Modelling and Forecasting Electricity Load in Secondary Substations

Emma Sjöborg

Department of Mathematical Statistics
Faculty of Engineering at Lund University

January 2017

LUND
UNIVERSITY

Abstract

In the energy sector a transition towards *smart grid* is now taking place as a step towards a sustainable energy distribution. In addition to many other solutions, this transition will depend upon extended measurements and data management to increase the knowledge about load flows (i.e electricity use) in the network. This thesis will concentrate on data management and statistical analysis of measurements based on a smart grid project in Hyllie - Malmö's largest development area with extensive environmental goals. The purpose of this thesis is to develop models for describing and forecasting the load in secondary substations as accurate as possible.

In order to fulfill this purpose, measurements have been collected from secondary substations in Hyllie. For comparison, data is also collected from Figeholm, where measuring has been going on for a longer time. Together with weather data from SMHI, the load data has been used to create a statistical model - a generalized additive model (GAM). GAM is a type of regression model describing the load (active power) in the secondary substation based on a number of explanatory variables. These parameters are mainly weather variables, such as temperature and wind speed, and calendar variables; as time of day, day of week and time of year. The models also take into account the load and temperature of the days before, by including the lagged values as explanatory parameters. The data from Figeholm has been used for better detection of the annual pattern, since this data covers a whole year.

All of the models show a significant relation between the load and the time of the day as well as day of week. For the stations in Figeholm a distinct annual pattern is also visible. This is not as pronounced for Hyllie, due to the shorter measure period. Furthermore, the lagged values also seem to have influence on the load. Considering the weather dependence, all stations show a significant relation between the load and the temperature. For the Hyllie stations, there also exist a relation between the load and the wind speed and global radiation. For all station in Hyllie, the same model, using the same explanatory variables has been used. This shows good model flexibility, as the load profiles of the Hyllie stations differs a lot between the stations. The models have also been tested for prediction of the load one day ahead with relatively good results. Lastly, this thesis will discuss the problems with load modelling and prediction, and how it can be improved with more information and longer measure periods.

Keywords: Load Modelling, Load Forecasting, Generalized Additive Models (GAM) Semi-parametric Regression, Smart Grid, Load Profiles, Secondary Substation.

Sammanfattning

Som ett steg mot en hållbar utveckling talas det inom energisektorn just nu om en övergång till smarta elnät - *smart grid*. Övergången till smart grid kommer, bland många andra åtgärder, kräva en utökad mätning och mätdatahantering för att öka kunskapen om hur lasterna (d.v.s. effektuttaget) i nätet egentligen ser ut. Detta examensarbete har fokuserat på mätdatahantering och statistikbehandling med utgångspunkt i ett smart grid projekt i Hyllie - ett nybyggnadsområde i Malmö som har som mål att bli Öresundsregionens klimatsmartaste stadsdel. Det övergripande målet är att ta fram modeller som kan beskriva samt prediktera lasten i nätstationer på bästa möjliga sätt.

För att uppnå målet har mätdata samlats in från nätstationer i Hyllieområdet samt, för jämförelse även från två nätstationer i Figeholm, där mätning pågått under en längre tid. Dessa data har sedan, tillsammans med väderdata från SMHI, används för att bygga upp en statistisk modell – en så kallad generaliserad additiv modell (GAM), för hur lasten i nätstationen ser ut. Detta är en typ av regressionsmodell som beskriver lasten (i form av aktiv effekt) i nätstationen utifrån ett antal förklarande parametrar. De förklarande parametrarna är främst väderdata; såsom utomhustemperatur och vindhastighet, och kalenderparametrar; vilken tid på dygnet, vilken veckodag och vilken tid på året det är. Modellerna tar även hänsyn till hur lasten och temperaturen har sett ut dagarna innan genom att ta med de eftersläpande värdena som förklarande variabler. Figeholmsdata har använts för att tydligare kunna påvisa säsongsvariationen, då mätdata från Figeholm täcker ett helt år.

Samtliga modeller visar att det finns ett signifikant samband mellan lasten i nätstationen och tid på dygnet och veckodag. Figeholmsdatan visar även på relation mellan tid på år och lasten, medan detta inte är lika tydligt i Hylliedatan på grund av den kortare mättiden. Även hur lasten såg ut 24 timmar innan verkar ha en inverkan. Alla modeller visar även signifikanta samband mellan lasten och temperaturen. Hylliedata påvisar även samband mellan lasten och vindhastighet och globalinstrålning. Samma modell, med samma förklarande parametrar, har använts för samtliga Hylliestationer och visar på en god flexibilitet i modellen, då lastprofilerna i de olika Hylliestationerna skiljer sig väldigt mycket åt. Modellerna har även använts för att prediktera lasten en dag i förväg med relativt goda resultat. Examensarbetet diskuterar även svårigheterna med att prediktera laster, samt hur predikteringar skulle kunna förbättras med mer information och längre mätperioder.

Acknowledgements

This thesis was written during the fall of 2016 at the Centre of Mathematical Sciences at Lund University, in cooperation with the Asset Development team at E.ON Elnät, Malmö.

I would first like to express my sincere gratitude towards my supervisors at E.ON Elnät, Ola Ivarsson and Ingmar Leisse, for giving me the opportunity to write this thesis and their engagement in the project. I am very grateful for their good advice and patient support during this whole process. Further, I would like to express my appreciation to all other co-workers at the office department at E.ON Elnät. I would also like to thank my academic supervisor Erik Lindström for his valuable suggestions and inputs throughout this project.

I would also like to direct a great thank you to the employees at Metrum and Netcontrol, for their cooperation and interest in this project.

Finally, I would like to address a tremendous thanks to all my friends and my family for their great support throughout my entire studies at Lund Technical University. I would like to dedicate a special thank you to Julia Ripa and Josef Kruber, for always believing in me.

Lund, December 2016
Emma Sjöborg

Frequently used expressions

Secondary Substation	A part of the electrical power system, where the power is transferred from higher distribution voltages as e.g. 10 kV or 20 kV to the voltage at the customer side, i.e. 0.4 kV.
Distribution System Operator (DSO)	A company that owns and operates the medium and low voltage power grid in a specific area. The DSO is responsible for transportation and distribution of the power to the customers in this area, but the company does not sell or produce any power.

Contents

Chapter 1

Introduction

"Those who do not remember the past are condemned to repeat it"

– George Santayana

This chapter will introduce the background of the thesis and description of the problem which the work is framed upon. Description of the purpose, questions at issue and delimitations will also be provided.

1.1 Background

Sweden is conducting an ambitious climate politic in the move toward a sustainable society. This requires changes is many different parts of the society. Today the dependence of a continuous energy distribution is very important. At the same time it contributes to a greater environmental impact. A more sustainable use of energy will require a transition to renewable energy sources and optimization of the grid. Increasingly dependence on renewable energy sources, more electric vehicles and micro production will require a network that can manage the irregular and small-scale production at the same time as it enables for a more controlled and adjustable consumption. This type of energy distribution and consumption is often summarized as *smart grids*, and will probably require some reconstruction of the network, but primarily it will depend upon increased integration of technical solutions, data management and IT-structure.

The transitions to smart grids will take time and require a lot of investments and inventions from the companies operating the electricity networks, also called distribution system operators (DSO). To manage a more uncertain production as well as enable for the consumer to influence their consumption, more information and automation is required. A better picture of the load flow in the networks and transformations is needed as well as information about when, why and where problems occur. These changes rely on more measurements of higher frequency and quality as well as better data management and analysis. Part of the solution is to develop more accurate load profiles for the consumption and better models for load prediction. This thesis will discuss methods for extracting more information from the data by statistical treatments and analysis. Accordingly a generalized additive regression model will be introduced, to model the timestamped data of the load in the secondary substations, together with weather data from SMHI.

1.2 Subject Foundation

The work has been carried out in collaboration with the asset development team at E.ON Elnät in Malmö. E.ON Elnät is one of Sweden's largest distribution system operator. The company is working with solutions for future sustainable energy distribution in many different projects, one of which is a smart grid project in Hyllie. This project includes 15 secondary substations. 14 stations are equipped with advanced measuring instrument from a company called Metrum and one has a combined measure and automation instrument from another company - Netcontrol. Having advanced measurements in the secondary substations is something rather new in the business and thus it still remains to evaluate the opportunities this can bring to the company. The software included in the measuring instruments provides information about power quality, events and interruptions, but does not keep any statistics of the measured parameters. This thesis is part of the smart grid project in Hyllie as it will discuss statistical analysis of the data received from these instruments, and examine to which extend this data can be used for better load profiles and prediction of the load.

1.3 Purpose

The main purpose of this thesis is to derive a regression model that can describe the load profiles in secondary substations as accurate as possible, depending on both calendar effects and weather conditions. The model should also have the capability to forecast the loads in the substations. To evaluate the models predictions will be made and goodness of fit will be tested.

1.4 Delimitations

To be able to carry through the thesis in plausible time, some delimitation had to be made. Firstly just one type of model is used - a generalized additive model. The models will use the load from the day before as a describing parameter, which means that the predictions can only been carried out on a one day ahead basis. (However it is possible to use the predicted values as inputs). The predictions has been made using know weather data, rather than projections, which is not the case if the models were to be used in practice and may affect the quality of the predictions.

1.5 Methodology

To collect the load data needed for the analysis, software from Metrum and Netcontrol has been used to assemble power parameters. Data from SMHI has been used to provide weather parameters on temperature, wind and sun. The software programs Python and R has been used to perform analysis of the data. Python has been used for management of the large datasets and basic statistical analysis as correlation between parameters, distribution and mean loads over different time periods. Further analysis has been carried out in R on the partly processed data. Studies has been made in the subject of load modelling and thereafter

a generalized additive model has been chosen, as it has the ability to capture the relation between the load and many different explaining variables and it provides a lot of flexibility. The model has been fitted to the datasets belonging to the different substations. The used parameters for the load is P (active power in kW) together with weather-, calendar- and lagged load parameters, discussed in Section 5.2. The models have been evaluated both depending on the goodness of fit and the prediction performance.

1.6 Outline

The outline for this thesis is:

Chapter 2: This chapter explains how the power system and distribution of electricity works today. Further some theory of the future network and smart grids is given, along with problems that may arise and possibilities to solve them. Information about E.ON and how load forecasting is used today is also provided.

Chapter 3: All statistical theory is given in this chapter along with description of the model that has been used - a generalized additive model (GAM) together with methods for model fitting.

Chapter 4: This chapter provides a description of the data that have been used in this thesis.

Chapter 5: In this chapter, justification of the model design and the used model parameters is given. The two models used for Figeholm and the model used for Hyllie is also presented.

Chapter 6: The results from fitting the models and using them for predictions is provided in this chapter.

Chapter 7: In this chapter, the results from the model fit and predictions is discussed. A comparison of the different models and how well they perform for the different stations is given. The chapter also review some problems with accurate load prediction and how the models and predictions can be improved in the future.

Chapter 8: This chapter provides a short summary of the thesis together with some conclusions of the results.

Chapter 2

Theory - Power Systems

"We believe that electricity exists, because the electric company keeps sending us bills for it, but we cannot figure out how it travels inside wires."

– Dave Barry

This chapter will provide information about E.ON Elnät and the theory needed for understanding how the power system works and why extended measuring and data analysis might be needed in the future. First the electric power system is explained, followed by a section describing how distribution system operators calculate the load when planning a network. The section also includes information about regulations for a distribution company. Lastly the concept of future networks, smart grids and the challenges it entails is described.

2.1 Electric power systems and energy distribution

Due to the lack of cheap storage possibilities when it comes to electricity, all electricity must be produces in the same instant as it is consumed and be transported to the locations where it is needed. Therefore, the production must correspond to the consumption at all times. The primary function of the power grid is consequently the transportation of electricity from the production to the end user. Further, the transportation must keep the losses within reasonable limits. Motivated by the fact that electrical power depends both on the voltage and the current, the power system is divided in different voltage levels. Assuming that only active power P_{load} is transferred and the power line is purely resistive, P_{load} can be calculated as $P_{load} = U \cdot I$, where U is the voltage level in the grid and I is the current. Therefore, to maintain a constant power in the grid, higher voltage requires less current and low voltage requires high current. Transmitting electricity will always cause some losses due to the resistance in the cable. The transmission losses is given by $P_{losses} = R_{cable} \cdot I^2$, where R_{cable} is the resistance in the cable and I still denotes the current. As both the transmitted power and the resistance can be seen as constant, the current will be the factor causing most losses. Thus the losses will be smaller in a high voltage grid, were the same amount of power can be transmitted at a lower current level. However, high voltage requires better insulation and larger components, which is more expensive. According to this fact, high voltage is used for long distance transportation, and low voltage is used for shorter transportation

near the end user. Most power systems is divided in three-phase for power transportation and consequentially requires three parallel conductors in the power line. This is due to cost efficiency, as a three-phase systems use less conductor material to transport the same amount of power than an equivalent single-phase line would. (Grauers, 2000)

2.1.1 The Swedish power system

Sweden needs a very large network to be able to secure a reliable energy distribution, ensure the balance between production and consumption and equalize the variation in energy demand. It also enables production in places where it is most cost efficient, reliable and have least impact on the environment. Most of the energy is produced in the north of Sweden, where large hydropower plants are installed, and most consumption takes place in the southern parts of Sweden. Most of the power production in Sweden derives from hydropower and nuclear power, but an increasing amount of wind power is used as well as biofuel for electricity and heat production. A map of the Swedish power system can be seen in Figure 2.1. (Grauers, 2000)

FIGURE 2.1: Map of the electricity grid in Sweden, from (Svenska Kraftnät, 2016b).

The Swedish grid can be divided in three levels - the local grid, the regional grid and the

national grid (220 kV and 400kV). The total length of the whole network is over 312 000 km, whereof 246 000 km is underground cable and the rest is overhead lines. In Sweden it exists around 160 different grid operators, however, the majority of the regional network is owned by three companies - E.ON Elnät, Vattenfall and Ellevio (former Fortum). Since there is only one distribution system operator (DSO) operating in an area, the companies has local monopoly in the region where they own the grid. (Svensk Energi, 2012).

The national grid is owned by the Swedish government and administered by Svenska Kraftnät (SVK) and consists of 15 000 km of power lines, 160 substations and switching stations and 17 overseas connections. The national grid can be compared to the highways in the electrical system and transport the electricity from the big power plants to the lower levels in the network. In additions to maintaining and monitoring the national grid, SVK is also responsible for securing that the production and consumption is balanced at all times. (Svenska Kraftnät, 2016a)

Most customers are connected to a local network, which is in turn connected to the regional network. The regional networks are connected to the national grid. Thereafter the regional network, with a voltage level between 40kV and 130kV, branches out to the local network. The local network is often divided in low (400/230V) and high voltage (10-20kV). 5,4 million users is connected to the low voltage side and only 7000 are connected to the high voltage side. The delivery performance in the Swedish grid is around 99,98%. (Svensk Energi, 2012)

2.1.2 Energy distribution in Sweden

The design of the Swedish power system as described in above section is based on a hierarchically energy distribution. It is built to transfer energy from the large power plants, through the national grid and further out through the regional and local network to the customers. The electricity production is heavily dependent on hydropower and nuclear power. In the last couple of years the amount of wind power is increasing in the Swedish system. Large wind power plants are normally installed in the medium network, while single windmills might be installed in the low voltage network. As more renewable energy sources are installed in the power system there is a need for some changes in the infrastructures of the Swedish electricity grid. The principle of the Swedish production and distribution can be seen in Figure 2.2. (Energymyndigheten/Swedish Energy Agency, 2015)

2.2 E.ON Elnät

E.ON Elnät is a distribution system operator (DSO) that is part of E.ON Sweden AB which in turn is part of E.ON SE (E.ON Societas Euopaea). EON Sweden AB has an annual turnover of approximately 34 billion Swedish krona, around 3300 employees and 1 million customers. (E.ON Elnät, 2016b) E.ON Elnät is responsible for the grid and the distribution of the annual 36 TWh electricity reaching around 1 million customers. These customers include over 860 000 private customers, 147 000 companies and 171 adjacent grids. The adjacent grids comprise 2000 production sites and 1000 micro production site. The total amount of annual transported electricity - around 42 TWh is transferred via 134 000 km of cable. 8000 km of this is regional networks of high voltage >24 kV, and 126 000 km is local networks of 0.4 - 24 kV. The grid also contains 685 primary substations and 44 000

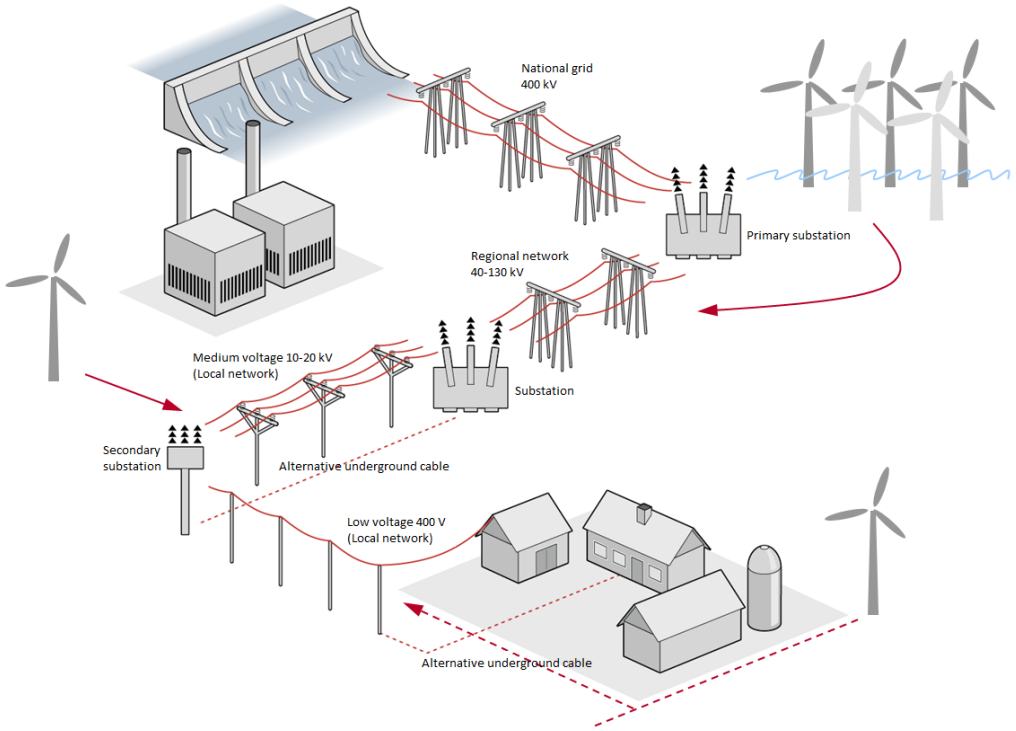


FIGURE 2.2: Explanatory picture of the Swedish energy distribution system, from (E.ON Elnät, 2016a).

secondary substations. (Ivarsson, 2016)

Ever since the large storms that affected Sweden a couple of years ago, E.ON Elnät has been working a lot with weatherproofing of the grid to decrease the downtime for the customers. The company takes a positive attitude to the transition to more renewable energy sources and local production. As a result, they invest in several development projects as local energy systems and smart grid solutions. (E.ON Sverige, 2016a)

2.3 Load forecasting

Load forecasting has been a relevant issue for many years. Even though it exists a large amount of studies on several different mathematical methods to use, few of these studies are made by DSO's and therefore few of the research presents actual benefits for the industry. (Hong and Fan, 2016)

Because the load varies over the day and over the year, the load in the network will vary a lot during its lifetime, and the network has to be able to tolerate all combinations of loads that could be expected from the consumers connected to the grid. When planning the grid, one is planning for the grid to tolerate the highest expected load - peak load, and the lowest, seeing that operation of the grid will also be possible for all other loads. Low loads are normally not a problem for the system and thus the peak load of the whole system is the most critical occasion. However low loads is a possible critical situation in a grid with a lot of installed micro production. The forecasting model also has to take into account that it is not very likely that all customers reach peak load at the same. This often dealt with in

something called coincidence factor and is based in the hypothesis that the maximum load of all customers will not sum up to the maximum load in the system. (Brännlund, 2011)

2.3.1 Peak load forecasting with Velander's formula

To calculate the maximum demand for a consumer, Velander's formula is widely used in Scandinavia (Dickert and Schegner, 2010). The formula approximate the peak load based on the consumers annual electricity usage W. Velander's formula assumes that the load for each customer is normally distributed and that the loads from different customers in a network are independent of each other. The customers are divided into categories and the load from customers in one category is assumed to be more or less similar. Velander formula can be written as:

$$\hat{P}_{i,max} = k_1 \cdot W_i + \sqrt{k_2^2 \cdot W_i}, \tag{2.1}$$

where $\hat{P}_{i,max}$ is the expected annual peak load for customer i and W_i the annual energy consumption. The linear term describes the effect of average loads and the square-root term accounts for the individual variations. k_1 and k_2 is the Velander constants that will vary depending on the category of the customer. These constant are statistically derived by regression for different types of consumption. In practice most DSO's use constants defined by the Swedish industry association - Svensk Energi.

If a network consists of n customers of the same category, then their coincident peak load can be calculated as:

$$\hat{P}_{max}(n) = n \cdot k_1 \cdot W + \sqrt{k_2^2 \cdot n \cdot W}. \tag{2.2}$$

(Dickert and Schegner, 2010)

Most networks consist of several different types of customers belonging to different kinds of categories. When calculating the total aggregated annual maximum load in the system the following formula is used:

$$\hat{P}_{tot,max} = \sum_{j=1}^{N} k_{1,j} W_j + \sqrt{\sum_{j=1}^{N} k_{2,j}^2 W_j}, \tag{2.3}$$

where $\hat{P}$ describes the annual peak load for the whole system, consisting of N different load categories. W_j is the annual energy consumption for the consumer load category j, and $k_{1,j}$ and $k_{2,j}$ is the Velander constants for category j.

The formula does not account for the variability during day, but different Velander constants can be used for different periods of the year to represent the season variability. When aggregating peak loads of groups where the peak occurs at different time of the day/week/year, the result of calculations can be a bit misleading and result in over dimensioning of the grid (see simplified concept in picture 2.3). (Brännlund, 2011)

2.3.2 Planning of a low voltage network

This section will describe how E.ON works with estimating the load in a local network, and is based on an interview with a network planner at E.ON Elnät. Estimation of the load is required when a new area is being built, to be able to plan how many secondary substations has to be built and what level of load flows the network has to be able to handle.

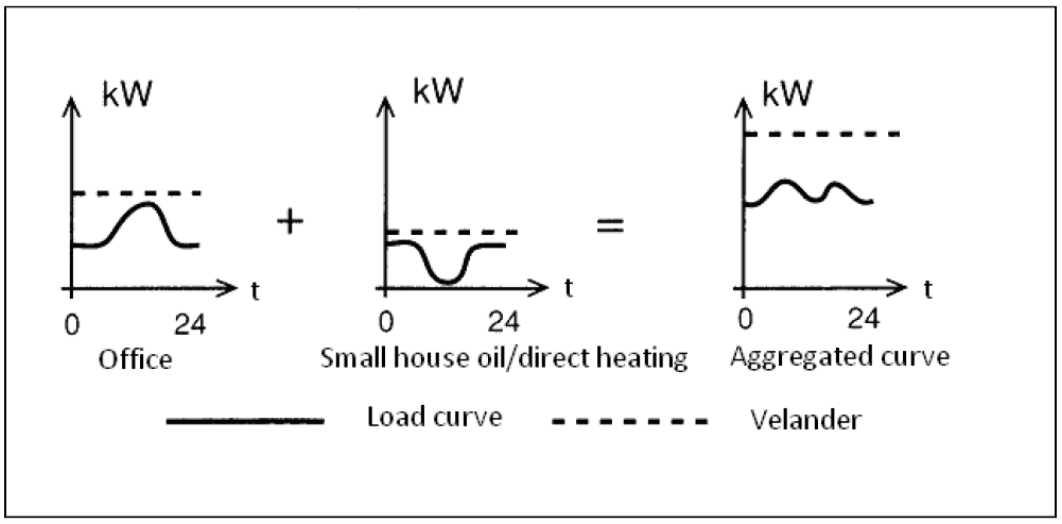


FIGURE 2.3: Comparison between actual power demand and expected demand using Velander formula, from (Svenska Elverksföreningen, 1991).

When planning for a new local network E.ON uses guidelines proposed by the industry organization (now called Svensk Energi). The theory behind those guidelines is stated in the book *Dimensionering av jordkabelnät* and *DAL-rapporten*. These ground rules is collected in E.ON's technical guidelines, and states the recommended load for planning local networks. The given load values in these reports is the aggregated coincidence load in the secondary substation. The reports states template for energy usage for different kind of buildings such as school, apartments, houses, stores, industries, etc. (It also takes into account if the buildings is heated by electric power or not.) These templates are based on several calculations and state how much power for example an apartment uses in kW. This value is based on the average value of power during peak load.

When planning a new grid, the network planner first takes a look at the zoning of the new area and then uses the fact about what type of building is to be made, together with the templates, to calculate the expected maximum load that the secondary substation has to tolerate. Later on, when they start to draw cables to the different buildings, the company pretty much has to give the customer what they ordered for the building. Since the planner of the building also wants to be sure to never get overloads in the network, they probably do some over dimensions, just to be sure.

If the new area is built quite slowly it gives the company time to get operating experience, before the whole area is expanded. This can give clues on the expected load – if there is still marginal for more capacity in the network or if it is close to overloads. However, when the whole area is built at the same time, there are no possibilities for this, and they have to rely on the first calculations.

If a secondary substation and/or the grid in an area has to be renewed or replaced, the company can look at old data for estimating the expected load. Historically, Velander's formula is used for these calculations. As explained earlier in Section 2.3.1, this formula is based on the category of the customers, rather than measurements of the actual load. However, there is a possibility to tweak these constants a little bit based on the measured load.

There might be a need for the industry to change the recommendations and recalculate the constants, but for a grid operator like E.ON, these marginal is still on the right side, it would have been worse if the electric use was larger than given by the formula. For the company it would be expensive with too excessive over-dimensioning, since the cost for the secondary substations is not entirely covered by the connection fee and E.ON still has to transport power to be able to make money in the long run. (Gustafsson, 2016)

2.4 Regulations and requirements on power security and power quality

This thesis will concentrate on the load flows in the power grid and the importance of better knowledge of these in the future smart grid. However, a great issue for the DSO's is the power quality problems that the extreme load flows may induce. (Some of these issues is discussed further in Section 2.6.1.) Swedish Energy Markets Inspectorate (Energimarknadsinspektionen) provides a lot of requirements that should be fulfilled by the Swedish DSO's regarding power quality. Power quality includes voltage quality - concerning the way the network impact customers and customer equipment, and current quality - concerning the customers impact on the network, (Bollen et al., 2008). E.ON Elnät must follow the following regulations from the Swedish Energy Market Inspectorate:

1. EIFS 2013:1 provides regulation on voltage quality and recommendations that need to be full-filled to satisfy a distribution of electricity of good quality.
2. EIFS 2013:2 provides regulations on supervision of delivery performance and responsibility to report outages.
3. EIFS 2013:3, providing regulations on delivery performance, including the functional requirement that no customer should have an outage longer than 24 hours, along with rules for hazard analysis.

Along with these regulations E.ON Elnät also follows some Swedish standards, which of most is EU-harmonized. Unlike the above regulations, standards are not cogent. In Sweden there are no rules specifying measuring in substations yet, it does however exist in other countries and it is a possibility that it will also be a part of the Swedish electricity regulations in the future. (Ivarsson, 2016)

2.5 Energy market

The price on electricity is known to have some impact on the usage. To decrease the peak loads, different electricity prices as well as different net tariffs may have to be implemented in the future. An example of how different tariffs in daily peak and peak off periods affect the daily load profiles is shown in Figure 2.4. These results comes from a research by Goude et al (Goude et al., 2014). The introduction of smart electric devices together with different net tariffs for the electricity price may have the same effect in Sweden. In this thesis the electricity prices will not be discussed any further, but this section is provided to demonstrate the impact that different net tariffs may have in the future.

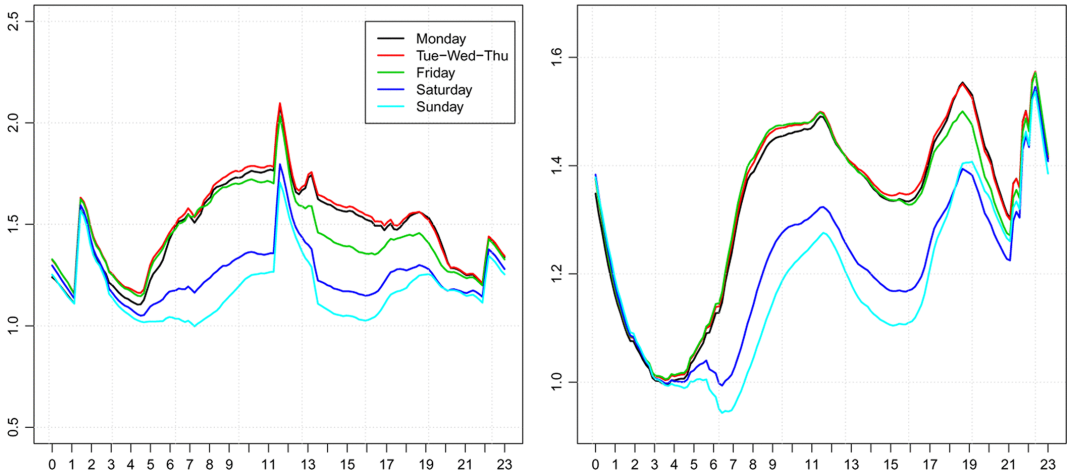


FIGURE 2.4: Figure showing the estimated load from two substations in France. The effect from different tariffs is clearly visible in the daily profiles. In order to reduce the consumption during peak hours, the residential customers may choose to subscribe to special tariffs - peak and peak off. The periods differs from different areas but usually peak off is during midday and at night time, which is visible in the figure as an increase in the load an midday and after 10 p.m., provoked by automatic tripping of devices as water heating and washing machines.

2.6 Smart grid - the future energy network

"Smart Grid is an electricity network that can cost efficiently integrate the behaviour and actions of all users connected to it – generators, consumers and those that do both – in order to ensure economically efficient, sustainable power system with low losses and high levels of quality and security of supply and safety."(European Regulators Group of Electricity & Gas, 2009)

As described in Section 2.1, the power distribution is designed for a radial operation, not considering the presence of power generation in the distribution level. In the future it is predicted that energy will be produced in a local and smaller scale, called micro production, extracted from renewable energy sources - where production depends on sun and wind. Furthermore, the customers are expected to be more engaged in their own consumptions and maybe even produce energy them self - so called prosumers. The produced energy should be possible to be sold and delivered back to the grid.

The transitions towards smart grid will require some changes in the infrastructure of the grid. Already available technology can be used for some of these changes, but it will sometimes require strengthening of the network. Strengthen the network by building new cables and substations can take several years, while changes on the customer side could happen much quicker. Since the future of the smart grid generation is still unclear it is important with a high flexibility in the network by introducing new technology. New technologies including load management and electricity storage together with control algorithms can provide this flexibility and prepare the network on large scale introduction of renewable energy production. These new technologies are often encapsulated as smart grid.

Smart grid is a broad concept covering the whole field from new technologies in transmission network, power electronics, IT solutions and understanding of load flows and opportu-

nities of load management on customer level. (Bollen, 2010)

2.6.1 Challenges and uncertainties with smart grid

As mentioned before, the power system is built for a hierarchic power flow with reliable and controllable production. Renewable energy sources will induce an intermittent energy production, as these type of energy sources is highly dependent on the weather. Consequently, it can be hard to adjust the production after energy demand. For example, the amount of radiation reaching a solar panel depends on the sun's position in the sky, which is deterministic, but also on the cloud cover and the temperature, both of which are stochastic. Thus the expected produced solar energy can be hard to calculate. The amount of consumption is also an unpredictable factor. Furthermore it is hard to predict how many customers that will install solar power or other renewables.

Not only does high integration of renewables induce a lot of uncertainty, it also affects the voltage quality in the network when new production is connected to the grid. Therefore there is a limit on how much local production that can be installed in low and medium voltage grid. If this limit is exceeded it may result in damaged equipment for the customers, which will be billed on the network operator. Thus it is of high importance for the DSO's to operate a grid that can cope with the transitions to local production and smart grid. Hence the company has to develop accurate load profiles for the substations. To make useful calculation on how the load flow will look like in a network a lot of information has to be collected about the customers, some assumptions probably has to be made, but most important a lot of data has to be collected and analysed. Calculating after "worst case scenario" is a solution giving low risks, but will also give over dimensioned estimations, entailing larger costs than necessary. Data collection and analysis will reduce the uncertainty, which can increase the capacity without increasing the risks. (Bollen, 2016)

2.6.2 Hosting capacity

A large-scale integration of renewable energy sources will have some consequences for the Swedish power system. The hosting capacity is the point where the amount of installed production exceeds the limit the network can tolerate without endangering the voltage quality or reliability of other customers. If the limit is to be exceeded some new investments is required. The hosting capacity limit is different depending on where in the grid the production is installed - it can vary from a few percent to more than 50 percent. The implementation of smart grids is supposed to increase the upper limit for the hosting capacity. The principle of hosting capacity is shown in Figure 2.5. (Bollen, 2010)

2.6.3 Measuring in secondary substation

Historically, measuring in secondary substations is not a common practice among DSO's, but it has become more common in the last years. Normally a portable measuring system is only set up if a problem is detected in the local network. A master thesis on this subject was written by Lindeberg (2009), stating that secondary substations is the most relevant and strategic place for measuring the local grid and gives better information about the network. It provides the ability to measure losses and power quality in a local network and can be

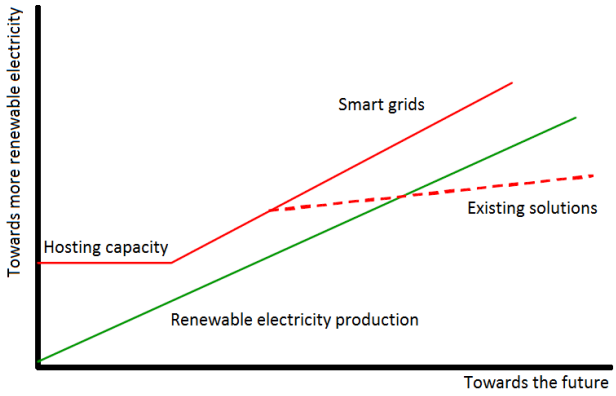


FIGURE 2.5: Graph showing the principle of hosting capacity (how much renewable energy the grid can tolerate), and how the integration of smart grid can heighten the limit for the hosting capacity. Picture borrowed from Swedish Energy Market Inspectorate (Bollen, 2010).

used to reduce interruption time for customers. Compiling and analysing of the data also facilitates decision support for the DSO. Better load profiles of the local network also gives the opportunity for the DSO's to pass on this information to prosumers and small local producers, that can use this to understand the possibility to sell electricity back to the grid.

Using so called smart secondary substations with installed automation and fault detections can reduce the outage time for the customers effectively and improve the asset life cycle management. (Kumpulainen et al., 2011)

2.6.4 E.ON's work towards smart grid

There exist several different definitions of smart grid depending on the industry describing the concept. There are a number of challenges that will affect the DSO's when it comes to smart grid. The main idea for the DSO's is to ensure the network security and guaranteeing a good power quality and a continuous supply. (European Regulators Group of Electricity & Gas, 2009)

For a regional grid operator like E.ON Elnät, where transmission operates partly in high voltage grid with long distances, smart grid can involve some sort of solutions for this type of grid. Some of these solutions could be effective regulation between wind power plants and the DSO, extended automation in switchgears, extended remote-control, intelligent data management and visualisation of the data in the operation centre and a more efficient error management and recovery of the power network.

However, the biggest part of E.ON's network includes distribution in lower voltage, where smart grids have a little different meaning. On this level automation and fault detection will also be a key in the future smart grid, but here the focus lies even more on information and control technology, distributed data management, ensuring the capacity demand of the customers and utilization of the micro production. The DSO's also has to manage an increasing

amount of prosumers in the local grids. One presumption to fully exploit the network and to provide fast information about for example capacity margins, is to develop an extended data management and data analysis with satisfactory control of the load profiles of the customer and the substations.

The predicted change of behaviour among the customers force the DSO's to get a better understanding of the actual load in the whole network and the expected development. Thus there exist an increasing need for extended measurements, data compilation and data analysing. To understand how the integration of local micro production affects the surrounding grid, it is especially important with measuring in secondary substations, along with measurements at regional and customer level. Extended measuring can yield load flow picture of the whole grid to help understand and provide information of dynamical capacity margins on a daily, weekly and yearly basis. (Ivarsson, 2016)

Since the use of reserve power often induce higher cost and more CO_2 emissions, the DSO's want to cut the power peaks in order to minimize the demand for ancillary service. This requires good understanding of when the power peaks occur, storage capacity and probably also adjustable net tariffs. With a stronger time-dependency of the electricity prices, the customers will demand more information about their own usage, to be able to influence the electricity cost. (European Regulators Group of Electricity & Gas, 2009)

2.6.5 Climate smart Hyllie

Hyllie, Malmös largest development area, will lead the way towards Malmö becoming a sustainable city. By as early as 2020, Hyllie will be 100-percent sustained by renewable or recycled energy. (Malmö Stad, 2011)

Malmö, Sweden's third largest city, has set up extensive environmental goals - 2020 the organisation of the city will be climate neutral and by 2030 the city will be supplied with 100% renewable energy. To reach this goal, Malmö's largest development area - Hyllie, will be set up as a model of a high-class sustainable city. Hyllie is supposed the lead the way for Malmö's future development as a sustainable city and be at the forefront in terms of innovation and the ability to link distribution with the usage and behaviour of the consumers. The goal for Hyllie is a 100% supply of renewable or recycled energy by 2020. Hyllie is still growing in size and fully developed it is expected to include about 9000 homes and almost as many workplaces. To ensure this development, the municipality of Malmö - City of Malmö, together with VA SYD and E.ON signed a climate contract for Hyllie in 2011. In this project VA SYD is responsible for the waste management development and E.ON is responsible for energy management development. E.ON Elnät will develop the smart grids network that will secure distribution of electricity as well as manage the local micro production. The smart grids project is supported by the Swedish Energy Agency and will include partnership with City of Malmö and other companies involved in the Hyllie project.(E.ON et al., 2013) The smart grids project is composed of several pilot projects, one of which is to develop real-time measurement, central data management and automation of the medium voltage network. This thesis is part of the data management as it will discuss the implementation of automatic analysis models treating the data from the secondary substations. (E.ON Sverige, 2016b)

FIGURE 2.6: Layout of the smart city of Hyllie, picture from (E.ON Sverige, 2016b).

Chapter 3

Theory - Model description and model selection methods

"All models are wrong, but some are useful."

– George E. P. Box

This chapter will provide all theory behind the mathematical model used for describing and forecasting the load in the secondary substation. The model used in this thesis is a generalized additive regression model (also called GAM) which is a type of a semiparametric model. A semiparametric regression model can investigate the influence from several explanatory variables on some response variable. The explanatory covariates could both have a linear and non-linear relation to the response. To understand the concept of GAM, the general understanding of linear regression models and additive models has to be introduced. The chapter is introduced with some load forecasting history using mathematical models, followed by some basic statistic theory used in this thesis.

3.1 Load modelling and forecasting based on mathematical models

Ever since the inception of the electric power industry, load forecasting has been a relevant business problem, trying to be solved both by research effort and industry practice. Often point load forecasting has been causing the greatest efforts and the forecasting generally refers to the forecasting of expected electricity demand at aggregated levels. When there is measurement data available, it is possible to make forecast based on actual observations rather than assumptions and standard templates.

A lot of studies has been made on this subject, predicting loads from customer level to international level, regarding both short term and long term forecasting. Models that are used for accurate load forecasts often regard the daily, weekly and yearly patterns as well as the impact of holidays. Weather data in terms of temperature is widely used in the models, since a clear pattern of the relation between temperature and load usually exists. Long term forecast needs to regard electricity policies, system planning and eventual climate changes,

while short term forecast concerns the daily and weekly patterns more carefully along with weather influence on a short term basis.

In a paper written by Tao Hong and Shu Fan (2016), the authors describe the advantage of probabilistic load forecasts rather than point load forecast. The probabilistic forecast may be in the form of quantiles, intervals or density functions and can be based on scenarios with assigned probabilities. The paper also describes load forecasting technologies and methodologies that has been used and classifies the technologies in statistical techniques and artificial intelligence techniques. Among the statistical approaches multiple linear regression, semiparametric additive models, autoregressive moving average (ARMA) and exponential smoothing has been used. Artificial intelligence technologies include artificial neural networks (ANN), fuzzy regression models, support vector machines and gradient boosting machines. Some of the results from earlier researches is described in Section 5.1.

Regression analysis is used to estimate the relationship between parameters and is built on statistical science. Multiple regression is used to describe the relations between a dependent variable (in this subject, usually the load) and several parameters that affects the response (for example weather parameters). Semiparametric additive models are regarded among regression models and are designed to comprehend both linear and non-linear relations to the dependent variable, within the framework of additive models. The goal for the model is to estimate the connection between the load and explanatory variables like weather and calendar parameters. A lot of theories behind these models is based on results from the book *Semiparametric Regression* (Ruppert et al., 2003), written by David Ruppert, M.P Wand and R.J Carrol.

If renewable energy sources will compose larger part of the power resources in the future, as expected, a need for using more explanatory variables will be introduced. Since sun radiation and wind speed will then have a greater impact on the load flow in the system, new models have to be developed that account for a greater dependence on the weather. The load models should also interpret for a higher integration of electric vehicles. (Hong and Fan, 2016)

In this thesis a semiparametric additive model will be used to describe the load profiles of secondary substations, regarding the relationship to both weather and calendar variables. All theory behind the model is described in Section 3 and the design of the model is described in Section 5.

3.2 Basic statistical methods

In this section some of the basic statistical models are described. For the readers familiar with statistical framework, this section can be skimmed or skipped completely.

To understand the structure of the models used to fit the data, the introduction of a stochastic variable and a stochastic process is given (see Section 3.2.1), followed by the appearance of autocorrelation in such processes (see Section 3.2.2). Furthermore the concept of degrees of freedom is explained in Section 3.2.3.

When using a model to fit a data set, there has to be some methods for evaluation if the model is a good fit for the data, i.e. describes the data satisfyingly. One of the most common methods used for model fitting is maximum likelihood estimation (see Section 3.2.4). In this

thesis regression models will be used. A common way to find the best fit for a regression model is by using least squares (see Section 3.2.5). However, when fitting more complex models, some other methods has to be used that are somewhat based on these more basic approaches. In addition, there are many different ways to express and measure the goodness of fit for these models. The methods used in this thesis will be MAPE (see Section 3.2.6), AIC (see Section 3.2.7) and R^2 (see Section 3.2.8).

These section will be followed by the introduction of linear models in Section 3.3 extended to generalized linear model in Section 3.4 (these type of models is not used in this thesis but included for completeness). Then the additive models and splines are introduced in Section 3.5 and then extended to generalized additive models in Section 3.6 and how the GAM is fitted in Section 3.7.

3.2.1 Stochastic variables and stochastic processes

A stochastic variable, is a random variable whose value is unknown and depend on a set of random events. The random variable belongs to some probability distribution. The probability distribution states the probability of occurrence of different values of the stochastic variable. A stochastic process is a collection of stochastic variables that belongs to the same state space. The variables in the process are indexed by, or depends on some set of numbers, usually seen as time. In this thesis the load in the secondary substations will be a stochastic process depending on time. (Lindgren et al., 2013)

3.2.2 Autocorrelation

Autocorrelation describes the correlation in a stochastic process, that is the correlation between the variable and it self at different points in time. It can be seen as the similarity in the data between the observations as a function of the time lags between them. The autocorrelation function can be studied to find seasonal patterns in data. As many models assumes that there is no autocorrelation in the dataset, when using these models the autocorrelation has to be accounted for. The autocorrelation between time t_1 and t_2 can be written as:

$$R(t_1, t_2) = \frac{\mathbb{E}[(X_{t_2} - \mu_{t_2})(X_{t_1} - \mu_{t_1})]}{\sigma_{t_2}\sigma_{t_1}}, \tag{3.1}$$

where X_{t_1} and X_{t_2} are the values at time t_1 and t_2. (Shumway and Stoffer, 2010)

3.2.3 Degrees of freedom

The degrees of freedom in a model fit can simply be explained as the number of values in the model that are allowed to vary. In (Walker, 1940), Walker defines the degrees of freedom as ”the number of observations minus the number of necessary relations among these observations”. In the GAM models used in this thesis the *effective degrees of freedom* - EDF will be used to measure the flexibility in the fitted model (further explained in Section 3.5.6).

3.2.4 Maximum likelihood estimation (MLE)

Maximum likelihood estimation is a method for estimating the parameters of a statistical model based on observations. The estimation is done by maximizing the likelihood of receiving those observations given the parameters.

The joint density function for independent and identically distributed observations x_i can be written as:

$$f(x_1, x_2, \ldots, x_n|\theta) = f(x_1|\theta) \times f(x_2|\theta) \times \cdots \times f(x_n|\theta).$$

Considering the observations x_i to be fixed "parameters" and θ to be the variable of a function. Now this function is called the likelihood:

$$\mathcal{L}(\theta; x_1, x_2, \ldots, x_n) = f(x_1, x_2, \ldots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta). \tag{3.2}$$

Since it is often more convenient to work with sums rather than products, usually the log-likelihood is used, which is the natural logarithm of the likelihood function:

$$\ln \mathcal{L}(\theta; x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} \ln f(x_i|\theta), \tag{3.3}$$

which gives the average log-likelihood:

$$\hat{\ell} = \frac{1}{n} \ln \mathcal{L}.$$

Finally the value of θ that maximizes $\hat{\ell}(\theta; x)$ is estimated (usually by derivation). The general definition of the maximum likelihood estimator (MLE), denoted θ_{mle} is given by:

$$\left\{\hat{\theta}_{mle}\right\} \subseteq \left\{\arg\max_{\theta \in \Theta} \hat{\ell}(\theta; x_1, x_2, \ldots, x_n)\right\}, \tag{3.4}$$

if maximum exists. (Blom et al., 2005)

3.2.5 Least squares estimation

Least square estimation, is a method for fitting point estimates of linear model parameters. The name comes from the fact that the sum of squared errors are minimized in the overall solution. Considering a data set (x_i, y_i) of n observations, where x_i is the explaining variable and y_i is the response variable. Considering a model on the form $f(x_i, \beta)$ describing the relation between the response y_i and the variable x_i, depending on some parameters β. The point estimates of β can then be obtained by minimizing the least squares:

$$S = \sum_{i=1}^{n} \left((y_i - f(x_i, \beta)\right)^2 \tag{3.5}$$

A more precis description of how this is used for linear regression models is found in Section 3.3.1. (Wood, 2006)

3.2.6 Mean absolute percentage error (MAPE)

MAPE is a measure used in model fitting and prediction methods and express the accuracy as a percentage. MAPE is usually described with the formula:

$$M = \frac{100}{n} \sum_{i=1}^{n} \left| \frac{A_i - F_i}{A_i} \right|, \tag{3.6}$$

where A_i is the actual value and F_i is the fitted value/forecast value at each time step i and n is the number of observations. MAPE is thereby a measure based only on how well the fitted values/predicted values follow the true values. It does not consider the parameters or degrees of freedom used in the model fit. Still, it is widely used in prediction analysis and thereby included in this thesis. (De Myttenaere et al., 2016)

3.2.7 Akaike information criterion (AIC)

AIC measure the goodness of fit for models fitted to some dataset and can be used for model selection. It measures both the goodness of fit for the different models as well as the complexity of the model using penalization for the numbers of parameters (or degrees of freedom). Penalizing for the number of parameters in the model is important, since usually adding more parameters will improve the model fit, but it will also make the model more complex and can result in overfitting. The AIC formula is:

$$AIC = 2k - 2\ln(\mathcal{L}), \tag{3.7}$$

where $\ln(\mathcal{L})$ is the maximized value of the log-likelihood described in Section 3.2.4, and k is the number of parameters or degrees of freedoms used in the model fit. The preferred model is the one with the minimum AIC. The theory behind the criterion requires that the data is fitted by maximum likelihood to the same data, and thus models with different transformation of the response cannot be compared directly. For the additive model used in this thesis, AIC should be calculated using the appropriate degrees of freedom that accounts for penalization. For penalized models this indicates using the effective degrees of freedom (see Section 3.5.6) rather than the number of parameters. (Wood, 2016)

3.2.8 R-squared (R^2)

R-squared (hereafter denoted R^2), also called the coefficient of determination, is a number indicating how much of the variance in the response variable that can be explained by the predicting variables. R-squared can take values from 0 to 1 and describes the square of the coefficient of multiple correlation. If R-squared has the value of 1, it means that all the variability in the data is explained by the model. Nevertheless is should be noted that high correlation does not necessarily means that the model term is significant. Also, too high values of R-squared could mean that the model overfits the data. R^2 can be written as:

$$R^2 = 1 - \frac{\sum_i^n \epsilon_i^2}{\sum_i^n (y_i - \bar{y})^2}, \tag{3.8}$$

where the numerator is the residual sum of squares and the denominator denotes the total sum of squares.

One problem with R^2 is that adding more parameters to the model will almost always increase the value of R^2, even if the parameter turns out to be of little significance. In this thesis the adjusted R^2 will be used as a measure for the goodness of fit. The adjusted R^2 is always lower than R^2 as it only increase its value if the added new predictor improves the model more than expected by chance. In other words the adjusted R^2 is modified for the number of predictors. The adjusted R^2, here denoted $\bar{R}^2$ can be written as:

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}, \tag{3.9}$$

where R^2 is the regular R-squared as in equation (3.8), n is the sample size and p is the number of explaining variables (excluding the constant term). Thus $n - 1$ is the degrees of freedom of the estimate of the dependent variable and $n - p - 1$ is the degrees of freedom for the error variance. In addition, there is a relation between the adjusted R^2 and AIC. For the normal case, using the same number of predictors, maximizing the adjusted R^2 should also minimize AIC. Again, a high value of adjusted R^2 indicates that the model explains the variability in the response well. (Wood, 2006)

3.3 Linear models

Regression is a statistical process for estimating the relationship among variables. The goal is to develop a function that fits the data in the best possible way. A regression model can be used to find a relation between some dependent (response) variable and one or more independent variables (predictors). Regression analysis includes many different techniques for modelling and analysis. The simplest models are the linear ones in which a linear relation between the response variable y and one or more predictor variables denoted x at every point (in this thesis timestamp) i is used. The definition of a linear model for a dataset $(y_i, x_{1,i}, \cdots, x_{p,i}), i = 1, \cdots, n$ is:

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_p x_{p,i} + \epsilon_i, \tag{3.10}$$

where p denotes the number of explaining parameters used. β_i is called the model parameters and needs to be estimated (β_0 is often called the intercept). ϵ denotes the error term (noise). The errors should have zero mean, be similar to white noise and be uncorrelated.

In vector form, equation (3.10) is written:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon} \tag{3.11}$$

Despite the name, linear models does not restrict the relationship between y and x to be linear. It only requires that the errors ϵ_i and the model parameters β_i enters the model in a linear way, but the predictor variables can enter the model non-linearly. Both the predictor variables and the response is allowed to enter the model as parametric transformations of the original parameter. For example, the following models are all linear and can be written on the general form in (3.10) :

1. $y_i = \beta_0 + x_i\beta_1 + \epsilon_i$
2. $y_i = \beta_0 + x_i\beta_1 + x_i^2\beta_2 + x_i^3\beta_3 + \epsilon_i$

3. $y_i = \beta_0 + x_{1,i}\beta_1 + x_{2,i}^2\beta_2 + \log(x_{1,i}x_{2,i})\beta_3 + \epsilon_i$

In linear models it is assumed that the response is continuous, and that the data can be modelled to be normal. Even though some relations between the response and the explaining variables can be estimated as linear, there are sometimes needs for more complex models describing these relations. Those models will be discussed further in Section 3.5.

The model is estimated by finding the model parameters β_i that fits the data in the best way. There exists a lot of different ways to estimate these parameters, one of the standard methods is the least square method described in Section 3.3.1. (Wood, 2006)

3.3.1 Model fitting for linear models: Least square estimation of regression parameters

When using regression models, least squares is a commonly used method to find the best fit for the data. Considering a regression model on vector form $\boldsymbol{\mu} = \boldsymbol{X\beta}$ where $\boldsymbol{\mu} = \mathbb{E}(\boldsymbol{y})$. By first recalling the link between the Euclidean length of a vector and the sum of squares of the elements of the vector. The results from least squares estimation in Section 3.2.5, can be used on linear models by minimizing the squared 2-norm between the observations and the corresponding predictions or fitted values:

$$S = \|\boldsymbol{y} - \boldsymbol{X\beta}\|_2^2 \tag{3.12}$$

Note that the resulting value will not change if $\boldsymbol{y} - \boldsymbol{X\beta}$ is rotated, since the expression directly describes the squared (Euclidean) length of the vector. This observation constitutes for the practical method of finding the model estimators $\boldsymbol{\beta}$. (Wood, 2006)

3.4 Generalized Linear Models

Often it is the case that the response variable cannot be modelled as normal, thus generalized linear models are introduced, which includes regression models that can handle non-Gaussian responses. Examples of response distributions that can be used are binary, Poisson and Gamma. The only requirement is that the distribution belongs to an exponential family. The basic structure of a GLM is:

$$g(\mu_i) = \boldsymbol{X}_i\boldsymbol{\beta} \tag{3.13}$$

where

$$\mu_i \equiv \mathbb{E}(Y_i) \quad \text{and} \quad Y_i \sim \text{some exponential family distribution.}$$

g is a smooth monotonic so called *link function*, $\boldsymbol{X}_i$ is the i^{th} row of a model matrix $\boldsymbol{X}$ and $\boldsymbol{\beta}$ is a vector of unknown parameters needed to be estimated.

Since GLM also is described in term of the linear predictors $\boldsymbol{X\beta}$, many of the concepts and result of linear modelling can be applied here too, with only a little configuration. The big difference is that a link function and distribution has to be chosen. Because the GLM is a more complex model, the model fitting will also be more complex than in simple linear models. Estimation and inference with GLM is based on maximum likelihood estimation (MLE) theory. But the maximization of the likelihood needs to be addressed with iterative least squares. (Wood, 2006) The link function makes the GLM nonlinear, however it can still

be described with a finite number of parameters and is therefore still defined as parametric. (Ruppert et al., 2003)

GLM's are not used in this thesis but is included in the theory section for a complete picture of generalized models.

3.5 Additive models

Additive models is a type of a nonparametric regression method that adds flexibility to the linear models, as the response can now depend on the prediction variables in other ways than detailed parametric relationships. A simple additive model can be written as:

$$y_i = f_1(x_{1,i}) + f_2(x_{2,i}) + \epsilon_i, \tag{3.14}$$

where f_j is unknown functions. These functions can be modelled as parametric, non-parametric or semiparametric (giving the non-parametric feature of the additive model). In this thesis the function will be so called smooth functions, modelled as regression splines. Consequently the theory behind splines and how to model them has to be introduced.

3.5.1 Splines

Normally the observations y_i is measured with noise, and thus it is generally a good idea to smooth the data rather than using interpolation (see theory behind linear interpolation in Appendix A.1). The non-parametric smooth functions f_n used in this thesis is estimated using different kinds of splines. A spline denotes a numeric function that is piecewise defined by polynomial functions. The spline order is defined by the highest order of the polynomial used. A common approach is to use polynomials of degree three, and thus these splines are called *cubic splines*. At places where the pieces connect, called knots, the function holds a high degree of smoothness and thus polynomials has to be chosen so that sufficient smoothing is guaranteed. The definition of a spline is:

$$S : [a, b] \to \mathbb{R}$$

on an interval $[a, b]$ that is made up of n subintervals $[t_{i-1}, t_i]$ where

$$a = t_0 < t_1 < \cdots < t_{n-1} < t_n = b.$$

On the interval i, S is restricted to a polynomial so that:

$$S(t) = P_i(t) \text{ for all } n \text{ subintervals}$$

where the polynomial has to fulfill:

$$P_i : [t_{i-1}, t_i] \to \mathbb{R}$$

If too many knots are used there is a risk for overfitting the data, meaning that the spline will follow small and random variations in the data as well as the main features, resulting in a noisy graph (and non-smooth). However, if the knots are too few, there is a possibility

that the main effects is not satisfactory described (over smoothed). Thus when using regression splines for smoothing, using least square estimation as in Section 3.3.1 could result in overfitting.

If all smooth where allowed when fitting a model, the maximum likelihood estimation of these models would also result in a constant overfitting of the parameter estimates. Thus the models are often fit by penalized likelihood maximization, meaning that the model likelihood is adjusted by adding a penalty for each smooth function which will penalize the "wiggliness" of the model. Thereafter the penalty is multiplied be an associated smoothing parameter to control the compromise between penalizing the wiggliness and penalizing the variance of residuals.

Denoting $h(x)$ as a natural cubic spline interpolating the points $x_i, y_i : i = 1, \cdots, n$ where $x_i < x_{i+1}$. Now rather than setting $h(x_i) = y_i$, $h(x_i)$ can be treated as n free parameters of the cubic spline. Then the estimation of these parameters is found by minimizing:

$$\sum_{i=1}^{n} \{y_i - h(x_i)\}^2 + \lambda \int h''(x)^2 dx, \tag{3.15}$$

where λ is a called the *smoothing parameter* and can be tuned to balance the two goals of matching the data and constructing a smooth. The resulting $h(x)$ is a *smoothing spline*. It can be shown that, for all functions f that are continuous on $[x_1, x_n]$ in addition to having absolute continuous first derivatives, the smoothing spline $h(x)$ is the function minimizing :

$$\sum_{i=1}^{n} \{y_i - f(x_i)\}^2 + \lambda \int f''(x)^2 dx. \tag{3.16}$$

For a regression spline, expression (3.16) can be written as:

$$\|\boldsymbol{y} - \boldsymbol{X\beta}\|^2 + \lambda \int_0^1 [f''(x)]^2 dx \tag{3.17}$$

If $\lambda = 0$, the regression spline will be un-penalized and thus tend to overfitting, while if $\lambda \to \infty$, f will be estimated as a straight line and result in over smoothing.

Due to the fact that f is linear in the parameters β_i, the penalty term (second term in (3.17)) can be written in quadratic form:

$$\int_0^1 [f''(x)]^2 dx = \boldsymbol{\beta}^T \boldsymbol{S} \boldsymbol{\beta}$$

where $\boldsymbol{S}$ is a matrix of know coefficients. Thus the goal for the penalized regression spline fit is to minimize:

$$\|\boldsymbol{y} - \boldsymbol{X\beta}\|^2 + \lambda \boldsymbol{\beta}^T \boldsymbol{S} \boldsymbol{\beta} \tag{3.18}$$

The problem with estimating the degree of smoothness of the spline thereby comes down to estimation of λ.The two most common methods for smoothing parameter selection is generalized cross-validation (GCV) and restricted maximum likelihood (REML). Both methods proceed by optimizing a function of λ. (Wood, 2006)

3.5.2 Choosing λ by GCV

GCV is based on prediction error criteria. When estimating splines, the goal is to get the spline estimate $\hat{f}$ as close to the true function f as possible. As described in the above section the principle of smoothing estimation is about choosing λ, and thus the aim is to choose λ so that $\hat{f}$ is as close to f as possible, by minimizing:

$$M = \frac{1}{n}\sum_{i=1}^{n}(\hat{f}_i - f_i)^2$$

where $\hat{f}_i \equiv \hat{f}(x_i)$ and $f_i \equiv f(x_i)$. M cannot be calculated directly since f is unknown, however an estimate of $\mathbb{E}(M) + \sigma^2$ can be developed that is the expected squared error of a prediction. Denoting $\hat{f}^{[-i]}$ as the model fitted to all data except y_i, then the ordinary cross validation (OCV) score is defined as:

$$\mathcal{V}_o = \frac{1}{n}\sum_{i=1}^{n}(\hat{f}_i^{[-i]} - y_i)^2. \tag{3.19}$$

This is a result from fitting the model to the remaining part of the data while leaving out each y_i in turn and then calculating the squared difference between the prediction and the left out data point. Replacing $y_i = f_i + \epsilon_i$, the average of the squared differences is given by:

$$\mathcal{V}_o = \frac{1}{n}\sum_{i=1}^{n}(\hat{f}_i^{[-i]} - f_i - \epsilon_i)^2 = \frac{1}{n}\sum_{i=1}^{n}(\hat{f}_i^{[-i]} - f_i)^2 - (\hat{f}_i^{[-i]} - f_i)\epsilon_i + \epsilon_i^2$$

Since $\mathbb{E}(\epsilon_i) = 0$ and ϵ_i and $\hat{f}_i^{[-i]}$ are independent, the expected score will be:

$$\mathbb{E}(\mathcal{V}_o) = \frac{1}{n}\mathbb{E}\Big(\sum_{i=1}^{n}(\hat{f}_i^{[-i]} - f_i)^2\Big) + \sigma^2$$

For large datasets $\hat{f}_i^{[-i]} \approx \hat{f}$ and thus $\mathbb{E}(\mathcal{V}_o) \approx \mathbb{E}(M) + \sigma^2$ (equality is given when datasets are very big). OCV is thereby the procedure of choosing λ to minimize $\mathcal{V}_o$.

Calculation $\mathcal{V}_o$ as above is rather inefficient, since the model has to be fitted n times, auspiciously it can be shown that:

$$\mathcal{V}_o = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}_i)^2/(1 - A_{ii})^2,$$

where $\boldsymbol{A}$ is the influence matrix corresponding to the model. By substitution the weights $(1 - A_{ii})$ by the mean weight, given by $\mathrm{tr}(\boldsymbol{I} - \boldsymbol{A})/n$, the generalized cross validation (GCV) score is obtained:

$$\mathcal{V}_g = \frac{n\sum_{i=1}^{n}(y_i - \hat{f}_i)^2}{[\mathrm{tr}(\boldsymbol{I} - \boldsymbol{A})]^2}. \tag{3.20}$$

Compared to OCV, GCV is computationally faster and has advantage in terms of invariance. For deeper knowledge on how GCV is used for modelling splines, the reader is refereed to (Wood, 2006).

3.5.3 Choosing λ by REML

Restricted maximum likelihood is based on the theory of MLE described in Section 3.2.4, but applies the principle of MLE to the least-squares residuals. In a paper written by Wood (2011), the writer closely discuss the smoothing parameter selection, in which he suggest that the REML method both provide computational efficiency, are less likely to cycle between local minima and provides the same reliability as GCV. Accordingly, REML has been chosen for estimation of the smoothing parameters in this thesis. In this section only the principles of REML will be provided. The justification for why REML is an adequate method for estimation of λ is provided in (Wood et al., 2015) and is based on the assumption that large datasets are used.

REML is used for mixed effect models, i.e. a model with both fixed effects and random effects. Regarding a mixed model, the parameters for the model is the fixed effect components, the variance of the random effect and the variance of the error. REML provides a way to estimate the variance components. A simple description is that the effects of the fixed variables are first removed by removing the unknown mean. Compared to MLE, the REML estimation is unbiased.

When using REML to estimate the smoothing parameter of a spline the following is employed. (Compare to expression in (3.18).) If S_i is treated as precision matrices of Gaussian random effects, it can be shown that λ_i can be estimated by REML. That is, choose λ_i to minimize:

$$\mathcal{V}_r(\boldsymbol{\lambda}) = \frac{\left\|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_\lambda\right\|^2 + \hat{\boldsymbol{\beta}}_\lambda^T \boldsymbol{S}_\lambda \hat{\boldsymbol{\beta}}_\lambda^T}{2\phi} + \frac{n - M_p}{2}\log(2\pi\phi) + \frac{\log|\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{S}_\lambda| - \log|\boldsymbol{S}_\lambda|_+}{2} \tag{3.21}$$

where $|\boldsymbol{S}_\lambda|_+$ is the product of the strictly positive eigenvalues of $\boldsymbol{S}_\lambda$, M_p is the degree of rank deficiency of $\boldsymbol{S}_\lambda$ and ϕ is the scale parameter. For further details the reader is referred to Chapter 6 (6.2) in (Wood, 2006) or Section 3.2 in (Wood et al., 2015).

3.5.4 Different types of splines

Using splines to estimate the smooth functions in a GAM requires both choice of spline and choice of knots for the spline. (Ruppert et al., 2003)

The different smoothing splines that will be used in this thesis is all based on cubic splines. The cubic spline $f(t)$ is a function where the sections $[t_i, t_{i+1}]$ is made up of cubic polynomials. The sections are joined so that the whole spline is continuous up to and including the second derivative and the end knots have zero second derivative i.e. $f''(t_0) = f''(t_n) = 0$. The result will be a natural cubic spline through the values at the knots. The different kinds of cubic spline used are:

1. Penalized cubic regression spline (cr), which can be written on the form:

$$f(x) = a_j^-(x)\beta_j + a_j^+\beta_{j+1} + c_j^-(x)\delta_j + c_j^+(x)\delta_{j+1}, \text{ if } x_j \le x \le x_{j+1} \tag{3.22}$$

   where $f(x)$ denotes the spline function, $x_1, \dots x_k$ the k knots, $\beta_j = f'(x_j)$ and $\delta_j = f''(x_j)$. $a_j^-, a_j^+, c_j^-, c_j^+$ describe the basis functions. The conditions described above implies that the spline should be continuous to second derivative in each x_j and that x_1 and x_k should have zero second derivative.

2. Cyclic penalized cubic regression spline (cc) is used when it is applicable for a model smooth to be cyclic, that is having the same values in the lower and upper boundaries. This is often true for seasonal data as time of the day, time of the week and time of the year. The penalized cubic regression spline on the form (3.22) can be modified to be cyclic by adding the conditions $\beta_1 = \beta_k$ and $\delta_1 = \delta_k$.

(Wood, 2006)

3.5.5 Smooths of more than one covariate

Modelling GAM also requires choosing the way the covariates is implemented in the model. It is possible to model smooths of more than one covariate, but this requires that the representation of these is considered. The smoothing basis of the covariates does not necessarily have to be the same, but the interaction between them needs to be chosen.

1. Isotropic smoothing can be used when the covariates are naturally on similar scales (for example coordinates). Penalty of the wiggliness is then treated the same way in all directions. The fitted spline will also be invariant to rotation. See 4.1.5 in (Wood, 2006) for further details.

2. Tensor products can be used to produce smooths of one or several covariates. They are especially useful when representing covariates that have different units. This method divides the bases in marginal bases with associated model matrices and penalty matrices. Thereafter these are combined using Kronecker product to one single model matrix for the smooth, but still with one penalty matrix for each marginal basis. See 4.1.8 in (Wood, 2006) for further details.

3. Varying coefficient models can be used when there exist a factor or parametric term that affects the smooth function in different ways depending on the factor. For example when having data from different locations, the location can be the factor that requires the smooth to be estimated with different coefficients for different locations. In this thesis factors will be used to denote different days of the week, since the day of week effect will depend on what type of day it is rather than the number (1-7) of the day.

(Wood, 2006)

3.5.6 Effective degrees of freedom

In the GAM models used in this thesis the effective degrees of freedom (EDF) will be used to measure the flexibility in the fitted model. The effective degrees of freedom for a GAM is given by $\mathrm{tr}(A)$ where A is the influence matrix of a model (see Section 1.3.5 in (Wood, 2006)). The maximum value of $\mathrm{tr}(A)$ is the number of parameters minus the number of constraints. Usually the effective degrees of freedom is broken down to the effective degrees of freedom for each smooth, since the degrees of freedom of the model is reduced by the penalty of each smoothing term. (Wood, 2006)

3.5.7 Choosing knots of the spline

Choosing the knots of the spline will balance the influence of the squared bias respective to variance. The basis dimension k in practice also sets up the limit for the degrees of freedom. If k is chosen to be too small there is a risk that the main features are not captured well, but if k is too big there is a risk of computational ineffectiveness. However, in the model used in this thesis, only the maximum possible k has to be chosen. The actual effective degrees of freedom will be estimated by the degree of penalization selected during fitting, using REML. Thus the effective degrees of freedom can be much less than the chosen k.

After fitting a GAM model the number of knots for each spline should be checked so that there is no risk for using too few knots and thus force over smoothing. This could be done by checking that the effective degrees of freedom for each smooth in the model fit is not too close to the upper limit of k for the same smooth. (Wood, 2016)

3.6 Generalized additive models

A generalized additive model (hereafter denoted GAM) follows from the additive models described in the previous section. It also has the relation to a generalized linear model (GLM) where the linear predictor involves a sum of smooth functions (specified by the user) of the covariates and a regular parametric component of the linear prediction. Hence the GAM includes both parametric and non-parametric model components and the model can specify the dependence of the response on the covariates in a flexible way, not depending on the presumption that all relations can be modelled as linear. The flexibility does however require a need of the theory of representing the smooth functions and the choice of how smooth they should be (see Section 3.5). The general structure of a GAM is:

$$g(\mu_i) = \boldsymbol{X}_i^*\boldsymbol{\theta} + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}, x_{4i}) + \dots \tag{3.23}$$

where

$$\mu_i \equiv \mathbb{E}(Y_i) \quad \text{and} \quad Y_i \sim \text{some exponential family distribution.}$$

A simple example of a GAM is:

$$\log(\mathbb{E}(y_i)) = f_1(x_{1i}) + f_2(x_{2i})$$

where in this case the independent response variable $y_i \sim$ Poisson and f_1 and f_2 are smooth function of the covariates x_1 and x_2. The logarithm function is an example of a link function that can be used. (Wood, 2006) The GAM used in this thesis can be written on the form:

$$y_i = f_1(x_{1,i}) + f_2(x_{2,i}) + \cdots + f_p(x_{p,i}) + \epsilon_i. \tag{3.24}$$

In all models described from earlier researches in Section 5.1, as well as models used in this thesis, the response variable y_i will describe the load and the covariates $x_{n,i}$ will describe the covariates that drive y_i, mostly calendar effects and weather parameters. ϵ_i denotes the model error at time i. The non-linear functions f_n are the smooth functions that needs to be estimated and theory on this subject was presented in Section 3.5.1.

For further understanding, let's adjust the general structure of (3.23) to:

$$g\left\{\mathbb{E}(y_i)\right\} = \boldsymbol{A}_i\boldsymbol{\theta} + \sum_j L_{ij} f_j \tag{3.25}$$

Where g is a known link function, $\boldsymbol{A}$ is now the n-row model matrix and $\boldsymbol{\theta}$ is still the parameter vector to be estimated, f_j is still the unknown smooth function of one or more variables (described in Section 3.5.1) and L_{ij} is a known linear functional. For each f_j there exist some measure of departure from the smoothness, denoted $J_j(f)$. When the L_{ij} are evaluation functional, expression (3.25) defines a GAM. The method for GAM fitting is rather extensive and is described in Section 3.7. (Wood et al., 2015)

3.6.1 Autocorrelation when using a GAM model

An important prerequisite of the GAM is that the data is correctly ordered and that there is no structure or seasonal trend in the residuals of the fit. If it exists some pattern in the residuals, this can be interpreted as the model fails to capture some important pattern in the data. The models also requires that the there is no autocorrelation in the residuals. When there is a lot of autocorrelation in the original data set, then the residuals have a high risk of being autocorrelated. Thus, the autocorrelation in the errors has to be dealt with. This can be done by adding more parameters in the fit that account for the seasonal trend in the data set. However, eliminating all autocorrelation by adding more parameters can be hard, and then an AR-process (or other timeseries processes) can be added to the model. (van Rij, 2016)

Modelling the residuals as an AR-process, can be done by letting the error term be an AR1 process. Consider the GAM model:

$$y_i = f_1(x_{1,i}) + f_2(x_{2,i}) + \cdots + f_p(x_{p,i}) + e_i$$

where the error term, e_i is modelled as an AR(1) process:

$$e_i = \rho e_{i-1} + \varepsilon_i, \tag{3.26}$$

and ε_i is independent identically distributed $N(0, \sigma^2)$ random variables. (Wood, 2006)

The value of ρ states the correlation of the residuals. This value can be found by first fitting a model ignoring the autocorrelation. ρ is thereafter estimated by calculating the correlation of the residuals for a certain lag (usually the second lag). ρ can also be estimated using an REML. The value of ρ used in this thesis will be the value of correlation at the second lag. (van Rij, 2016)

3.7 Model fitting of a GAM

The fitting of the GAM is performed by penalized iteratively re-weighted lest squares (PIRLS) (described in Section 3.7.2) and the smoothing parameter selection implemented by REML (described in Section 3.5.3). That is, λ is selected by using REML and β is estimated using PIRLS. In the model fitting process, the choice of which type of splines to use and the knots of the spline has to be decided. In addition, the prediction variables to use and how to model them (parametric or non-parametric), have to be decided, as well as the potential interaction between them. When fitting a model, the autocorrelation of the residuals also has to be considered and accounted for.

In practice the GAM model has been fitted to the data by using the function *bam* from the *mgcv*-package in R. This function uses a performance-oriented iteration for large datasets in every step of the PIRLS (described in Section 3.7.2).

3.7.1 Basis of model fitting for Gaussian case

First a description of the Gaussian identity link case will be explained as an introduction.

Considering the general structure of GAM described in equation (3.25), with a Gaussian link function g resulting in independently normally distributed y_i with variance ϕ. If every f_j is represented by a linear basis expansion (for example a spline) and for each f_j the measure departure of smoothness, J_j is quadratic in the basis coefficients, equation (3.25) can be written as:

$$\mathbb{E}(y) = \boldsymbol{X}\boldsymbol{\beta},$$

where $\boldsymbol{\beta}$ contains $\boldsymbol{\theta}$ (the parameter vector) and all the basis coefficients. The model fitting considering estimation of $\boldsymbol{\beta}$ can be done by minimizing:

$$\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \sum_j \lambda_j J_j(f_j) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \sum_j \lambda_j \boldsymbol{\beta}^T \boldsymbol{S}_j \boldsymbol{\beta} \tag{3.27}$$

where $J_j(f_j) = \boldsymbol{\beta}^T \boldsymbol{S}_j \boldsymbol{\beta}$ is used, in which $\boldsymbol{S}_j$ is a known $p \times p$ matrix. $\boldsymbol{X}$ contains all dependent variables and the basis functions, and $\boldsymbol{\beta}$ contains model parameters and the basis coefficients. Model fit according to (3.27) assumes that $n > p$, that is that the number of observations used in the model are greater than the number of smoothing bases (rank) of the model. For this thesis, this assumption is justifiable, since the models are based on tens of thousands of observations.

Estimation of λ is described in Section 3.5.2 and 3.5.3. When knowing λ, expression (3.27) can be minimized to obtain the parameters of $\boldsymbol{\beta}$. Consider the case where the model matrix $\boldsymbol{X}$ is first QR-decomposed into $\boldsymbol{Q}$ - a column orthogonal $n \times p$ part and $\boldsymbol{R}$ - an upper triangular $p \times p$ part, yielding $\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{R}$. By forming $\boldsymbol{f} = \boldsymbol{Q}^T\boldsymbol{y}$ and $\|\boldsymbol{r}\|^2 = \|\boldsymbol{y}\|^2 - \|\boldsymbol{f}\|^2$ expression (3.27) can be written as:

$$\|\boldsymbol{f} - \boldsymbol{R}\boldsymbol{\beta}\|^2 + \|\boldsymbol{r}\|^2 + \sum_j \lambda_j \boldsymbol{\beta}^T \boldsymbol{S}_j \boldsymbol{\beta}. \tag{3.28}$$

Using this result, smoothing parameter selections by REML according to (3.21) can be rewritten to:

$$\mathcal{V}_r(\boldsymbol{\lambda}) = \frac{\left\|\boldsymbol{f} - \boldsymbol{R}\hat{\boldsymbol{\beta}}_\lambda\right\|^2 + \|\boldsymbol{r}\|^2 + \hat{\boldsymbol{\beta}}_\lambda^T \boldsymbol{S}_\lambda \hat{\boldsymbol{\beta}}_\lambda^T}{2\phi} + \frac{n - M_p}{2}\log(2\pi\phi) + \frac{\log|\boldsymbol{R}^T\boldsymbol{R} + \boldsymbol{S}_\lambda| - \log|\boldsymbol{S}_\lambda|_+}{2} \tag{3.29}$$

The importance in this conclusion is that once $\boldsymbol{R}, \boldsymbol{f}$ and $\|\boldsymbol{r}\|^2$ is obtained $\boldsymbol{X}$ is no longer of importance for fitting and accordingly model fitting can be done without forming $\boldsymbol{X}$ completely, resulting in computational efficiency.

It can be shown that, using methods based on iterative updating of a QR- or Choleski-decomposition, only a small sub block of $\boldsymbol{X}$ is needed to compute $\boldsymbol{R}, \boldsymbol{f}$ and $\|\boldsymbol{r}\|^2$. For the interested reader this is further discussed in Appendix B of (Wood et al., 2015).

3.7.2 Penalized iteratively re-weighted least squares (PIRLS)

In the generalized case the model is given by:

$$g\{\mathbb{E}(y_i)\} = \boldsymbol{X}_i\boldsymbol{\beta}, \tag{3.30}$$

but the unknown function and their penalties can be found in the same way as in 3.7.1. However, in this case the model has to be estimated using a penalized maximum likelihood rather than penalized least squares. This is done in the PIRLS-algorithm:

Denote V as a function so that $\text{Var}(y_i) = \phi V(\mu_i)$ and $\mu_i = \mathbb{E}(y_i)$. Initialize $\hat{\mu}_i = y_i + \xi_i$ and $\hat{\eta}_i = g(\hat{\mu}_i)$ (where ξ_i is only added to make sure $g(\hat{\mu}_i)$ exists, and thus is often small). The following PIRLS-scheme is then iterated until convergence:

1. Set $z_i = g'(\hat{\mu}_i)(y_i - \hat{\mu}_i) + \hat{\eta}_i$ and $w_i = V(\hat{\mu}_i)^{-1/2} g'(\hat{\mu}_i)^{-1}$.
2. By putting w_i in a diagonal matrix $\boldsymbol{W}$ the weighted version of expression (3.27) can be minimized according:

$$\|\boldsymbol{W}\boldsymbol{z} - \boldsymbol{W}\boldsymbol{X}\boldsymbol{\beta}\|^2 + \sum_j \lambda_j \boldsymbol{\beta}^T \boldsymbol{S}_j \boldsymbol{\beta}$$

Using this method for large data sets requires a lot of storage and take long time since both a penalized least square problem has to be solved and optimization of model smoothing parameters should be done for each iteration step. For efficiency, performance oriented iteration with an outer optimization can be used when working with large datasets. In this method a QR-update is used on the $\boldsymbol{W}\boldsymbol{X}$ matrix rather than a QR-decomposition (described in Appendix A.2) for each step in the PIRLS.

For further details the reader is refereed to Chapter 3 and 4 in (Wood, 2006) and Section 3 and Appendix B in (Wood et al., 2015).

3.7.3 Testing for significance of model terms

After fitting a model, the significance for each smoothing term likewise the intercept has to be examined. The significance of the parametric terms can be evaluated by the p-value calculated by the Bayesian estimated covariance matrix of the estimated parameters. Significance of the smooth term can be calculated using the p-value based on test statistic in which the components are weighted by iterative fitting weights, (see (Wood, 2012) for more details). Because the p-value describes the probability that the smoothing term is equal to zero, the lower the p-value is, the more significant is the term.

Even though a term is designated with high significance, it may not have any influence in the prediction or fitting of the model, thus several aspects has to be considered in parameter selection.

3.7.4 Goodness of fit

There are a lot of different methods for evaluating the goodness of fit for a model. The models will also be evaluated on the performance of predictions. For comparing the model fit AIC, MAPE and R^2 will be used (described in Section 3.2). For comparing the predictions, the model that provides the best predictions according to MAPE and the residual diagnostic will be chosen (residual mean close to zero, the distribution close to normal, and no apparent structure in the residuals).

Chapter 4

Data description

"Errors using inadequate data are much less than those using no data at all."

– Charles Babbage

This chapter will describe the data used in this thesis. The data is collected both from internal sources and external sources. The external data is in terms of historical weather parameters as wind speed and temperature. The data of the electricity usage i.e. the load is collected from different secondary substations in Hyllie area in Malmö. The measurements are provided from measuring units from Metrum (in 14 of the stations) and Netcontrol (in one of the stations). Both types of units also measure a lot of other parameters as well as keeping track of power quality issues and other events. The unit from Netcontrol also has implemented automation that can be operated from the control room, making this station a so called smart substation.

Since Hyllie is a developing area, it is hard to say if the load is representing an "ordinary" usage or if it is dependent on periods of construction or people still moving in. Seeing that no parameters are available for describing this change, it can be hard to model the load in a satisfying way. One of the Hyllie stations in this thesis is know to have solar production. In the future it is expected to be more local production from sun and wind affecting many of the stations. In order to compare the results using the Hyllie data and in an attempt to make a better model for the load, data from two secondary substations in Figeholm has also been used. The measurements here are collected from a whole year and the usage here is presumed to be more stable. The stations in Figeholm belongs to an area including a large conference site and some smaller houses. This area is chosen since the secondary substations in this area also is equipped with measuring units. The measuring equipment used in Figeholm is also provided by Metrum.

4.1 Load data

The measuring in the Hyllie stations started at different times during the spring 2016 and extend until December 2016, meaning that the datasets consists of data from 5-8 month. The parameter used for denoting the load is active power (P), described in kW. The data from Hyllie is stored every ten minute as mean value since the last stored timestamp, resulting

in 144 timestamps per day/1008 timestamps per week. The measurements from Figeholm are stored once every minute. Since all the models used in this thesis will be based on ten minute values, the data from Figeholm will be compressed to ten minute data, simply taking the mean value over the nine preceding timestamps. The weather data used in the models are in a lower frequency (one hour basis) and will be transformed to ten minute values, using linear interpolation (see Appendix A.1). Since data is from Figeholm is collected for a whole year (October 2015 to October 2016), which happened to be a leap year, the total size of the dataset used for fitting is $144 \cdot 366 = 52704$. One station from Figeholm and three stations from Hyllie has been chosen for demonstration in this thesis. The data used from the different secondary substations is described in Table 4.1.

Station	Whole dataset	Nbr fit	Nbr Pred	Nbr of missing values
Figeholm	52704 (366 days)	52704	6192	16 [1]
Hyllie1	29664 (206 days)	22248	7416	25
Hyllie2	26496 (184 days)	19872	6624	2
Hyllie3	29952 (208 days)	22464	7488	20

TABLE 4.1: Description of the data used for fitting the models and the data used for predictions. The number of missing values is the total number for the whole dataset.

4.1.1 Missing data and data quality

Both for Hyllie and Figeholm the data have few missing values. Load data often contains outliers due to events that are unknown and hard to measure or predict. These outliers will affect the models, but is not deleted, due to the risk of loosing the information they contain. However, some extreme outliers caused by maintenance work in the secondary substation and in the overhead substation have been deleted, and is included in the number of missing values (see Table 4.1). The deleted data and the missing values have been approximated with linear interpolation (see Appendix A.1). These periods of missing or deleted data are no longer than a few hours. Due to the large size of the datasets this is not assumed to have any significant effect on the results. Since no customer data were available, it was not possible to do include any test for customer outages, which otherwise would have been applicable.

4.2 Weather data

Historical weather data is collected from an open database provided by the Swedish Meteorological and Hydrological Institute (SMHI). The weather stations are chosen according to the closest in distance; for Hyllie resulting in data from station Malmö A less than 10 kilometres east from Hyllie. In one of the models for Hyllie a parameter for global radiation has also been used. This parameter is collected from a sun station in Lund, almost 30 kilometres north-east from Hyllie. In Figeholm the temperature is measured at a station at the north end point of Öland, approximately 30 kilometres east of Figeholm. The wind speed data is received from a weather station Gladhammar A around 40 km north of Figeholm.

[1] There were 156 missing values in original dataset of 527 040 one minute values.

The reason for not using the same station here is absence of data and the hypothesis that the wind speed on the north end point on Öland will not be representable for the wind speed on the main land in Figeholm.

Most of the weather data is quality assured by SMHI and is thus regarded as trustworthy. Since the data provided by SMHI is hourly mean values, the data has been sub-sampled for the ten minute model using linear interpolation, described in Appendix A.1.

4.3 Calendar data

The GAM includes the relation to several calendar variables. These have been extracted directly from the timestamp extracted from the timeseries of the load. These include day of the year, month, week, day type and time of the day. To get the holiday dependence, public holidays have been labelled as Sunday. Furthermore bridge days and ”de facto holidays” has been labelled as Saturday.

Chapter 5

Model Design

"Statisticians, like artists, have the bad habit of falling in love with their models."

– George E. P. Box

This chapter will be introduced by some results from earlier researches using semiparametric additive models for load forecasting. Thereafter, the model design is described, starting with some analysis from plotting the data. From this analysis, some conclusion of how to design the model is drawn, following a description of the models that has been used.

Management of the data, with merging of the weather data and the data from the substations as well as some data analysis has been performed in the software program Python (Pyt, 2014). For deeper analysis of the data and all model design and analysis of results, the software program R (R Core Team, 2013) has been used. For the implementation of the models the R package mgcv (Wood, 2016) has been used.

Through the execution of this thesis, a lot of different GAM model fits has been tried out in order to find the model that describes the data in the best way. Models has been fitted for the different stations in order to find the best model for the specific stations. Since the data from Figeholm covers a whole year, the models fitted to the Figeholm-data are assumed to be more correct and hence results based on these models will be introduced first. In the sections 5.6 and 5.7 the two best models for Figeholm will be described. Further the comparison between them will be described in Section 6. The model used for the stations in Hyllie will be described in Section 5.8 and results from model fitting is presented in Section 6.2.

The GAM model has been fitted according to the performance oriented iteration version of the PIRLS (from Section 3.7, with an REML model to choose the smoothing parameter λ. After fitting the different models to data from different substations, the models has also been tested for prediction.

5.1 Results from earlier researches

The researches described in this section is mainly published from IEEE and semiparametric additive models have been successfully used in several researches analysing both short term and long term load forecasting. The models are declared to provide a good trade-off between the ability to congregate the complex relations in the data as well as an automatic estimation process, not heavily dependent on human intervention. (Goude et al., 2014).

Many of the researches focus on forecasting for a large area (typically a whole country) and uses data from several years back in the models, for example the paper by Goude et al (2014) modelling predictions for the load in 2200 substations in France, on a short and middle term perspective. The model includes data recorded every ten minute from 1900 of the substations (excluding those with series including too many outliers) for year 2006 to 2011 and weather data from 63 weather station recorded every third hour. In France they also have special tariffs for peak on and peak off periods which also are included in the model and shows interesting result in the load profiles (see Figure 2.4 in Chapter 2). As many other researches, this include calendar variables and weather parameters as explaining covariates. After all data have been transformed to ten minute data, 144 semiparametric additive models is fitted - one for each time step of the day. The writers compare the forecast model using both known weather data, and using weather forecast, which demonstrate the significant addition in model uncertainty that the uncertainty in the weather forecasts brings. They also highlight the necessity of high quality weather forecast. The writers also note the complexity in middle term load forecasting at substation level, since the load is highly affected by changes in the area, for example a mall being built. If there is no information about these changes, there is no way to predict them. For this thesis this is also a relevant issue since the substations from Hyllie covers an area with a lot of changes and a lot of new and ongoing constructions.

A paper written by Hyndman and Fan (2010) models the load in South Australia. The report focuses on peak electricity demand and long-term forecasts and thus only includes hours 12:00 to 20:30, and the months November to March, were the electricity usage reach its highest levels. This report also uses one model for each time period of the day, in this case half hourly intervals. Another interesting aspect of this report is the inclusion of demographic and economic variables for the long term predictions. Because these factor will highly influence the electricity demand growth in the future. Used parameters are for example residential population, persons per household, household sector per capita disposable income, average electricity price, etc. Since all these unknown factors are supposed to be significant, this brings a lot of uncertainty to the model. The writers solve this by assuming three different scenarios. Another issue with long-term forecast is the lack of any precise and credible weather predictions. To simulate future temperatures the writers use a double season block bootstrap that aims to capture both the daily and yearly variations in the temperature. Further, the writers choose to divide the model in two parts, one accounting for calendar and weather influence as 48 semiparametric additive models for each time stamp, and one linear regression model based on the annual effect of economic and demographic variables. The results show that the uncertainties related to the economic forecasts are very small compared to those associated with the weather forecast. Nevertheless it should be noticed that the economic relations can change over time as it is dependent on politics, new technologies and other factors.

In the above papers, the data is fitted with one model for each time step of the day. The

reason for this is stated to be a better goodness of fit and prediction accuracy. The issue with this approach is discussed in the a paper written by Wood et al (2015). Here a new fitting iteration is introduced that can effectively handle large datasets and thus only require one single model. The described approach is called PIRLS, and is explained in more detail in Section 3.7.2. The used explaining variables in this paper are also based on calendar effects and weather parameters. Using one model for the whole data set instead of dividing it in instant of the day, entails larger autocorrelation in the data. This is dealt with by using an AR(1) model on the residuals.

Inspiration has been used from some of these studies, but implemented in a way that is supposed to describe the underlying data for this thesis in the best way.

5.2 Parameters

In the models the response variable y_i will always describe the load data for the different secondary substations. The collection of describing variables used in the models differs a bit between the models, but is always some of the following:

1. *DayType* (1-7), where 1 = Monday and 7 = Sunday, where the day type is set as a factor rather than a number.
2. *Yearday* (1-365), describing day of the year.
3. *Month* (1-12), where 1=Januari, 12 = December.
4. *Timestamp* (1-144), denoting the time of the day according to ten minute intervals. 1 = 00:00 and 144 = 23:50.
5. *Temp* (°C), describing the mean temperature for the last ten minutes for each time *i*.
6. *Wind* (m/s), describing the mean wind speed for the last ten minutes for each time *i*.
7. *GB* (Watt/m^2), describing the mean global radiation for the last ten minutes for each time *i*.
8. P_{i-144} (kW), describing the lag 144 load, i.e. the load of the day before.
9. P_max_{i-144} (kW), describing the rolling max value for last day, lagged by 144 timestamps.
10. P_meanD_{i-144} (kW), describing the rolling mean value for last day, lagged by 144 timestamps.
11. P_meanW_{i-144} (kW), describing the rolling mean value for last week, lagged by 144 timestamps.
12. $Temp_{i-144}$ (°C), describing the lag 144 temperature, i.e. the temperature of the day before.
13. $Temp_{i-288}$ (°C), describing the lag 288 temperature, i.e. the temperature of two days before.

5.3 Parameter selection

Even if a model term is highly significant, it does not necessarily mean that the model produce accurate predictions, thus it is of importance to choose parameters which is both of significance and results in the best forecasts. Before choosing a model, the data is plotted in search for which parameters that will have influence on the load.

5.3.1 Figeholm data

First the load is studied according to different time horizons. In Figure 5.1 the load for the whole timeserie and the daily profiles based on data from station Figeholm is seen. The daily profiles are studied by simply grouping the data according to day of the week and then taking the mean values over the time of the day. For comparison the data has also been grouped according to the month of the year to study if there is a seasonal change in the daily profiles.

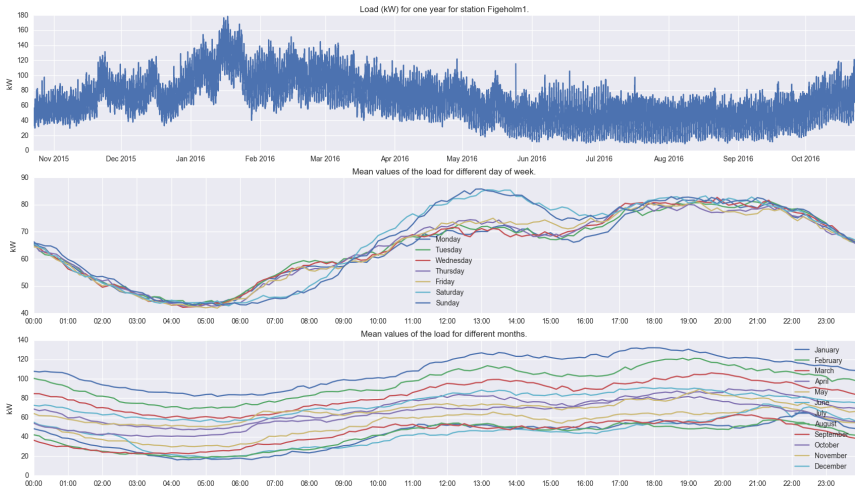


FIGURE 5.1: Load data and daily profiles based on the mean values over the whole dataset for station Figeholm. (OBS different y-axis is used for a distinguishable demonstration of the differences.)

As seen in the figures, there is a seasonal pattern of a lower usage in the summer month and a higher usage in the winter. There also exists a clear difference between the daily profiles for weekdays and weekends. Since a significant evening peak can be seen in the figures, as well as a morning peak (which occurs later in weekends), guesses can be made that the station has many resident customers connected to the substation. Furthermore, due to the relative high loads during daytime, the load is probably also influenced by a load from offices and other buildings and functions with day-time activity. Since a large conference site is connected to the station, this behavior seems rather expected. To understand more about the correlation between the parameters, some scatterplots is also studied. In Figure 5.2, the scatterplot matrix for station Figeholm is shown, with histograms for the different parameters on the diagonal.

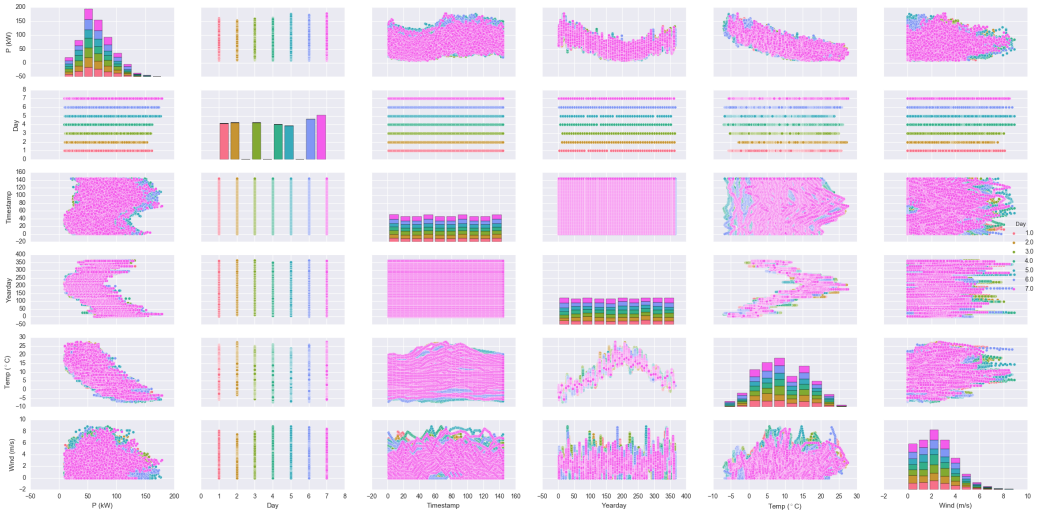


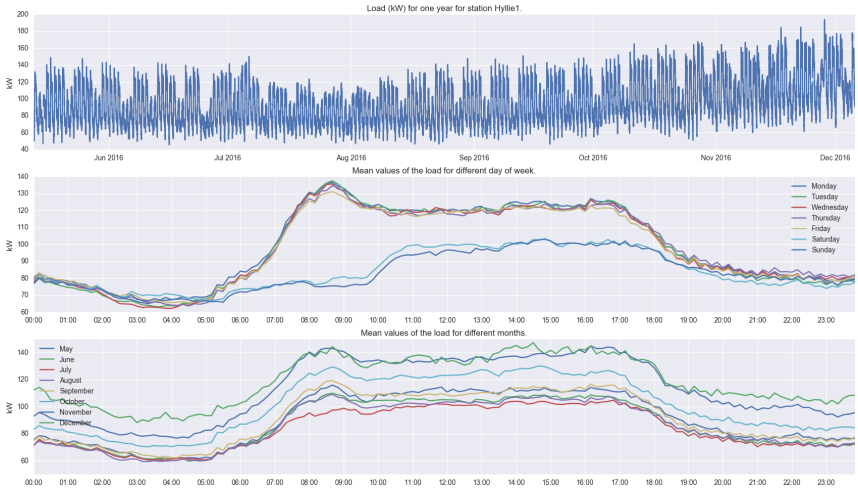
FIGURE 5.2: Scatterplot matrix for Figeholm. Data is colored according to day of the week (1-7, for Monday to Sunday). In the distribution for the day types there are more Saturdays and Sundays due to labelling of the holidays.

From the scatterplot matrix some guesses can be made on which parameters that may have influence on the load. As stated earlier the day of year and the time of the day obviously have some influence on the load. It also appears to exist a dependency between the temperature and the load too, (though this can also be due to the effect that less electricity is used in the night, when the temperature also is lower). The dependence between the wind speed and the load is hard to state in this plot, but will still be tested in the models.

5.3.2 Hyllie data

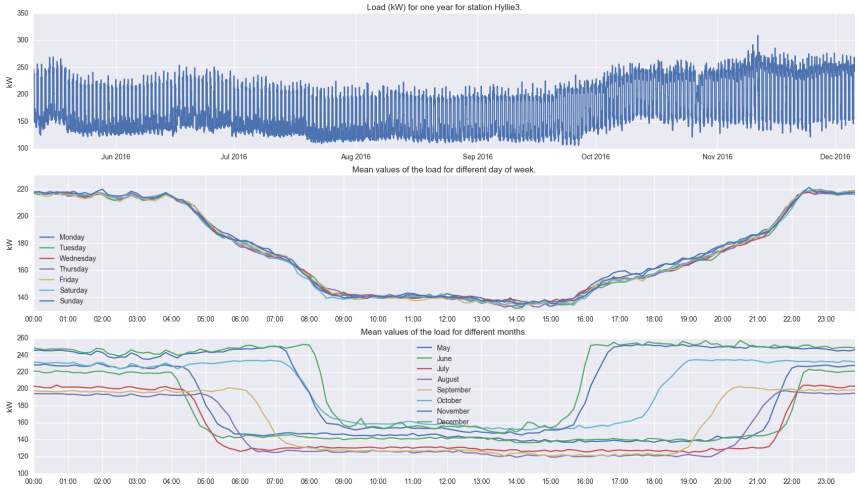
The load from the secondary substations in Hyllie varies a lot depending on the station. To make this thesis easy to grasp and to evaluate the flexibility of the model, three of the substations has been selected that all displays different types of loads. The loads and the daily profiles is shown in Figure 5.3.

From the plots it is possible to make some assumptions about the loads. Both Hyllie1 and Hyllie2 seems to be stations with high influence of offices and other operations with activity concentrated to daytime activity during weekdays. Hyllie2 is connected to a large building, almost exclusively containing offices. The station also includes production from solar power. Hyllie1 includes a car park and based on the load profiles it can be assumed that some electricity is turned on when cars arrive or leave, because of the higher load in the daytime and the morning peak. Hyllie3 shows a completely different load than the other stations. As can be seen in the lower of the plots in 5.3(c) the load has an on/off behaviour, with high load during night time and lower during day time, showing almost no difference depending on the day of the week. It is also clear that the high load period during night-time is longer during spring and autumn than in the summer month. This behaviour is supposed to be caused by illumination of a train-station that is connected to this station and is turned on and turned off at a certain time every day.



(A) Hyllie1

(B) Hyllie2



(C) Hyllie3

FIGURE 5.3: Load data and daily profiles based on mean values, for the Hyllie stations.

For the stations in Hyllie, data is also collected from SMHI describing the global radiation. This measurement is from Lund quite far from Hyllie, but in the scatterplot matrix in Figure 5.4 some dependence seems to be visible for this parameter and thus it will also be included in the models for the Hyllie stations. The scatterplot matrix will look different for the different stations, but the same conclusions on which factors that affects the load, can be drawn for most stations.

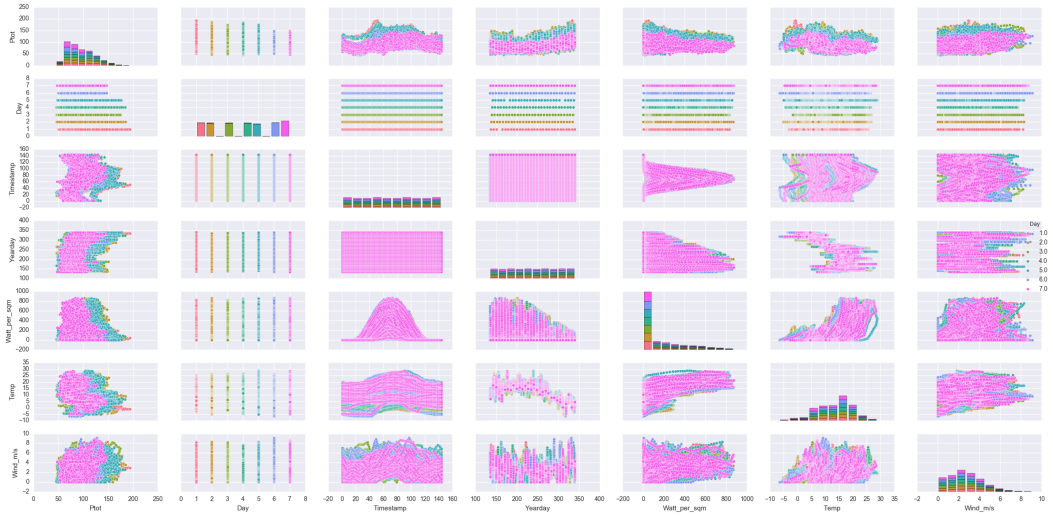


FIGURE 5.4: Scatterplot matrix for Hyllie1. Data is colored according to day of the week (1-7, for Monday to Sunday).

5.4 Distribution of the response

The GAM model can be used based on some different distribution of the response variable. Thus it has to be examined which distribution to use and if there is a need for transformations of the response. Since the time period for collection of the data, and the profiles for the load vary between the stations, the distribution of the response also varies. For most stations the distribution has some degree of skewness. For some stations the distributions seems to be a combination of two normal distributions, (probably due to the distinction between periods of low loads and high loads.) Since the data collected in Hyllie do not cover a whole year, the distribution cannot be determined as easily.

The histogram and the normal QQ-plot for the load P at the station in Figeholm is shown in Figure 5.5. Using the models require that the type of distribution is stated, but no further analysis of the parameters of the distributions are needed.

As seen in the Figure 5.5 the response variable is reasonably close to normal. The models for Figeholm will thus assume that P is normally distributed, without any transformation of the response. For simplicity, the load in Hyllie will also be treated as normal.

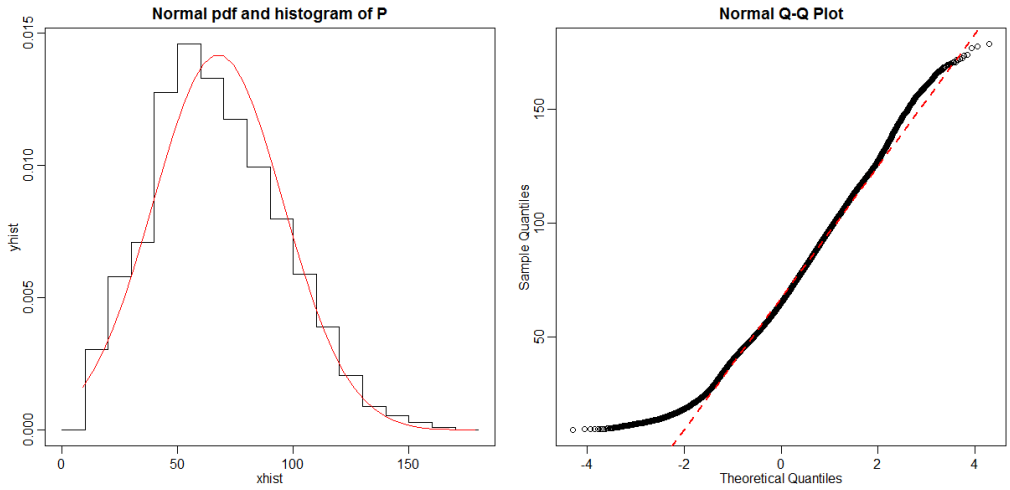


FIGURE 5.5: Histogram over P and a QQ-plot for normality for station Figeholm.

5.5 Weighting the data

Since the data (especially from Figeholm) is collected over a seemingly long time period, it might be a desideratum to weight the data so that newer data is of higher importance for the model fit than older data. This might be included in the model fit as prior weight on the contribution of the data to the log likelihood. The data models are fitted both using no prior weighting of the data (i.e a constant prior weight equal to 1) and including weights. Among the tried out weights, the best model for the weights has been shown to be:

$$W_i = \frac{n_i}{2N} \cdot \frac{P_i}{\bar{P}}, \tag{5.1}$$

where N is the length of the dataset, n_i is the number of the observation (from 1 to N) and $\bar{P}$ is the mean value of all loads P_i. For some stations, including the weights give better predictions, while for some it does not add any improvement.

5.6 Figeholm - Model 1

Since the value of the load is expected to be dependent on the value of the load from the days before, this model uses the load from one day before as an explaining variable. The dependence is however expected to vary depending on the time of the day and if the day before is a weekday or weekend. Thus this term will include the timestamp of the day TS_i to model the interaction between the load 24 hours before and the time of the day. To handle the ”day of week - effect”, seven different smooths will be used for each day of the week (modeled as a factor). The load is also expected to depend on the temperature at the same time, but also on the temperatures from the days before, which is also considered in the model. In this model the yearly seasonal trend has been modelled as terms depending on the month. The model also includes two terms describing the rolling mean and max values for the past day as well as the rolling mean value for the past week. In addition to parameters describing the seasonal pattern, the residuals will be modelled as an AR-process, given by

equation (3.26). This is necessary to accommodate for the correlation of the closest lying measurements according to timestamp. The model is fitted both with and without prior weighting of the data. When weights are used they are implemented as in equation (5.1).

Model M1 can be written:

$$\begin{aligned} P_i = {} & f_{1,D\_type(i)}(TS_i, P_{i-144}) + f_2(TS_i, M_i) \\ & + f_3(T_i, TS_i) + f_4(T_{i-144}, T_{i-288}) + f_5(P\_meanW_{i-144}) \\ & + f_{2,D\_type(i)}(P\_max_{i-144}, P\_meanD_{i-144}) + \varepsilon_i \end{aligned} \tag{5.2}$$

where

- P_i denotes the load (kW) for each ten minute interval of the day.
- $f_{1,D\_type(i)}$ is a tensor product of a cubic regression spline for P_{i-144}, denoting the load, lagged by 144 timestamps (i.e. 24 hours before), and a cyclic cubic regression spline for TS_i describing the ten minute period of the day. The smooth is estimated for each day of the week indicated by $D\_type(i)$.
- f_2 is a tensor product of two cubic regression splines for timestamp TS_i and the month of the timestamp M_i (cyclic for TS_i).
- f_3 is a tensor product of two cubic regression splines for timestamp TS_i and the temperature T_i for timestamp i. (Cyclic cubic regression spline for TS_i.)
- f_4 denotes a tensor product of two cubic regression splines for the temperature T_{i-144} lagged by 144 timestamps (one day before), and temperature T_{i-288} lagged by 288 (two days before) .
- f_5 denotes a cubic regression spline for P_meanW_{1-144} denoting the mean load for the past one week, lagged by 144 timestamps.
- $f_{2,D\_type(i)}$ is a tensor product of two cubic regression splines for P_max_{i-144}, denoting the daily maximum load lagged by 144 timestamps, and P_meanD_{i-144} denoting the daily mean load lagged by 144 timestamps. The smooth is estimated for each day of the week indicated by $D\_type(i)$.
- ε_i denotes the error term, modelled as an AR-process, where $\rho = 0.722$

5.7 Figeholm - Model 2

In this model there will be no parameters describing the rolling mean or max values. Also this model is fitted both with and without prior weighting of the data. When weights are used they are implemented as in equation (5.1).

Model M2 can be written:

$$\begin{aligned} P_i = {} & f_{1,D\_type(i)}(TS_i, P_{i-144}) + f_2(YD_i) + f_3(TS_i, YD_i) \\ & + f_4(T_i, TS_i) + f_5(T_{i-144}, T_{i-288}) + f_6(W_i) + \varepsilon_i \end{aligned} \tag{5.3}$$

where the P_i, $f_{1,D\_type(i)}$ is modelled in same way as in M1. f_4 is equal to f_3 in M1 and f_5 is equal to f_4 in M1. The other terms are:

- f_2 is a smoothing function depending on which day of the year i belongs to YD_i.
- f_3 is a tensor product of two cubic regression splines for timestamp TS_i and day of the year for the timestamp YD_i (cyclic for TS_i).
- f_6 is a smoothing spline describing the relation to the wind speed W_i at timestamp i.
- ε_i denotes the error term, modelled as an AR-process, where $\rho = 0.736$

Unlike M1, M2 includes the relation to the wind speed. The reason why this term was not included in M1 is since it showed no significance in this model.

5.8 Hyllie - Model 3

For the Hyllie stations one model seems to perform better than the others both according to MAPE and predictions. This model is referred to as model M3 and is described in equation (5.4):

$$\begin{aligned} P_i = {} & f_{1,D\_type(i)}(TS_i, P_{i-144}) + f_2(T_i, TS_i) \\ & + f_3(T_{i-144}, T_{i-288}) + f_4(W_i) + f_5(GB_i, TS_i) + \varepsilon_i \end{aligned} \tag{5.4}$$

where the P_i, $f_{1,D\_type(i)}$ and ε_i is modelled in same way as in M1 and M2. f_2 is equal to f_3 in M1 and f_3 is equal to f_4 in M1. The other terms are:

- f_4 is a smoothing spline describing the relation to the wind speed W_i, at timestamp i.
- f_5 is a tensor product of two cubic regression splines (cyclig in TS) for timestamp TS_i and the global radiation GB_i at time i.

Since the data do not cover a whole year, no parameter describing the time of the year is used in this model. Also the rolling mean and max parameter has been abolished since it did not result in better fits. Model M3 has been fitted both with and without weights. When weights are used, they have been given the following values:

$$W_i = \frac{n_i}{2N} \cdot \frac{P_i}{\bar{P}},$$

where N is the length of the dataset, n_i is the number of the observation (from 1 to N) and $\bar{P}$ is the mean value of all loads P_i.

Chapter 6

Results

"If you torture the data long enough, it will confess."

– Ronald Coase

This section will describe the results from fitting and prediction for the different models. Section 6.1 will submit the results for Figeholm and Section 6.2 will present the results for Hyllie.

6.1 Model fitting for Figeholm

This is the results for model fitting based on data from station Figeholm. M1 is the model described in Section 5.6 and M2 is the model described in Section 5.7. M1 AR and M2 AR is the same models including the AR-process. The result of the fitting for each parameter using model M1 AR is shown in Table 6.1. The p-values describe how significant the each model term is (see Section 3.7.3) and since the p-value is very small for all terms, all terms is expected to be of significance. The result for fitting M2 AR also shows high significance for all terms as can be seen in Table 6.2.

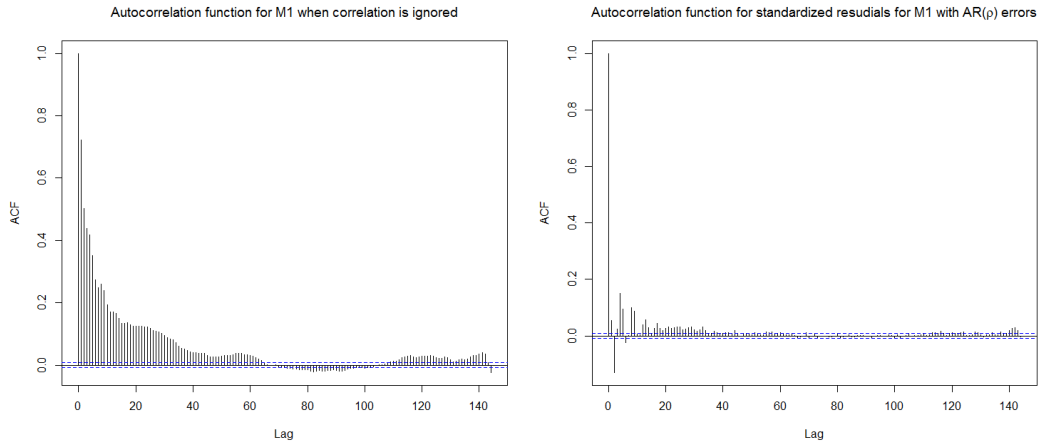
As described in the theory section, the autocorrelation of the residuals has to be considered in the model fitting process. The autocorrelation plot for the residuals are shown in Figure 6.1 together with the lag plot for the residuals. As can be seen in this figure including the AR-process for the residuals will lower the autocorrelation significantly. Those plots shows the fit for model M1 versus model M1 AR. Thus it can be stated that modelling the residuals as an AR-process shows satisfactory results in decreasing the autocorrelation, and thus the AR version of the models will be used for further results.

Parametric coefficients	**Estimate**	**Std. Error**	**t-value**	***p*-value**
(Intercept)	77.1340	2.4158	31.9286	< 0.0001
Smooth terms	**edf**	**Ref.df**	**F-value**	***p*-value**
te(TS,Ptot_144):D_type1	8.5735	12.4171	1.7496	0.0463
te(TS,Ptot_144):D_type2	1.0027	1.0053	39.9118	< 0.0001
te(TS,Ptot_144):D_type3	11.3618	17.1195	2.6914	0.0002
te(TS,Ptot_144):D_type4	11.0180	16.4530	2.7192	0.0002
te(TS,Ptot_144):D_type5	17.7058	26.0302	2.5620	< 0.0001
te(TS,Ptot_144):D_type6	14.8475	16.9824	30.1063	< 0.0001
te(TS,Ptot_144):D_type7	25.1649	32.4907	15.3276	< 0.0001
te(TS,Month)	68.2046	76.6130	19.5168	< 0.0001
te(Temp_i,TS)	52.9619	63.6820	19.0654	< 0.0001
te(Temp_144,Temp_288)	24.5526	31.0569	8.5467	< 0.0001
te(Max,Mean_i):D_type1	10.5277	12.0114	16.9334	< 0.0001
te(Max,Mean_i):D_type2	7.4035	8.7295	20.1406	< 0.0001
te(Max,Mean_i):D_type3	11.2040	12.5616	16.3641	< 0.0001
te(Max,Mean_i):D_type4	10.7841	12.3051	15.5974	< 0.0001
te(Max,Mean_i):D_type5	7.2363	8.5226	21.2342	< 0.0001
te(Max,Mean_i):D_type6	12.2583	13.8595	14.8570	< 0.0001
te(Max,Mean_i):D_type7	11.1201	12.8042	14.7746	< 0.0001
s(Mean_w)	16.5948	17.9248	13.9615	< 0.0001
Total estimated degrees of freedom:	349			

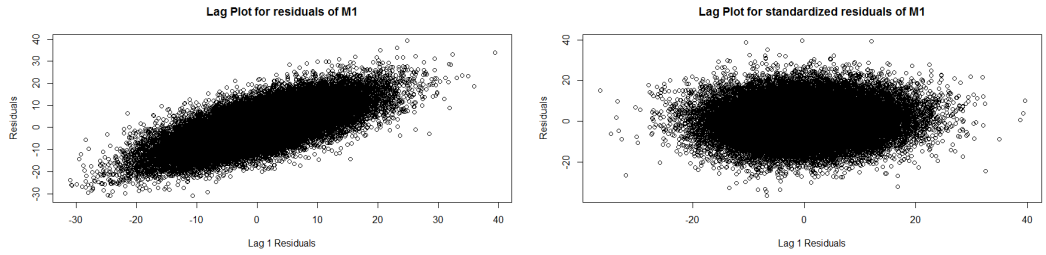
TABLE 6.1: Resulting fit for each parameter and the smoothing terms for M1 AR described in equation 5.2, based on data from Figeholm. (te describes the tensor product of two regression splines and s denotes a smoothing spline.)

Parametric coefficients	**Estimate**	**Std. Error**	**t-value**	***p*-value**
(Intercept)	68.8733	1.9622	35.1002	< 0.0001
Smooth terms	**edf**	**Ref.df**	**F-value**	***p*-value**
te(TS,Ptot_144):D_type1	10.3362	14.7109	5.6502	< 0.0001
te(TS,Ptot_144):D_type2	3.0126	4.1250	15.7613	< 0.0001
te(TS,Ptot_144):D_type3	11.4045	17.0415	7.2850	< 0.0001
te(TS,Ptot_144):D_type4	11.5644	17.1828	6.8048	< 0.0001
te(TS,Ptot_144):D_type5	20.9539	30.3160	4.4533	< 0.0001
te(TS,Ptot_144):D_type6	20.2861	25.9151	20.7664	< 0.0001
te(TS,Ptot_144):D_type7	25.2774	32.5950	16.1436	< 0.0001
s(YD)	18.5992	18.8202	38.4884	< 0.0001
te(TS,YD)	68.2795	88.0000	17.7603	< 0.0001
te(Temp_i,TS)	63.0895	82.0000	17.6467	< 0.0001
te(Temp_144,Temp_288)	25.4254	31.9943	20.6665	< 0.0001
s(Wind)	1.1630	1.3097	8.5575	0.0056
Total estimated degrees of freedom:	298			

TABLE 6.2: Resulting fit for each parameter and the smoothing terms for M2 AR described in equation 5.3, based on data from Figeholm. (te describes the tensor of product two regression splines and s denotes a smoothing spline.)



(A) Autocorrelation

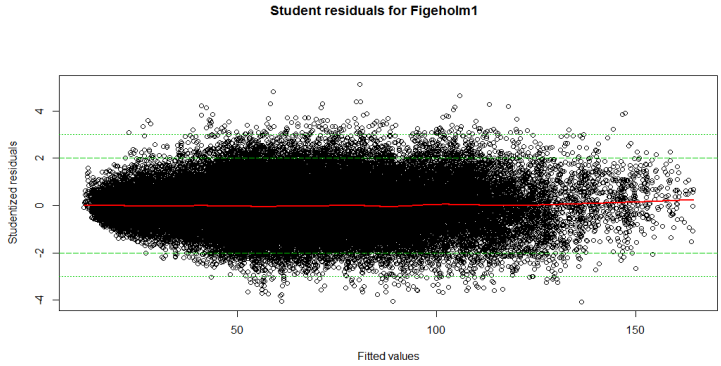


(B) Lag plot ignoring autocorrelation

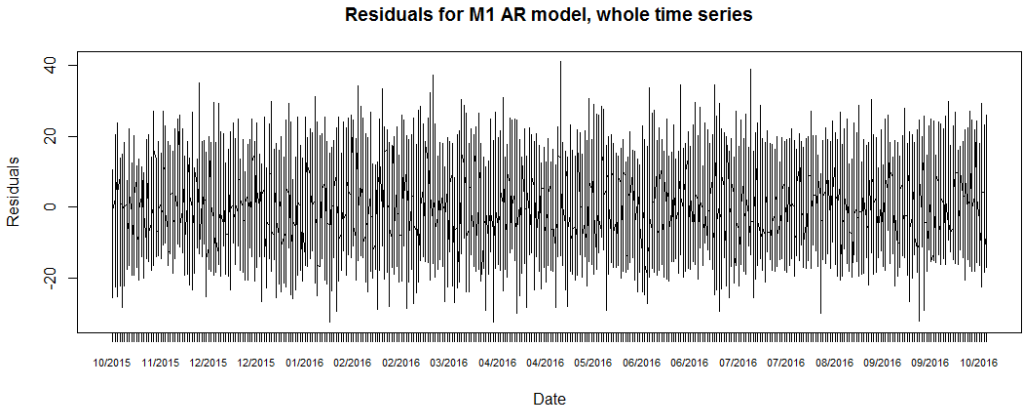
(C) Lag plot regarding autocorrelation

FIGURE 6.1: Some plots showing the difference in the correlation of the residuals when using an AR-process for the residuals. Plot (a) shows the difference in autocorrelation. Plot (b) shows the lag 1 plot of the residuals when ignoring the correlation, which is clearly visibly in this plot. Plot (c) shows the standardized residuals when correlation is modelled as an AR-process of the residuals. These residuals are supposed to be approximately uncorrelated under correct model, which they seems to be.

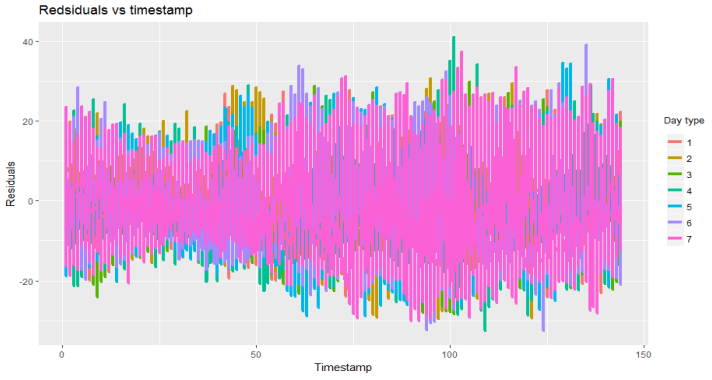
Another important aspect in model fitting is that there are no patterns among the residuals. In Figure 6.2(a) the student residuals is plotted in relation to the fitted values, for model M1 AR. The red line in the plot is the trend modelled as a spline and this line shows no clear pattern in the residuals. The conclusion from this plot is that the model does not under- or over-estimate any order of the fitted values. The dashed and dotted green lines show two standard deviations and four standard deviations respectively. This implies that there are a still some existing variance in the data that the model cannot capture. Figure 6.2(b) shows the residuals over the whole time period, and this plot indicates that it does not exist any typical pattern over the year in the residuals. The spikes in the residuals are probably due to outliers in the original data. The last plot 6.2(c) shows the residuals plotted against the time of the day, colored according to day of the week. This shows that fitting is more accurate during night time than during day time. This is most certainly due to the higher variation and more outliers in the load during daytime.



(A)



(B)



(C)

FIGURE 6.2: Plot (a) shows student residuals for M1 AR model fitted for Figeholm 1. (b) shows the residuals for the same fit for the whole dataset. (c) shows the residuals against time of the day, colored according to weekday.

Figure 6.3 shows an additional four standard diagnostic plots for the residuals typically used for evaluating the fit. A good model fit requires that the residuals are normally distributed and shows no pattern in the predicted values, as seems to be fulfilled by model M1 AR. Model M2 AR shows similar result for the student residuals and the autocorrelation as well

as the residual diagnostics.

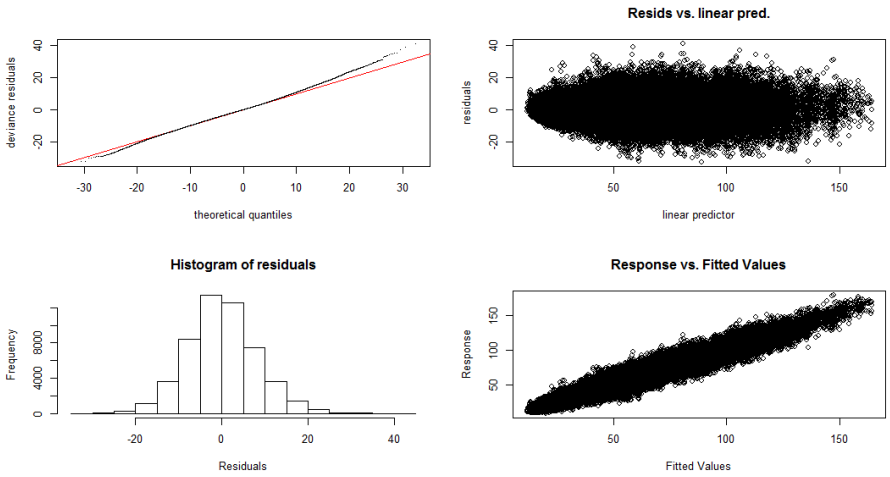


FIGURE 6.3: Diagnostic plots for the residuals for model M1 AR. The upper and lower left plot shows that the residuals are close to a normal distribution. The upper right plot is the raw residuals vs the linear predictor, which shows no pronounced pattern. The lower right plot shows the response vs the fitted values and this plot shows no sign of under or over estimation in some interval.

In Table 6.3 the results are shown both including the AR term and fitting ignoring the autocorrelation. The errors are smaller for the model ignoring the autocorrelation, as can be seen in the MAPE values, however the values of the adjusted R^2 gets better. In addition the effective degrees of freedom for the AR models is around 50% of the EDF for the models ignoring autocorrelation, showing the practical importance of using the AR-process. In the table R-sq(adj) describe the adjusted R-squared.

Model	MAPE (%)	AIC	R-sq(adj)
M1, $\rho = 0$	10.67	363348	0.93
M1, $\rho = 0.722$	11.13	323115	0.92
M2, $\rho = 0$	10.94	367041	0.92
M2, $\rho = 0.736$	11.16	322985	0.92

TABLE 6.3: Indications of goodness of fit for Figeholm.

6.1.1 Relation between the load and the prediction variables

To understand how the load is affected by the different explaining variables i.e. the relation between the response and the smooth terms, some plotting can be made. In Figure 6.4 the plots for model M1 AR is shown. In these plots all other parameters than the ones showed on the x-axis and y-axis are kept fixed.

To get a more interpretable result for the daily profiles, these are plotted in 2 dimensions, shown in Figure 6.5.

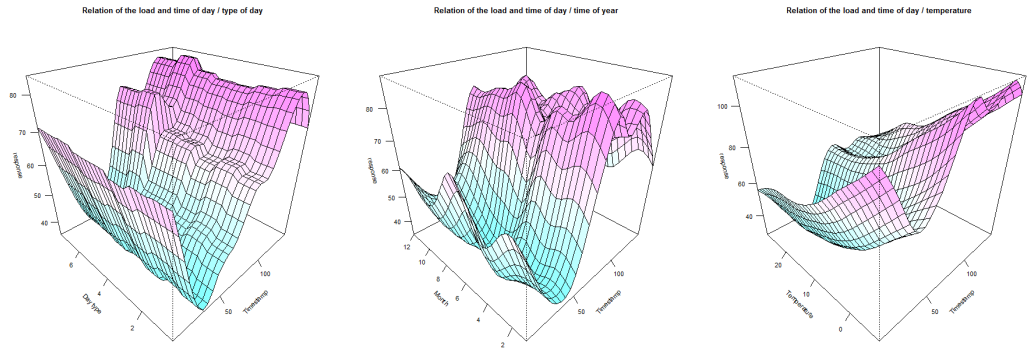


FIGURE 6.4: Relation between the response variable and some explaining variables for model M1 AR, fitted on the data from Figeholm.

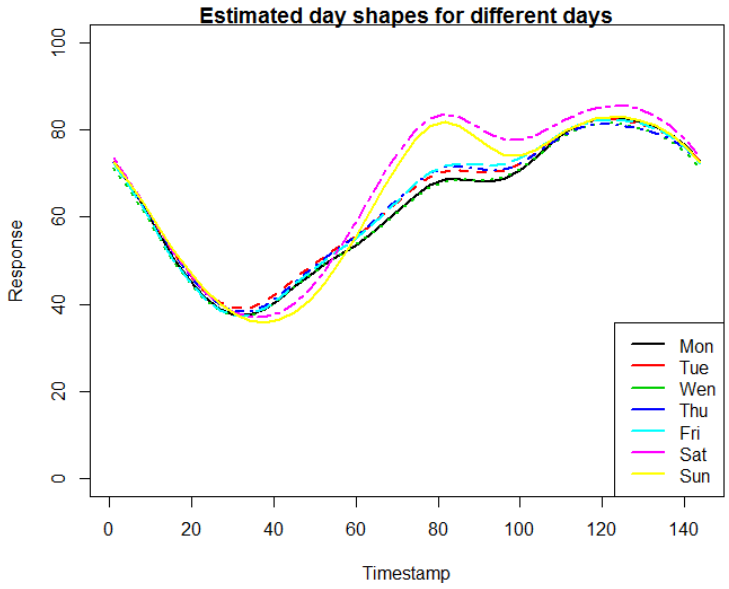


FIGURE 6.5: Estimated day shapes for Figeholm using M1 AR model.

6.1.2 Prediction results

The two models including the AR-process is now used for predicting the load for 41 days from the 25th of October to the 4th of December 2016. The predictions has been made using known weather parameters. The prediction results for model M1 AR and M2 AR are shown in Figure 6.6. As can be seen in this plot, both models seem to have problem with underestimation of the higher load values, however model M1 AR yields somewhat better predictions than model M2.

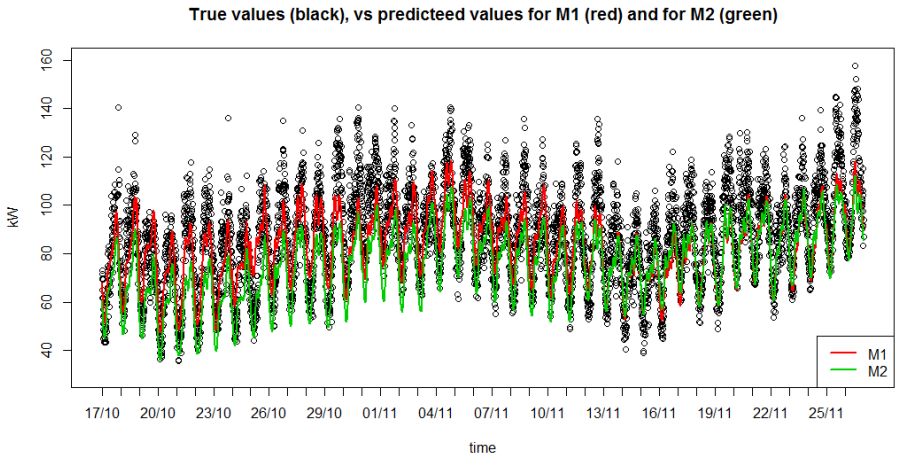


FIGURE 6.6: Predictions for Figeholm data using M1 AR (red line) and M2 AR (green line). Plotted against the true values (black circles).

As the prediction of the upper quantiles is biased, this might be solved by adding prior weights to the models. The result in Table 6.4 shows the result from fitting and prediction using three different kinds of model M1 and M2. First one ignoring the autocorrelation, second one using an AR-process for the residuals and the third one using prior weights and the AR-process. Thus, only the AR models are used when including weights and the autocorrelation will not be affected by the weights.

The table also shows the mean value of the residuals for the prediction, which indicates that all models underestimates the load, but the ones using prior weights performs better. The prediction errors for model M1 AR are lower than for M1 without the AR model and the mismatch (MAPE) is also lower. This may indicate that there exists some overfitting when correlation is ignored. In the table Mean(res) explains the mean value of the residuals in the prediction.

Model	MAPE (%) fit	MAPE (%) prediction	Mean (res) prediction
M1, $\rho = 0$	10.67	11.05	-4.503
M1, $\rho = 0.7222$	11.13	10.67	-3.908
M1W, $\rho = 0.7112$	13.72	9.97	-2.622
M2, $\rho = 0$	10.94	10.98	-7.484
M2, $\rho = 0.7358$	11.16	14.14	-11.59
M2W, $\rho = 0.7213$	14.45	9.88	-5.05

TABLE 6.4: Result of predictions for Figeholm for three different kind of model M1 and M2. First one is ignoring the autocorrelation, second one using an AR-process for the residuals and the last one also including prior weights.

To visualize the effect of using prior weights in prediction, Figure 6.7 shows the predictions for M1 AR with and without weights. As can be seen both in this figure and in Table 6.4, model M1 AR using weights performs best when it comes to predictions.

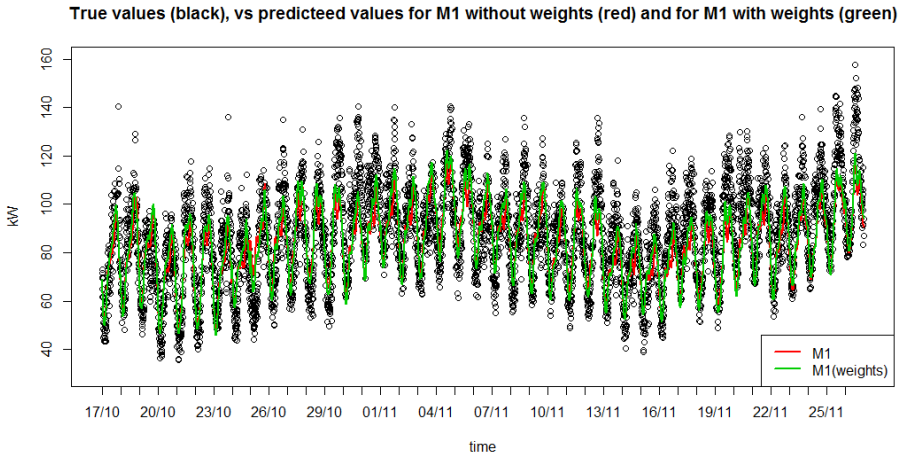


FIGURE 6.7: Predictions for Figeholm data using M1 AR, with and without weights. Plotted against the true values.

6.2 Model fitting results for Hyllie

As shown in the above section the AR-process is required to decrease the autocorrelation in the residuals and therefore only the AR version of model M3 will be used in this section. Model M3 is fitted to the three Hyllie stations - Hyllie1, Hyllie2 and Hyllie3. A table over the time series used in fitting and prediction is shown in Table 6.5.

Station	Data used for model fit	Data used for prediction
Hyllie1	2016-05-14 - 2016-10-15	2016-10-15 - 2016-12-05
Hyllie2	2016-06-05 - 2016-10-20	2016-10-21 - 2016-12-05
Hyllie3	2016-05-12 - 2016-10-14	2016-10-15 - 2016-12-05

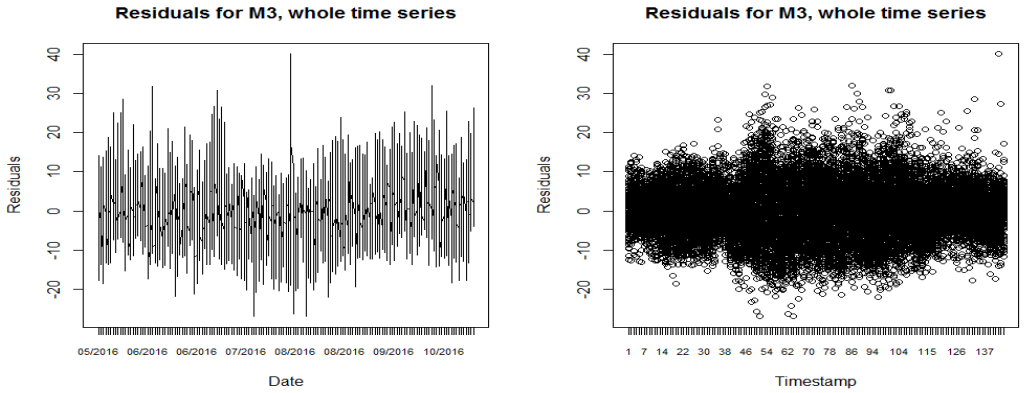
TABLE 6.5: Data sets used for fitting model M3 to Hyllie station.

The tables with fitting results is found in appendix B. As described before the Hyllie data differs a lot, thus model M3 performs variously well on the different stations. The model has been fitted both with and without weights, as described in Section 5.8. For all three stations the model without weights displays the best performance both in goodness of fit and in prediction. The result from the fitting and the predictions are shown in Table 6.6. In the table R^2 describe the adjusted R^2 and Mean(res) is the mean value of the residuals for the prediction.

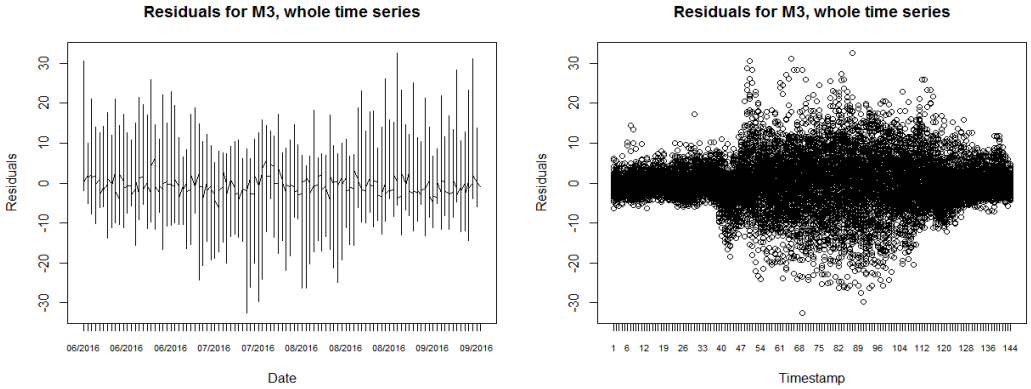
Station	Model	Weights	AIC	R-sq	MAPE fit	MAPE pred	Mean(res)
Hyllie1	M3, $\rho = 0.73$	No	123783	0.907	5.83%	9.74%	-2.669
Hyllie1	M3, $\rho = 0.70$	Yes	133842	0.911	6.16%	10.65%	-2,933
Hyllie2	M3, $\rho = 0.71$	No	103676	0.94	7.59%	7.88%	-2.193
Hyllie2	M3, $\rho = 0.68$	Yes	114425	0.95	8.51%	10.50%	0.870
Hyllie3	M3, $\rho = 0.65$	No	142925	0.942	4.13%	8.24%	0.477
Hyllie3	M3, $\rho = 0.62$	Yes	149157	0.955	4.28%	9.46%	8.587

TABLE 6.6: Result for fit and prediction for all three Hyllie stations. All model include the AR-process and are fitted both with and without prior weights.

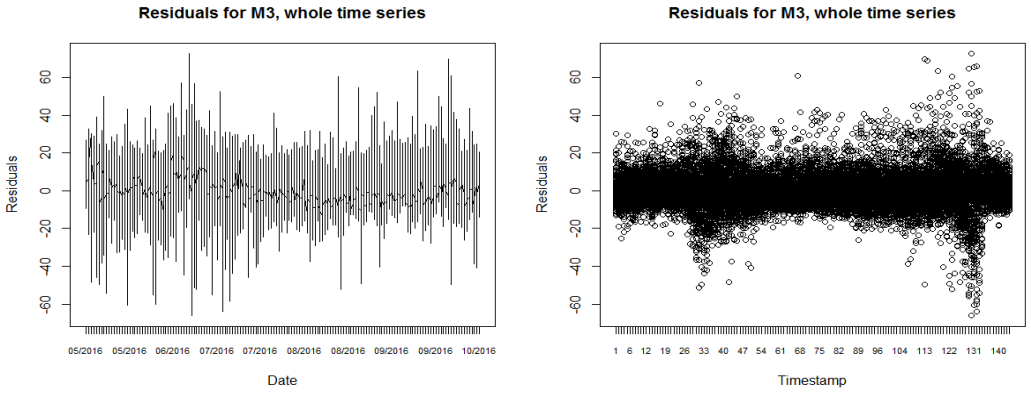
Examining the residual plots gives a hint of the problems in model fitting. The residuals plotted over the whole time series as well as over the day is shown in Figure 6.8. Note that the magnitude of the residuals is not comparable since the size of the load differs for the stations. There are some interesting patterns that is worth discussion. For Hyllie1 and Hyllie2 the major errors when fitting the model occur during daytime when the variations in loads are higher. Hyllie3 is a station with on/off load producing errors for the hours were the ”switch” takes place, i.e. in the morning and in the early evening. This is a common problem in modelling and predicting electricity usage, since there are a lot of outliers in the dataset, caused by unpredictable events.



(A) Hyllie1



(B) Hyllie2



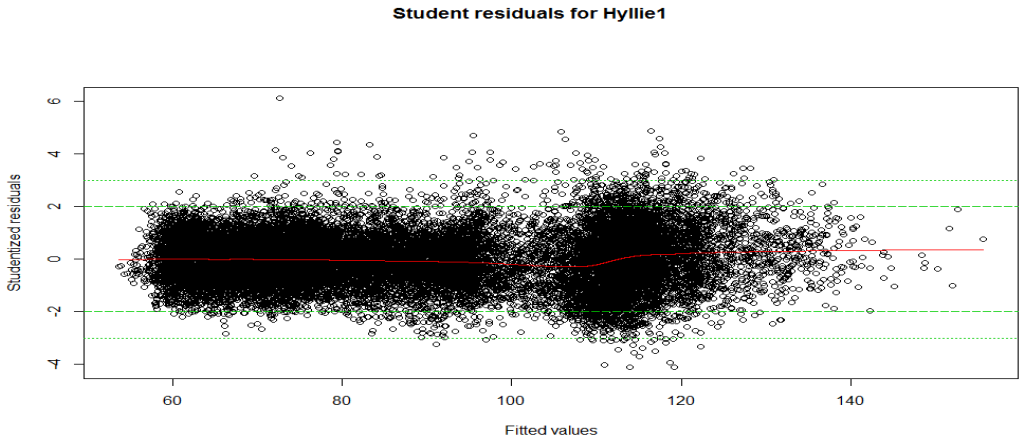
(C) Hyllie3

FIGURE 6.8: Residual plots for the three Hyllie stations using model M3 without prior weights. The plots on the left hand side show the residuals against the whole time series. The plots on the right hand side show the residuals against time of the day.

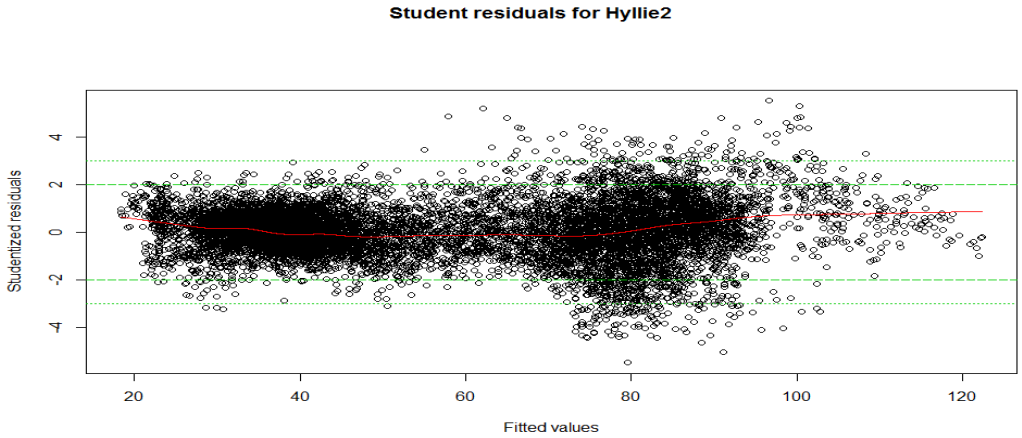
Also the residuals plotted against the fitted values are examined. The results are shown in Figure 6.9. Compare to the same plot for Figeholm (Figure 6.2(a)) it can be stated that for the Hyllie station the residuals shows more patterns. Especially for Hyllie2 and Hyllie3, there exist two ”groups” of residuals depending on low loads and high loads. For Hyllie1 and Hyllie2 it can again be seen that the high loads (during the daytime) are harder to model than the low loads and thus give larger residuals.

By plotting the response variable against some of the explaining variables, it is possible to understand how they affect the load. This is done for the three stations in Figure 6.10. Note that the middle plot for each station is showing the global radiation, and since there is no sun at nighttime, these timestamps has been deleted. There seems to be some relation between the temperature and the load and between the global radiation and the load. The dependence of the wind speed is not as obvious, but there might be a small increase in the load due to increased wind speed. Note that for Hyllie2 the temperature effect is reversed compare to the other station, with an increasing load for increasing temperatures. This might be an effect of air conditioning in the summer, as well as the fact that the measurements does not yet cover a whole year and thus the temperature effect might still be indistinct.

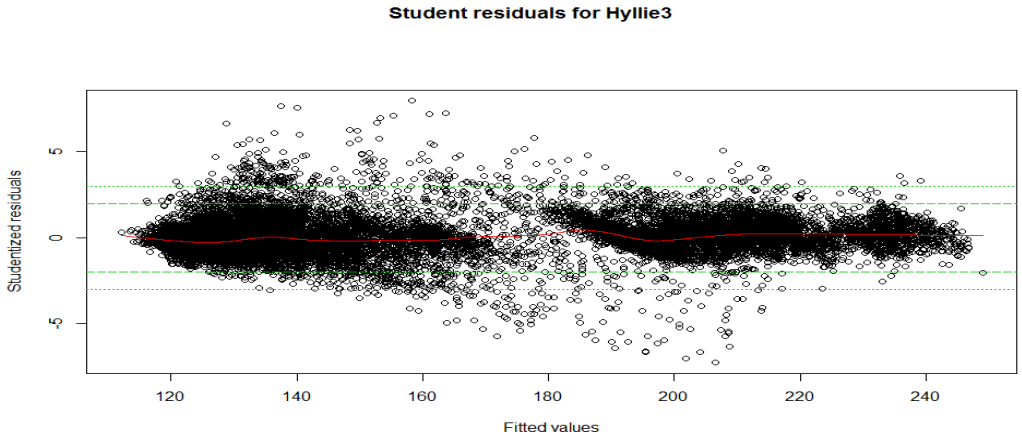
The figures in 6.11 shows the estimated daily profiles for the Hyllie station. Comparing these with the plotted mean values for different days in Figure 5.3, a similarity can be seen. The day shapes for Hyllie3 is however a bit misleading, since the on/off load is not so visibly when plotting the day shapes for the whole period.



(A) Hyllie1

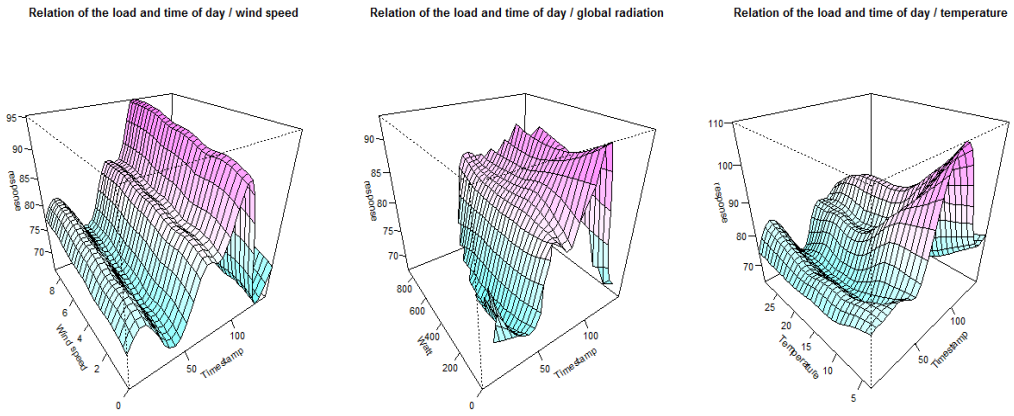


(B) Hyllie2

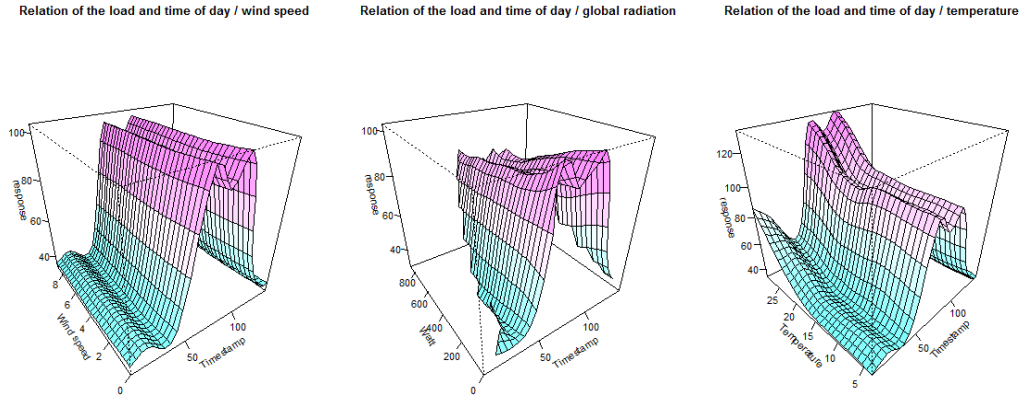


(C) Hyllie3

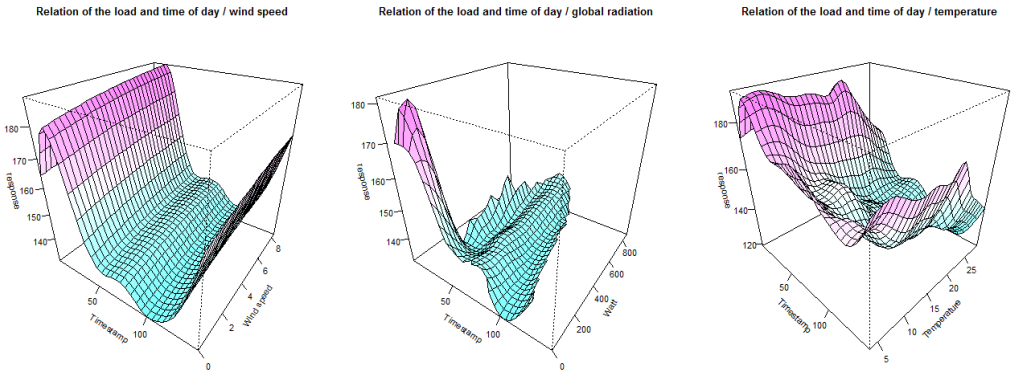
FIGURE 6.9: Students residual versus fitted values for the three Hyllie stations using model M3.



(A) Hyllie1

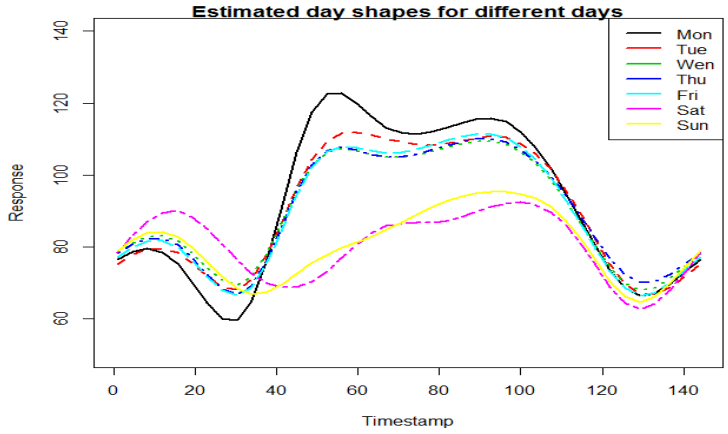


(B) Hyllie2

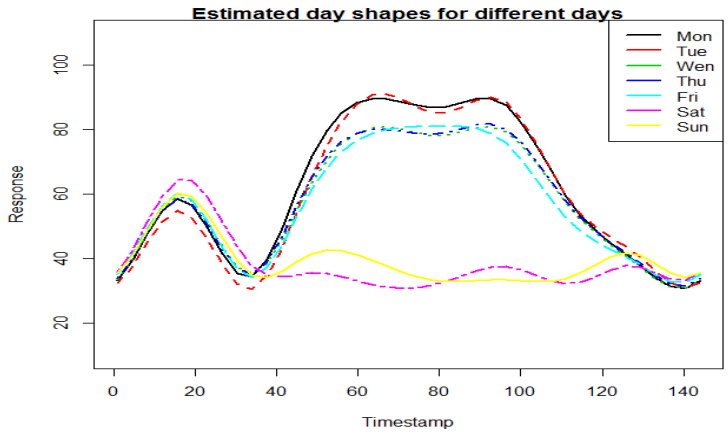


(C) Hyllie3

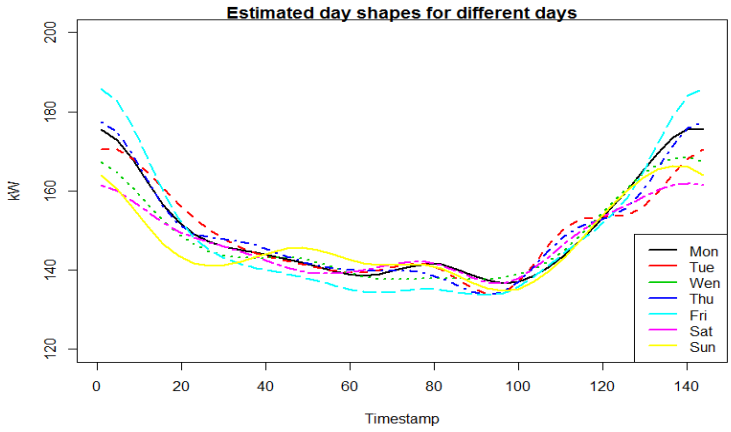
FIGURE 6.10: Response variable against some predictors for fitting model M3 to the three different data sets from Hyllie. Note that the the axis are rotated differently for Hyllie3.



(A) Hyllie1



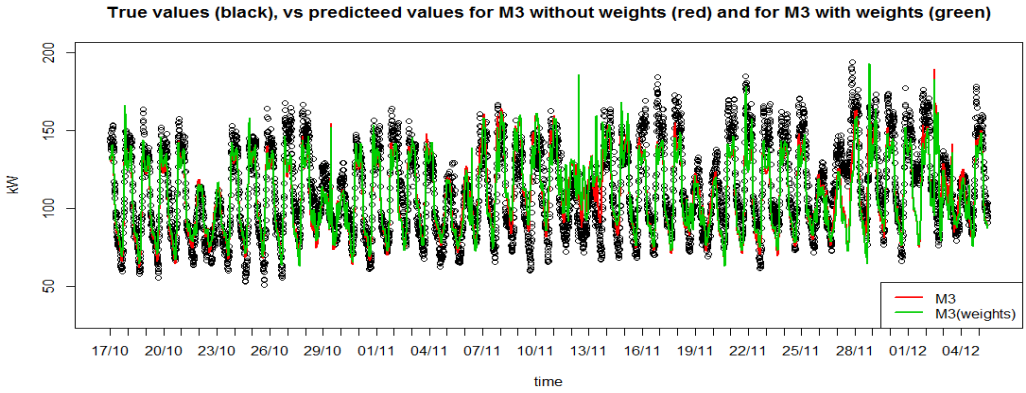
(B) Hyllie2



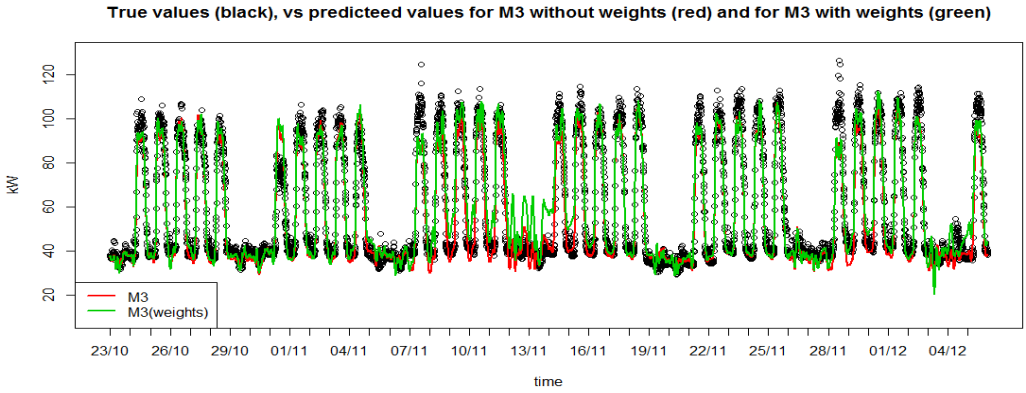
(C) Hyllie3

FIGURE 6.11: Estimated load profiles for different days for the three Hyllie stations using model M3.

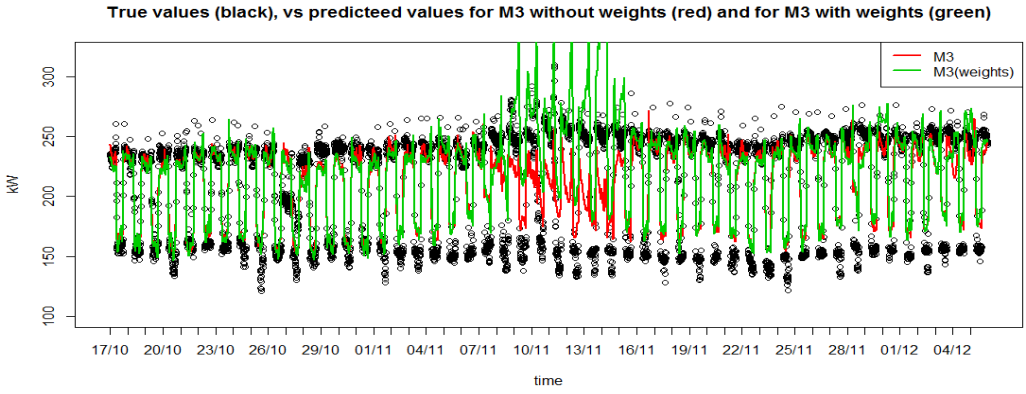
Predictions has been made according to the the time series given in Table 6.5. Figure 6.12 shows the prediction for the three stations using model M3 with and without weights.



(A) Hyllie1



(B) Hyllie2



(C) Hyllie3

FIGURE 6.12: Predictions for the three Hyllie stations using model M3 (with and without prior weights).

It is clear, both regarding the value of MAPE and adjusted R^2 for the predictions (given in Table 6.6), and by looking at the plots in Figure 6.12 that for all stations model M3 without any prior weights will yield the best results. The predictions for all stations shows some large errors starting around the 10^{th} of November and giving especially large error the weekend 12^{th}-13^{th} November. The reason for this error is unknown but may be caused by some maintenance work in the regional substation. Using weights will also produce larger errors, probably due to the fact that the prior weights will assign resent values higher importance.

Chapter 7

Discussion

"If you can't explain it simply, you don't understand it well enough."

– Albert Einstein

The results given in Chapter 6 and the potential usage and development of the models will be discussed in this chapter. In this thesis a lot of focus has been on the mathematical models and thus some of the results, that is not of importance for the models, has been left out of this report.

Since measuring in secondary substations is new to E.ON Elnät, it does not currently exist any analysis or modelling of the load in secondary substations. As a consequence, it is hard to compare the models and results in this thesis with any existing data or models. As mentioned in the theory section about Velander's formula (Section 2.3.1) and planning of a local network (Section 2.3.2), some models are used for prediction of the maximum load when new local network is being built, but afterwards, usually no measurements is collected as long as no problem arises. Hence, discussion will be based on the opportunities these models provide rather than a comparison, including some problems and development potentials.

7.1 Choice of model and model parameters

The goal for the models was to explain how the load varies according to different variables. The time of the day, time of the week and time of the year was of high interest to capture, as well as the impact of the weather. The generalized additive model was therefore chosen thanks to the ability to capture all these different relations in the same model. When choosing which parameters to include and how to model them, many different alternatives has been tried out. The models described in this thesis were selected since they showed the best results according to goodness of fit and because the included parameters, showed high significance and are supposed to have impact on the load.

The temperature relation was expected to be of importance. Traditionally the electricity usage is higher in the colder month, than during summertime. Some buildings still depend on electricity for heating and even when other heating solutions is used, electronic devices can help increasing the temperature, for example in an office. Some of the load - temperate

dependence can also be explained by the fact that more electricity is used in the winter, due to longer periods of darkness. This increase in electricity usage is not directly caused by a colder climate, but will still capture this relation, as it is often colder in winter time. An interesting result is that, for Hyllie2 there seems to be an increase in electricity usage due to temperature increases. This is the opposite of an expected relation for Swedish electricity usage, but might be explained by air conditioning and ventilation of the office buildings connected to this station. Since the data used in model fitting for this station is from the beginning of June to the middle of October, this relation might look different as more data is included.

The relation to wind speed is a bit unclear and as seen in the figures to the left in Figure 6.10 it seems that, for the data used in this thesis, it could be modelled as linear. It can increase the load as more power is needed for heating due to strong winds, but with newer building this effect is assumed to be less pronounced. However it will be important in a grid containing local wind-power and thus it is included for the Hyllie station. Also the relation to the global radiation is included for the model used in Hyllie, since this parameter is expected to be of importance as more solar power is installed. The measurements for global radiation used in this thesis come from a station quite far from Hyllie. However the relation still seems to be significant for all three station (see tables B.1, B.2, B.3 in Appendix B), even if the cloud coverage probably differs in these two places. The explanation for this effect can be that the difference is cancelled out in the long run. In Hyllie2 it is known to be some production from solar power installed. As visible in Figure 6.10, there might be a stronger dependence of the global radiation for this station compared to the others.

7.2 Comparison between the different stations

Even if the stations used in this thesis have very different types of loads, the models perform reasonably well for all of them. The value of the adjusted R^2 is high for all model fits, indicating that the model captures the variability in the data in a satisfying way. For Hyllie, the same model gives good results for all stations.

The model used for the Hyllie station seems to perform better and yields better predictions than the models used for Figeholm. An explanation to this might be that the load in Figeholm has more variability than the load in Hyllie. This might be a result of too much data used for the model fits and is probably the reason why models using prior weights perform better for the Figeholm data. Thus it remains to update the models as new data arrives and reduce the impact of older data that is no longer relevant for the predictions.

Even if the value of MAPE are lower and the predictions seem better for Hyllie, the residual diagnoses and the autocorrelation is still better for Figeholm. This is presumably a result from the response not being normally distributed and due to the fact that the load is still changing a lot over time in Hyllie.

For the Hyllie stations, where the measurements do not cover a whole year, it is hard to model the annual dependence. Try-outs have been made including a yearday or month variable, but this will result in overfitting the dependence of these parameters. This is probably due to the fact that the model captures the relation of increase in load due to increase in the yearday parameter too strongly. Having data from many successive years gives the advantage of modelling the dependence of seasonal changes. On the other hand, for an area

like Hyllie where the load is affected by the changes in the area this could possibly result in misleading conclusions. Nevertheless, the load in Hyllie will have some annual patterns and thus the seasonal dependence should be modelled when more data is available. In Figure 6.8 (a) and (b) there could be some seasonal patterns visible in the residuals caused by the exclusion of this parameter.

Since the stations differs quite a lot it would be a good idea to make the models more flexible to the used data. Even though the coefficients and smoothing parameters is estimated for each dataset, an improvement would be to add a generic parameter selection so that the model includes only the most significant parameters for that station. Furthermore, the maximum number of knots in the smoothing splines is set to be constant, which may affect the result as it limits the effective degrees of freedom for the explaining variables.

7.3 Autocorrelation

Parts of the autocorrelation are reduced by adding the dependence of the load and temperature from the previous days. Furthermore, modelling the errors as an AR-process reduces the autocorrelation to acceptable levels. However there will still be some correlation between the residuals. Especially the model fit for Hyllie stations still show some autocorrelation (see Appendix C). This is probably due to the fact that the data is collected in such high frequency that the autocorrelation is very high. Perhaps a more advance time series model could be used to manage the autocorrelation better.

7.4 Problems with modelling and prediction of data containing outliers

Data with many outliers are regarded as ”bad quality data”, but for the company these outliers can be very interesting and important to understand. However, since these outliers can have a remarkably effect of the performance of the predictions, some extreme outliers has been removed from the dataset. These removed outliers can be assumed to be caused by some abnormal events and thereby non-representative for the load profile models. Large power measurements like the ones used in this thesis will always contain a lot of outliers, complicating the modelling. But since the outliers will exist also in future loads it might be of interest to include them in the models. As a result, the model fits and the predictions will have some error spikes. Especially in Figure 6.8, it is visible how the outliers and variance of the data affects the performance of the models. As mentioned before, the major errors occur when the variance in the loads are high (during daytime for Hyllie1 and Hyllie2, and for the hours were the ”switch” takes place for Hyllie3). This pattern can also be caused by the fact that the smoothing splines strive for being as smooth as possible, at the expense of flattening out the outliers. This might also be one of the reasons why the models underestimates the peak loads in the predictions. The outliers are also suspected to cause spikes in the autocorrelation, due to sudden dependencies.

All kinds of unusual events will affect the forecast as there is no way for the model to predict these unusual events. An attempt to statistically describe the logged events has been tried out. However due to the fact that the events are too few to show anything of statistical significance, these parameters have not been included in the models. With more event-data

and by using new input parameters there might be opportunities to include this in the models. By having a parameter indicating if there is something unusual going on above or below the station, parts of these problems can be solved. This will require more measuring and a fast attendance of the data. It may also be possible to predict the probability for unusual events when more data is collected, and if some connections to when these events happens can be detected. Thus some works consider the outliers and probability of exceedance of lower and higher limits should be of interest.

7.5 Holiday effect

Through the work of this thesis it has been noticed that the event of holidays, day before a holiday and clamping days can have very different impact on the loads depending on what type of customers the secondary substation is connected to. For example the load for station Hyllie2 where the load differs a lot between weekdays and weekends, an event of a holiday may have big impact if the businesses connected to the station are closed on holidays. Furthermore, it is not only the public holidays that will affect the load. Also school holidays, summer holidays, Christmas holidays, etc. will have an impact on the load that will vary between different types of stations. In conclusion, further work with the holiday effect is of importance to satisfactory predict the loads in different stations.

7.6 Need of more information

The effect of change of power consumption in an area, based on the changes in this area, is something that is assumed to have a lot of impact, especially for long term predictions. The possibility to measure and model this effect might be a presumption for better predictions, especially for an area like Hyllie, where major changes are expected in the close future. Also in the long term the load might be affected by the customer behaviour. As this is a pilot project for smart grid and smart homes, the customers are expected to be more conscious about their own electricity consumption. As they install solar power, buy electric vehicles and use electricity in smarter ways, their changing behaviour will have impact on the load in the substation. How to include this in the models is a question that has to be discussed. There might be a possibly input parameter for this behaviour, or maybe the models should be more adaptive. In addition, the studying of load profiles on customer level may be required to understand the customer behavior better. This can also provide information about why events occur and if some customers have a greater impact on the load in the substation.

Figure 6.9(c) shows the students residuals for a fit using model M3 on station Hyllie3, connected to the train station. In this figure it can be seen that there exists some variability in the data that the model is unable to capture. As the load in this station is strongly affected by the on- and off-switch of the illumination of the train station, some parameter describing this might have to be introduced to improve the model fit.

When modelling many secondary substations in the same area, the load at the different substation should probably not be modelled as independent since there will exist some correlation between the stations. Part of this correlation will be captured by the fact that they will all depend on the same weather parameters, but probably not all. To further model the relation between the stations a vectorial GAM model can be used.

7.7 Opportunities with data analysis in future smart grids

Challenges and opportunities that smart grids bring for the DSO's where discussed in Section 2.6. Most of these challenges require a lot of different solutions, and data management and analysis is just one of them. One of the main concerns are power quality issues regarding the voltage levels in the grid, and the voltage problems that occur for very high and very low load flows. This issue will require solutions that have not been discussed in this thesis, like more automation and frequent measure of the power quality. However increased data analysis can help the DSO's to understand when these problems occur, why they occur and when the risk is higher for different issues. More accurate short term predictions may also be of interest for the operations departments at the DSO's.

As described in Section 2.3.1 the formula (Velanders) used today when predicting load flow in local networks is very simplified and built on old results. Increased data analysis of real data may give the opportunity to do better dimensions in the future that are based on more resent investigations. Today there is a lot of focus on the high load, since this is the greatest issue for the DSO. However, the low loads may also be a problem in the future with a lot of micro production and thus it is of importance to start focusing on load other than the high loads. In addition, to be able to equalize the differences between low load and high load, information about the load flows in the grid has to be available and thus some sort of data analysis has to be implemented.

Better knowledge about the loads flows in the local network and precise prediction might also bring another advantage for the DSO, as there will be an increasing interest for this information from other parties. For example a prosumer wanting to install a solar power plant that brings the opportunity to sell the excess power back to grid. This prosumer may want to have information about the times when there is an increasing demand in the surrounding grid, to get a better picture of the market. This might open up for completely new business opportunities for the DSO's.

Chapter 8

Summary and Conclusion

"Any powerful idea is absolutely useless until we choose to use it."

– Richard Bach

The purpose of this thesis was to investigate if there is a possibility to adequately model the load profiles for different secondary substations in a local network and use these models to forecast the loads depending on both calendar effects and weather conditions. The models described in this thesis focus on short term forecasting one day ahead.

Increased measurements in secondary substation may be of high importance in the future smart grid. Not only does it provide an instrument for faster fault detections and knowledge about events. Through the work with this thesis it is shown that the simple compilation and management of the data can increase the knowledge for the company, as the information in historical data is scarcely used otherwise. It is also clear that, even though the model fitting and predictions still needs refining, it provides opportunities, both as tools for better load profiles and as a prediction instrument. Since one of the goals with smart grids, is to even out the load during the hours of the day, there has to be available information about how the load looks like. In a future smart grid network this information will be necessary as it gives information about the capacity margins in the grid, likewise it provides knowledge about the influences a local production may have for the area.

If the model shows a more elaborate relation between day of week and time of day, rather than weather parameters, one can assume that the transition to renewable energy sources in the network is not completed. It is expected that in a smart grid network the difference between high load and low load will be smaller. For example an increasing usage during nighttime when the demand is lower and thus electricity is cheaper, and smaller morning and evening peaks. With more renewable energy, the models will probably show a stronger relation between the load and the wind and sun, with a higher usage during strong winds and sunny periods.

As the weather dependence in the model is assumed to be of importance in a future smart grid area, E.ON Elnät will install a weather station in Hyllie that measures the weather parameters on a local level. This weather station will also include global radiation as this variable varies a lot and can differ on relatively short distance. As more solar power systems are installed in Hyllie, this parameter will have an increasing importance for correct modelling of the loads.

Bibliography

Gunnar Blom, Jan Enger, Gunnar Englund, and Holst Lars. *Sannolikhetsteori och statistikteori med tillämpningar*, volume 5:5. 2005.

Math Bollen. EI R2010:8 Anpassning av elnäten till ett uthålligt energisystem - Smarta mätare och intelligenta nät, 2010.

Math Bollen. Old and new uncertainty in power systems. In *Proceedings of IEEE PES Membership meeting with Distinguished Guest Speakers*. Lulea University of Technology, sept 2016.

Math Bollen, Y. Yang, and F. Hassan. Integration of Distributed Generation in the Power System - A Power Quality Approach, 2008.

Greta Brännlund. Evaluation of two load forecasting methods used at fortum. Master's thesis, KTH Electrical Engineering, 2011.

Arnaud De Myttenaere, Boris Golden, Bénédicte Le Grand, and Fabrice Rossi. Mean absolute percentage error for regression models. *Neurocomputing*, 192:38–48, 2016.

J Dickert and P Schegner. Residential load models for network planning purposes. In *Modern Electric Power Systems (MEPS), 2010 Proceedings of the International Symposium*, pages 1–6. IEEE, 2010.

Energymyndigheten/Swedish Energy Agency. Energy in Sweden 2015. `https://www.energimyndigheten.se/globalassets/statistik/overgripande-rapporter/energy-in-sweden-till-webben.pdf`, 2015. [Online; accessed 14-November-2016].

E.ON, VA SYD, Malmö Stad, and Swedish Energy Agency. Climate smart-Hyllie - testing the sustainable solutions of the future. `http://malmo.se/download/18.760b3241144f4d60d3b69cd/1397120343885/Hyllie+klimatkontrakt_broschyr_EN_2013.pdf`, 2013. [Online; accessed 21-October-2016].

E.ON Elnät. Kraftöverförings-bild, 2016a. [Picture internally used at E.ON].

E.ON Elnät. E.ON Elnät Sverige AB. `https://www.eon.se/eldistribution`, 2016b. [Online; accessed 1-November-2016].

E.ON Sverige. E.ON Sverige i siffror. `https://www.eon.se/om-e-on/om-foeretaget/nyckeltal.html`, 2016a. [Online; accessed 1-November-2016].

E.ON Sverige. Smart city Hyllie, Setting up real-time business, 2016b. [Presentation internally used at E.ON].

European Regulators Group of Electricity & Gas. Position Paper on Smart Grids - An ERGEG Public Consultation Paper, 2009.

Yannig Goude, Raphael Nedellec, and Nicolas Kong. Local short and middle term electricity load forecasting with semi-parametric additive models. *IEEE transactions on smart grid*, 5(1):440–446, 2014.

Anders Grauers. *Elteknik*. Chalmers Tekniska Högskola, 2000.

Anders Gustafsson. Interview with a local network planner at E.ON Elnät. Personal communication, 2016.

Tao Hong and Shu Fan. Probabilistic electric load forecasting: A tutorial review. *International Journal of Forecasting*, 32(3):914–938, 2016.

Rob J Hyndman and Shu Fan. Density forecasting for long-term peak electricity demand. *IEEE Transactions on Power Systems*, 25(2):1142–1153, 2010.

Ola Ivarsson. Interview with a specialist on E.ON Elnät. Personal communication, 2016.

Lauri Kumpulainen, Petri Trygg, Kim Malmberg, M Hyvärinen, S Pettissalo, and M Loukkalahti. Secondary substation monitoring and control–practical benefits through intelligent components and systems. In *International Conference on Electricity Distribution*, pages 1–4, 2011.

Annelie Lindeberg. Mätning i nätstationer - nyttor och problem, En studie i samarbete med Göteborg Energi AB. Master's thesis, Chalmers tekniska högskola, 2009.

Georg Lindgren, Holger Rootzén, and Maria Sandsten. *Stationary stochastic processes for scientists and engineers*. CRC press, 2013.

Malmö Stad. Climate smart Hyllie. `http://malmo.se/Nice-to-know-about-Malmo/Sustainable-Malmo-/Sustainable-City-Development-2016/Sustainable-City-Development/Climate-smart-Hyllie.html`, 2011. [Online; accessed 21-October-2016].

E Meijering. A chronology of interpolation: from ancient astronomy to modern signal and image processing. *Proceedings of the IEEE*, 90(3):319–342, Mar 2002. ISSN 0018-9219. doi: 10.1109/5.993400.

Python 3.5 programming language. Python Software Foundation, 2014. URL `http://www.python.org/`.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. URL `http://www.R-project.org/`.

David Ruppert, Matt P Wand, and Raymond J Carroll. *Semiparametric regression*. Cambridge university press, 2003.

Robert H Shumway and David S Stoffer. *Time series analysis and its applications: with R examples*. Springer Science & Business Media, 2010.

Svensk Energi. Elnätet - nära 14 varv runt joden. `http://www.svenskenergi.se/Elfakta/Elnatet/`, 2012. [Online; accessed 1-November-2016].

Svenska Elverksföreningen. *Belastingsberäkning med typkurvor*. Svenska Elverklföreningen, 1991.

Svenska Kraftnät. National grid. `http://www.svk.se/en/national-grid/`, 2016a. [Online; accessed 1-November-2016].

Svenska Kraftnät. The electricity grid. `http://www.svk.se/drift-av-stamnatet/stamnatskarta/`, 2016b. [Online; accessed 1-November-2016].

Jacolien van Rij. *itsadug: Interpreting Time Series, Autocorrelated Data Using GAMMs (itsadug)*, 2016. R package version 2.2.

Helen M Walker. Degrees of freedom. *Journal of Educational Psychology*, 31(4):253, 1940.

Simon N Wood. *Generalized Additive Models - An Introduction with R*. CRC press, 2006.

Simon N Wood. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(1):3–36, 2011.

Simon N Wood. On p-values for smooth components of an extended generalized additive model. *Biometrika*, page ass048, 2012.

Simon N Wood. *mgcv: Mixed GAM Computation Vehicle with GCV/AIC/REML Smoothness Estimation*, 2016. R package version 1.8-12.

Simon N Wood, Yannig Goude, and Simon Shaw. Generalized additive models for large data sets. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 64(1): 139–155, 2015.

Appendix A

Additional mathematical methods

A.1 Linear interpolation

To be able to use a ten minute model, when only hourly mean values is provided linear interpolation has been used. Linear interpolation approximates the unknown point y for x, between two known points (x_o, y_o) and (x_1, y_1) according to:

$$\frac{y - y_0}{x - x_0} = \frac{y_1 - y_0}{x_1 - x_0}. \tag{A.1}$$

Basically the formula states that the slope should be the same between x and x_0 as it is between x_0 and x_1. To be able to use a ten minute model, when only hourly mean values is provided linear interpolation has been used in this thesis. (Meijering, 2002)

A.2 QR-decomposition

QR-decomposition is used in linear algebra and is the method of decomposition of a real square matrix A into the product $A = QR$, where Q is an orthogonal matrix (i.e. $Q^T Q = I$) and R is an upper triangular matrix. The factorization will be unique if A is non-singular. It exist several methods for QR-decomposition.

In this thesis a QR-decomposition will be used in fitting of the GAM, as this will lower the computational burden. (Wood, 2006)

Appendix B

Result of model fits for Hyllie

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	96.2167	1.1098	86.6943	< 0.0001
B. smooth terms	edf	Ref.df	F-value	p-value
te(TS,Ptot_144):D_type1	23.0480	28.8558	7.6738	< 0.0001
te(TS,Ptot_144):D_type2	27.0102	34.1082	12.8868	< 0.0001
te(TS,Ptot_144):D_type3	21.4312	28.7016	13.8215	< 0.0001
te(TS,Ptot_144):D_type4	16.7524	22.8558	15.8251	< 0.0001
te(TS,Ptot_144):D_type5	15.4007	20.7462	17.2852	< 0.0001
te(TS,Ptot_144):D_type6	38.9352	46.0015	32.1705	< 0.0001
te(TS,Ptot_144):D_type7	36.0669	42.8387	37.3858	< 0.0001
te(Temp_i,TS)	42.4730	51.6677	7.8676	< 0.0001
te(Temp_144,Temp_288)	19.4640	26.2469	4.8652	< 0.0001
s(Wind)	6.7244	8.8068	8.6593	< 0.0001
te(Watt,TS)	20.7507	76.0000	1.5792	< 0.0001

TABLE B.1: Summary of results for fitting model M3 without weights to Hyllie1.

A. parametric coefficients	Estimate	Std. Error	t-value	*p*-value
(Intercept)	67.6726	1.3177	51.3570	< 0.0001
B. smooth terms	edf	Ref.df	F-value	*p*-value
te(TS,Ptot_144):D_type1	14.1536	16.6060	5.5866	< 0.0001
te(TS,Ptot_144):D_type2	39.3015	45.9447	11.3408	< 0.0001
te(TS,Ptot_144):D_type3	32.0999	39.5565	17.0971	< 0.0001
te(TS,Ptot_144):D_type4	33.4468	40.3022	17.2264	< 0.0001
te(TS,Ptot_144):D_type5	30.5704	37.2672	21.4929	< 0.0001
te(TS,Ptot_144):D_type6	46.3333	52.6201	32.3391	< 0.0001
te(TS,Ptot_144):D_type7	34.5377	40.2987	44.8260	< 0.0001
te(Temp_i,TS)	49.6735	59.2223	19.5180	< 0.0001
te(Temp_144,Temp_288)	35.0225	43.3562	3.8298	< 0.0001
s(Wind)	3.7068	4.8660	9.3675	< 0.0001
te(Watt,TS)	35.9839	76.0000	10.3527	< 0.0001

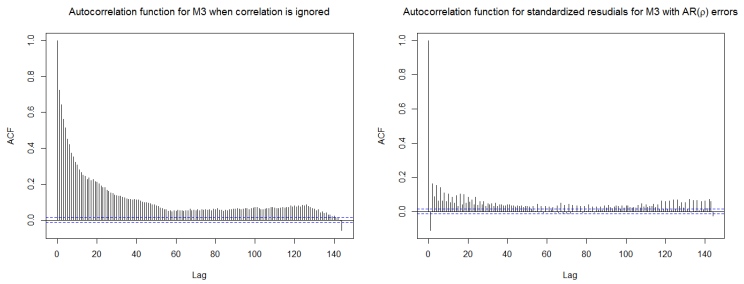
TABLE B.2: Summary of results for fitting model M3 without weights to Hyllie2. (te describes the tensor product two regression splines and s denotes a smoothing spline.)

A. parametric coefficients	Estimate	Std. Error	t-value	*p*-value
(Intercept)	176.8545	1.8157	97.4048	< 0.0001
B. smooth terms	edf	Ref.df	F-value	*p*-value
te(TS,Ptot_144):D_type1	45.4144	55.2363	45.5336	< 0.0001
te(TS,Ptot_144):D_type2	45.0882	54.8517	55.0564	< 0.0001
te(TS,Ptot_144):D_type3	38.5810	48.9373	54.6312	< 0.0001
te(TS,Ptot_144):D_type4	43.5311	53.2010	52.3752	< 0.0001
te(TS,Ptot_144):D_type5	46.5968	56.4020	49.3889	< 0.0001
te(TS,Ptot_144):D_type6	38.4427	47.9391	55.8235	< 0.0001
te(TS,Ptot_144):D_type7	34.1306	43.5294	63.1450	< 0.0001
te(Temp_i,TS)	47.9978	56.9744	11.4698	< 0.0001
te(Temp_144,Temp_288)	29.2225	37.3843	6.8662	< 0.0001
s(Wind)	2.6919	3.4916	36.9378	< 0.0001
te(Watt,TS)	42.0659	75.0000	10.2886	< 0.0001

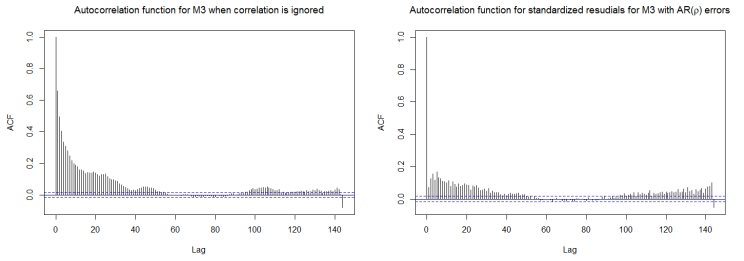
TABLE B.3: Summary of results for fitting model M3 without weights to Hyllie3.

Appendix C

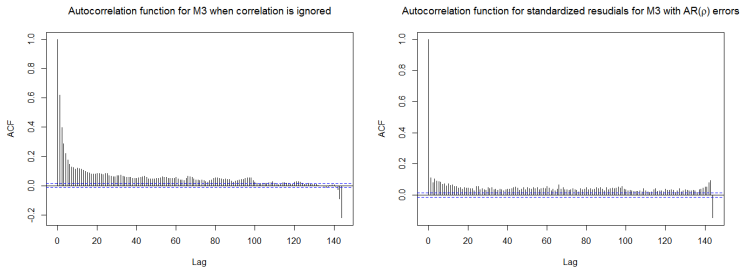
ACF for Hyllie stations



(A) Hyllie1



(B) Hyllie2



(C) Hyllie3

FIGURE C.1: Autocorrelation for model M3 without weights fitted to the three different Hyllie stations. Left figures showing model fits when autocorrelation is ignored and right figures showing model when the residuals are modelled with an AR term.