Ektor Sotiropoulos

Modeling of German Electricity Load for Pricing of Forward Contracts

Master Thesis
PSL-1122
Preface

Deciding to carry out my master thesis at the Power Systems Lab at ETH Zurich and in the direction of power markets is a decision I never regretted. Firstly, because the sector that interests me most is power systems, but also because I was introduced to a fascinating topic for which I had little experience in the past, the power markets. During those last 6 months, my every-day life was in and around this lab. I would like to thank all people in the lab for their help and more specifically Prof. Andersson and my supervisors Yang He and Marcus Hildmann for giving me the opportunity to make this project real and for guiding me throughout its whole course. I could never neglect to mention my beloved flat-mates, that lived together with me the course of this project and helped me whenever needed, Marios and Evdokia, my family, parents, brother and grandmother, Spyros, Sophia, Fivos and Maria for supporting me throughout my entire student career, my friends in Switzerland and in Greece for being the salt and pepper in my life and finally Mary who was always on my side during this trip. Thank you all.
Abstract

Nowadays, a large part of the traded volume in power markets represents energy to be consumed or produced in the future as a forward or future product. Hence, the need for precise forecasting of load demand and prices arises for all market players. This project proposes an hourly load model for Germany with focus on long term forecast using tools provided from the econometrics theory, such as regression analysis, and time series analysis with stochastic calculus. The main purpose of this model is for applications in forward markets, such as fundamental modeling of spot prices, pricing of forward contracts and the formulation of the hourly price forward curve (HPFC).

The proposed model captures the deterministic component of the load, as well as its stochastic variations. Decomposition of daily load level and hourly load was made in order to achieve more accurate results. Then, each part was modeled separately and aggregated in the end. A linear regression approach was used for capturing the deterministic patterns of the load such as yearly, weekly seasonality and intraday patterns, taking into account holidays, weather and economic trends. An autoregressive model (AR) was used for the modeling of the stochastic variations.

Parameter calibration was made once in a 5 year period (2006-2010) and once in a 4 year period (2006-2009). Then, the model was evaluated and validated out of sample for 2011 and 2010-2011 respectively. The results presented in Chapter 5 were promising and led us to produce a future forecast for the period 2012-2014.
Contents

List of Acronyms viii
List of Symbols x

1 Introduction 1
  1.1 Literature Review 2
  1.2 Motivation 3

2 Methodology 5
  2.1 Econometrics 5
  2.2 Time series analysis 8

3 Data and Analysis 13
  3.1 Data 13
  3.2 Analysis 15

4 Modeling 25
  4.1 Daily load level modeling 26
    4.1.1 Deterministic part 26
    4.1.2 Stochastic part 28
    4.1.3 Overall daily load level model 30
  4.2 Hourly load modeling 33
    4.2.1 Deterministic part 33
    4.2.2 Stochastic part 35
    4.2.3 Overall hourly load model 38
  4.3 Overall load model 40
    4.3.1 Model parameters 41

5 Results 45
  5.1 In sample results 46
  5.2 Out of sample results 46
    5.2.1 Out of sample 2011 47
    5.2.2 Out of sample 2010-2011 50
  5.3 Future forecast 54
<table>
<thead>
<tr>
<th>CONTENTS</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>6 Conclusions</td>
<td>57</td>
</tr>
<tr>
<td>Bibliography</td>
<td>63</td>
</tr>
</tbody>
</table>
### List of Acronyms

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACF</td>
<td>Autocorrelation Function</td>
</tr>
<tr>
<td>ANN</td>
<td>Artificial Neural Networks</td>
</tr>
<tr>
<td>AR</td>
<td>Autoregression</td>
</tr>
<tr>
<td>ARIMA</td>
<td>Autoregressive Integrated Moving Average</td>
</tr>
<tr>
<td>ARMA</td>
<td>Autoregressive Moving Average</td>
</tr>
<tr>
<td>Cov</td>
<td>Covariance</td>
</tr>
<tr>
<td>ECDF</td>
<td>Empirical Cumulative Density Function</td>
</tr>
<tr>
<td>EEX</td>
<td>European Energy Exchange</td>
</tr>
<tr>
<td>ENTSO-E</td>
<td>European Network of Transmission Operators for Electricity</td>
</tr>
<tr>
<td>EPDF</td>
<td>Empirical Probability Density Function</td>
</tr>
<tr>
<td>FFT</td>
<td>Fast Fourier Transform</td>
</tr>
<tr>
<td>FL</td>
<td>Fuzzy Logic</td>
</tr>
<tr>
<td>GDP</td>
<td>Gross Domestic Product</td>
</tr>
<tr>
<td>HPFC</td>
<td>Hourly Price Forward Curve</td>
</tr>
<tr>
<td>MAE</td>
<td>Mean Absolute Error</td>
</tr>
<tr>
<td>MAPE</td>
<td>Mean Absolute Percentage Error</td>
</tr>
<tr>
<td>OECD</td>
<td>Organization for Economic Co-operation and Development</td>
</tr>
<tr>
<td>OLS</td>
<td>Ordinary Least Squares</td>
</tr>
<tr>
<td>PACF</td>
<td>Partial Autocorrelation Function</td>
</tr>
<tr>
<td>RMSE</td>
<td>Root Mean Square Error</td>
</tr>
<tr>
<td>TSO</td>
<td>Transmission System Operator</td>
</tr>
</tbody>
</table>
# List of Symbols

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha_{d,i}$</td>
<td>Daily load model deterministic model parameters</td>
</tr>
<tr>
<td>$\beta_{d,i}$</td>
<td>Autocovariance parameters</td>
</tr>
<tr>
<td>$\gamma_{d,i}$</td>
<td>Daily model autoregressive parameters</td>
</tr>
<tr>
<td>$\delta_{d,i}$</td>
<td>Hourly model autoregressive parameters</td>
</tr>
<tr>
<td>$d_{i,j}$</td>
<td>Hourly model deterministic part parameters</td>
</tr>
<tr>
<td>$\zeta_{i,j,k}$</td>
<td>Error terms</td>
</tr>
<tr>
<td>$\phi_i$</td>
<td>Autoregression parameters</td>
</tr>
<tr>
<td>$f$</td>
<td>Deterministic part</td>
</tr>
<tr>
<td>$L$</td>
<td>Total load demand</td>
</tr>
<tr>
<td>$L_d$</td>
<td>Average daily load demand</td>
</tr>
<tr>
<td>$L_h$</td>
<td>Hourly load demand</td>
</tr>
<tr>
<td>$\mu$</td>
<td>Mean value</td>
</tr>
<tr>
<td>$P_k$</td>
<td>Partial autocorrelation parameters</td>
</tr>
<tr>
<td>$p$</td>
<td>Partial autocorrelation function order</td>
</tr>
<tr>
<td>$R$</td>
<td>Coefficient of determination</td>
</tr>
<tr>
<td>$\rho_k$</td>
<td>Autocorrelation parameters</td>
</tr>
<tr>
<td>$\sigma$</td>
<td>Standard deviation</td>
</tr>
<tr>
<td>$\text{Var}$</td>
<td>Variance</td>
</tr>
<tr>
<td>$X_i$</td>
<td>Independent model variable - regressor</td>
</tr>
<tr>
<td>$\tilde{x}$</td>
<td>Stochastic model part</td>
</tr>
<tr>
<td>$Y$</td>
<td>Dependent model variable</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

The liberalization of power markets that started spreading widely during the last decade of the 20th century has significantly changed the energy transaction landscape. Nowadays, electricity is traded in the electricity exchange and in over-the-counter markets. A large part of the traded volume represents energy to be consumed or produced in the future as a forward or future product. Hence, the need for precise forecasting of load demand and prices arises for all market players.

An important attribute of electricity, which differentiates it from other commodities, is that demand and supply have to be constantly in balance at all moments [1]. Therefore, from the power system point of view it has always been necessary to predict future load demand in order to match it as much as possible with the supply. Load forecasting is a crucial task for the operation of power systems. It is a necessary tool from the viewpoint of the utilities, the system operators, the government and the market players. Load forecasting, though, can be distinguished by the different time-scales and the different application purposes [2], [3]. Long-term forecasts of the overall load, of the peak load or for the load demand trend are necessary for reliability issues and planning of the power system, such as construction and decommission of plants and maintenance scheduling. Week and day-ahead forecasts are used in combination with the merit order curve for unit commitment, hour-ahead forecasts for the economic dispatch of the system and 10-minutes ahead for the correct load frequency control scheduling. Therefore, load forecasting can be very useful in the operation of a power system.

From the market point of view, and in our particular case the European power markets, the load forecast plays an important role in the definition of electricity prices. Short-term load forecasts, such as day ahead forecasts, are frequently used along with the merit order curve for the forecasts of spot market prices. However, mid- to long-term forecasts, with a time horizon of few weeks to several years, are necessary for the pricing of forward contracts.
and futures prices.

Nowadays, with the evolution of the liberalized power markets, most of the transactions are made in the forward market rather than the spot market \([4]\). Firstly, the transaction costs are generally higher in the spot market. Furthermore, the risk involved for market players in the forward markets is less because they have the possibility to foresee a potential risk and hedge against it. On the contrary, spot market players are constrained to the prices imposed by the market procedures and, therefore, their risk is higher \([4]\).

\subsection*{1.1 Literature Review}

The topic of load forecasting has gained importance in recent years because of its increasing weight in the reformed power markets. After carefully examining literature related to the topic of load forecasting, we came across many detailed publications, using various methodologies, concerning short- and long-term forecasts. However, we came across a limited number of references in detailed hourly long-term load models.

The objective in most of the publications was to model the short-term effects on the spot prices, mostly day up to a week ahead \([1, 5, 6, 7, 8, 9]\). Publications with focus on medium- and long-term modeling usually aimed in capturing the overall load trend through the years, including monthly or daily average and peak load \([3, 10, 11, 12, 13, 14]\). Only a few contained long-term hourly load models but even those had the weekends and holidays removed for simplicity reasons \([15, 16]\).

According to the purpose of each of the studied works, various forecasting methods were used. One of the main methodologies we came across was the linear regression which is the most popular and most simple in terms of computational effort \([13, 14, 15]\). There was also a very common use of autoregressive (AR), autoregressive moving average (ARMA) and autoregressive integrated moving average (ARIMA) models \([6, 7, 8, 9, 14]\). \([13]\) proposes the combination of fuzzy logic (FL) and artificial neural networks (ANN) for the long term forecast of monthly peak load and compares it to a multiple regression model and a simple ANN model, while \([5]\) evaluates these methods independently and compares them with an AR model. Parametric and non-parametric regression is another approach we came across in literature, usually followed by a stochastic model such as an AR \([3, 9, 11, 12, 16]\). Other mentioned approaches include the use of a Kalman filter, transfer function models, curve fitting and principal component analysis.

It is important to note that only a number of these models included exogenous variables such as weather data for the model formulation \([3, 7, 9, 11, 12, 13]\). All the other publications studied, included processing of the historic load data alone. It is straightforward to say that weather factors
and weather seasonality is already embedded in historic load data [10]. Finally, no model of those examined made use of economic data as exogenous variable of load demand mainly because the time horizon of their forecasts and calibrating samples was not very broad to incorporate such long term effects. However, taking into account the economic conjuncture of our time, it is impossible to exclude such a factor.

1.2 Motivation

The highest motivation for the realization of this master thesis was this insufficiency in the literature with clear focus on long-term hourly load modeling. In particular, objective of this project was to construct an hourly load prediction of Germany with a horizon of several years ahead. However, it is important to note that the focus of this project is on the total load demand of Germany regardless of congestion issues, power flows and other limitations of the power system. We assumed that this demand is always covered by the TSOs.

Purpose of this load model is to be used in applications of fundamental modeling of power markets, and in particular for pricing of forward contracts, modeling of spot prices and the construction of the hourly price forward curve (HPFC), with the use of a supply model, as seen in Figure (1.1). Therefore, we selected the German EEX market because it is one of the most important and well-established power markets in Europe, besides Nordic pool and the French market. In addition, the availability of data for the German case was also an important factor for its selection.

![Figure 1.1: Simple example of supply and demand curves of the power market](image)

The main goal of our approach is to define both the deterministic com-
ponent and the stochastic variations of the load. Historic data show that the load follows a similar pattern in yearly scale, but also in the weekly and daily scale. Therefore our model needs to contain this deterministic patterns of the load in various time horizons. Furthermore, we wanted to test the significance of weather factors in load modeling, since part of it is already embedded in the load data. In addition, the deterministic part had to contain the effect of economic factors such as the Gross Domestic Product (GDP) and the industrial production indicator. Finally, the exact hourly load demand follows a random nature. Therefore, an accurate model should also include the stochastic variations of the load and incorporate this uncertainty. This nature of load demand can be explained with the use time series analysis and stochastic calculus.
Chapter 2

Methodology

Main goal of this project is to create a load model in order to be able to perform long-term forecasts. This model will then be used for the pricing of forward and futures contracts. Electricity prices vary in an hourly basis and therefore we are interested in hourly load demand. This data, as we will see later in this chapter, are time series and should be analyzed as such.

In this chapter, the main concepts of the forecast methodologies that were used in this master thesis will be presented. Firstly, the reader will be introduced to the econometric theory. Then, some basic aspects of time series analysis will introduce some forecasting tools, which, with the addition of the use of stochastic calculus will form the basis of our model.

2.1 Econometrics

Econometrics can be defined as the science of testing economic theories. It provides tools for the formulation of mathematical models that describe economic phenomena and forecast future values of economic variables. All in all, econometrics use economic theory and statistical techniques to analyze economic data [17].

The most common question that the econometrics try to answer is: ‘What is the effect of variable $a$ on variable $b$’. One of the most simple and frequent methods used to provide an answer to that question is linear regression.

Multiple linear regression

In this project, we used linear regression in order to describe the deterministic model for the load demand for Germany. However, since the load demand depends on several factors, the use of multiple linear regression, which estimates the effect of various variables on the dependent variable, was compelling.
The goal of the multiple linear regression model is to estimate the effect on the dependent variable $Y$ of each of the independent variables $X_1, ..., X_n$, or regressors, while holding the others constant. Equations (2.1) and (2.2) show the equation representing the multiple linear regression model in the analytic and matrix notations [17].

$$ Y_t = \beta_0 + \beta_1 X_{1,t} + \beta_2 X_{2,t} + \ldots + \beta_n X_{n,t} + u_t \quad (2.1) $$

for all $t$, where $X_1, ..., X_n$ are the regressors and $u$ the error terms. The first part of Equation (2.1) without the error terms, represent the estimation, $\hat{Y}_t$, that this model produces for the actual value of $Y_t$. In matrix notation, multiple linear regression can be written as follows:

$$ Y = X\beta + u \quad (2.2) $$

Ideally we want our model to be exactly equal to the actual measured time series. However, this is impossible to achieve and we need to minimize the difference between the measured values and the modeled ones according to some norm. In Equations (2.3) we can see the general equation for the minimization we want to achieve and the special case of the ordinary least squares (OLS) for $p = 2$ which is the one used widely in literature and in the case of this thesis.

$$ \min \sum_t \|Y_t - \hat{Y}_t\|_p \quad (2.3a) $$

$$ \min \sum_t \|Y_t - \hat{Y}_t\|_2 = \min \sum_t (Y_t - \hat{Y}_t)^2 \quad (2.3b) $$

The coefficients of the multiple regression model will be estimated using the ordinary least squares (OLS) method. The objective of this method is to estimate the coefficients of the regressors, $\beta$, by minimizing the sum of squared prediction errors as in Equation (2.4).

$$ \min_{\beta} \sum_t (Y_t - \beta_0 - \beta_1 X_{1,t} - \ldots - \beta_n X_{n,t})^2 = \sum_t (Y_t - \hat{Y}_t)^2 \quad (2.4) $$

OLS is the dominant method used in practice for estimating the parameters of linear regression. Some of the theoretical properties of the OLS estimators are highly desirable for the statistical analysis. The most important of which is the fact that they are consistent, efficient and unbiased estimators [17]. In addition, the OLS is successful in estimating the model parameters if the mean of the residuals, $E(u)$, is equal to zero. Furthermore, the variables $X_1, ..., X_n, Y$ need to be independently and identically distributed random variables. Finally, large outliers should be unlikely, because the coefficients might be sensitive to such disturbances, and there
CHAPTER 2. METHODOLOGY

should not be perfect multicollinearity between the regressors since it would lead to a division by zero in the parameter estimation process [17].

A similar approach can be used for non-linear regression with the use of a polynomial, logarithmic, a variable interaction or a sinusoidal [17]. As it will be shown later on, in this project we examined the efficiency of load models that included sinusoidal regressions of several frequencies and their harmonics. However, the use of factor models, that included the influence of specific regressors such as temperature, with the combination of sample splitting and decomposition produced significantly better results.

Measure of model fit

There are several quantities that are used in the econometrics approach to statistical analysis that can measure the success of a method into fitting a model. One of the most important ones it the $R^2$, or coefficient of determination.

The coefficient of determination represents the fraction of the sample variance of $Y$ that is explained by the regressors $X$. It is equivalent to one minus the fraction of the variance in $Y$ not explained by the regressors [17]. The $R^2$ ranges between 0 and 1. A low value of the coefficient of determination may mean that there are still some regressors that were not used in the modeling process and are needed in order to better explain the dependent variable $Y$. Usually, the adjusted $R^2$ is used for linear regression with multiple regressors. The expression is given in Equation (2.5). For the non-adjusted $R^2$, the reader can recur to the econometrics bibliography [17].

$$R^2 = 1 - \frac{Var(u)}{Var(Y)} = 1 - \frac{\sigma_u^2}{\sigma_Y^2}$$

(2.5)

Another measure that evaluates the estimated coefficients, $\hat{\beta}$, given a certain hypothesis or assuming a null hypothesis, is the t-statistic. This is defined as the ratio of the difference between the estimated parameter $\hat{\beta}_i$ and a hypothetical value of that parameter over its standard deviation. Equation (2.6) shows this relation [17].

$$t-stat_{\beta_i} = \frac{\hat{\beta}_i - \beta_{i,0}}{\sigma_{\beta_i}}$$

(2.6)

where $\beta_{i,0}$ corresponds to the hypothesis made for the $\beta_i$.

In the case of this project we always assumed $\beta_{i,0} = 0$. Therefore, the higher the value of t-stat, in our particular case, the higher the accuracy of the estimated parameter. However, too high a value would imply that the data points are serially correlated [17].

Some evaluation quantities of the resulting model are the mean absolute error, MAE, the mean absolute percentage error, MAPE, and the root mean
squared error, RMSE. These quantities do not take into account the under or over prediction of the model but measure the overall error, i.e. the difference between the modeled and the actual value of the dependent variable. This error derives from the error terms of the estimation, $u_t$, but also from the error in estimating the coefficients $\beta$ in the regressive and autoregressive parts of the model [17].

MAE is given by Equation (2.7) and is the average absolute error:

$$MAE = \frac{1}{n} \sum_{t} |u_t|$$  \hspace{1cm} (2.7)

MAPE is similar to MAE but, as shown in Equation (2.8), each difference is divided by the actual value of the dependent variable.

$$MAPE = \frac{100\%}{n} \sum_{t} \frac{|u_t|}{Y_t}$$  \hspace{1cm} (2.8)

Finally, RMSE is the root of the mean of the squared errors, as shown in Equation (2.9).

$$RMSE = \sqrt{E(Y - \hat{Y})^2}$$  \hspace{1cm} (2.9)

### 2.2 Time series analysis

A time series is an ordered sequence of observations [18]. The ordering is usually made through time as implied by its name, but also spatial ordering is possible. Examples of time series can be found in economics, engineering, natural and social sciences. Time series can be either continuous or discrete, however, discrete values can be managed more easily in terms of computational complexity. The main objective of time series analysis is the better understanding and description of a mechanism, the forecast of future values and the improvement in the control a system [18, 19].

There are several different types of time series in terms of their characteristics. Stationary time series vary around a fixed level. Non-stationary ones may exhibit a trend or a seasonal variation. The latter are called seasonal time series and in the case of this project load time series belong to this category [18].

Scientists, after carefully observing a physical phenomenon, try to create a mathematical model in order to describe it. Such models are called deterministic and if no other perturbation is present it can describe accurately the phenomenon. However, there is no phenomenon in real life that is completely deterministic. Unobservable factors play an important role in the distortion, to a small or large extent, of the final outcome. Therefore, it is necessary to derive a model that calculates the probability that the outcome is within a certain range. Such models are the stochastic models [19].
Stochastic processes

A stochastic process is a family of time indexed random variables which describes the probability structure of the observations \[18, 19\]. The main goal of such models is to identify the main properties of the phenomenon based on the observations.

Assuming a stochastic process \(Z_t\) with mean \(E(Z_t) = \mu\) and variance \(Var(Z_t) = E(Z_t - \mu)^2 = \sigma^2\) the covariance between \(Z_t\) and \(Z_{t+k}\) is called the autocovariance function \[18, 19\]:

\[
\gamma_k = Cov(Z_t, Z_{t+k}) = E(Z_t - \mu)(Z_{t+k} - \mu) \quad (2.10)
\]

The autocorrelation function (ACF) is then defined as the correlation between \(Z_t\) and \(Z_{t+k}\) \[18, 19\]:

\[
\rho_k = \frac{Cov(Z_t, Z_{t+k})}{\sqrt{Var(Z_t)}\sqrt{Var(Z_{t+k})}} \quad (2.11)
\]

Another important quantity in terms of identifying seasonal correlations within a sample is the partial autocorrelation function (PACF). This function represents the correlation between \(Z_t\) and \(Z_{t+k}\) with the linear dependency on the intervening variables in the time steps from \(t + 1\) to \(t + k - 1\) removed \[18, 19\].

\[
P_k = \frac{Cov([Z_t - \hat{Z}_t], (Z_{t+k} - \hat{Z}_{t+k})]}{\sqrt{Var(Z_t - \hat{Z}_t)}\sqrt{Var(Z_{t+k} - \hat{Z}_{t+k})}} \quad (2.12)
\]

where \(\hat{Z}_t\) and \(\hat{Z}_{t+k}\) are estimates of \(Z_t\) and \(Z_{t+k}\) as a linear function of \(Z_{t+1}, ..., Z_{t+k-1}\) \[18\].

An important stochastic process which we will see later on in this project for the description of the error terms is the white noise. A stochastic process is a white noise process if it is a sequence of non-correlated random variables from a fixed distribution with constant mean, variance and covariance. This process is a Gaussian white noise if its distribution is normal \[18\].

Autoregressive models

A stationary stochastic process can be modeled with the use of an autoregressive (AR) model. In this AR model, the current value of the process is represented as a linear combination of previous values. In the following equation, an autoregressive model of order \(p\) is shown \[19\]:

\[
\hat{Z}_t = \phi_1 \hat{Z}_{t-1} + \phi_2 \hat{Z}_{t-2} + ... + \phi_p \hat{Z}_{t-p} + \epsilon_t \quad (2.13)
\]

The parameters \(\phi_i\) must satisfy the condition \(\sum |\phi_i| < 1\) to ensure stationarity.
CHAPTER 2. METHODOLOGY

However, the choice of the order $p$ should not be arbitrary. A useful tool for the selection of the order can be the PACF of the sample [19]. If the resulting values for lags greater than $p$ are 0, then $p$ should be the order of the AR model of the sample. In many cases the resulting PACF values are essentially (but not exactly) zero after a certain lag, if the approximate of the PACF is zero at a significance level of 5% [19].

Non-stationary stochastic processes have to be transformed into stationary ones with the extraction of the deterministic part in order to be modeled with the use of an autoregressive model. In case there is a linear upward trend, for example, it can be modeled separately and then extracted from the full time series. Generally, there are several different types of deterministic trends. The most common ones are the $k$-th order polynomial trend and the sinusoidal trend. These types can be easily modeled with the use of regression analysis. Many non-stationary processes, though, can be sufficiently modeled with the combined use of autoregression and moving average approaches as explained in [18, 19], such as ARMA (autoregressive moving average) and ARIMA (autoregressive integrated moving average) models. However, it is not in the purpose of this project to further analyze them since other methods were used.

OLS estimation in time series analysis

Regression analysis is the most commonly used method in statistical analysis of time series. Linear regression, which is the most simple approach of regression analysis, uses the ordinary least squares estimation as the best unbiased estimator for the parameters of the model [17, 18].

It is very frequent for statistical time series to follow a seasonal pattern. Decomposition of the data is necessary in order to model correctly the different patterns, the seasonal ($S_t$), the trend ($T_t$) and the stochastic patterns or deterministic approach error ($\epsilon_t$). The seasonal ones can be modeled with the use of a linear combination of sinusoidal functions for various frequencies. Then, the linear trend can be modeled as a polynomial time function or as a function of a different regressors that drive this trend [19].

\[
Z_t = T_t + S_t + \epsilon_t \\
S_t = \beta_0 + \sum_{i=1}^{n} \beta_{1,i} \sin(2\pi if) + \beta_{2,i} \cos(2\pi if) \\
T_t = \alpha_0 + \sum_{j=1}^{m} \alpha_j t^j
\] (2.14)

where $f$ represents the fundamental frequency of the sample.

It should be noted that, a regression model in time series data, as the ones in (2.1) and (2.2) that model the effect of some variables on the dependent
variable, may produce error terms which still contain autocorrelation. This means that the error terms need further modeling in order to extract further seasonal properties that were left out [19]. AR models can offer a sufficient solution to this problem if the selection of order $p$ is correctly selected. As mentioned above, PACF offers a simple and efficient tool for the selection of the order of an AR model that can produce error terms that have negligible autocorrelation.

Other similar approaches make use of moving average and seasonal ARIMA models, but the integration would cancel out the seasonality that we wanted to model. ARMA models may perform slightly better than simple AR models in the stochastic part but we decided to use the latter for simplicity reasons. [19] and [18] offer sufficient literature on ARMA and ARIMA models.

Forecasting

The objective of this project is to create a model that can describe well the load data of the past and that can predict as accurately as possible the load demand of the future. A forecast model needs first to be calibrated within the available sample using the methodologies mentioned in this chapter. In order to obtain a confident model, its in-sample performance should be sufficiently successful. However, it is obvious that testing a model within the same data that were used for calibration and fitting cannot constitute its validation. Therefore, it is necessary to leave a part of the available sample, usually the most recent part, outside the calibration procedure. Then, the resulting model can produce a pseudo-forecast for that period and the comparison of this prediction with the actual data that were unused constitute the out-of-sample evaluation of the model. This method can be used as a validation of the model because it is more objective than the in-sample evaluation, even though it is still quite biased, since during the construction of the model there is knowledge of these pseudo-future values. Finally, the real evaluation and validation of a forecast model is made in real time, for data that are selected after the formulation of the model. This comparison is completely objective and unbiased but takes a lot of time for long-term forecasts such as ours [17, 19, 18].

Quantities presented before, such as $R^2$, MAE, MAPE and RMSE can be used for the evaluation and the performance assessment of the model. Another comparison that can be made in terms of the performance of the model is between the distribution function of the actual data and the modeled data.
Chapter 3

Data and Analysis

In this chapter, the data that was used for the formulation of the load model will be introduced. Then, the main drivers of load demand will be discussed and the particular case of Germany will be further analyzed. The effect of each factor on load demand will be separately studied in order to better understand their relation. Finally, a good insight on the seasonality of load demand will be given with focus on its fundamental frequencies.

3.1 Data

In order to create a concrete model, reliable data are necessary. Since load demand is a complex process that depends on several different factors, load data alone were not enough to perform an accurate forecast model. In order to analyze the effects of these factors and better understand the nature of load demand, electric load data were used along with weather, economic and calendar data.

The most important information necessary for the formulation of such a model is the load data. Historic data of hourly load for Germany were available from the European Network of Transmission System Operators for Electricity (ENTSO-E) [20]. We used data from 2006 until 2011. An overview of that can be seen in Figure (3.1) along with a smaller scale example that shows the weekly and hourly load patterns. The seasonality of the load can be better understood by studying its moving average for different time windows. Figure (3.2) shows the hourly load along with the semester and monthly moving average. It is obvious that the load is higher during the winter and lower during the summer. Another observation is that there is a steep change especially in the monthly moving average at the end of each year which is caused by the Christmas vacation. Finally, it should be noted that for the year 2009 when the economic crisis became more apparent, the load demand is significantly reduced in comparison to the preceding and succeeding years.
Following this last remark, it became crucial that for the formulation of an accurate model, it would be necessary to include economic data. GDP and industrial production data were used for the better understanding of the relation between the economic trends and load demand, but also for the formulation of the model. These data were obtained from the Organization
for Economic Co-operation and Development (OECD) [21]. Figure (3.3) shows the quarterly and monthly data respectively for the GDP and industrial production indicator for Germany. The significant drop of these two indicators for the year 2009 because of the financial crisis should be noted.

![Figure 3.3: Industrial production indicator and GDP for Germany](image)

Furthermore, temperature data was obtained from 61 different weather stations from the German weather agency [22]. The average daily temperature from 1973 up to date was used. In Figure (3.4) an overview of these temperature data for our sample is shown. In addition, Figure (3.5) depicts the historic daily average temperature data for Germany. The monthly moving average is used for this figure in order to smooth the resulting curve.

Finally, data on national public holidays were necessary for the correct calibration and prediction of the load demand. These data were obtained from [23].

### 3.2 Analysis

In the previous section, an overview of the used data was presented. The analysis of this data will be shown hereafter. Firstly, the factors that influence the load demand will be discussed. Then, our analysis for the particular case of Germany will be presented in order to better understand the selection of the regressors for the formulation of the load model.

A first feeling on the main factors that stir the electricity load demand could lead to the weather and economic activity. The profile of the consumers, as well, plays an important role. The living standards of the popu-
The participation alongside with the participation of the commercial, industrial, transportation and other sectors in the consuming mix will determine the factors and to what extent they affect the particular case of the load demand in an area or a country.

Literature review of [3, 13] and analysis of the load data available for
CHAPTER 3. DATA AND ANALYSIS

Germany led us to the conclusion that the following factors influence the load demand:

- Weather (temperature, humidity, sunlight, wind speed)
- Season
- Economic trends
- Day of the week
- Hour of day
- Public holidays and vacations
- Unpredictable events
- Population growth

Weather

Careful study of the effect of various weather properties in electric load demand is necessary in order to be able to extract a relation that could be used as an input for a load model. According to [24] most electricity load forecast models have made use of temperature alone, some have included humidity, fewer have included even further weather properties and others have included none. In this master thesis we decided to use only temperature for the modeling of German load.

[25, 26, 27] indicate that temperature is inversely proportional to load demand for temperatures below $\sim 16-17^\circ C$, depending on the geographic region, and directly proportional for temperatures above $\sim 20^\circ C$. In the particular case of Germany, Figure (3.6) depicts the relation between total daily consumption and average daily temperature for working days. It is clear, in this case, that load demand decreases with the increase of temperature. Even though the relation is not directly linear and the quadratic fit has been included in the figure, a linear relation could very well fit the plotted data. This is because the average daily temperature in Germany raises only rarely above $20^\circ C$ that would drive the cooling needs to increase the total load demand. Therefore, as we will see later on, our model assumes, for simplicity, a linear relation between temperature and load demand and the results are promising.

Furthermore, the inversely proportional effect of temperature and load can be seen in Figure (3.7). In winter time, when load is maximum, the minimum temperatures occur. Similarly, during the summer, when higher temperatures are present, the load is minimum.

Season

Concerning the seasonalities evident in the hourly load one could state that they are already present in the temperature time series. However, other
seasonal events such as the length of the day and other seasonal economic and social activities, that depend on the consumption profile of the region in question, need to be modeled as well.
Figure (3.8) shows how the average hourly load pattern varies throughout the year. It is evident, as mentioned previously, that the load demand is higher in the winter. However, the peak demand in the summer is at midday while in the winter it is later in the evening.

We have already seen in Figure (3.2) the long term seasonal effects in load demand. However, it should be noted that around Christmas and New Year’s eve, every year, there is a significant decrease in the load demand because many people go on holidays and the economic activities other than the commercial are not as active as the preceding and following weeks. This is more obvious in the monthly moving average load curve.

Finally, Figure (3.9) shows the overall seasonalities that are present in the German hourly load data. The yearly, weekly and daily fundamental frequencies are shown along with their harmonics. Yearly and weekly harmonics of greater order are more significant than the daily harmonics, but the daily frequency amplitude is greater than those for the yearly and weekly ones.

Economic trends

One of the most important factors that influences load demand in the longer term is the economic situation in the region in question. The trend in the past few years in most countries, especially those in OECD and China, was that electricity consumption was slowly growing [28]. So were the economic indicators and especially GDP [29].

2009 was a year when the economic crisis led to a recession in many of the developed countries. This is also evident in Germany, as shown in
Figure 3.9: FFT on German hourly load time series. Fundamental frequencies and their harmonics

Figure (3.3). The fact that at the same time the electricity demand drops, as depicted in Figure (3.2), is interesting. This ‘coincidence’ can be observed in Figure (3.10), where GDP, industrial production indicator and load semester moving averages are shown for Germany.

The question that arises, though, is which of the two economic indicators is more suitable to describe this macroscopic effect of the economy on load demand. The answer to that question depends on the consumption profile of the region in question. Germany is a country whose industrial sector constitutes more than 40% of the final electricity consumption [30]. Both indicators were tried for modeling of the German hourly electricity load, but in the end, industrial production performed much better. This can be explained also by a careful examination of Figure (3.11), where the relation between monthly average load and industrial production in Germany is nearly perfectly linear.

Day of the week, hour of day, holidays and other factors

In smaller scale, the factors that determine the load level and load pattern depend on the type of day in question, that is whether it is a regular working day, weekend or holiday. Then, in even smaller scale, the hour of day of a specific day determines the actual hourly load level.

It is obvious that load on a non-working day will be less than that on a working one, but also the load shape will be significantly different. The latter is because the consumption mix will be different in such a day. In
CHAPTER 3. DATA AND ANALYSIS

Figure 3.10: Semester moving averages for load, industrial production and GDP for Germany

Figure 3.11: Monthly average load against industrial production indicator in Germany for 2006-2011

addition, load demand during night time and early morning will be much less than during the day.

All the above can be further examined and supported in Figure (3.12). An interesting observation is that 3rd of October which is the German Unity Day, a public holiday for Germany, has the load level and shape of a Sunday. Figure (3.13) shows another example of a smaller scale load plot, this time in
the month of April, which includes the Easter holidays. In Germany, Friday and Monday before and after Easter constitute public holidays. However, in this second figure it is not only the holiday load level and shape that is of interest, but also the fact that the load shape during weekdays in April and October is significantly different. In addition, a difference in the load shape between Saturdays and Sundays can be noticed, especially in the first image for 27th and 28th September 2008 and in the second image for the 31st of March and 1st of April 2007. The yearly seasonality of the hourly load shape throughout the year has already been observed in Figure (3.8).

Finally, unpredictable events, such as big sport events or elections, can influence the load demand. Usually such cases affect the load demand in a regional rather than a national level. Furthermore, population growth could be considered as an important factor concerning load demand, but in developed countries this growth rate is not so significantly high and we could assume that it is already taken into account in the economic indicators. For load demand forecasts of longer terms, such as 5 years or more, it should be also taken into account.
Figure 3.13: Example of weekly and daily load pattern for Germany during Easter of 2007
Chapter 4

Modeling

The load data, analyzed in this thesis, is assumed to be a time series. Therefore, it consists of a deterministic and a stochastic part as shown in Equation (4.1) [19, 18],

$$dL_t = f(L_t, t)dt + g(L_t, t)dW_t,$$

where $L$ is the load, $f$ is the deterministic part, $g$ the stochastic and $dW_t$ denotes the differential form of the Brownian motion, described by a Wiener process.

If we simplify the above equation we obtain the following form to describe the hourly load demand model:

$$L(t) = f(t) + \tilde{x}(t),$$

where $\tilde{x}$ is the stochastic part of the load.

There are several ways to model the deterministic part. The main concepts considered were the sinusoidal models and the factor models. After carefully examining both, it was decided to use the factor models for various reasons. On the one hand, sinusoidal approaches can model only periodic events and situations that do not have a seasonality or occur every year on a different date - such as moving holidays - cannot be modeled correctly. On the other hand, it is important to understand which factors influence the load and to what extent, not only which frequencies are more dominant.

For the factor models, multiple linear regression was used in order to calibrate the model parameters. As mentioned in chapter 2, the value of the dependent variable is estimated as a linear combination of the independent variables - regressors, as in Equation (4.3) [17].

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \ldots + \beta_k x_{kt} + u_t = \hat{y}_t + u_t, \quad t = 1, \ldots, T$$

where $y$ is the dependent variable, $\hat{y}$ is an estimation of the dependent variable, $x$ the regressors, $u_t$ the error term and $T$ the length of the sample.

The calculation of the parameters $\beta$ is done using the ordinary least squares method [17, 18]:

25
\[ \min_\beta \sum u_t^2 = \sum (y_t - \hat{y}_t)^2 \]  

(4.4)

The load data used for the calibration of our model ranged from January 2006 until December 2010 [20]. Our initial approach was to model the complete time series of the load data. The inclusion of the most important driving factors of load demand, such as temperature, economic trends, the difference between working and no working days and between each hour of the day, was necessary. After careful study of these initial results two things became obvious. On the one side, there was need to better capture the seasonal yearly effect on the load level and shape, and on the other, the different load shape between different day types. Therefore, we decided to decompose the data into daily load level and hourly load, as can be seen in Equation (4.5). In addition, we added the monthly and day type effect on the load level and shape in order to incorporate all the drivers of load demand. This idea was also proposed in [15].

\[ \tilde{L}(t) = \tilde{L}_d(t) + \tilde{L}_h(t) \]  

(4.5)

### 4.1 Daily load level modeling

In order to model the daily load level we had to form the daily load level time series from our hourly data. This was done by taking the average load for every 24 hours. Hence, our time series consisted of 365 values per year instead of 8760. In addition, it is clear that these time series consist of a deterministic and a stochastic part.

\[ \tilde{L}_d(t) = f_d(t) + \tilde{x}_d(t) \]  

(4.6)

where \( f_d \) is the deterministic part and \( \tilde{x}_d \) the stochastic.

#### 4.1.1 Deterministic part

For the modeling of the deterministic part of these time series, it was necessary to include all long term effects on load such as seasonality, economic trends, temperature and of course type of day. After careful observation of our initial trial and error approaches the following equation was used for the deterministic part.
\[ f_d(t) = \alpha_{d0} + \alpha_{d1} \cdot IndustrialProduction(t) + \alpha_{d2} \cdot Temperature(t) \]
\[ + \sum_{i=1}^{12} \beta_{d,i} \cdot Month_i(t) + \alpha_{d3} \cdot WorkingDay(t) \]
\[ + \alpha_{d4} \cdot Holiday(t) + \alpha_{d5} \cdot HalfHoliday(t) + \sum_{j=1}^{7} \gamma_{d,j} \cdot Weekday_j(t) \]

where \( \alpha, \beta, \gamma \) are parameters to be calculated, \( IndustrialProduction \) and \( Temperature \) are obvious time series, \( Month \) is a 12 column binary matrix where for each time step (row) the column of the corresponding month has a value of one and the remaining ones zero, \( WorkingDay, Holiday \) and \( HalfHoliday \) are binary vectors whose value is one on a given time step if it is a working day, a holiday or a so called half-holiday respectively, and \( Weekday \) is a seven column binary matrix that follows the same pattern as \( Month \) matrix this time for the seven weekdays. It is important to note here for better understanding of our regressors that the column sum of the matrices \( Month \) and \( Weekday \) is one for all rows. All parameters were calculated with the use of multiple linear regression (‘regress’ Matlab command) and OLS estimation.

The concept of a half-holiday will be hereby defined in order to clarify all the regressors used. After careful observation of the load demand throughout the five years in our sample - 2006 to 2010 - we came across some days whose load level was less than a usual working day and more than a weekend’s usual load. Therefore, we decided to define the half-holidays. Days that were included in this day type were those in between Christmas and New Year’s day when many people are away on vacation and also days used to “bridge” public holidays with the weekends to form long weekends, such as Fridays after a public holiday on a Thursday and Mondays before a public holiday on a Tuesday.

An overview of the in-sample results to this initial approach can be seen in Figure (4.1) and on a smaller scale in Figure (4.2). An important observation from the first figure is that for the year 2009, when the economic crisis is more evident, the load level is significantly less than the preceding and succeeding years. However, the inclusion of the industrial production in the regressors has successfully captured this abnormality.

The residual of the deterministic model - the difference between the actual daily load and the deterministic part - is according to Equation (4.6) the stochastic part of the daily level time series. An overview of this residual can be seen in Figure (4.3). It can be observed that this residual is significantly greater during Christmas vacation. This can be explained by the fact that each year the load level around this time significantly depends on which day is Christmas and New Year’s day. This affects the number of
days people decide to take off from their work and the difference in load level was not predicted accurately enough. It is important to study better and understand the mechanisms that drive people’s decisions on their vacation dates depending on the day that a public holiday occurs.

4.1.2 Stochastic part

The actual load level on a given day is a fairly random value because it depends on many factors that cannot be predicted accurately, such as the behaviour of people. Therefore, this value depends a lot on the previous values of the daily load level. We decided to use an autoregressive (AR) model in order to correctly capture these temporal correlations on the daily
load. In Equation (4.8), we can see the basic idea of this method [19].

\[ \tilde{x}(t) = c + \sum_{i=1}^{p} \delta_i \cdot \tilde{x}(t - i\Delta t) + \tilde{\epsilon}(t) \] (4.8)

where \( c \) is the constant term, \( \delta_i \) are the effects of the previous values in the present value and \( \epsilon \) is the error term.

In Figure (4.4) the autocorrelation of the stochastic part is depicted. It is obvious that there is quite a significant correlation in the first few lags which drops as the lags are increased.

Figure 4.3: Daily level load stochastic part

\[ \tilde{x}(t) = c + \sum_{i=1}^{p} \delta_i \cdot \tilde{x}(t - i\Delta t) + \tilde{\epsilon}(t) \]

where \( c \) is the constant term, \( \delta_i \) are the effects of the previous values in the present value and \( \epsilon \) is the error term.

In Figure (4.4) the autocorrelation of the stochastic part is depicted. It is obvious that there is quite a significant correlation in the first few lags which drops as the lags are increased.

Figure 4.4: Autocorrelation of the stochastic part of the daily load level

We used the partial autocorrelation function in order to determine how
many lags it was necessary to include for the appropriate model the stochastic part [19]. In Figure (4.5), it can be seen that the first three lags are significant enough and therefore they were the ones used. The AR model of the stochastic part of the daily load level is shown in Equation (4.9).

\[
\tilde{x}_d(t) = c_d + \sum_{i=1}^{3} \delta_{d,i} \cdot \tilde{x}_d(t - i\Delta t) + \tilde{\epsilon}_d(t)
\]  

(4.9)

4.1.3 Overall daily load level model

In Figure (4.6) the overview of the overall model for the daily load level is shown. Even though it is a large scale image, it can be seen that most of the patterns of the daily load level have been captured. Especially, the “tricky” year 2009 where the financial crisis led to much less load demand than the preceding and succeeding years, is well covered with the use of the industrial production indicator. A closer look of smaller scale is shown in Figure (4.7). The residuals now have been reduced significantly.

The overall residual of the daily load level model is shown in Figure (4.8). The standard deviation of the residuals of the deterministic part is \(\sigma_{dd} = 1.79GW\) while the one for the stochastic part is \(\sigma_d = 1.40GW\). The standard deviation of the stochastic model, calculated by solving the system of equations in (4.10) [19] for a 3rd order AR model, is \(\tilde{\sigma}_d = 1.78GW\) which is almost the same as \(\sigma_{dd}\).
Figure 4.6: Overview of the complete daily load level model (yearly scale)

Figure 4.7: Daily load level model (weekly scale)

\[
\begin{align*}
\rho_1 &= \delta_{d,1} + \delta_{d,2} \rho_2 + \delta_{d,3} \rho_2 \\
\rho_2 &= \delta_{d,1} \rho_1 + \delta_{d,2} + \delta_{d,3} \rho_1 \\
\rho_3 &= \delta_{d,1} \rho_2 + \delta_{d,2} \rho_1 + \delta_{d,3} \\
\sigma_{x_d} &= \frac{\sigma_d^2}{1 - \rho_1 \delta_{d,1} - \rho_2 \delta_{d,2} - \rho_3 \delta_{d,3}}
\end{align*}
\]

where \( \sigma_d \) is the standard deviation of the residual of the AR model, \( \delta_{d,i} \) the parameters of the process and \( \rho_i \) the parameters of the autocorrelation.

In Figure (4.9), where the autocorrelation of the residual is depicted, it is obvious that most temporal correlations of the load have been adequately modeled. Furthermore, as Figure (4.10) shows the distribution of the residuals can be considered fairly normal. Therefore, this residual can
be modeled as white noise with standard normal distribution. Finally, the overall equation of the model is shown in Equation (4.11).

$$\tilde{L}_d(t) = f_d(t) + c_d + \sum_{i=1}^{3} \delta_{d,i} \cdot \tilde{x}_d(t - i \Delta t) + \sigma_{d} \tilde{\epsilon}_{d,w}(t)$$  \hspace{1cm} (4.11)
CHAPTER 4. MODELING

Figure 4.10: Histogram of the residual of the overall residual of the daily load level model

After the completion of the daily load level model, we proceed with the modeling of the hourly part in the next section.

4.2 Hourly load modeling

After capturing the average load of every day, we obtain the residual hourly load with the use of Equation (4.5) by subtracting the daily load from the total load. The resulting time series consists of the hourly load data which have a mean of practically zero and capture only the hourly pattern of the load of every day relative to the daily load level.

\[
\tilde{L}_h(t) = \tilde{L}(t) - \tilde{L}_d(t) \tag{4.12}
\]

The hourly load data will again consist of a deterministic and a stochastic part, as it can be observed in Figure (4.11). In this figure we see that there is a daily seasonality, especially for working days, and a weekly seasonality because of the weekends. In addition, there is a change in the daily load pattern depending on the time of year. Therefore, this deterministic part can be captured in a similar way as we did for the daily load and the residual of that with the use of an autoregressive model.

\[
\tilde{L}_h(t) = f_h(t) + \tilde{x}_h(t) \tag{4.13}
\]

4.2.1 Deterministic part

As discussed previously and after careful examination of the hourly load data, it became obvious that it is mainly influenced by the season, the type
of day and the corresponding hour of day. Thus, the deterministic part was modeled as per Equation (4.14).

$$f_h(t) = \sum_{i=1}^{3} \sum_{j=1}^{12} (d_{i,j} + \sum_{k=1}^{24} \zeta_{i,j,k}(t) \cdot \text{Hour}_{i,j,k}(t))$$  \hspace{1cm} (4.14)

In order to obtain better results we performed linear regression for each type of day and each month separately. The first summation in Equation (4.14) corresponds to the different types of day. Observation of the data indicated that there was a difference in the load patterns between working days, Saturdays and Sundays. Thus, we defined three different day types: Working days, Saturdays (including half-holidays) and Sundays (including holidays). The second summation obviously corresponds to the twelve months, because it was observed that the load pattern depends on the temperature and sunlight duration and changes from month to month. $c_{i,j}$ is the constant term for each regression, $\zeta_{i,j,k}$ is the effect of hour $k$ on the hourly load of a day of type $i$ on month $j$ and $\text{Hour}_{i,j,k}$ is a binary variable that takes the value of one in case $t$ corresponds to hour $k$ and zero otherwise.

In Figure (4.12) a close view of the result of the deterministic approach can be seen. The difference between the actual values of the hourly load and those of the modeled deterministic part form the stochastic part. This is shown in Figure (4.13).
4.2.2 Stochastic part

For the modeling of the stochastic part of the hourly load we used an AR model as we did for the daily load level stochastic part. The temporal correlation of the hourly load stochastic part can be seen in Figures (4.14) and (4.15) where its autocorrelation for a working day in December and June is depicted. The autocorrelation of the residuals for different types of day and months follow a similar pattern to that. In these figures we can see that there is a high correlation in the first few lags (hours) which is gradually reduced as the lags increase.

In order to determine the number of lags to be used for the AR model, we used again the partial autocorrelation function. In Figures (4.16) and
Figure 4.14: Autocorrelation of the hourly load model stochastic part for a working day in December

Figure 4.15: Autocorrelation of the hourly load model stochastic part for a working day in June

(4.17) we can see the partial autocorrelation function for working days in December and June. The other months and types of days follow a similar pattern. Therefore, we used the first two lags for the AR modeling of the stochastic part as in Equation (4.15).
Figure 4.16: Partial autocorrelation function of the hourly load model stochastic part for a working day in December

Figure 4.17: Partial autocorrelation function of the hourly load model stochastic part for a working day in June

\[
\tilde{x}_h(t) = \sum_{i=1}^{3} \sum_{j=1}^{12} (c_{h,i,j} + \sum_{k=1}^{2} \delta_{h,i,j,k} \cdot \tilde{x}_{h,i,j}(t - k\Delta t) + \tilde{e}_{h,i,j}(t)) \quad (4.15)
\]
4.2.3 Overall hourly load model

In Figure (4.18) a part of the overall model for the hourly load is depicted. It is clear that the residuals have been reduced significantly. The standard deviation of the residual of the deterministic part of the hourly load is $\sigma_{hd} = 1.314 \text{ GW}$, while the standard deviation of the stochastic residual is $\sigma_h = 639.7 \text{ MW}$. The standard deviation of the stochastic model, calculated as per Equation (4.16) [19] for a 2nd order AR model in a similar way as for the daily stochastic part, is $\sigma_{\tilde{x}_h} = 1.31\text{ GW}$ which is almost the same as $\sigma_{hd}$.

\[
\sigma_{\tilde{x}_h}^2 = \frac{(1 - \delta_{h,2})\sigma_h^2}{(1 + \delta_{h,2})((1 - \delta_{h,2})^2 - \delta_{h,1}^2)} \tag{4.16}
\]

where $\sigma_h$ is the standard deviation of the residual of the AR model and $\delta_{h,i}$ the parameters of the process.

Furthermore, in Figures (4.19) and (4.20) the autocorrelation of the residual of the stochastic part is shown for the working days of December and June and it is clear that the temporal correlations have been successfully modeled. The remaining months and types of days have a similar result. In addition, Figures (4.21) and (4.22) show that the distribution of the overall residual is quite normal and it can be modeled as white noise with standard normal distribution. The overall equation of the hourly model is shown in Equation (4.17).

\[
\tilde{L}_h(t) = f_h(t) + \sum_{i=1}^{3} \sum_{j=1}^{12} (c_{h,i,j} + \sum_{k=1}^{2} \delta_{h,i,j,k} \cdot \tilde{x}_{h,i,j}(t - k\Delta t) + \sigma_{h,i,j} \tilde{\epsilon}_{h,i,j,w}(t))
\tag{4.17}
\]
Finally, we proceed with the complete model of the load in the next section.
CHAPTER 4. MODELING

4.3 Overall load model

Incorporating the daily load level values to the hourly load modeled in the previous section we obtain the complete model or in-sample model for the German electricity load. An overview of all the equations used to create our load model can be seen in Equation (4.18). Therefore we include in Equation (4.18a) the Equations (4.18c), and (4.18d), containing the daily
deterministic and stochastic part, and the Equations (4.18f) and (4.18g), containing the hourly deterministic and stochastic part.

**Overall model:**
\[ \tilde{L}(t) = \tilde{L}_d(t) + \tilde{L}_h(t) \]  
(4.18a)

**Daily part model:**
\[ \tilde{L}_d(t) = f_d(t) + \tilde{x}_d(t) \]  
(4.18b)

\[ f_d(t) = \alpha_{d0} + \alpha_{d1} \cdot \text{IndustrialProduction}(t) + \alpha_{d2} \cdot \text{Temperature}(t) \]
\[ + \sum_{i=1}^{12} \beta_{d,i} \cdot \text{Month}_i(t) + \alpha_{d3} \cdot \text{WorkingDay}(t) \]  
(4.18c)

\[ + \alpha_{d4} \cdot \text{Holiday}(t) + \alpha_{d5} \cdot \text{HalfHoliday}(t) + \sum_{j=1}^{7} \gamma_{d,j} \cdot \text{Weekday}_j(t) \]
\[ \tilde{x}_d(t) = \tilde{c}_d + \sum_{i=1}^{3} \delta_{d,i} \cdot \tilde{x}_d(t - i\Delta t) + \sigma_{d,w}(t) \]  
(4.18d)

**Hourly part model:**
\[ \tilde{L}_h(t) = f_h(t) + \tilde{x}_h(t) \]  
(4.18e)

\[ f_h(t) = \sum_{i=1}^{3} \sum_{j=1}^{12} (d_{i,j} + \sum_{k=1}^{24} \zeta_{i,j,k} \cdot \text{Hour}_{i,j,k}(t)) \]  
(4.18f)

\[ \tilde{x}_h(t) = \sum_{i=1}^{3} \sum_{j=1}^{12} (c_{h,i,j} + \sum_{k=1}^{2} \delta_{h,i,j,k} \cdot \tilde{x}_{h,i,j}(t - k\Delta t) + \sigma_{h,i,j}\tilde{\epsilon}_{h,i,j,w}(t)) \]  
(4.18g)

An overview of the result of the complete model can be seen in Figure (4.23) and a more closer view in Figure (4.24). The residual of the complete model is depicted in Figure (4.25). The mean of the residual is practically zero (\(-3\text{MW}\) while the load average is around 55\text{GW} - 0.055\%) and its standard deviation 1.54\text{GW}.

The overall computational complexity for the model calibration and parameter estimation is relatively low. The time necessary for the completion of all computations and calculation of the parameters estimations was no more than one minute using MATLAB\textsuperscript{®} on a standard PC.

### 4.3.1 Model parameters

Table (4.1) shows the calibrated parameters of Equation (4.18c). \(\alpha_{d0}\) corresponds to the constant term of the daily load level, \(\alpha_{d1}\) to the industrial production vector, \(\alpha_{d2}\) to the temperature, \(\alpha_{d3}\) to the working-day vector, \(\alpha_{d4}\) to the holiday and \(\alpha_{d5}\) to the half-holiday vector. It should be noted
that the significance of the holiday vector is surprisingly low ($t - \text{stat value}$). However, this can be explained by the fact that the effect on load demand of these days is partially captured also by the working-day vector and that each holiday occurs in a different season and therefore the load demand may be significantly different. $\beta_{d,6}$ corresponds to the regressor for the month of June, $\beta_{d,12}$ for the month of December, $\gamma_{d,1}$ to Sundays and $\gamma_{d,5}$ to Thursdays. Then, in Table (4.2) the stochastic parameters of the daily load level model from Equation (4.18d) are shown.

In Table (4.3) the calibrated parameters for the hourly deterministic part of Equation (4.18f) are shown. $\zeta_{1,6,12}$ and $\zeta_{1,6,24}$ correspond to the regressors
Table 4.1: Parameters of the daily load level deterministic part

<table>
<thead>
<tr>
<th>α_{d0}</th>
<th>α_{d1}</th>
<th>α_{d2}</th>
<th>α_{d3}</th>
<th>α_{d4}</th>
<th>α_{d5}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>5.54 \cdot 10^3</td>
<td>206.45</td>
<td>-151.79</td>
<td>1.19 \cdot 10^4</td>
<td>-395.91</td>
</tr>
<tr>
<td>t-stat</td>
<td>NaN</td>
<td>38.06</td>
<td>12.46</td>
<td>13.44</td>
<td>0.45</td>
</tr>
</tbody>
</table>

Table 4.2: Parameters of the daily load level stochastic part modeling

<table>
<thead>
<tr>
<th>c_{d0}</th>
<th>δ_{d1}</th>
<th>δ_{d2}</th>
<th>δ_{d3}</th>
<th>σ_{d}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>0.18</td>
<td>0.52</td>
<td>-4.5 \cdot 10^{-3}</td>
<td>0.1835</td>
</tr>
<tr>
<td>t-stat</td>
<td>5.5 \cdot 10^{-3}</td>
<td>22.53</td>
<td>0.17</td>
<td>8.00</td>
</tr>
<tr>
<td>R^2</td>
<td>0.384</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

for noon (12.00) and midnight (00.00) respectively for working days in the month of June. Similarly, ζ_{1,12,12} and ζ_{1,12,24} correspond to noon (12.00) and midnight (00.00) for working days in the month of December. Then, in Table (4.4) the stochastic parameters of the hourly load from Equation (4.18g) are shown. The first value of the subscript of the stochastic parameters corresponds to the type of day (1 for working days, 2 for Saturdays and 3 for Sundays), the second corresponds to the month and the third to the number of lag in the autoregressive model.
### Table 4.3: Parameters of the hourly load deterministic part modeling

<table>
<thead>
<tr>
<th></th>
<th>( d_{1.6} )</th>
<th>( \zeta_{1.6,12} )</th>
<th>( \zeta_{1.6,24} )</th>
<th>( d_{1.12} )</th>
<th>( \zeta_{1.12,12} )</th>
<th>( \zeta_{1.12,24} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>(-2.26 \cdot 10^3)</td>
<td>(1.23 \cdot 10^3)</td>
<td>(-4.33 \cdot 10^4)</td>
<td>(-1.28 \cdot 10^4)</td>
<td>(1.97 \cdot 10^4)</td>
<td>(6.52 \cdot 10^4)</td>
</tr>
<tr>
<td>t-stat</td>
<td>18.62</td>
<td>71.67</td>
<td>25.25</td>
<td>64.47</td>
<td>70.13</td>
<td>23.23</td>
</tr>
<tr>
<td>(R^2)</td>
<td>June - 0.979</td>
<td></td>
<td></td>
<td>December - 0.955</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 4.4: Parameters of the hourly load stochastic part modeling

<table>
<thead>
<tr>
<th></th>
<th>( c_{h,1.6} )</th>
<th>( \delta_{h,1.6,1} )</th>
<th>( \delta_{h,1.6,2} )</th>
<th>( \sigma_{h,1.6} )</th>
<th>( c_{h,1.12} )</th>
<th>( \delta_{h,1.12,1} )</th>
<th>( \delta_{h,1.12,2} )</th>
<th>( \sigma_{h,1.12} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>0.052</td>
<td>-0.24</td>
<td>560.9</td>
<td>-0.18</td>
<td>1.18</td>
<td>-0.37</td>
<td>810.1</td>
<td></td>
</tr>
<tr>
<td>t-stat</td>
<td>0.0045</td>
<td>54.32</td>
<td>12.15</td>
<td>0.0096</td>
<td>54.75</td>
<td>17.20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(R^2)</td>
<td>June - 0.779</td>
<td></td>
<td></td>
<td>December - 0.780</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Chapter 5

Results

In this chapter we will present the evaluation of the performance of the model in the calibrating sample (2006-2010) and out of sample for the year 2011. In addition, the results for an extended out of sample performance will be shown and in particular for the years 2010-2011. Finally, an attempt to forecast the future hourly load of Germany for the period 2012-2014 will be shown in the last section.

For the evaluation of the model we used the following indicators which are better explained in Equation (5.1):

- MAE - mean absolute error
- MAPE - mean absolute percentage error
- RMSE - root mean square error
- $R^2$ - coefficient of determination

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |\hat{y}_i - y_i| = \frac{1}{n} \sum_{i=1}^{n} |\epsilon_i|$$
$$MAPE = \frac{100}{n} \sum_{i=1}^{n} \left| \frac{\hat{y}_i - y_i}{y_i} \right| = \frac{100}{n} \sum_{i=1}^{n} \left| \frac{\epsilon_i}{y_i} \right|$$
$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \epsilon_i^2}$$

$$R^2 = 1 - \frac{Var(\hat{y} - y)}{Var(y)} = 1 - \frac{Var(\epsilon)}{Var(y)},$$

where $Var(y) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2$

where ‘barred’ variables represent the mean value, ‘hatted’ ones represent the modeled or forecasted value of a quantity, $\epsilon$ is the error of the model and $Var(y)$ is the variance of $y$. 

45
5.1 In sample results

Figures (4.23) and (4.24), in the previous chapter, show schematically the result of the application of the calibrated model against the calibrating data set. It is clear that our model fits well the hourly load data for Germany.

In Table (5.1), the evaluation of the in-sample model is shown. The results are promising since the MAPE is low and the MAE is in the order of 1 GW while the average load for Germany throughout the whole sample is 55.4 GW. Finally, the high value of $R^2$ shows that the model performed well and captured most of the driving factors of load demand.

<table>
<thead>
<tr>
<th>In-Sample Performance (complete model)</th>
<th>2006-2010</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAE</td>
<td>1.09 GW</td>
</tr>
<tr>
<td>MAPE</td>
<td>2.05%</td>
</tr>
<tr>
<td>RMSE</td>
<td>1.54 GW</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.975</td>
</tr>
</tbody>
</table>

However, for the construction of this model we used all available data. Even the values of the previous days and hours, for the stochastic part of the daily and hourly part of the load, were used. Table (5.2) shows the evaluation of the deterministic part of the model alone. It is obvious that the change is not dramatic and that the deterministic part models the German hourly load quite well. Finally, it should be noted that the run time for the in sample model is relatively low, in the order of ten seconds. Similar computational times also arise in the out of sample model construction.

<table>
<thead>
<tr>
<th>In-Sample Performance (deterministic part)</th>
<th>2006-2010</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAE</td>
<td>1.32 GW</td>
</tr>
<tr>
<td>MAPE</td>
<td>2.44%</td>
</tr>
<tr>
<td>RMSE</td>
<td>1.79 GW</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.948</td>
</tr>
</tbody>
</table>

5.2 Out of sample results

As discussed earlier, the period 2006-2010 was used for the calibration of the model, i.e. for the selection of the parameters. In this section, the results of the model outside its selected sample will be presented.
Firstly, the model will be tested for the year 2011 assuming perfect knowledge of the future, i.e. actual temperatures and industrial production values, in order to evaluate the performance of the model alone. Then, we will use forecasts and assumptions for these factors so that we can test the model on quasi-real forecast effort. In order to better test the validity of our model, we will use a shorter period for the calibration of the model (2006-2009) and a longer for the evaluation by performing a two-year out of sample test for 2010-2011 and analyze the results.

5.2.1 Out of sample 2011

In this section we will evaluate the performance of the model calibrated as per Chapter 4 against actual load data for a one-year period, 2011, out of the sample used for calibration. As mentioned earlier, at first we assumed perfect knowledge of the “future” in order to test the model performance regardless of the success of the weather forecast and economic indicator’s estimation. Therefore, actual values of temperature and industrial production were used. However, we also tested the performance of the model without using this knowledge, assuming historical average daily temperatures for Germany and a linearly increasing industrial production. The slope of the industrial production increase was selected by the data used for the 2006-2010 period.

Figures (5.1) and (5.2) show the overview and a closer look of the performance of the deterministic part of the model. One can easily observe that the fit is relatively good in terms of capturing the yearly seasonality and the daily load level and shape for weekdays and weekends. In addition, in the second figure, the Easter of 2011 is depicted, along with the preceding and succeeding weeks, and it can be seen that our model has successfully predicted the fact that these days are holidays. However, it is important to note that every year Easter falls on different dates, therefore the weather and daylight hours might be significantly different from year to year. This explains why the load level is not 100% accurate and the load shape a little bit different. Finally, we do not show here the plots for the case without knowledge of the future because they are almost identical.

Tables (5.3) and (5.4) show the evaluation of the out of sample performance of our model. It is important to notice that by removing the “perfect knowledge of the future” the results are not significantly effected. Therefore the prediction was fairly accurate.

Simulation paths

Figure (5.3) shows a close look of the results for 2011 including the 95% confidence interval, the actual load, the deterministic forecast and a simulation including the stochastic and error terms. The deterministic forecast of
CHAPTER 5. RESULTS

48

Figure 5.1: Out of sample load forecast for 2011 (yearly scale)

Figure 5.2: Out of sample load forecast for 2011 (weekly scale)

this figure did not include perfect knowledge of temperature and industrial indicator. The sample paths are generated with use of the stochastic parameters calculated in Equations (4.18d) and (4.18g) with the inclusion of the simulated error terms and the 95% confidence interval curves are created from the deterministic one with the inclusion of ± 2 standard deviations for the stochastic part of the model, including the daily and the hourly parts

\( \sigma = \sqrt{\sigma_{2d}^2 + \sigma_{2h}^2} \).

In order to further validate our model, we examined the similarity between the empirical probability distributions. Figure (5.4) shows the em-
CHAPTER 5. RESULTS

Table 5.3: Evaluation of the out of sample performance of the model for 2011 assuming perfect future knowledge

<table>
<thead>
<tr>
<th>Out-of-Sample Performance (perfect future knowledge)</th>
<th>2011</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAE</td>
<td>2.07 GW</td>
</tr>
<tr>
<td>MAPE</td>
<td>4.00%</td>
</tr>
<tr>
<td>RMSE</td>
<td>2.77 GW</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.945</td>
</tr>
</tbody>
</table>

Table 5.4: Evaluation of the out of sample performance of the model for 2011 without future knowledge

<table>
<thead>
<tr>
<th>Out-of-Sample Performance (no future knowledge)</th>
<th>2011</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAE</td>
<td>2.23 GW</td>
</tr>
<tr>
<td>MAPE</td>
<td>4.35%</td>
</tr>
<tr>
<td>RMSE</td>
<td>2.99 GW</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.935</td>
</tr>
</tbody>
</table>

Figure 5.3: Out of sample load forecast and simulation for 2011 (weekly scale)

Empirical probability density function plot, on top, and empirical cumulative density function plot, below. The red lines correspond to the actual load data of our sample and the blue lines to 10 simulations. The statistical distributions of the simulations and the actual data are very much alike which is a further validation of our model. There are some load ranges were we slightly over-predict the actual output and others where we under-predict. However, this difference is always in the range of 2-3% and it could be ex-
plained by the difference in the actual industrial production indicator and our assumption.

Figure 5.4: EPDF and ECDF of the actual load data of 10 simulations for 2011

Finally, in terms of quantitative measures, Table 5.5 shows the comparison of the central moments of the distribution of the actual data and of 10 simulations. It is clear that the simulations are approximating the nature of the actual load demand appropriately. There is a slight overestimation of 2.7%, as seen in the mean value, but the standard deviation, skewness and kurtosis show that these distributions have very similar characteristics.

Table 5.5: Central moments of the actual and out of sample simulated paths for 2011

<table>
<thead>
<tr>
<th>Central Moments</th>
<th>Mean (GW)</th>
<th>std (GW)</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual</td>
<td>55.96</td>
<td>10.02</td>
<td>-0.567</td>
<td>1.85</td>
</tr>
<tr>
<td>10 Simulations</td>
<td>57.44</td>
<td>9.61</td>
<td>-0.735</td>
<td>1.95</td>
</tr>
</tbody>
</table>

5.2.2 Out of sample 2010-2011

In order to test the validity of our model for longer time periods we decided to do an out of sample test for the period 2010-2011. For this prediction we had to perform an in sample calibration for the period 2006-2009. We used the same approach as for the previous one.

In this section as well, we performed the test two times, one assuming perfect knowledge of the “future” and one without. For the second case we
CHAPTER 5. RESULTS

proceeded as in the previous section; we used historical daily temperature data and a linearly increasing industrial production factor. Figures (5.5) and (5.6) show schematically the results to this approach. The results are promising even though there is a slight under-prediction of the load especially in the year 2010.

Figure 5.5: Out of sample load forecast for 2010-2011 (yearly scale)

Figure 5.6: Out of sample load forecast for 2010-2011 (weekly scale)

In Tables (5.6) and (5.7) the evaluation of the out of sample performance of our model is shown. Again here, the difference between the two approaches are not significant and this evaluation shows that our model can stand out of sample for a further period than just one year.
Table 5.6: Evaluation of the out of sample performance of the model for 2010-2011 assuming perfect future knowledge

<table>
<thead>
<tr>
<th>Out-of-Sample Performance (perfect future knowledge)</th>
<th>2010-2011</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAE</td>
<td>2.01 GW</td>
</tr>
<tr>
<td>MAPE</td>
<td>3.79%</td>
</tr>
<tr>
<td>RMSE</td>
<td>2.66 GW</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.925</td>
</tr>
</tbody>
</table>

Table 5.7: Evaluation of the out of sample performance of the model for 2011 without future knowledge

<table>
<thead>
<tr>
<th>Out-of-Sample Performance (no future knowledge)</th>
<th>2010-2011</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAE</td>
<td>2.13 GW</td>
</tr>
<tr>
<td>MAPE</td>
<td>3.92%</td>
</tr>
<tr>
<td>RMSE</td>
<td>2.80 GW</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.926</td>
</tr>
</tbody>
</table>

**Simulation paths**

Figure (5.7) shows a close view of the result of 10 simulations of the German hourly load for the period 2010-2011. These simulations include the stochastic part with the error term as we did in the 2011 prediction in the previous section. In addition, the confidence interval of two standard deviations of the stochastic part is included - which corresponds to a 95% confidence interval. It is clear that all simulations are almost in all time-steps within these limits. Another interesting observation is that the maximum difference of this range is less than 10 GW.

We examined in this section also the similarity between the empirical probability distributions. Figure (5.8) shows the empirical probability density function plot, on top, and empirical cumulative density function plot, below. The red lines correspond to the actual load data of our sample and the blue lines to 10 simulations. The statistical distributions of the simulations and the actual data are, also in this case, very much alike which is a further validation of our model, even though there is a small scale (in the range of 2-3%) over- and under-prediction in certain ranges of load demand.

Finally, concerning the quantitative measures for this out of sample results, Table (5.8) shows the comparison of the central moments of the distribution of the actual data and of 10 simulations for the period 2010-11. In this case as well the simulations are approximating the nature of the actual load demand appropriately. There is a slight overestimation also in this case in the range of 3%, as seen in the mean value, and again the standard
deviation and kurtosis are almost identical, while the skewness is a little bit more different but still showing that both distribution have a similar shape. Therefore, the simulations offer a good estimation of the characteristics of load demand.
Table 5.8: Central moments of the actual and out of sample simulated paths for 2010-11

<table>
<thead>
<tr>
<th>Central Moments</th>
<th>Mean (GW)</th>
<th>std (GW)</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual</td>
<td>55.59</td>
<td>9.67</td>
<td>-0.0347</td>
<td>1.92</td>
</tr>
<tr>
<td>10 Simulations</td>
<td>57.34</td>
<td>9.70</td>
<td>-0.0686</td>
<td>1.96</td>
</tr>
</tbody>
</table>

5.3 Future forecast

After calibrating the model and succeeding in its out of sample evaluation assessment, we performed a future forecast for the period of 2012-2014. For the temperature forecast we used the historical daily average values of the last 40 years and for the industrial production we assumed a linear increasing trend.

Figures (5.9) and (5.10) show the overview of the resulting forecast in a yearly and a daily scale. The second figure also includes a sample path, calculated using the stochastic terms with the inclusion of the error terms. It depicts the time around Easter of 2013 and we can see that there is a difference in the load level and shape during these holidays.

Figure 5.9: Future load forecast for 2012-2014 (yearly scale)
Figure 5.10: Future load forecast for 2012-2014 (daily scale)
Chapter 6

Conclusions

In this chapter we will briefly discuss the main results of our work, including its main contribution. Then, some potential applications of the presented methodology will be proposed. Finally, some important remarks on the findings and potential improvements of the model will be made.

Summary

In this master thesis report we presented the formulation, results and evaluation of a long term hourly load model for Germany for pricing of forward contracts. At first, we decomposed the model into a daily and hourly part. Then, we modeled the deterministic and stochastic part separately for each one of the parts by calibrating the model parameters once in the 2006-2010 and once in the 2006-2009 period. For the deterministic part, linear regression was used and for the stochastic part autoregressive models of up to three steps. After obtaining the values of the model parameters we performed two out of sample tests, one for 2011 and one for 2010-2011 by using the parameters from the calibrating periods respectively. The main results concerning the performance of the model are shown in the previous chapter. These results are promising and show that even a simple approach to a complex issue can have an accurate performance. Finally, using the parameters obtained from the first calibration we performed a forecast for the hourly load of Germany for the next three years (2012-2014).

Contribution

This work’s main contribution is a proposed procedure for the formulation of long term load demand forecast models. According to the reviewed literature by the author on the topic of load forecast models, there were many detailed ones using various methodologies but most were focusing on short-term predictions [1, 5, 6, 7, 8, 9]. Goal of these was usually the forecast of day ahead up to a week ahead spot prices. Other models that had a more
medium to long term approach focused mainly on the load level forecast rather than a detailed hourly model [3, 10, 11, 12, 13, 14]. Another part of the reviewed literature contained only a forecast for working days excluding weekends and holidays [15, 16]. Therefore, we believe that this work constitutes a contribution to the research of load models in the effort to add a missing piece in the existing load model spectrum.

This master thesis offers a long term hourly load model for Germany which can be extended for different areas. The main idea of this work was to better understand the drivers of load demand and its daily, weekly and yearly seasonality. Even though load demand is a stochastic time series, there is a clear deterministic pattern which can predict the actual load very accurately. This is because people will usually behave in a similar way in two almost identical days.

**Application**

The proposed detailed hourly load model for Germany can have several applications, technical and economic. At first, it could be a useful tool in order to obtain an overall picture of the demand and consumption in Germany. In addition, one could monitor the evolution of the peak load through the years in order to better understand, in combination with the eventual decommissioning of several plants, how much installed capacity should be added in the years to come. Therefore, it is a useful tool for the monitor and understanding of the evolution of the power system within Germany, and also concerning potential needs of electricity imports, when studied in combination with an electricity supply forecast.

In regard to the market applications of the model, as it can be derived by the title of this work, the most important application of this model is for the pricing of forward curves. With the use of fundamental modeling and with the inclusion of a dispatch and fuel price forecasts for the determination of the merit order curve, one can predict the hourly price of electricity in the future. Our model can therefore contribute to the formulation of a long term Hourly Price Forward Curve (HPFC). We are confident that the model can be useful for prediction of hourly load up to several years ahead and perform relatively accurately.

Furthermore, it should be noted that there is no limitation for the use of this model even for short-term forecasts. Thus, it can also be used in the short term and perform forecast of hourly load demand for Germany for several days or weeks ahead. Hence, our model could be used also for the spot price forecast.

Another interesting observation that our model can offer is the prediction of days with very low or very high load. The former are candidates for very low (even negative) in case there is a high wind and solar energy infeed and the latter for very high prices respectively. Therefore, from a market point
of view, it is crucial to know beforehand which days can offer a higher risk for market players long time before they arrive.

Finally, this model can be applied in different countries or areas other than Germany. A little fine tuning by the user would be necessary of course, mainly concerning the economic trends and temperature relation with the load demand. Availability of data is a crucial factor as well. All in all, we are confident that this model will have equivalently good results for other areas as well.

Discussion

Concerning the overall evaluation of the proposed model, we are confident that it can produce a fairly accurate forecast for the hourly electricity load for Germany. However, there were some assumptions that need to be discussed in order to be able to extend the model for different countries and larger or smaller areas.

One of the most important issues was to take into account the economic trend. As we saw in the previous sections, the year 2009 was not a usual one for Germany in terms of load demand. There was a significant drop in comparison with the previous and following years. At first we tried to model this particular issue with the inclusion of the GDP in the model. However, a large part of the GDP (roughly 70%) contains services and the primary sector [29], which were not affected by the crisis immediately. In addition, the industrial sector of Germany, which constitutes roughly 29% of the German GDP [29], corresponds to almost half of the electricity consumption (41%, 2009) [30]. A change in industrial production cannot be directly included in the GDP, since it is an economic indicator which does not correspond to immediate changes. For example, if a company that produces automobiles stops production for a few months it means that it stops consuming big amount of energy but it continues to sell from its existing stock. Therefore, the GDP will not be highly affected immediately from this change. However, the industrial production indicator can immediately incorporate these changes. In addition, the economic crisis affects the consumption habits of other users, such as services, transportation, commercial users, households etc., in longer term than it does for the industry. Thus, the industrial production is a more representative indicator for the economic trends affects on electricity demand load for Germany.

It should be hereby noted that this indicator is representative and can incorporate the major economic influence on load demand for Germany which is a large industrial country. In order to construct a model for a different country or area where the industrial sector is not such a big driver of load demand, then, different indicators would be more suitable. Therefore, it would be preferable to first study and identify the main drivers of electricity load of the area in question and then select the ideal indicator that follows
CHAPTER 6. CONCLUSIONS

the changes in load as closely as possible.

Another important issue is the use of the temperature as a regressor. By separating the calibration for each month, for the daily load level but also for the hourly load pattern, we have already included in our model the yearly seasonality of the temperature. However, it should be noted that temperature is not uniform across wide areas. In addition, load demand is also not uniform. Therefore, for wide areas it would be more accurate to use a weighted temperature vector depending on the consumption of each area. Another solution would be to separate the model in different areas, in case of course this will significantly improve the results. In our case, the average temperature across the whole of Germany was good enough. Wider areas, though, or areas where there is a high deviation between minimum, maximum and average temperature across the country with a generally colder and a warmer part, such as Italy or France, might need one of the mentioned approaches in order to achieve better accuracy.

Furthermore, the temperature’s relation with the load was fairly linear in the particular case of Germany. However, in countries such as Spain or Greece high temperatures during the summer lead to the use of air conditioning for cooling which is a high domestic load. This relation, then, will not be linear anymore and it should be corrected in order to include this non-linearity. This adds weight to the conclusion that some tuning is necessary for our model, when one wants to apply it in an area other than Germany.

Outlook

While observing the load characteristics for Germany during vacation and public holiday periods we came across a factor that influences significantly the load demand. People’s behaviour and decision on when to take a leave from their work influences significantly the overall load demand. In particular, public holidays that occur on Tuesdays or Thursdays create an incentive for workers to take Monday or Friday respectively as days off. In addition, during Christmas time we noticed a high variation from year to year in the load level of days between Christmas day and New Year’s eve. The reason for this is that the day of the week that these two holidays occur influences the decision of people to take a full week off or not. Therefore, it is necessary to study thoroughly the decision making for this period in order to be able to incorporate it more accurately in the model, since people will generally react in the same way in two years that have Christmas and New Year’s day on the same week day.

Furthermore, even though the results of the model are promising, several minor issues can lead to further improvements, and especially during holidays when our prediction’s precision and the corresponding parameters’ significance is somewhat reduced. An inclusion of a separate indicator for
Christmas and Easter could produce a significant increase in the prediction performance for these days. It is also evident that the load during Easter is similar to that of other public holidays. However, Easter holiday is a moving holiday and might occur from late March to late April. The weather can vary significantly between those dates, as well as the sunlight hours, and thus the load level. Therefore, it would be necessary to include more years in the calibrating sample in order to be able to predict more accurately the load level of Easter depending on the date of occurrence. In addition, we believe that grouping separately the holidays of the year depending on the season they occur, can increase slightly the accuracy of the results for these specific days.

It could be proposed as well, in order to achieve further improvements in the detail of this model, to include the influence of humidity, wind speed and sunshine in the load demand. First, it is necessary to study the influence depending on the climate of the region in question and the combined influence of temperature and humidity. If a clear relationship is identified then this weather factors could also be included in the load model.

Finally, as an extension to this work, it would be useful to perform several different scenarios for the future and evaluate them. Given the fact that we performed a detailed forecast for the period 2012-2014, with some arbitrary but realistic assumptions, different ones could be introduced, concerning the temperature, the economic trends and even the inclusion of a population growth factor. Then, the study of such different forecasts could be useful for the evaluation of the uncertainty of the prediction or for its verification in case it is performed in time periods for which we have load data.
Bibliography


