

B.Sc. Thesis  
Bachelor of Science in Mathematics and Technology

**DTU Compute**

Department of Applied Mathematics and Computer Science

# Statistical models for analysis of frequent readings of electricity, water and heat consumption from smart meters

In cooperation with SEAS-NVE

Anton Stockmarr (s164170)

Ida Riis Jensen (s161777)

Mikkel Laursen (s164199)

Kongens Lyngby 2019



**DTU Compute**  
**Department of Applied Mathematics and Computer Science**  
**Technical University of Denmark**

Matematiktorvet  
Building 303B  
2800 Kongens Lyngby, Denmark  
Phone +45 4525 3031  
[compute@compute.dtu.dk](mailto:compute@compute.dtu.dk)  
[www.compute.dtu.dk](http://www.compute.dtu.dk)

# Summary

---

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.



# Resumé

---

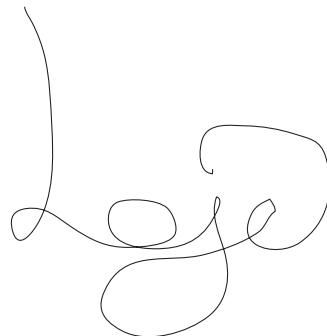


# Preface

---

This xxx thesis was prepared at the department of Applied Mathematics and Computer Science at the Technical University of Denmark in fulfillment of the requirements for acquiring a yyy degree in zzz.

Kongens Lyngby, June 14, 2019



Anton Stockmarr (s164170)  
Ida Riis Jensen (s161777)  
Mikkel Laursen (s164199)



# Contents

---

<b>Summary</b>	i
<b>Resumé</b>	iii
<b>Preface</b>	v
<b>Contents</b>	vii
<b>1 Introduction</b>	1
1.1 Motivation . . . . .	1
1.2 Introduction to WATTS app . . . . .	1
<b>2 Data</b>	3
2.1 Original data . . . . .	3
2.2 Cleaning and preparation . . . . .	5
2.2.1 Missing values . . . . .	5
2.2.2 The sun and the wind . . . . .	6
2.2.3 Data checking . . . . .	6
<b>3 Exploratory Analysis</b>	7
3.1 Examination of the Heat Consumption . . . . .	7
3.1.1 Weather data . . . . .	8
3.1.2 BBR data . . . . .	9
3.2 Multicollinearity . . . . .	11
3.3 Data segmentation . . . . .	12
3.3.1 Segmentation by piece-wise optimization . . . . .	14
3.3.2 Segmentation by significant deviations . . . . .	14
<b>4 Models on the Daily Consumption</b>	19
4.1 Linear regression . . . . .	19
4.1.1 Model assumptions . . . . .	19
4.2 Simple linear regression model . . . . .	21
4.2.1 Validation . . . . .	21
4.2.2 Results . . . . .	22
4.3 Multiple linear regression model . . . . .	23

4.3.1	Splines . . . . .	25
4.3.2	Results . . . . .	26
4.4	Regression model for comparing houses . . . . .	28
4.4.1	Validation . . . . .	28
4.4.2	Results . . . . .	28
4.4.3	Predictions . . . . .	32
4.5	Comparison . . . . .	32
4.6	Visualization of the results . . . . .	32
<b>5</b>	<b>Models on the Hourly Consumption</b>	<b>35</b>
5.1	Description of the Hourly Consumption . . . . .	35
5.2	The ARMA Models and Their Extensions . . . . .	38
5.3	Applying the models . . . . .	41
5.3.1	Seasonal model . . . . .	41
5.3.2	Physical model . . . . .	42
<b>A</b>	<b>Tables</b>	<b>45</b>
A.1	Estimates and test of the simple linear regression model . . . . .	45
A.2	Estimates and test of the multiple linear regression model . . . . .	45
A.2.1	Significance of parameters for full model . . . . .	45
<b>B</b>	<b>Figures</b>	<b>51</b>
<b>Bibliography</b>		<b>55</b>

# CHAPTER 1

# Introduction

---

According to the International Energy Agency [3], heat is the largest energy end-use. Providing heating for homes and industrial purposes accounts for around 50% of the total energy consumption. Renewable heat consumption in the form of bioenergy contribution is expected to grow which will be a better solution for the climate. In relation to the individual consumer, it makes sense to become aware of one's heat consumption, e.g many consumers pay more for their heat consumption than they could. This can be solved by making small adjustments such as replacing radiators with more efficient cooling, replacement of leaky windows, improve insulation of the house etc. Which factors that can influence the heat consumption, are not known to most consumers and thus it can be a challenge to know how to minimize the consumption.

Heat consumption can be described using mathematical models, namely statistical models, and this can lead to an optimization/minimization of the consumption.

## 1.1 Motivation

The aim of this report is to investigate the tap water consumption and thereby provide possible extensions to the app created by SEAS-NVE, WATTS. By illustrating these features in the app, customers can become aware of their heat consumption and at the same time get a sense of what physical phenomena affect their house.

Our approach is to develop statistical models in order to analyse which factors influence the heat consumption.

## 1.2 Introduction to WATTS app

The app WATTS is designed and created by the danish energy and optical fibre broadband concern, SEAS-NVE. The app provides an overview of the energy consumption to the consumer by showing the actual consumption and predicting the expected consumption. In addition, the app keeps track of the consumers budget and give the user the opportunity to compare their consumption with similar customers. The energy consumption in relation to the expected consumption is visualised with the colours

green, yellow and red. The colours are used to indicate whether the consumption is expected to be lower than expected, to exceed the budget by 0-30%, or to exceed the budget by more than 30%. The app is under expansion such that users are offered the same applications for their heat consumption. In continuation of this, it is possible to add a feature showing the wind dependency on the heat consumption.

# CHAPTER 2

# Data

---

Idas far kan lige skrive her om fjernvarmesystemet. tilføj hvordan fjernvarmesystemet er opbygget og hvordan målingerne er foretaget. når vandet kommer ind i huset, sidder der en måler, som mäter Temperaturen på det der kommer ind og det der kommer ud igen samt mængden af det.

The data is provided by SEAS-NVE in three data sets. The house data consists of 71 .csv-files containing 8 attributes for each house which is 513877 data points in all. The second data set includes weather data containing 10,140 observations and predictions of the next 2283 data points, all with 11 attributes. Furthermore, the third data set is from Bygnings- og Boligregistret (BBR) and contain details for each of the houses e.g. total area, year of construction and type of house. The main focus of this section will be how this data is cleaned and prepared for the further analysis.

## 2.1 Original data

The original house and weather data include hourly observations from the period 31-12-2017 23:00 to 7-02-2019 10:00. The time period varies in the house data, i.e there exists data for houses containing observations for approximately six months and there are also houses that contain observations for about 13 months. This will be taken into account when cleaning the data and in the analysis the two groups of houses will be referred to as the long and short houses. Before the data was provided to this project, it had already been slightly altered. When Aalborg Forsyning, who provided the data, received the observations from the hardware installed in the houses of their clients, it is generally not divided into precise one hour intervals. Aalborg Forsyning interprets the time stamps on the observations and changes them to one hour intervals. If observations are missing, an interpolation method is used to simulate the data. The exact procedure is described in [2].

Table 2.1 below shows the attributes from the house data set which is used to define the Heat Consumption as

$$Q = c \cdot m \cdot \Delta T, \quad (2.1)$$

where  $c$  is the specific heat capacity for water which is  $4.186 \text{ kJ/kg} \cdot \text{C}^\circ$ ,  $m$  is the water density in  $\text{kg}$  and is calculated from the volume as  $1\text{m}^3 = 1000 \text{ kg}$  and  $\Delta T$  is

the cooling degree in  $C^\circ$ . In this way, the heat consumption is converted from  $kJ$  to  $kwh$  (since  $1\ kwh = 3600\ kJ$ ) which is the desired unit for the heat consumption.

Variable	Description
StartTime	Start time and date for measurements.
EndTime	End time and date for measurements.
Energy	Consumption in $kWh$ .
Flow	Amount of water passed through meter in $m^3/hour$ .
Volume	in $m^3$ .
TemperatureIn	Temp. of the water flowing into a house in Degrees/C.
TemperatureOut	Temp. of the water flowing out of a house in Degrees/C.
CoolingDegree	Difference between Temp.In and Temp.Out in Degrees/C.

**Table 2.1:** Attributes from the original house data..

Table 2.2 shows the attributes in the weather data and Table 2.3 shows which attributes the BBR data consists of.

Variable	Description
StartTime	Start time and date for measurements. Hourly values.
Temperature	Temperature outside in Degrees/C.
WindSpeed	Wind speed in $m/s$
WindDirection	Wind direction i degrees from 0 to 360, 0 being North
SunHour	The level of sunshine in the hour in a scale from 0 to 1
Condition	The weather condition given in numbers described in [5]
UltravioletIndex	The UV index level
MeanSeaLevelPressure	
DewPoint	
Humidity	
PrecipitationProbability	
IsHistoricalEstimated	Binary variable, true if the datapoint is a prediction

**Table 2.2:** Attributes from the original weather data..

Variable	Description
Key	The house ID key
HouseType	Type of house: Apartment, house, industrial etc.
TotalArea	The total area of the house in $m^2$
Floors	The number of floors in the house
Basement	How many $m^2$ basement there is in the house
Attic	How many $m^2$ attic there is in the house
ConstructionYear	The year of construction for the house
Surfaces	The material on the surface of the outdoor walls of the house
ReconstructionYear	The year of the latest reconstruction of the house
AdditionalHeating	If there are any additional heating installed in the house. Fireplace etc.

**Table 2.3:** Attributes from the BBR data..

## 2.2 Cleaning and preparation

In this section, it is described how the raw data is cleaned and prepared for the statistical analysis.

Due to the fact, that `StartTime` and `EndTime` is always one hour apart, it is redundant to use both of the attributes. The observations of most of the attributes are made at time `EndTime`, and for that reason it is used as `ObsTime` for the observations. For the weather data set, the observations is made at time `StartTime`, and there is no `EndTime` for this data set. When merging these data sets, `ObsTime` is alligned with `StartTime`. The format of these attributes is changed to a `Posixct` value with d-m-Y H:min:sec as the structure.

Every now and then, one or more data points in a row are missing. When this happen, a data point with NA-values for all of the attributes except `ObsTime`, is placed in the data set, which makes the data set easier to use in the modelling process. In some cases, the models that the data set is "complete", i.e. that there are no missing values. This is true for the time series models that will be described in section [citer section](#) In the data sets there are no indication of whether or not it is weekend. This attribute is added as well as the school holidays.

Both weather data and the house data are aggregated with mean values for each day in order to convert hourly values into daily values since there are of interest when modelling in chapter 3, two of the attributes is aggregated in a different way, which is explained later.

### 2.2.1 Missing values

In the vast majority of our houses, the data is consistent with data every hour from a given starttime until the end of January 2019. Then for many of the houses, there is

a gap of 6 whole days in the start of February, and then 2 more days of data. For this reason, the February data is cut out completely. The rest of the missing readings are handled so that they are based on the values of readings around the gap, i.e. they are interpolated by calculating the average of the readings around the missing reading. The purpose of filling out the missing values with interpolation of data, is to ensure that there exist observations for every hour. This makes it possible to develop time series models to describe data and also predict customers' consumption on an hourly basis.

### 2.2.2 The sun and the wind

A physical factor that could possibly affect the heat consumption is the sun. In raw data, the attributes `Condition`, `SunHour`, and `UltraVioletIndex` can be seen as explanatory variables for the sun. Instead, an attribute, `Radiation`, is added to calculate the solar radiation for a given day. This attribute is determined with use of the R function `calcSol` from the library `solaR`. The ultraviolet index is a measurement of the strength of ultraviolet radiation and since the attribute `Radiation` is more exact, `UltraVioletIndex` is removed from the weather data set.

Another physical factor that might be of importance is the wind. There are data available for both the wind direction in degrees and the wind speed. When the data is aggregated into daily values, it is important to pay special attention to the wind attributes, since it is not logical to take the average of degree values. For example, the average wind direction of 359 degrees and 0 degrees is not 179.5 degrees. Instead the wind direction and wind speed are interpreted as polar coordinates in a coordinate system. They are converted to rectangular coordinates. Then they are aggregated from hourly values into daily values, and returned to polar coordinates. When the wind is aggregated this way, wind directions with high wind speeds are weighted higher than wind directions with low wind speeds. Also the problem with the periodicity of the wind direction is solved.

### 2.2.3 Data checking

As mentioned, there are some measurements missing in the house data and it can therefore be difficult to do modelling for the houses in question. To avoid these difficulties, a so called "Data Checking" function has been made in order to check whether several constraints for the data are fulfilled. There must be a certain number of observations and the amount of missing data should not exceed a certain fraction of the data observation period.

# CHAPTER 3

# Exploratory Analysis

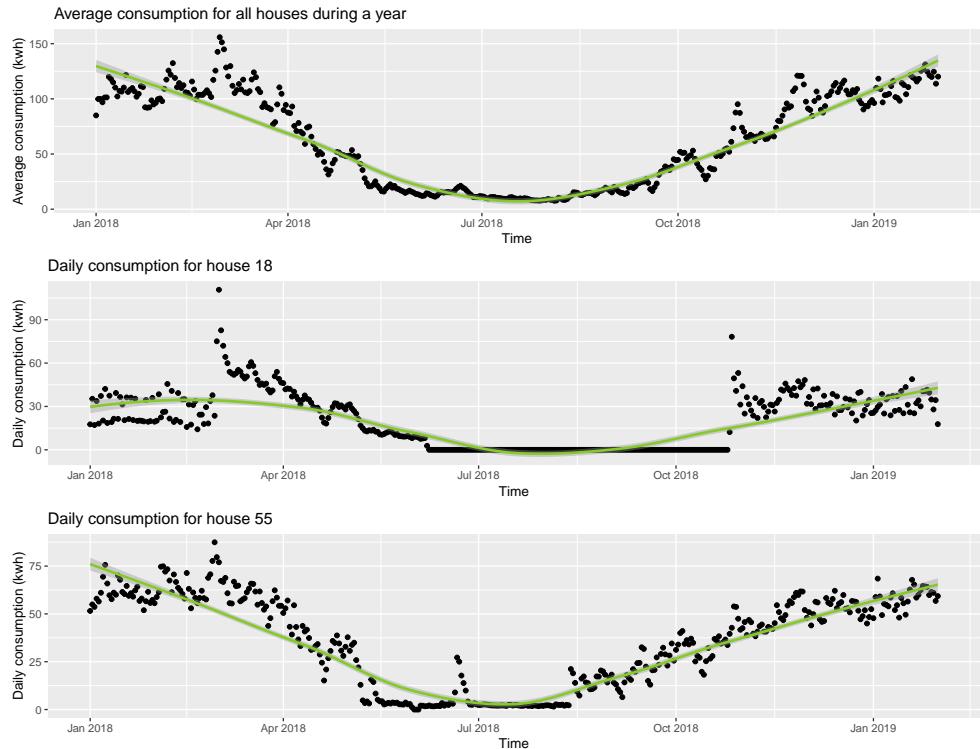
---

First part of the analysis is to explore the different attributes in the data, in order to detect possible patterns or correlations. The exploratory analysis is also used to get an understanding of data and its behaviour. Hence, this chapter is about visualizing the different attributes focusing on their influence on the heat consumption. As the heat in each house is turned off in the summer period, data is segmented such that the summer period is excluded from the data used for modeling.

## 3.1 Examination of the Heat Consumption

To get an overview of the heat consumption for each house, the daily average heat consumption for each house is investigated. Figure 3.1 shows the daily average consumption for all the houses and the daily consumption of two houses - one that follows the trend and one that deviates. These two houses are chosen to be visualised throughout the report and their specifications are presented in section 3.1.2. It can be seen that the slopes around the summer months are close to 0. As mentioned, the data in focus in this project is where the heat is turned on, hence the period where the heat consumption is close to 0 needs to be removed. Exactly how this is done will be explained and discussed in the data segmentation section. All three plots show some unusual high data points around April 2018. This can be due to the fact that it was snowing in Denmark at that time which is supported by the article found in [1].

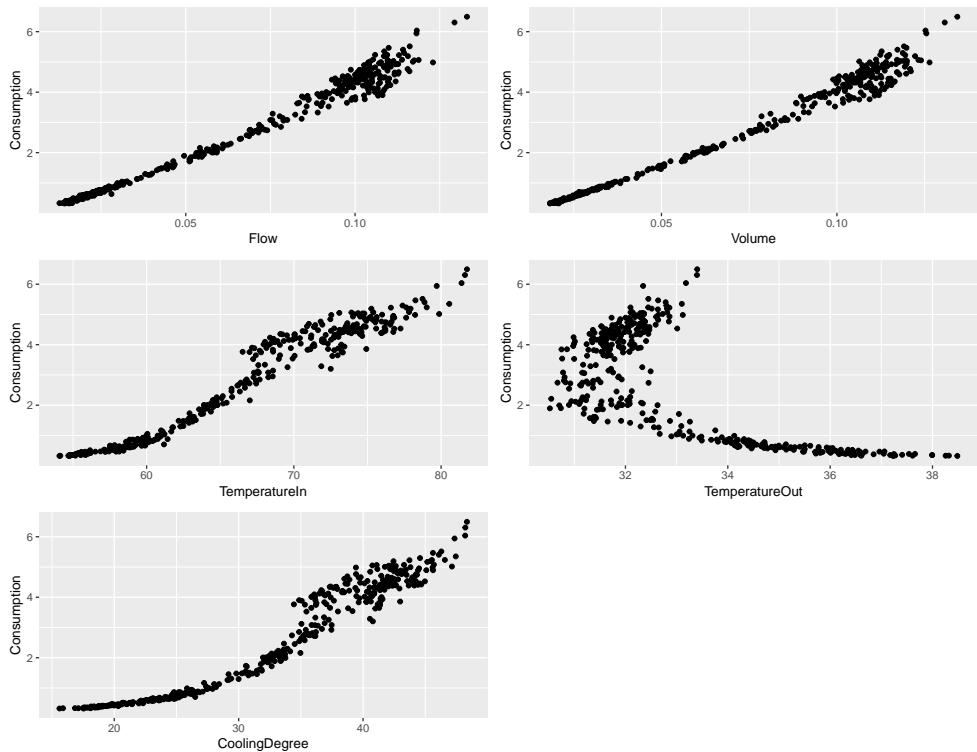
The remaining attributes from the house data is examined through a scatterplot shown in Figure 3.2 and Figure B.1, in order to find possible linear relationships with the consumption. There are clear linear relationships between the consumption and the flow, the volume, the cooling degree and the temperature going in respectively. It is expected that the consumption depends linearly on the volume and cooling degree, cf. the main equation given in (2.1). The relationships between consumption and the flow and volume are quite similar which is in line with the description of the two attributes given in Table 2.1. It is also seen that the temperature of the water coming out of the system depends on how much is used. So if the return water is quite hot, the house has not fully utilized the heat for the heat consumption.



**Figure 3.1:** Daily consumption during a year (2018). The top plot shows the average consumption for all the houses. The plot in the middle shows an example of a house that deviates from the trend and the last plot shows a house that follows the trend.

### 3.1.1 Weather data

The weather data is also examined through scatterplots given in Figure 3.3 and Figure B.2 in order to detect dependencies between the average consumption of the houses and the weather attributes. It is already known that the outside temperature has a significant influence on the consumption which is in line with the linearly relationship between the temperature and the consumption in Figure 3.3. Furthermore, consumption is approximately influenced by the dew point which is probably explained by the linear relationship between Temperature and DewPoint illustrated in Figure B.2. The scatterplots show that the consumption overall is independent of the attributes WindSpeed, WindDirection and MeanSeaLevelPressure. **Synes I der mangler noget her?**



**Figure 3.2:** Scatterplots of the daily values of relevant house attributes. There are clear linearly dependencies between Consumption and e.g. Flow.

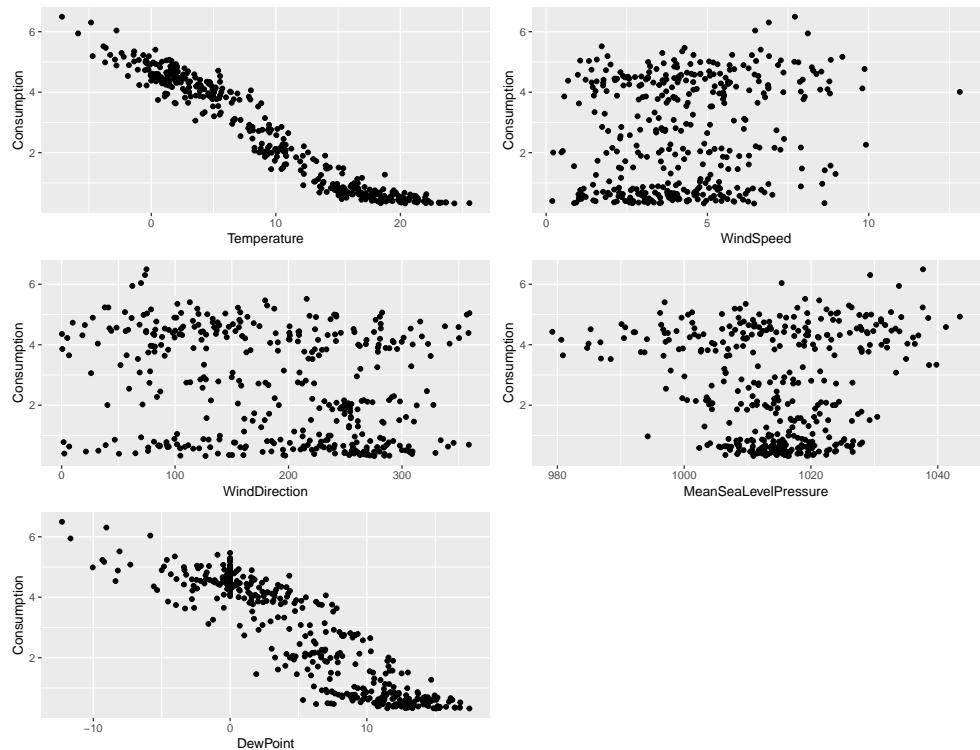
### 3.1.2 BBR data

Presumably, the BBR data has influence on the heat consumption in particular the total area and year of construction. Hence, Figure 3.4 illustrates the house specifications of all houses focusing on the type of house, the total area, the year of construction and reconstruction. [Lige knytte nogle kommentarer til histogrammerne](#). Since the two houses, 18 and 55, showed in Figure 3.1 will be used throughout the report, it is important to know their specifications:

#### House 18:

- House type: Parcel
- Area:  $128 \text{ m}^2 + \text{attic of } 34 \text{ m}^2$
- Year of construction: 1927

- Reconstructed in 1998



**Figure 3.3:** Scatterplots of the daily values of relevant weather attributes. There are clear linearly dependencies between Consumption and Temperature, as was expected.

#### House 55:

- House type: Parcel
- Area:  $160\ m^2$

- Year of construction: 1971

- Has wood-burning stove

The areas of the houses are somewhat similar, they are of the same house type and their ages are close to each other. However, house 18 has an odd behaviour which will be evident in the later results.

The average of the heat consumption for each house is found/determined for the winter period. By dividing the average consumption with the total area of the house the consumption pr.  $m^2$  is calculated. Figure 3.5 shows the year of construction and the consumption for each of the houses. The year of construction is here determined

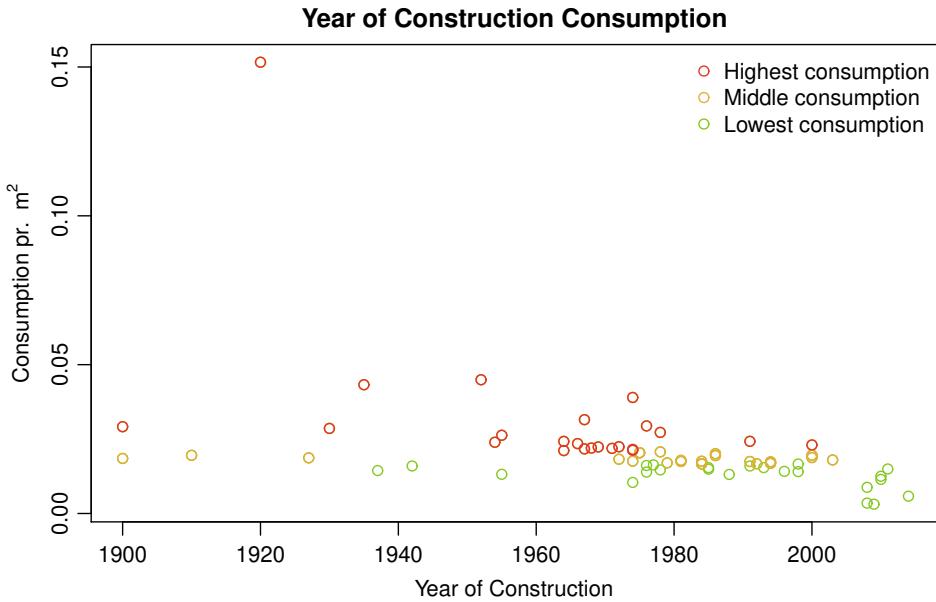


**Figure 3.4:** Details of the houses from BBR data focusing on the type of house, the total area in  $m^2$ , the year of construction and reconstruction respectively.

by either the year of construction or the year of the latest reconstruction of a house. Figure 3.5 clearly shows that the later a house is constructed (or reconstructed), the better is the insulation of the house as the consumption decreases with the year of construction. Furthermore, there is a clear outlier in the figure which has a remarkable high consumption pr.  $m^2$ . When looking up the house in the BBR data, it is seen that the outlier is an apartment of 61  $m^2$  build in 1920.

## 3.2 Multicollinearity

Multicollinearity occurs when two or more explanatory variables are highly correlated. In linear regression, multicollinearity ... Multicollinearity can be investigated by calculating the correlation using the function `cor()` in R.



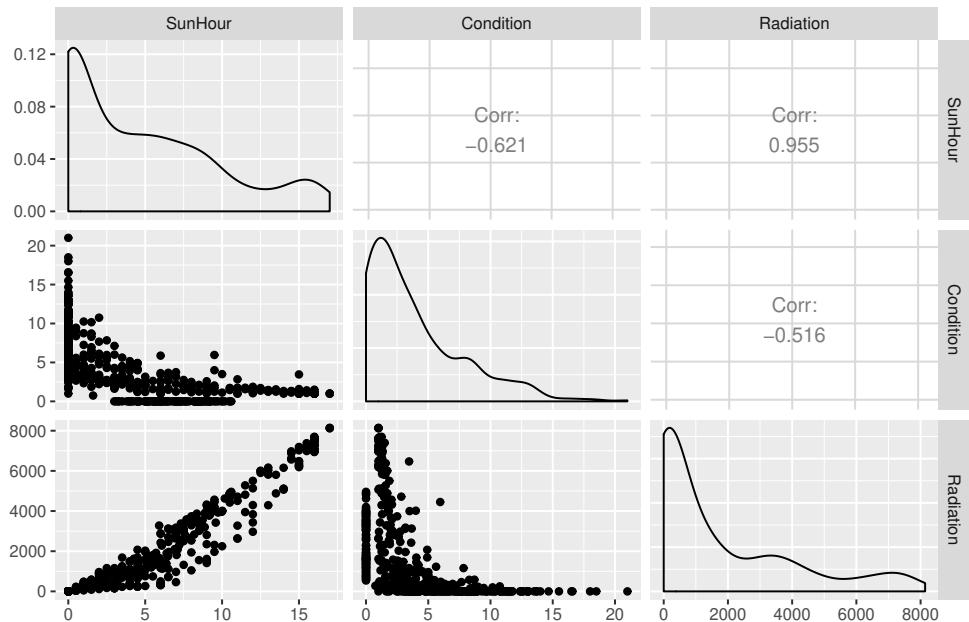
**Figure 3.5:** Plot showing the year of construction and the average consumption pr.  $m^2$  for each house. It is seen that there is a tendency that the later a house is built or reconstructed, the better is the insulation of the house.

Figure B.2 clearly shows that there is a high correlation between **Temperature** and **Dewpoint**. The exact correlation between the two attributes is calculated at 0.936, hence it is decided to remove **Dewpoint**. Furthermore, it is assumed that **Radiation** is a replacement for the attributes describing the sun, namely **Condition** and **SunHour**. This is the basis for expecting a correlation between the radiation and the sun attributes. Figure 3.6 shows a plot of the correlation matrix between the abovementioned attributes. There is a high correlation between **Radiation** and **SunHour** at 0.955, thus **SunHour** is removed from the weather data set.

The complete data set used for modeling in chapter 4 can be seen in table 3.1. **Not done**

### 3.3 Data segmentation

Since one of the focuses of this paper is to estimate how much energy a house uses for heating depending on different outside temperatures, it is important to distinguish



**Figure 3.6:** Scatterplot showing the correlations between the three attributes Condition, Radiation and SunHour. It is clearly seen that the radiation and the sun hour are highly correlated.

Variable	Description
Date	End time and date for measurements. Hourly values.
Temperature	Temperature outside in Degrees/C.
WindSpeed	
WindDirection	
Condition	
MeanSeaLevelPressure	Avg. atmospheric pressure at mean sea level in mbar.
PrecipitationProbability	Measure of the probability that precipitation will occur.
Observation	The number of observations for each day for each house.
Consumption	CoolingDegree times Volume from House data
Holiday	A categorical attribute with 6 levels: Working day, Weekend, Autumn break, Christmas break, Winter break and Spring break.

**Table 3.1:** Attributes used for modeling.

between when the house is actually being heated, and when the water is just being used for tap water consumption. If the inhabitants are not home for a longer period, there will probably be low consumption, even though it might be cold outside. This does not necessarily mean that the house is well isolated. And if there is consumption in warm periods, it is likely to be tap water consumption, and not heating. The data can be seen as part of two different distributions. One where the heating is turned off, and one where it is turned on. In this section different approaches will be examined on how to distinguish between the two distributions. The goal is to find some temperature, where it can be assumed that all data points below it belongs to the distribution with heating turned on. Two approaches will be described below, together with their pros and cons.

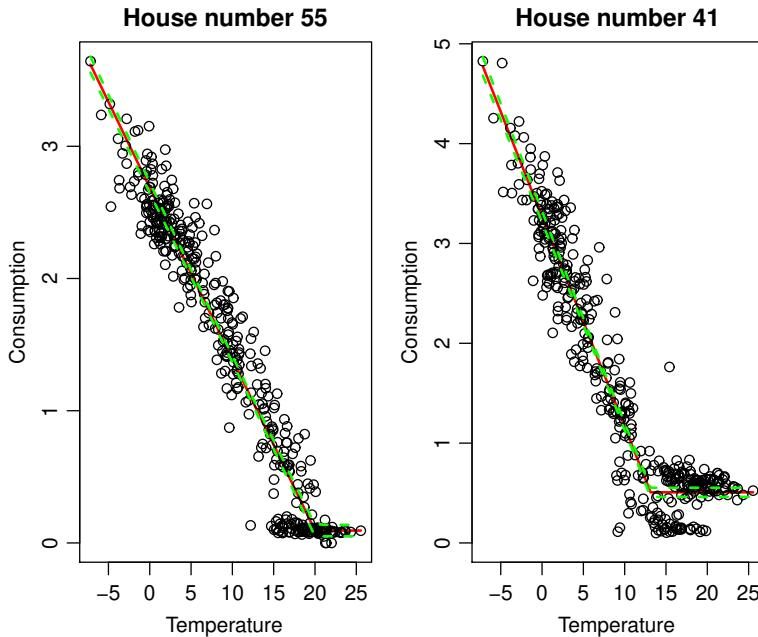
### 3.3.1 Segmentation by piece-wise optimization

The first approach is to make a linear regression on the data with two segments. A breakpoint  $\alpha$  is found, such that the SSE is as small as possible. The second segment is restricted to being constant. This way the breakpoint illustrates when the consumption goes from being linearly dependent on the temperature, to having a constant value. This method was tested on every available house, where a new breakpoint was found for each house.

Figure 3.7 shows the regression for two different houses. On both houses the line fits rather well with the low-temperature data points. But it is not very accurate around the breakpoint. The house on the left shows very clearly, that the assumption that all points below the breakpoint belong to the distribution without heating, is not accurate. Even though this approach can easily take out a lot of data where there is clearly no heating, it will in many cases set the breakpoint too high. The "tail" of the low consumption distribution might still be included, causing a bias in the model, and some variation that is not accounted for. The method is also not very robust. Depending on how the points are spread out, the breakpoint is sometimes as high as 20 degrees, which is not desirable.

### 3.3.2 Segmentation by significant deviations

In the second approach, the data points are examined from high temperatures to low. First, all data points from above 20 degrees are assumed to belong to the distribution without heating. If a data point is more than two standard deviations above from the mean of this distribution, it is assumed to belong to the distribution with heating. Now the data points are divided by temperature into one degree intervals. For each interval, starting from above and moving down, all data points in that interval are examined. The last interval where at least 20% of the data points are less than two standard deviations away, is chosen as the breakpoint of that house. An example of the approach is seen on figure 3.8. On the left the data points are plotted with standard deviations on the y-axis. The red line highlights the two standard deviations.



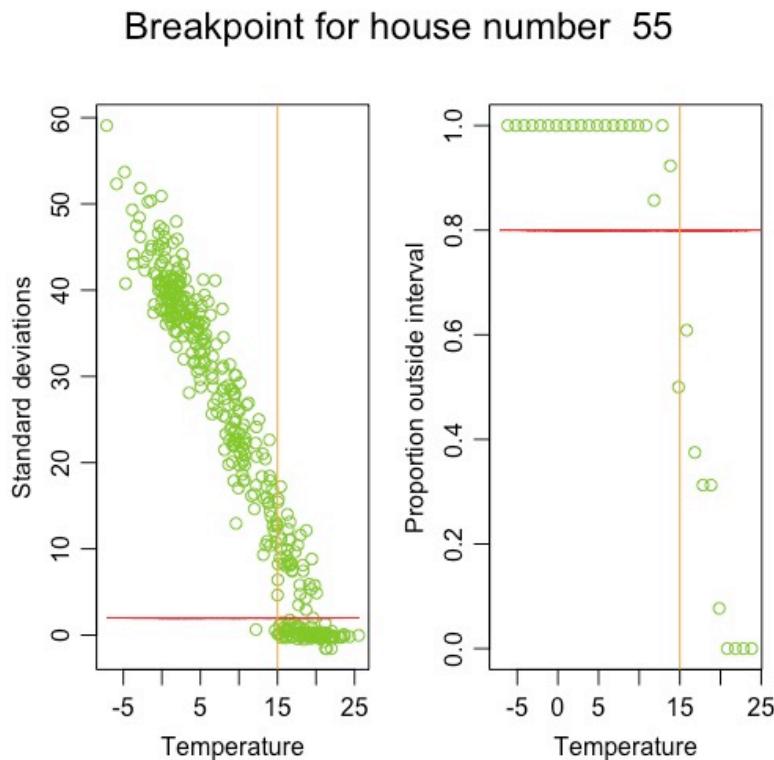
**Figure 3.7:** Piece-wise optimization of the consumption. The red line is the regression line and the green line is the confidence interval..

On the right there is a plot showing how many of the data points that are outside the interval. Here, the red line shows the 80% that determine the breakpoint. The orange line shows the breakpoint.

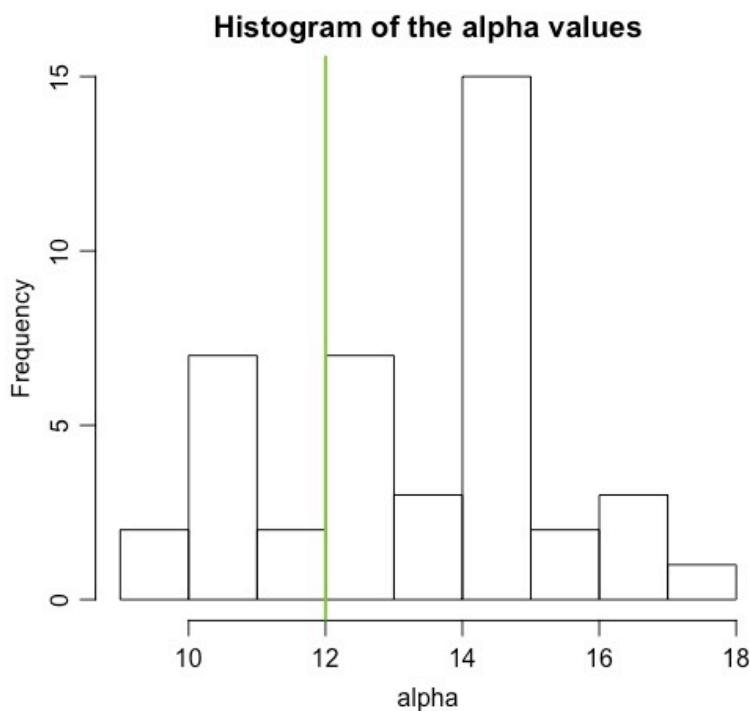
This model is more robust than the first. It is more selective, and provides a good way to set the breakpoint on the correct side of the mentioned "tail" that may occur at temperatures both with and without heating. When comparing figure 3.8 to figure 3.7, one can see that this method sets the breakpoint a bit lower, removing more points without heating. If the consumption data behaves badly, and chunks of datapoints are low enough to be within the two standard deviation, then a lot of data can potentially be removed, and there might be too little data left.

Until now the focus has been to find a breakpoint for every individual house. But it might be preferable to have a single breakpoint all houses. This way the segmentation becomes more robust to houses with unforeseen heat consumption. Figure 3.9 shows a histogram of the breakpoint values for every house in the data set. The global breakpoint should be in the low end of the scale. It is better to remove data points that could have been used, than to include too many points that belong to a different distribution with a different variation, which could make the assumptions of the model worse. It would not be good to choose the minimum breakpoint, since that would be

very vulnerable. A single house with a very low breakpoint might make the model bad for all the other houses. So the breakpoint that is chosen is the first quantile. As it is shown on the figure, this is 12 degrees. All models in the following sections will only be considering data where the temperature less than or equal to 12 degrees.



**Figure 3.8:** An illustration of how the breakpoint is found using segmentation by significant deviations. On the left figure the line illustrates two standard deviations from the high temperature distribution. The right figure shows how many points are outside the two standard deviations. The last point below 80% is the chosen breakpoint.



**Figure 3.9:** A histogram of the alpha values for every house in the third segmentation method. The first quantile is chosen as the overall breakpoint. It is 12 degrees, illustrated by the green line.

# CHAPTER 4

# Models on the Daily Consumption

---

Now that data is cleaned and prepared, a statistical analysis consisting of data segmentation and linear regression models can be made. The purpose of the analysis is to detect which attributes affects the performance of a specific house.

## 4.1 Linear regression

Linear regression is a method to model the relationship between a dependent variable and one or more independent variables, where the unknown model parameters are estimated from the data. With the dependent variable  $Y$  and the independent variables  $x_1, \dots, x_n$ , the linear regression model is formulated as

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \dots + \beta_p x_{i,p} + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \quad i = 1, \dots, n. \quad (4.1)$$

The variables  $\varepsilon_i$  are errors which are assumed to be white noise while also being i.i.d (independent and identically distributed). Equation (4.1) shows a multiple linear regression model as it contains more than one explanatory variable. In this section both a simple linear model and a multiple linear model has been fitted to data given in table 3.1.

As the best linear model  $Y_i$  is desired, the total deviation from the data has to be as small as possible. The least squares method given as

$$\text{SSE} = \sum_{i=1}^n (Y_i - (\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \dots + \beta_p x_{i,p}))^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \quad (4.2)$$

is chosen for estimating the model. The parameters  $\beta_j$  are optimized to minimize the sum of squared errors of prediction (SSE).

### 4.1.1 Model assumptions

When SSE is minimized the model needs to be validated by checking whether the underlying model assumptions are fulfilled.

**1** Normality of residuals

**2** Variance homogeneity

**3** Variance should be independent of location

**4** Linear relationship between  $x_j$  and  $Y$

Chapter 7 in [4] explains the model assumptions listed above. To summarize, a model can be checked by looking at diagnostic plots of the residuals. The first plot considered is the Residuals vs Fitted, where the residuals are expected to be randomly scattered. To test whether the residuals are normally distributed a normal quantile–quantile plot is used. Here the residuals are expected to follow a straight line. A Scale-Location plot shows normalized and weighted residuals by sample leverage where the residuals are also expected to be randomly scattered. The last diagnostic plot shows the Residuals vs Leverage which should be a straight line and there should not be any clear patterns. If these assumptions are not met, it can influence the parameter estimates and with it the significance of the parameters. In addition, a Shapiro-Wilk test is performed to check if the residuals of the models are i.i.d with  $\mathcal{N}(0, \sigma^2)$ .

The fitting of the regression models is carried out by using the method stepwise regression which updates the model in each step. In each step it is considered whether a variable is added or subtracted from the set of explanatory variables based on specific criteria. This process is called variable selection and Chapter 7 in [8] explains how this process can be done by using either forward or backward selection. In this project a modified version of backward selection is applied. The models are used for comparing which explanatory variables influence each house. Therefore, the models are not reduced using the R function `step`. Instead the significance of the parameters are investigated and then the parameters which are significant for the majority of the houses will be used in an updated linear regression model. Thus, the variable selection is done manually which can be said to be a modified form of backward selection. The level of significance is determined by an F-test where the variables selected have a p-value below a threshold which is chosen at 0.05. [Missing reference](#)

Both a simple linear and a multiple linear regression model will be implemented in order to detect which attributes affect the performance of a specific house. This will be done by interpreting the estimates of the relation between the different explanatory attributes and **Consumption**. As mentioned, the p-value of the estimates of the explanatory variables will be the main focus when investigating which attributes influence the performance. Moreover, transformation of data is not considered since the purpose is to interpret the results and a transformation would ...

## 4.2 Simple linear regression model

A simple linear regression model is fitted to each house with `Consumption` as a function of `Temperature`. Since it is expected that the temperature is the physical phenomenon with the greatest influence on the heat consumption, it is chosen as the independent variable. Hence, the simple linear regression model applied to each house is

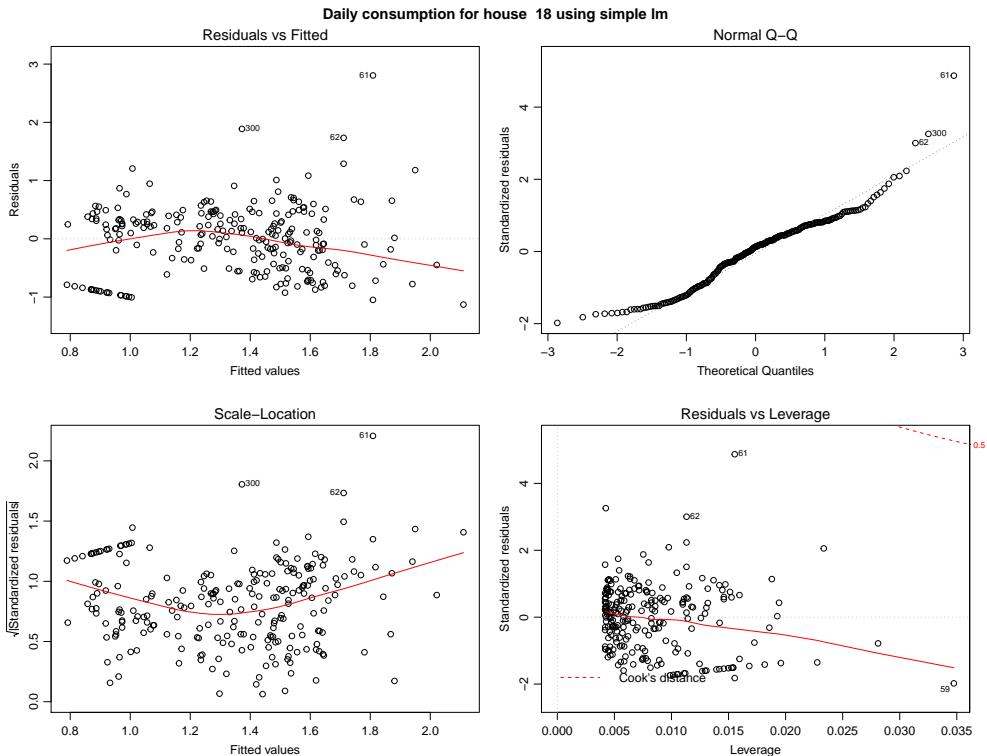
$$Y_Q = \mu + \beta_T \cdot x_T + \varepsilon \quad (4.3)$$

The models are performed by using the `lm()` function in R. The models will then be validated by examining whether the model assumptions in Chapter 4.1.1 are met and different tests on normality of residuals are performed. The simple model only includes one explanatory variable, thus a variable selection is not performed.

### 4.2.1 Validation

To validate the model, different methods are used. The abovementioned model assumptions are checked and furthermore tests of normally distributed residuals are performed. If the model assumptions are fulfilled and the residuals are i.i.d with  $\mathcal{N}(0, \sigma^2)$ , the model is said to be valid.

Figure 4.1 and Figure 4.2 shows examples of the model applied on two of the houses, where one does not fulfill the assumptions and another model that overall can be said to fulfill the assumptions. The Residuals vs. Fitted plot in Figure 4.1 clearly shows that the residuals are not randomly scattered around mean 0 indicating that the variance are not constant and an odd behaviour appears in the bottom left corner. The QQ-plot shows tails and the residuals do not follow a straight line. The Scale-Location and Residuals vs. Leverage plots also show that the residuals can not be said to be i.i.d. In contrast to the model of house 18, the behavior of the residuals in Figure 4.2 seems more normally distributed. Overall, they are randomly scattered around mean 0 and the QQ-plot shows that the majority of the residuals lie on a straight line. However, the majority of the models do not fulfill the assumptions. In addition, a Shapiro-Wilk test and a sign test is performed. The hypothesis tested in the Shapiro-Wilk test is that the residuals are i.i.d and if the p-value  $> 0.05$  the hypothesis can not be rejected i.e. the residuals are normally distributed. In the sign test the hypothesis is that the number of positive signs are equal to the number of negative signs, which is one of the requirements when checking for normality. If the model assumptions are also fulfilled, then the model is concluded to be valid. The p-values from both of the tests are found in Table A.2 and it is clearly seen that the majority of the simple linear regression models have residuals with p-value above the significance level at 0.05. Thus, the hypothesis is rejected and it can be concluded that the residuals are not i.i.d.



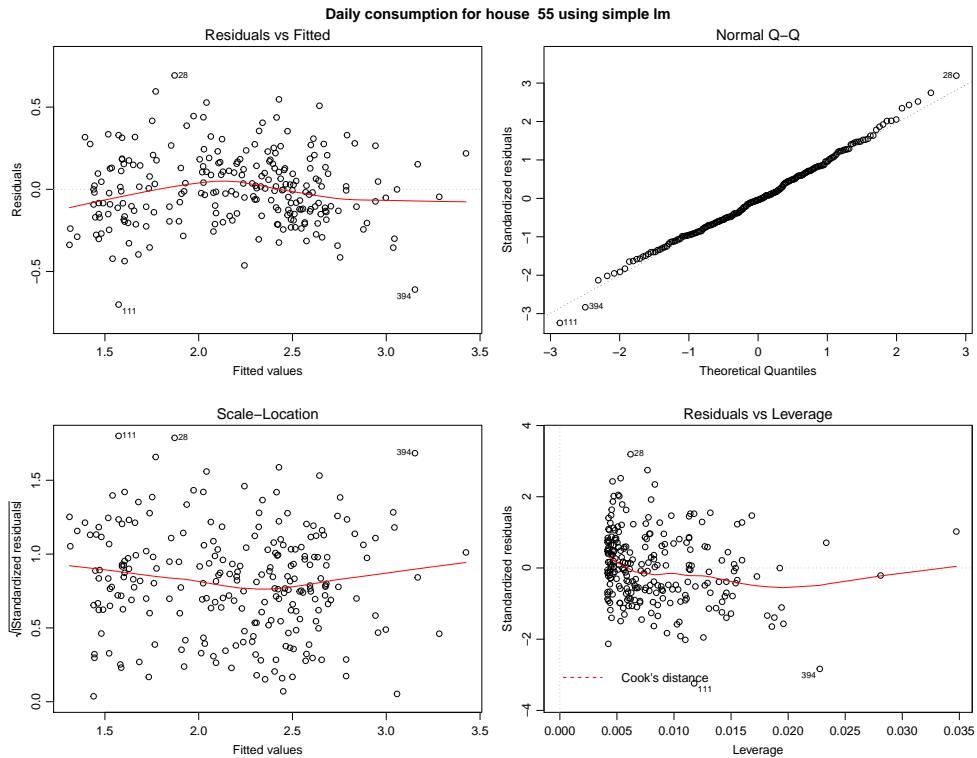
**Figure 4.1:** Residual plots of house 18 based on the simple linear regression model given in (4.3). The model assumptions of a linear regression model are not fulfilled for this specific house.

## 4.2.2 Results

Table A.1 shows the estimates from all the simple linear regression models and they are also visualized in ?? including a 95% confidence interval.

Overordnet kan den simple lineære regressionsmodel ikke beskrive tenden. Den antager, at temperaturen er den eneste faktor der påvirker husenes varmeforbrug. Men ved at undersøge hvorvidt model assumptions er opfyldt, så 'fails' modellen i de fleste tilfælde. Dette tyder på, at der findes flere faktorer, der påvirker varmeforbruget, hvilket selvfølgelig er forventet. Der vendes tilbage til dette i sammenligningen af de to regressionsmodeller. Når vi fitter en lineær model med en attribut, så vil man lægge linjen så summen af alle residualerne er lig 0, derfor består sign testen, men shapiro og model assumptions er ikke opfyldt, så det er ikke nok.

The conclusion that the temperature is highly significant implies that this at-



**Figure 4.2:** Residual plots of house 55 based on the simple linear regression model given in (4.3). The model assumptions of a linear regression model are overall fulfilled.

tribute will be included in the multiple linear regression model.

### 4.3 Multiple linear regression model

The linear regression model is extended to a multiple linear regression model as the inclusion of several independent variables is expected to improve the model. The simple model clearly showed that the heat consumption is affected by other physical factors than temperature. Hence, a full multiple linear regression model containing the attributes given in Table 3.1 is performed on the model data. The backward selection of the variables are perfomed such that the highly significant parameters from the full multiple regression model are chosen and used to develop an updated model. Since **Condition** and **PrecipitationProbability** are not normalised, they

are excluded from the model. In addition, it is mentioned in Chapter 2 that the house data consists of house with observations for approximately a year and house with observations for approximately six months. Thus, the two distinct lengths of observations are modeled slightly different. There do not exist observations for winter break and spring break in the data containing the short houses. The parameters will be denoted as follows: Intercept (I), Temperature (T), North (N), East (E), South (S), West (W), Mean Sea Level (MSL), Solar Radiation (SR), Winter Break (WB), Spring Break (SB), Autumn Break (AB), Christmas Break (CB), Weekend (WKND), the interaction between the temperature and the different wind directions (T:N, T:E, T:S, T:W). This lead to the following two multiple linear regression models:

$$\begin{aligned} Y_{Q,L} = & \mu + \beta_T \cdot x_T + \beta_N \cdot x_N + \beta_E \cdot x_E + \beta_S \cdot x_S + \beta_W \cdot x_W + \beta_{MSL} \cdot x_{MSL} \\ & + \beta_{SR} \cdot x_{SR} + \beta_{WB} \cdot x_{WB} + \beta_{SB} \cdot x_{SB} + \beta_{AB} \cdot x_{AB} + \beta_{CB} \cdot x_{CB} \\ & + \beta_{WKND} \cdot x_{WKND} + (\beta_{T:N} + \beta_{T:E} + \beta_{T:S} + \beta_{T:W}) \cdot x_T + \varepsilon \end{aligned} \quad (4.4)$$

$$\begin{aligned} Y_{Q,S} = & \mu + \beta_T \cdot x_T + \beta_N \cdot x_N + \beta_E \cdot x_E + \beta_S \cdot x_S + \beta_W \cdot x_W + \beta_{MSL} \cdot x_{MSL} \\ & + \beta_{SR} \cdot x_{SR} + \beta_{AB} \cdot x_{AB} + \beta_{CB} \cdot x_{CB} + \beta_{WKND} \cdot x_{WKND} \\ & + (\beta_{T:N} + \beta_{T:E} + \beta_{T:S} + \beta_{T:W}) \cdot x_T + \varepsilon \end{aligned} \quad (4.5)$$

The models include the interaction between the temperature and the wind since it is expected that this interaction has an influence on the heat consumption. That is, it is expected that the influence of the wind is greater when the temperature is lower. In linear regression, it is desired to have as simple a model as possible, and since the models would be too complex when including all the interactions, only the interaction between the temperature and the wind is included.

The models show that the interactions between the attribute *Holiday* and the other attributes are chosen to be excluded. The reason is that *Holiday* is used to investigate how the consumption changes during holiday periods. The parameters will be denoted as follows: Intercept (I), Temperature (T), North (N), East (E), South (S), West (W), Mean Sea Level (MSL), Solar Radiation (SR), Winter Break (WB), Spring Break (SB), Autumn Break (AB), Christmas Break (CB), Weekend (WKND) and the interaction between the temperature and the different wind directions (T:N, T:E, T:S, T:W). The wind directions are not inserted directly into the model. As mentioned in Table 2.2, the wind directions are given as degrees from 0 to 360. Instead of just classifying degree intervals as north, south, east or west, splines are used to model the directions. How this is done is described below. As for the simple model, the multiple models must also be validated by investigating diagnostic plots and testing for normally distributed residuals. Before performing the models, the wind will be modified which the following section explains.

The models show that the interactions between the attribute *Holiday* and the other attributes are chosen to be excluded. The reason is that *Holiday* is used to investigate how the consumption changes during holiday periods. The parameters will be denoted

as follows: Intercept (I), Temperature (T), North (N), East (E), South (S), West (W), Mean Sea Level (MSL), Solar Radiation (SR), Winter Break (WB), Spring Break (SB), Autumn Break (AB), Christmas Break (CB), Weekend (WKND) and the interaction between the temperature and the different wind directions (T:N, T:E, T:S, T:W). The wind directions are not inserted directly into the model. As mentioned in Table 2.2, the wind directions are given as degrees from 0 to 360. But in the model the wind direction is not quantified in this way, rather it should be an effect on each of the four major directions: north, south, east and west. These will in the following be called wind direction categories. One way to classify the wind direction could be to use indicator variables, and assign each possible wind direction to a category. In this project spline functions will be used to determine the effect a certain wind direction has on the categories. As a consequence, a wind direction can have different effects on different categories. For example wind coming directly from north might have a large effect on the north category, but also a lesser effect on east and west. In real life the wind direction is often distorted close to the ground because of turbulence. Also wind directions on the border between two categories should have an effect on both. For these reasons, splines seem like a good choice for modelling the wind.

### 4.3.1 Splines

Each major direction is described by a basis spline. Together they form a spline basis, that spans the space of the wind direction. The theory on splines introduced here is based on [7]. A spline basis is defined by a knot vector  $\Xi$  and a polynomial degree  $q \in \mathbb{N}$ . The  $i$ 'th basis spline is a function  $\mathcal{N}_{\Xi,i}^q : \mathbb{R} \rightarrow \mathbb{R}$ . If the spline basis should contain  $n$  basic splines, then  $\Xi = \{\xi_1, \xi_2, \dots, \xi_{n+q+1}\}$ . When the knots are equidistant, the spline is uniform. Then the spline is defined by the Cox-de Boor recursion formula:

$$\mathcal{N}_{\Xi,i}^0(\xi) = \begin{cases} 1 & \text{if } \xi_i \leq \xi \leq \xi_{i+1} \\ 0 & \text{otherwise,} \end{cases} \quad (4.6)$$

for the splines of degree 0, and for higher degrees

$$\mathcal{N}_{\Xi,i}^j(\xi) = \frac{\xi - \xi_i}{\xi_{i+j} - \xi_i} \mathcal{N}_{\Xi,i}^{j-1}(\xi) + \frac{\xi_{i+j+1} - \xi}{\xi_{i+j+1} - \xi_{i+1}} \mathcal{N}_{\Xi,i+1}^{j-1}(\xi), \quad (4.7)$$

for  $j = 1, 2, \dots, q$  and  $i = 1, 2, \dots, n + q - j$ . When the splines are uniform, the continuity of the basis splines across a knot is  $q - 1$ . This means that the  $q - 1$ 'th derivative exists and is continuous. In this project splines of degree 2 is used, meaning that the first derivative is continuous at every point on the spline. The splines used here are periodic. That means that after the first  $n$  knots, the knot sequence starts over. For modelling the wind direction, four knots are used for the knot vector, each associated with a wind direction category. Here they are defined as *northeast*, *southeast*, *southwest* and *northwest*. The reason for this will be explained later in

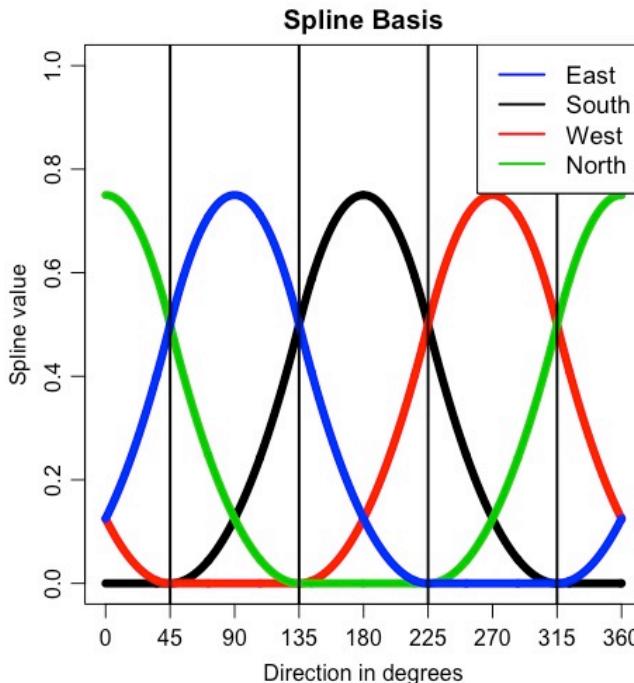
this section. For each of these knots there is a basis spline. When the degree of the splines is 2, it means that the spline for a given knot is zero at the opposite knot. For example, when using four knots the spline peaking in the north would always be zero in the south. Higher degrees would not maintain this property. The spline basis can be seen on Figure 4.3. Notice that the sum of the entire spline basis at a given point always adds up to one. The figure shows how the  $i$ 'th spline, associated with the  $i$ 'th knot does not peak at that knot, but in the following interval. This is the reason why the knots are chosen in this way. By choosing the knots to be between the main directions north, east, south and west, this is where the basic splines peek.

Now the splines can be used to represent their category in the linear regression model. At a given data point, the wind direction is given as input to each basic spline. The result is then weighted by the windspeed. The result is used as the effect of the effect of the category that spline represents. As an example, a windspeed of 2 from the angle 135 degrees would give the following result for the variables in the regression model:  $x_N = 0$ ,  $x_E = 0.5$ ,  $x_S = 0.5$  and  $x_W = 0$ . These results can be derived from Figure 4.3.

### 4.3.2 Results

The full multiple regression models are steps of the backward selection which is why they are not validated. The purpose of this step is, as mentioned, to determine which parameters are found to be significant in the majority of the models. Hereafter, these parameters are included in a general regression model that can be used for comparing the houses performance.

When performing the two models given in (4.4) and (4.5), without reduction, the significance of the parameters are determined and can be found in Table A.3-A.4. In addition, Table 4.1 and Table 4.2 are generated in order to determine which parameters are significant for the majority of the houses. The tables clearly show that the total of significance of the intercept, temperature, east and west as well as the interaction between temperature and west and the solar radiation occurs in more than half of the models. Thus, these are included in the general regression model. But when the influence of the wind on the consumption is examined, it does not make much sense to exclude some of these parameters. Therefore, both north, south, east and west are included as well as their interaction with the temperature. Each parameter from the different holiday attributes is not significant for enough of the models. There might have been some patterns in when the households vacationed, which the tables also indicate, but the impact on consumption is not large enough for the holidays to be included in the general model.



**Figure 4.3:** The spline basis used to model the wind direction. Each color is a different basis spline, and the vertical lines mark the knots.

	I	T	N	E	S	W	MSL	SR	WB	SB	AB	CB	WKND	T:N	T:E	T:S	T:W
Sum of ***	5	41	0	18	2	24	6	22	3	3	1	5	4	0	1	7	9
Sum of **	6	1	1	9	5	12	4	10	6	2	0	2	0	1	1	9	9
Sum of *	6	1	5	7	7	2	2	2	3	6	5	3	8	2	3	5	11
Total of 43	17	43	6	34	14	38	12	34	12	11	6	10	12	3	5	21	29

**Table 4.1:** The distribution of significant parameters from the multiple linear regression model for long houses. There are 43 long houses, thus the total of the significance of each parameter for each house is in relation to the number of long houses.

	I	T	N	E	S	W	MSL	SR	AB	CB	WKND	T:N	T:E	T:S	T:W	
Sum of ***	0	27	0	4	0	15	0	5	2	0	0	0	0	0	0	3
Sum of **	2	0	0	6	2	5	2	6	0	0	3	0	0	2	5	
Sum of *	2	0	1	8	4	4	4	4	2	5	2	1	1	3	9	
Total of 27	4	27	1	18	6	24	6	15	4	5	5	1	1	6	17	

**Table 4.2:** The distribution of significant parameters from the multiple linear regression model for short houses. As for the long houses, the total of the significance is in relation to the number of long houses.

## 4.4 Regression model for comparing houses

Based on the tables illustrating the significant parameters for the long and short houses, an updated multiple linear regression model is made. The purpose of this more general model is to compare which parameters influence each house. Furthermore, houses with e.g. same area, construction year etc. can be compared. The comparison model derived from Table 4.1 and Table 4.2, becomes

$$Y_Q = \mu + \beta_T \cdot x_T + \beta_N \cdot x_N + \beta_E \cdot x_E + \beta_S \cdot x_S + \beta_W \cdot x_W \\ + \beta_{SR} \cdot x_{SR} + (\beta_{T:N} + \beta_{T:E} + \beta_{T:S} + \beta_{T:W}) \cdot x_T + \varepsilon. \quad (4.8)$$

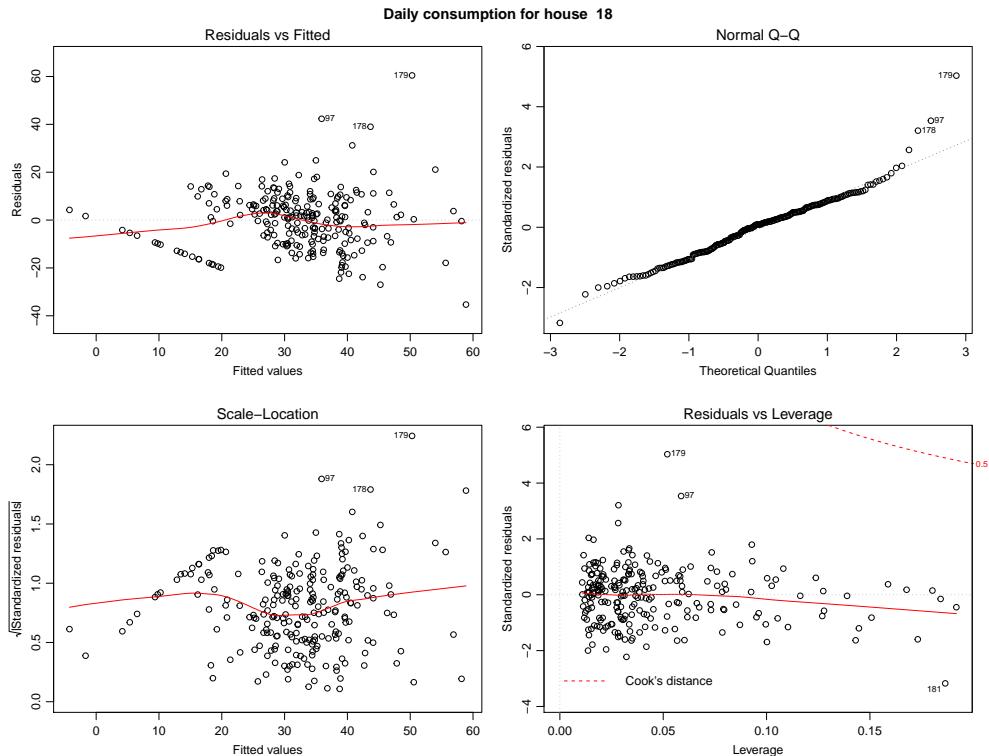
Likewise, the models are validated and if the assumptions are fulfilled the models can be used for comparison and predictions. The model in (4.8) is expected to explain the heat consumption of the houses in a more accurate way as it takes several significant parameters into account. As mentioned, the purpose of the general model is to compare the houses for what influences their specific heat consumption and thereby the house's performance. Thus, they are not further reduced.

### 4.4.1 Validation

In order to determine whether the general model is valid, the diagnostic plots in Figure 4.4 and Figure 4.5 are investigated. The residuals from the model for house 18 behave quite odd and are not randomly scattered. It seems like this specific house can not be described by a linear regression model, which will be discussed further in Chapter [diskussionen](#). On the other hand, the diagnostic plots of the residuals from house 55 are definitely improved and the residuals show a randomly behavior. Overall, the diagnostic plots can be said to be fulfilled. They are not perfect but that is not expected either. There are some houses which residuals have strange patterns, but the majority of the models can be said to be valid. Furthermore, the Shapiro-Wilk tests and the sign tests in Table A.5 are used to determine if the residuals are normally distributed. The Shapiro-Wilk test shows that the hypothesis of normally distributed residuals for most of the models cannot be accepted. Similarly, the results of the sign tests do not indicate that the residuals can be concluded to be normally distributed. However, this can be said to be contrary to the diagnostic plots, which is quite remarkable. Therefore, this will be discussed in more detail in the comparison of the two regression models. In spite of the abovementioned, the models are concluded to somewhat fulfill the model assumptions.

### 4.4.2 Results

The results of the general regression model performed on each house, are illustrated in Table 4.3 and Table 4.4 with the significance of each parameter. As expected, the intercept and the temperature are highly significant in all models and with the "correct" sign. That is to say, the expected sign of the temperature is negative since



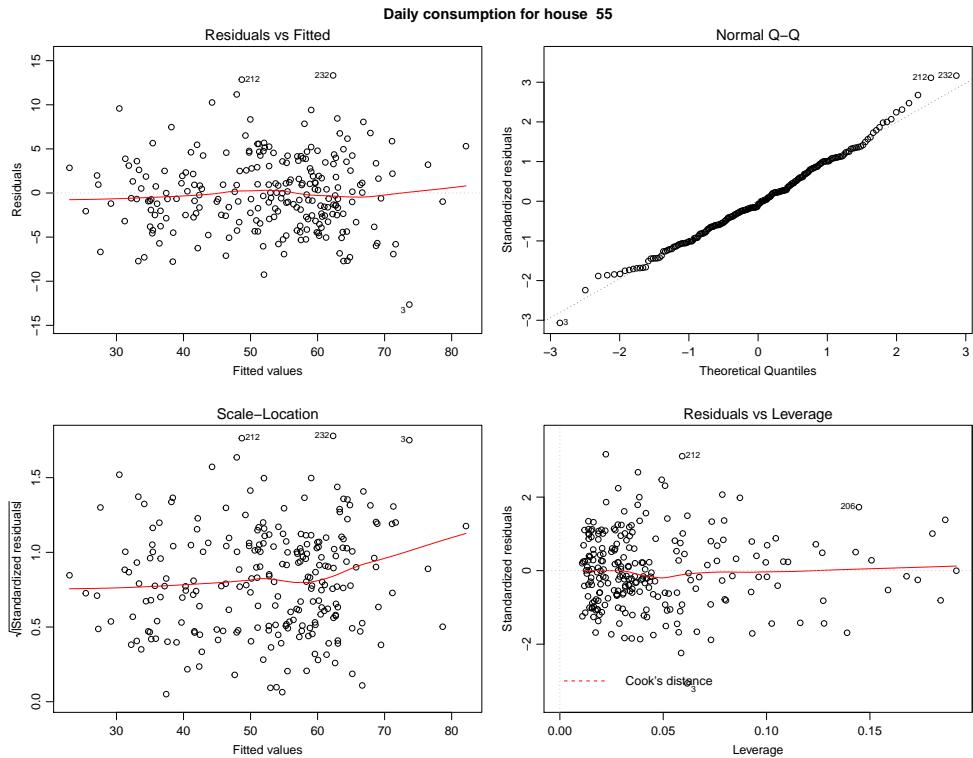
**Figure 4.4:** Diagnostic plot of the general multiple linear regression model for long house 18.

the increase in the temperature will result in a decrease in the heat consumption .... The attribute **SolarRadiation** effects the majority of the houses. The solar radiation warms the house up through the windows, so it makes sense that this has significance for how large the heat consumption is.

The wind directed from East and West also have great influence on the consumption of most houses. Not done Kommentér på tabellerne i forhold til sammenligningen af huse

Index	I	T	N	E	S	W	SR	T:N	T:E	T:S	T:W
1	+***	-***				+**			+. -**		
2	+***	-***		+**		+***	-***			-***	
3	+***	-***		+***		+***	-***		+*	-*	
4	+***	-***	+		+***		-***				
5	+***	-***		+***		+***	-***	+	+***	-**	
7	+***	-***	-*	+***	-.	+**	+***	-**	+	-**	
11	+***	-***		+***		+***	-***			+	
12	+***	-***		+		+*					
14	+***	-***		+***		+***	-**			+**	
18	+***	-**		+***	-**	+**	+		+***	-***	
21	+***	-***		+		+***				-***	
22	+***	-***			+***	+**	-***			-*	
23	+***	-***	-.	+		+***	-***		+***	-***	
28	+***	-***		+	+		-**				
29	+***	-***		+***		+***	-**	+	+**	-***	
30	+***	-***		+	+	+***	-***			-*	
31	+***	-***		+***		+***	-***	+	+***	-***	
32	+***	-**			+	+**		+		-.	
33	+***	-***		+***	+	+***	-***			-**	
34	+***	-***		+***	+	+***	-**		+	-**	
36	+***	-***		+***		+***	-***		+**		
37	+***	-***		+***	+	+**	-***				
38	+***	-***		+		+***	-***			-***	
40	+***	-***		+	+	+**	-***	+	+	-*	
41	+***	-***		+***	+	+***		+	+	-.	
42	+***	-***		+***		+***	-***			-*	
44	+***	-***	-*				-**			+*	
45	+***	-***			+	+***	-*	+	+	-***	
46	+***	-***		+	+	+***	-**			-**	
47	+***	-***		+		+***	-**	-**		-***	
48	+***	-***		+	*	+***	-**			-**	
49	+***	-***					-**				
50	+***	-***	-.	+***	-*	+**	-***	+	+***	-.	
52	+***	-***	-.		-**	-.	-***			+***	
54	+***	-***		++		+**	-***		+*	-*	
55	+***	-***		+	+	+***	-***		+**	-*	
56	+***	-***		++		+***	-***		+**	-*	
57	+***	-***		+***		+***				-**	
58	+***	-***	-.	+		+***	-***	+	+***	-**	
61	+***	-***		+	+	+		-.		-.	
64	+***	-***	-*	+***		+***	-***	-.	+**	-**	
65	+***	-***		+		+***	-***		+	-**	
66	+***	-***	+	+***	++	+***	-***		+**	-**	

**Table 4.3:** Significance of parameters for 'long' houses.



**Figure 4.5:** Diagnostic plot of the general multiple linear regression model for long house 55.

Index	I	T	N	E	S	W	SR	T:N	T:E	T:S	T:W
6	+***	-***		+. .		+**	-*				
8	+***	-***		+*		+***	-*			+*	-.
9	+***	-***	-.			+*	-***	+		+*	-**
10	+***	-***		+***		+**	-**				-.
13	+***	-***		+. .		+**	-.	+		+**	-**
15	+***	-***		+**		+**					
16	+***	-***		+. .		+***	-*				-**
17	+***	-***		+*	+***	+***	-**				-***
19	+***	-***		+*	-.			+		+**	
20	+***	-***			+*						
24	+***	-***		+*		+***					-***
25	+***	-***		+**		+***				+*	-*
26	+***	-***				+***	-*	+		+*	-*
27	+***	-***		+. .	+**	+**					-*
35	+***	-***		+**	+. .	+***	-**				-**
39	+***	-***				+*	-*			+	.
43	+***	-***		+**	+. .	+***	-**				-**
51	+***	-***			+. .	+***	-**			+*	-**
53	+***	-***			+*	+*					-.
59	+***	-***	-*	+**	+*	+***	-*	+		+*	-.
60	+***	-***	-*	+**		+***	-***	+**		+**	-***

### 4.4.3 Predictions

In addition, the models are used to predict 31 days ahead. It must be known that the model data has been divided into training and test sets in order to make predictions. The test set consists of the last 31 days and the remaining observations constitute the training set. The predictions are made using the function `predict` in R. In consistency with the other results, it is the two houses, 18 and 55, whose predictions are illustrated in Figure 4.6 and Figure 4.7. It can be seen that the predictions for house 55 lie close to the actual data and captures the overall behaviour. The 95% prediction interval are narrow, which indicates that the standard deviations of this model are fairly small. The general regression model performs as expected. Contrary, the model for house 18 has problems with capturing the oscillating behaviour and the 95% prediction interval are quite wide, i.e. the corresponding standard deviations are large. But since the actual data lie within the 95% prediction interval the model can be used for describing the heat consumption in house 18.

## 4.5 Comparison

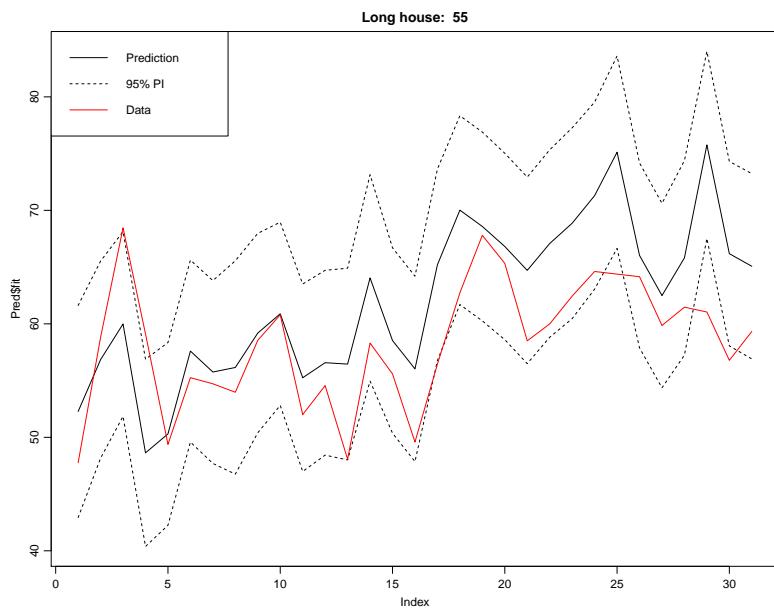
When looking at the results of the different tests in Table A.2-A.5, the simple model passes the tests in more cases than the multiple model. This might be explained by the inclusion of the wind in the multiple regression model. The so-called directions are not all significant for all houses. Thus, this can affect the behavior of the residuals.

On the other hand, the interpretation of the simple model shows that the temperature has a great influence on the consumption. But from a physical perspective it is known that the heat consumption is influenced by other physical phenomena as well. Hence, the simple regression model is indeed too simple.

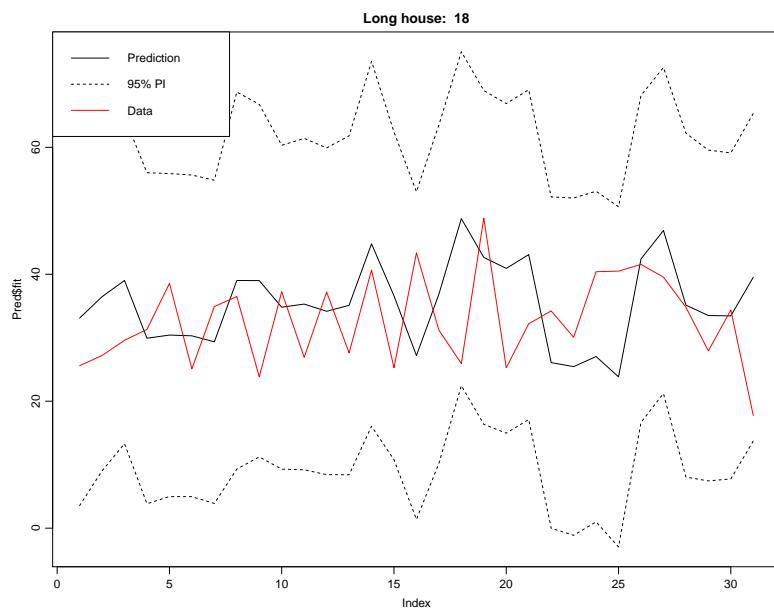
## 4.6 Visualization of the results

Part of this project is also to give suggestions on how the results of the analyses can be illustrated to the users of the WATTS app.

The wind dependency showed in [ref](#) and [ref](#) can be illustrated in a more user-friendly way by coloring according to the wind's influence on the house. The coloring is scaled according to the estimate of the wind, and that is relative to the user's own house's performance. [ref](#) shows that the heat consumption for house 55 is most affected when the wind comes from west. **Kan godt være boys skal kommentere lidt på det her.**



**Figure 4.6:** Predictions of the last 31 days of the model data for house 55 including a 95% prediction interval. The model captures the behavior of this house quite well.



**Figure 4.7:** Predictions of the last 31 days of the model data for house 18 including a 95% prediction interval. The model does not capture the behavior of this house at all.

# CHAPTER 5

# Models on the Hourly Consumption

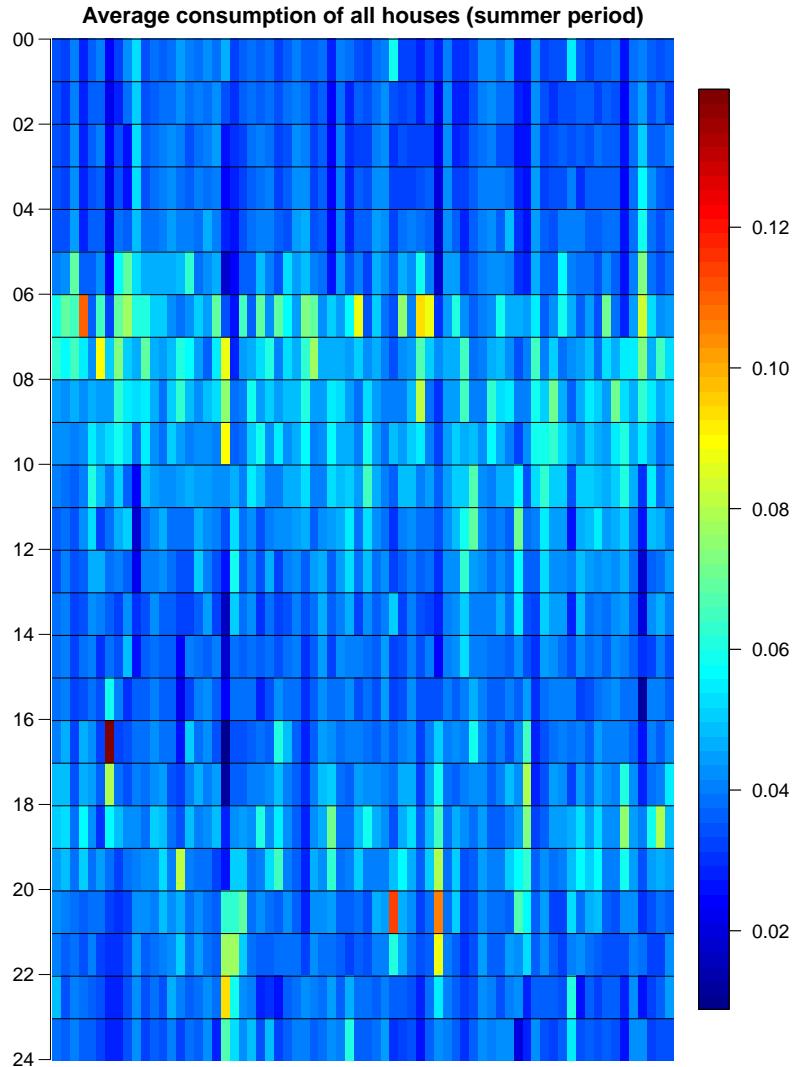
---

In this section, a more detailed look at the hourly consumption will be provided. One of the goals is to get a better understanding of the tap-water consumption during the summer period, such that it can be used when looking at the winter period. This can be done by looking at the distributions of the consumption during the day in the two periods. Another goal of this section is to model the hourly consumption as a time series. An ARIMAX model will be applied to give a better understanding of the data. The ARIMAX model can also be used to give short term predictions of the consumption. These predictions will be compared to the ones provided by the multiple regression model.

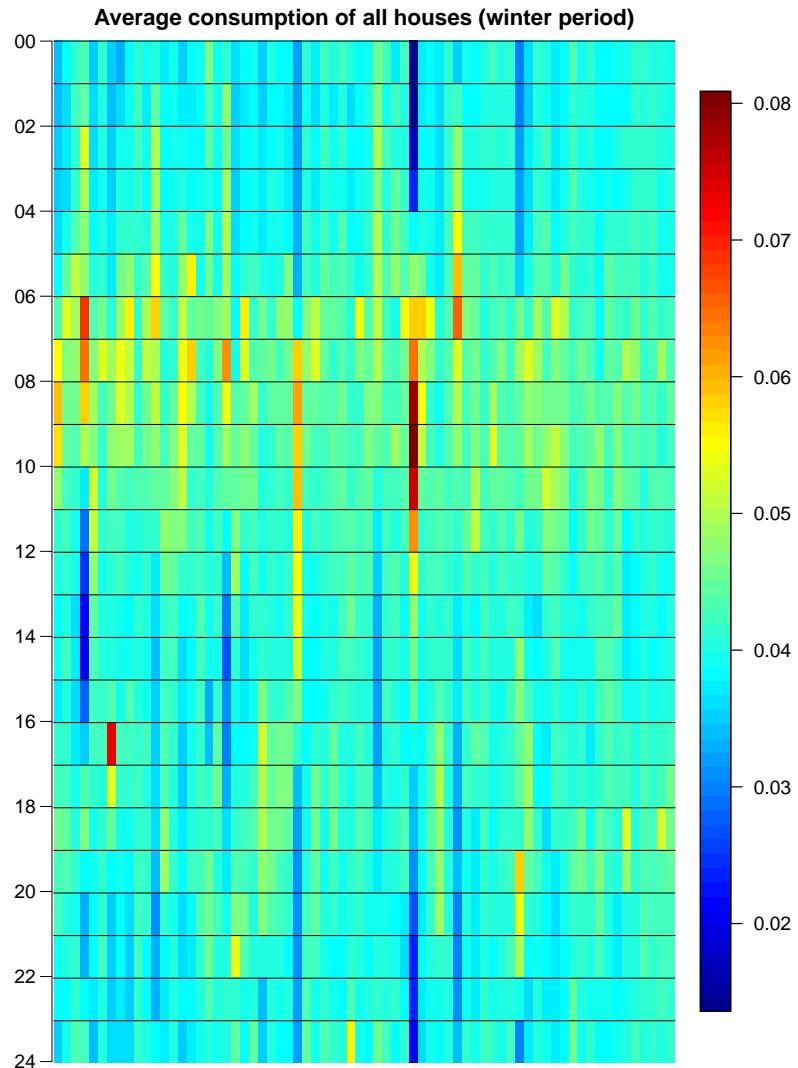
## 5.1 Description of the Hourly Consumption

Figure 5.1 and Figure 5.2 show the average consumption of each house during the day for the summer period and the winter period respectively. The different hours can be seen on the y-axis, and the colour coding show the fraction of that house's consumption in that hour interval. Each vertical strip of colours is a single house. Each strip sum up to 1. Looking at the summer period in Figure 5.1, a general trend is apparent: the consumption is usually larger around 7 AM and to some degree around 7 PM. Almost every house peaks in one of these periods, and some peaks go up to 12% of the daily consumption. On the other hand, there is almost no consumption between 11 PM and 5 AM. The same goes for the afternoon between 1 PM and 4 PM. Figure 5.4 shows the average distribution of all houses together with lines indicating the quantiles. On the figure it can be seen that the intervals 06 – 11 and 18 – 19 are in the top quantile. 00 – 05 and 15 – 16 is where the consumption is lowest.

These trends make sense. In the summer period, not much energy is used for heating the house. There is usually a significant amount of tap water consumption in the morning, when people take warm baths and make breakfast. Sometimes a dishwasher might be running as well. Then there is not much consumption while people are at work or school. When they get home in the late afternoon the consumption



**Figure 5.1:** The normalized average consumption of every house during the day in the summer period. This is characterized by the days where the average outside temperature is above 15 degrees. The horizontal lines indicate the hours and each vertical strip is a house. The scale indicates the fraction of the total consumption during the day.



**Figure 5.2:** This figure shows the same as Figure 3.1, but only in the winter period, characterized by an outside temperature below 12 degrees.

rises again as they prepare for dinner or use hot water in other ways. During the night time the consumption becomes low again.

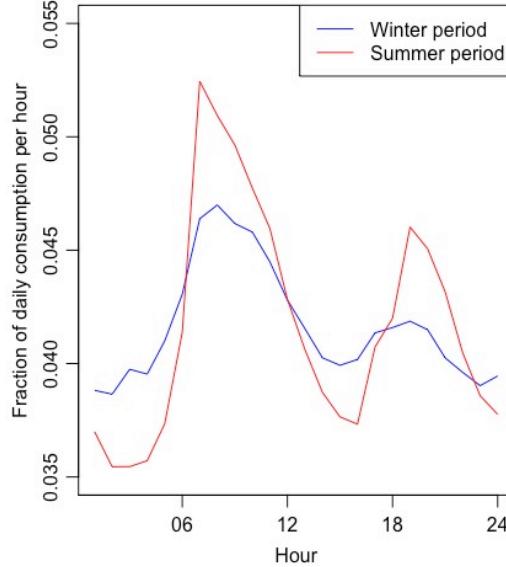
The winter period on Figure 5.2 is a bit different. There are still significant peaks in the morning, and to some extent in the evening as well, but in general the consumption is more spread out on the entire day. This is mostly because of the heating consumption in the winter period. While people are not at home or while they are sleeping, the heating is still turned on. The highest peaks only go to 8% of the daily consumption here. One house stands out in this plot. A bit to the left of the middle there is a house where the consumption is several times higher between 8 AM and 12 noon. This house has almost no consumption during the night. But the house is not a commercial building and its area is only  $138 m^2$ . So this house appears to have an efficient night time drop for their thermostat.

These figures illustrate the general trend of the houses, but it is hard to compare them in a meaningful way. But Figure 5.3 shows the average distribution of all houses during the day. Both the winter season and the summer season show the same trends that was discussed above. But this plot also shows how the winter period is more smoothed out than the summer period. Keep in mind that the lines only show the relative distribution, and they do not take into account that the consumption in the winter period is significantly higher. As one can see on the y-axis, the difference between the two curves is very small. A night time period can be defined as the hours 23 – 05. This is the period after the consumption drops in the evening, and before it rises in the morning. In this period, the houses on average use 21,9% of their daily consumption in the summer period, and 23,7% in the winter period. A completely uniform consumption would be 25%. It is not surprising that the consumption in the night hours is lower than the average. Neither is it surprising that the consumption at night in the summer period is relatively smaller than in the winter period. The extra cost of heating the house makes the consumption more spread out on the 24 hours of the day. But it is surprising that the difference between the summer period and the winter period is only 1.8 percentage points. With this in mind, the time series modelling will now be introduced.

In addition, Figure 5.4 is used to detect the significant intervals during a day in relation to the consumption. It is clearly seen that the time interval with highest consumption is 06-11 am and around 6 pm. **Not done**

## 5.2 The ARMA Models and Their Extensions

The consumption of a house during a certain period with hour intervals is a time series. A time series is a realization of a stochastic process. In this section the ARMA model will be introduced, and an extended ARMA model, the ARIMAX model, will be fitted to the consumption. The theory of the ARMA model is based on chapter 5 from the book "Time Series Analysis" by Henrik Madsen [6]. The ARMA model fits the data to a linear stochastic process, with an autoregression part (AR) and a moving average



**Figure 5.3:** The average distribution of the heat consumption during the day for the winter period and the summer period respectively. The winter period is more smoothed out, but they are very similar.

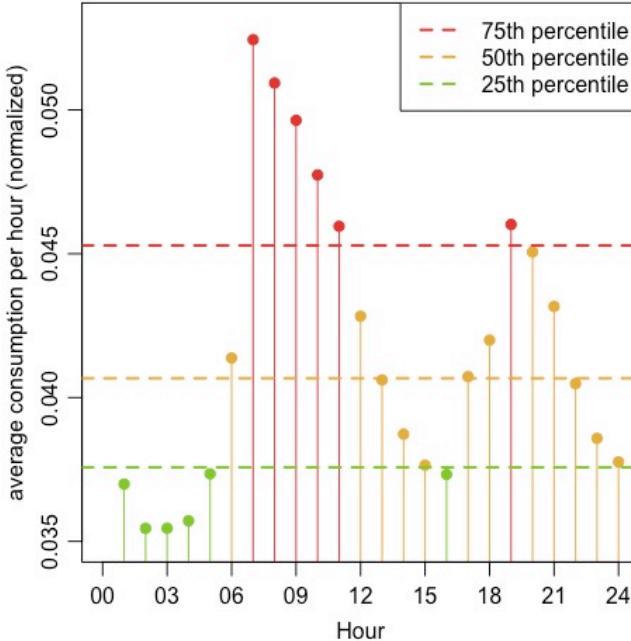
part (MA). A linear process  $\{Y_t\}$  is a process that can be written as

$$Y_t - \mu = \sum_{i=0}^{\infty} \psi_i \epsilon_{t-i}, \quad (5.1)$$

where  $\mu$  is the mean of the process,  $\{\epsilon_i\}$  is white noise and  $\{\psi_i\}$  is the weights. For now, the mean  $\mu$  is assumed to be zero. To define the ARMA model, the backwards shift operator  $B$  is first introduced as  $B(Y_t) = Y_{t-1}$ . An ARMA process has the form

$$\phi(B)Y_t = \theta(B)\epsilon_t, \quad (5.2)$$

where  $\phi$  and  $\theta$  are polynomials on the shift operator  $B$  with degree  $p$  and  $q$  respectively.  $\theta(B)$  is the autoregressive part and  $\phi(B)$  is the moving average part. The process is denoted as an  $ARMA(p, q)$  process. ARMA processes are linear. If one applies  $\psi(B)$  to  $Y_t$  and substitutes  $Y_{t-1}$ , then  $Y_{t-2}$  and so forth, the form in (5.1) is obtained. An ARMA process is stationary if all the roots of  $\phi(z^{-1})$  are within the unit circle. Stationarity is a very desirable property. In a stationary process, the mean and variance does not change over time. But often, processes will not be stationary due to long term trends. For example the mean consumption of a house has a periodic trend during the year. This was clearly illustrated on Figure 3.1. But long term trends



**Figure 5.4:** The average consumption of all houses during the day. Each point indicate the average consumption in the previous hour interval. The hours in the 75th percentile is 06 – 11 and 18 – 19.

can be eliminated by introducing differencing. Instead of modelling the process  $\{Y_t\}$ , one can model the process  $\{Y_t - Y_{t-1}\}$ , i.e. the difference between observations. This is formalized with the difference operator  $\Delta = (1 - B)$ . The differenced ARMA model is called the *ARIMA*( $p, d, q$ ) model, or the autoregressive integrated moving average model. It has the form

$$\phi(B)\Delta^d Y_t = \theta(B)\epsilon_t, \quad (5.3)$$

where  $d \in \mathbb{N}$  is the differencing factor. Apart from the long term trends, the model might also have short term periodic trends. In this case a 24 hour periodic trend would be expected. The ARIMA model can be expanded to a seasonal ARIMA with season  $s$ , such that

$$\phi(B)\Phi(B^s)\Delta^d \Delta_s^D Y_t = \theta(B)\Theta(B^s)\epsilon_t. \quad (5.4)$$

This is a seasonal *ARIMA*( $p, d, q$ )  $\times$  ( $P, D, Q$ ) $_s$ .  $\phi, B, \Delta$  and  $\theta$  are defined as in (5.3). But here  $\Phi$  and  $\Theta$  are also included. They are polynomials in  $B^s$  of degree  $P$  and  $Q$  respectively.  $D$  is the differencing of the seasonal component of the model.

In this particular project we have access to the consumption, but also to the weather data. The exploratory analysis showed that there was a significant correlation between consumption and temperature. The temperature is also a time series, and it can be used as input to the ARIMA model to make a better fit. This is called using an exogenous variable. When an exogenous variable is used, the model is called an ARIMAX model. The following part is based on chapter 8 in [6] and the *R* documentation. The model looks like this

$$\phi(B)\Phi(B^s)\Delta^d\Delta_s^D Y_t = \theta(B)\Theta(B^s)\epsilon_t + \omega(B)X_t, \quad (5.5)$$

where  $X_t$  is the exogenous variable and  $\omega(B)$  is a polynomial in  $B$ . The exogenous part can also be differenced or have seasonal components, but in this project those extensions will not be explored. In fact only  $\omega(B) = \omega_0$  will be used in the modelling process. The software used for the arima processes is of course *R*, but in particular the *arima* function. This function does not estimate the exogenous parameters according to (5.5). It actually starts out by making a regression of the series  $\{Y_t\}$  on the exogenous variable  $\{X_t\}$ . Then this fit is substituted for  $Y_t$  in (5.4). This approach is less precise, since it executes the parameter estimation in two steps, first for the exogenous variable, then for the rest of the variables. An alternative method is to use the *MARIMA* package in *R*. This package computes the estimates in (5.5) by using different approximation methods than the *arima* function. Neither methods should be considered "correct", but they produce different results.

## 5.3 Applying the models

In this section, different ARIMAX models will be applied to the data. First the *arima* function in *R* will be used to test different models. The models will include a season of 24 hours. Then the chosen model will be compared with the *MARIMA* function. In the end a model will be developed that relies only on the physical dependencies of the consumption, and does not have any seasonal component.

### 5.3.1 Seasonal model

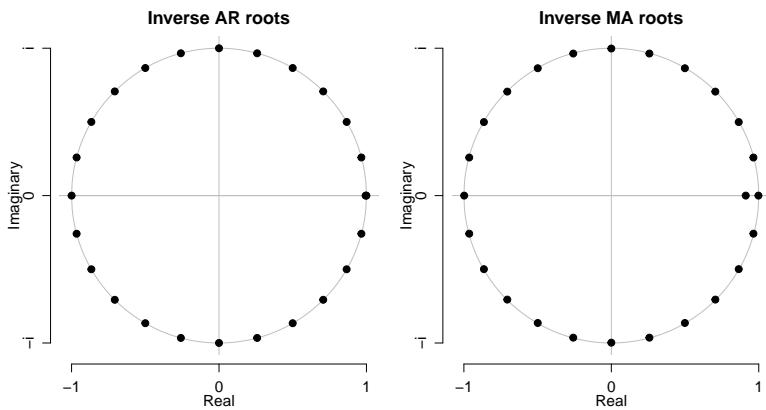
To generate a model, one must decide a model order. It is important that the model is not too complex. With the amount of data available for each house, too many parameters in the model causes the running time to increase drastically. More importantly, sometimes the optimization method used in the *arima* function does not converge. This happens more often when there are more parameters, and the model should be applicable to as many different houses as possible. As mentioned in the section above, stationarity is an important property. If the order of the autoregressive part (both non-seasonal and seasonal) is too high, the model will not be stationary unless differencing is used. Differencing on the other hand makes the model more obscure and harder to interpret. If the model is only differenced once, it means that

it models the difference between one hour and the one before that. If the seasonal part is differenced, it models the difference between an hour and the same hour the day before.

These considerations have led to the conclusion that a  $(1, 0, 1) \times (1, 0, 1)$  model is a good starting point. But for most houses this model is not stationary and the optimization does not converge. Figure 5.5 shows the roots of the model when applied to house 55. Here it can be seen how every inverse AR root is on the edge of the unit circle. This model introduces 25 roots in  $\phi$ , 1 from the non-seasonal part, and 24 from the seasonal part. As many of the inverses of these roots as possible should be inside the unit circle. For this reason, it makes the most sense to difference the seasonal part of the model. The one root from the non-seasonal part might still be on the edge of the unit circle, but the rest will most likely not be.

### 5.3.2 Physical model

Fysisk model. Ingen season. En del AR og lidt MA. Fx  $(3,0,1)$ . Så med reg.var temperatur + sin, cos, sin med ny periode, cos med ny periode



**Figure 5.5:** The roots of the first model when applied to house 55. All the inverse AR roots are on the edge of the unit circle.



# APPENDIX A

## Tables

---

- A.1 Estimates and test of the simple linear regression model
- A.2 Estimates and test of the multiple linear regression model
  - A.2.1 Significance of parameters for full model

House	Intercept	Temp.	House	Intercept	Temp.
1	2.398533	-0.105537	36	2.852860	-0.122672
2	3.398808	-0.137495	37	6.27161	-0.36085
3	4.144182	-0.203752	38	3.58106	-0.20111
4	2.884192	-0.186361	39	6.32565	-0.38797
5	3.876744	-0.192553	40	4.333811	-0.223071
6	2.525637	-0.111327	41	3.270357	-0.209849
7	2.153872	-0.119296	42	2.655654	-0.112816
8	6.49637	-0.28241	43	3.462811	-0.180196
9	1.411321	-0.077829	44	1.785110	-0.082274
10	16.80251	-0.76470	45	25.1895	-1.1268
11	2.853030	-0.121927	46	2.860330	-0.153524
12	4.588960	-0.210243	47	1.299755	-0.088525
13	4.322333	-0.171578	48	3.070152	-0.162541
14	5.975906	-0.292424	49	3.99257	-0.18209
15	3.022654	-0.160148	50	4.463920	-0.195064
16	4.041873	-0.224724	51	4.501192	-0.204155
17	8.20034	-0.40646	52	4.94798	-0.33593
18	1.616462	-0.068947	53	2.753387	-0.152706
19	4.51546	-0.19795	54	3.123352	-0.166824
20	1.796715	-0.111773	55	2.634592	-0.110295
21	2.881053	-0.123945	56	2.19092	-0.09952
22	2.600607	-0.154493	57	3.007890	-0.128622
23	4.001741	-0.180271	58	2.603941	-0.115700
24	4.073325	-0.183069	59	3.581131	-0.181395
25	7.58839	-0.30079	60	4.61306	-0.25118
26	5.05844	-0.22397	61	3.71516	-0.17677
27	2.742857	-0.121458	62	3.906559	-0.173636
28	30.38456	-1.64664	63	11.79446	-0.50298
29	4.985258	-0.231925	64	2.327872	-0.122545
30	4.451597	-0.227997	65	2.375351	-0.128330
31	4.844060	-0.253319	66	5.364626	-0.264278
32	2.408233	-0.082787	67	3.712356	-0.145792
33	3.584799	-0.158777	68	4.61927	-0.22110
34	3.778189	-0.151325	69	4.022828	-0.180830
35	3.462811	-0.180196	70	3.625038	-0.167733

**Table A.1:** Estimates of the simple linear regression model.

House	Shapiro-Wilk	Sign	House	Shapiro-Wilk	Sign
1	7.458816e-09	0.154571	36	0.6551522	0.604916
2	0.03683475	0.897101	37	0.00211908	0.437686
3	0.1522292	0.897101	38	0.09384231	0.244235
4	0.1385351	0.897101	39	0.720376	0.853408
5	0.4889633	0.897101	40	0.0814062	0.698022
6	0.06605537	0.579290	41	0.7499187	0.698022
7	1.77266e-08	0.000096	42	0.5544048	0.195667
8	0.001297956	0.579290	43	0.9688807	0.853408
9	0.002104755	0.355280	44	0.05668529	0.795902
10	0.130679	0.704379	45	0.678235	0.517817
11	0.709406	0.698022	46	0.2546043	1
12	.319923e-07	0.437686	47	0.001065575	0.002847
13	0.0001239323	0.853408	48	4.674662e-09	0.897101
14	0.2859703	0.517817	49	0.0352531	0.698022
15	0.877681	1	50	0.9042102	0.604916
16	0.03106733	1	51	0.2314035	0.459691
17	0.4268815	0.579290	52	0.02903943	0.795902
18	1.30237e-05	0.300679	53	0.4957589	1
19	0.809942	0.355280	54	0.04954421	0.195667
20	0.3582973	0.711703	55	0.2428003	0.437686
21	0.1858593	0.517817	56	0.09574403	1
22	0.07530409	1	57	0.0213735	0.795902
23	0.006563705	0.300679	58	0.7065081	0.897101
24	0.9487515	0.579290	59	0.07945885	0.711703
25	0.2231339	0.459691	60	0.5896005	1
26	0.002878431	0.355280	61	5.313489e-06	0.019687
27	0.5797924	1	62	0.001550156	0.355280
28	0.008843725	0.300679	63	0.5771746	0.853408
29	0.006802018	0.365187	64	0.006368894	0.517817
30	0.08408281	0.897101	65	0.01968257	0.195667
31	0.001679153	1	66	0.0006757623	0.604916
32	8.699853e-07	0.013802	67	0.1761617	1
33	0.917481	0.517817	68	0.2359113	0.579290
34	4.238303e-05	0.517817	69	0.008934653	0.579290
35	0.9688807	0.853408	70	0.8963474	0.267182

**Table A.2:** P-values from Shapiro-Wilk test and sign test on the simple linear regression model.

Index	I	T	N	E	S	W	MSL	SR	WB	SB	AB	CB	WKND	T:N	T:E	T:S	T:W
1		-***	.	+***	+*			-***	-***					+. .	-***		
2	**	-***		**		+***		-*	-*							-***	
3	-***		+***			+***	+***	-***	+	-*					+**	-*	
4	-*	-***	+	*		+***		+*	-***								
5	-***		+***			+***	+***	-***	+*	+*	+***				+***	-**	
7	***	-***	-**	+***	-**	+**	-***	+**	+*	+***				-***	+*	-**	
11	**	-***		+***		+***		-***	+***						+**		
12	-***					+								-*			
14	+	-***		+***		+***		-**	+***		-*				+**		
18	***	-*	+***	-*	*	+**	-**	+	.		-.				+**	-***	
21	**	-***		+		+***										-***	
22	-***			+***	+	**		-***	-**						-*		
23	+	-***	-.	+		+***		-***			-.				+***	-***	
28	-***			+		+		-***	-*		-***	-***					
29	+	-***		+***		+***		-**	-*	+	.			+	+**	-***	
30	+	-***		+***		+***		-***	+*	-*					+	-.	
31	-***			+***		+***	+**	-***	+	.	-.				+***	-***	
32	-**			+		+***		-**	-*						+*	-*	
33	-***							-**									
34	+	-***		+***	+	+***		-**		-*	+***	-*				*	
36	**	-***		+***		+		-***							+**		
37	-***			+***	+	**		-***									
38	-***	-***		**		+***	+***	-***	+*	-.			-*	+	+	-***	
40	+	-***		+	*	+**		-***		-**	-*				+	-*	
41	***	-***		+***		+***		-.	+	.	-.				+	-.	
42	-***			+***		+***	+***	-***	+	.			-***		+*	-*	
44	-***	-*						-**							+**	+**	
45	**	-***			+	+***		-.							+	+	-**
46	-***			+	+	**		-**						-*		-*	
47	-***			+	+	+***	+	-*	-.				+***	+***	-*	-***	
48	+	-***		+	**	+**		-**	+**					-***		-**	
49	+	-***						-**									
50	-***	-*	+	***	-**	+		-***	+**					+		+***	
52	+	-***	-*	+	-*			-**	-***	-***		-***		+		+***	
54	-***			**		**		-***			-.	-*			+**	-*	
55	-***			+		+***	+*	-***			-.	-*			+***	-*	
56	**	-***		**		+***		-**	-.		-*				+*	-*	
57	+	-***		+***		**										-*	
58	-***	-.	+	.		+***		-***								-*	
61	***	-***		**		**	-***	+	**				-.		+**	-**	
64	-***	-*	+	***		**		-***	+				-*		-*	+**	
65	-***		+	*		+***		-***							+*	-**	
66	-***		+***	+	*	+***		-***		-.	-.				+**	-**	

**Table A.3:** Significance of parameters from the full multiple linear regression model performed on 'long' houses.

Index	I	T	N	E	S	W	MSL	SR	AB	CB	WKND	T:N	T:E	T:S	T:W
6		-***		+*		+**		-*	-.						-*
8	++*	-***		+*		+***		-*					+		-*
9	-***	-.			+*	+**	-***	-*		-*	+		+*	-**	
10	-***		+***		+***	+	-**			-**	+			-*	
13	-***	+*		+**	+*	-*				-.	+		++*	-*	
15	-***	+***			+**		-.		-.	-.				-.	
16	-***	+*		+***	+	-*								-**	
17	-***	+**	+**	+***			-**			-.				-***	
19	-***	+**	-.		+		-.		-.	-.	+		++*		
20	-***			+*											
24	-***	+*		+***		+*								-***	
25	-***	+**		+***								+		-*	
26	-***	+	+	+***	+*		-**	-***						-*	
27	-***	+*	+**	+***						-*				-*	
35	-***	+**	+*	+***			-**			+*				-**	
39	++*	-***			+*	-*	-.								
43	-***	+**	+*	+***			-**			+*				-**	
51	-***	+		+***			-**			-*		+		-*	
53	-***	+*		+*										-.	
59	-***	-.	+***	+*	+***		-.				+	+		-.	
60	-***	-*	+**	+	+***		-***				+*	+	+	-***	
62	++*	-***			+*		-.	-*							
63	+	-***		+*						-**			+		
67	++*	-***	+***		+***		-***	+	+*			-*	+	-**	
68	-***			+*		-***			+*	-**			+*	-*	
69	-***			+	+**		-.	-***				+			
70	-***		+	+	+***	+**	+***							-*	

**Table A.4:** Significance of parameters from the full multiple linear regression model performed on 'short' houses.

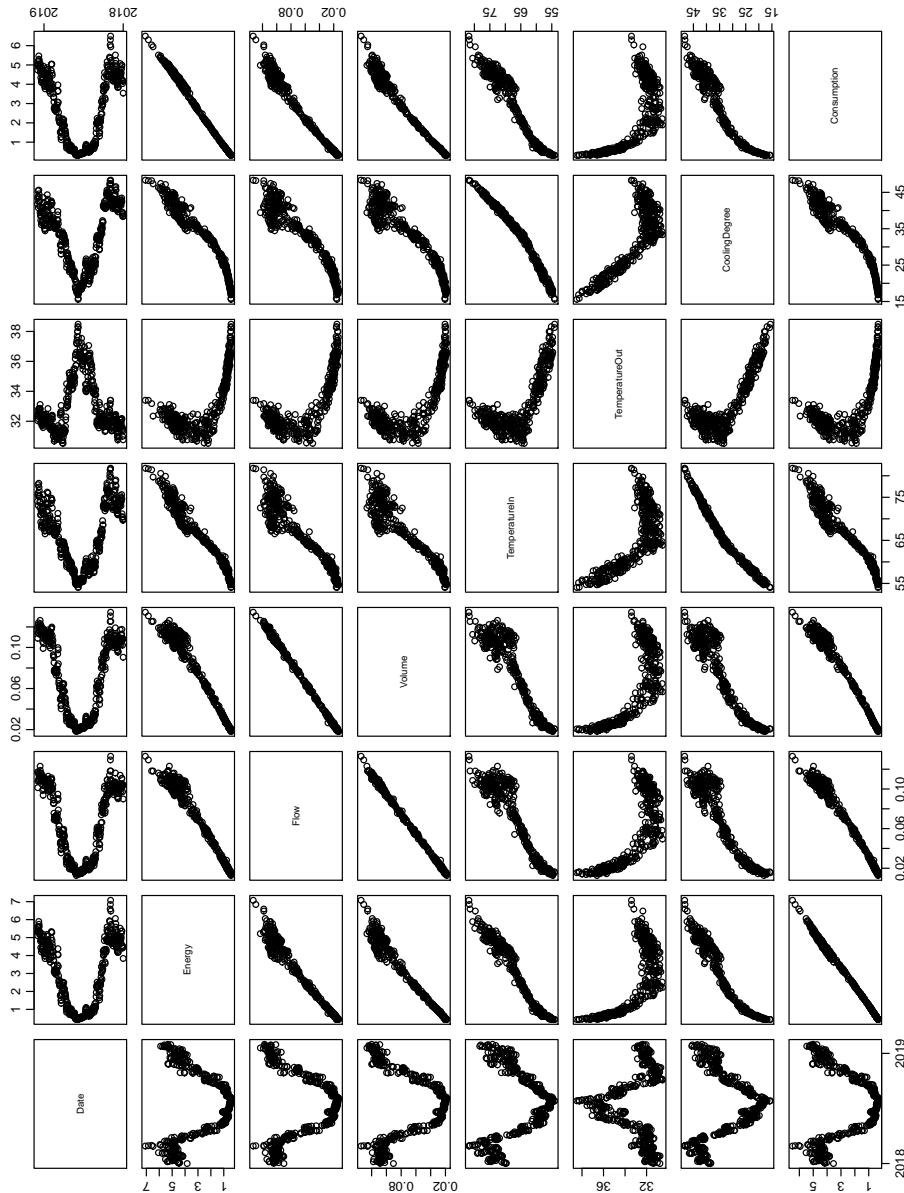
House	P-value	Sign	House	P-value	Sign
1	1.269662e-07	0.604916	36	0.0318435	0.897101
2	0.2120532	0.365187	37	0.5588932	0.897101
3	0.06289541	0.897101	38	0.002469713	0.244235
4	0.1693807	0.897101	39	0.1319804	0.579290
5	0.9020079	1	40	0.2151643	1
6	0.511858	1	41	0.7165372	0.604916
7	2.400099e-10	0.006472	42	0.5484956	1
8	0.3662207	0.355280	43	0.8539095	0.853408
9	0.01292524	0.195349	44	0. 09057111	0.517817
10	0.5039661	0.254605	45	0.8307638	0.897101
11	0.2445798	1	46	0.7925841	1
12	5.838099e-06	0.092399	47	1.763434e-05	0.120377
13	0.01173622	0.853408	48	4.364622e-05	0.517817
14	0.2712269	0.517817	49	0.09749062	0.517817
15	0.9682545	0.853408	50	0.6649854	0.897101
16	0.07409811	1	51	0.6341959	1
17	0.9973268	1	52	4.159959e-05	0.437686
18	1.017533e-06	0.365187	53	0.4897047	0.853408
19	0.2113022	1	54	0.008756327	0.154571
20	0.6677595	0.711703	55	0.5201494	0.517817
21	0.155591	0.897101	56	0.4827258	0.795902
22	0.003089513	1	57	0.005584645	0.795902
23	0.09877236	0.698022	58	0.4809962	0.698022
24	0.461353	0.853408	59	0.6763879	1
25	0.9855876	0.459691	60	0.6583508	0.711703
26	0.01226589	0.459691	61	2.841694e-05	0.052087
27	0.5094042	0.853408	62	0.009033716	0.267182
28	0.5752663	0.019687	63	0.6070834	0.355280
29	0.1056845	0.698022	64	0.002157001	0.795902
30	0.4512082	0.517817	65	0.09290001	0.437686
31	0.4481276	0.300679	66	0.002398244	0.795902
32	4.134475e-08	0.038238	67	0.735418	1
33	0.4464175	0.795902	68	0.2093689	0.853408
34	5.011367e-06	0.604916	69	0.0001833683	0.195349
35	0.8539095	0.853408	70	0.1393319	1

**Table A.5:** P-values from Shapiro-Wilk test for normality on the general multiple linear regression model.

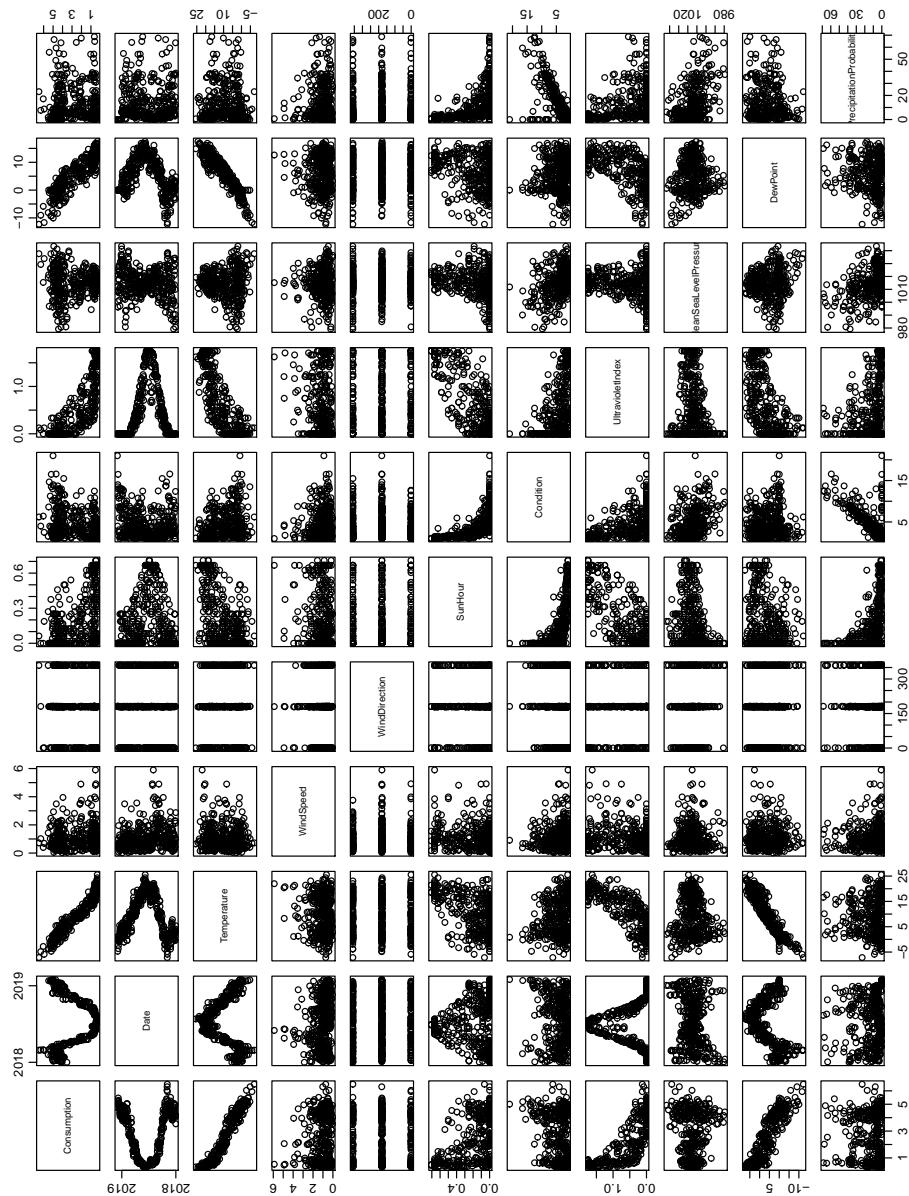
## APPENDIX **B**

# Figures

---



**Figure B.1:** Scatterplot showing the average of relevant attributes from house data.



**Figure B.2:** Scatterplot showing the average of relevant attributes from weather data.



# Bibliography

---

- [1] TV 2. *Se billedeerne - Danmark igen ramt af vintervejr.* URL: <http://vejr.tv2.dk/2018-04-03-se-billedeerne-danmark-igen-ramt-af-vintervejr>.
- [2] Aalborg Forsyning. *Udfyldning af huller i tidsserier og udglatning af Kamstrup time aflæsninger.* 2019.
- [3] International Energy Agency. *Renewables 2018 - Market analysis and forecast from 2018 to 2023.* URL: <https://www.iea.org/renewables2018/heat/>.
- [4] Michael J. Crawley. *Statistics: An introduction using R.*
- [5] *Description of the attribute Condition.* URL: [https://www.wunderground.com/weather/api/d/docs?d=resources%2Fphrase-glossary&\\_ga=2.166768631.87242339.1555497846-195014723.1555497846&fbclid=IwAR3NANmRzhGE0uS4m5YqMu06\\_qJDJFQ7-Wd3xWjKPV68Njtso\\_4PSVuVk9o](https://www.wunderground.com/weather/api/d/docs?d=resources%2Fphrase-glossary&_ga=2.166768631.87242339.1555497846-195014723.1555497846&fbclid=IwAR3NANmRzhGE0uS4m5YqMu06_qJDJFQ7-Wd3xWjKPV68Njtso_4PSVuVk9o).
- [6] Henrik Madsen. *Time Series Analysis.*
- [7] Nielsen, P. N., Gravesen, J., Gersborg, A. R., & Pedersen, N. L. *Isogeometric Analysis and Shape Optimization in Fluid Mechanics.* 2012.
- [8] Simon J. Sheather. *A Modern Approach to Regression with R.*

