CHALMERS

Development of a Tool to Determine the Correlations between Transmission Losses and Influencing Factors

Master's Thesis in Sustainable Energy Systems

Kristina Nilsson & Jenny Jeppsson

Department of Energy and Environment
Division of Electric Power Engineering
Chalmers University of Technology
Gothenburg, Sweden 2014

Abstract

Investigations on which factors that affect the power transmission losses have been performed several times before. However, increasing the size of the studied system rapidly increases the complexity and the problem becomes computationally heavy to solve. In this thesis a statistical approach has been used in order to be able to investigate what factors that affect the power transmission losses in a large and complex network. The area investigated is Regionnät Syd, a medium voltage network with over 200 substations, owned by E.ON Elnät Sverige AB. Special focus in the thesis is on how losses are affected by the increasing share of distributed generation, such as wind power production, in the system. The statistical calculations were performed in Lavastorm, a data analysis software which can support statistical models written in the programming language R.

The result gave the correlations between the power losses and the parameters included in the study and was presented as a marginal effect analysis. The main finding is that distributed wind power production affects the power losses significantly. For the majority of the data an increasing share of wind power decreases the power losses, but after a critical value the losses start increasing. To use a statistical approach proved to be a feasible method to study a power network. However, it is important to keep in mind that a correlation between a factor and a variable are not necessarily the same thing as causation.

The final model presented in the results has some defects that are accounted for in the discussion. Possible improvements, in order to get reliable results, are presented as recommendations for future work.

key words: transmission losses; wind power; regression analysis; big data

Acknowledgements

We would first of all like to thank Sezgin Kadir and E.ON Elnät Sverige AB for the opportunity to do this Master thesis work at the company. We would like to express our gratitude for the warm welcome and inclusion in the Energicontrolling group where we were placed.

We would like to acknowledge and thank our supervisors at Energicontrolling; Wilhelm Schånberg and Sofia Nivhede, for all help and discussions throughout the progress of the work. Another person that has been of great importance in this work at E.ON is Torbjörn Stenström. Thank you for all help with Lavastorm and the possibility to use R within the software. Moreover, we would like to thank Claes-Håkan Månsson as well as Johan Nilsson and Andreas Brorsson for all additional data collecting help and valid discussions.

We would like to thank Malmö Moderna Museum, for giving us the opportunity to have a workshop there, and to all participants at E.ON for taking their time to attend. It was an important step in out project work. Furthermore, we have been given the chance to do a couple of study visits during our time at E.ON. We would like to thank Ola Borrström, also at E.ON Elnät Sverige AB for the visit to the electrical substation and to Per Rosén at E.ON Business Innovation Sverige AB for taking the time to be our guide when visiting Hållbarheten, a sustainable project-house of E.ON in Västra Hamnen, Malmö. Both the flats and the discussions were very interesting.

We would not like to forget all the help we have had from Chalmers; We would like to acknowledge Vera Lisovskaja at Mathematical Sciences for all the help with the statistics throughout the project. Thank you Vera, your guidance and help has been crucial for us in this project. Another thanks goes to our examiner Ola Carlson, Professor in Sustainable Electric Power Production.

Finally, we would like to thank our families for all the support, not forgetting all the help in the changeovers between Göteborg and Malmö.

Jenny Jeppsson and Kristina Nilsson, Malmö 09/06/2014

Contents

1

Introduction

To meet the challenges that the power grids face today, with a constantly increasing share of renewable power production in the system, the power grids must evolve. The grids must become smarter, as the term smart grid implies. One of the goals of developing the smart grids is that the electricity supply system shall become more energy efficient. By allowing power production of all sizes, encouraging renewable power production and allowing customers to participate as producers, the regional production will play an increasingly important role in power production in the future.

But how shall the power grids be developed? What changes are needed in order to handle these new challenges? How much regional production, such as wind power, is beneficial before the losses are too high? Is it cost efficient to reinforce the grid to lower losses? To understand how the new development influences the power grids is of interest and there is a lot of research going on in the field. One way to study the system is by looking at it from a statistical point of view, to see how the power grid is answering to recent development.

In this thesis, the impact on the losses in the power grid is studied using a statistical approach. This introduction chapter will cover the aim, problem description and goal of the study as well as a short introduction to the company where the study is performed and the statistical approach. The boundaries for the study are covered as well.

1.1 The Aim of the Project

The aim of this project is to investigate if a top-down statistical approach is feasible for studying the losses in a power distribution network. The aim of studying the losses is to quantify which parameters that do effect the system losses and at what level.

1.2 Losses in Power Grids

The losses occurring when transmitting power does not only decrease the energy efficiency of the transmission, they also incur increased costs. The power losses are unwanted but unavoidable. However, the share of transmitted power that is lost along the way between production site and end user varies significantly between different countries and areas. In Sweden 7% of the produced electricity is dissipated before it reaches the customer according to statistics from The World Bank. In Germany, the country with the most efficent network in the world, the same figure is 4%. The figure differs a lot in between countries, some have loss rates as high as 55%. Note that in countries with numbers as high as 55 % there is a lot of electricity theft, which appears in the statistics as losses. [1]

The divergent performances of different power grids can be explained. Some key factors are known, others are not. Knowledge of what factors increase and decrease the power losses makes it possible to optimise the grid with respect to the power losses. The stability and reliability of power delivery have historically been given higher priority than the power losses, but of late the losses have started to attract more attention.

Today in Sweden, the grid owners operate as a regulated monopoly since the customers can't choose which company that shall deliver their energy. The grid owner's revenues are regulated by the Energy Markets Inspectorate (Energimarknadsinspektionen). The grid owner's expenses due to power losses are considered inelastic and can be forwarded directly to their customers. But these rules are about to change, and the losses will in the future be considered partly responsive of efforts from the grid owner.[2]

1.3 E.ON Elnät Sverige AB

The study was performed at the department Energicontrolling at E.ON Elnät Sverige AB. E.ON Elnät Sverige AB a subsidiary of E.ON Sverige which is part of the German energy company E.ON AG. E.ON Elnät Sverige AB owns parts of the Swedish distribution network and is responsible for the operation and maintenance of these. They provide the service of transporting the power from connecting grids to end users. They also transport power from one grid owner to another, only letting the power pass through their network.

Prior to the project, an investigation had been performed at Energicontrolling regarding the network Regionnät Syd, showing that there is a potential correlation between losses in the medium voltage net and the wind power production. Preliminary results show that losses decrease with increased wind power. It is of interest for E.ON Elnät Sverige AB to explore the matter further, by both validating the preliminary result, and finding correlations between other relevant parameters and the power losses in the net.

1.4 Problem Description

Apart from that the power losses vary with the total amount of power being transported in the distribution network there are other factors that affect the losses in Regionnät Syd. Today, there is a lack of knowledge about which of these other factors that are most important and how they influence the losses. Without this knowledge it is hard to explain why the losses are greater one month compared to another, and it is also difficult to predict the level of losses to use for budgeting. The lack of knowledge obstructs the development of the network towards becoming more efficient. A more efficient network brings both environmental and economic benefits.

1.5 The Statistical Approach

Instead of using the collected data with the physical conditions of the system, as is used when for example weather is forecasted it is possible to use a statistical approach to find possible correlations [3]. This is especially powerful when trends are wanted.

Measurements are generally done by a reason, and the data is collected to fulfil a certain purpose. But the collected data can also be used to find out more about how a system works. Today there is a trend that more and more data is collected and stored, thus opportunities to reuse the data for different tasks arises. Moreover, as the size of the data collection increases, more accurate correlations can be obtained. Using the statistical approach gives an opportunity to study a system which is large and complex, without ending up with numerical operations which are too computationally heavy.

1.6 Objectives

The objective is to create a tool for analysis of power losses in a network. This tool will bring better understanding of what takes place in a medium voltage net and simplify the procedure of finding measurement errors as well as assist in the process of forecasting losses.

1.7 Project Boundaries

In this project, historical data will be handled in order to analyse the behaviour of a power network and its losses. The project focuses on finding the key parameters that affect the losses. The power network that was studied is a part of the medium voltage grid, owned by E.ON Elnät Sverige AB, called Regionnät Syd. Figure 1.1 shows the area for Regionnät Syd, it approximately includes everything south of the black line. The power losses that are studied in this project is the measured difference in energy from input to output in the medium voltage network i.e. of voltage levels of $40kV$ to $130kV$. The losses due to transformaton of power from the high voltage grid to the medium voltage grid are not included. Neither are losses occuring in the high voltage and low

voltage networks included. The available data is collected on an hourly basis and cover the period July 2012 to December 2013. The variables that will be analysed are; total amount of power fed into the system, different weather parameters, the influence of the input from some specific substations and regional production i.e. wind power, hydro power and thermal power production.

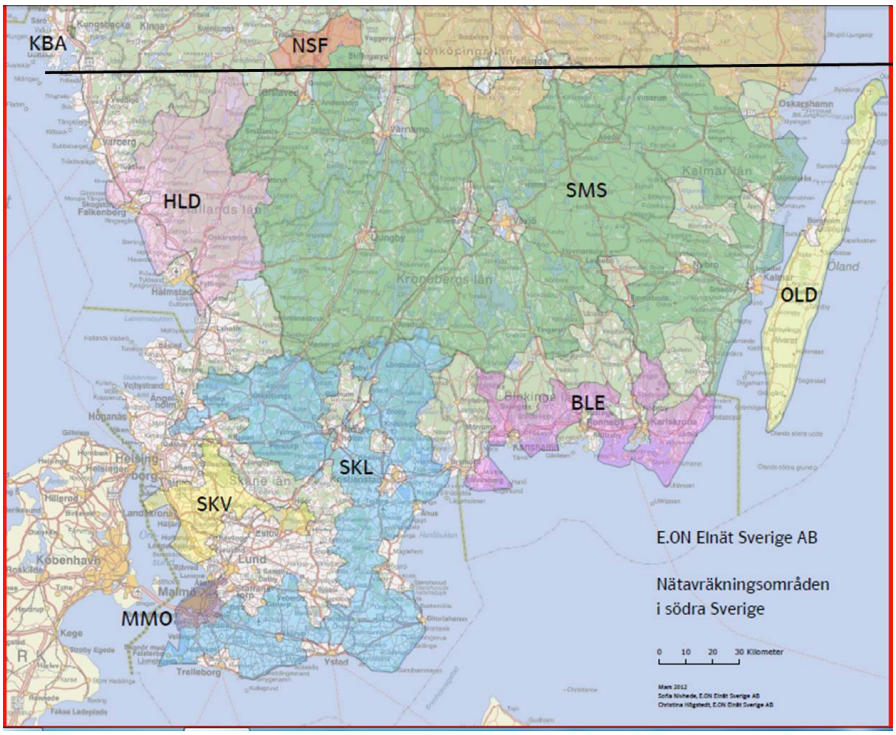


Figure 1.1: Regionnät Syd stretches from the northern border of the low voltage networks SMS and HLD and southwards. [4]

1.8 Report Structure

The report will henceforth go through the parts; Background, Development of the Tool, The Investigated Parameters, The Regression Analysis and then Discussion and Conclusion. The background chapter will cover the basic knowledge of power transmission losses and describe the concept of regression analysis. The next chapter, Development of the Tool, will focus on the method used for the study and for the development of the tool. After that, the chapter The Investigated Parameters will describe the task of choosing parameters and the collection of data in detail. The regression analysis chapter will present the results. All of this will then be discussed and the report ends with presenting the drawn conclusions. In the Appendices the reader can find relevant material, e.g. long tables and figures, that did not fit into the report.

2

Background

This chapter covers the theoretical background which is the platform for the development of the method and the analysis of the results of the study. The chapter consists of two parts; the first part covers the basic knowledge about the electrical power grid and its transmission losses and the second part will cover the theoretical background to regression analysis and explain how to work with it.

2.1 Why There are Transmission Losses

When energy is transported in the distribution wires from the electricity producers, substations or adjacent distribution networks to the end users there is always a power loss. The aim of this chapter is to explain that phenomena and why it occurs. At first it will be described from the viewpoint of a single wire, followed by a description of losses in power transformers and after that losses from the viewpoint of a full network.

2.1.1 A Single Wire

The power transmission losses on a line, between a location A and a location B, depend on the difference in voltage between A and B and the magnitude of the current. The relationship is stated in Equation (2.1) with power losses, P, the voltage difference, V, and the current, I. Rearranging Ohm's law gives Equation (2.2), given that the temperature is constant, which leads to that Equation (2.1) can be expressed as Equation (2.3).

$$P = VI \tag{2.1}$$

$$V = IR \tag{2.2}$$

$$P = I^2R \tag{2.3}$$

This means that the transmission losses depend on two entities; the current and the resistance. In the following sections the different circumstances which affect the current and resistance will be discussed.

Resistance

The behavior of electricity can in this situation be analogous to flowing water. The friction between for example the water and the riverbanks or rocks at the bottom of the river hinders the water from flowing freely. Some of the potential energy of the water is lost to heat. The resistance can be explained as an equivalent to friction and during the transportation some of the electric energy is dissipated as heat. The resistance is dependent on the material, the shape and the temperature of the wire. Materials have a specific conduction capability, where good conduction leads to low resistance. The shape of the wire means both length and cross-sectional area, where the losses increase with transportation distance and decrease with larger cross-sectional area. Finally, the temperature affects the conductivity, which means that as the temperature in the material rises the resistance increases. [5]

Current

The losses are correlated exponentially to the current. The amount of power to be transported can be described by the Equation (2.1), the difference is that here V is the voltage of the line relative to ground. A single wire has a prespecified value for the voltage level, and this level is more or less always held constant, which means that the current has to be increased in order to increase the amount of transported power.

With different combinations of voltage and current the same power output can be achieved. A low current and high voltage difference leads to significantly lower losses than in the reversed case. In the distribution network the voltage levels are limited, from 130kV down to 0.4kV, because of the safety aspect. The closer the wires are to the customer, the lower voltage levels are used and thus higher current. [5]

Reactance

To simplify the explanation to start with it was in the description above assumed that the lines, loads, generators and transformers were purely resistive. That would be the case if a direct current, d.c., was used. Since an alternating current, a.c., is used in most parts of the modern power networks there is something called reactance that has to be taken into consideration as well. Reactance is the property of a device to influence the timing between the current and the voltage. If there is only resistance, the current and voltage reaches maximum and minimum simultaneously. A device can either cause inductive or capacitive reactance.

Inductive Reactance

When the direction of the current is constantly changing, the magnetic field caused by the current will alternate at the same frequency. This continuously changing magnetic field will in turn induce another current in the wire. This induced current will work in the opposite direction from the original current, causing the oscillation back and forth to lag. In Figure 2.1 the phase shift between current and voltage is 90°, and represents a case with zero resistance and pure inductive reactance. The inductance occurs in coil windings in for example electrical equipment, as well as in the wires for transmitting power.

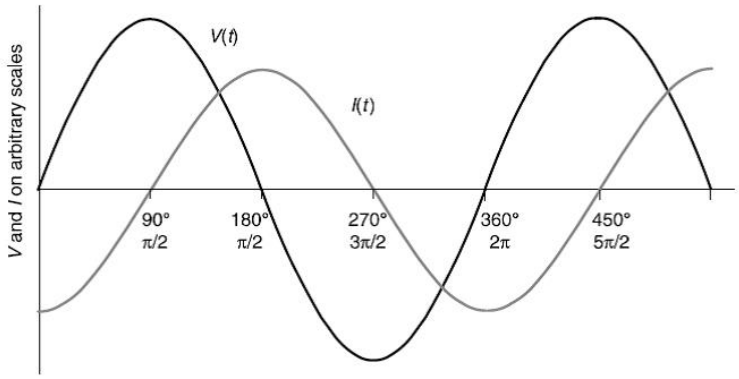


Figure 2.1: When the reactance is inductive, the current lags behind the voltage [5]

The inductive reactance depends on the inductance of the device, L, which depends on the shape of the wire. The shape of the wire refers to for example how many windings a coil consists of. The impeding effect on the current increases as the frequency, ω, of the a.c. increases.

$$X_L = \omega L \tag{2.4}$$

Capacitive Reactance

Capacitance is another type of reactance. A capacitor is a device that has the reversed influence on the current as compared to an inductor; it makes the current lead the voltage instead. The capacitive reactance is inversely dependent on the frequency and the capacitance, C.

$$X_C = -\left(\frac{1}{\omega C}\right) \tag{2.5}$$

The Definition of Power

The definition of electric power stated in Equation (2.1) is a simplified explanation of how the entities voltage, current and resistance interact with each other. It holds true at

an instant and for a purely resistive load. The interaction is more accurately expressed in Equation (2.6). The power varies between a maximum, when both I and V are at their maximum, and zero. The average power is calculated using the average values of I and V as in Equation (2.7).

$$P(t) = V(t) \times I(t) \tag{2.6}$$

$$\bar{P} = \bar{V} \times \bar{I} \tag{2.7}$$

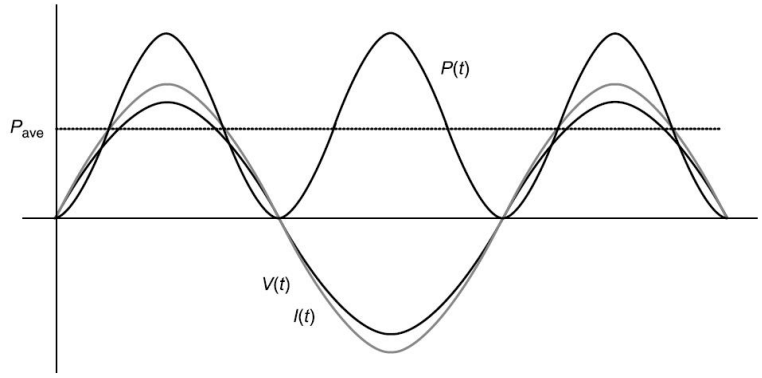


Figure 2.2: When the load is purely resistive, the phase shift between I and V is 0 [5]

But since the load is very seldom purely resistive, the reactance must also be taken into consideration. The reactance is the sum of the inductive and the capacitive reactance and since they work counteractively they partly cancel each other out. Just as it is desirable to hold the resistance of the wires low, the reactance should be contained as well [5]. The reactance causes a phase shift between the current and the voltage which, during parts of the cycle, produces imginary power, or reactive power as it is also called. The equation for reactive power, Q, can be seen in (2.8). A larger phase shift, ϕ, gives a larger Q. The real power, P, is the actual power that can be utilized, for definition of P see Equation (2.9). P is larger when ϕ is smaller. The combination of the two gives the apparent power, S, see Equation (2.10). The apparent power is what counts when congestion on the power lines is considered or when losses are calculated. The reactance does not lead to a power loss directly since the power produced is the real power. But a large share of reactive power causes the current to increase in order to provide the same amount of power. A higher current leads to higher losses.

$$\bar{Q} = \bar{V} \times \bar{I} \sin \phi \tag{2.8}$$

$$\bar{P} = \bar{V} \times \bar{I} \cos \phi \tag{2.9}$$

$$\bar{S} = \bar{V} \times \bar{I} \tag{2.10}$$

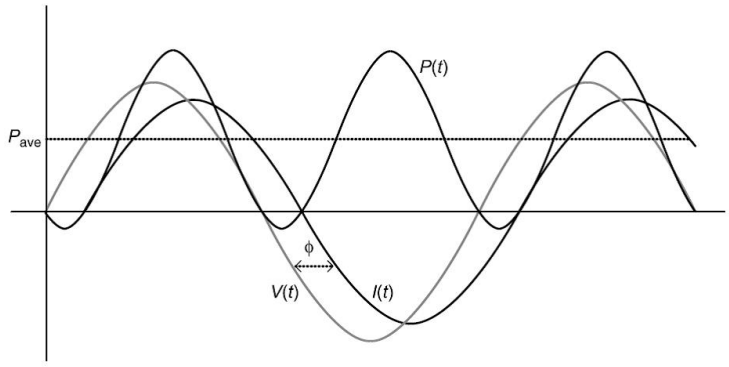


Figure 2.3: The phase shift between I and V creates reactive power [5]

2.1.2 Transformers

The electricity that reaches a customer has on average passed through four transformation steps. [6]. In the transformers, a small percentage of heat will be dissipated likewise the heat dissipation along the wires. One part of it is due to the resistance in the conductor windings, called the copper losses. The larger part is the losses that occur in the iron core, called the iron losses. The a.c in the conductor windings causes the direction of the magnetic field to change rapidly. This implies that iron particles on the microscopic level must constantly realign themselves in the direction of the electromagnetic field. This realignment encounters friction on the microscopic level which leads to heat dissipation.[5]

2.1.3 A Full Network

The basic explanation to what causes the resistive losses in a wire, accounted for in a previous section, allow an understanding of what causes the losses in a full network. However, when studying a network the complexity is increased, since a lot of different factors together influence the total losses. Examples of these factors are weather conditions, types of generation, load, distance between generation and consumption and placement of wires; in the ground or in the air.

Load

A greater load will incur increased current, thus higher losses. The increased current will also rise the temperature in the wire, which affects the resistance of the wire. In that way, the increased resistance makes the losses increase even further.

Weather

Apart from that the weather affects the demand from the users it has an effect on the resistance of the wires as well. When the weather is cold and windy more heat is

transported away from the wires, keeping the temperature of them down. And since resistance increase with increasing temperature, low temperatures and high wind speeds are favorable. [5] For overhead wires; the wind speed, air temperature and the solar radiation is important. For underground cables the temperature, humidity and the heat conductivity of the soil are the features of importance. [6]

According to Klas Roudén, at Svenska Kraftnät, an air temperature of $-5\,^{\circ}\mathrm{C}$ can reduce the losses by 10 % as compared to an air temperature of $20\,^{\circ}\mathrm{C}$. Further Roudén describes how extreme weather such as rime, fog, heavy rain and wet snow can cause a significant increase of corona losses. [7] Corona losses occur when there is a small electric current flowing to the ground through the air. These arcs that discharge into the air occur mostly when the voltage is high, from 130 to 400 kV. [5]

Regional Production

The increasing share of wind power has two effects on the network. The wind power that is produced locally and counts as distributed generation has a shorter path from generation to consumption. It has passed through fewer transformers which also reduces the losses. This holds true as long as the power produced by the wind farms is consumed in the same region [8]. The other effect, and one of the concerns with wind power, is that the induction generators of the wind turbines do not supply, only consume, reactive power [5].

2.2 Regression Analysis

In the previous section the background information about losses in a power line were covered. As described, it is fairly straight forward to say what affect a line, and which losses that specific line has due to the surrounding environment and load. But it gets harder if a larger system is considered, and simulations are needed. Even then, very soon the calculations get too complex, and a whole regional net is too big for this kind of analysis. Therefor this work's approach has been to investigate if it is possible to obtain the same knowledge about the system with a different methodology. Instead of simulating the system, a model built using regression has been used. The model can then be used to find out how strongly correlated different parameters in the system are with the losses, and can potentially also be used to predict the losses in the near future.

This section will cover the mathematical background to regression. It will start off with linear regression and move on towards non-linear regression and multiple correlations, where more than one parameter is allowed to influence the losses in the system. This gives a model that is closer to the reality. It is possible to fit any kind of model to a dataset but to use the dataset to predict new outcomes the model must be good. There are ways to check how well the model fits to the dataset but also how good the predictions are, and the section will cover some of these tests.

Be aware that the statistical approach is a simplification of the complex network, and a statistic approach is a method to find trends and potential correlation, but true correlations can never be found if not the system is built up from the mathematical relations and simulated. Still, the statistical approach has shown that even when simulation cannot be done, it is possible to see how a complex system does interact in the big perspective.

2.2.1 Linear Regression

In the simple linear regression model the dataset consists of two measured parameters, variable x and y, where y is defined as the response variable of variable x. The variables are correlated through a straight line between the data points, the equation of the line can be seen below: [9]

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{2.11}$$

Each data point (x_i, y_i) in the set is correlated by the y-intercept β_0 and the slope β_1. The error in the linear relationship is given by the last term, ϵ_i. This error value is said to have a normal distribution around the regression line. To fit a regression line to the dataset a least-square fit is used. This is done when the quantity Q, in Equation (2.12) is minimised. In other words, when the sum of the distances squared between the collected data point y_i and the fitted value is minimised, or mathematically when the

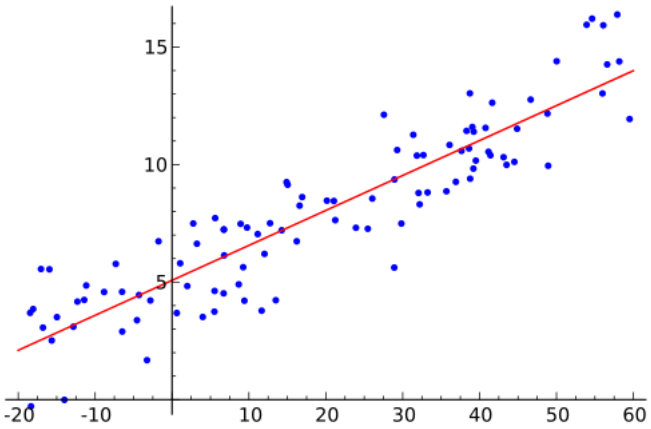


Figure 2.4: An example of a linear regression, where the data points and the line given by Equation (2.17) is seen. Here $\beta_0 = 5$ and $\beta_1 = 0.05$. [10]

partial derivative is zero.

$$Q\left(\beta_0, \beta_1\right) = \sum_{i=1}^{n} \left(y_i - \left(\beta_0 + \beta_1 x_i\right)\right)^2 \Rightarrow \frac{\partial Q}{\partial \beta_0} = \frac{\partial Q}{\partial \beta_1} = 0 \tag{2.12}$$

Here n gives the number of observations in the dataset. The resulting expressions of β_0 and β_1 are known as the normal equations, and can be seen in Equations (2.13) and (2.14).

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{2.13}$$

$$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} \tag{2.14}$$

The hat on the factors indicates that these values are fitted values, calculated from the input dataset, and $\bar{x}$ and $\bar{y}$ are the mean values. Because the error, ϵ_i is normally distributed, so are the fitted parameters. S_{XY} and S_{XX} are defined in Equations (2.15) and (2.16)

$$S_{XY} = \sum_{i=1}^{n} \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right) \tag{2.15}$$

$$S_{XX} = \sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2 \tag{2.16}$$

Thus, the fitted regression line, originated from the dataset is given by the normal equations and the x-values. The hat on y once more indicates that the value is the calculated y-value, as a response of the x-values and the regression model.

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \tag{2.17}$$

The difference between Equation (2.11) and Equation (2.17) is that the β-parameters take their true values in Equation (2.11) whereas in Equation (2.17) they take the fitted

values given by the normal equations. Another difference is that in Equation (2.11) the error value is included. In the fitted regression line there is instead an error variance, $\hat{\sigma}^2$. It depends on the difference between the measured data y_i, and the calculated value $\hat{y}_i(x_i)$.

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{n-2} = \frac{\sum_{i=1}^{n} y_i^2 - \hat{\beta}_0 \sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i y_i}{n-2} \tag{2.18}$$

With the equations presented in this section it is possible to fit a line to any dataset, but that does not mean it is a good model. It is important to see if the variables correlate to each other, and in what way. Thus visualisation of the regression line together with the raw data is important, as seen in Figure 2.4. That is the first check to see if the model works. [9]

2.2.2 How Do We Know That We Can Trust the Parameters?

In the previous section a simple regression model was fitted to the dataset. The output were the $\hat{\beta}$-parameters. In this section the confidence level of these parameters will be explored. This is done by constructing a confidence interval around the regression line.

The Slope Parameter β_1

β_1 is true unknown slope parameter and $\hat{\beta}_1$ is the estimate from fitting the regression line with the given data. The standard error, *s.e.* of the estimate is approximated to be:

$$s.e.\left(\hat{\beta}_1\right) \approx \frac{\hat{\sigma}}{\sqrt{S_{XX}}} \tag{2.19}$$

The denominator in Equation (2.19) shows that $s.e.\left(\hat{\beta}_1\right)$ is decreasing with increasing S_{XX}. From the definition of S_{XX}, see Equation (2.14), it can be seen that if the x_i-values are more spread out, that is, further away from the mean value, the larger S_{XX} is. Thus, a greater spread in x_i results in a more accurate estimate of $\hat{\beta}_1$.

With distribution of both the slope parameter and the error variance the resulting confidence interval is obtained from a t-distribution with $(n-2)$ degrees of freedom. The confidence interval is the interval of in which the true value of the β_1 parameter is, with $(1-\alpha)$ certainty. Usually α is set to 0.05 giving a 95% certainty. A derivation of the confidence interval and more information about t-distributions can be found in [9]. The $(1-\alpha)$ confidence interval for the slope parameter is:

$$\beta_1 \in \left(\hat{\beta}_1 \mp \frac{\hat{\sigma}}{\sqrt{S_{XX}}}\, t_{\alpha/2,n-2}\right) \tag{2.20}$$

Hypothesis Testing for Individual β Parameters

The null hypothesis, H_0, is set up to test the statistical significance of the β-values. If H_0 is rejected, it means that the estimated value $\hat{\beta}_1$ is separated from the expected value.

H_0 and the alternative hypothesis, H_A, for the regression analysis is formulated as

$$H_0 : \beta_1 = 0 \tag{2.21}$$

$$H_A : \beta_1 \neq 0 \tag{2.22}$$

H_A is stating that β is separated from 0, which means it could be a value either above or below 0. Thus, the t-distribution for the expected β-value is two-tailed. The t-statistic is defined as:

$$t = \frac{\hat{\beta}_1}{s.e.\left(\hat{\beta}_1\right)} \tag{2.23}$$

The t-statistics describes the relation between $\hat{\beta}_n$ and the estimated standard deviation. A high value of the t-statistic means that the coefficient β is much larger than the variance and that leads to a low p-value. The value of $\hat{\beta}_n$ is the mean value of the normal distribution.

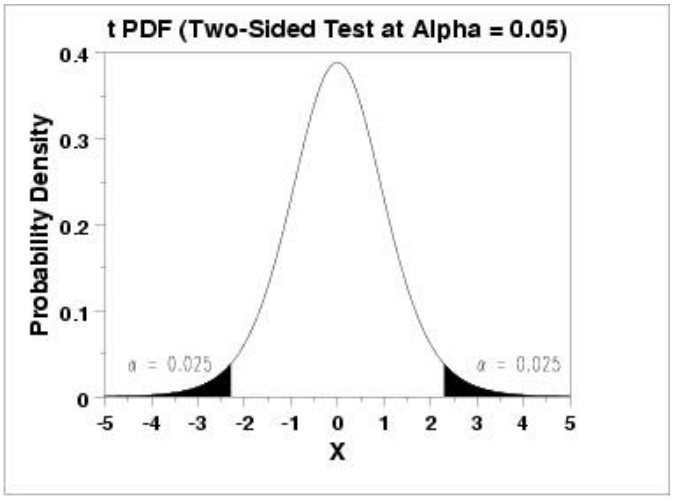


Figure 2.5: Normal distribution with α-value [11]

The α-value describes the critical value for when H_0 should be rejected. If the α-value is 0.05 it means that if H_0 is rejected, there is a 5 % risk that it was rejected even though it was true. The α-value can be visualized, as in Figure (2.5), as an area of the tails of the distribution. The p-value also describes a probability, and can be visualized the same way as α. If the p-value is smaller than α the H_0 is rejected. If the null-hypothesis is rejected it is proof of that there is a correlation between the dependent and the independent variable.

The Confidence Interval

In a previous section the slope parameter was looked into, and an expression of where to find the true parameter was given, with respect to the fitted parameter. Similar can

be done to create a confidence interval for the whole regression line, which this section will do. The fitted regression line gives the correlation between the specific measured value x and an estimate of y, as can be seen in Equation (2.17). But the true line, for a specific value x^* (with no error variance) is given by:

$$\hat{y}|_{x^*} = \beta_0 + \beta_1 x^* \tag{2.24}$$

With the expression for $\hat{\beta}_0$ and $\hat{\beta}_1$ in the normal equations (Equations (2.14) and (2.13)) the expected value for y (i.e. $\hat{y}|_{x^*}$) is witin the interval described by:

$$\beta_0 + \beta_1 x^* \in \left(\hat{\beta}_0 + \hat{\beta}_1 x^* \mp s.e.\left(\hat{\beta}_0 + \hat{\beta}_1 x^*\right)\ t_{\alpha/2,n-2}\right) \tag{2.25}$$

where

$$s.e.\left(\hat{\beta}_0 + \hat{\beta}_1 x^*\right) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{XX}}} \tag{2.26}$$

and by using the t-statistics as for the slope parameter, β_1, with $(n-2)$ degrees of freedom it is possible to do the hypothesis tests on the actual value (Equation (2.24)).

$$t = \frac{(\hat{\beta}_0 + \hat{\beta}_1 x^*) - (\beta_0 + \beta_1 x^*)}{s.e.(\hat{\beta}_0 + \hat{\beta}_1 x^*)} \tag{2.27}$$

Thus it can be seen that the standard error, given in Equation (2.27) depends on the point x^*, and has a minimum when $x^* = \bar{x}$. For a larger number of samples the fitted line gets closer to the true line, as can be seen in Equation (2.26), where the standard error decreases with increasing n, and is zero for an infinite n. From Equation (2.25) it is possible to construct confidence bands around the regression line for all values x. The case $x^* = 0$ gives the point where the line intercept the $y-$axis, i.e. at β_0. From Equation (2.25) the confidence interval of where to find the true β_0 value can be created:

$$\beta_0 \in \left(\hat{\beta}_0 \mp \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}}\ t_{\alpha/2,n-2}\right) \tag{2.28}$$

As for the slope parameter the confidence bands for the regression line can be visualised, and an example is shown in Figure 2.6. [9]

Now the fit of the parameters has been examined, and confidence intervals have been drawn up. The confidence interval does give the interval where the true parameter values can be found, or where the true regression line can be found. The smaller the interval is, the more reliable is the values of the fitted parameters. This does not mean that all collected points will be within this interval, it does only give the interval to where the regression line could be, within 95 % certainty (if $\alpha = 0.05$). In the next section the prediction interval will be derived.

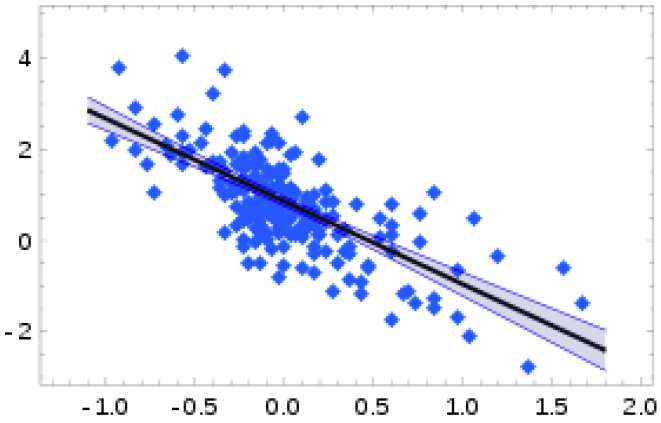


Figure 2.6: Data with the regression line together with the confidence bands. [12]

Prediction Interval

When a prediction of the y-value is made from the fitted regression line, not only is the variance from the fit taken into consideration but the variability of the error term in Equation (2.11) is also taken into consideration. Thus, within $(1 - \alpha)$ confidence, for a value x^*, not included in the dataset creating the model, the $y|_{x^*}$ value is within the limits:

$$y|_{x^*} \in \left(\hat{\beta}_0 + \hat{\beta}_1 x^* \mp \hat{\sigma} \sqrt{\frac{n+1}{n} + \frac{(x^* - \bar{x})^2}{S_{XX}}}\, t_{\alpha/2,n-2} \right) \tag{2.29}$$

As for the confidence band the minimum prediction interval is at $x^* = \bar{x}$, and does increase for values further from the mean. But, even for an infinite n there will be an interval as the variance of the error term, $\hat{\sigma}^2$ given in Equation (2.18), will always be present. [9]

2.2.3 Residual Analysis

A straight line can be fitted by the regression analysis to any set of data. That does not mean that there is a linear relationship. The confidence and prediciton bands explained in the previous section do give some insight if this is the case, but can not only be considered. In this section other methods to check the reliability of the model will be accounted for.

Residuals

The definition of the residuals, e_i, are the difference between the data value and the value given by the fit, for all the values included in the dataset, $1 \leq i \leq n$:

$$e_i = y_i - \hat{y}_i \tag{2.30}$$

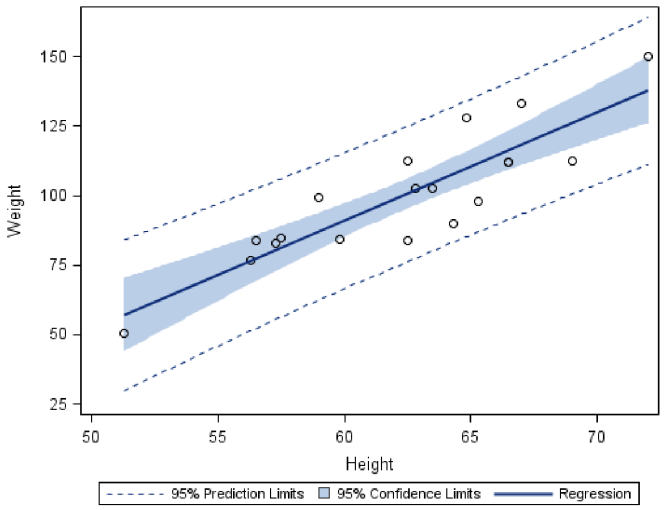


Figure 2.7: The prediction interval is outmost. The confidence interval is given by the light blue area around the regression line. [13]

The residuals of a fitted regression model are important when evaluating the model, and if the assumption used are valid.

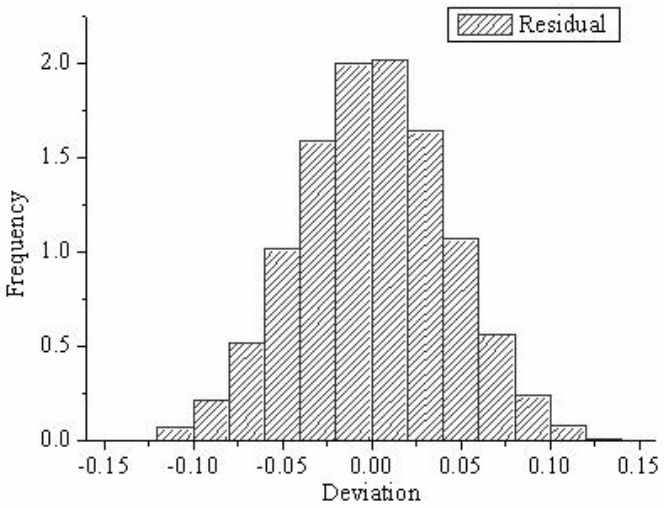


Figure 2.8: Histogram over the residuals. [14]

Checking for Outliers

Moreover, by studying the residuals, potential outliers of the model can be found. Outliers are data points that lies far away from the fitted regression line. The behaviour of the error variances can also be studied, for a good model the variances are fairly constant

in size and are normally distributed over the whole data span, as can be seen in Figure 2.8. It is best to find the behaviour of the residuals by studying a plot, but potential outliers can be identified by calculate the ratio $\frac{|e_i|}{\hat{\sigma}}$, and if that is greater than three the point is a potential outlier and further action might need to be considered.

$$\frac{|e_i|}{\hat{\sigma}} > 3, \text{ for outliers} \tag{2.31}$$

Even though outliers are identified, they should not automatically be removed from the dataset. The amount of outliers and their affect on the fitted regression line need to be analysed before careful consideration of their future place in the dataset. [9]

Homoscedastic, Unbiased and Normally Distributed Residuals

If the positive and negative residuals are grouped together in different groups there are indications that a the model for the fit is not correct. If there is any kind of pattern of the residuals at all, the model needs to be looked into. The residuals should be spread out randomly, but even. See Figure 2.9 for examples of patterns to look out for, where (a) is the ideal case.

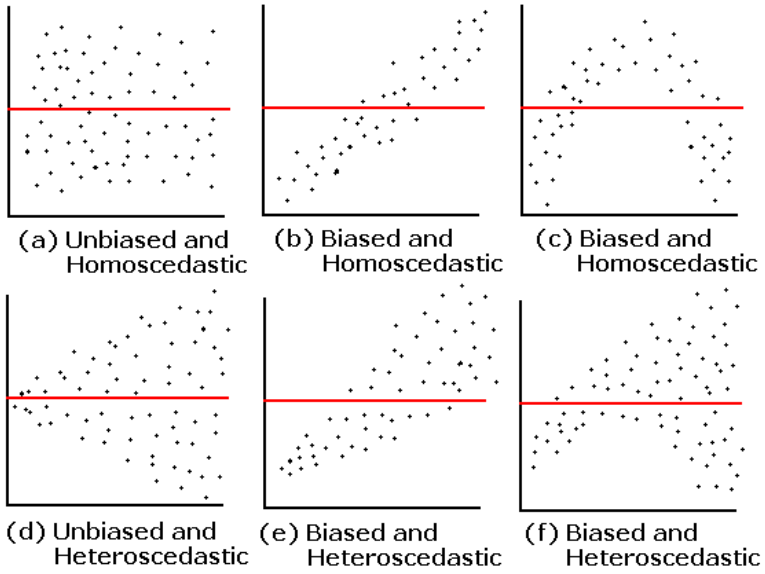


Figure 2.9: Looking at residuals, where (a) is the ideal case. [15]

Futhermore, a normal probability plot of the residuals also called a normal score plot, shows if the error terms are indeed normally distributed, as assumed. If everything is as it should, the normal score plot will look similar to Figure 2.10. [9]

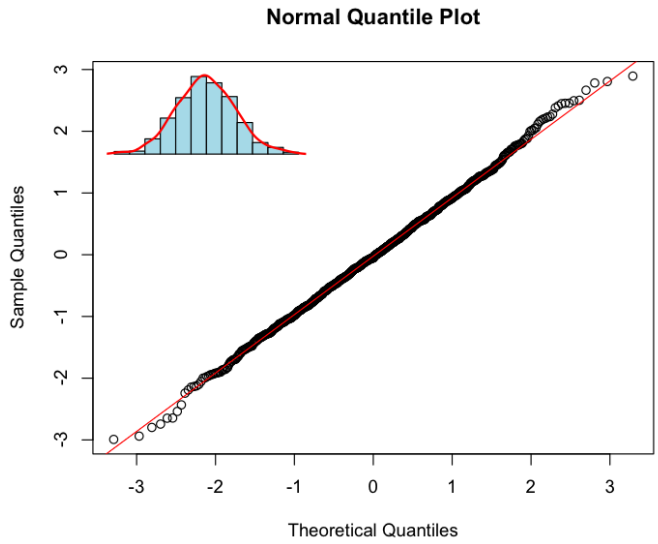


Figure 2.10: The normal score plot testing if the residuals are normally distributed as assumed, which they are in this figure. [16]

Autocorrelation

In time series, that is a dataset that has been accumulated over time, it is very likely that the data measurement a time step earlier does correlate to the data measurement a time step after. In other words, there is a natural order in the observations. An example is the power load in the grid, the load does change throughout, but the change is smooth, thus there is most likely a correlation from measurement to measurement. In regression analysis this can cause problems if the correlation between data measurements are not accounted for. The most common way to find out if there is a problem is to look at the residuals and then search for autocorrelation. Autocorrelation is the correlation between adjacent residuals; if there is a structure in the data, and it looks like the residual at a time step is correlated to the residual at a close by time step. An illustration of autocorrelation is given in Figure 2.11 where the first figure does not suffer from autocorrelation but the second and third does. [17]

Why Autocorrelation is Important to Look Out For

In Equation (2.11) the error term included, ϵ_i, is assumed to be normally distributed and random. When there is autocorrelation in the results, this is not the case. The error terms can still be normally distributed as in Figure 2.11 but they are not random. This results in that the significance tests, for example the p-values, together with the confidence- and prediction intervals are no longer reliable. With autocorrelation the regression model gives too small confidence- and prediction intervals, so in reality these intervals should be larger. Thus, if the model shall be used to predictions, the validity of the model can be questionable. [17]

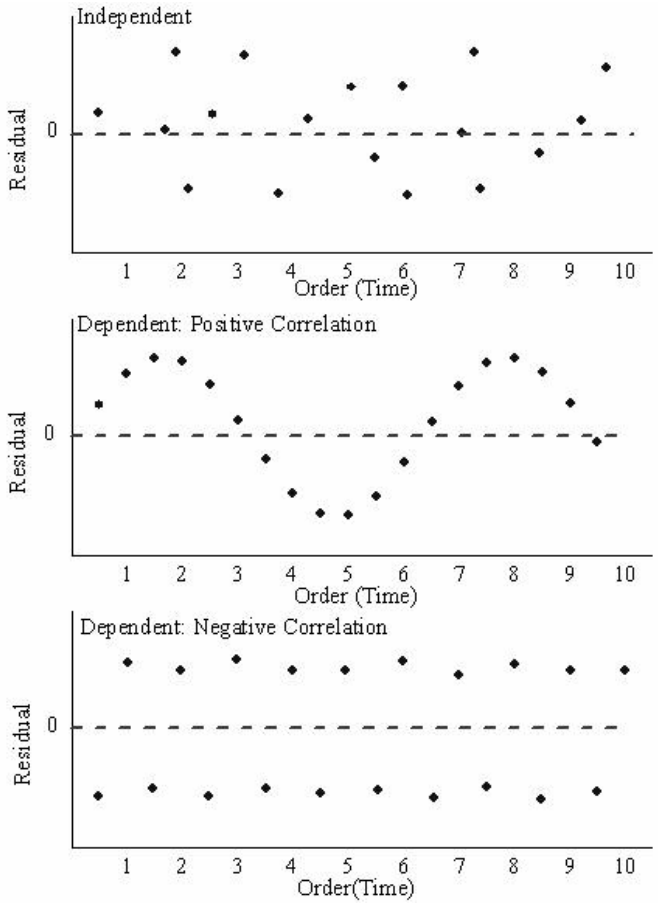


Figure 2.11: Autocorrelation check. The top figure suffer no autocorrelation. [14]

How to Get Around Autocorrelation

Autocorrelation can vanish from the model if the appropriate parameters are added to the model. These parameters are most likely time series. It can be very difficult to find all the appropriate parameters, resulting in no autocorrelation in the residuals at all, but in theory it is achievable. For example weather parameters can be included to explain the hourly and seasonal changes in other parameters.

Another way to correct the residuals from the autocorrelation is to correct the used input parameters from time series dependence. For example, if income is a parameter, and the input income is taken from many years, the used input could be the inflation corrected income. Then time itself, can of course be a parameter, to account for time trends and indicator variables can be added for seasonable effect. A final suggestion of how to get around autocorrelation is to consider lagging. Subsequently $y(t-1)$ is used as a dependent parameter when the model is fitted for the $\hat{y}(t)$ calculation. But, then the model itself is a differentiating model, and it is not always an applicable result. [17]

To Identify Autocorrelation

The first and simplest way to see if there is any autocorrelation in the residuals is to look at the index plot of the residuals, as seen in Figure 2.11. Another simple way is to try to figure out if there is an autocorrelating trend is to plot the residual versus the lagged residuals, as seen in Figure 2.12. A diagonal pattern does imply a possible autocorrelation. [18]

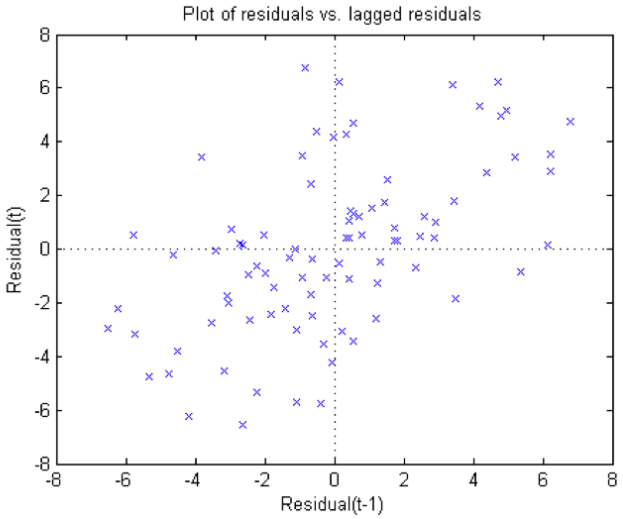


Figure 2.12: Lagged residuals: the residual of time t is plotted against the value one time step before. [18]

Two test to check if there is autocorrelation is the Durbin-Watson test and to plot the autocorrelation function, the ACF. Another test is to do a run test. The first two tests still assume that the error is independent and normally distributed, whereas the last test does not. In the run test the number of times the residuals swaps from positive to negative values is calculated. This number should be t-distributed if there is no autocorrelation. [17]

2.2.4 Linearising Data

Even if a dataset in a plot shows a non-linear relationship, linear regression methology can still be used when the data is analysed. A non-linear relationship can nearly always be linearised. For example:

$$M = \gamma_0 \, e^{\gamma_1 x} \epsilon_i^* \Rightarrow \ \ln(M_i) = \ln(\gamma_0) + \gamma_1 x_i + \epsilon_i \tag{2.32}$$

Then, comparing to Equation (2.11):

$$y_i = \ln(M_i), \ \beta_0 = \ln(\gamma_0), \ \beta_1 = \gamma_1, \ \epsilon_i^* = e^{\epsilon_i} \tag{2.33}$$

Thus, by including linearisation of the data the regression line can be fitted as a stright line to the linearised data, and finally transfered back with the fitted line with the confindence and prediction bands. [9]

2.2.5 Multiple Correlations

When the simple regression model does not apply to the data, it is still possible to use regression to find the correlations in the input data. This could be the polynomial dependence in x_i. By using more than one input variable it is possible to find multiple correlations in the data. Equation (2.11) is expanded with a $\beta_j x_{ji}$ term for each extra input variable. Where $1 < j < k$ and k is the total number of input variables. For $k = 1$ it is a simple regression, as expressed by earlier sections. In a quadratic regression model an input variable is the square of another input variabe, for example $x_{2i} = (x_{1i})^2$. Matrix formulation simplify the expressions of the linear regression equations for larger k.

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \ \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{pmatrix}, \ \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix}, \ \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix} \tag{2.34}$$

$\mathbf{Y}$ and ϵ are $n \times 1$ matrices, $\mathbf{X}$ an $n \times (k+1)$ matrix and β is a $(k+1) \times 1$ matrix. $(k+1)$ comes from the fact that there is an intercept included. In matrix form Equation (2.11) becomes:

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon \tag{2.35}$$

As the error is multivariate normally distributed, the covariancce matrix is $\sigma^2 I_n$, where I_n is the identity matrix. Thus the response variable has the normal distribution $\mathbf{Y} \sim N_n\left(\mathbf{0}, \, \sigma^2 \mathbf{I}_n\right)$. The normal equations can be found by minimising the quantity Q similar to Equation (2.12). In matrix form the normal equations are:

$$\hat{\beta} = \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{Y} \tag{2.36}$$

Where the transpose and the inverse of matrix $\mathbf{X}$ is included. Similar to the simple regression model, it is possible to get confidence and predictions intervals to the multiple linear regression model. The regression model gives a fitted value of $\hat{y}|_{\mathbf{x}^*} = \mathbf{x}^{*T}\hat{\beta}$, where $\mathbf{x}^*$ is a set of variables ($\mathbf{x}^* = x_1^*, \ldots, x_k^*$). If $k = 1$, $\hat{y}|_{\mathbf{x}^*}$ is the regression line (or curve if there is a non-linear relationship between the variables), and if there are two variables $\hat{y}|_{\mathbf{x}^*}$ is a surface mapping out the fit of the regression model. The actual value of the response variable for $\mathbf{x}^*$ is $y|_{\mathbf{x}^*} = \mathbf{x}^{*T}\beta$. This true value is with $(1-\alpha)$ confidence within the boundaries given by:

$$y|_{\mathbf{x}^*} \in \left(\hat{y}|_{\mathbf{x}^*} \mp s.e.\left(\hat{y}|_{\mathbf{x}^*}\right)\ t_{\alpha/2,\, n-k-1}\right) \tag{2.37}$$

The standard error of the fit is given by:

$$s.e.\left(\hat{y}|_{\mathbf{x}^*}\right) = \hat{\sigma}\sqrt{\mathbf{x}^{*T}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{x}^*} \tag{2.38}$$

Similar to the linear regression model, when the prediction interval is calculated the variable $\mathbf{x}^*$ is not part of matrix $\mathbf{X}$. The prediction interval of where $y|_{\mathbf{x}^*}$ can be is given, with corresponding standard error, below:

$$y|_{\mathbf{x}^*} \in \left(\hat{y}|_{\mathbf{x}^*} \mp s.e.\left(\hat{y}|_{\mathbf{x}^*} + \epsilon\right)\ t_{\alpha/2,\, n-k-1}\right) \tag{2.39}$$

$$s.e.\left(\hat{y}|_{\mathbf{x}^*} + \epsilon\right) = \hat{\sigma}\sqrt{1 + \mathbf{x}^{*T}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{x}^*} \tag{2.40}$$

Residual Analysis

How to do the residual analysis, and what to look out for is similar to the simple linear regression model. The equations looks different to allow for the increased number of variables. The residuals are in matrix form $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}$, giving $SSE = \mathbf{e}^T\mathbf{e}$ and the error covariance to be $\hat{\sigma}^2 = SSE/\left(n-k-1\right)$. The fitted regression solution $\hat{\mathbf{Y}}$ can be expressed in terms of the observed data $\mathbf{Y}$, and the 'hat matrix' $\mathbf{H}$.

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y} = \mathbf{H}\mathbf{Y} \tag{2.41}$$

With the 'hat matrix' and the identity matrix I_n the residuals can be expressed as

$$\mathbf{e} = \left(\mathbf{I}_n - \mathbf{H}\right)\mathbf{Y} \tag{2.42}$$

In the previos section about the residual analysis Equation (2.31) is given. In matrix notation the standardised residual for the ith residual is given by:

$$e_i^* = \frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}} \tag{2.43}$$

Where h_{ii} is the i^{th} diagonal element of $\mathbf{H}$. Potential outliers are when $e_i^* > 3$, as before, and as the value of h_{ii} is small, the above equation can often be approximated to Equation (2.31). [9] Another importance of h_{ii} is the measure of leverage of the point.

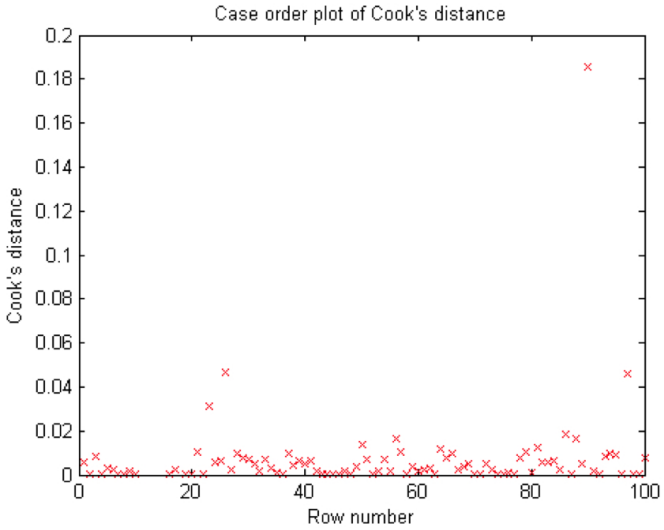


Figure 2.13: A measure of the Cook's distances for a dataset of 100 points. It can be seen that there is one point that has a much larger value than the rest. This point is both a potential outlier and a point with high leverage, thus probably should be removed. [18]

With a high value the leverage is high, hence the risk that the fitted regression line gets distorted due to this point is high. Cook's distance combine the leverage of a point and the outlier measure. Hence the distance is a measure of influence of the observation to the fitted model. [17] Figure 2.13 shows an example of a Cook's distance plot, with one point that stands out. This point has both a high value on e_i^* and a high leverage, thus the risk that this point affect the regression model is therefor high, and is strongly suggested to be removed from the dataset. [18]

2.2.6 Correlations Between the Variables

When performing a multiple regression analysis it is important to check whether the parameters in matrix $\mathbf{X}$ truly are independent. If they are not, the results will be more difficult to interpret. A parameter may then seem to be correlated to the variable of interest when it is actually only correlated to one other parameter which in turn is the true cause for increase or decrease of the variable of interest.

A correlation matrix shows if there are correlations among the independent variables. One way of measuring correlation is the sample correlation coefficient, r, defined below.

$$r = \frac{S_{XY}}{\sqrt{S_{XX}}\sqrt{S_{YY}}} \quad (-1 \leq r \leq 1) \tag{2.44}$$

The definitions of the S_{ij} functions can be found in Equations (2.15) and (2.16). The coefficient of determination, R^2, discussed in next section, is linked to the sample correlation coefficient as $R^2 = r^2$.

For $r = 0$ it is said that there is no *linear* correlation between x and y. For a non-linear relationship x and y can visually clearly be correlated but r can still be zero. For $r > 0$ there is a positive association between x and y. Then an increase in x results in an increase in y. The case is the opposite for $r < 0$. [9]

When r exceeds a critical value the parameters might be considered dependent. There are no exact rules for what critical value that is right to consider, and the critical value varies depending on whom is asked. Some say a good critical value for r is $\sqrt{0.5} \approx 0.7$ [19] and others say $\sqrt{0.8} \approx 0.9$ [20].

2.2.7 Variable Selection Method

The variable selection for regression analysis is the hardest part in regression model building. The model should not include all parameters that are available, but only those that are correlated to the dependent variable. There are several different methods for variable selection, where stepwise selection is one of the most commonly used. Despite the weaknesses of the methods, statisticians still use them [20]. Probably since there are few other ways of selecting the variables, when the amount of variables are large, and since the method is considered to be sufficient despite its weaknesses.

The stepwise selection of variables is a modified version of the forward selection method. By choosing different set-ups for $\mathbf{X}$, several alternative models are created. These are compared by a selection criteria, which most often is R^2, and one model is chosen above the others. This is done stepwise, where a higher number of variables are added over time. The stepwise method avoids testing every possible combination of variables. At the same time it avoids the problem occurring in the forward selection model, where a variable stays in the model just because it was not discarded in an earlier step. In the stepwise optimisation of the model, the relevance of a variable is evaluated again and again. [21]

The problem with the stepwise selection method is that it may lead to an over-fitted model. A variable may contribute to lowering the error, even if it has no true influence over the dependent variable. The model is trained on the dataset but the prediction performance is poor. There is also a risk that the results are not entirely reliable. E.g. the p-values don't have the proper meaning, the confidence intervals are falsely narrow and the $\hat{\beta}$-values are too big. [20]

To avoid some of the problems caused by the stepwise model optimisation, more steps can be added in the model selection process. What is often neglected by the statistician is the fact that his/her own knowledge is greater than the computer's. The stepwise selection model can be used as long as the results are not trusted blindly. The risk of over-fitting can be reduced by choosing a better selection criteria, like Akaike's information criterion, AIC, instead of R^2. AIC punishes a model with many variables, which R^2 do not [22].

When the regression model has been optimised from the variable selection method the prediction performance can be analysed. The next section will tackle this part.

2.2.8 Prediction Performance

The regression analysis results in a model that describes the behavior of the variable or variables of interest. This model can be very good at describing the outcome in the given dataset, but that does not mean that the true correlations have been found. If a random series of numbers is added as a parameter to the model, this will most likely increase the accuracy of the model, even though the data series is pure noise and has nothing to do with the variable of interest [23]. There is a risk that by adding more and more variables and thereby increasing the complexity of the model, it ends up over-fitted. To avoid over-fitting, the prediction performance of the model is tested. Here, in this thesis, one method will be described.

A part of the data, for example one quarter, is separated from the rest and the model is trained on the remaining three quarters. This gives an opportunity to test the prediction performance of the created model by comparing the mean squared error, MSE in Equation (2.45), of the training set and the test set.

$$MSE = \overline{(y_i - \hat{y}_i)^2} \tag{2.45}$$

The error is always greater for the test set, which is natural since the model is adjusted to the training set. When developing the model it is important to see that the mean squared error of the training set decreases. But the mean squared error of the test data have to decrease simultaneously, otherwise the model will end up over-fitted. [23]

The process of adding variables and testing the goodness of fit of the model as well as the prediction performance is an iterative process. A new test dataset has to be excluded from the training data each time. Otherwise there is a risk that the model is only adjusted to the two datasets and still does not describe the actual relationships between the dependent variable, $\mathbf{Y}$ and the independent variables $\mathbf{X}$. [23]

This concludes the background chapter of the thesis. In the next section the software development will be considered.

3

Tool Development

One of the aims in this project was to create a platform, in which Energicontrolling at E.ON Elnät could perform statistical analysis on their collected data, but also to use other data sources and merge to their analysis. The previous chapter went trough the necessary background information needed for the project. This chapter describes how the tool for investigating correlations was developed. It starts by introducing Lavastorm Analytics, the software that was used to create the tool, and continues with describing the working process stepwise. The data collection will only be described briefly since it is more thoroughly described in the following chapter.

3.1 Lavastorm

The big data analytics software Lavastorm was a newly addition to the Energicontrolling's software programs and used in the development of the statistical platform.

Lavastorm uses graphs which makes it easy to follow each step in the data handling and data analysis. Each process consists of one or a group of nodes, where the calculations take place. In the graph it is possible to specify which nodes to run, and the outputs from previous nodes are saved in the memory. Thus, it is possible to develop a single node, without having to load the input file for every test run. Lavastorm provides a good overview of the data collection, data treatment and the building of the regression analysis model. [24] The straightforward and intuitive interface makes the software ideal for the purpose of this project; a variety of users can easily get updated to what's happening in the graph and use it or develop it further. With an extra add-on, Lavastorm Analytics platform can supports statistical calculations. Then an R-node is available, and within this node the open source programming language R is used for the calculations [25]. For this project a server was used, where the added on R-node package was accessed. This resulted in the Lavastorm program called on the server when the specific nodes were run.

3.2 Data Collection

Data was gathered for the selected parameters of interest. The data was modified to an overall format and the quality of the data was investigated. The data format for the regression analysis consisted of a single matrix. One column specified the time for the measurement and one column specified $\mathbf{Y}$, and the rest of the columns consisted of $\mathbf{X}$. See Section 2.2.5 for a reminder of the definition of both $\mathbf{X}$ and $\mathbf{Y}$. Each row in the data matrix is a specific measurement. In Lavastorm all data collected was modified to fulfil the correct format. The data handling and the regression analysis were done in separate graphs, mainly so that both graphs could be developed separately.

The input data in this project was loaded into the Lavastorm graph both from normal Excel files (.xlsx) and .csv files. If the data was spread out on several sheets in the Excel document, an import-node for each sheet was necessary in Lavastorm. One of the big advantages with Lavastorm is the data handling. It is easy to merge different sources of data in different formats, and thus it was possible to produce data in the format wanted for the regression analysis.

For each step in the process of modifying data a separate node was used, making the process easy to follow. In a final step in the data formation graph all data was merged together and sorted by the time stamp. This resulted in some data was not used, as all parameters needed to have data for each time stamp in the final data matrix.

3.3 Visualisation

Before performing the regression analysis, the data was plotted to visualise the correlation between the dependent variable y and one independent variable x_i. In this report the dependent variable is the losses in the distribution network Regionnät Syd. The difficult part with regression analysis is not to produce results, but to interpret them. Viewing the diagrams was a good way to be able to better understand the results from the upcoming regression analysis. The diagrams also gave a hint on whether there was a possible dependency between the variables. If such dependency did not exist, the variable might be redundant and should then be excluded from the dataset for the regression analysis. It's hard to know however, since that x_i in collaboration with another variable x_j may affect y, even though there is no correlation between only x_i and y. The variable should also be viewed as a time series, i.e. by creating a plot of the independent variable vs time. A variable may be good to include in the model as a means to manage an eventual problem with autocorrelation, even though it is not a causal variable.

Lavastorm provided several visualisation nodes, which were straightforward and easy to use without software-development needed. But the qualities of these plots were not as good as wanted, as it was hard to customise them. With the R-node in Lavastorm it was possible to develop visualisation nodes as well as customise these plots. The output of the R-node was a plot that either could be visualised in Lavastorm or linked to another node where it could be exported as a picture file.

For each parameter considered, two plots were done. In one of the plots the data

collected was plotted against the measured losses and possible trends were looked for. The second plot showed the time series trends for each parameter.

3.4 Correlation in Data

Section 2.2.6 covered how to check if the included parameters are independent as assumed. As a final test before performing the regression analysis the correlation between the parameters were analysed in a separate R-node. The resulting matrix was sent back to Lavastorm and then published in a .xlsx format.

3.5 Regression Analysis

In the second graph in Lavastorm the actual regression analysis was developed. The graph consisted of nodes before the main R-node doing the regression analysis, and nodes after, visualising the result as well as analysing the result further. In the nodes before the main R-node final alterations of the input data was done, together with additional parameter analysis, such as correlation analysis, mentioned in the previous section. The regression analysis took place in a R-node in the graph. This node was also where the main software development took place.

In the first part of the R-node the regression model was optimised, and in the second part the model was partly analysed. The regression model was in itself optimised in two steps. In the first regression step a base model was created from the data using a linear model function in R. From the visualisation of the data, it was realised that all the parameters were not linearly correlated with the losses. Linearisation of non-linear data is described in Section 2.2.4. To allow for non-linearity in the parameters, and still use the linear regression analysis function in R the input data was modified in a node prior to the main R-node. The $\mathbf{X}$ matrix was expanded with the square of each parameter.

3.5.1 Model Optimisation

In the second regression step the model was optimised using the base model as starting point. To restrict the amount of parameters, to only include parameters that clearly did correlate to the losses, a variable selection method was used. To be more specific; the stepwise variable selection was used, see Section 2.2.7. The function included in R's statistical toolbox, called step, took care of choosing which parameters to include in the model. The selection criterion which R uses to compare the models is AIC, see Section 2.2.7. In the main R-node, the step function added and removed parameters to optimize the model in an iterative manner. Furthermore, in the step function the effect of two variables increasing or decreasing simultaneously is also considered and is assigned a specific parameter. As these combinations, or cross correlations, of the variables come into play the number of parameters grow very fast. Thus, the iterations over all the parameters given and a potential mix between each parameter made the step function the most computational demanding part in the whole graph.

After the two regression steps R had calculated an optimised regression model for the given parameters. The step function allowed for cross correlation terms, but further parameterisation of each individual parameter had to be done prior to the node.

3.5.2 The Resulting Model

The resulting model derived from the regression analysis was presented as a list of the parameters together with the corresponding $\hat{\beta}$ and p-values. The p-values are an indication on whether a β-value, i.e. the correlation between two parameters, is significant or not, see Section 2.2.2. The $\hat{\beta}$-values with the corresponding standard error and p-value were sent back to the main Lavastorm graph from the R-node and then published in an .xlsx format.

3.6 Residual Analysis

When executing a regression analysis, some assumptions concerning the data are made. It has to be checked whether these assumptions were made correctly, otherwise the result of the regression analysis is not authentic. The three assumptions that were made are: the parameters are independent; the residuals are normally distributed with a homogenous variance over the sequence of the independent variable and there is no autocorrelation in the residuals. Section 3.4 covers the first assumption, and the other two assumptions are covered in the residual analysis of the model.

The foundation for the residual analysis was created in the main R-node. In R the fitted regression model is an object, and as part of the object there are several analysis packages including a few figures to help with the residual analysis [25]. From the fitted model the residuals were accessed and returned to the Lavastorm graph for visualisation together with figures for the actual residual analysis. Several different figures were studied to check for autocorrelation, as can be seen in Section 2.2.3.

3.7 Prediction Performance

Due to an awareness of the weaknesses of the variable selection method, it was complemented with prediction performance tests of the model, see Section 2.2.8. Following the procedure of the example in the section the main R-node was given the functionality of dividing the dataset into two sets before performing the regression analysis.

In the main R-node 20% of the data was randomly selected to be parted from the dataset and form a test set. The remaining 80% formed the training set which created the base model first, then the optimised model from the step function. The expected value $\hat{\mathbf{Y}}$ is compared to the actual value of $\mathbf{Y}$ in the test set to check the prediction performance of the model.

This was done by histograms in separate Lavastorm nodes and by calculating the mean square error, see Equation (2.45). The mean squared error value of the datasets are supposed to be as low as possible, and as close to each other as possible. At the same

time, the residuals has to be normally distributed for both datasets. In other words; both the mean square error and the distribution of the residuals was a fast way to investigate the goodness of fit of the model before conducting a deeper analysis.

3.8 Marginal Effect Analysis

A marginal effect analysis was performed to simplify the interpretation of the results, i.e. the $\hat{\beta}$-values. The effect of each factor is represented by several parameters. As an example the effect of temperature is described by the variable Temperature, $(\text{Temperature})^2$ and several combined parameters like for example Temperature·Import.

In the marginal effect analysis each factor was investigated individually. For each investigation a new $\mathbf{X}$ was created. All the parameters not dependent on the factor of investigation were set to their mean value. The parameters dependent on the factor of the study were varied, as the factor itself was varied from its minima to its maxima. For each $\mathbf{X}$ a prediction calculation was performed in the main R-node. The result of the calculation was $\hat{\mathbf{Y}}$ togther with the confidence- and the prediction intervals of the loss rate. Another R-node was used to visualise the result which were sent back to the Lavastorm graph and then saved at the working directory at the server. A figure per factor were produced, analysed and compared to the collected raw data.

This concludes the section about how the development of the tool in Lavastorm and R was done. In the next section the parameters used in this project will be introduced and examined.

4

Investigated Parameters

The previous chapter described how the statistical tool was developed. This chapter will cover the method used for data collection. It will account for how data for the parameters included in the model was gathered and why the parameters were chosen.

4.1 The Input Data

The loss rate is the variable of investigation and the houry measurements forms vector **Y**. Matrix **X** consists of the hourly measurements of the 11 parameters given in Table 4.1.

Table 4.1: Parameters included in the regression analysis, with mean values within the Regionnät Syd.

Parameter	Mean	Unit
LossRate	1.321	%
Feed.In	2.615	GWh/h
WindP	0.168	GWh/h
Temp	8.599	°C
WindSpeed	3.926	m/s
RH	84.46	%

Parameter	Mean	Unit
Precipitation	0.068	mm/h
WindInstEffect	0.230	GW
HydroP	0.131	GWh/h
ThermalP	0.119	GWh/h
Export	0.042	GWh/h
Import	0.035	GWh/h

Where LossRate is the power losses in the region compared to the load. Feed.In is the total power fed into the region; WindP, HydroP and ThermalP are the regional production of wind power, hydro power and thermal power; Temp is the average temperature considered; RH the average relative humidity of the air in the region; WindInstEffect

the installed capacity of wind power production in the region and finally, Import and Export is the import and export to Denmark.

The Workshop

A regression analysis of losses, feed-in power and wind power production was performed and the results from this analysis were presented at a workshop, arranged half way through the project. The aim of the workshop was to gather people at E.ON Elnät Sverige AB with different knowledge about the behavior of the distribution network and its losses in order to find new inputs and ideas to the project. The goal was to make use of this knowledge in order to develop the model and improve its predictive performance. The idea was that the outcome of the workshop would be a list of potential new variables to add to the model as well as guidance to how to retrieve the necessary data for each variable. The inputs from the participants of the workshop can be seen in Appendix A. This list was used for the data collection step in the second half of the project.

4.1.1 The Collection of Data

Data concerning the distribution network, such as total feed-in of energy into the region, the losses in the region, the amount of power produced from hydro, thermal and wind power within the region and feed-in at substations connected to Denmark was available through the data base and data treatment software Pomax. Pomax is used for the daily operations at Energicontrolling. The information on how much energy that has passed through one substation during one hour is saved in Pomax and is therefore available, thus the unit is GWh/h. It was a great advantage to be able to make use of large amounts of data that was already collected to fulfil other purposes.

The total feed-in of energy per hour into the region can be calculated as the sum of values from each substation with an input value together with input from distributed generation in the region. Similarly, to find the total output from the region per hour, the values from all substations with an output value were summed up. The difference between these two is representing the power losses in the system, the dependent variable that was to be studied.

The demanded reliability of the data, for its original purpose, is high. Therefore the quality was deemed as sufficient for this study. There are errors in the data, caused by mistakes occurring in the measurement process or the collection of measurement data, but these errors are estimated to be small enough to not influence the result of the study. The period covered in the data set was July 2012 until March 2014. The reason for not including earlier observations was that the quality of data was lower and the system had overall gone through big changes before the summer of 2012. Data for the period July 2012 to December 2013 was then used, since more recent data was not available for all 11 parameters.

The wind power was the sum of the wind power production in the region, but also the production in neighbouring medium and low voltage networks, if these neighbouring networks consisted of 100% wind power and the production was close to Regionnät Syd's

boarders. The production in the neigbouring networks does affects the overall losses in Regionnät Syd, hence should be included in the wind power production studied.

Weather data were collected from SMHI's database [26]. At their webpage hourly measurements of air temperature, wind speed, precipitation and relative humidity are available for a number of locations. Data for five different locations were downloaded and treated in Lavastorm for all weather parameters. A mean value was calculated from the five values for each hour and for each parameter. To only use one location to represent the weather conditions in the region was considered insufficient and therefore a mean from the values of many locations was used instead. The locations were Helsingborg, Malmö, Skillinge, Öland's southern headland and Ljungby, all marked in the map in Figure 4.1 with red dots.

The data for the installed wind power capacity was gathered from the information sys-

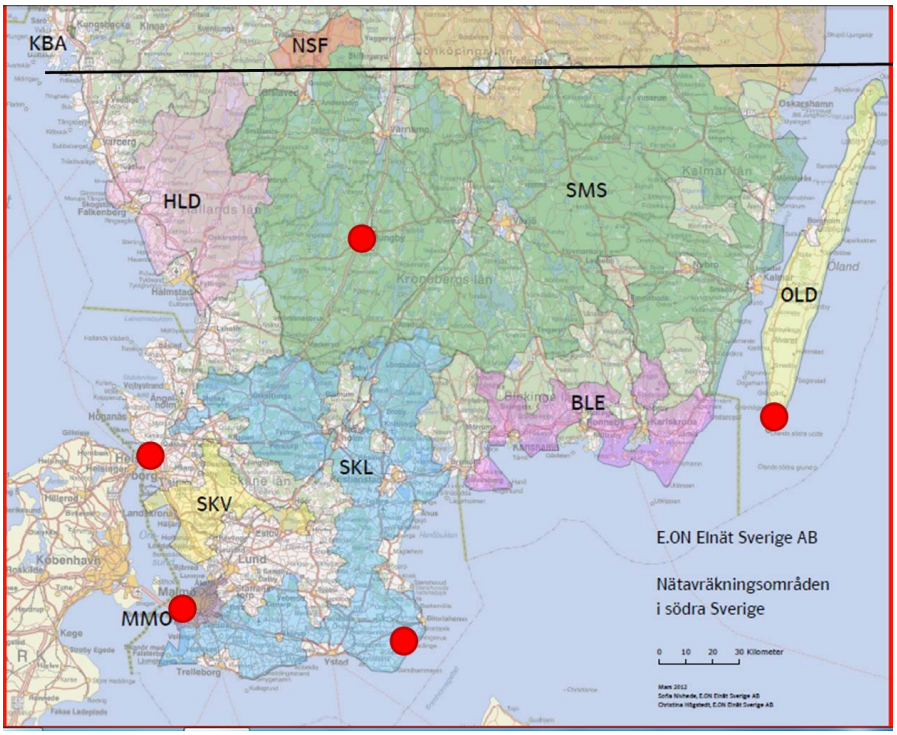


Figure 4.1: The five locations used to represent the weather in the region marked with red dots. [4]

tem SAP, but there were only monthly values available. However, the installed capacity does not fluctuate over short time periods as some of the other parameters do. A value per month was therefore deemed sufficient to represent the steady growth of installed capacity in the region.

4.1.2 Why Parameters were Chosen

In this section indata for each parameter will be analysed. The analysis is based on both a scatter plot of losses as a function of each parameter and two time series plots for each parameter. These are placed in Appendix *B* since there was not room to fit them into this chapter.

Power Losses

The data for loss rate, i.e. losses in relation to feed-in power, is a time series that shows no distinct seasonal dependency. Looking at the time series for a shorter period reveals that the loss rate for one hour is dependent on the loss rate in the previous hour. The time series for both the entire period and the shorter period can be seen in Figures B.1a and B.1b.

Total Feed-in Power and Wind Power Production

The first parameters to be included in the matrix $\mathbf{X}$ were total feed-in power and wind power production. This was since the first task at hand was to investigate if the regression analysis would give the same result as the previously performed investigation, mentioned in the introduction chapter, concerning the correlation between wind power production and transmission losses. In Figure B.3a it is visible that to start with there is a negative correlation between the loss rate and the wind power production. But over a certain point, when wind power production is above around 0.35 GWh/h the correlation is instead positive. A more thorough study of the parameter showed that most data points in the area above 0.35 GWh/h are from October-December 2013.

Total feed-in power was included since the losses are very strongly correlated with it. This is visualised in Figure B.2a where even the loss rate increases with increased feed-in power. This is due to the exponential relationship between the losses and the current.

Weather

The decision to include weather parameters was done partly due to the fact that some weather conditions may have an impact on the losses. These impacts are described in the background chapter, Section 2.1.3, and were also mentioned during the brainstorming session at the workshop. Some weather parameters were not expected to have any influence over the losses directly, but including the parameters in the data set matrix could help reduce the autocorrelation problem, as discussed in Section 2.2.3. The autocorrelation problem was discovered in the model with two parameters, wind power production and feed-in power, that was presented at the workshop.

The weather conditions of interest were temperature, wind speed, precipitation and relative humidity of the air. Temperature is negatively correlated with the loss rate, see Figure B.10a. This is expected since the feed-in power increases as the power demand increases which is correlated to cold weather. Wind speed has a neative correlation

with the loss rate, see Figure B.5a. This is expected since the wind speed is strongly corrrelated to the wind power production as well as the fact that the wind transports heat from the wires which lowers the resistance. The precipitation and the relative humidity of air seems to have a small or no effect on the loss rate, see Figures B.11a and B.12a.

Installed Wind Power Capacity

The grid is constantly being developed, and one thing that causes changes in the system is the development of new wind power capacity. Adding installed wind power capacity as a parameter gives better information on how wind power affects the system. It is then possible to not only consider the amount of wind power that is produced but also how big share of the available capacity that is being utilized.

The data for the installed wind power capacity is visualised in Figure B.4. (a) shows the data for the losses as a function of the parameter. It is difficult to interpret the diagram, the main point of the datapoints is not clear which makes it look like there is no correlation. When using a mean value for the loss rate for each month the correlation is visible. (b) shows the steady growth of wind power capacity over time.

4.1.3 Import, Export and Regional Production

The parameters import and export were discussed at the workshop. They were deemed to have an impact since the power coming from and going to Denmark passes through the distribution network. The correlation test, discussed in the next section, showed that import and export were not correlated to feed-in power. But they proved to be slightly correlated to the losses. If instead looking at their time series, it is seen that the import is high while the export is low and vice versa. See Figures B.8a to B.9c.

The regional production, consisting of the two parameters thermal power production and hydro power production, were of interest since they were expected to have a similar impact on the losses as wind power production since they are also distributed generation. The visualisation showed weak correlations between the parameters and the loss rate. See Figures B.6a and B.7a. The hydro power is positively correlated, which might be explained by studying the time series in Figure B.6b. This diagram shows that the hydro power production has its peak during winter, which means that the high loss rate that depends on high feed-in power coincides with the hydro power production peak. Hydro power production means the production in the region and should not be confused with imported power from the hydro power plants in northern Sweden. There are no dams connected to the hydro power production, the production reacts to the precipitation with small delay.

The scatter plot for thermal power showed almost a flat parabola. Thermal power production has its peak during the coldest winter months, when the feed-in power has its peak as well. High thermal power production can therefore be associated with high losses even though it is not causing the losses to increase. The time series in Figure B.7b shows that the production is high during the winter months only.

4.1.4 Parameters Considered but Not Included

In addition to the weather data that is available for free, E.ON has procured data of hourly measurements of temperature, wind speed, wind direction and cloudiness at 19 locations. The package also included data of wind speed at 100 m above sea level. The origin of that data was simulations of the wind speed at that high level, and not actual measurements. But the data for the parameters cloudiness and wind speed at 100 m above sea level were incomplete, with gaps in the time series and it did not stretch back as far as the data for other parameters. Thus, these parameters were excluded from the data set even though it would have been interesting to study the effects of these parameters.

Data for wind direction was also available at SMHI's webpage. At the workshop it was mentioned that the parameter has an effect on the temperature of the wires. It was said that not only the speed of the wind but also the direction matters. However, the effect of this parameter is a local phenomenon and the parameter is not relevant on a system level.

4.2 Correlated Parameters

Strong correlation between the independent parameters may cause problems, as was discussed in Section 2.2.6. An analysis of the correlation among the parameters gave resulting matrix shown in Table 4.2. The values in the matrix represents the correlation coefficient r.

Table 4.2: The collinearity matrix, where yellow fields indicate strong correlation as $r \geq 0.7$.

	LossRate	Feed-In	WindP	Temp	Wind Speed	RH	Precipitation	WindInstEffect	HydroP	ThermalP	Export	Import
LossRate	1	0.522	-0.275	-0.145	-0.236	0.043	-0.015	-0.140	0.294	0.200	0.280	-0.231
Feed-In	0.522	1	0.005	-0.664	0.148	-0.017	-0.032	0.026	0.549	0.702	-0.104	-0.028
WindP	-0.275	0.005	1	-0.153	0.771	-0.024	0.057	0.379	-0.033	0.039	-0.386	0.307
Temp	-0.145	-0.664	-0.153	1	-0.117	-0.249	0.042	-0.123	-0.444	-0.775	0.393	-0.308
Wind Speed	-0.236	0.148	0.771	-0.117	1	-0.251	0.078	0.110	0.170	0.125	-0.292	0.234
RH	0.043	-0.017	-0.024	-0.249	-0.251	1	0.195	0.002	0.103	0.003	-0.020	0.113
Precipitation	-0.015	-0.032	0.057	0.042	0.078	0.195	1	0.004	0.012	-0.073	-0.021	0.075
WindInstEffect	-0.140	0.026	0.379	-0.123	0.110	0.002	0.004	1	-0.454	0.087	-0.573	0.335
HydroP	0.294	0.549	-0.033	-0.444	0.170	0.103	0.012	-0.454	1	0.369	0.111	-0.062
ThermalP	0.200	0.702	0.039	-0.775	0.125	0.003	-0.073	0.087	0.369	1	-0.360	0.235
Export	0.280	-0.104	-0.386	0.393	-0.292	-0.020	-0.021	-0.573	0.111	-0.360	1	-0.581
Import	-0.231	-0.028	0.307	-0.308	0.234	0.113	0.075	0.335	-0.062	0.235	-0.581	1

The thermal power production is correlated both with the power fed into the system

and the temperature of the air. Wind speed is correlated to wind power production, which is not a large surprise. The values of r are not extreme in these cases since they are on the border of or above the stricter of the two suggested critical values, i.e. 0.7. It is of interest to examine the matter further in future studies, but that will not be done in this thesis. For this work all parameters in the data set are judged to be sufficiently uncorrelated.

5

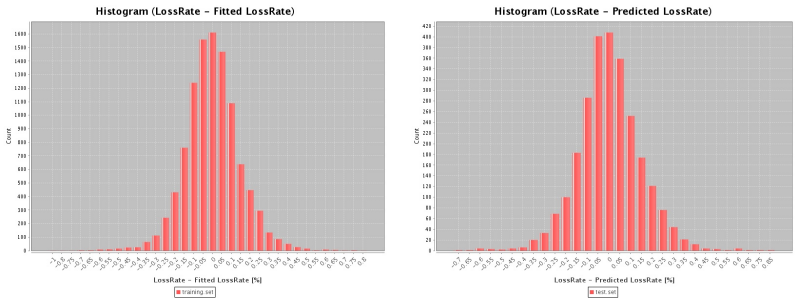
Regression Analysis

Both the development of the tool to perform the regression analysis and the method for collecting data have been presented in the two previous chapters. In this chapter both deliverables and final results from the regression analysis will be presented. There is not one predetermined course of action when it comes to regression analysis. It was consequently necessary to adapt the work along the process through trial-and-error. By first performing a regression analysis and then analyse the results, the decision to alter the data set, and thus redo the regression analysis, or to keep it was made. This iterative process led to the model, which produced the final results.

5.1 All Data

With the limitations in the data from the different sources, the complete data spans from July 2012 to December 2013. With hourly measurements there are a total of 12 975 measurements of each variable. 80% of the data, the training set, was used for fitting the model and 20% was used to check how well the prediction performance of the model is, e.g. the test set of the data. The running time of the regression analysis of the dataset was over two hours. The general background to residual analysis was introduced in Section 2.2.3 and there patterns to look out for were highlighted. In this section the regression model will be evaluated with the outcomes from a residual analysis.

The error in the model, that is the fit of the loss rate compared to the measured values, can be seen in histograms, similar to Figure 2.8. The histograms for the test- and the training data sets can be seen in Figure 5.1. They are very similar in shape, and both looks like they have a normal distribution, centralised at zero, i.e. where the fitted values coincide with the measured values of the loss rate. Thus, from these plots it looks like the model with the parameters given in Table 4.1 does have a relation with the losses in the power system. Next the mean squared error is compared between the



(a) Training (80%) data set **(b)** Test (20%) data set

Figure 5.1: Histograms of the spread of the residuals. A negative number means that the regression model estimate the loss rate to be lower than measured value. The regression model was created with the complete data set. The histograms should be normally distributed with the symmetry around zero.

Table 5.1: Mean squared errors, MSEs, for the training and test data sets. The complete data set was used in the regression.

Data set		**MSE**
Training	(80%)	0.021
Test	(20%)	0.022

training and test data sets. From Table 5.1 it is possible to see that the mean square error is small and as expected smaller for the training data set than the test set.

In Figure 5.2a it is possible to see that the residuals do not seem to suffer from heteroscedasticity. Figure 5.2b shows that the residuals does not have a perfect normal distribution, the tails of the distribution are somewhat heavy. In (d), the Residual versus Leverage plot, it is possible to see that there are no potential outliers. The two plots to the left, (a) and (c) are not clear enough to see if the residuals does have a pattern.

To see if there was any autocorrelation in the residuals, the residuals were compared to the residuals a time step before. The resulting plot can be seen in Figure 5.3a where e_{i-1} is plotted versus e_i. Another method used to identify autocorrelation was the ACF function. The correlation between the residuals of different lags are calculated, and the resulting plot can be seen in Figure 5.3b. From both plots in Figure 5.3 it is straight forward to see that the model suffer from autocorrelation. In Figure 5.3a the fact that the point are accumulated on the line $e_i = e_{i-1}$ shows the correlation in the residuals with the residuals one lag behind, whereas in Figure 5.3b potential autocorrelation has been identified for the first 40 lags, as none is below the dotted blue line. It is also possible to see that the correlation does decrease at longer time lags.

The high correlation in the residuals shows that the model, with the included pa-

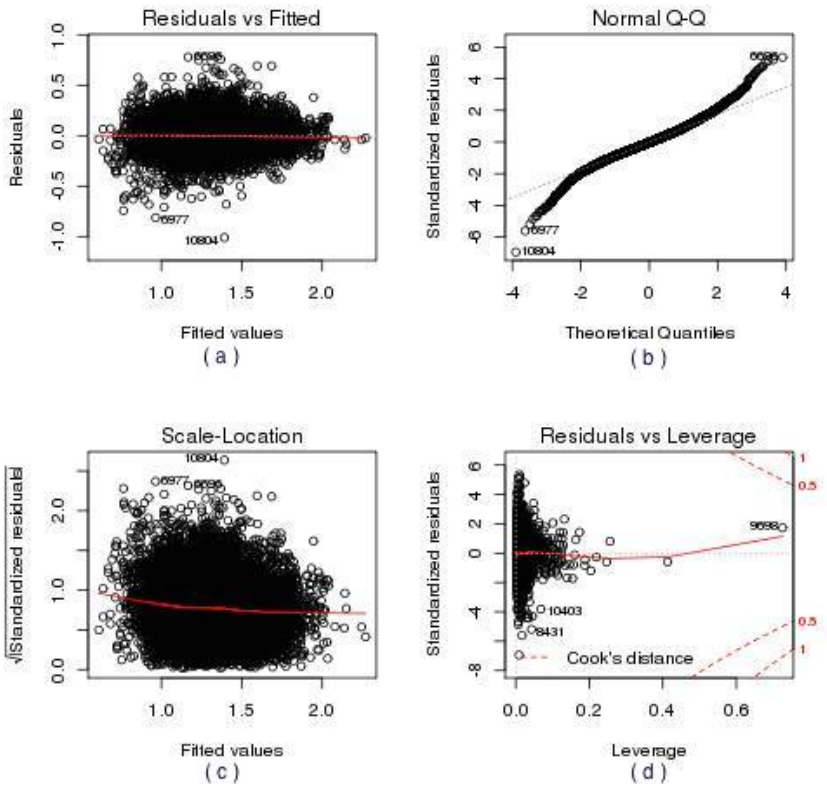
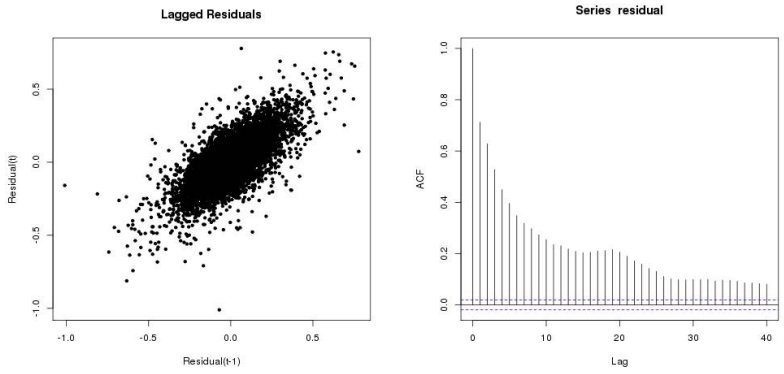


Figure 5.2: Residual analysis of the model created with the complete data set. (b) shows that the residuals are not completely normal distributed and (d) shows that there are no outliers in the set.



(a) **(b)**

Figure 5.3: Residual analysis for the regression model created with the whole data set.

rameters do not adequatly describe the loss rate in the power system studied.

5.2 Data Reduction

The best way to address the autocorrelation in the residuals is to add new parameters, that are time series, to the model. But, that is outside the scope of this master thesis. A quick-fix is instead to cut down the data used in the regression analysis. By selecting a measurement every n^{th} time step the used measurements end up being independent if the value n is large enough. With independent measurements there should not be any autocorrelation. A further developed selection method for this reduction is to split the complete data set into m sets, each set consisting of n values. From each of the m sets a measurement is randomly chosen and the resulting data set is $1/n^{th}$ smaller than the original data set.

In the first attempt $n = 30$ was used, resulting in a decrease from 12 975 measurements down to 433. Figure 5.4 (a) and (b), shows the improvement in autocorrelation with respect to Figure 5.3.

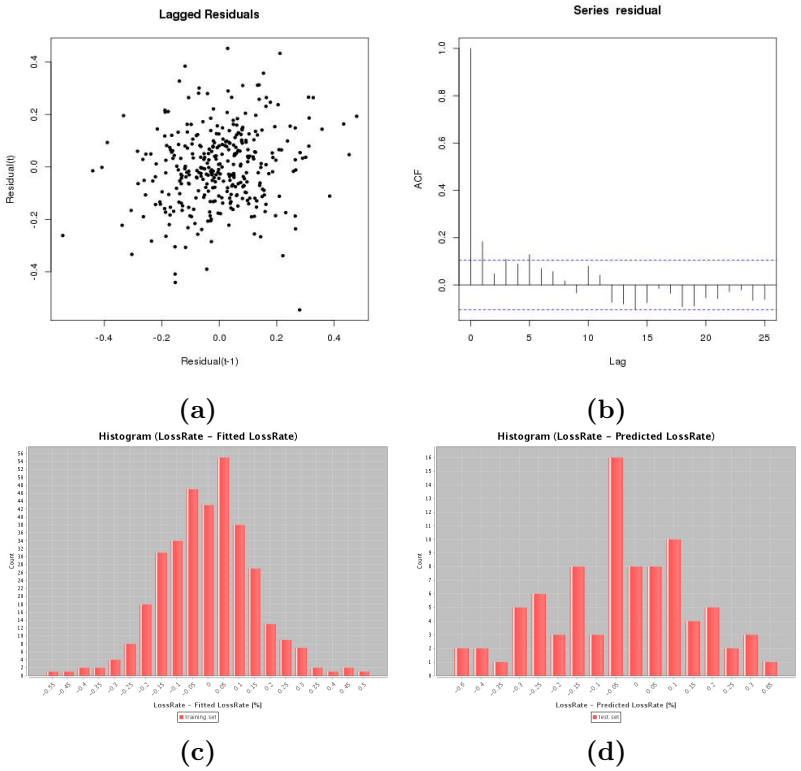


Figure 5.4: Residuals of a regression model where $n = 30$ has been used.

But when the residuals are plotted in a histogram done in (c) and (d), and compared

to Figure 5.1, it is possible to see (from Figure 5.4d) that the model is not feasible to predict the loss rate. In either (c) or (d) are the maxima at zero, and the shape of the histogram of the test set, in Figure 5.4d, does not have a normal distribution. Thus, there is not enough data collected to allow for an autocorrelation correction by excluding data.

However, the autocorrelation can still be improved by a reduction of data, with a lower n-value than 30. When $n = 5$ the autocorrelation is somewhat improved, see Figure 5.5 (a) and (b). Moreover, the test data shows an adequate result, as can be seen in Figure 5.5d.

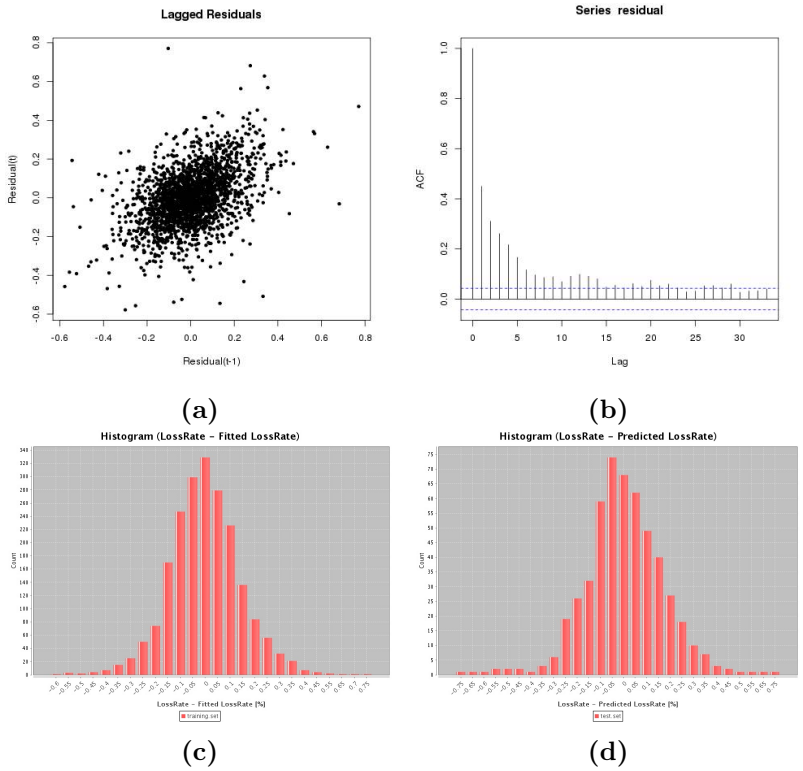


Figure 5.5: Residuals of a regression model where $n = 5$ has been used.

Therefore, when the regression model was analysed further a reduced set of data, with $n = 5$, was used to minimise the autocorrelation problem, even if this caused a much smaller data set. However, this also results in less reliable confidence and prediction intervals, as these will be underestimated due to the autocorrelation still present in the residuals.

5.3 Performance of the Model

The dataset created by semi random selection of one out of five data points and then excluding 20% of that data was used to create the final model. The actual result of a regression analysis is a list of $\hat{\beta}$-values together with their standard errors and p-values. But the complete list is too bulky to be presented here, instead the complete list of the $\hat{\beta}$-values can be found in the Appendix, in Table C.1.

The performance of the model can be evaluated from how accurate it is, i.e. how small the average error is. The MSE for the training set is representing the accuracy of the model and can be seen in Table 5.2. The MSE for the test set is presented in the table as well. That result indicates that the model is not overfitted to the training set, since it is able to produce as small errors for the test set as for the training set.

Table 5.2: MSE of the training set and test set for the final model with the 11 parameters.

Data set		**MSE**
Training	(80%)	0.022
Test	(20%)	0.022

5.3.1 Model with Month Indicators

As a way to investigate whether the seasonal dependencies were described by the included parameters or not, 12 month indicators were added. These were binary parameters which created 12 different models, one for each month, within the model. The factor for each month was added to replace the single intercept β_0 used earlier and was also included in the cross correlations. The hypothesis was that the month indicators could handle the monthly variations that none of the parameters covered, and thus the factors would prevent the parameters from falsely trying to explain the outcome of the dependent variable. By introducing the factors, the correlations between parameters and the dependent variable would hopefully be described more correctly in the model.

The new model with 11 parameters and 12 month factors gave the results presented in Table 5.3.

Table 5.3: MSE of the training set and test set for the third run. 11 parameters and 12 month factors are included in the model.

Data set		**MSE**
Training	(80%)	0.016
Test	(20%)	0.023

The MSE-values for the training set is lower than for the final model described earlier

in this section. The MSE for the test set is slightly higher than in the final model which indicates that it might be overfitted. However, month indicators was not included in the final model for further analysis. This was since the factors only represents some seasonally dependent parameters missing, it does not give any information on which the actual influencing parameters are.

5.4 Marginal Effect

The marginal effect analysis was performed to simplify the interpretation of the results, i.e. the $\hat{\beta}$-values. The analysis were performed for the model that used every fifth value from the data set and all 11 parameters, but without the month indicators. The results from that analysis will be presented in this chapter. The $\hat{\beta}$-values from the model can be found in Table C.1 as already mentioned. In the figures section, Section 5.5, the result of the marginal effect analysis for each parameter is visualised.

The analysis of the results in this chapter is based on what is found in these diagrams. Be aware, as there is still autocorrelation present, the confidence and prediction intervals are underestimated. Figure 5.6 shows the results for feed-in power, wind power production, installed wind power and wind speed. Figure 5.7 shows the results for hydro power, thermal power, import and export. Figure 5.8 shows the results for temperature, precipitation and relative humidity. For all parameters the confidence intervals are quite narrow, thus the uncertainties for the results are low. Note that it does not mean that it is certain that the parameters in reality affect the losses in the way that is presented here. It only means that correlation found is quite certainly right.

As the amount of feed-in power increases, the losses increase a lot. As described in Section 2.1.1 this is expected, since more power on the lines increases the current and thereby the losses. The effect of feed-in power is seen in Figure 5.6a.

Wind Power Production within the region has a negative correlation with the loss rate, to start with, and after a certain point the relationship is the opposite. Since the wind power production is not correlated to the feed-in power it is more certain that it is the wind power production that causes the effect on losses. See Figure 5.6b. For the larger part of the hours included in the study the wind power production seems to be decreasing the losses. This is visible in the underlying scatter plot where it is apparent that the density of data is higher in the left end of the wind power production range than in the right end. The result is backed up by what was mentioned in Section 2.1.3. It was said that regional production, such as wind power, can decrease the distance between production site and end user and thereby the losses.

The amount of installed wind power has negative correlation with the loss rate, to start with, and after a certain point the relationship is the opposite, just as for wind power production. This can be seen in Figure 5.6c. The result can be interpreted as

the expansion of wind power capacity to start with helped lower losses, but the most recently built wind farms has had the opposite effect. By looking back at the parameter wind power production and its scatter plot it is possible to see that the curve reaches a minima in loss rate somewhere around 0.35 GWh/h. That would mean that if wind power produces more energy than that the losses are slowly increasing again. It is more probable that more capacity in a non-optimal area is causing higher losses than that certain amount of wind power in the system causes higher loss rates. A lot of new wind power capacity was installed in the final three months in the data set. At the same time the capacity factor for wind was very high. Therefore the effect of the two parameters wind power production and wind power capacity might be hard to differentiate for the regression model.

The parameter wind speed is negatively correlated with the loss rate. See Figure 5.6d. As discussed in Section 4.1.2 the wind cools the wires and lowers the resistance and thereby the losses. However, the result shows a steeper slope than expected and that is probably due to the correlation between wind speed and wind power production. It leads to that some of the effect of wind power production is falsely allocated to wind speed.

The correlation between hydro power production and the loss rate is weakly positive and can be seen in Figure 5.7a. The scatter plot indicates a similar relationship. This might be a misleading correlation and not the true effect of hydro power production, as discussed earlier when analysing input data in Section 4.1.3. However, the correlation analysis for input data, see Section 4.2, showed that the correlation between hydro power production and feed-in power was not too high. Hydro power is regionally produced just as wind power production and could therefore be expected to help lowering the losses. But there are fewer power plants than for wind power production, which might cause the effect to decrease, since then the production is not always close to the end user. To conclude, it is hard to know how hydro power production affects the losses.

The thermal power production, as seen in Figure 5.7b, seems to be decreasing the losses slightly at first and then increase them. The scatter plot shows that the thermal plants are either run low or high and the result of the correlation analysis in Section 4.2 shows that thermal power is correlated with feed-in power. The negative trend at low values of thermal power production might be explained by the regional production character of the plants. There are even fewer thermal power plants than hydro power plants in the region but a small effect of regional production could be the reason for the losses decreasing at low values of thermal power production. To conclude, it is hard to know whether the arguments in this text is valid. The correlation between thermal power production and feed-in power makes it hard to know how thermal power production affects the losses.

There is a positive correlation between import and the losses, but the scatter plot indicate the contrary. See Figure 5.7c. It is hard to say why that is so. Since the share of import from the total feed-in power is rather small, the slope of the curve seems unreasonably high. Further investigation is needed to discover the reason. Perhaps import is correlated with another parameter which in turn is positively correlated with the losses?

For export the relationship, seen in Figure 5.7d, is similar to the one between import and losses, a positive correlation. In this case it corresponds better to the pattern in the scatter plot. The exported power has been transported through the whole region, since the area is a net importer of energy and the connections to the high voltage grid is far from the connections to Denmark. It is reasonable that the losses increase with increased export since the mean distance of transport increases which casues higher losses. But the slope is too high compared to the size of the exported power. Further investigation is needed to understand the impact of export. Perhaps export, similar to import, is correlated with another parameter which in turn is positively correlated with the losses?

The temperature is negatively correlated with the loss rate until around 0 °C, then the curve flattens and in the end of the temperature range the correlation is positive. However, the correlations are weak. See Figure 5.8a. In Section 4.1.2 it was described that low temperature cools the wires and decreses the resistance and thereby the losses. Therefore a positive correlation between temperature and loss rate is expected. It is seen that for temperatures above approximately 5 °C that is true for the model. But for colder temperatures that is not the case and that is probably due to that temperature is correlated with feed-in power. At very cold temperatures, the feed-in power inceases a lot.

The precipitation has no visible effect on the losses, as seen in Figure 5.8b. This is contradicting to what was described in the background in Section 2.1.3; that heavy rain or snowfall can cause corona losses. The spread of the values for precipitation was suffering from the method of creating a mean value from five locations. The range of values might have become too narrow to see the trends. The result could also mean that there are no corona losses on lower voltage levels than $400kV$.

A weak positive correlation between relative humidity of the air and the loss rate can be seen in Figure 5.8c. This could be explained by the corona losses at very high relative humidity. Apart from heavy rain or snowfall; thick fog can also affect the power lines. Since a mean value for the relative humidity for five locations is used, the value is seldom as high as 100. However there can be fog in some areas of the distribution network even when the mean value is below 100. To explain the correlation with corona losses is a bit uncertain since the same correlation was not found for the precipitation case.

Summary

Feed-in power has the strongest positive correlation, and the biggest impact on the losses. Wind speed is the only factor which has a negative correlation over the whole range. Wind speed together with wind power production and installed wind power has the most significant impact on the losses after feed-in power. Temperature and relative humidity seems to have an impact on the losses as well. For hydro power and thermal power weak correlations to the losses have been found, but it is hard to tell what the true impact from those two factors are. Import and export has positive correlations that seem too big compared to the amount of power that is imported and exported and therefore the result is questionable. Precipitation has no correlation at all.

5.5 Marginal Effect Figures

The result for each parameter is presented as a line diagram showing the result from the marginal effect analysis for each parameter. The lines are placed over a scatter plot of the loss rate as a function of the same parameter. The five lines in the marginal effect diagram represents the line describing the marginal effect in the middle, outside of it its confidence bounds and outermost its prediction bounds. Confidence bounds and prediction bounds serves to describe the uncertainty of the result. The concepts are explained more thoroughly in the background chapter, Section 2.2.2.

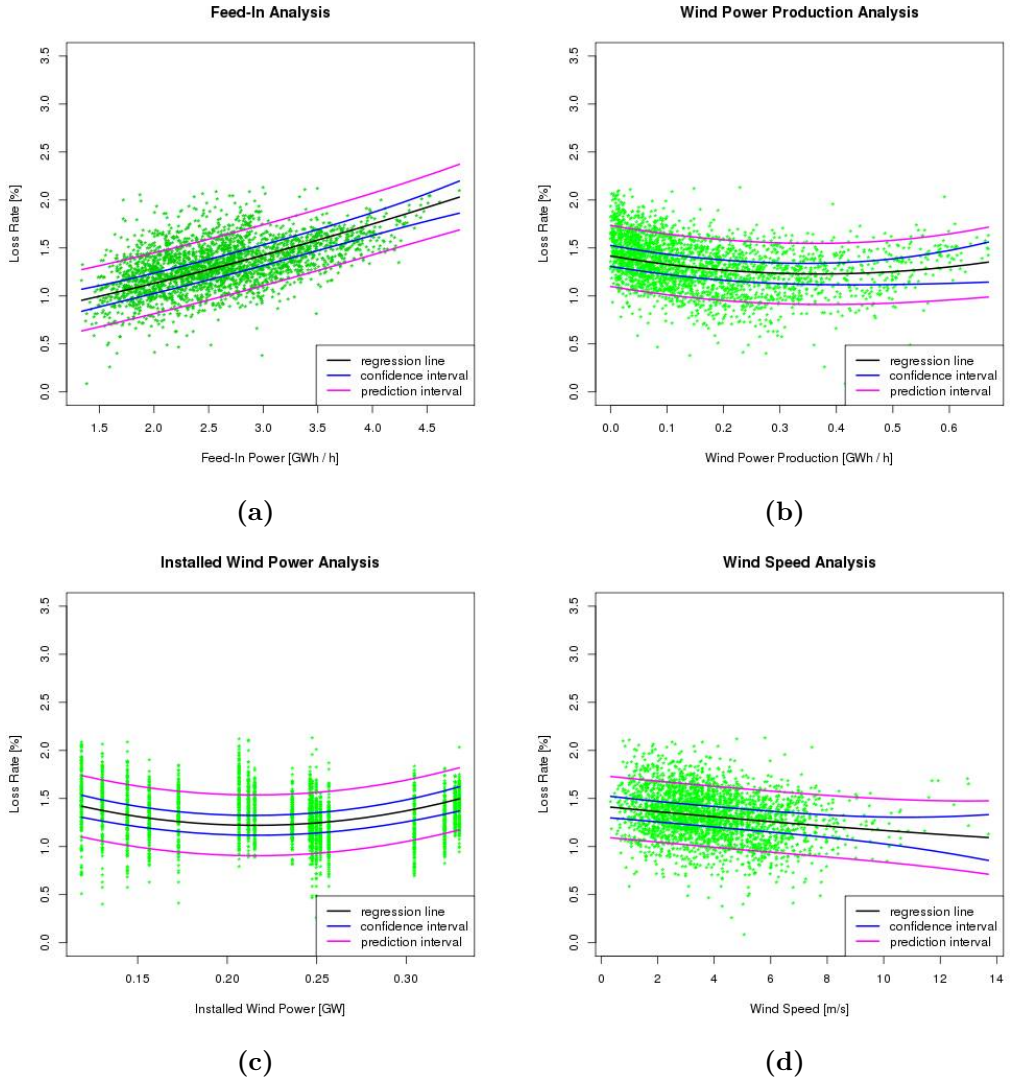


Figure 5.6: Scatter plots and marginal effect analysis diagram for the parameters feed-in power, wind power production, installed wind power and wind speed.

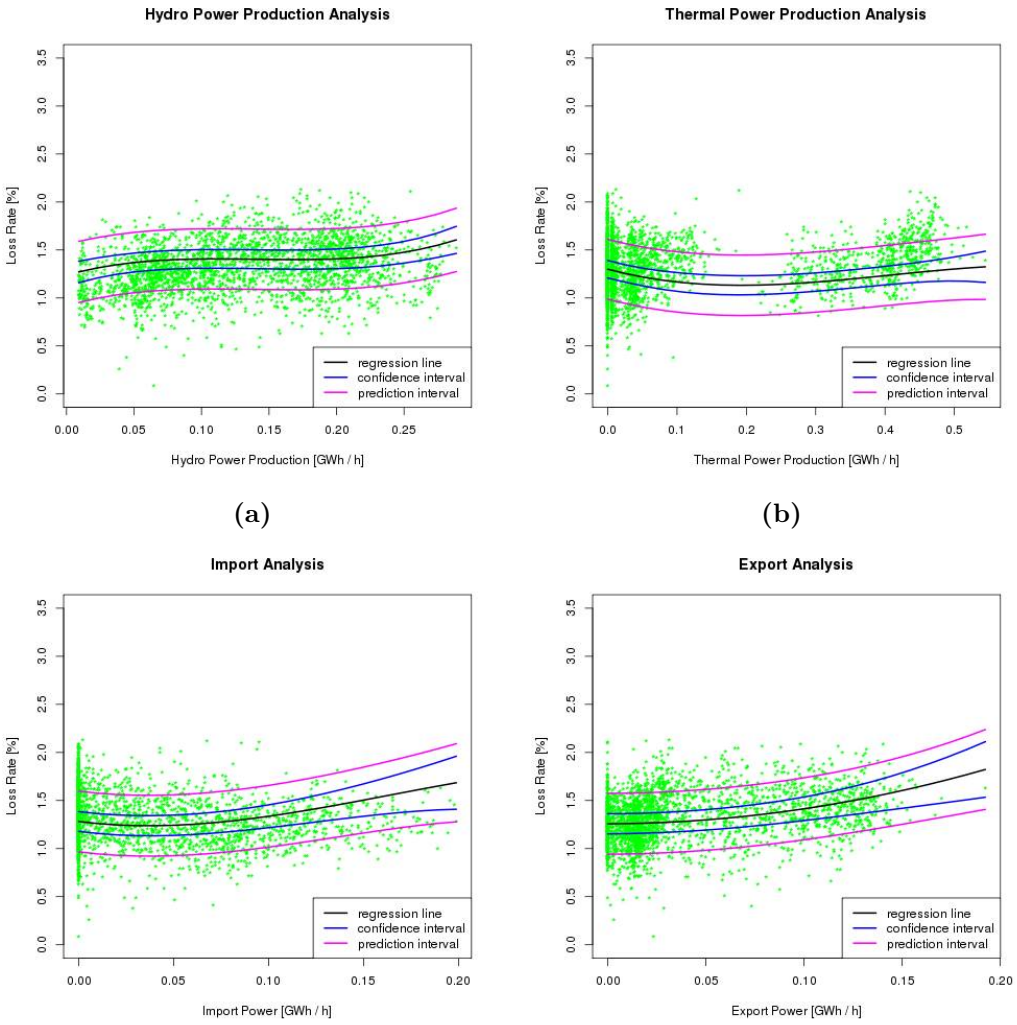


Figure 5.7: Scatter plots and marginal effect analysis diagram for the parameters hydro power production, thermal power production, import and export.

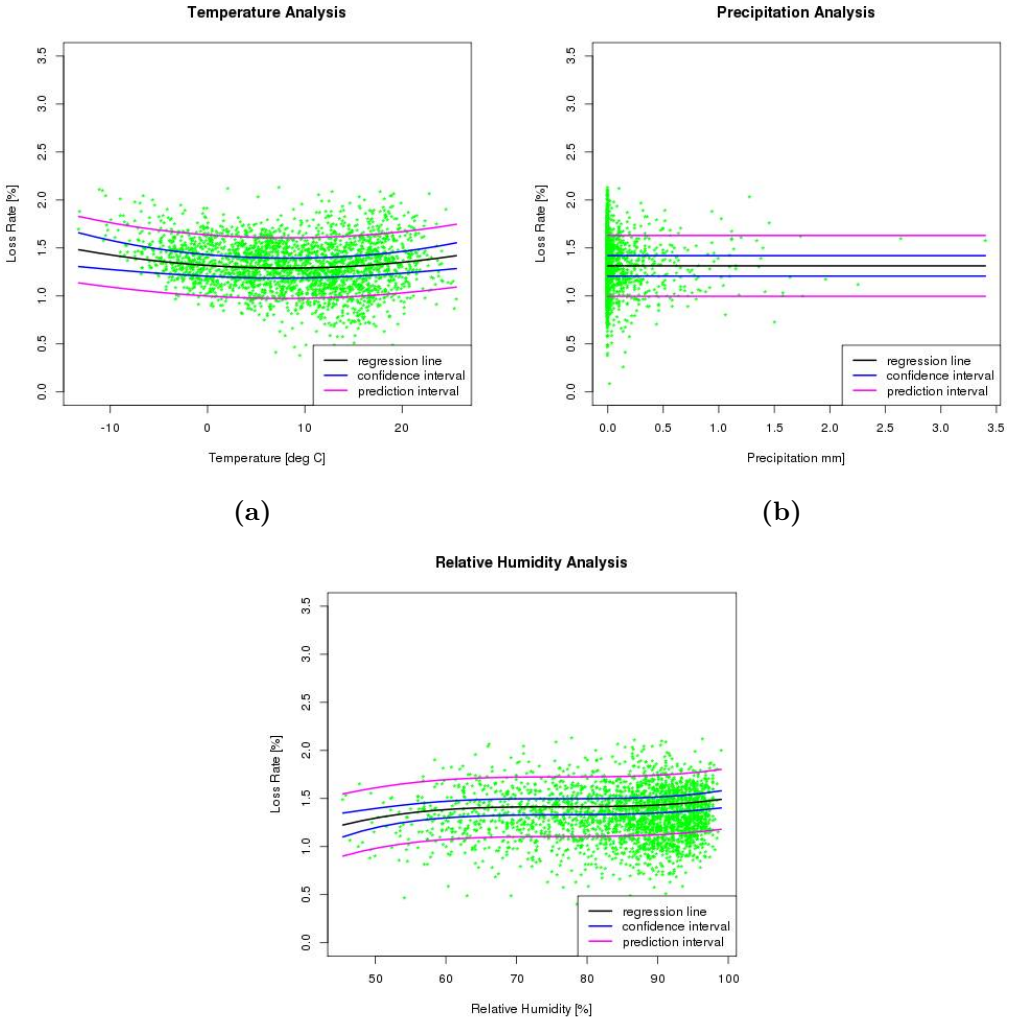


Figure 5.8: Scatter plots and marginal effect analysis diagram for the parameters temperature, precipitation and relative humidity.

5.6 Stability of the Model

The stability of the model was checked both by comparing MSE values and by comparing marginal effects between different runs. The same setup as for the final model was used in seven different runs. In each run a new data set was used. To start with, the performance and prediction performance of the model was studied. The results showed that they deviate depending on the training set, since in each run a new training and test set is created. The results from the seven different runs are presented in Table 5.4.

Most important to note is the difference in results. The differences indicate that the model is unstable and that the variable selection process is sensitive to which data set that is chosen. The values from the sixth run is what was presented as the accuracy of the model in Section 5.3.

Table 5.4: MSE of the training set and test set for five different runs with the 11 parameters.

Mean Square Error

Data Set	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6	Run 7
Training (80%)	0.023	0.023	0.022	0.022	0.022	0.022	0.022
Test (20%)	0.023	0.025	0.027	0.025	0.030	0.022	0.026

In the sixth and seventh run the marginal effect analysis was perfomed using their corresponding $\hat{\beta}$-values. The result for the marginal effect analysis of the final model, presented in the previous section, is the result from run number six. When running the marginal effect analysis for two different data sets it became clear that even the impact from each parameter changed from one run to another. The marginal effect of the parameters changed, and the most distinctive difference was for the three parameters connected to wind power; wind power production, installed wind power and wind speed.

In Figure 5.9 the results for a marginal effect analysis for the three parameters connected to wind power for two different runs are presented. To the left in the figure are the same results as was presented in the previous section, coming from the sixth run, to the right in the figure are the results from the seventh run. It is clear that changing the training set has a large impact on the result of the model.

A reduction of data was made in order to handle the problem with autocorrelation, discussed in Section 5.2. This might be the reason for the problem with instbility that occured.

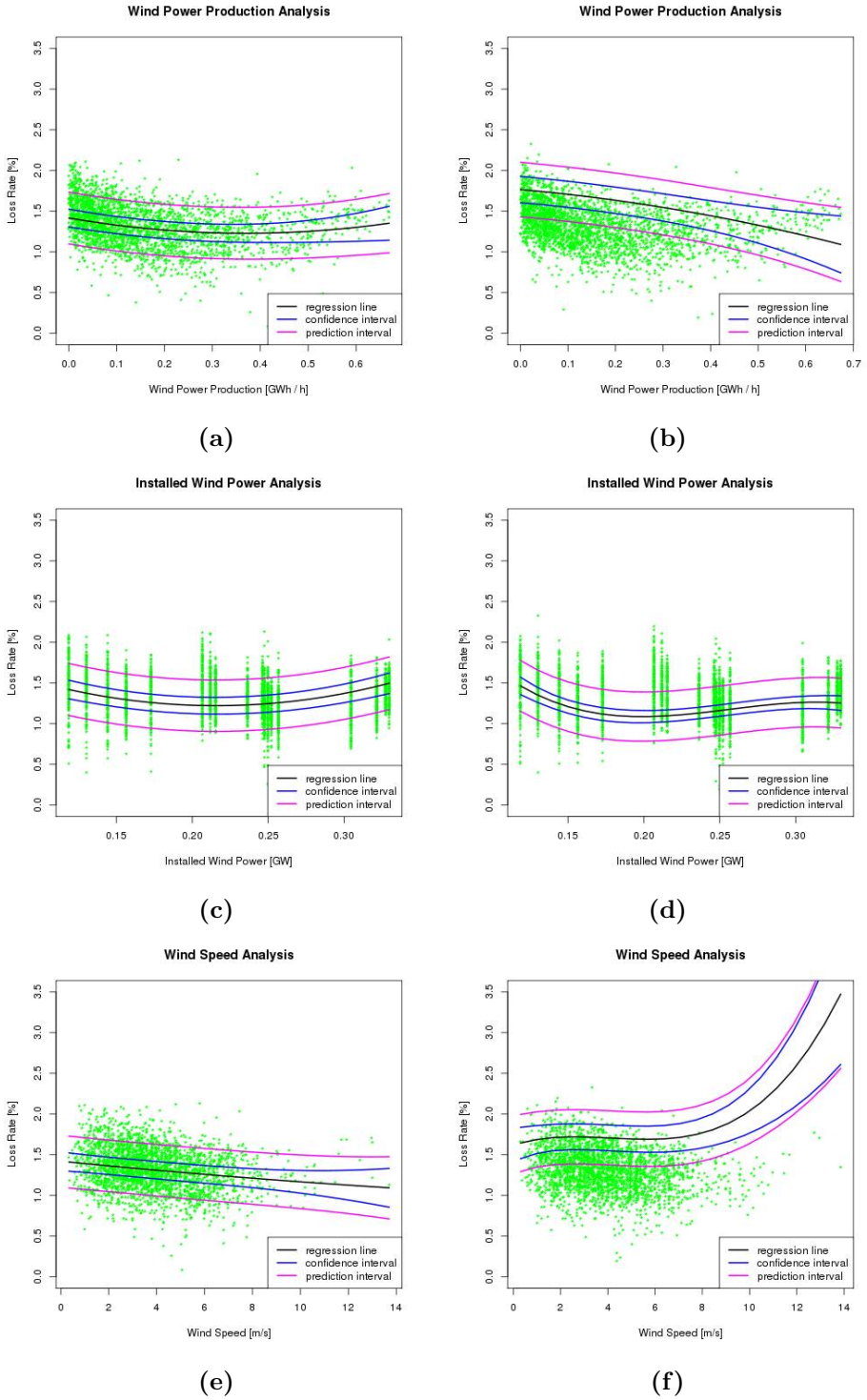


Figure 5.9: The three diagrams to the left are the final result and the three diarams to the right are the result from another run.

6

Discussion

In the previous chapters the method, the collected data and the result from the regression analysis has been presented. In this section the results of this thesis work will be discussed. It will discuss each part of the project and give recommendations for further work. A summary of the recommended work can then be found in Appendix D.

With the threats of today's climate change energy saving and energy efficiency has had a space in media. But it is possible to try to mitigate the trends the climate change has in many different ways. One way is to increase the efficiency in power transmission as utilising the system more effectively can save energy. One example is next generation's smart grids. An important question to ask is what does impact the losses in the system. The pros and cons of lowering the losses at the expense of other tings must be evaluated. For example, the environmental impacts of network developments must be compared to the costs associated with it. E.g. how much regional production such as wind power is beneficial in a net import region? Wind power production is an intermittent energy source and can consequently push the stability limits of the power grid. Moreover, regional production is generally inserted at a low or medium voltage level, thus it must be transformed up to higher voltage levels and back down if the power needs to be transported long distances before consumed. Therefore, low-carbon emission regional power production alternatives can incur a higher environmental impact than first assumed, but the impact is still very small compared to higher emission production alternatives.

This master's thesis project did consist of different parts. One of the aims was to investigate if a statistical approach is at all feasible to use on a distribution network, to study the losses, and to get a better understanding in the dynamics of the network on a system level. Another aim was to, by studying the losses in the network, understand what factors they do depend on. The goal was to develop a method suitable for Energicontrolling at E.ON Elnät Sverige AB, to do statistical analyses, with the software

the group had access to. The goal was also to identify potential parameters and using regression analysis study how these parameters correlated with the losses in the system. A statistical approach such as regression analysis can find the correlation between the losses in the power network and potential parameters, but correlations in the data is not necessary causation, and the results must be examined and used with caution.

6.1 Working with Lavastorm and R

Lavastorm is a big data analysis software, and therefor in a software-group that will probably prove to be a great tool in the future where there will be more and more big data environments. In Lavastorm it is straightforward to see the data management processes, hence it is suitable to use if many people shall be involved in the same project.

In the project Lavastorm and R was run on a server. Hence the running time of the nodes in the graphs was not affected by the computers' performances. But still the running times were long considering that it was not a big data case. Another, but bigger disadvantage was that only one process using R could be done at a time. Thus, only one person at a time could to statistical calculations at the server.

Lavastorm proved to be an easy way to handle the data from different sources, and if the raw data was updated, it was easy to refresh the graphs and to quality-check the updated data. But using Lavastom had its limitations as well. The Excel read-in node did not work if the characters Å, Ä, Ö and : was included in the column titles. This was not ideal with the interaction with Pomax as these characters had to be changed in the Excel sheet each time after refreshed data had been loaded from Pomax. Nevertheless, it is possible to get data directly from Pomax into Lavastorm in the future and then the automation of the data-update would doubtless be simpler.

There are many advantages with Lavastorm, and most of the problems encountered were solved. The problems that could not be solved during the development of this project were all connected to the R-node. The used R-node is a new add-on in Lavastorm, and the development is still, as writing, a continuing process. That resulted in that there was no documented guide of how to use it, only a template graph. Moreover the outputs from the R-node were not always complete. For example, if the output was a matrix the output missed the row names. It was necessary to have R installed as a separate program to access the support documents of possible functions, but also to see how to get the right outputs from the built-in R functions. In Lavastorm, when the R-node did not run successfully the error messages were hard to understand, and they did not always give indication to where in the code the error occurred. Furthermore, if there was an error in the script, the time it took for Lavastorm to abort and produce an error message was much greater than the running time for a functioning R script.

Moreover, as only certain formats were allowed to output from the R-node to the Lavastorm graph it was not possible to see the model summary, which can be easily accessed from an R script in a R software. As the R-node is created in Lavastorm to do

these kind of analyses, it is strongly suggested that the allowed outputs from the node are looked into, and that a summary model would be considered as a new output format.

A big disadvantage with the R-node was that the model itself could not be passed on to another node. The result of the model could be passed on, e.g. the $\hat{\beta}$- values, but not the model itself, thus all the potential analysis had to be done in the main R-node. This made the script in this node long, and working against the benefits of using Lavastorm, where in general all steps are easily followed by individual nodes. Moreover, when the second part, with the model analysis, was developed, the first part had to be run for every trial run. This was a problem since in the first part where the model was optimised, R's step function was located, which is the most time consuming part in the script. To get around these problems the base model was used when the latter part was developed, and not until the model analysis was done the step function was re-activated in the script.

To conclude: Lavastorm proved to be a good data handling software with great potential in taking on the big data problems correlated to the power networks in the future. It will allow for more cooperation between different groups within the company, as it is possible to share each individual groups data, even though it has different formats. The Lavastorm R add-on will in the close by future be an important part in analysis of the accumulated data. Nevertheless, more development in the Lavastorm-R interaction is needed to make the R-node easier to work with and steps to shorten the running times should be considered. From Energicontrolling's side, using R from a server location might not be ideal. First, the limitation of users running the node is not fitting, more employers should be able to use Lavastorm to do analysis, and each user should be able to use R both for more advanced calculations than the built-in nodes could handle, and also for statistical analysis.

6.2 Evaluation of the Method

From the three-parameter model presented at the workshop another nine parameters were added in the second half of the project. Weather parameters were added both due to their correlations with losses and the fact that they were time series. For example temperature would hypothetically correlate with the seasonal and daily variations in the losses. More regional power production, apart from wind power, was added to see if a negative trend line would emerge, as the overall transmission distance of the power would be reduced.

The model used in R was a stepwise function, optimising the model by iterations where a parameter was added and removed depending on the improved fit of the model. This allowed cross correlation between parameters, resulting in a model with a large number of $\hat{\beta}$-values. The prediction performance, was analysed using the MSE-values, and the histograms of the residuals.

R offers another function where the iterative steps optimising the model also take these values into consideration. This function could be further investigated, but would

make the iterative step more computational heavy.

The residuals of the loss rate suffered from autocorrelation. As the project was about to see the possibilities of the regression analysis method and not only to do a regression analysis on the collected data, the autocorrelation problem could not be investigated properly and addressed correctly. Instead an improvement of the residuals could be done by a semi-random selection of the raw data. Thus, the data was reduced to make each measurement more independent, but could not be reduced to overcome the problem. As then the amount of data used in the model fit was too small to produce a decent model. The effect of the autocorrelation is that the confidence and prediction bounds are underestimated. Investigation on another way to decrease the autocorrelation in the model is of great importance, without decreasing the number of measurements included in the raw data.

Furthermore, as more data will be available over time, the fit of the model will better describe the reality; in this project the data only spanned 1.5 years. Still, the model will only find correlation in the data, and each correlation needs to be checked with reality as correlation and causation is not the same thing. Therefor the purpose of the model is to raise flags of likely relationships, more than determine the connections.

6.3 Evaluation of the Model

With the tools created in Lavastorm to analyse the models derived it was possible to evaluate if the fitted models were feasible. The different tests showed different effects, and the result chapter shows that one test can show that the model is good whereas another test shows the opposite. Hence it is important to stress the usage of all the different tests when a model is analysed.

The consistency of the model returned in different runs was evaluated and the outcome was that the stepwise function used in R did optimise the model with different sets of parameters each run. To conclude; using only part of the raw data resulted in a model that was dependent on which selection of the raw data that was used. This can indicate that the size of the dataset used in the stepwise function was too small, and a larger dataset size might potentially, without increasing the autocorrelation or the measurements dependence, get a more stable model. With time the dataset will naturally increase, as measurements are done every hour and added to both Pomax and SMHI, but there are statistical methods that could be of interest as well. Bootstrapping is one example, or looking into stochastic processes could be another option. With stochastic processes, like random walks or Brownian motions, the model could in the future be used to foresee the future months losses.

6.3.1 Monthly Factors

Adding monthly factors can be a way to identify missing parameters with monthly trends. In this project monthly factors were added but the effect was such that they

were later dismissed. The factors gave each month a separate intercept, but with the cross correlation parameters, the factors resulted in different regression models, i.e. one per month. Be aware, the month factors in themselves do not mean anything. It is only an indication of that there are other factors that influence the losses that are not yet included in the model. In the beginning of this project, when only the feed-in power and the wind power production were compared to the loss rate, month indicators did improve the model significantly. But with the final set of parameters in the $\mathbf{X}$ matrix the difference in the models between having monthly factors or not is no longer large. Still, the MSE- values are improved in the training set for the case with month indicators compared to the case when they where not included. The number of parameters in the model with month indicators is roughly twice compared to without these factors, as cross correlation with parameters are included. This results in a much longer running time for the factor, since the stepwise function has a lot more options to choose between. As the study was developed the monthly differences were dismissed. Further studies can be made in how the variation of each parameter change each month, over time, to see if there are seasonal changes missed so far, or if the model with monthly factors do fit better to the measured result and possibly make better predictions. Therefore, in the marginal analysis in this report the model without month indicators were used.

6.3.2 Marginal Effect Analysis

In the marginal analysis all but one parameter were given their mean value. The parameter of investigation was varied from their minima to their maxima. The model is given in Table C.1 in Appendix C. As discussed, there is still autocorrelation in the residuals, but the model is, within this thesis work, as optimised as could be. The resulting images of the marginal analysis can be found in Section 5.5. As there is autocorrelation included in the model the confidence and prediction bounds shown are underestimated.

Ohm's law gives the theoretical relationship between the *losses* and the current. The *loss rate* should therefor be linearly dependent on the feed-in power. As expected, the loss rate do increase for larger feed-in powers.

Wind Analysis

The measurement scatter for wind power production shows a tendency as a parabola, with a minimum at around 0.35 GWh/h. The fitted correlation of the model shows a negative trend for the majority of data. I.e. that wind power produced within the region does lower the system losses. It is the only regional power produced studied that shows a strong negative trend. In other words; the more wind power produced the lower the losses are in the system, but only up to a certain limit, after which losses increases.

Nevertheless, there are other parameters that need to be taken into consideration if this kind of statement should be considered. Wind speed is also correlated to the losses negatively. That is, the windier it is the lower losses. Wind speed does affect the system in many ways. First, when it is windy more wind power is produced. This link is seen in the correlation table at page 37. But also, the wind cools the cables, resulting in lower

losses. Another studied parameter that has a link to the wind power production is the installed effect.

In the marginal analysis the installed wind power results in a parabola with a minima at 0.20 GW. As the installed capacity is only increasing, see Figure B.4, the trend is that more installed wind capacity does increase the overall losses. Thus the benefit from the wind power production and the wind speed must be weighted in an investigation of when the system cannot include more wind power without increasing the losses. Furthermore, the placement of the wind power productions within the area is of great importance when considering the losses. The benefit of regional production is questionable if the production is much larger than the load. Not only is the transported distance increased, but for each transformer the power have to pass, the higher the loss rate gets.

Regional Production, Import and Export

The model gives a small positive correlation between the regional production of hydro power and the loss rate. This is not expected, as regional production should lower the losses. For thermal power the correlation with respect to the losses is the opposite. How much the weak correlation to both the feed-in power and the temperature affects the result can be further studied. Next are the import and export parameters. Both show that a higher level does result in a higher loss rate. For import, comparing the result to the scatter plot, with the raw data, cannot validate the result.

All these parameters need to be further analysed so that the result make sense. Still, the model has a better prediction performance with them than without, thus they seem to play an important role. However if it's them, or a parameter not included in the study is hard to express.

Weather Parameters

The marginal analysis shows that the lowest loss rate is when the temperature is around 5 °C. The potentially biggest benefit of including temperature in the regression analysis cannot be visualised in the marginal effect analysis, as it is the fact that temperature is a time series. One major benefit is that the model gets a seasonal parameter, making it possible to use the model without including monthly factors. Besides, temperature, and other parameters do better describe the reality than monthly factors would ever do.

The precipitation analysis does indicate that the precipitation parameter does not influence the mode, as the fitted line is visually flat. In further studies the existence of the parameter in the model could be investigated. As discussed with the temperature, the precipitation parameter might have other uses to the model, as it is as well a time series. Present, in the marginal analysis of precipitation, all other parameters are at their means, but maybe the precipitation parameter does play an important role at different values of the other parameters?

Relative humidity shows a weak positive correlation with the losses. Once again the inclusion of the parameter might be more important for the model than seen in the marginal analysis. Additionally, as discussed in the result section, the relative humidity

is a very local parameter, resulting in effects on a very local level and to get an average value within the whole region might not be the best way to study how the relative humidity does affect the loss rate in the system.

In Section 2.1.3 the corona effect at high voltage networks was discussed. This study shows that there might be a corona effect even at networks with lower voltages. However, it was not visible in the precipitation analysis, but could be seen in the relative humidity analysis. Could it be that the corona effects do not increase with increased precipitation? I.e. as long as it rain corona effects will take place. Further analysis would be motivating.

Limitations of the Analysis

In the marginal analysis one parameter at the time was analysed. The other parameters were kept at their mean value from the measured data. With the time limitations of this project it was a general and fast way to see how the parameters were affecting the loss rate. If this is a good approach or not is questionable. Especially when considering the parameters that are closely linked to the wind power production, i.e. the installed wind power and the wind speed. E.g. should not the effect of both increases in wind power and installed capacity be studied? As the installed capacity does in reality limit the possible amount of wind power that can be produced.

A more important question to discuss is the choice of constant value for the parameters. Does it make sense to have the mean value of installed wind power as the constant value for any of the analyses? Moreover, the feed-in power has a big impact on the losses, and in this analysis the feed-in power was set to its mean value. What impacts does the other parameters have on the loss rate when the load is high; hence the system is under a much higher stress?

Most of these questions could be addressed in another marginal analysis, where the problem description would be different. That is one of the advantages with this approach to study the transmission network. When the regression model has been optimised, it is possible to answer questions to many different scenarios. Hence it is strongly recommended that further work is done to do further analyses on the model, where different scenarios are created and different problem descriptions evaluated.

6.4 Concluding Discussion

To conclude this section the aims of the master's thesis project should be answered.

By doing a marginal analysis the effect on the loss rate the different parameters had were investigated. Insight in how the loss rate was affected by the regional production such as wind power was gained. The study showed that wind power does for most hours decrease the losses in the system, but for some hours, an increasing trend has been identified. The causation of this has not been found here. The analysis showed that further insight could be obtained, with a broaden analysis.

With Lavastorm it was possible to deliver a method for which the company can continue to do statistical analysis on their data. There is also huge further development

potential.

The final aim was to see if a statistical approach was feasible to use when studying a transmission network. Regionnät Syd is a large area to consider, with many local phenomena that cancel out. Still, it was possible to gain much insight in correlations to the losses and potential causations for this. Thus, it is reasonable to study the network in this manner. Including more parameters would improve the model, as would an increase in data. Still, the model is suitable enough to get more insight into how the losses are affected.

However, if a smaller area was studied, more local phenomena could be included and correlations to these and the losses could be studied. There will be a break-even point where the statistical approach no longer outweighs the simulation approach in gaining knowledge. Hence recommendations are to change the system boundaries; to zoom in to smaller areas; study lower voltage networks but also to study different medium voltage networks, for example in north of Sweden where the networks are net producers.

7

Conclusion

It is concluded that working with Lavatorm and R as software for statistical calculations had both advantages and disadvantages compared to working directly in e.g. R. To use Lavastorm with its graph where all nodes were visible provided a good overview of all the steps in the regression analysis. Thus, it is easier to teach someone else to use the model, and develop it. The drawback was the shortcomings considering the interaction between Lavastorm and R which gave long running times and made the search for errors in the R-scripts difficult. However, Lavastorm and R has great potential and most of the problems that ocurred in this project will most likely be solved in the near future.

The final model is not perfect and there are still things that can be improved. For example, all the influencing parameters are not found and there are too few data points at the moment. Both things lead to a somewhat instable model with autocorrelations in the data. The MSE might become lower than it is now, to give an even more accurate model. The ideas of how to improve the model can be found in the Appendix D, Future Work.

The result gave the correlations between each factor and the loss rate. The most interesting finding was the correlation between losses and wind power production, installed wind power and wind speed. These correlations can be explained which gives them credibility and the information can for example be a support when deciding how to continue the development of Regionnät Syd. Other interesting findings are that relative humidity and temperature seems to have a positive impact on the losses, higher temperatures and high relative humidity increases the losses. The impact of feed-in power was already known previous to the study. The other factor's impact are more uncertain and more information is needed before conclusions can be drawn.

The statistical approach can help finding the factors that influence losses. But there are shortcomings of this approach. Many factors are very local phenomena and since one value for the whole region is used the range of values for a factor can be too narrow. The differences are evened out and the influence of a factor may be lost in the statistical

noise. However, this property can also be the advantage of the approach. Relationships can be found that are valid for the system level which would not be visible otherwise. Something that is a problem on a local level might be something positive on a system level.

It is of importance to keep in mind that the statistical approach can only provide information on which parameters that are correlated to the losses. This does not prove that the losses are really influenced by the factors.

Appendices

A

Workshop

Date 14^{th} March, 2014

Place Moderna Museet, Malmö

Time 09.00 - 11.30

The aim of the workshop was to gather people at E.ON Elnät with different knowledge about the behaviour of the distribution network and its losses in order to find new inputs and ideas to the project. The goal was to make use of this knowledge in order to develop the model further and improve its predictive performance. The idea was that the outcome of the workshop would be a list of new variables to add to the model as well as guidance to how to retrieve the necessary data for each variable.

A.1 Workshop Schedule

The workshop was scheduled for the 9^{th} week of the project. This was a good time to invite people from outside the project to present the deliverables and then let them give inputs to the project. The timing was good since we had learned the method of regression analysis by this time, and was able to estimate how much could be done during the remaining weeks. The scheduled used on the workshop can be seen in Table A.1. The presentation covered the introduction and background of the project and the results and conclusions available at the time. The purpose of the presentation was to give the participants a brief overview. The brainstorm session was divided into three parts with one question for the participants to answer in each part.

Table A.1: Schedule for the workshop that took place on the Moderna Museum in Malmö as part of the project development.

Workshop Schedule 14^{th} March, 2014	
09.00-09.35	Presentation of the deliverables
09.35-09.50	Time for questions and answers about the project
	Coffee Break
10.10-11.30	Brainstorm session

A.2 Brainstorm

The discussion was performed in small groups and then the findings were presented in front of the whole group. The inputs from the participants of the workshop was documented and saved to be used later on. Both as input to what the project would be focused at at the remaining time of the thesis work, and for future works in the Energicontrolling group. The three questions discussed were:

1. How can a statistical model for power losses be used at E.ON?
2. What factors affect power losses? How can data for these variables be found?
3. How to proceed with the project after the master thesis is finished?

Question number two is most relevant for the thesis and therefore the findings from that brainstorming are presented next.

A.2.1 Finding from Workshop

What factors affect power losses? How can data for these variables be found?

1. Weather
   (a) Wind – wind speed especially, but also wind direction
   (b) Temperature
   (c) Rain – heavy rain or snowfall may cause corona-losses
   (d) Solar insolation – The temperature of the wires are affected by insolation. Higher temperature gives higher resistance
      i. Cloud cover – data from Andreas Brorsson
      ii. SMHI's database?
   (e) Relative humidity of the air – Thick fog can have the same effect as heavy rain.

i. SMHI's database?

2. Geography
   (a) Feed-in at every substation is one variable.
   (b) Output at every substation is one variable.
   (c) Distance between feed-in/production and end users.
      i. Data is found in Pomax.
      ii. DPpower makes it possible to locate every substation.
3. Standard vs non-standard couplings
   (a) Binary data, 0 for standard and 1 for non-standard couplings.
   (b) Contact person: Fredrik Appenrodt
4. Öresundverket- ÖVT
   (a) Losses are affected when this thermal power plant is switched on and off.
   (b) A potential case study.
5. Reactive power
   (a) Causes variations in current, higher current leads to higher losses.
   (b) Accumulated data partly available in Pomax, the quality of this data gathered today is much better than before. In the future more reactive power data will be available in Pomax.
   (c) Data could be gathered from UDW, where it is possible to find data from two years back. Names for substations different than in Pomax, may cause a problem.
6. Hour of the day
   (a) Correlated to weather and the energy feed-in (the consumption).
7. Area and quality of cables
   (a) The cables are replaced over time.
   (b) Data is available in DPpower.
8. Prices at the spotmarket
   (a) Correlated to energy feed-in.
   (b) Data is available at Nordpool's website.
9. Import/Export to Danmark

(a) May increase the amount of power that is only passing through the distribution network.

(b) Import: Danish wind power.

(c) Export: Energy exported to Denmark normally comes from the transmission grid and passes through Regionnät Syd.

(d) Data available in Pomax /UDW /Nordpool

10. The ongoing development of the distribution network

(a) Increased capacity, bottlenecks are found and avoided.

(b) Installed wind power capacity.

A.3 Participants

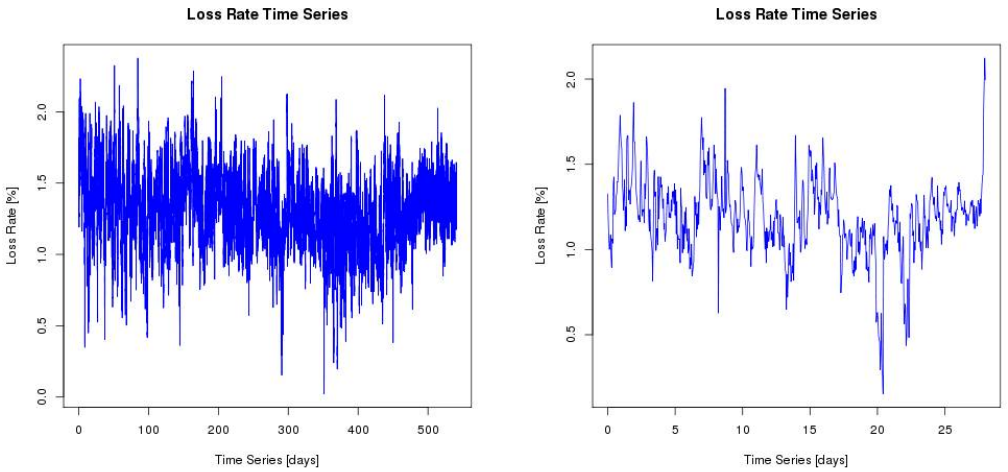
Table A.2: Workshop Participants

Workshop Participants	
Brant, Linda	E.ON Elnät
Brönmark, Torsten	E.ON Elnät
Brorsson, Andreas	E.ON Elnät
Hellman, Linus	E.ON Elnät
Jeppsson, Jenny	Chalmers
Kadir, Sezgin	E.ON Elnät
Larsson, Martin	E.ON Elnät
Lemvall, Ronny	E.ON Elnät
Lund, Elin	E.ON Elnät
Lundsgård, Anna	E.ON Elnät
Månsson, Claes-Håkan	E.ON Elnät
Music, Esad	E.ON Elnät
Nilsson, Johan	E.ON Elnät
Nilsson, Kristina	Chalmers
Nivhede, Sofia	E.ON Elnät
Norin, Simon	E.ON Elnät
Schånberg, Wilhelm	E.ON Elnät
Suleman, Zahoor	LTH
Svensson, Andreas	E.ON Elnät

B

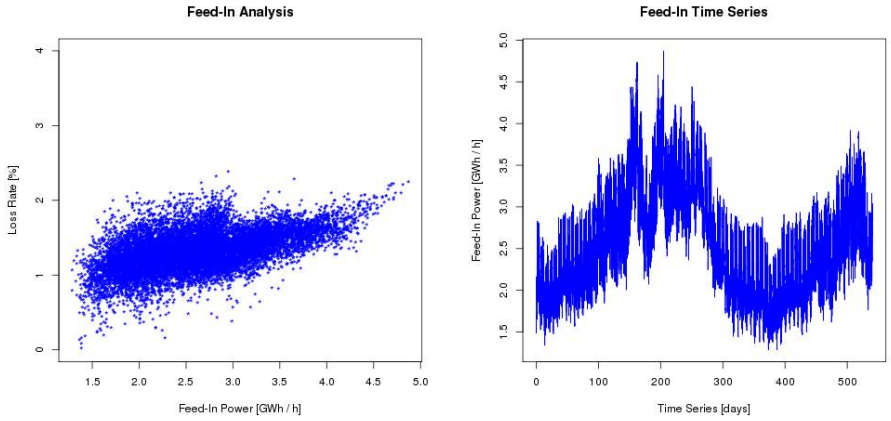
The Input Data Diagrams

In this appendix the figures that did not fit into Section 4.1.2 are presented. See that section for the figure analysis.

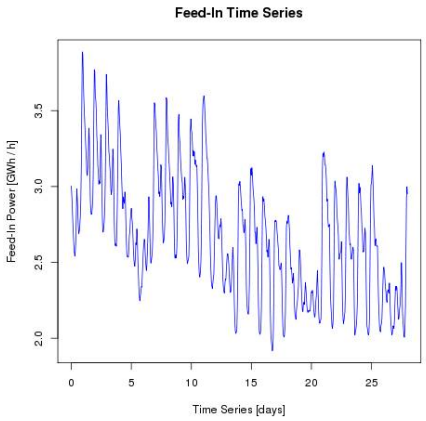


(a) Complete dataset **(b)** April 2013

Figure B.1: The loss rate over time. Both for the whole data set and a snap shot of four weeks in April 2014

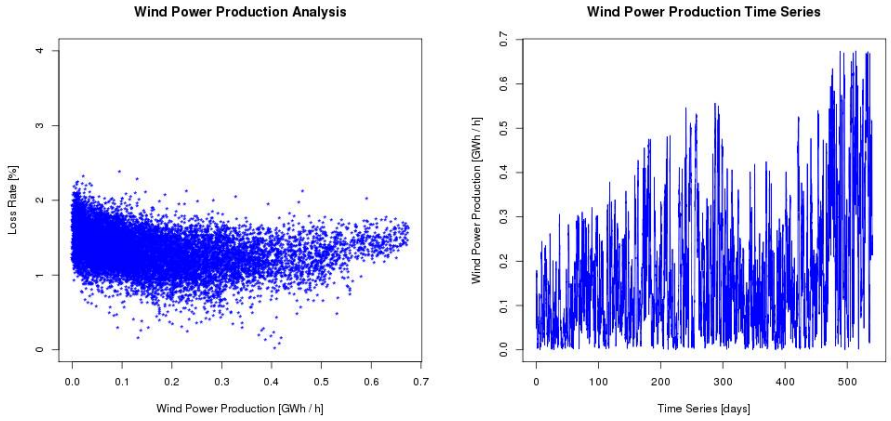


(a) **(b)** Complete dataset



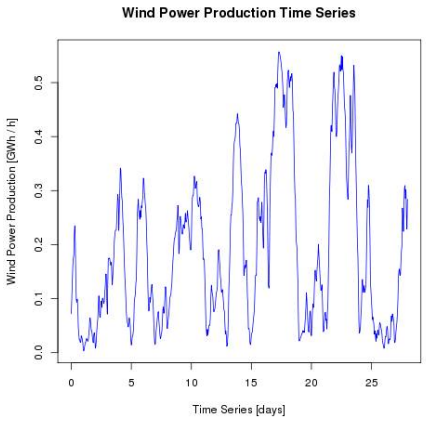
(c) April 2013

Figure B.2: The top left diagram show the losses as a function of feed-in power. The two other diagrams show the feed-in power over time; both for the whole data set and a snap shot of four weeks in April 2014



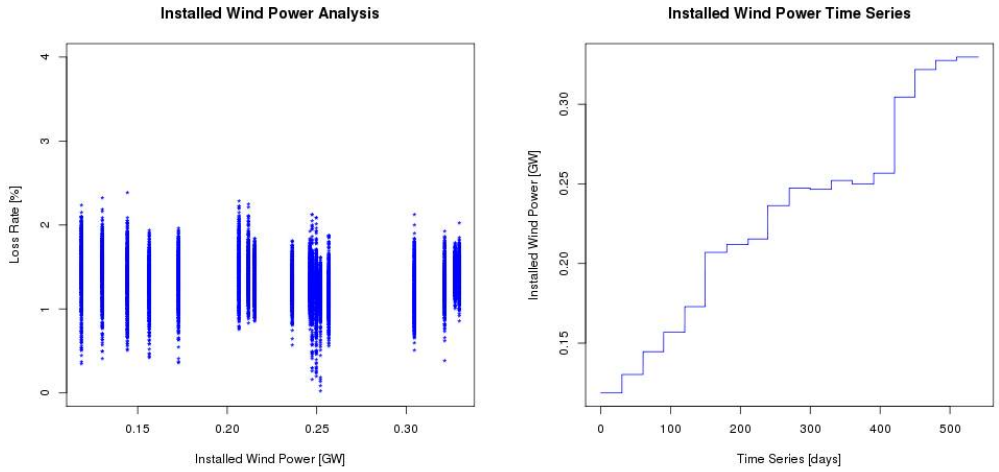
(a)

(b)



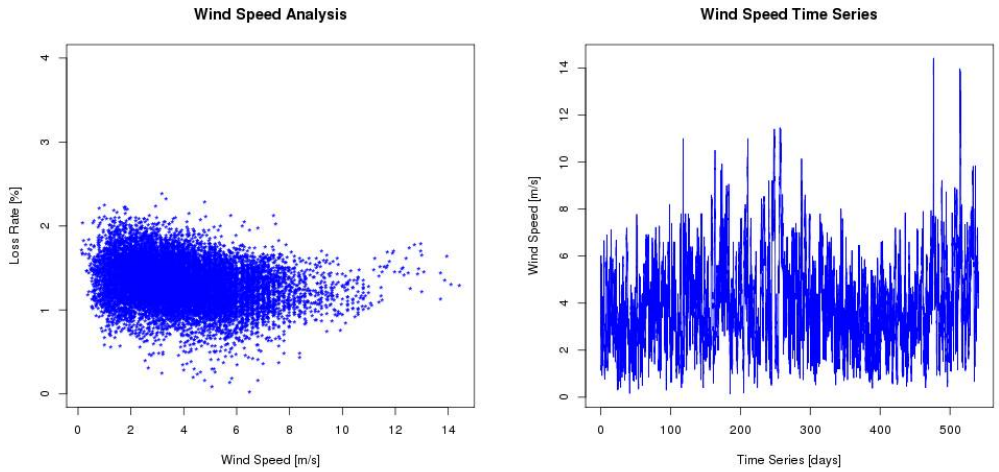
(c)

Figure B.3: The top left diagram show the losses as a function of wind power production. The two other diagrams show the wind power production over time; both for the whole data set and a snap shot of four weeks in April 2014



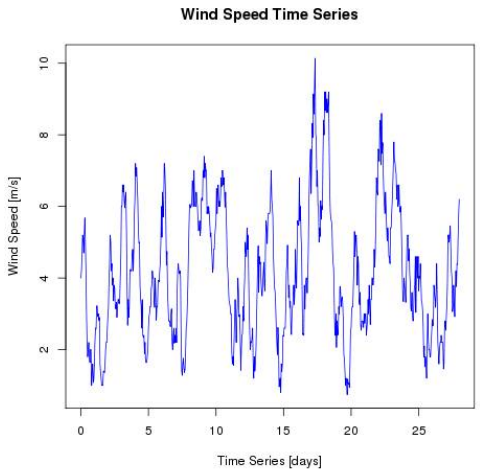
(a) **(b)**

Figure B.4: The diagram to left show the losses as a function of installed wind power. The diagram to the right show the installed wind power over time.



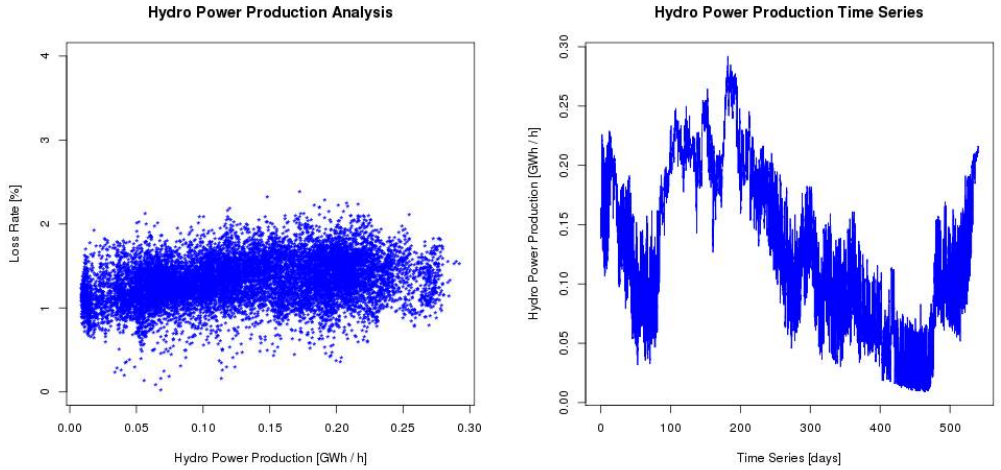
(a)

(b)



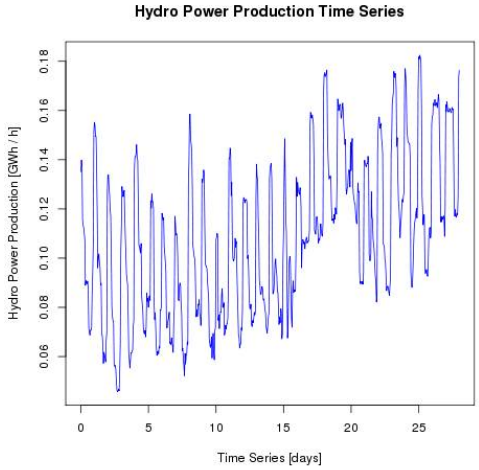
(c) April 2013

Figure B.5: The top left diagram show the losses as a function of wind speed. The two other diagrams show the wind speed over time; both for the whole data set and a snap shot of four weeks in April 2014



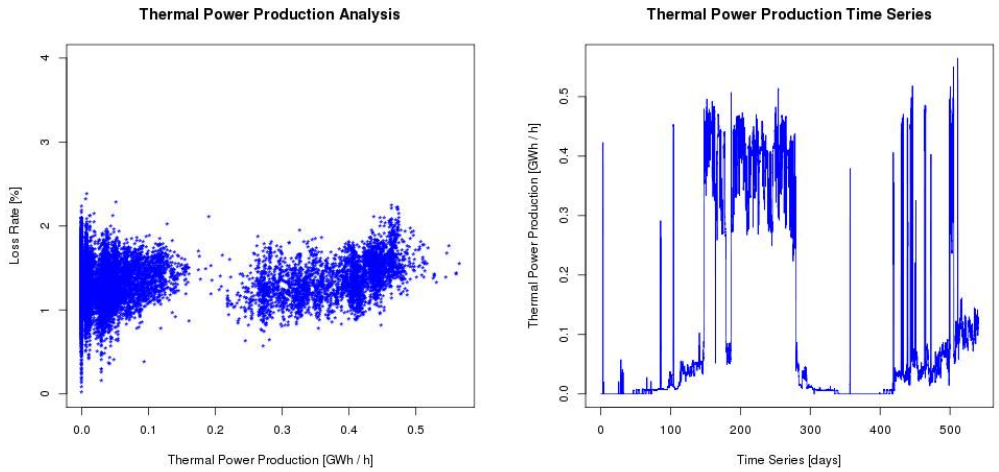
(a)

(b)



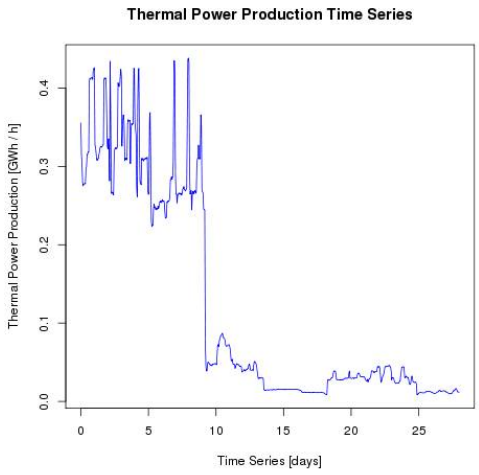
(c) April 2013

Figure B.6: The top left diagram show the losses as a function of hydro power production. The two diagrams below show the hydro power production over time; both for the whole data set and a snap shot of four weeks in April 2014



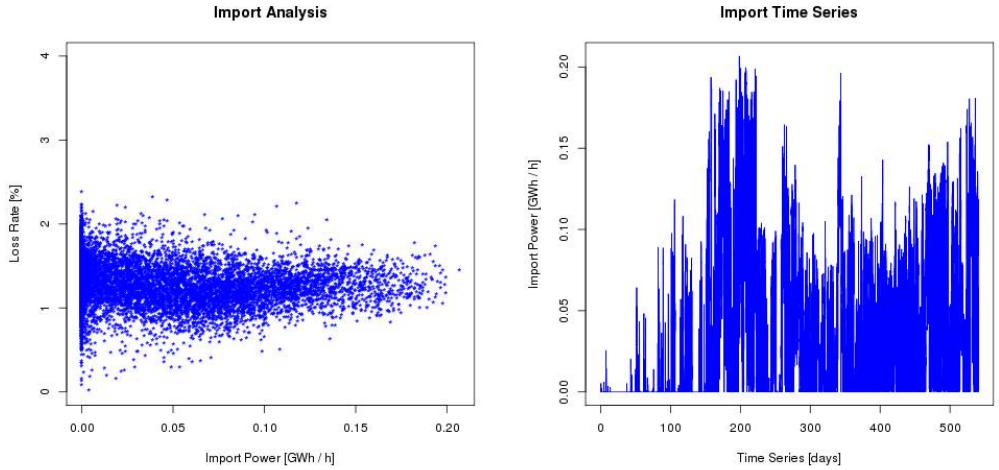
(a)

(b)



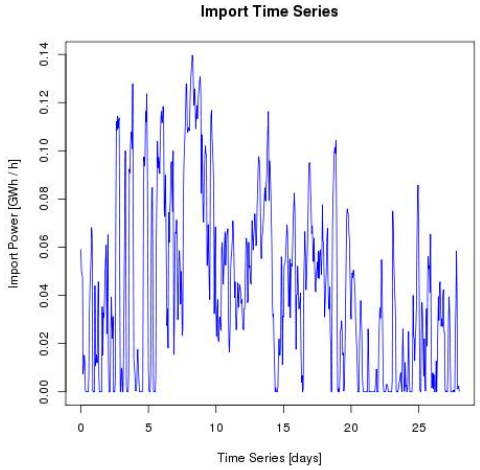
(c) April 2013

Figure B.7: The top left diagram show the losses as a function of thermal power production. The two other diagrams show the thermal power production over time; both for the whole data set and a snap shot of four weeks in April 2014



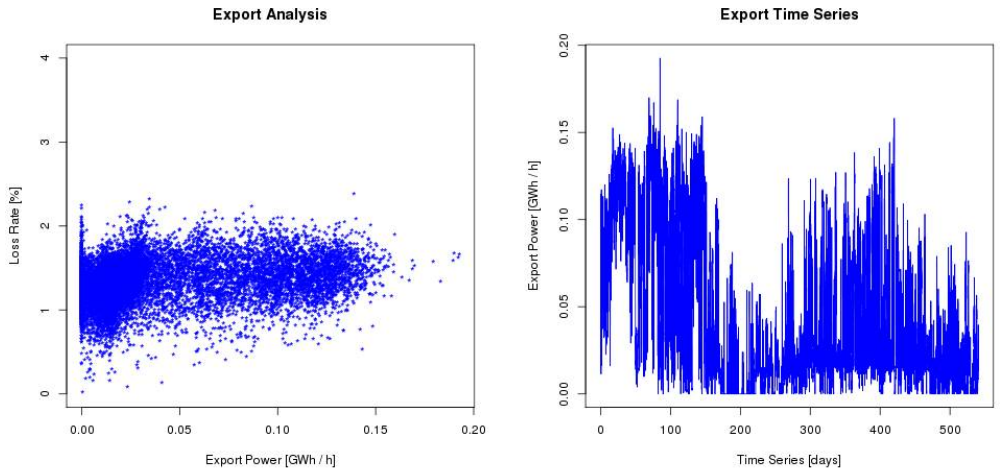
(a)

(b)



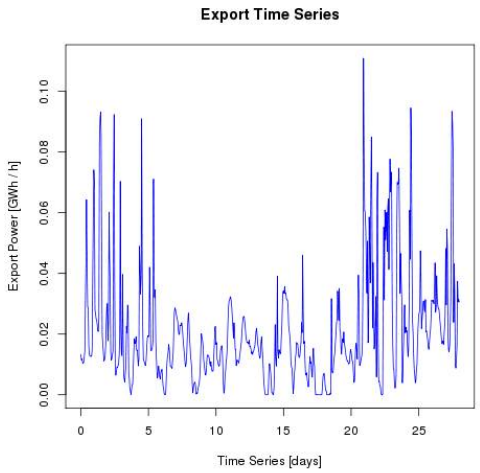
(c) April 2013

Figure B.8: The top left diagram show the losses as a function of import. The two other diagrams show the import over time; both for the whole data set and a snap shot of four weeks in April 2014



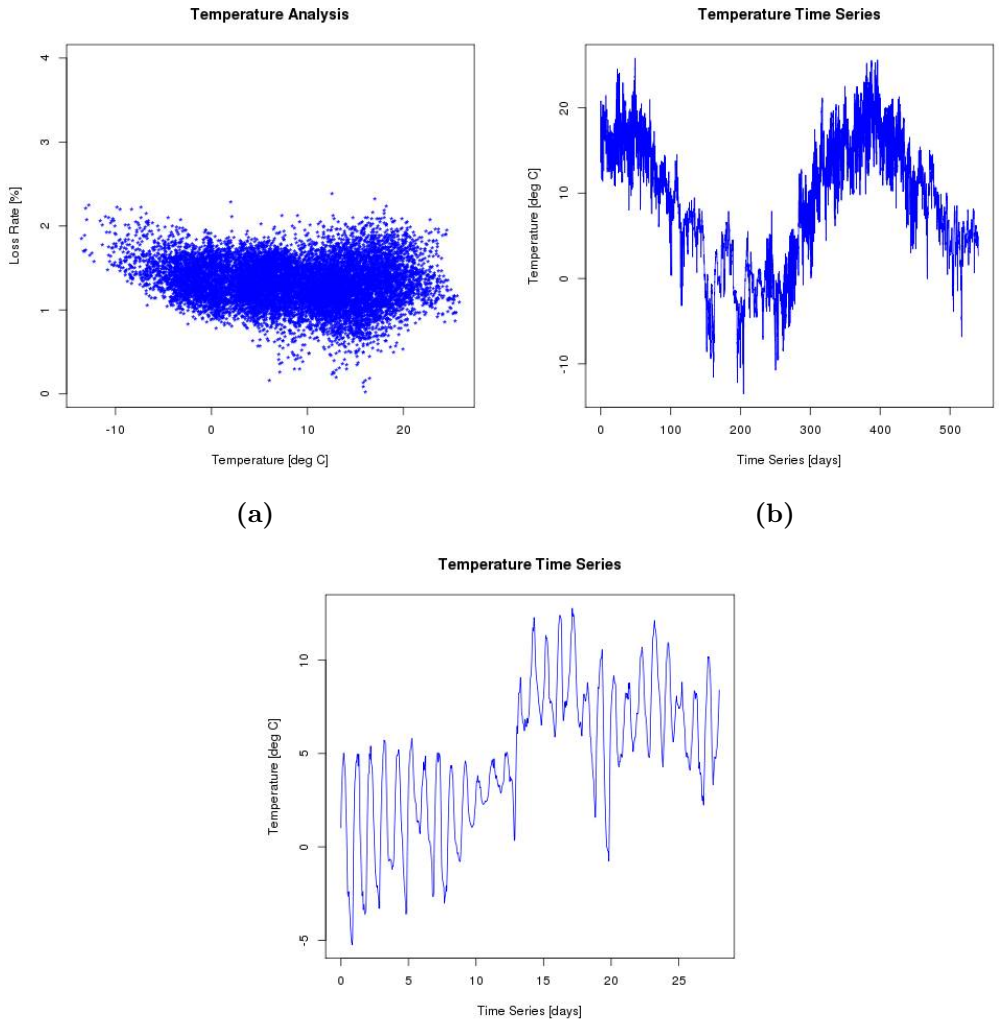
(a)

(b)



(c) April 2013

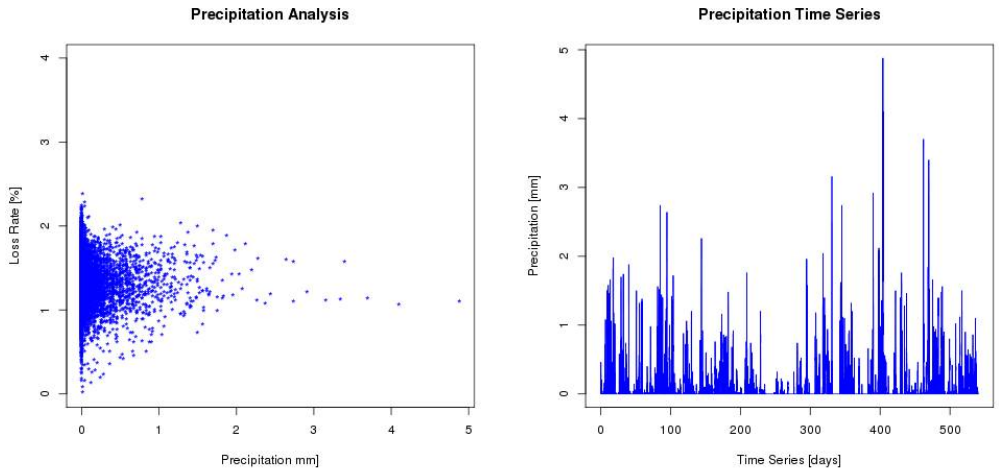
Figure B.9: The top left diagram show the losses as a function of export. The two other diagrams show the export over time; both for the whole data set and a snap shot of four weeks in April 2014



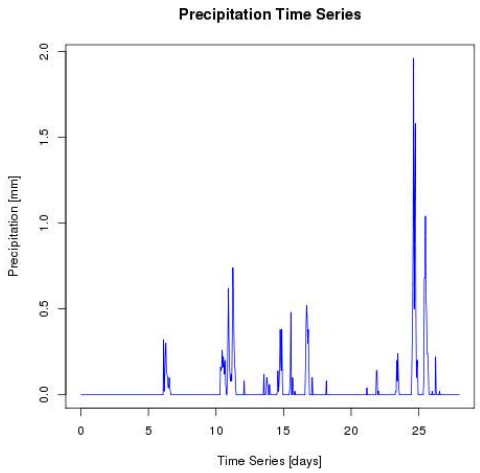
(a) **(b)**

(c)

Figure B.10: The top left diagram show the losses as a function of temperature. The two other diagrams show the temperature over time; both for the whole data set and a snap shot of four weeks in April 2014

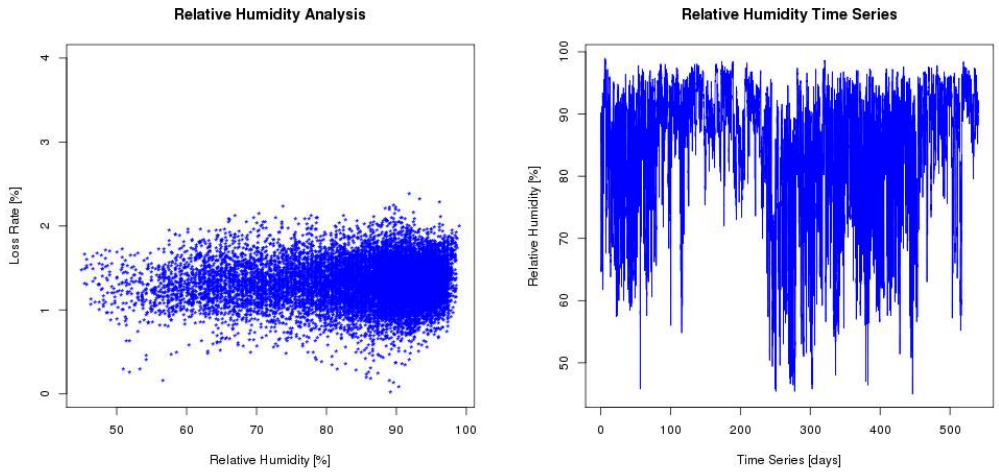


(a) **(b)**



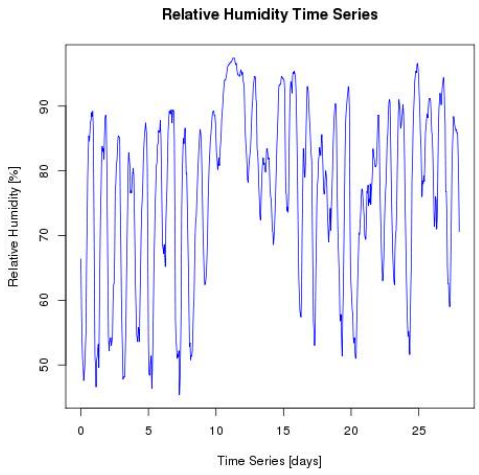
(c) April 2013

Figure B.11: The top left diagram show the losses as a function of precipitation. The two other diagrams show the precipitation over time; both for the whole data set and a snap shot of four weeks in April 2014



(a)

(b)



(c) April 2013

Figure B.12: The top left diagram show the losses as a function of relative humidity. The two other diagrams show the relative humidity over time; both for the whole data set and a snap shot of four weeks in April 2014

C

Regression Model Result

In this appendix the result of the regression analysis is presented as a list of $\hat{\beta}$- values.

Table C.1: Final model's parameters with the corresponding $\hat{\beta}$-values.

Parameter	$\hat{\beta}$ **-value**	**S.E.**	p**-value**
Intercept	1.77E+00	4.97E -01	3.76E-04
Feed-In	4.08E -01	1.27E -01	1.30E-03
Feed-In2	8.63E -02	2.31E -02	1.97E-04
WindP	-1.24E+00	4.64E -01	7.39E-03
WindP2	-7.78E -01	6.37E -01	2.22E-01
Temp	-4.32E -02	1.33E -02	1.19E-03
Temp2	9.65E -04	3.58E -04	7.10E-03
WindSpeed	1.75E -03	2.07E -02	9.33E-01
WindSpeed2	-4.54E -02	5.23E -03	8.07E-18
RH	7.35E -01	4.66E -01	1.15E-01
RH2	-7.06E -03	2.84E -01	9.80E-01
WindInstEffect	-2.22E+01	5.56E+00	6.89E-05
WindInstEffect2	8.95E+01	2.50E+01	3.58E-04
HydroP	5.74E+00	1.25E+00	4.45E-06
HydroP2	-2.48E+01	5.50E+00	7.21E-06
ThermalP	-1.41E+00	4.36E -01	1.29E-03

Continued on next page

Table C.1 – continued from previous page

Parameter	$\hat{\beta}$ **-value**	**S.E.**	**p-value**
ThermalP2	5.89E+00	2.12E+00	5.49E-03
Export	4.61E+00	1.22E+00	1.61E-04
Export2	-2.60E+01	1.41E+01	6.52E-02
Import	1.66E+00	7.55E -01	2.84E-02
Import2	4.93E+01	1.47E+01	8.26E-04
WindP.WindInstEffect	4.22E+00	1.68E+00	1.22E-02
WindP.Temp	-3.28E -02	6.81E -03	1.56E-06
Feed-In2.WindP2	-5.03E -02	2.55E -02	4.85E-02
WindP2.WindInstEffect2	1.76E+01	6.52E+00	6.91E-03
HydroP2.Export	-1.64E+01	8.04E+00	4.13E-02
Export.Export2	1.61E+02	5.99E+01	7.40E-03
Temp2.HydroP	4.48E -03	1.03E -03	1.35E-05
Temp2.Export	-5.41E -03	1.22E -03	9.98E-06
WindInstEffect.Import	-1.19E+01	2.83E+00	2.65E-05
WindSpeed.WindInstEffect2	-5.88E -01	3.00E -01	5.04E-02
Feed-In.Import2	-2.87E+01	1.04E+01	5.81E-03
Feed-In2.ThermalP2	-7.68E -02	3.27E -02	1.88E-02
WindSpeed2.WindInstEffect	3.65E -01	4.13E -02	2.30E-18
WindInstEffect2.HydroP2	5.02E+01	1.47E+01	6.72E-04
WindInstEffect2.Export	1.82E+01	6.95E+00	9.04E-03
Feed-In2.Export2	-1.36E+00	4.09E -01	9.00E-04
WindSpeed2.WindInstEffect2	-6.59E -01	8.42E -02	8.42E-15
RH2.Export	-2.53E+00	7.36E -01	6.08E-04
RH.WindInstEffect2	-7.47E+00	1.86E+00	6.25E-05
Feed-In2.Import2	4.06E+00	1.87E+00	3.03E-02
ThermalP.ThermalP2	-5.71E+00	2.79E+00	4.11E-02
Feed-In.HydroP	-2.03E+00	6.12E -01	9.49E-04
Feed-In.HydroP2	5.50E+00	1.93E+00	4.34E-03
Feed-In.WindInstEffect	-1.70E+00	2.24E -01	5.43E-14
WindP.Import	1.87E+00	7.59E -01	1.36E-02

Continued on next page

Table C.1 – continued from previous page

Parameter	$\hat{\beta}$ **-value**	**S.E.**	**p-value**
Export.Import2	2.18E+02	8.99E+01	1.54E-02
Feed-In.Temp	5.36E -03	2.58E -03	3.79E-02
Temp.WindInstEffect	1.29E -01	4.46E -02	3.75E-03
ThermalP2.Import2	3.14E+01	1.52E+01	3.95E-02
WindInstEffect.WindInstEffect2	-7.92E+01	3.56E+01	2.63E-02
WindSpeed2.ThermalP2	-6.32E -03	3.97E -03	1.11E-01
HydroP.HydroP2	2.30E+01	1.51E+01	1.28E-01
Temp2.WindInstEffect2	-1.60E -02	4.83E -03	9.21E-04

D

Future Work

This section is a summary of the recommended future analyses. It was further discussed in Chapter 6.

D.1 The Software and the Model

D.1.1 Lavastorm

The advantages using Lavastorm and R at Energicontrolling out weight the disadvantages. All problems with Lavastorm were connected with the R-node, except one. The exception was that the input reader, for .xlsx files, was not compatible with the characters Å, Ä. Ö and semicolon. A better solution to get around this issue could be to read in from the Pomax database directly, instead of going past Excel.

With time and conversations with Lavastorm will the R-node hopefully be easier to use. There are a few suggestions in how to get the node to work better. When running the graph with the regression analysis it is the R-nodes that take up most of the time. This is the case even for R-nodes with little code within. Thus, the first recommended investigation is to look into the possibilities to make the R-node run faster.

The second suggestion with respect to Lavastorm and the R-node would be to improve the outputs of the R-node. That the matrices are not complete, is a big disadvantage for Lavastorm, but even the restriction in allowed outputs. As the idea with the R-node is to do regression analysis and other statistical calculation the fact that the most general output in R, the summary of the model, is not allowed as an output is another big disadvantage with Lavastorm. The fact that the model could not be sent to another node, made it not possible to split up the model optimising process with the model analysis into two different modes. Recommendation would be either to be able to pass a model to another R-node, or that it would be easier to output the necessary information from the main R-node to another node where the analysis could take place. An example

in the template would be sufficient.

A third suggestion is to develop all R code outside the Lavastorm environment, in a R-program, and then when the code does what is desired, to create the R-node. By doing this, debugging would be much easier, as an R-program has mode purposely developed error messages. Moreover, using the R-program would make it easier for the programmer to find documentaries about certain built-in functions in R. It would be quicker to check if the code work as the running time would be shorter. On the other hand, the raw data would need to be extracted from Lavastorm and read into the R-program. The benefit of using Lavastorm might then be undermined.

D.1.2 The Model

Further model optimising is needed for regression model. It is recommended to investigate if a different function in R than step is suited for the optimisation of the model. As it is the number of $\hat{\beta}$ -values differs between each run. Another R-function could be the cross validation.

The result shows that the residuals still suffer from the autocorrelation, even when only a fifth of the data is used to make each measurement more independent. Thus further investigation is needed to improve the residuals. This projects limitation was that a quick-fix was used to decrease the problem. If more data was available a further decrease might be sufficient. More data will be available after time, or methods such as bootstrapping can be used. Another way to address the problems would be to look into the parameters. The autocorrelation is present as there are important parameters missing. Investigation in finding these parameters is highly recommended, as the best way to get solve autocorrelation is to add the missing parameters.

A third recommended investigation is to see how stochastic process will improve the model, and hence decrease the autocorrelation. A benefit with the approach using stochastic process would be that the resulting model would be better suited to predict future event, if wanted.

D.2 The System Boundaries

The region studied can be seen in Figure 1.1 on page 4. The region is one of the largest medium voltage networks in Sweden, as well as the most southern. It is strongly recommended to investigate how the model change with changing regions.

If different regions were studied, the geographic dependence of the losses and the parameters could be investigated. One difference between regions is that in north of Sweden the regions are net producers, whereas in the south they are net importers. Thus the effect from regional productions would most probable differ.

In a scaled down study, to a smaller region, but still within the same voltage level, would see how the affect from the local parameters, such as weather, would have on the losses. An example of a case study is to isolate the island Öland, also part of

Regionnät Syd. Another scaled down study that is recommended is to investigate how the parameters affect the losses on a low voltage level, closer to the consumers.

D.3 The Parameters

Both introduce more and remove parameters can be of interest for an improved model. As already discussed in this section, by adding the ”missing parameters” the autocorrelation in the residuals will disappear. But, which these parameters might be is harder to know. A suggestion is to try with different time series and see how the residuals react by there addition, individual and together. Reduction in parameters would be due to correlation between the parameters. As it is the wind speed is strongly correlated to the wind power production. Thus it is hard to identify the effects of a parameter on the losses when the other is present in the set. Further reasoning is needed when the correlation between losses and wind power production is studied, as the wind power capacity level does also affect the loss rate. Hence it is recommended to look into the correlation between each parameter further.

When there is a higher share of regional renewable energy producers in the system it is advised to study their affect on the losses. E.g. how solar and wave power does relate to the losses. Do they decrease the losses as the wind power seems to, or do they correlate to an increase as hydro and thermal power production?

The workshop's aim was to look into which parameters should and could be added to the model. See Appendix A for the specific parameters and suggestions to where to get the input data. Some preliminary research has been done in the inclusion of how the system is operated, by looking at the number of non-standard couplings. In the next section this preliminary research is presented.

D.3.1 Standard vs Non-Standard Couplings in Regionnät Syd

This factor may have an impact in the losses and it would be interesting to investigate the matter further. The parameter was not included in the model in the thesis since time was scarce and collecting the data requires time. The person that was contacted and who would be able to gather the necessary data is Fredrik Appenrodt at E.ON Elnät.

Why It Is of Interest

Having the couplings in normal mode is the most optimal way to run the system. But during maintenance work some couplings are rearranged and put in a non-standard mode. The hypothesis is that running the system with non-standard couplings will lead to increased power transmission losses. The non-standard couplings can e.g. lead to that the power gets longer transmission distances, and thereby increase the losses. Adding this variable to the statistical model might increase the accuracy and make it better

at describing the system or predicting the future. It will also give an opportunity to understand how much the losses are affected by this parameter.

Where Can the Information Be Found?

It is possible to get information about the standard and non-standard couplings in the UDW at E.ON Elnät. Every change or event in the system is registered and saved there. Lists covering all these events are available for the last month. For previous months, historical data files can be gathered. The system was upgraded in the autumn 2012 therefor this is the hard limit in the historical data collection.

Suggestions of How to Adapt the Data

The coupling mode can be given a binary value, e.g. 0 for standard and 1 for non-standard coupling. To be able to introduce this variable to the model, the data has to be provided as one value per hour for every hour preferable during the studied period from 2012-07-01. The data value will represent how many non-standard couplings there are during one specific hour. The data should be provided as an .xlsx file or .csv file to simplify the process of merging it with the other input data for the model. What is required is a list with two columns; one for the date and hour (YYYYMMHHTT) and one for the value of interest. The data available contains information about what changes that has occurred in the system and not the actual system mode at a specific hour. Therefore the required data has to be created out of the available information. This has to be performed backwards, by using the information about the actual system mode in real time and going backwards step by step to create the required data. One problem is that the switching between standard and non-standard coupling can happen quite fast and the data is gathered every second rather than every hour. Within one hour the coupling can be switched several times and then it is hard to say whether that hour should be given the value 0 or 1. The format of the value thus needs to be evaluated, and potentially to include a time-variable, to allow for non-standard coupling time.

The possibility of creating the data needed was discussed with Fredrik Appenrodt. He believed that it was achievable by creating a report in UDW Explorer or by using Oracle. However, it would require several days for him to do it. It is then necessary to limit the project to focus on a specific voltage level ($130kV$) and a specific area (Region Syd).

Bibliography

[1] International Energy Agency, Electric power transmission and distribution losses (% of output) (2013).
URL http://wdi.worldbank.org/table/5.11

[2] S. Pandur, Löpande kostnader i förhandsregleringen - grundprinciper vid beräkning (2010).
URL http://www.energimarknadsinspektionen.se/Documents/Forhandsreglering_el/Viktiga_dokument/Lopande_kostnader_i_forhandsregleringen_grundprinciper_vid_berakning.pdf

[3] N. Silver, The Signal and the Noise: Why So Many Predictions Fail-but Some Don't, Penguin Press, 2012.

[4] S. Nivhede, C. Högstedt, Nätavräkningsområden i södra sverige, E.ON Elnät Sverige AB (2012).

[5] A. von Meier, Electric power systems: a conceptual introduction, John Wiley and Sons, 2006.

[6] A. Sumper, A. Baggini, Electrical energy efficiency: technologies and applications, Wiley. com, 2012.

[7] K. Sempler, Så höga är förlusterna i elnäten (November 2009).
URL http://www.nyteknik.se/popular_teknik/teknikfragan/article265620.ece

[8] S. K. Injeti, D. N. P. Kumar, Optimal planning of distributed generation for improved voltage stability and loss reduction, International Journal of Computer Applications 15 (1) (2011) 40–41.

[9] A. Hayter, Probability and Statistics for Engineers and Scientists, PWS Publishing Company, 2006.

[10] Sewaqu, Linear regression plot - wikipedia (November 2010).
URL http://en.wikipedia.org/wiki/File:Linear_regression.svg

[11] NIST/SEMATECH, E-handbook of statistical methods (April 2012).
URL http://www.itl.nist.gov/div898/handbook/eda/section3/eda3672.htm

[12] Stpasha, Confidence bands - wikipedia (July 2009).
URL http://en.wikipedia.org/wiki/File:Okuns_law_with_confidence_bands.svg

[13] SAS Institute Inc., Adding prediction and confidence bands to a regression plot (March 2014).
URL http://support.sas.com/documentation/cdl/en/grstatproc/62603/HTML/default/a003155517.htm

[14] OriginLab Corporation, Graphic residual analysis (April 2014).
URL http://www.originlab.com/doc/Origin-Help/Residual-Plot-Analysis

[15] S. Jost, Linear regression (April 2014).
URL http://condor.depaul.edu/sjost/it223/documents/regress.htm

[16] STATS4STEM.ORG, R: Normal distribution (2013).
URL http://www.stats4stem.org/r-normal-distribution.html

[17] S. Chatterjee, J. S. Simonoff, Handbook of Regression Analysis, Handbook of Regression Analysis, Wiley, 2013.
URL http://books.google.se/books?id=X77JngEACAAJ

[18] MathWorks, Statistics Toolbox^{TM} User's Guide, The MathWorks, Inc, 2013.

[19] P. Chennamaneni, R. Echambadi, J. D. Hess, N. Syam, How do you properly diagnose harmful collinearity in moderated regressions? (2008).
URL http://jindal.utdallas.edu/files/19.pdf

[20] B. Ratner, Statistical and machine-learning data mining : techniques for better predictive modeling and analysis of big data, Taylor & Francis, 2012.

[21] J. S. Milton, C. A. Jesse, Introduction to Probability and Statistics, McGraw-Hill, 2003.

[22] H. Akaike, G. Kitagawa, The Practice of Time Series Analysis, Springer, 1999.

[23] S. Fortmann-Roe, Accurately measuring model prediction error (May 2011).
URL http://scott.fortmann-roe.com/docs/MeasuringError.html

[24] Lavastorm Analytics, Lavastorm analytics (2014).
URL http://www.lavastorm.com

[25] Institute for Statistics and Mathematics of WU (Wirtschaftsuniversität Wien), The r project for statistical computing (2014).
URL `http://www.r-project.org`

[26] SMHI, Meteorologiska observationer (April 2014).
URL `http://opendata-download-metobs.smhi.se/explore/`