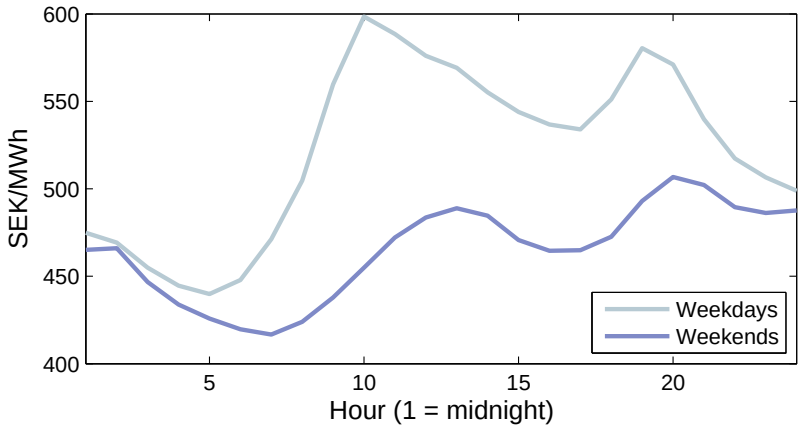

Figure 1.6: Average hourly electricity prices for 2010.

On average, the prices are higher during the weekdays than during the weekends. The price begins to increase at roughly 6 in the morning. A peak sets in at around 9 in the morning and is sustained a bit into lunch time. As the workday ends the electricity prices begins to decline, the demand begins to shift over to residential use and another peak occurs at dinner time in the evening. In **Figure 1.7** we see a 10 day sample of hourly price and demand data for the time periods 1 – 11 January 2010 and 30 June – 10 July 2010.

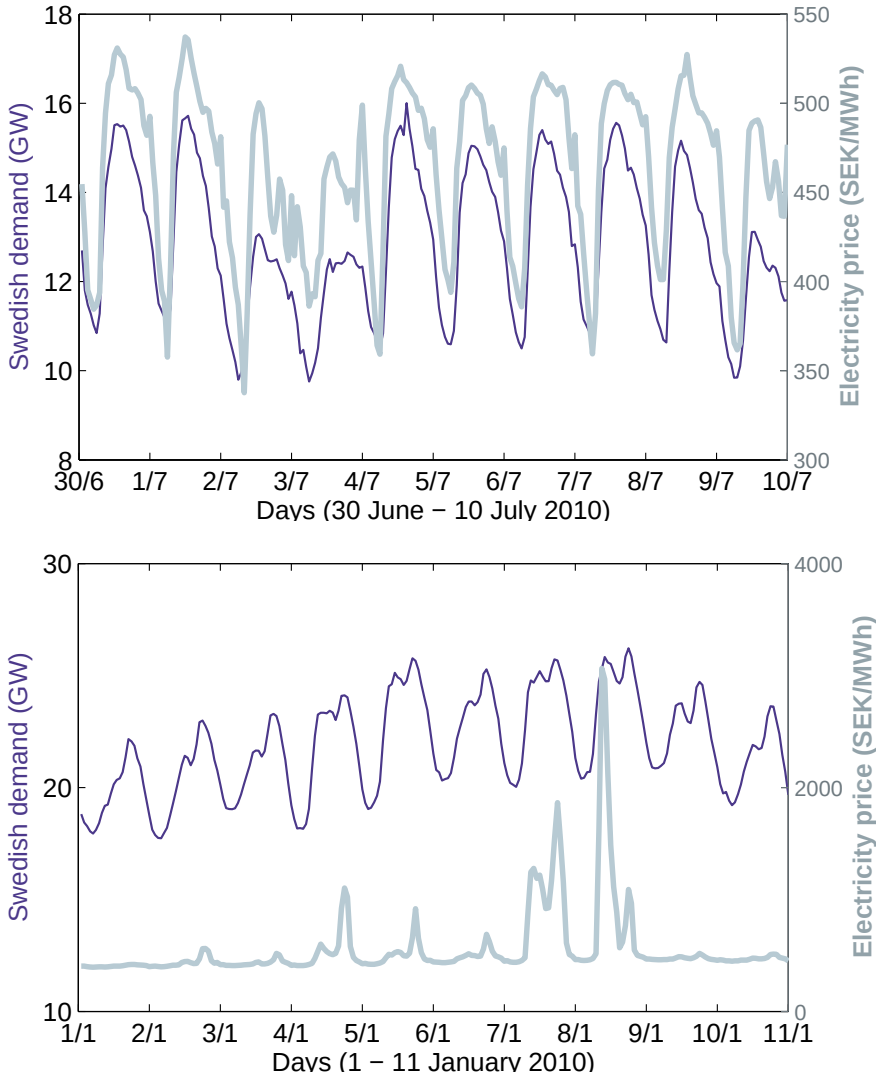

Figure 1.7: Samples of hourly electric prices for the periods 30 June – 10 July 2010 and 1 – 11 January 2010.

The left vertical axis shows the demand in gigawatts (GWh/h = GW) and the right vertical axis shows the electricity price in SEK/MWh. The horizontal axis corresponds to midnight of that particular day. These figures give a clear view over the daily usage patterns and their persistence over time. It is apparent that the summer prices mimic the pattern of the demand movements. During the winter days it seems that when the demand reaches 25 GW it appears to over–saturate the supply as this demand level seems to trig price spikes in the electricity market.

It is also apparent that the electricity prices have a strong seasonal component that chiefly reflects the residential and commercial heating needs during the

winters. This is illustrated in **Figure 1.8** that shows hourly weekday averages for each of the four seasons.

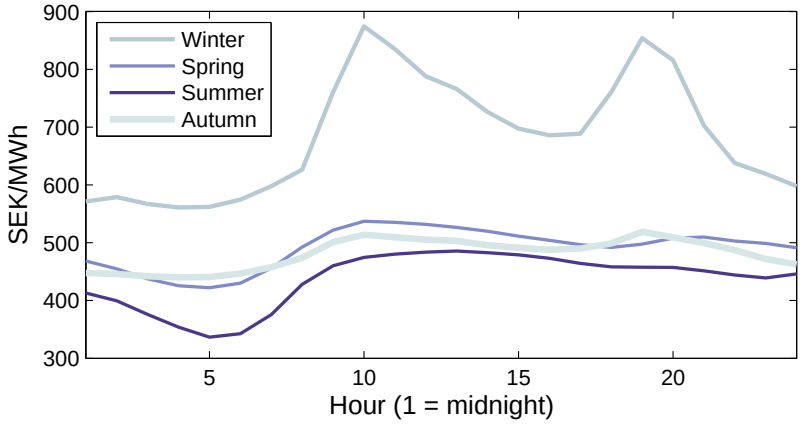

Figure 1.8: Average weekday hourly electricity prices by season.

Compared to southern countries there is no significant need for air conditioning during the summer as is clearly illustrated in **Figure 1.5**.

Figure 1.9 presents a 5–day OHLC chart and a daily OHLC chart showing hourly electricity price developments during 2010. It illustrates how prices make dramatic swings which tend to occur in clusters. This is a result of demand saturating and even exceeding the capacity of the electricity system.

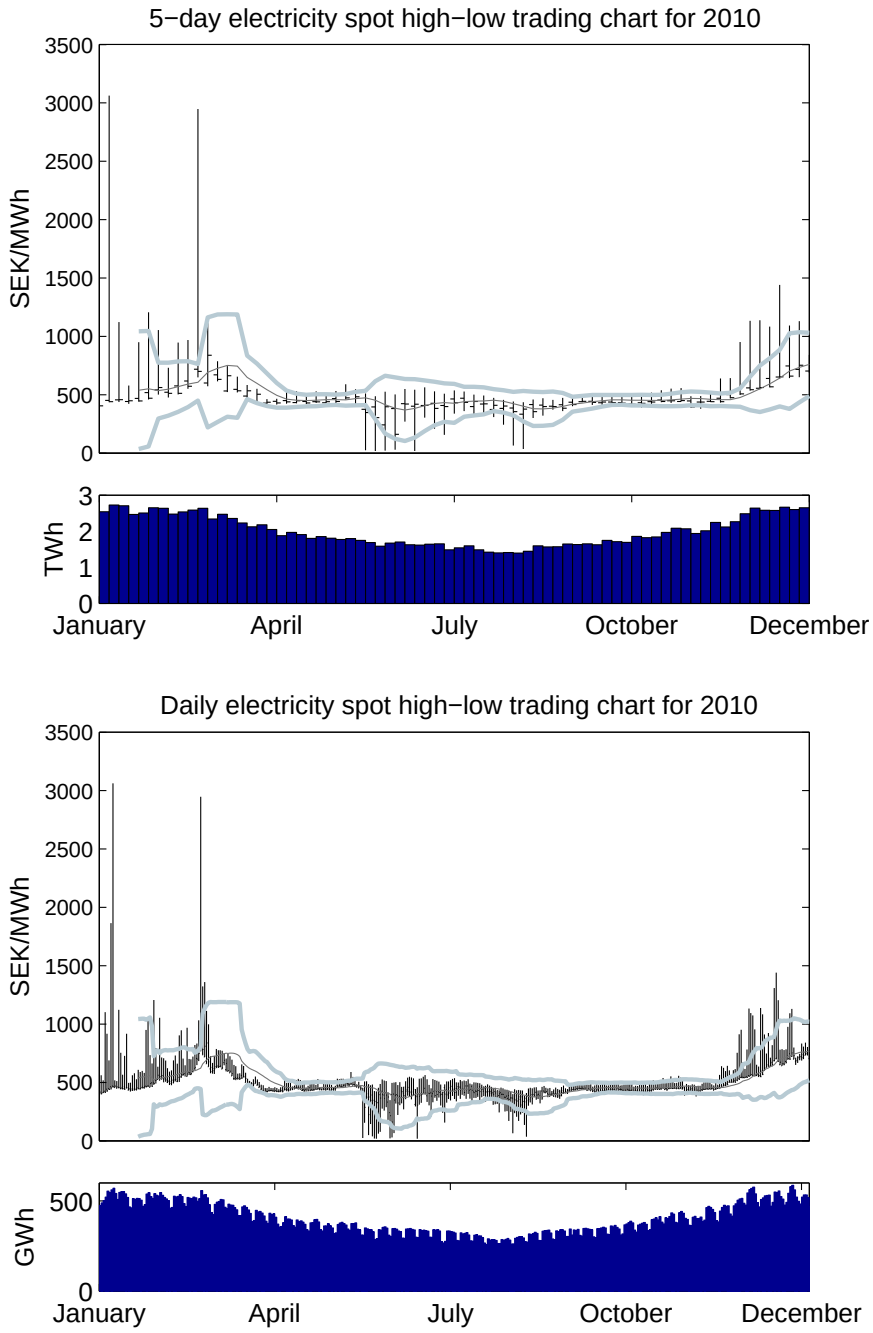

Figure 1.9: The upper pane of the upper figure is a 5–day OHLC chart with 20–day Bollinger bands of second order superimposed and the lower figure shows a daily OHLC chart with the same Bollinger bands. The lower panes show the amount of electricity delivered during each period, i.e. each 5–day period in the upper figure and 1–day period in the lower figure.

We end this Section with a brief discussion about the distributional properties of the electricity price data. In **Figure 1.10** we have an empirical 100–bin histogram of the electricity price data with the normal distribution superimposed on it. The figure shows that the deviations from the normal distribution are significant.

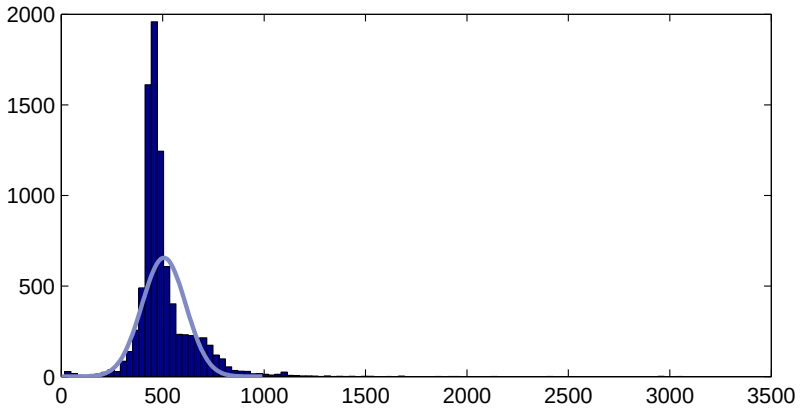

Figure 1.10: Empirical histogram of electricity prices with the normal density superimposed.

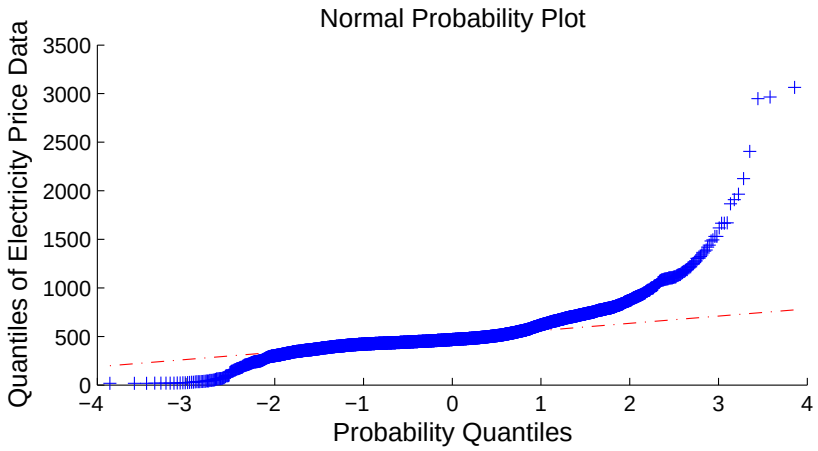

Figure 1.11: QQ–plot of electricity prices.

In **Figure 1.11** we have a QQ–plot of the data. The superimposed line represents the normal distribution and if the empirical distribution were perfectly normal, all points would coincide with that line. We see that the empirical distribution shows a fairly normal behaviour within the first positive and the first two negative quantiles but deviates from that behaviour beyond those quantiles. Both of these figures suggest that the empirical distribution is more heavily tailed than the normal distribution. **Table 1.1** shows some summary statistics of the electricity price data. It also indicates heavily tailed properties showing a considerably higher kurtosis than the normal distribution

and a positive skewness. Something to note is that we don't have negative prices in the data.

Statistic	Value
Mean	505.9123
Minimum	17.1900
Maximum	3063.3100
Standard Deviation	152.4489
Skewness	3.5065
Kurtosis	38.9342

Table 1.1: Electricity price summary statistics.

In **Figure 1.12** we have the autocorrelation function for the price levels and they are statistically significant even beyond 1000 lags.

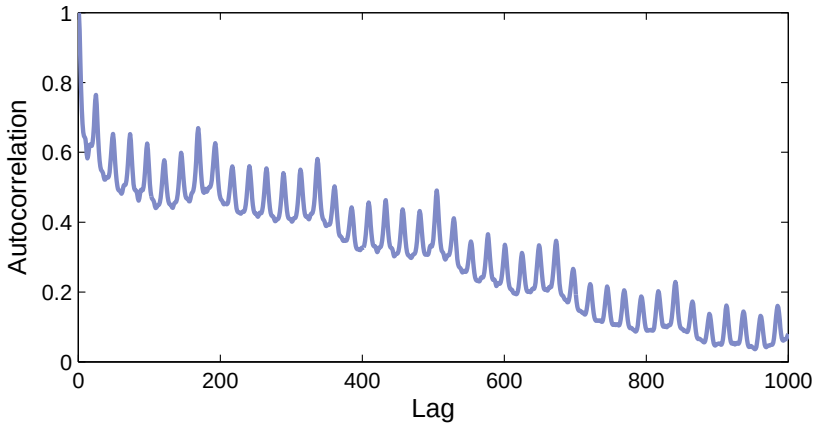

Figure 1.12: Autocorrelation function for electricity prices.

The behaviour of the correlogram reflects the recurring daily usage patterns and the weekend/weekday cycle. This is quite distinct from the behaviour of e.g. equity prices that are commonly assumed to follow a random walk with drift. Although there is a field of research in behavioural finance that suggests that there are some predictable movements in equity prices and they in many cases also have a statistically significant autocorrelation [Ding Granger 1996], they are considerably smaller than the autocorrelation inherent in the electricity price data. In **Figure 1.13** we have the autocorrelation function for the square of the prices which is quite similar to the correlogram in **Figure 1.12**.

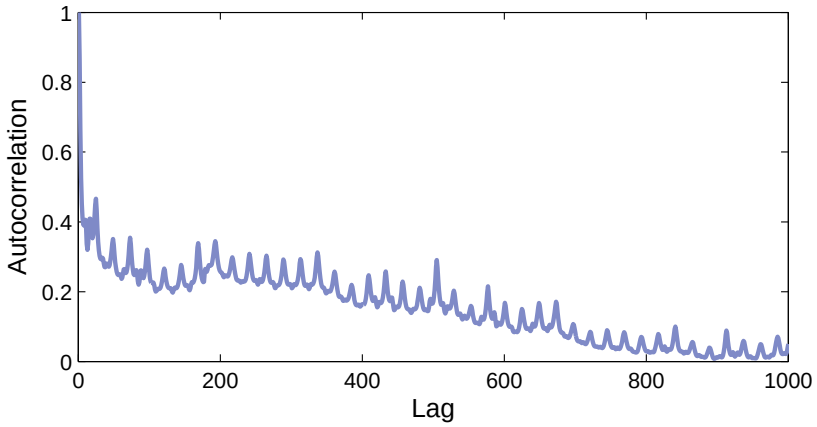

Figure 1.13: Autocorrelation function for the square of electricity prices.

We have a high degree of persistence of the second order moment even after 100 lags which implies high volatility persistence over time. From the analysis of the data we conclude just like Knittel, Roberts [Knittel Roberts 2005] that it is suggestible that any modeling effort should account for the following characteristics of the price series:

- Mean reversion
- Seasonal effects for the short-term (i.e. time of day, weekend vs. weekdays)
- Seasonal effect for the time of the year and/or correlation with temperatures
- Time-varying volatility
- Extreme values

1.5 The ARMA model

The ARMA model is a linear model to describe a time–series data that is assumed to be a linear combination of a purely autoregressive process and the moving average of some white noise process [Kitagawa 2010]. An autoregressive process of order p, or in short AR(p) is described by

$$y_t = \sum_{i=1}^{p} a_i y_{t-i} + \varepsilon_t , \tag{1.1}$$

where the coefficients $a_1, \ldots, a_p$ are the parameters of the model and ε_t is a Gaussian white–noise process with variance σ^2 that is stochastically independent of $y_{t-1}, y_{t-2}, \ldots, y_{t-p}$. A process that can be entirely described by this model is called an autoregressive process. The model can also be looked upon as a linear filter where the observations y_t is the input signal and the process ε_t is an output signal [Albin 2003 p99]. The impulse response of that filter is then defined as $h(k) = a_k$. The autoregressive model can also be looked upon as a difference equation [Eriksson Larsson Wahde 2000 p35] which has analogies with linear differential equations. If we define the lag operator B where $By_t = y_{t-1}$ and $B^k y_t = y_{t-k}$ for any integer $k \geq 0$, then we have

$$\sum_{k=0}^{p} a_k B^k y_t = \hat{\varepsilon}_t , \tag{1.2}$$

where a_0 is -1 and $\hat{\varepsilon}_t = -\varepsilon_t$. The lag operator is analogous to the differential operator of differential equations and this difference equation has the characteristic equation

$$z^p + a_1 z^{p-1} + a_2 z^{p-2} + \ldots + a_p = 0 , \tag{1.3}$$

where the roots describe the characteristics of the autoregressive model. If we want the model to be stationary, the roots of its characteristic equation must lie within the unit circle, i.e. each root z_i must satisfy the relation $|z_i| < 1$. Autoregressive processes with unit roots are called integrated processes and detection of integration can be done with e.g. the Augmented Dickey–Fuller test [Greene 2011]. An example of a non–stationary autoregressive process is illustrated in **Figure 1.14**. We assume that the electricity price data follows a stationary process. The two most common ways of fitting an autoregressive model to a stochastic process are the Least–Square method and the Yule–Walker method.

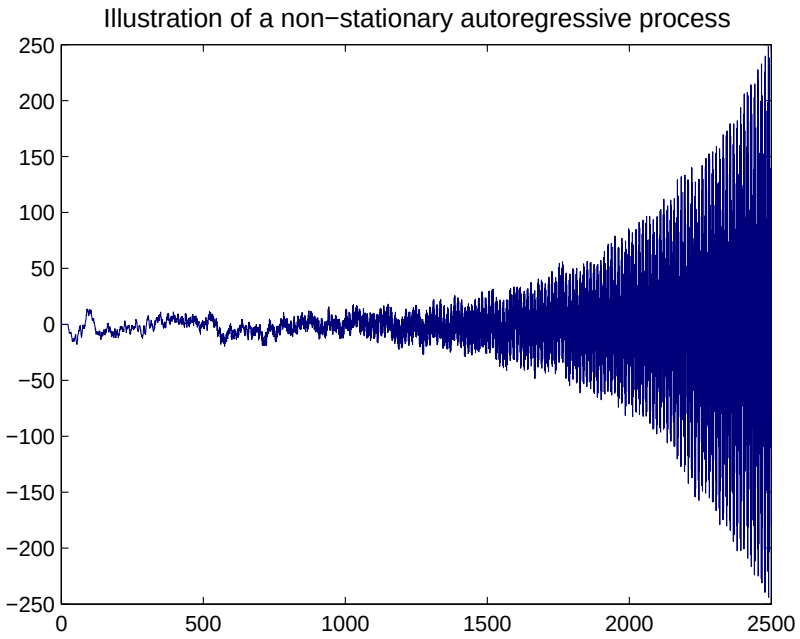

Figure 1.14: Estimations of autoregressive processes can easily turn non–stationary, this figure shows a simulation of such a process.

A moving average model of order q, or in short MA(q) is defined by

$$y_t = \sum_{k=0}^{q} b_k \varepsilon_{t-k} \,, \tag{1.4}$$

where $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ is a set of stochastically independent Gaussian variables with mean 0 and variance σ^2, which are also called error terms. Also this model can be seen as a filter where the error terms represent the signal in and y_t the filtered signal out. The impulse response of that filter is then defined as $h(k) = b_k$. We don't have to worry about integration as this model yields only stationary processes [Albin 2003 p109]. **Figure 1.15** illustrates how the order of a moving average process affects its behaviour.

An ARMA(p, q) model with autoregressive order p and a moving average order q is described by

$$y_t = \sum_{i=1}^{p} a_i y_{t-i} + \varepsilon_t - \sum_{i=1}^{q} b_i \varepsilon_{t-i} \,, \tag{1.5}$$

where $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ is a set of Gaussian variables with a mean 0 and variance σ^2 just like in the case with the pure moving average model. These error terms are stochastically independent of each other and the past observations in the autoregressive part of the model.

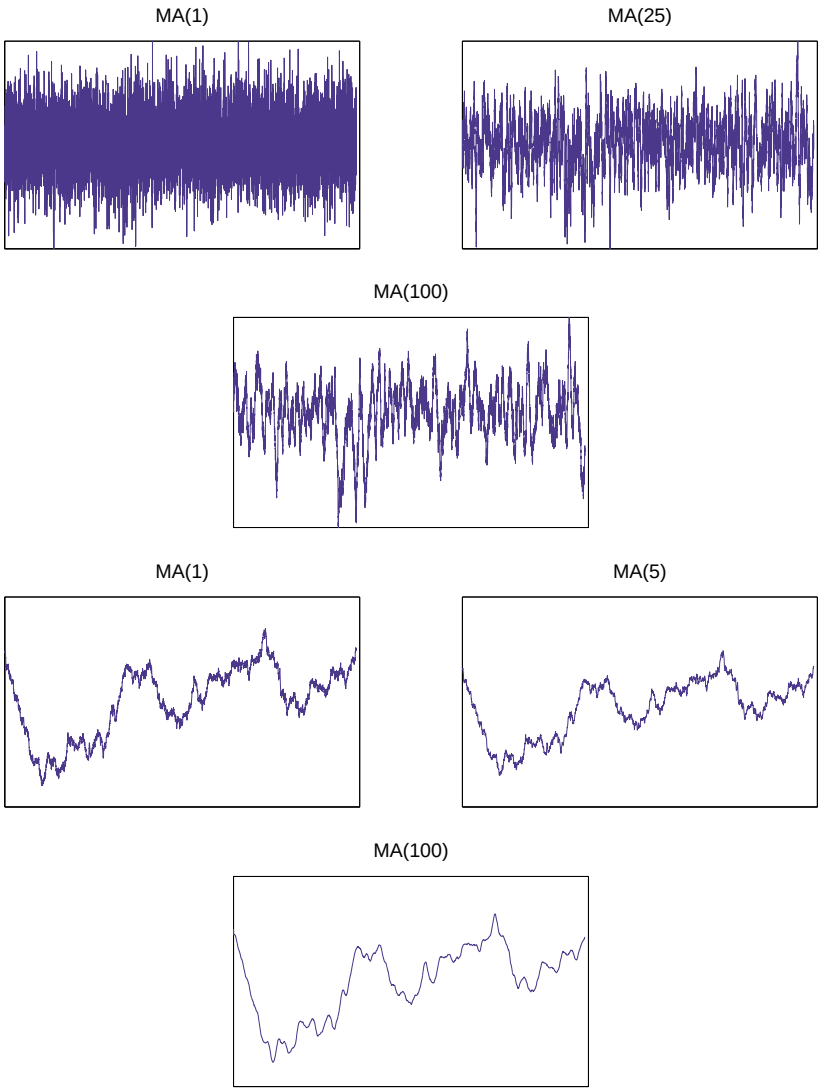

Figure 1.15: An illustration on moving average processes of different orders with a white noise process (upper triad) and a Wiener trajectory (lower triad) as a background process. A higher order yields a lower "detail" in the output which suggests that it has low–pass like filter properties.

This model can be seen as a generalization of the pure autoregressive model and the pure moving average model where each of them is a special case where the AR(p) model actually is an ARMA(p, 0) model and the MA(q) model is an ARMA(0, q) model. If we let $a(B) = 1 - \Sigma_{i=1}^{p} a_i B^i$ and $b(B) = 1 - \Sigma_{i=1}^{q} b_i B^i$ then the ARMA model can be expressed more concisely as

$$a(B)\, y_t = b(B)\, \varepsilon_t , \qquad (1.6)$$

and the ARMA model can actually be expressed as a moving average model with infinite order. Let $g(B) = a(B)^{-1}b(B)$ then

$$y_t = g(B)\varepsilon_t = \sum_{i=0}^{\infty} g_i \varepsilon_{t-i} \,. \tag{1.7}$$

It can also be looked upon as a filter just as the pure AR(p) and MA(q) models and its impulse response function is then defined as $h(k) = g_k$. ARMA models go under the Box–Jenkins framework and are sometimes called Box–Jenkins models [Brockwell Davis 2006]. The Box–Jenkins framework is a framework for fitting ARMA and ARIMA models in the presence of trend and seasonalities, this will be covered in more depth in the methodology Chapter of this paper. Also, filter theory becomes significant when one wants to filter out trend and seasonality variations. An ARIMA model is a model that deals with integration as discussed above. Another way to generalize the ARMA model is to account for exogeneous influences such as temperature variations. The ARMAX(p, q, r) model where r is the order of exogeneous influence is described by

$$y_t = \sum_{i=1}^{p} a_i y_{t-i} + \sum_{j=1}^{r} b_j u_{t-d-j} + \varepsilon_t + \sum_{k=1}^{q} c_k \varepsilon_{t-k} \,, \tag{1.8}$$

where d is the delay in discrete time–steps, or samples that occur before the exogeneous influence affects the output y_t.

2. Methodology

This Chapter presents the methodology behind estimation and evaluation of the models as presented in Chapter 1. The first Section discusses the general principles behind model estimation and validation. Section 2.2 and 2.3 discusses general methods for estimating pure autoregressive processes whereas subsection 2.4 and 2.5 provides an estimation framework for more complex ARMA and ARMAX models. Section 2.6 introduces and discusses the Box–Jenkins framework for identifying and removing trend and seasonalities within observed data to be used for model estimation. Section 2.7 introduces the concept of goodness–of–fit and provides criteria for validating model estimations.

2.1 Principles behind model estimation and evaluation

The basic approach is to define a set of models or a set of candidates to fit to observed data. The accuracy of the fit is usually compared against a criterion or tested using e.g. a χ^2 based test criterion such as the Ljung–Box test or an information criterion such as the Akaike information criterion (AIC). These test criteria get less reliable the more complexity there is in the model and even less reliable when comparing different types of models. So, a more rigorous approach is to test the fit of a model is by making a simulation based on the model and comparing it to the empirical distribution of the observed sample data. Information criteria can on the other hand be used to estimate the optimal order of e.g. an autoregressive model. The general approach is illustrated in **Figure 2.1** [Ljung 1999].

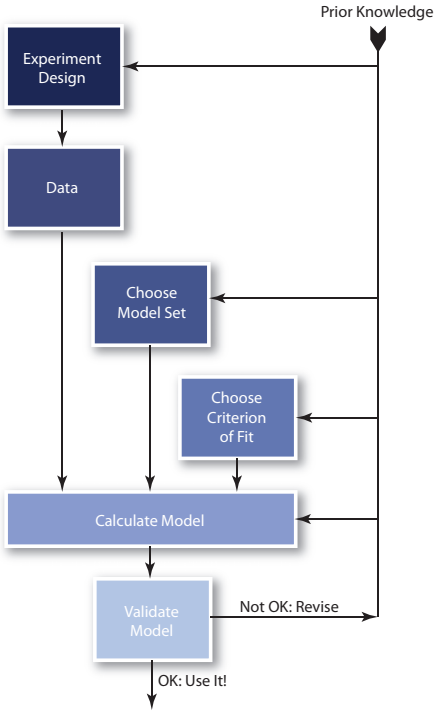

Figure 2.1: The system identification loop used to find the model that best represents the observed data.

Source: Ljung, 1999

The experiment design involves fitting ARMA and ARMAX models to observed electricity price data (that has trends and seasonal cycles removed according to the Box–Jenkins framework) and temperature data. The model set comprises of ARMA(p, q) and ARMA(p, q, r) models over a parameter space (p, q) and (p, q, r) that is computationally viable. Each of these models is fit to the electricity price data using a least–squares approach which is inherent in the recursive iterative Kalman filtration process. The estimated model is then checked for non–stationarity and validated against an information criterion and/or the measurement criteria as defined in the Kolmogorov–Smirnov framework.

2.2 Fitting AR models using least–squares

The least square method can be used for estimating linear models such as the autoregressive model and the moving average model. If we let $\tilde{y}_t$ be the value described by the model to be fitted, then the equation to be minimized is

$$S(\Theta) = \sum_{t=t_0}^{T} (y_t - \tilde{y}_t)^2 \qquad (2.1)$$

for a series of T observations. The parameter vector Θ contains the parameters $\{\theta_1,...,\theta_N\}$ that are to be estimated. The starting time t_0 is set so that the equation is defined for the parameters of the highest order in the model. For an ARMAX(p, q, r) model, the starting point is $t_0 = \max(p, q, r + d) + 1$. So we have a minimization problem with the partial derivatives

$$\frac{\partial \Theta}{\partial \theta_i} = 0 . \tag{2.2}$$

For simpler models such as the AR(p) model, we can formulate a linear vector equation

$$A\vec{x} = \vec{b} \tag{2.3}$$

where $\vec{b}$ is a vector of the observed values at each time point t in the time–series, $\vec{x}$ is a vector with the model parameters a_1, …, a_r and A is a matrix with prior observations $y_{t-1},\ldots,y_{t-r}$ at each time point t. So the first equation is at time point $t = p + 1$ if the first observation in the time–series has taken place at 1. So we have a system with more equations than unknown and it is minimized by multiplying both sides of the transpose of the matrix A from the left and then multiplying by the inverse of $A^T A$ on both sides

$$\vec{x} = \left(A^T A\right)^{-1} \vec{b} . \tag{2.4}$$

The projection inherent in the scalar product automatically finds the shortest distance between the higher dimensional "equation" space and the lower dimensional parameter space. This makes it a computationally efficient and yet powerful method to minimize the errors in the estimation. The error terms are then estimated from

$$\varepsilon_t = y_t - \tilde{y}_t \tag{2.5}$$

where

$$\tilde{y}_t = \sum_{k=1}^{p} a_k y_{t-k} \tag{2.6}$$

and they are assumed to be $N\left(0,\sigma^2\right)$ distributed. Another important result of the least–squares method is that the error terms are orthogonal to the observations. One can actually show that the least–squares solution is optimal if and only if the error terms are orthogonal to y_t through the Wiener–Hopf equations [Sorenson 1970].

2.3 Fitting AR models using the Yule–Walker method

Another approach to estimate an autoregressive model is through the Yule–Walker equations [Albin 2003 p112]. The autocovariance function for a stochastic process $\{X(t)\}_{t\in T}$ is defined as $r_X(s, t) = \mathrm{E}[X(s)\ X(t)]$ for $s, t\in T$ if $\mathrm{E}[X(t)^2] < \infty$. Let $e(t)$ be an AR(p)–process where $e(t) = \Sigma_{k=0}^{p} a_k Z(t-k)$ with a characteristic polynomial $A(z) = \Sigma_{k=0}^{p} a_k z^k$ such that $A(1) \neq 0$, then (see [Albin 2003 p113] for proof)

$$\begin{cases} r_X(\ell) + a_1 r_X(\ell-1) + \ldots + a_p r_X(\ell-p) = 0 \text{ for } \ell \geq 1 \\ r_X(0) + r_X(1) + \ldots + r(p) = \sigma^2 \end{cases} \tag{2.7}$$

where $r_X(k) = r_X(t, t–k)$. If we let $\sigma_k = r_X(k)$ then

$$\begin{bmatrix} \sigma_1 \\ \sigma_2 \\ \sigma_3 \\ \vdots \\ \sigma_p \end{bmatrix} = \begin{bmatrix} \sigma_0 & \sigma_1 & \sigma_2 & \cdots & \sigma_p \\ \sigma_1 & \sigma_0 & \sigma_1 & \cdots & \sigma_{p-1} \\ \sigma_2 & \sigma_1 & \sigma_0 & \cdots & \sigma_{p-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_p & \sigma_{p-1} & \sigma_{p-2} & \cdots & \sigma_0 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_p \end{bmatrix} \tag{2.8}$$

or, in short

$$\vec{\sigma} = \Sigma \vec{a}\,, \tag{2.9}$$

where $\vec{\sigma}$ is a vector of autocorrelation values. Σ is the autocovariance matrix and $\vec{a}$ is the parameter vector of the AR(p) model [Brockwell Davis 2006 p239]. The elements σ_0 in the diagonal in the matrix are the variance σ^2 of the observations. So, by constructing the covariance matrix we can estimate the AR(q)–coefficients from $\vec{a} = \Sigma^{-1}\vec{\sigma}$. This method is similar to the least–square method in that we try to minimize the prediction errors. The error terms are estimated as in (2.5) by using the Yule–Walker estimated parameters in (2.6).

2.4 State–space modeling

Linear models such as the autoregressive model and the autoregressive moving average model can also be expressed through a state–space representation and be treated entirely within the state–space framework. The state–space framework is essential when estimating linear models using the Kalman filter. The state–space representation is divided in two parts

$$\vec{x}_n = F_n\vec{x}_{n-1} + G_n\vec{v}_n \quad \text{(system model)} \tag{2.10}$$
$$\vec{y}_n = H_n\vec{x}_n + \vec{w}_n \quad \text{(observation model)} \tag{2.11}$$

where $\vec{x}_n$ is a k–dimensional unobservable vector commonly referred to as the state, $\vec{v}_n$ is the system noise, or state noise; an *m*–dimensional vector with zero mean and a covariance matrix Q_n. The vector $\vec{w}_n$ is an ℓ–dimensional vector with Gaussian white noise just like the vector $\vec{v}_n$ but with covariance matrix R_n. So, the dimensions of the matrices F_n, G_n, and H_n and the vectors $\vec{x}_n$, $\vec{y}_n$, $\vec{v}_n$ and $\vec{w}_n$ looks like this

$$[k\times 1] = [k\times k][k\times 1] + [k\times m][m\times 1]$$
$$[\ell\times 1] = [\ell\times k][k\times 1] + [\ell\times 1]$$

Generally, the vector $\vec{y}_n$ in the observation model (2.11) represents observations, or measurements of some data that is unobservable due to some uncertainty that is inherent in the process of retrieving that data. If the observation model in (2.11) is considered to be a regression model that represents a mechanism for obtaining the time–series $\vec{y}_n$, then the vector $\vec{x}_n$ corresponds to the regression coefficients of that model and the system model in (2.10) represents the time–change of these coefficients. If the vector $\vec{x}_n$ is considered to be an unknown input signal, then the system model represents the generation mechanism of that signal, and the observation model represents the structure of the observed signal.

2.4.1 State–space representation of the AR model

If we consider the autoregressive model just as in (1.1)

$$y_t = \sum_{i=1}^{p} a_i y_{t-i} + v_t \tag{2.12}$$

then the state–vector is defined as $\vec{x}_t = (y_t, y_{t-1}, \ldots, y_{t-p+1})^{\mathrm{T}}$. Since the model is stationary, the parameters are time–independent and the system matrix in (2.10) is time–invariant, i.e. $F_n = F_t = F$. Thus F is a $p\times p$ matrix and the system noise matrix G is a time–invariant *p*–dimensional vector where

$$F = \begin{bmatrix} a_1 & a_2 & \cdots & a_{p-2} & a_{p-1} & a_p \\ 1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 & 0 \\ 0 & 0 & \cdots & 0 & 1 & 0 \end{bmatrix}, \quad G = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}. \tag{2.13}$$

Since the first element in the state–vector $\vec{x}_t$ actually is the observation $\vec{y}_t$, we can define the observation model by setting the observation matrix H as a time–invariant vector where $H = [1\ 0\ \ldots\ 0]$. Since we don't have any observation noise, we end up with the following models

$$\vec{x}_t = F\vec{x}_{t-1} + G\vec{v}_t \quad \text{(system model)} \tag{2.14}$$

$$y_t = H\vec{x}_t \quad \text{(observation model).} \tag{2.15}$$

The system noise is a univariate Gaussian stochastically independent variable as defined in the AR(p) model with a time–invariant 1×1 dimensional covariance matrix $Q = \sigma^2$ and the covariance matrix for the observation noise is $R = 0$. So, in the state–space framework, the autoregressive model is a special case where the state–vector $\vec{x}_t$ is completely determined up to time t without any observation noise. It should be noted that this state–space representation is by no means a unique representation of the autoregressive model and there are other state–space representations for the very same model. Another way to find a state–space representation for the autoregressive model is by defining a step–ahead predictor. So, given a series of observations up to time $t - 1$ and the autoregressive model we can find a step–ahead prediction $\tilde{y}_{t+i|t-1}$ at a future time $t + i$ by

$$\tilde{y}_{t+i|t-1} = \sum_{j=i+1}^{p} a_j y_{t+i-j} \tag{2.16}$$

that must be recursively evaluated one step at a time until time $t + i$. If we define the state–vector $\vec{x}_t$ in terms of step–ahead predictors as

$$\vec{x}_t = \begin{bmatrix} y_t \\ \tilde{y}_{t+1|t-1} \\ \vdots \\ \tilde{y}_{t+p-1|t-1} \end{bmatrix} \tag{2.17}$$

we can define the matrices in the system model and the observation model as

$$F = \begin{bmatrix} a_1 & 1 & 0 & \cdots & 0 & 0 \\ a_2 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{p-2} & 0 & 0 & \cdots & 1 & 0 \\ a_{p-1} & 0 & 0 & \cdots & 0 & 1 \\ a_p & 0 & 0 & \cdots & 0 & 0 \end{bmatrix}, \qquad (2.18)$$

$$G = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad H = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix},$$

which is another state–space representation of the linear autoregressive model.

2.4.2 State–space representation of the ARMA model

If we consider the autoregressive moving average model as in (1.5)

$$y_t = \sum_{i=1}^{p} a_i y_{t-i} + v_t + \sum_{kj=1}^{q} c_k v_{t-k} \qquad (2.19)$$

we can find a state–space representation for it using step–ahead predictors. The step–ahead based state–space representation considerably simplifies the moving average part of the system model which is why we choose this form for the ARMA model. The moving average coefficients can then be defined with the system noise as a time–invariant [1 × q+1] matrix G where

$$G = \begin{bmatrix} 1 \\ c_1 \\ \vdots \\ c_q \end{bmatrix} \qquad (2.20)$$

and the covariance matrix for the system noise is $Q = G \cdot G^{\mathrm{T}}$. The rest is the same as in the state–space representation for the autoregressive model, i.e.

$$F = \begin{bmatrix} a_1 & 1 & 0 & \cdots & 0 & 0 \\ a_2 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{p-2} & 0 & 0 & \cdots & 1 & 0 \\ a_{p-1} & 0 & 0 & \cdots & 0 & 1 \\ a_p & 0 & 0 & \cdots & 0 & 0 \end{bmatrix}, \text{ and} \qquad (2.21)$$

$$H = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}.$$

For the dimensions to add up, the matrices and vectors are set up as

$$[s \times 1] = [s \times s][s \times 1] + [s \times 1][1 \times 1]$$
$$[1 \times 1] = [1 \times s][s \times 1] + [1 \times 1]$$

where $s = \max(p, q + 1)$. If the order p is larger than $q + 1$ or the other way around, the overshooting elements will be zero and not distort the state–space representation of the model.

2.4.3 State–space representation of the ARMAX model

If we consider the autoregressive moving average model with an exogeneous influence as in (1.8)

$$y_t = \sum_{i=1}^{p} a_i y_{t-i} + \sum_{j=1}^{r} b_j u_{t-d-j} + v_t + \sum_{kj=1}^{q} c_k v_{t-k} \qquad (2.22)$$

we need to redefine the system model such that we also account for the external influence in the state–space representation

$$\vec{x}_t = F\vec{x}_{t-1} + E\vec{u}_{t-1} + G\vec{v}_{t-1} \quad \text{(system model)} \qquad (2.23)$$
$$y_t = H\vec{x}_t \quad \text{(observation model).} \qquad (2.24)$$

which could be seen as a more generic version of the state–space model. So the previous cases are then special cases of the state–space representation where $E = 0$. We once again assume that the transition matrices are time–invariant and that the parameters are stationary. The state–space representation for the ARMAX model is then the same as in the case for the previously mentioned state–space representation for the ARMA model but with

$$E = \begin{bmatrix} 0_d \\ b_1 \\ \vdots \\ b_r \end{bmatrix}, \quad 0_d = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}, \tag{2.25}$$

where 0_d is a delay vector with d zero–elements. So $\vec{u}_t$ is an observation of the external influence at time t and in fact, a scalar just like the $\vec{v}_t$ in the system noise term is a scalar. Here we have $s = \max(p, q + 1, r + d + 1)$. The validity of this state–space representation can be shown by deriving the equation line–by–line starting at the bottom row of the system model equation and applying a lag operator to get to the top equation. Through this procedure one ends up with Equation (2.22). A more thorough treatment can be found in [Hamilton 1994].

2.5 Kalman filtration

This Section outlines the Kalman filtration framework beginning by discussing the principles behind it. It begins by introducing the iterative prediction–filtration process and formulating it as a minimization problem. The second half of this Section discusses ways to set initial values for the parameters and the system covariance matrix before running the Kalman filtration.

2.5.1 The principles of Kalman filtration

The general approach to estimate the state in a state–space model is to obtain the conditional distribution $p(x_t|Y_j)$ of the state x_t for a given set of prior observations $Y_j = \{y_1, \ldots, y_j\}$ [Kitagawa 2010]. Since the state–space model is a linear model with Gaussian noises and an initial state x_0 that is normally distributed, the conditional distributions to be estimated also become normal distributions. Since the only parameters that completely determine such distributions are the mean and variance, it is sufficient to estimate the mean vectors and the variance–covariance matrices of the conditional distributions. However such estimations require a vast amount of computations. The Kalman filtration technique is a recursive state estimation technique given a set of past observations that greatly reduces the amount of computations required for such estimations. At large [Welch Bishop 2001] identifies four major operations that are provided by the Kalman framework

- **One–step prediction** – This operation predicts the state and the error covariance in the next time–step (say $t + 1$) given a set of assumptions (that are given in the matrices in the system part of the state–space model) and a series of past observations

- **Filter** – When a new observation has been collected in the next time–step, this operation extracts this observation from the observation part of the state–space model and adjusts the prior prediction accordingly. The adjustment will be weighted in favor of the factor with the lowest noise; if the system noise is lower than the observation noise the adjustment will be weighted towards the prediction.
- **Smoothing** – If there are missing past observations, this operation can estimate these missing observations given future observations. It is similar to prediction but running backwards.
- **Increasing horizon prediction** – This is similar to the one–step prediction, but this time there is no filtering in between and the same matrices in the system equation are reapplied repeatedly without adjustment.

In short, to estimate a linear model such as the ARMA(p, q) or ARMAX(p, q, r) model the procedure is as follows

1. Estimate an initial state and an initial covariance matrix
2. Take the first element of a time–series of given observations and adjust the initialization state with it using the Kalman–filter
3. Make a prediction for the next time–step using the adjusted state using the prediction operation
4. Adjust the prediction against the observation at that time–step using the Kalman–filter
5. Go back to the third calculation step until the last observation is reached

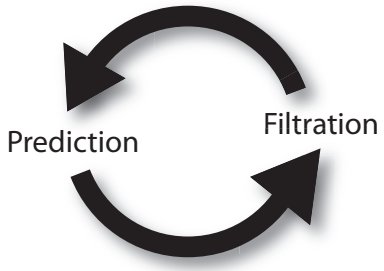

Figure 2.2: An illustration of the recursive iteration process behind Kalman–filtration.

Source: Kitagawa 2010

So we are essentially using a for–loop on calculation steps 3–5 that is visualized in **Figure 2.2**. If we let the conditional expectation and the variance–covariance matrix for the state x_{t+1} be denoted by

$$\begin{aligned}\hat{x}_{t+1|t} &= E\left[x_{t+1} \mid Y_t\right] \\ \Sigma_{t+1|t} &= E\left[\left(x_{t+1}-\hat{x}_{t+1|t}\right)\left(x_{t+1}-\hat{x}_{t+1|t}\right)^T\right],\end{aligned} \tag{2.26}$$

then we can define the prediction and filtration algorithms in the Kalman framework as follows [Welch Bishop 2006]:

Prediction

I. Project the next state

$$\hat{x}_{t+1|t} = F_{t+1}\hat{x}_{t|t} + E_{t+1}u_{t|t} \tag{2.27}$$

II. Project the next error–covariance

$$\Sigma_{t+1|t} = F_{t+1|t}\Sigma_{t|t}F_{t+1|t}^T + G_{t+1}Q_{t+1}G_{t+1}^T \tag{2.28}$$

Filtration

I. Compute the Kalman gain

$$K_{t+1} = \Sigma_{t+1|t}H_{t+1}^T\left(H_{t+1}\Sigma_{t+1|t}H_{t+1}^T + R_{t+1}\right)^{-1} \tag{2.29}$$

II. Update the prediction with an observation at y_{t+1}

$$\hat{x}_{t+1|t+1} = \hat{x}_{t+1|t} + K_{t+1}\left(y_{t+1} - H_{t+1}\hat{x}_{t+1|t}\right) \tag{2.30}$$

III. Update the error covariance

$$\Sigma_{t+1|t+1} = \Sigma_{t+1|t} - K_{t+1}H_{t+1}\Sigma_{t+1|t} \tag{2.31}$$

A proof of Equation (2.27) – (2.31) is provided in Appendix A.1. We are facing a minimization problem where we want to minimize the squared sums of the errors or innovations α_t which is equal to minimizing the following equation [Hamilton 1994 p132]

$$l = \sum_{t=0}^{N-1}\log(\omega_t) + N\sum_{t=0}^{N-1}\log\frac{\alpha_t^2}{\omega_t}, \tag{2.32}$$

where

$$\omega_t = H_t\Sigma_{t|t-1}H_t^T + R_t \tag{2.33}$$

and

$$\alpha_t = y_t - \hat{y}_{t|t-1} = y_t - H_t\hat{x}_{t|t-1} \tag{2.34}$$

for some observations $y_0, y_1, \ldots, y_{N-1}$. A proof is given in Appendix A.2.

2.5.2 Setting initial values for Kalman filtration

Selecting initial values for the Kalman procedure could be achieved by using a less precise method for estimating the ARMA or the ARMAX model. Hannan, McDougall [Hannan McDougall 1988] suggest using the least–squares method in two steps to estimate the ARMA model. An alternative is to use the Yule–Walker method as described in Section 2.3. In the first step, pure AR(p) models are estimated by selecting the order $p = 1, \ldots, \log(T)^{1.5}$ that yields the AR(p) model with the lowest Bayesian information criterion, let's call it p_{BIC}. Then in the next step AR(p) is estimated for order $p = p_{\text{BIC}} + \max(p, q)$. The first p elements of this estimation are the autoregression coefficients and the next q elements in this estimation are the moving average coefficients. The estimation of these parameters can then be used as initial values for $a_1, \ldots, a_p$ and for $c_1, \ldots, c_q$ in the Kalman estimation procedure. Finding initial values for ARMAX estimations using the methods of Hannan, McDougall and Hannan, Kavalieris [Hannan Kavalieris 1984] results according to them in a considerable increase in complexity. But since we only want to yield an estimation that is fairly close we make do with estimating AR(p) for $p = p_{\text{BIC}} + \max(p, q, r)$ and let the next q elements after the first p elements of the estimation be initial values for $c_1, \ldots, c_q$ and the r elements following the first p elements be initial values for $b_1, \ldots, b_q$ in the ARMAX model to be estimated with the Kalman method.

The system covariance should also be set before the Kalman iterations begin, a method to compute the initial system covariance is provided in Appendix A.3.

2.6 Trend, seasonality and the Box–Jenkins framework

The Box–Jenkins framework provides a methodology for identifying trend and seasonality components in a time–series. The work of Box–Jenkins is based on ARMA models which are sometimes are called Box–Jenkins models. In the basic Box–Jenkins framework, the trend and seasonality cycles are assumed to be deterministic and formulated as [Brockwell Davis 2006 p15]

$$X(t) = m(t) + \sum_i s_i(t) + Y(t), \tag{2.35}$$

where $Y(t)$ is a stationary stochastic process, $m(t)$ the trend component and $s_i(t)$ seasonality components. The trend component to be estimated could either be a linear function or a polynomial that is fitted to the data. Another more flexible way to estimate the trend is to use a moving average smoothing function. This moving average can be estimated by

$$\hat{m}(t) = \frac{1}{2q+1} \sum_{j=-q}^{q} X(t+j), \tag{2.36}$$

which is defined for $q + 1 \le t \le T - q$ for some interval $2q + 1$ which has an odd number of elements. If one wishes to use a moving average over an even interval, the following smoothing function could be used

$$\hat{m}(t) = \frac{1}{4q}(X(t-q) + X(t+q)) + \frac{1}{2q} \sum_{j=-q+1}^{q-1} X(t+j), \tag{2.37}$$

where the end–terms at $\pm q$ are given only half weight so as to give the moving average the same weight forward as backward when the interval has an even number of elements. This smoothing function can also be looked upon as a high–pass filter that removes the (slower moving) trend from the data. A seasonality cycle is a variation with a periodicity d where

$$s_i(t) = s_i(t+d). \tag{2.38}$$

This seasonality can have any shape as long as it is d–periodic. A more sophisticated approach is to assume that there is some level of interaction between the trend and seasonality cycles [Koopman Lee 2008] and [Chen Vidakovic Mavris 2004]. For simplicity we assume that the seasonality cycles are sinusoidal and non–interacting with the trend. This enables us to use Fourier methods to detect and filter out seasonality cycles using band–stop filters.

With Fourier analysis, and a proper frequency or period scaling, seasonality cycles can easily be detected and filtered out. Time–series data can be transformed to the frequency spectrum using the FFT algorithm. The FFT is an algorithm used to compute the discrete Fourier transform (DFT) using less computations. The DFT is defined as follows [Folland 1992]

$$\hat{f}(\omega) = \sum_{i=0}^{T} f(t_i) e^{-2\pi i \omega t_i}, \tag{2.39}$$

where ω is the frequency of the spectral density $\hat{f}(\omega)$. The highest frequency that we can truly see in the Fourier space is set by Nyquist's sampling theorem which says that if an input signal contains no frequencies higher than ξ Hz, then it will be fully represented if it is sampled at a rate of 2ξ Hz, i.e. at twice that frequency (for an unlimited time–period). So, we cannot expect to find true representations for frequencies at higher than half the sampling rate of the observations. The lowest frequency is limited by the time–period the measurements that are taken, i.e. periodicities that are close to the sampling period cannot be differentiated from the trend as their recurrences span beyond the sampling period.

A good way to filter out seasonality cycles, once they have been identified is to use a set of suitable filters. As we have alluded previously, the models to be fitted into the observed stochastic observations can be looked upon as a dynamical system. Since the AR, MA, ARMA and ARMAX models are all linear and time–invariant, they have an impulse response function where the output signal from the model can be described as a convolution between an input signal and its impulse response function [Petersson], i.e. for an input signal $x(t)$, the output signal $y(t)$ is

$$y(t) = h(t) * x(t) = \int_{-\infty}^{\infty} h(t)\, x(t-\tau)\, d\tau, \qquad (2.40)$$

where $h(t)$ is the impulse response function. The Fourier transform of a convolution turns into a product in the Fourier space which makes it particularly easy to find a suitable filter that removes the seasonality components from the data.

Some things to consider when applying Fourier analysis is that Fourier transforms are prone to artifacts that "color" the spectrum when there are discontinuities in the data. This stems from the so–called Gibbs phenomenon. There are less artifact–prone methods such as Wavelet analysis if such artifacts are a concern. Also, the filters work best when applying them on detrended data.

2.7 Evaluating goodness–of–fit

When finding and fitting a statistical model to observed data, we strive to make the model represent the observed data as closely as possible. For example, when fitting the parameters of the AR model, there is an optimization process that minimizes the errors of the estimations or in the case of ARMA estimations, the best estimation is the one with the highest likelihood, or log likelihood fit. There are different criteria than just the maximum likelihood or the least squared errors to use to determine goodness–of–fit of a particular model. The ideal way of measuring goodness–of–fit is by comparing the

underlying empirical distribution with the distribution of the estimation. The problem however, is that the underlying empirical distribution behind the observations is generally not known so one has to find a way to measure the goodness–of–fit without knowing it. The likelihood measure uses the law of large numbers which says that as the number observations tend to infinity, their distribution tends to the true underlying distribution. The downside to it is that it is only applicable to a model with a fixed number of parameters. Information criteria such as Akaike's Information Criterion or Bayes Information Criterion extend this likelihood function to models with different numbers of parameters. The basic assumption among these criteria is that the uncertainty increase as the number of parameters to be estimated increase. When using a model selection approach based on the information criteria, the estimation that yields the lowest value of the information criteria is then said to be the best fit. The Akaike Information criteria is defined as follows

$$\mathrm{AIC} = 2k - 2\log(L), \tag{2.41}$$

where k represents the number of parameters in the model and L the maximum likelihood measure of the model estimation. Another measure is Akaike's final prediction error, which is defined as follows

$$\mathrm{FPE} = V\frac{N+k}{N-k} \tag{2.42}$$

for some loss function V defined as

$$V = \det\left[\frac{1}{N}\sum_{i=1}^{N}\varepsilon_i\varepsilon_i^T\right] \tag{2.43}$$

where ε_i are the error terms of the model. The downside of such measures is that they don't actually tell the goodness–of–fit or whether the estimation truly follows anything that is similar to the empirical distribution. To evaluate the goodness–of–fit we need to make simulations of our estimations and compare the distribution of those simulations to the empirical distribution, an approach that requires considerably more computations than the information criteria above. The observation can then be assumed to approximately represent the empirical distribution. One such measure that makes these assumptions is the Kolmogorov–Smirnov distance which is defined as follows [Massey 1951]

$$D_n = \sup_x |F_n(x) - F(x)|, \tag{2.44}$$

where $F_n(x)$ is the cumulative distribution of the n observations and $F(x)$ is the cumulative distribution of the model simulations. Another approach to measure the closeness to the empirical distribution is the following

$$H_n = \int_{-\infty}^{\infty} |F_n(x) - F(x)| dx . \qquad (2.45)$$

The lower the value of the measure is, the better the fit.

3. Results

In this Chapter we present the results from our simulations and estimations of the ARMA(p, q) and ARMA(p, q, r) models on the electricity spot price data as retrieved from Nord Pool Spot A/S. The first Section evaluates the Box–Jenkins framework on well known and defined simulations. It also evaluates the Kalman filtration method and the method used by the internal Matlab ARMAX package for model estimation. Section 3.2 evaluates the presence of trend and seasonality cycles in the electricity price data. Section 3.3 introduces observed 3–hour temperature data for 2010 as retrieved from SMHI. This data is converted to hourly temperature data by using piecewise cubic Hermite interpolation and then used as exogeneous input in the ARMAX estimations. The actual estimations are made and evaluated in Section 3.4.

3.1 Simulations and evaluation

In the first step we test our tools provided by the Box–Jenkins framework and check that they properly remove trend and seasonality cycles from a time–series. We do that by taking a set of randomly generated values from a white noise process where we add trend and seasonality data synthetically. So by taking 1000 samples from a white noise process, adding a 24–hour and 168–hour periodicity (by using cosines) and a trend following the function exp(t/1000) we end up with the process $X_{sim}(t)$ as in **Figure 3.1**.

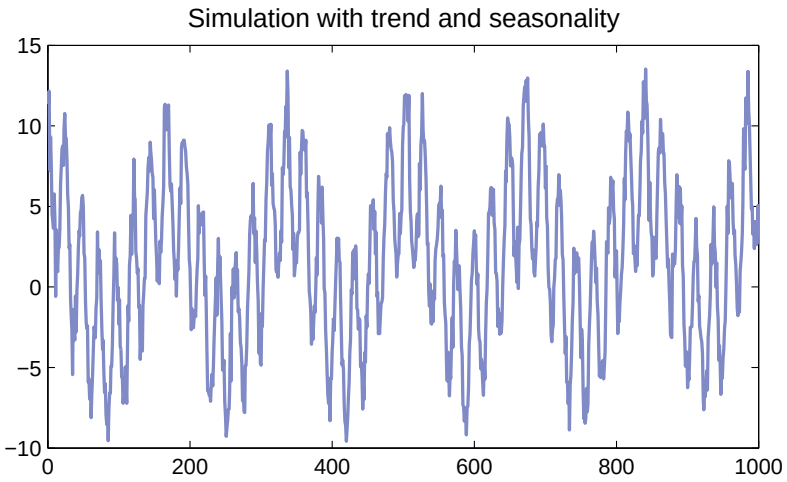

Figure 3.1: A simulated discrete stochastic process $X_{sim}(t)$ with an applied trend and seasonalities.

It is quite clear that the series is heavily influenced by periodicities. A Fourier analysis is shown in **Figure 3.2**.

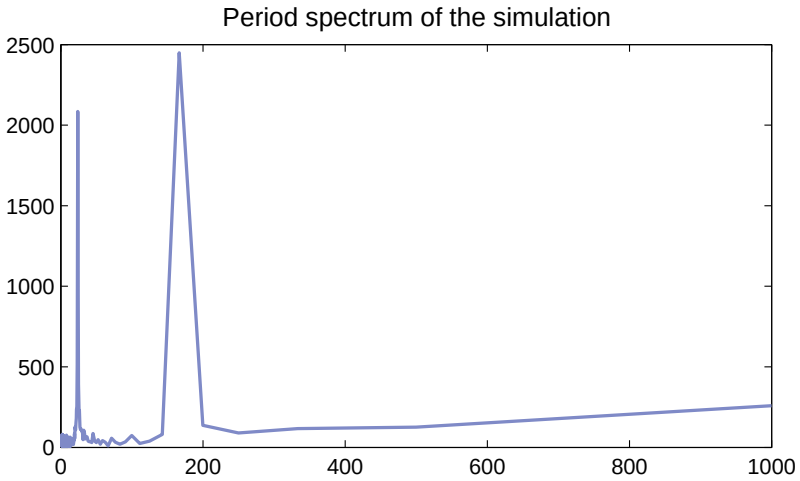

Figure 3.2: The process $X_{sim}(t)$ transformed with (2.39) using an FFT algorithm and scaled with respect to periodicity.

The trend is also easily detected by calculating the moving average for the process. The 168–hour trend is shown in **Figure 3.3**.

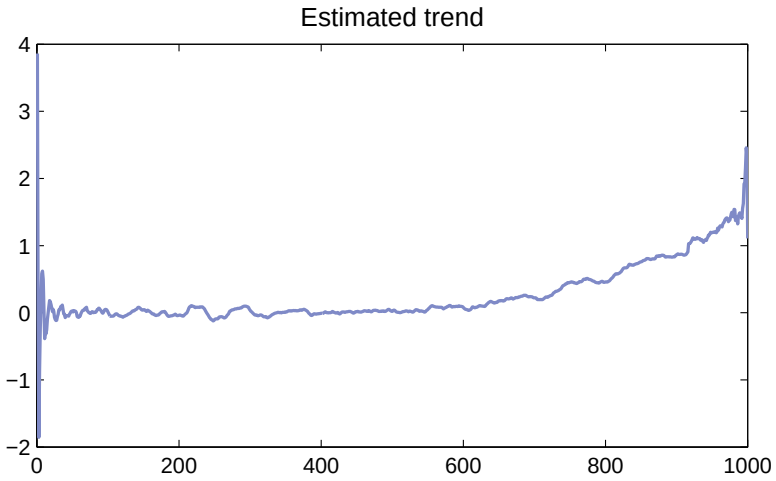

Figure 3.3: The 168–hour trend extracted from the simulated process $X_{sim}(t)$.

We see that once the trend and seasonality components are removed from our simulated data, once again it looks like a white noise process as shown in **Figure 3.4**.

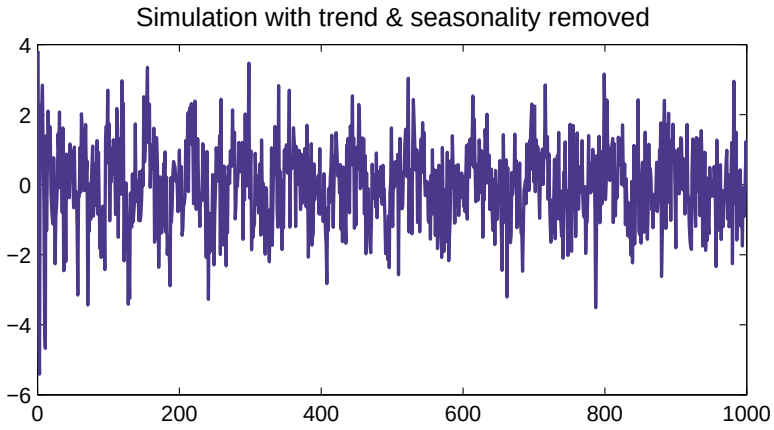

Figure 3.4: The resulting stochastic process extracted from $X_{sim}(t)$ with the trend and seasonalities removed from it.

So we conclude that our tools for deseasonalization and detrending do a proper job. In the next step we test how the Kalman filter and the internal Matlab function fits an ARMA(3, 3) model to the simulated data with trend and seasonality and the detrended/deseasonalized data.

$X_{sim}(t)$			
AR		MA	
Kalman	ARMAX	Kalman	ARMAX
2.9256	2.9261	-2.4534	-2.4572
-2.9195	-2.9206	1.9953	2.0166
0.9935	0.9941	-0.5133	-0.5336

$Y_{sim}(t)$			
AR		MA	
Kalman	ARMAX	Kalman	ARMAX
-0.6457	-0.5580	9.2819	0.7841
0.1401	0.3531	6.4188	-0.0869
0.1261	0.0183	0.3875	0.0574

Table 3.1: A comparison of estimated AR and MA coefficients of an ARMA(3, 3) model on simulated data using the Kalman–filtration algorithm and the internal ARMAX Matlab package. $Y_{sim}(t)$ is the detrended and deseasonalized process.

In **Table 3.1** we see that for the series with the trend and seasonalities present they yield roughly the same estimations for the ARMA coefficients whereas they yield different results when running the estimations on the filtered signals.

	Kalman		ARMAX	
	FPE	AIC	FPE	AIC
$X_{sim}(t)$	1.4372	0.4608	1.6728	0.6127
$Y_{sim}(t)$	0.0172	-3.9653	1.2627	0.3314

Table 3.2: The information criteria and future prediction errors as defined by (2.41) and (2.42) evaluated for the estimations in **Table 3.1**.

The information criteria in **Table 3.2** suggests that both methods are roughly the same but that the Kalman method produce "better" results when there is no trend and seasonality present. The Kalman filter takes about 100 times longer to run the same estimations as the internal Matlab function. So we choose to stick with the internal Matlab function as long as we can get sane estimations with it.

3.2 Trends and seasonalities in electricity price data

Now that we have established that our tools for deseasonalization and detrending are sane, we move on to analyze hourly spot prices on Nordpool spot for the year 2010. The first thing we do is to see what seasonalities there are in our data. **Figure 3.5** shows periodicities that are up to 500 hours long.

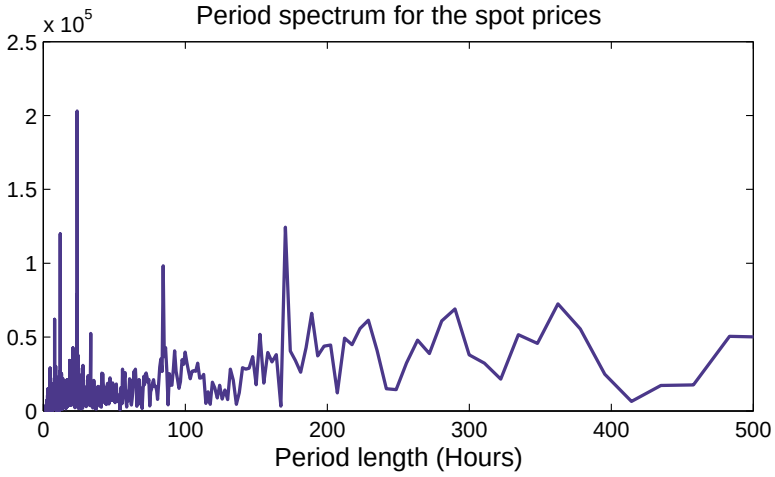

Figure 3.5: The estimated period spectrum for the hourly systemic spot prices during 2010 as noted on the Nord Pool Spot exchange.

We see several peaks that need to be removed from our data. A careful analysis indicates that we have periodicities at 4.77, 7.952, 11.93, 23.9 84.47, 170.6, and 2175 hours. We draw the conclusion that 11.93 and 23.9 are related to the 24 hour periodicity and 84.47 and 170.6 are related to the 168 hour periodicity.

The 4.77, 7.952 are a little harder to explain, perhaps they come from 8 hour working shifts. The most difficult periodicity to explain is the 2175 hour (or 90 day) periodicity. It seems like there is something that occurs roughly every 3 months. A set of Butterworth filters are designed as narrow bandstop filters and we successfully remove the periodicities from the time–series data. The results from the filtration vs. the original data are shown in **Figure 3.6**.

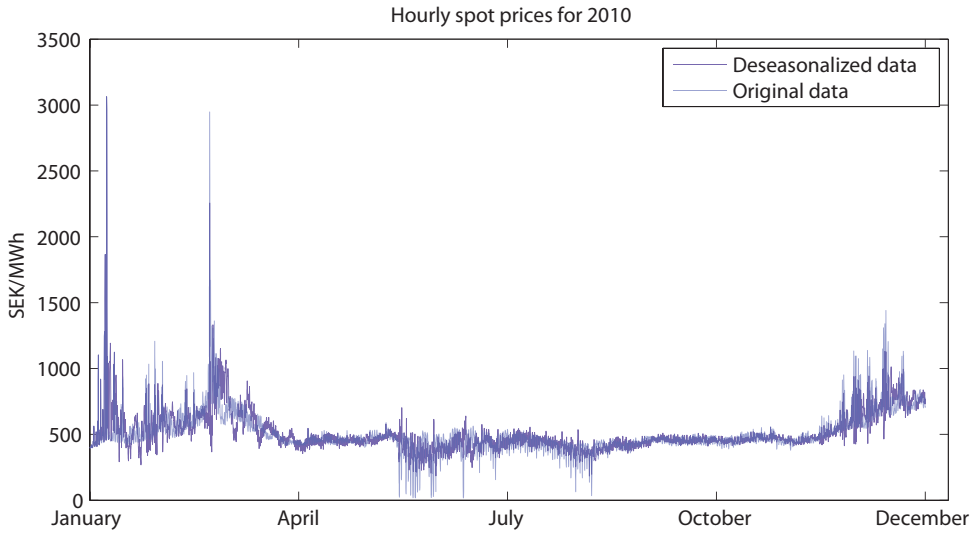

Figure 3.6: A comparison of deseasonalized electricity price data with the hourly systemic spot prices as noted on Nordpool Spot during the year 2010.
Source: Nord Pool Spot A/S

The electricity price data also has a trend. Different estimations for trend on the hourly spot price data are illustrated in **Figure 3.7**.

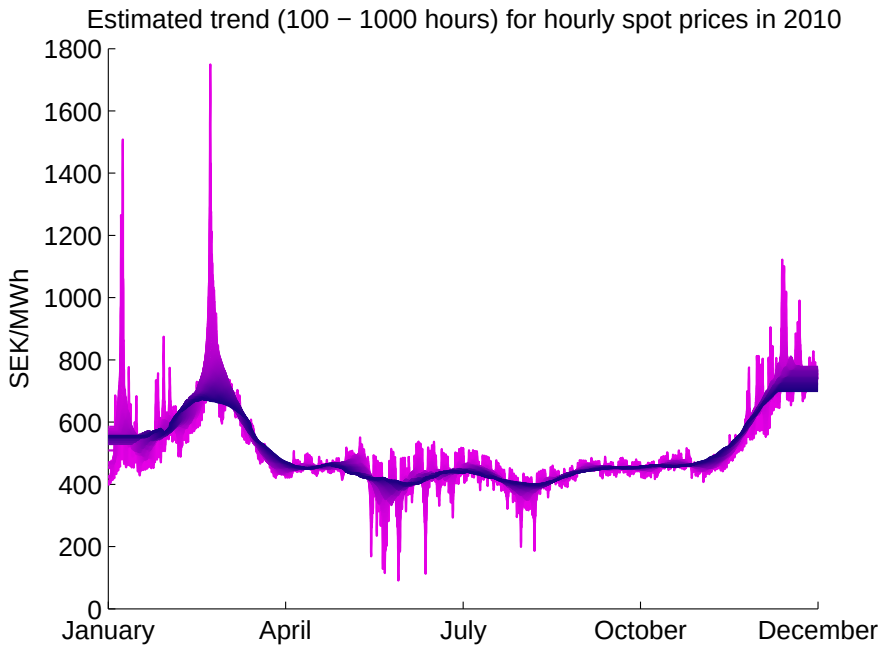

Figure 3.7: Moving average trends ranging from 100 to 1 000 hours added on top of each other using a hue that transitions from bright purple (100 hour moving average trend) to dark blue (1 000 hour moving average trend).

The trend was set as a 500 hours long moving average which was removed in the deseasonalization.

3.3 Using temperature data as exogeneous input

The temperature data used in the estimations come from SMHI (Sweden's Meteorological and Hydrological Institute) and they were sampled every third hour of the day during the year 2010. The weather–stations from where the data has been collected are illustrated in **Figure 3.8**.

Figure 3.8: The chosen weather stations from where the 3–hour temperature data for 2010 were collected.

Source: SMHI

From this data a national average temperature has been calculated. It is well known that most of the Swedish population lives in the south, so a 75% weight has been added to the temperature data retrieved from the three major cities in the south. So, the temperature data from the weather–stations in the north; Malung, Kuggören, Storlien, Hemavan, Haparanda, Jokkmokk and Katterjåkk only have a 25% weight. The data retrieved had been measured only every three hours and some temperature values were in fact missing. By using piecewise cubic Hermite interpolation, the missing values were filled in and the data were converted to hourly data. As we know, what is interesting is the difference between the average room temperature and outdoor temperature, assuming that the average room temperature is 20 °C, the temperature data were converted using $20 - x$. The national average temperature never exceeded this room temperature so we don't even need to deal with negative temperature differences. This modified temperature data constitutes the external variable that we use to fit the ARMAX models. **Figure 3.9** clearly shows that there is a

considerably higher need for heating during the winter months than for the summer months.

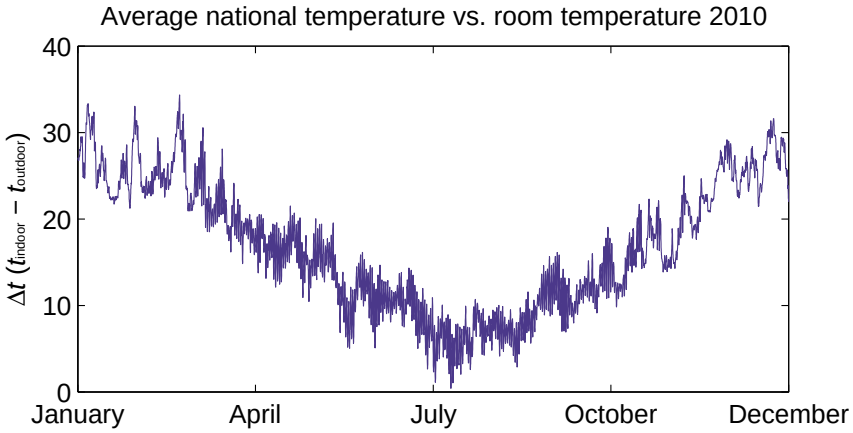

Figure 3.9: The hourly temperature as a weighted nationwide average where 75 % weight was put on the three major cities in the south and 25 % weight was put on the temperature data from the weather stations in the north.

Source: SMHI

3.4 Fitting ARMAX models on electricity price data

In this Section we fit ARMA(p, q) and ARMAX(p, q, r) models to hourly and daily observed electricity price data for 2010. We make the estimations over parameter spaces (p, q) and (p, q, r) respectively and chose the best fit according to a criterion. In the first two Sections we primarily focus on the Akaike information criteria (AIC) as defined by (2.41) and in the rest of the Sections we also use a modified Kolmogorov–Smirnov framework as defined by (2.45) to find the best fit on simulations of our estimations. Estimations with the lowest AIC and/or lowest K–S measure are considered to be "winners". We also explore how the quality of our estimations are affected by the period chosen; we compare estimations on the whole dataset with selected sub–periods such as the first three (winter) months and three months in the late summer / early autumn (July – October). Transformations of the observed data can also affect the estimability; we look into estimations on the log() – transformed price data and on price data that is divided by 1000.

3.4.1 Hourly data

In the first step we test how deseasonalization and detrending affects the goodness of the fit. Using the Akaike information criterion and prediction

errors we see that deseasonalization and detrending indeed makes the ARMAX model easier to fit to the data. We also compare fitting the whole year data to the model vs. only fitting the first 3 (winter) months of the data. This comparison shows that the winter months are more difficult to fit than the rest of the data. We try over a parameter space of p = 1, …, 20, q = 1, …, 20, r = 1,…,3 and see what solution yields the best/lowest Akaike Information Criteria. The results with the best p, q, r values are given in **Table 3.3**.

	p_{best}	q_{best}	r_{best}	AIC
$1y_{orig}$	20	11	3	7.2824
$1y_{deseason}$	17	12	3	7.2693
$1y_{deseason,\ detrend}$	20	5	3	7.2520
$3m_{orig}$	19	19	3	8.1876
$3m_{deseason}$	19	19	3	8.1793
$3m_{deseason,\ detrend}$	19	19	3	8.1723

Table 3.3: The ARMAX estimations with the lowest AIC on different variants of the spot price data as estimated on the whole year of data and the first 3 months (January–March) of the data.

In both cases we see that deseasonalization and detrending lowers the information criteria. We also see that we have corner solutions so we are not near an optimal fit in this parameter space. Attempts were made at fitting the models to data in the 3 month period after the winter period above. This resulted in halved information criteria. It stands clear that the spikes prevent the model from properly fitting to the data and that we need to deal with these spikes if we want to get proper fits. Also attempts at using higher orders for r were made. By setting p = 20 and q = 5, the optimal order for r was found at r = 27. The problem is illustrated in **Figure 3.10** and **Figure 3.11**.

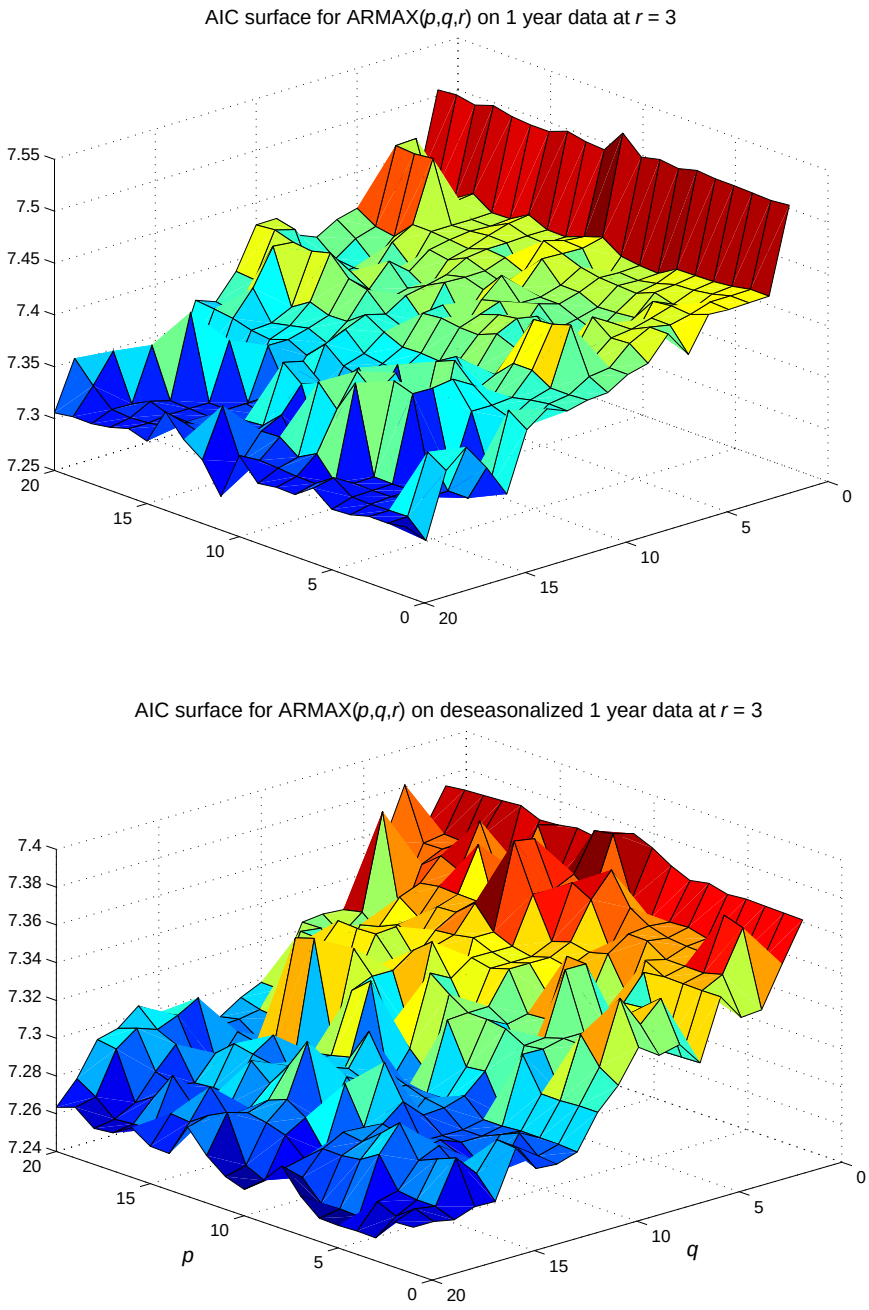

Figure 3.10: The AIC surfaces for ARMAX(p, q, 3) estimations on the whole dataset. The upper surface represents estimations run directly on the dataset whereas the lower surface represents estimations on detrended and deseasonalized data.

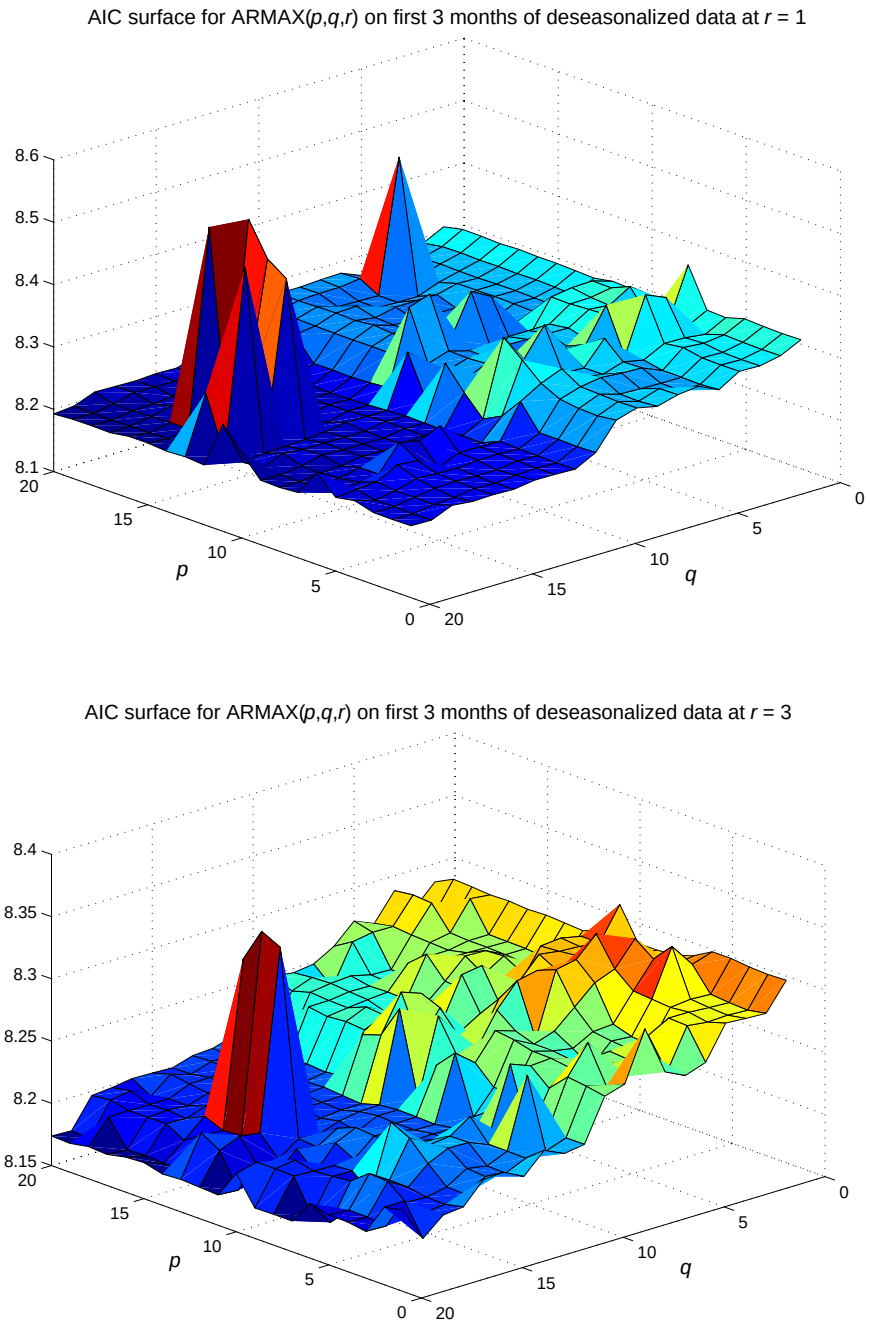

Figure 3.11: The AIC surfaces for ARMAX(p, q, 3) estimations on the first three months (January–March) of spot price data. The upper surface represents estimations run directly on the dataset whereas the lower surface represents estimations on detrended and deseasonalized data.

We see from **Figure 3.10** and **Figure 3.11** that the AIC is rather big and there is no obvious minimum to choose from. It also quite clearly suggests that there may be a higher order fit outside this parameter space that yield lower AIC values. The first row of the figure compares ARMAX estimations on the original data with ARMAX estimations on deseasonalized data and it is clear that the deseasonalization allow for better estimations. The lower row looks at the first three months of the data and compares ARMAX estimations using first order temperature influence with third order temperature influence and we also see that the third order influence is a better fit than the first order. Similar conclusions can be drawn from the FPE meshes which are omitted from this thesis.

3.4.2 Hourly log() – transformed data

In the next step we examined the possibility of fitting the ARMAX model to log() data. We looked at the whole year and the first three months using the same parameter space as before which yielded some interesting results which are concluded in **Table 3.4**.

	p_{best}	q_{best}	r_{best}	AIC
$1y_{orig}$	20	18	3	-5.0428
$1y_{deseason}$	20	8	3	-5.7913
$1y_{deseason,\ detrend}$	20	18	3	-5.7936
$3m_{orig}$	19	17	2	-5.6269
$3m_{deseason}$	18	15	3	-5.3446
$3m_{deseason,\ detrend}$	20	19	3	-5.3432

Table 3.4: The best estimations of ARMAX(p, q, r) on log() price data as chosen by AIC just like in **Table 3.3**.

We now see that the AIC values are negative which suggests that our estimations are better than without the log(). Unfortunately, the results once again suggest that we have corner solutions. Selecting sub–data for hours 2000 – 4400 instead of the first three months does not make a better fit than the whole data but the fit is not much worse either when tweaking the parameters. Further tweaking of the parameters yielded an even lower AIC value; a "winner" was found at p = 84, q = 18, r = 31, yielding AIC = –5.9698 which is even lower than the results in the table **Table 3.4**. **Figure 3.12** shows the FPE surfaces for the log() – transformed estimations.

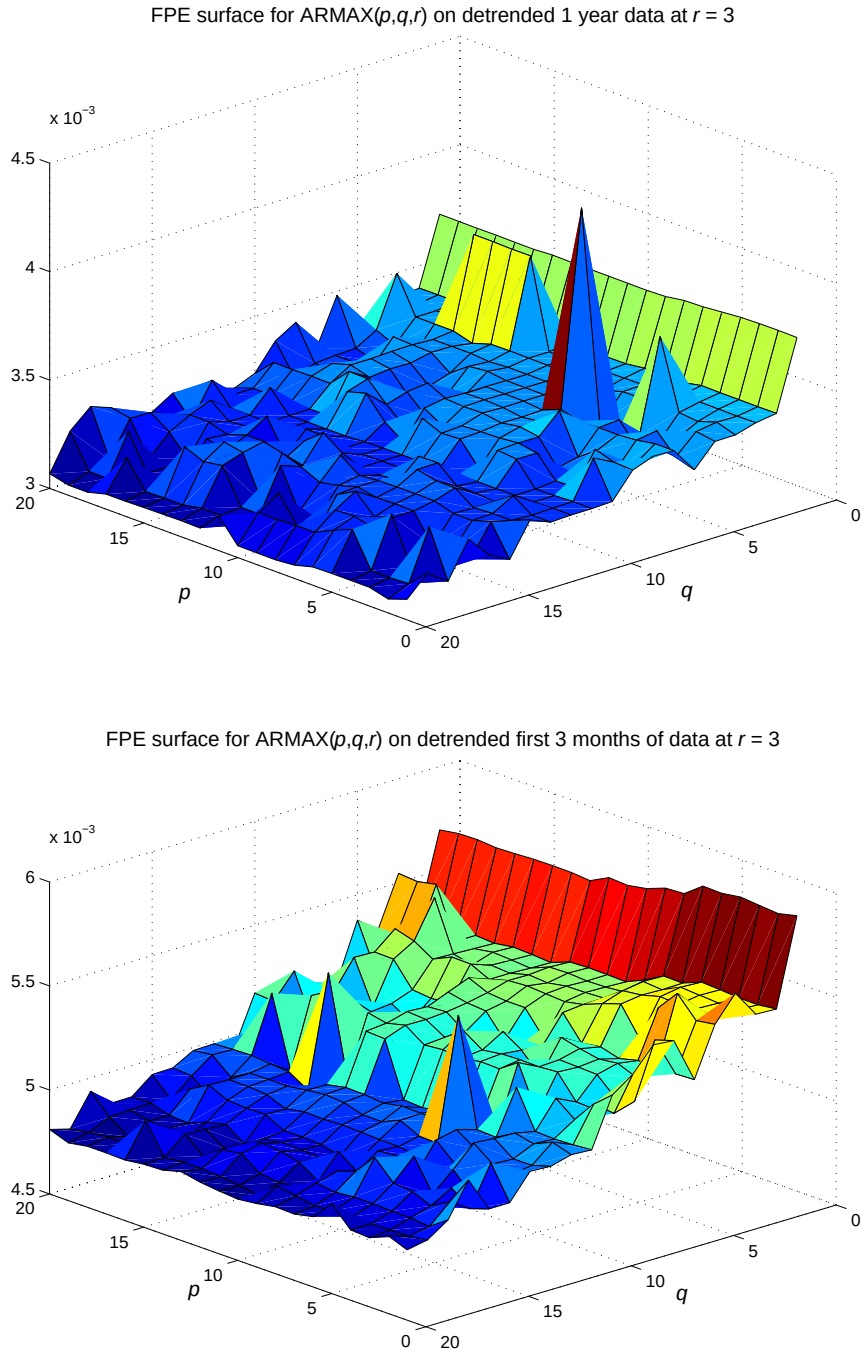

Figure 3.12: The AIC surfaces for ARMAX(p, q, 3) estimations on the detrended/deseasonalized and log() – transformed dataset. The upper figure represents estimations on the whole dataset whereas the lower figure represents estimations on the first three months of the dataset.

We can see in **Figure 3.12** that the surface for the three months data is steeper than when fitting the whole year. An attempt to fit a pure ARMA(p, q) model on the log() – transformed data was also made over the parameter space p = 1,…, 30 and q = 1,…, 30. The lowest AIC solution was achieved at p = 30 and q = 30 yielding AIC = –5.9337 which regrettably is a corner solution.

3.4.3 Hourly data and evaluation using simulations

In the next step we tried to fit the ARMAX(p, q, r) and simulate the estimated model. The cumulative distribution function of the simulation was compared to the estimated empirical distribution and the estimation that lies closest to the empirical distribution is the "winner". Also, we found another way to transform the electricity price data to make it more estimable. By dividing the data by 1000, the information criteria for the estimations went even lower than when using the logarithm. This time the estimations on the hourly spot price data were made for the period mid July to mid October so as to avoid the extreme behaviour that occurred during the winter months. The parameter space for these estimations were p = 1, …, 24, q = 1, …, 24, r = 1, …, 24. The results are concluded in **Table 3.5**.

	p_{best}	q_{best}	r_{best}	AIC	Δ
$ARMAX_{Jul–Oct}$	24	10	13	3.5331	0.0142
$ARMAX_{Jul–Oct}/1000$	3	17	13	-10.2824	1.1728
$ARMA_{Jul–Oct}$	1	8	–	3.5674	0.4592
$ARMA_{Jul–Oct}/1000$	2	7	–	-10.2481	1.1904

Table 3.5: ARMAX estimations on deseasonalized and detrended price data for the period July–October. The best candidates were chosen by the Kolmogorov–Smirnov measure in (2.45). "/1000" denotes that the data has been divided by 1000 before estimation.

The Kolmogorov–Smirnov comparisons for the estimations in **Table 3.5** are shown in **Figure 3.13**.

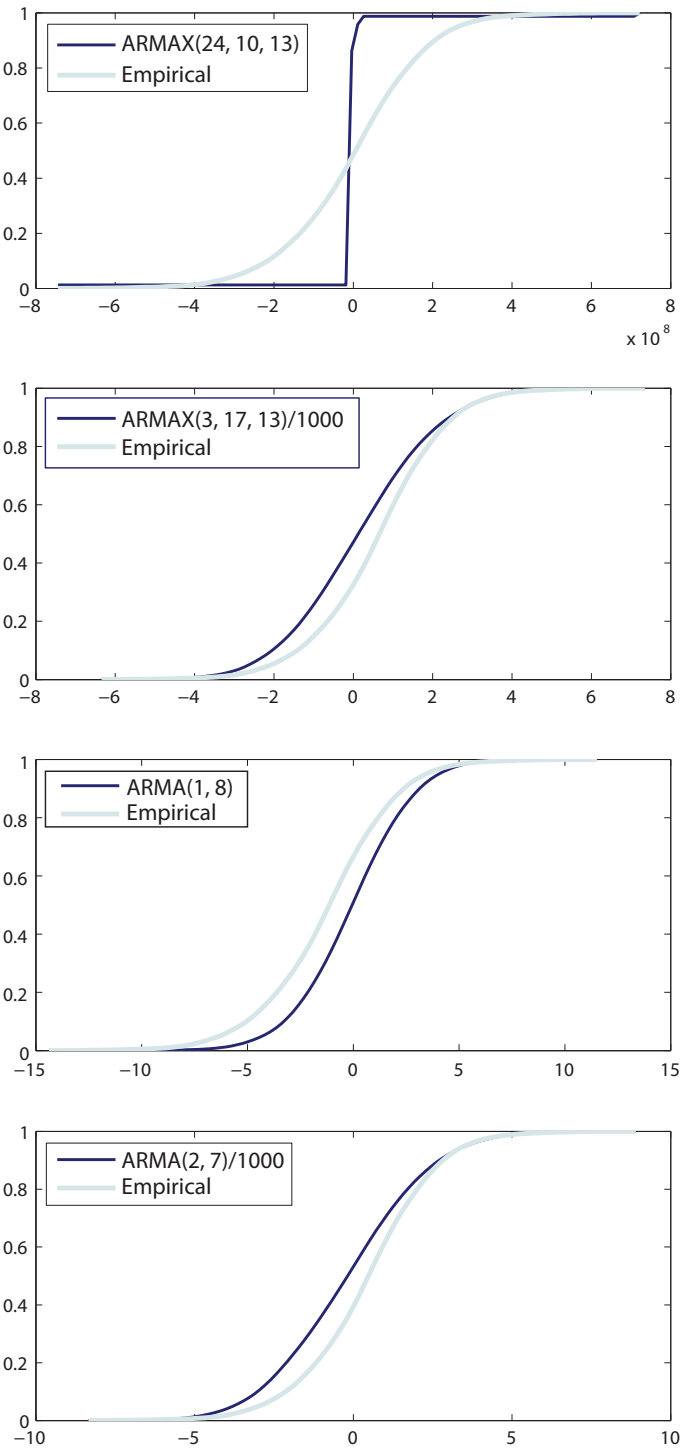

Figure 3.13: The cdf–functions of simulations of the estimations in **Table 3.5** compared to the cdf–function of the empirical distribution estimated from the observed data.

We can dismiss the first ARMAX estimation right away as it is non–stationary. A combined evaluation of simulations and the information criteria indicates that the ARMAX(3, 17, 13) is the best estimation and even better than the pure ARMA estimation. A realization of that estimation is visualized and compared to the original data in **Figure 3.14**.

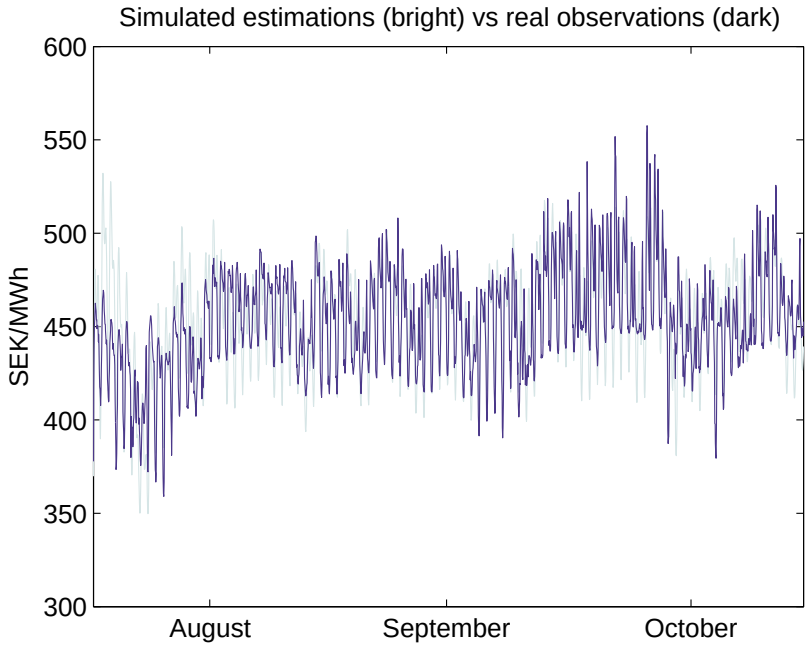

Figure 3.14: A simulation of the $ARMAX_{Jul–Oct}/1000$ (3, 17, 13) estimation that is compared directly to the real observed data.

3.4.4 Daily data and evaluation using simulations

In the next step we look at the daily electricity price data. The first step is to analyze for trend and periodicities. The period spectrum of the daily electricity spot price data is shown in **Figure 3.15**.

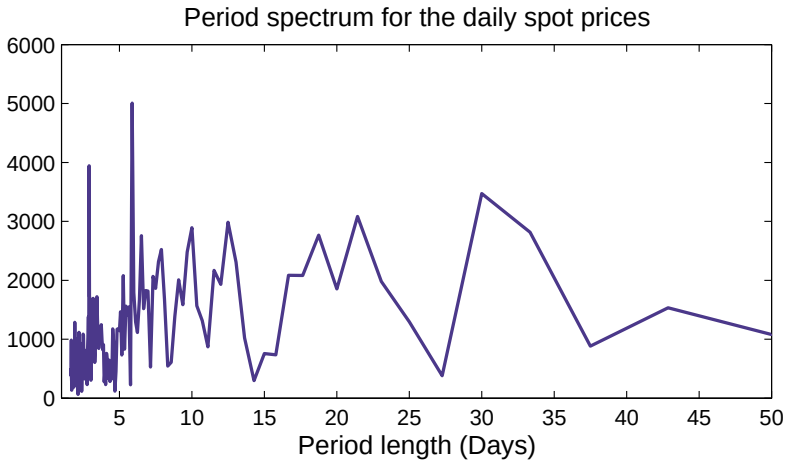

Figure 3.15: The estimated period spectrum for the observed daily systemic spot prices during 2010 as noted on the Nord Pool Spot exchange.

Source: Nord Pool Spot A/S

The largest peak is at 75 days that joins with a peak at 60 days and 100 days (not shown in **Figure 3.15**). Then there are two smaller peaks at 2.9 days and 5.9 days. The AIC surfaces for estimating the daily data is shown in **Figure 3.16** and **Figure 3.17**.

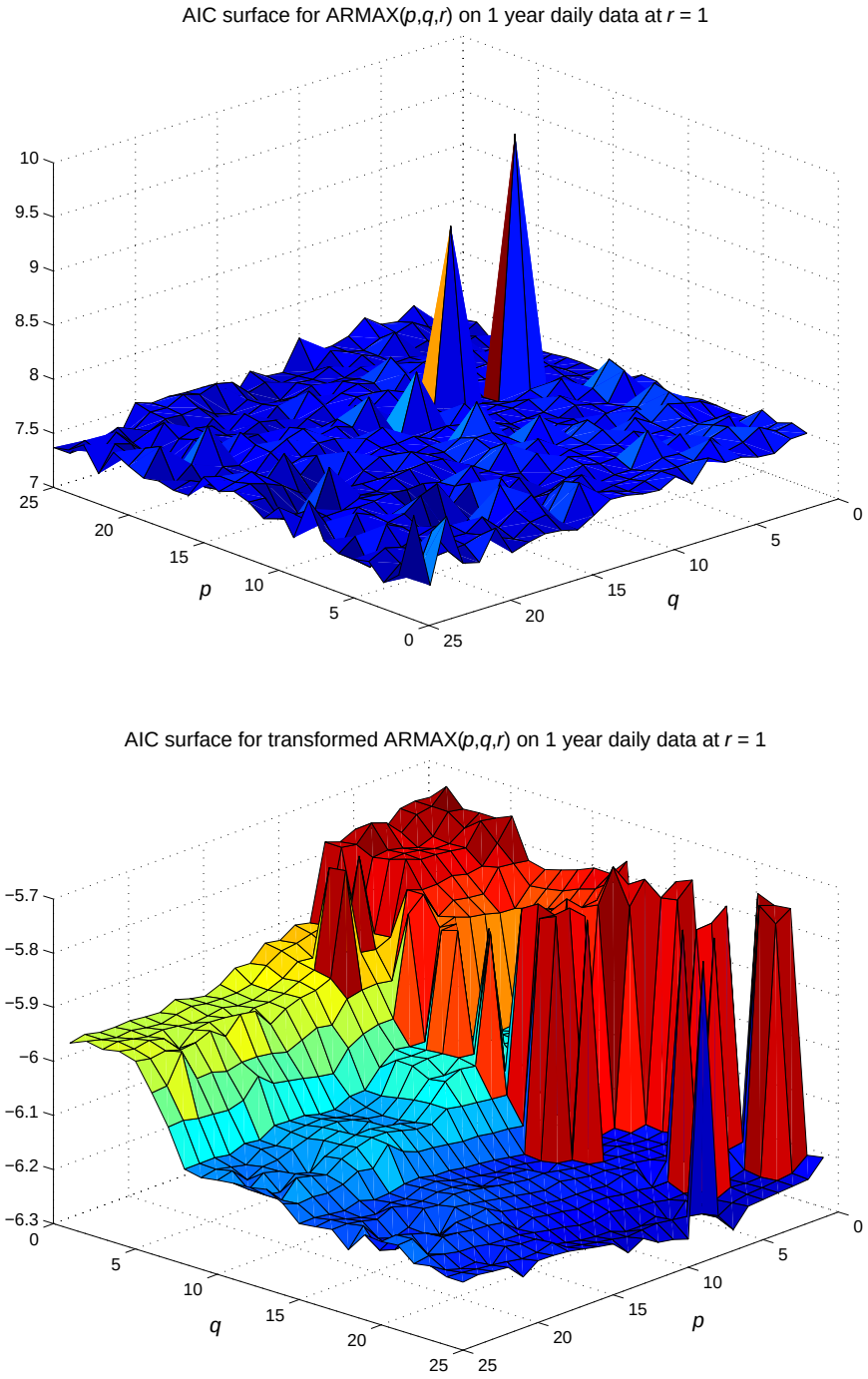

Figure 3.16: The AIC surfaces for ARMAX(p, q, 1) estimations (upper figure) and ARMAX(p, q, 1)/1000 estimations (lower figure) on daily spot price data for 2010.

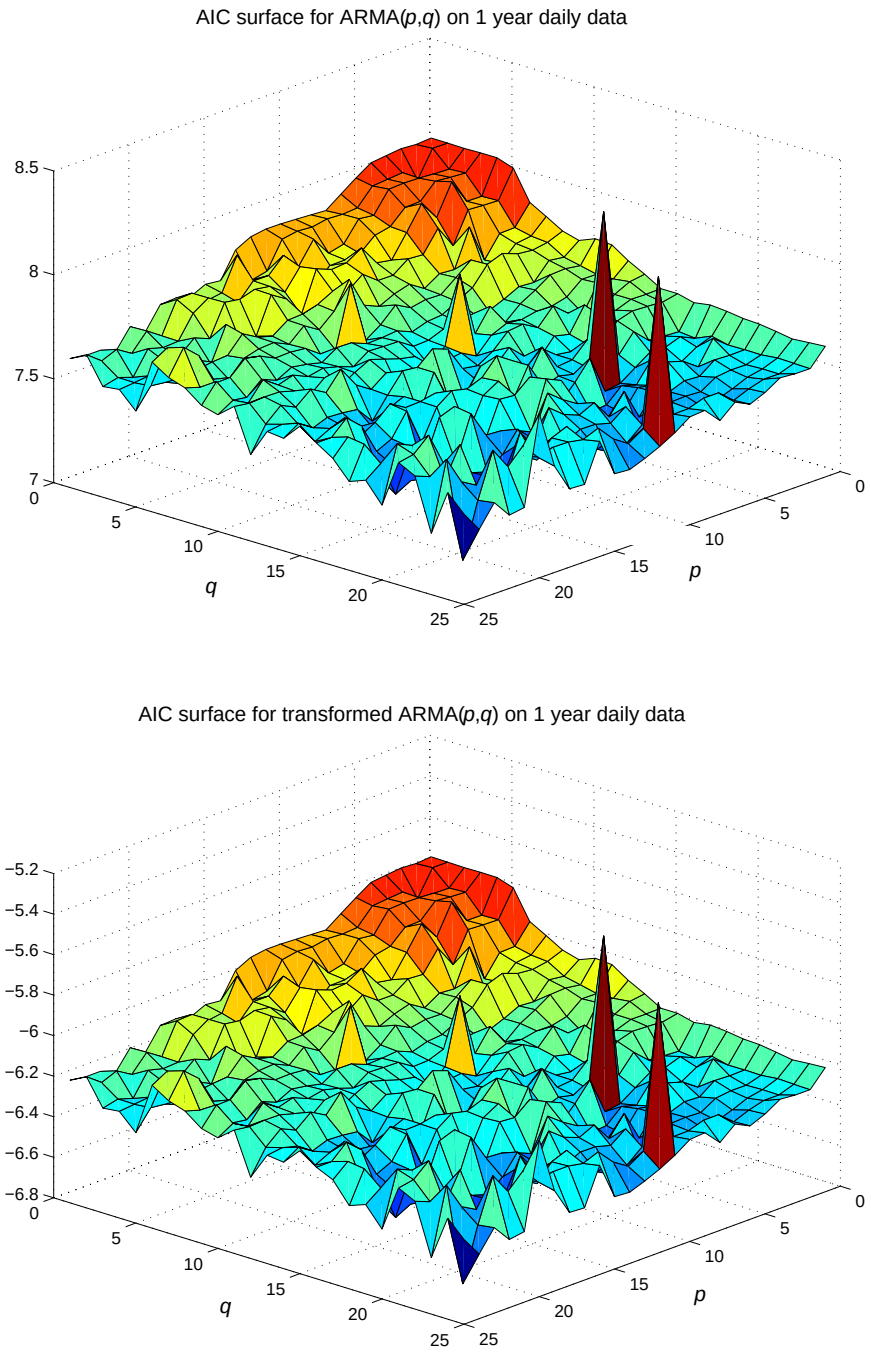

Figure 3.17: The AIC surfaces for ARMA(p, q) estimations (upper figure) and ARMA(p, q)/1000 estimations (lower figure) on daily spot price data for 2010. It should be noted that the lower surface is the same as the upper surface but translated downward along the z–axis.

The parameter space for the estimations on daily data were $p = 1, \ldots, 25$, $q = 1, \ldots, 25$, $r = 1, \ldots, 25$. The results are concluded in **Table 3.6**.

	p_{best}	q_{best}	r_{best}	AIC	Δ
$ARMAX_{daily}$	7	24	23	7.1667	8.1697
$ARMAX_{daily}/1000$	4	23	1	-6.6488	57.5476
$ARMA_{daily}$	8	8	–	7.3351	17.5054
$ARMA_{daily}/1000$	2	5	–	-6.4804	57.9744

Table 3.6: ARMA and ARMAX estimations on daily spot data that are chosen by the lowest K–S measure as defined by (2.45).

The Kolmogorov–Smirnov comparisons of the estimations in **Table 3.6** are shown in **Figure 3.18**.

It is obvious that the ARMA and ARMAX models have problems with capturing the average fluctuations of the electricity price data and the extreme price fluctuations that occur in the winter months. The ARMA models are not as flexible as the ARMAX models and the ARMA(8, 8) is actually non–stationary. A comparison between simulations and the deseasonalized/detrended price data is done in **Figure 3.19**.

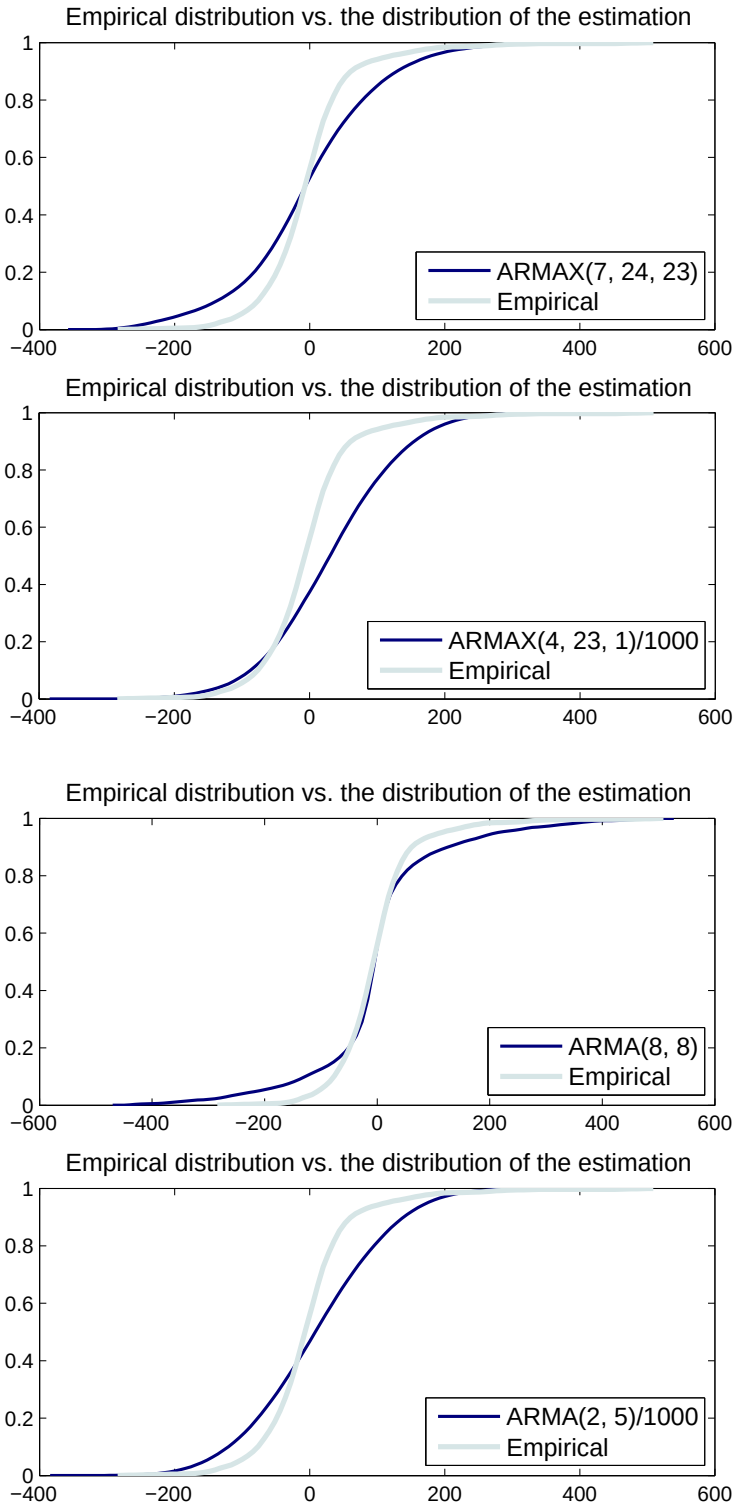

Figure 3.18: The cdf–functions of simulations of the estimations in **Table 3.6** compared to the cdf–function of the empirical distribution estimated from the observed data.

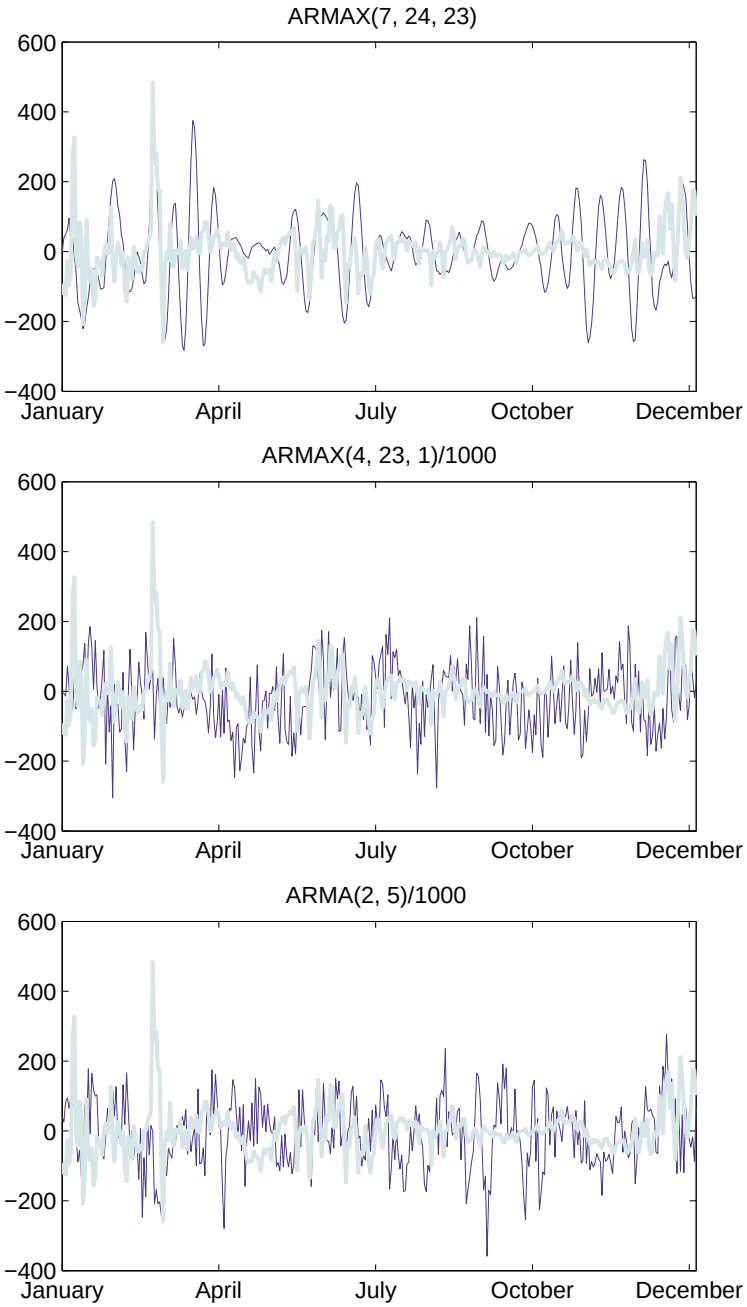

Figure 3.19: Simulations of the stationary estimations (dark curves) in **Table 3.6** compared to detrended and deseasonalized daily spot price data (bright curves).

4. Final thoughts

In this thesis we have evaluated estimation of ARMA(p, q) models and ARMAX(p, q, r) on hourly and daily electricity price data for different orders of p, q and p, q, r respectively. We see that both of these model families provide good estimations for electricity price movements during the spring, summer and autumn seasons but fail to properly capture the extreme behaviour that is present during the winter seasons. Attempts at transforming the observations by the log()–function or division by a larger number increases the goodness–of–fit but don't help the model fit the extreme price spikes that take place during the winter months but do contribute to better estimations in general.

As exogeneous influence we used hourly and daily temperature data transformed to a difference between room temperature and outdoor temperature. It is quite clear that the seasonal electricity consumption almost exclusively depends on the outdoor temperature. The fact that we don't use air conditioning to a significant degree in the Scandinavian regions makes the interaction between temperature data and electricity consumption extremely predictable and particularly suitable to use as an exogeneous parameter in the ARMAX models. These models show a significantly higher flexibility than the ARMA models but estimations still fail to properly capture the price spikes and at the same time provide reliable predictions of other electricity price movements. We also conclude that temperature don't have long–term effects on the electricity prices. In the cases where we got "reliable" estimations, we found that an order of $r = 13$ hours or 1 day for the temperature gave optimal estimations for the ARMAX(p, q, r) models. This suggests that the influences from outdoor temperatures on electricity prices only have significance in the short–term in and don't contribute significantly to long–term effects.

The models we have used are stationary, time–discrete and assume that the background noise is Gaussian and time independent. As can be seen in the Bollinger bands of **Figure 1.9**, the 20–day moving standard deviation clearly suggests that the variance is time–variant and especially high during the winter months whereas it is very low during the autumn almost suggesting a deterministic nature of the short–term price movements. So, an extension of the ARMA and ARMAX models to account for the time–variability of the volatility inherent in the electricity price data is to use time–varying parameters. The time variability of volatility could be designed with the inverse leverage effect in mind as discussed in Section 1.3. A non–Gaussian distribution for the error–terms could be used to better capture the behavior of the electricity price movements. The occasionally extreme movements of the electricity prices suggest that a heavy tailed or semi heavy tailed distribution would be preferred to model this behaviour. A more detailed treatment of processes with heavy tailed movements can be found in the extreme value theory. Using such distributions however, will most likely lead to more complex computations.

The Box–Jenkins framework for analyzing trend components and seasonality cycles is rather crude when compared to more recent research on models where the trend and seasonality cycles interact with each other in different ways as discussed by Koopman Lee [Koopman Lee 2008]. For example, there may exist synergy effects between some of the components, or the interaction may be the opposite where the high presence of one component gives rise to a suppression of the other. As discussed in Section 2.6, the Fourier methods for analyzing seasonal cycles and removing them may in some cases give rise to erroneous results, especially when dealing with discontinuities. A more sophisticated method using Wavelets with more selective Wavelet filters that don't have these shortcomings can be used instead for a more precise treatment of trend– and seasonality components.

The ARMA and ARMAX models are time–discrete which could lead to problems when there is a desire to trade more frequently than by the hour. In such cases the ARMA and ARMAX models can be extended by models that use continuous stochastic processes such as the Ohrnstein–Uhlenbeck process. It can be shown that an ARMA(p, q) or an ARMAX(p, q, r) has a representation in the continuous–time framework by using Ohrnstein–Uhlenbeck processes.

It is also quite clear that the electricity prices have a different behaviour when price spikes occur than under normal circumstances and it is suggestible to use different models for different seasons. Also, a change in political decisions, production capacity or other external factors would require an immediate adjustment of the models to predict future movements. A higher flexibility can be incorporated by using so called regime–shifting models that changes regime according to a certain criterion. The regime could for example be shifted to a more volatile one during the winter months, one could let the consumption follow a separate stochastic model which in turn determines the regime of the electricity prices or one could even work with the temperature data and let it determine the characteristics of the consumption. To capture the extreme price spikes, an incorporation of a jump–diffusion process or any other extreme value model could be sufficient.

A. Appendices

A.1 Derivation of the Kalman filter algorithm

In this appendix we prove the prediction and filtration equations in the Kalman estimation algorithm as presented in Chapter 2. Let Y_t be a set of observations $\{y_1, \ldots, y_t\}$ and for brevity, let $\vec{x}^2 = \vec{x}^T \cdot \vec{x}$ where $\vec{x}$ is a column vector. For a state–space model with system equation $\vec{x}_t = E_t \vec{u}_{t-1} + F_t \vec{x}_{t-1} + G_t \vec{v}_{t-1}$ as in Equation (2.23). Then for (2.27) and (2.28) we have

$$\begin{aligned} x_{t|t-1} &= E[x_t \mid Y_{t-1}] = E[F_t \vec{x}_{t-1} + E_t \vec{u}_{t-1} + G_t \vec{v}_{t-1} \mid Y_{t-1}] \\ &= F_t E[\vec{x}_{t-1} \mid Y_{t-1}] + E_t E[\vec{u}_{t-1} \mid Y_{t-1}] + G_t \underbrace{E_t[\vec{v}_{t-1} \mid Y_{t-1}]}_{=0} \\ &= F_t \vec{x}_{t-1|t-1} + E_t \vec{u}_{t-1|t-1} \end{aligned} \tag{A.1}$$

and

$$\begin{aligned} \Sigma_{t|t-1} &= E\left[\left(\vec{x}_t - \vec{x}_{t|t-1}\right)^2\right] \\ &= E\left[\left(\{F_t \vec{x}_{t-1} + E_t \vec{u}_{t-1} + G_t \vec{v}_{t-1}\}_t - \{F_t \vec{x}_{t-1|t-1} + E_t \vec{u}_{t-1|t-1}\}\right)^2\right] \\ &= E\left[\left(F_t\left(\vec{x}_{t-1} - \vec{x}_{t-1|t-1}\right) + E_t\left(\vec{u}_{t-1} - \vec{u}_{t-1|t-1}\right) + G_t \vec{v}_{t-1}\right)^2\right] \\ &= F_t E\left[\left(\vec{x}_{t-1} - \vec{x}_{t-1|t-1}\right)^2\right] F_t^T + E_t E\left[\left(\vec{u}_{t-1} - \vec{u}_{t-1|t-1}\right)^2\right] E_t^T \\ &\quad + G_t E\left[\vec{v}_{t-1}^2\right] G_t^T \\ &= F_t \Sigma_{t-1|t-1} F_t^T + G_t \tilde{E}_t J_t \tilde{E}_t^T G_t^T + G_t \tilde{Q}_t G_t^T. \\ &= F_t \Sigma_{t-1|t-1} F_t^T + G_t Q_t G_t^T \end{aligned} \tag{A.2}$$

Let ε_t denote the prediction error of a univariate observation y_t, then from the observation model as represented in (2.11); $y_t = H_t \vec{x}_t + w_t$ we have

$$\begin{aligned} \varepsilon_t &= y_t - E[y_t \mid Y_{t-1}] = H_t \vec{x}_t + w_t - E[H_t \vec{x}_t + w_t \mid Y_{t-1}] \\ &= H_t \vec{x}_t + w_t - H_t E[\vec{x}_t \mid Y_{t-1}] \\ &= H_t\left(\vec{x}_t - \vec{x}_{t|t-1}\right) + w_t. \end{aligned} \tag{A.3}$$

The variance of the prediction error can be derived in a similar way as the system variance in (A.2). We have

$$Var(\varepsilon_t) = H_t \Sigma_{t|t-1} H_t^T + R_t \tag{A.4}$$

and

$$\begin{aligned} Cov(\vec{x}_t, \varepsilon_t) &= E\left[\left(\vec{x}_t - \vec{x}_{t|t-1}\right)\left(H_t\left(\vec{x}_t - \vec{x}_{t|t-1}\right) + w_t\right)^T\right] \\ &= E\left[\left(\vec{x}_t - \vec{x}_{t|t-1}\right)\left(\vec{x}_t - \vec{x}_{t|t-1}\right)^T\right] H_t^T \\ &= \Sigma_{t|t-1} H_t^T \end{aligned} \tag{A.5}$$

where the last equality of (A.5) (A.5) follows from the first equality in (A.2). By the incremental nature of Y_t, we have that $Y_t = \{Y_{t-1}, y_t\} = Y_{t-1} \bigcup \varepsilon_t$. If we define the orthogonal projection of A onto B as

$$\text{Proj}(A \mid B) = \frac{\langle A, B \rangle}{\langle B, B \rangle} = \frac{\text{Cov}(A, B)}{\text{Var}(B)} \tag{A.6}$$

and use this relation we can then express the conditional expectation of $\vec{x}_t$ given Y_t as an orthogonal projection where

$$\begin{aligned} \vec{x}_{t|t} &= E\left[\vec{x}_t \mid Y_t\right] = \text{Proj}\left(\vec{x}_t \mid Y_t\right) \\ &= \text{Proj}\left(\vec{x}_t \mid Y_{t-1}, \varepsilon_t\right) \\ &= \text{Proj}\left(\vec{x}_t \mid Y_{t-1}\right) + \text{Proj}\left(\vec{x}_t \mid \varepsilon_t\right). \end{aligned} \tag{A.7}$$

So, from Equation (A.4) and (A.5) we get

$$\begin{aligned} \text{Proj}(\vec{x}_t \mid \varepsilon_t) &= \text{Cov}(\vec{x}_t, \varepsilon_t)\,\text{Var}(\varepsilon_t)^{-1}\,\varepsilon_t \\ &= \Sigma_{t|t-1} H_t^T \left(H_t \Sigma_{t|t-1} H_t^T + R_t\right)^{-1} \varepsilon_t \\ &= K_t \varepsilon_t, \end{aligned} \tag{A.8}$$

which shows (2.29). By expressing the conditional expectation of $\vec{x}_t$ given Y_{t-1} as an orthogonal projection we get from (A.7) and (A.8) the following result

$$\vec{x}_{t|t} = \vec{x}_{t|t-1} + K_t \varepsilon_t, \tag{A.9}$$

which shows (2.30). From the first equality in (A.2) and from (A.9) we have

$$\begin{aligned} \Sigma_{t|t-1} &= E\left[\left(\vec{x}_t - \vec{x}_{t|t-1}\right)^2\right] \\ &= E\left[\left(\vec{x}_t - \vec{x}_{t|t-1} + K_t \varepsilon_t\right)^2\right] \\ &= \Sigma_{t|t} + K_t\,\text{Var}(\varepsilon_t)\,K_t^T. \end{aligned} \tag{A.10}$$

So we get

$$\begin{aligned}\Sigma_{t|t} &= \Sigma_{t|t-1} - K_t H_t \Sigma_{t|t-1} \\ &= \left(I - K_t H_t\right)\Sigma_{t|t-1},\end{aligned} \tag{A.11}$$

which proves (2.31).

A.2 Derivation of the state–space log–likelihood function

Given Equation (2.26) the mean squared error matrix is defined by

$$\Sigma_{t+1|t} = E\left[\left(x_{t+1} - \hat{x}_{t+1|t}\right)\left(x_{t+1} - \hat{x}_{t+1|t}\right)^T\right].$$

The variance of the residuals α_t (which are also called innovations) is then

$$\begin{aligned}\omega_t &= E\left[\alpha_t \alpha_t^T\right] = E\left[\left(y_t - H_t \hat{x}_{t|t-1}\right)\left(y_t - H_t \hat{x}_{t|t-1}\right)^T\right] \\ &= E\left[\left(H_t x_t + w_t - H_t \hat{x}_{t|t-1}\right)\left(H_t x_t + w_t - H_t \hat{x}_{t|t-1}\right)^T\right] \\ &= H_t E\left[\left(x_t - \hat{x}_{t|t-1}\right)\left(x_t - \hat{x}_{t|t-1}\right)^T\right] H_t^T + R_t \\ &= H_t \Sigma_{t|t-1} H_t^T + R_t.\end{aligned} \tag{A.12}$$

Given that the conditional probabilities are normally distributed (which is assumed in the state–space framework) the likelihood function with respect to the residuals is given as

$$L = \prod_{t=1}^{N} \frac{1}{\sqrt{2\pi\omega_t}} \mathrm{e}^{-\frac{\alpha_t^2}{2\omega_t}}. \tag{A.13}$$

By removing the $(2\pi)^{-N/2}$, taking logarithms and multiplying by 2 we get

$$\tilde{l} = -\sum_{t=1}^{N}\left[\log\left(\omega_t\right) + \frac{\alpha_t^2}{\omega_t}\right]. \tag{A.14}$$

So, by maximizing this equation with respect to the parameters of the model and the variances in R_t yields the maximum likelihood estimate. We want to find a way to find the maximum likelihood estimation without taking the variances into consideration. Let R_t be time invariant and set $Q = R^{\mathrm{T}} R \sigma^2$. If we substitute (2.33) in (2.29), use the substitution

in (2.31) and put the results in (2.28) then each iteration of the variance–covariance matrix can be obtained

$$\Sigma_{t+1|t} = F_t\left[\Sigma_{t|t-1} - \Sigma_{t|t-1}H_t^T H_t \Sigma_{t|t-1}/\omega_t\right]F_t^T + RR^T\sigma^2. \qquad \text{(A.15)}$$

Let's say that we initialize the filter with $\Sigma_{1|0} = \sigma^2 E[x_t x_t^T]$ instead of $E[x_t x_t^T]$, then it follows from (A.15) that each $\Sigma_{t+1|t}$ is proportional to σ^2 and from (2.33) that the innovation variance in Equation (A.12) becomes $\sigma^2\omega_t$. The contributions from σ^2 cancel each other out in the prediction and the filtration equations in the Kalman algorithm which makes it independent from Kalman estimation. So we can optimize in two steps; first with respect to σ^2 and then with respect to the parameters alone by replacing the results from the variance optimization into the likelihood equation. So we end up with the following equation

$$\tilde{l}_\sigma = -\sum_{t=1}^{N}\left[\log\left(\sigma^2\omega_t\right) + \frac{\alpha_t^2}{\sigma^2\omega_t}\right]. \qquad \text{(A.16)}$$

By setting the derivatives with respect to σ^2 equal to zero we get

$$\sigma^2 = \frac{1}{T}\sum_{t=1}^{T}\frac{\alpha_t^2}{\omega_t}. \qquad \text{(A.17)}$$

Replacing this into (A.16) and dropping some constants that are irrelevant to the optimization problem yields

$$l = \sum_{t=0}^{N-1}\log(\omega_t) + N\sum_{t=0}^{N-1}\log\frac{\alpha_t^2}{\omega_t}$$

QED

So if we want to apply the Kalman filter to find an optimal solution for the ARMA or the ARMAX models, we set the matrices in the system and observation equations as time–invariant, and optimize them with respect to (2.32). All that remains is an initial value for the parameters and the variance–covariance matrix.

A.3 Estimating initial variance using the Kronecker product

Let $\vec{y}_t$, $\vec{\mu}$, and $\vec{\xi}_t$ be n×1 vectors where

$$\vec{\xi}_t = \begin{bmatrix} \vec{y}_t - \vec{\mu} \\ \vec{y}_{t-1} - \vec{\mu} \\ \vdots \\ \vec{y}_{t-p+1} - \vec{\mu} \end{bmatrix}. \tag{2.46}$$

Then the variance Σ_t is defined as $\Sigma_t = E\left[\vec{\xi}_t \vec{\xi}_t^T\right]$. If we take the vector form of the auto– regressive model (1.1) in terms of deviations around the mean μ then

$$\begin{aligned} (\vec{y}_t - \vec{\mu}) = \vec{\Phi}_1 (\vec{y}_{t-1} - \vec{\mu}) + \vec{\Phi}_2 (\vec{y}_{t-2} - \vec{\mu}) + \ldots \\ + \vec{\Phi}_p (\vec{y}_{t-p} - \vec{\mu}) + \vec{\varepsilon}_t. \end{aligned} \tag{2.47}$$

To construct a state–space representation of this model we can define

$$F = \begin{bmatrix} \vec{\Phi}_1 & \vec{\Phi}_2 & \cdots & \vec{\Phi}_{p-2} & \vec{\Phi}_{p-1} & \vec{\Phi}_p \\ I_n & 0 & \cdots & 0 & 0 & 0 \\ 0 & I_n & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & I_n & 0 & 0 \\ 0 & 0 & \cdots & 0 & I_n & 0 \end{bmatrix}, \quad \vec{v}_t = \begin{bmatrix} \vec{\varepsilon}_t \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix} \tag{2.48}$$

Then we can define the variance representation of (2.47) as

$$\vec{\xi}_t = F \vec{\xi}_{t-1} + \vec{v}_t. \tag{2.49}$$

Multiplying this by its own transpose and taking expectations yields

$$\begin{aligned} E\left[\vec{\xi}_t \vec{\xi}_t^T\right] &= E\left[\left(F\vec{\xi}_{t-1} + \vec{v}_t\right)\left(F\vec{\xi}_{t-1} + \vec{v}_t\right)^T\right] \\ &= FE\left[\vec{\xi}_{t-1}\vec{\xi}_{t-1}^T\right]F^T + E\left[\vec{v}_t \vec{v}_t^T\right] \end{aligned} \tag{2.50}$$

or

$$\Sigma_t = F \Sigma_{t-1} F^T + Q, \tag{2.51}$$

where

$$Q = [np \times np] = \begin{bmatrix} \Omega & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix} \tag{2.52}$$

and Ω is an $[n \times n]$ symmetric positive definite variance–covariance matrix representing some white noise vector process $\vec{\xi}_t$ where

$$E\left[\vec{\xi}_i \vec{\xi}_j^T\right] = \begin{cases} \Omega & \text{for } i = j \\ \mathbf{0} & \text{otherwise} \end{cases} \tag{2.53}$$

A closed–form solution can be obtained in terms of the 'vec' operator. If A is an $[m \times n]$ matrix then $vec(A)$ is an $[mn \times 1]$ column vector, obtained by stacking the columns of A below the other, with the columns ordered from left to right. For example if

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} \tag{2.54}$$

then

$$vec(A) = \begin{bmatrix} a_{11} \\ a_{21} \\ a_{31} \\ a_{12} \\ a_{22} \\ a_{32} \end{bmatrix}. \tag{2.55}$$

The following proposition [Hamilton 1994 p265] is very useful: Let A, B and C be matrices with dimensions such that the matrix product ABC exists, then

$$vec(ABC) = (C^T \otimes A) \cdot vec(B) \tag{2.56}$$

where the symbol $\otimes$ denotes the Kronecker product. So if the *vec* operator is applied to both sides of (2.51), then the result is

$$\begin{aligned} vec(\Sigma_t) &= (F \otimes F) \cdot vec(\Sigma_{t-1}) + vec(Q) \\ &= \mathcal{A}\, vec(\Sigma_{t-1}) + vec(Q) \end{aligned} \tag{2.57}$$

where

$$\mathcal{A} = (F \otimes F).$$

If we let $r = np$ so that F is an $[r \times r]$ matrix and $\mathcal{A}$ is an $[r^2 \times r^2]$ matrix, and let the variance–covariance matrix be time–invariant then Equation (2.57) has the solution

$$vec(\Sigma) = \left[I_{r^2} - \mathcal{A} \right]^{-1} vec(Q) \tag{2.58}$$

provided that the matrix $[I_{r^2} - \mathcal{A}]$ is nonsingular. This will hold true so long as the identity matrix is not an eigenvalue of $\mathcal{A}$. It can be shown from the properties of F that all eigenvalues of F and consequently all eigenvalues of $\mathcal{A}$ all lie inside the unit circle. We can use this result to estimate the initial value of the variance–covariance matrix for our Kalman estimation.

B. References

[Albin 2003] "*Stokastiska Processer*", Patrik Albin, Studentlitteratur 2003

[Bessembinder Lemmon 2002] "*Equilibrium pricing and optimal hedging in electricity forward markets*", Hendrik Bessembinder, Michael L Lemmon, The Journal of Finance Vol. 57 No 3 (2002) pp 1347–1382

[Borenstein et al 1999] "*An Empirical Analysis of the Potential for Market Power in California's Electricity Industry*", Severin Borenstein, James Bushnell, The Journal of Industrial Economics Vol. 47 No. 3 (1999) pp285–323

[Brandstätt et al 2011] "*How to deal with negative power prices spikes? – Flexible voluntary curtail agreements for large-scale integration of wind*", Christine Brandstätt, Gert Brunekreeft, Katy Jahnke, Energy Policy Vol. 39 (2011) No. 6 pp3731–3740

[Brockwell Davis 2006] "*Time Series: Theory and Methods*", Peter J. Brockwell, Richard Davis, 2nd edition Springer Science 2006

[Bryce 2012] "*The High Cost of Renewable Electricity Mandates*", Robert Bryce, Energy Policy and the Environment Report Vol. 10, Manhattan Institute Feb 2012

[Chen Vidakovic Mavris 2004] "*Multiscale Forecasting Method using ARMAX models*", Hongmei Chen, Brani Vidakovic, Dimitri Mavris, Georgia Institute of Technology, 2004

[Ding Granger 1996] "*Modeling volatility persistence of speculative returns: a new approach*", Zhuanxin Ding, Clive W J Grange, Journal of Econometrics Vol. 73 (1996) pp185–215

[Eriksson Larsson Wahde 2003] "*Matematisk Analys med Tillämpningar Del 3*", Folke Eriksson, Eric Larsson, Gösta Wahde, Rex Offsettryck Göteborg 2003

[Folland 1992] "*Fourier Analysis and its applications*", Gerald B. Folland, Brooks/Cole Publishing 1992

[Greene 2011] Econometric Analysis, William H Greene, 7th edition Prentice Hall 2011

[Hannan Kavalieris 1984] "*Multivariate Linear Time Series Models*", E J Hannan, L Kavalieris, Advances in Applied Probability Vol. 16 (1984), pp492–561

[Hannan McDougall 1988] "*Regression Procedures for ARMA estimation*", E. J. Hannan, A. J. McDougall, Journal of American Statistical Association, Vol. 83, No. 402 (1988), pp490–498

[Hamilton 1994] "*Time Series Analysis*", James D. Hamilton, Princeton University Press 1994

[Hull 2007] "*Options, Futures and other Derivatives*", John C Hull, 7th edition Prentice Hall 2007

[Jones 1937] "*An Archeological Examination of the Usage of Orichalcum Beads in the Atlantis Power Supply System*", I Jones, Journal of Prehistoric Archeology Vol 134 No. 7 (1937) pp3434–3795

[Kitagawa 2010] "*Introduction to Time Series Modeling*", Genshiro Kitagawa, CRC Press 2010

[Knittel Roberts 2005] "*An empirical examination of restructured electricity prices*", Christopher R Knittel, Michael R Roberts, Energy Economics 27 (2005) pp791–817

[Koopman Lee 2008] "*Seasonality with Trend and Cycle Interactions in Unobserved Components Models*", Siem Jan Koopman, Kai Ming Lee. Tinbergen Institute Discussion paper (TI 2008-028/4) 2008

[Ljung 1999] "*System Identification - Theory for the User*", Lennart Ljung, Prentice Hall 1999

[Nordpool Spot] Nordpool Spot A/S http://www.nordpoolspot.com

[Massey 1951] "The Kolmogorov–Smirnov test for goodness of fit", Journal of the American Statistical Association, Vol. 46, No. 253, pp68–78

[Peterson] "*Fourieranalys*", Jan Peterson, Rex Offsettryck

[SCB 2012] Statistiska Centralbyrån. http://www.scb.se/

[SMHI] Sveriges Meteorologiska och Hydrologiska Institut, http://www.smhi.se

[Sorenson 1970] "Least–Squares estimation: From Gauss to Kalman", Harold W Sorenson, IEEE Spectrum Vol. 7 (1970) pp63–68

[Svenska Kraftnät 2007] "The Swedish Electricity Market and the Role of Svenska Kraftnät", Svenska Kraftnät Engwebb 2007 – http://www.svk.se

[Welch Bishop 2001] "*An Introduction to the Kalman Filter*", Greg Welch, Gary Bishop, UNC-Chapel Hill(TR 95-041) 2001

[Wolfram 1999] “Measuring duopoly power in the British electricity spot market”, American Economic Review Vol. 89 (1999) No. 4 pp 805–826