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
www.compute.dtu.dk

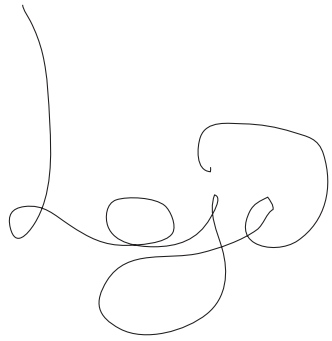
Abstract

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.

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 3, 2019



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

Acknowledgements

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.

Contents

CHAPTER 1

Introduction

1.1 Motivation

1.2 Introduction to WATTS app

CHAPTER 2

Data

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 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 which will be taken into account when cleaning the data.
Table 2.1 below shows the attributes from the house data set.

Variable	**Description**
StartDateTime	Start time and date for measurements.
EndDateTime	End time and date for measurements.
Energy	Electricity 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..

The Heat Consumption is defined as

$$Q = V \cdot \Delta T \tag{2.1}$$

The weather data set consists of the attributes seen in Table 2.2.

Variable	Description
StartDateTime	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	
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..

The BBR data set consists of the attributes seen in Table 2.3.

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 `StartDateTime` and `EndDateTime` is always one hour apart, it is redundant to use both of the attributes. The obervations of most of the attributes are made at time `EndDateTime`, and for that reason it is used as `ObsTime` for the observations. For the weather data set, the observations is made at time `StartDateTime`, and there is no `EndDateTime` for this data set. When merging these data sets, `ObsTime` is

alligned with `StartDateTime`. 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 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.

In the house data there are some measurements missing 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.

2.2.1 Missing values

Der hvor der mangler data har vi udregnet dem ved interpolation som givet i det der paper fra Anders. For at kunne lave tidsrække modeller, må der ikke være NA's i data

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.

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. 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 is 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. Det varme vand der kommer ud af systemet afhænger af hvor meget man bruger. Det ses, at hvis det vand der kommer ud er meget varmt, så er det fordi man ikke har brugt det, eller trukket sådan varmeforbrug ud af det.

It is already known that there is a dependency between the heat consumption and the time of the year. During the summer period there is almost no consumption. The

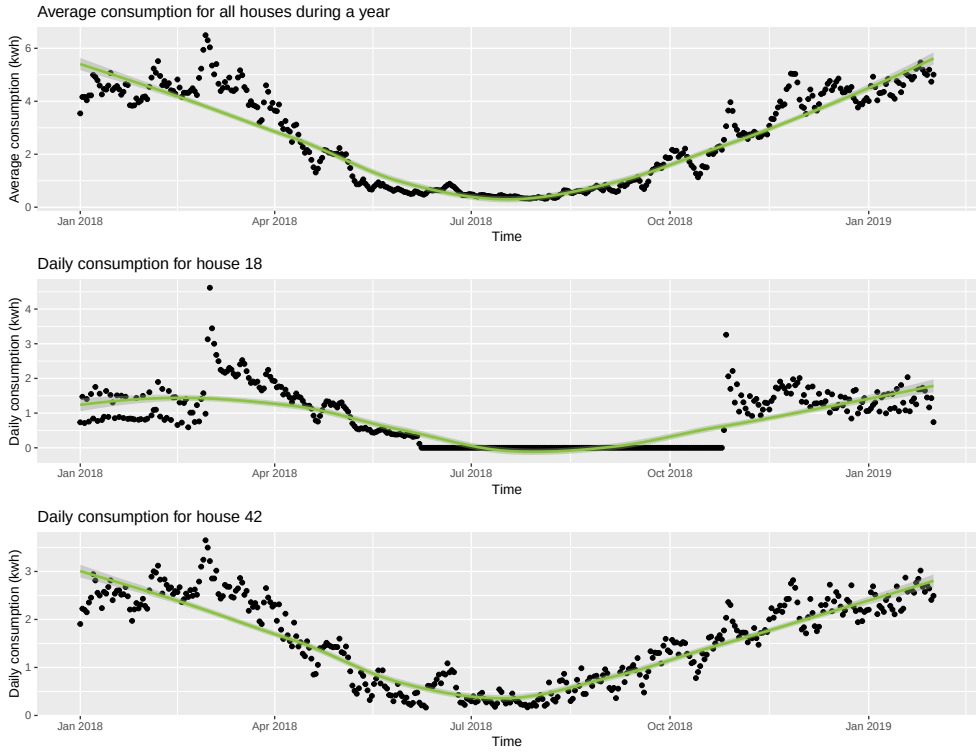


Figure 3.1: Daily consumption during a year (2018). The top plot shows the average consumption for all the houses. The plot in the middle shows an example of a house that follows the trend and the last plot shows a house that deviates from the trend.

consumption in this period is probably mostly tap water. The next important thing is the relation between temperature and consumption. High temperatures tend to imply a higher consumption. And the reason why the consumption depends so clearly on the time of year can be assumed to that certain periods have similar temperature levels. It can also be seen that there is a correlation between dewpoint and consumption. This can be due to the correlation between dewpoint and temperature.

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. The temperature outside has a high

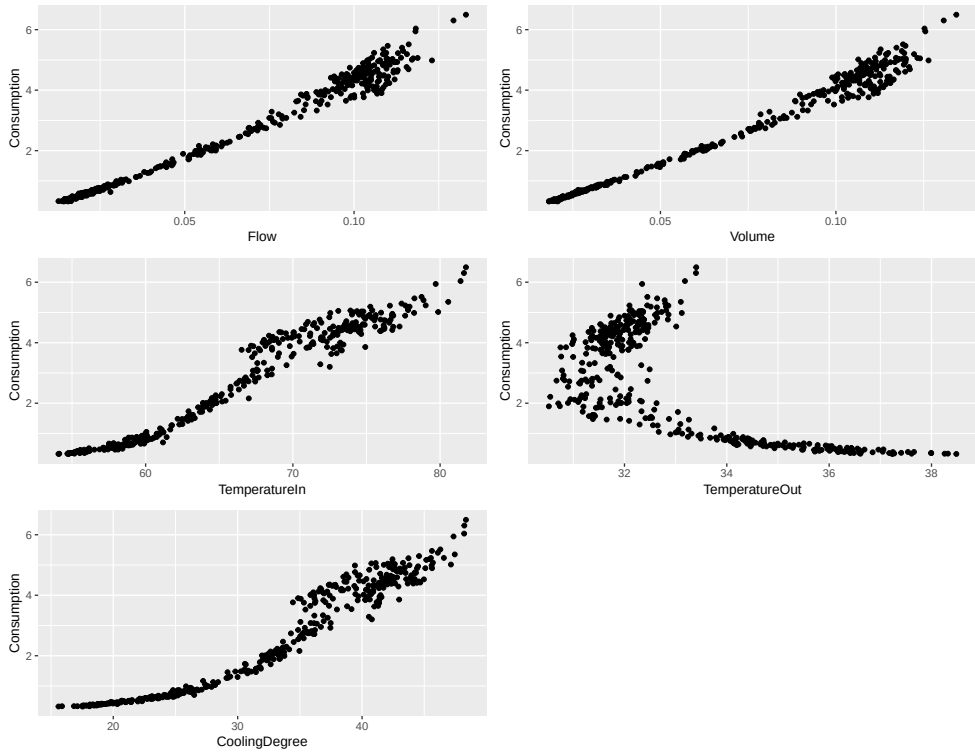


Figure 3.2: Scatterplots of the daily values of relevant house attributes. There are clear linearly dependencies between `Consumption` and e.g. `Flow`.

3.1.2 BBR data

Presumably, the BBR data has influence on the heat consumption in particular the total area and year of construction. Mangler lidt her.

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 3.4 shows the year of construction and the consumption for each of the houses. The year of construction is here determined by either the year of construction or the year of the latest reconstruction of a house.

Figure 3.4 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

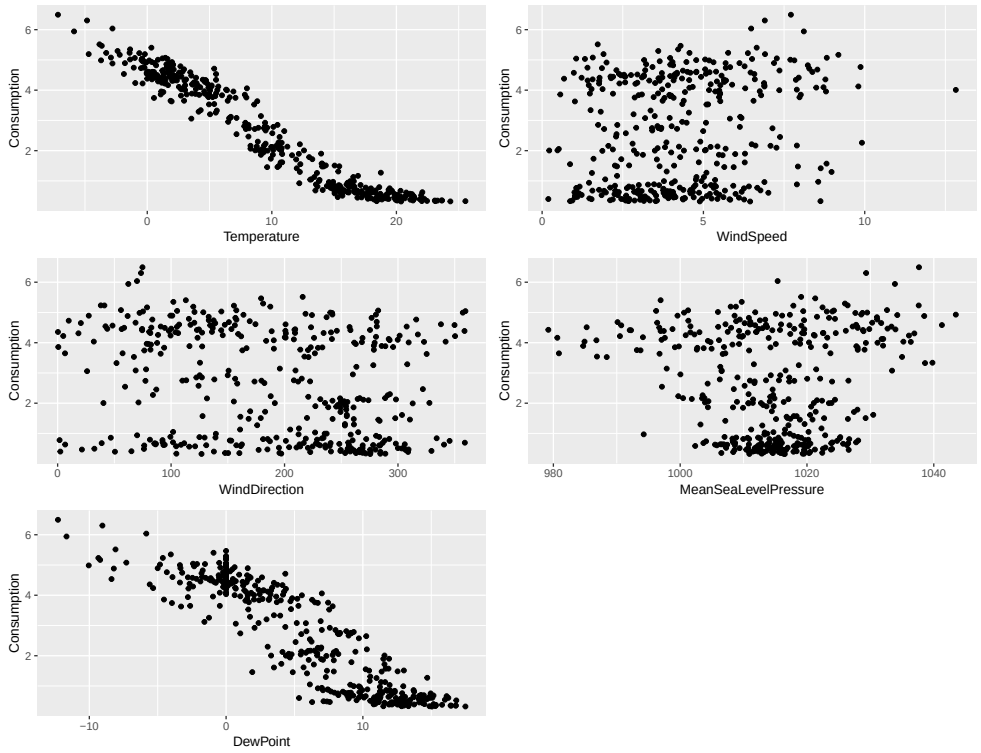


Figure 3.3: Scatterplots of the daily values of relevant weather attributes. There are clear linearly dependencys between `Consumption` and `Temperature` as was expected.

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 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`.

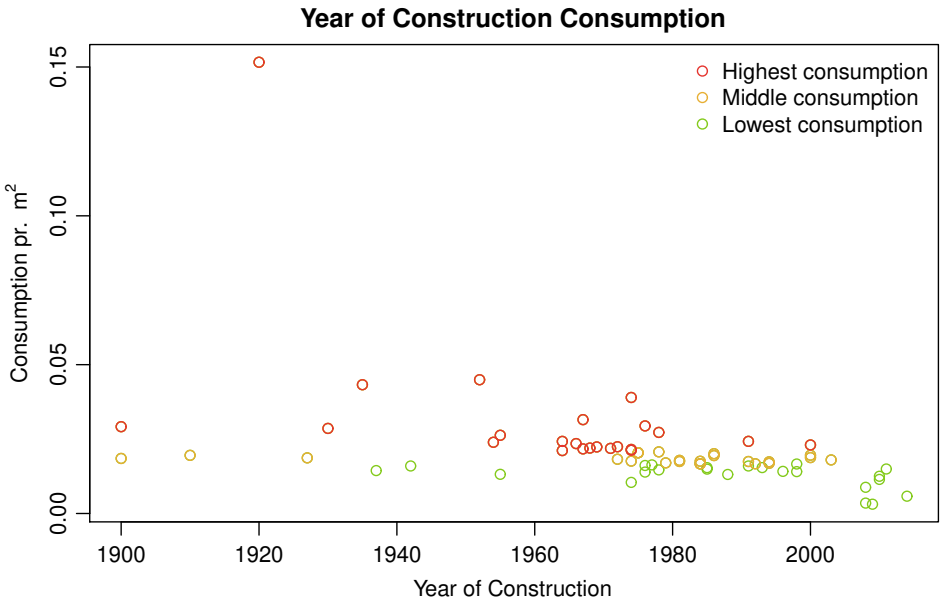


Figure 3.4: Plot showing the year of construction and the average consumption pr. m^2 for each house. It is clearly seen that there is a tendency that the later a house is built or reconstructed, the better is the insulation of the house.

This is the basis for expecting a correlation between the radiation and the sun attributes. Figure 3.5 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.

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 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

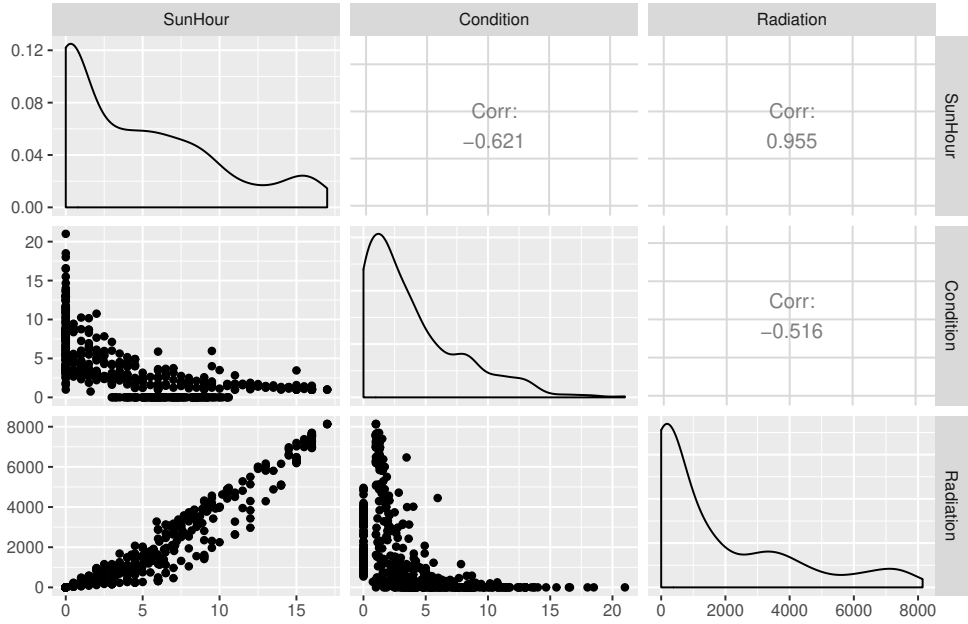


Figure 3.5: Scatterplot showing the correlations between the three attributes `Condition`, `Radiation` and `SunHour`. It is clearly seen that the radiation and the sun hour are highly correlated.

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.

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 α 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 3.6 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 3.7. On the left the data points are plotted with standard deviations on the y-axis. The red line highlights the two standard devations. 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

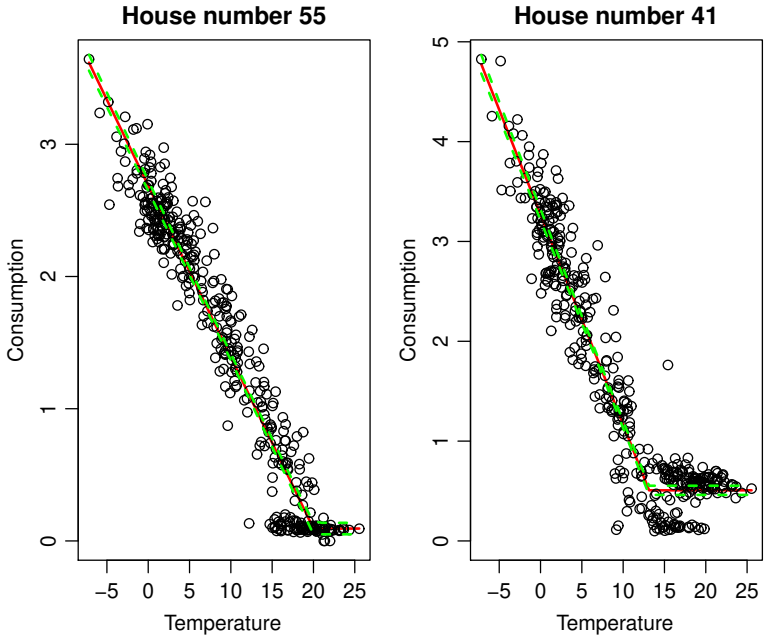


Figure 3.6: Piece-wise optimization of the consumption. The red line is the regression line and the green line is the confidence interval..

occur at temperatures both with and without heating. When comparing figure 3.7 to figure 3.6, 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 unforseen heat consumption. Figure 3.8 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.

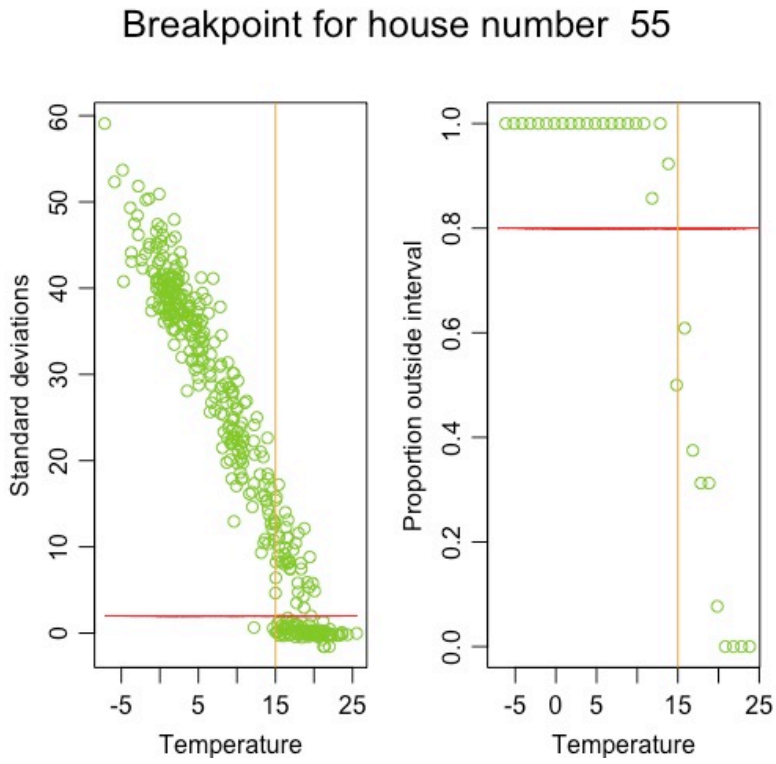


Figure 3.7: An illustration of how the breakpoint is found using segmentation by significant deviations. On the left figure the line illustrates two standard deviations from the high temperature distribution. The right figure shows how many points are outside the two standard deviations. The last point below 80% is the chosen breakpoint.

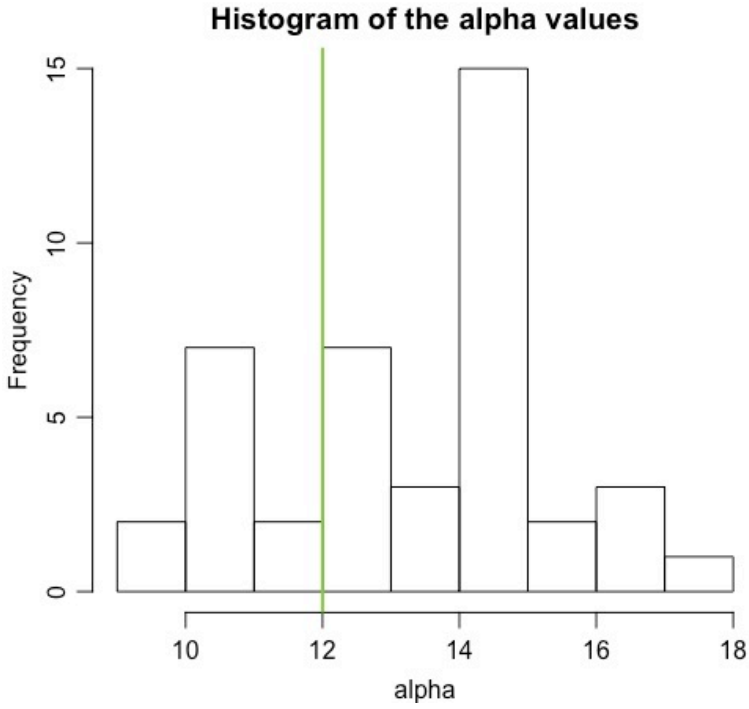


Figure 3.8: A histogram of the alpha values for every house in the third segmentation method. The first quantile is chosen as the overall breakpoint. It is 12 degrees, illustrated by the green line.

CHAPTER 4
Statistical models

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. Mangler nok lidt her. With the dependent variable Y and the independent variables $x_1, \ldots, x_n$, the linear regression model is formulated as

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

The variables ε_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

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

is chosen for estimating the model. The parameters β_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 homogenity

3 Variance should be independent of location

4 Linear relationship between x_j and Y

Chapter 7 in [2] explains the model assupmtions 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 [4] 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 signifance 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 signifance is determined by an F-test where the variables selected have a p-value below a threshold which is chosen at 0.05. Mangler nok en ref. til det her.

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.

4.2 Simple linear regression model

A simple linear regression model is fitted to each house with `Consumption` as a function of `Temperature` (here denoted HC and T). 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. The models are performed by using the

`lm()` function. The models will then be validated by examining whether the model assumptions in Chapter 4.1.1 are met. The simple model only includes one explanatory variable, thus a variable selection is not performed. Begrundelse for at vi ikke transformerer data.
Hence, the simple linear regression model applied to each house is

$$Y_{\mathrm{Q}} = \mu + \beta_{\mathrm{T}} \cdot x_{\mathrm{T}} + \varepsilon \tag{4.3}$$

4.2.1 Validation

To validate the model, different methods are used. The abovementioned model assumptions are checked and futhermore a test of normally distributed residuals is 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 a house that overall fulfill the assumptions skal nok lige skrives lidt bedre :-)). The top left plot in Figure 4.1 clearly shows that the residuals are not randomly scattered. Mangler noget her.
In addition, Table A.2 shows the p-values from the Shapiro-Wilk test and it is clear that the majority of the simple models have residuals where the p-value is above the significance level. Model assumptions kan ikke sige at være opfyldt, når de ikke er opfyldt i størstedelen af tilfældene

4.2.2 Results

Table A.1 shows the estimates from all the simple linear regression models.
Overordnet kan den simple lineære regressionsmodel ikke beskrive trenden. Den antager, at temperaturen er den eneste faktor der påvirker husenes varmeforbrug. Men ved at undersøge hvorvidt model assumptions er opfyldt, så 'failer' 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.

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. 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

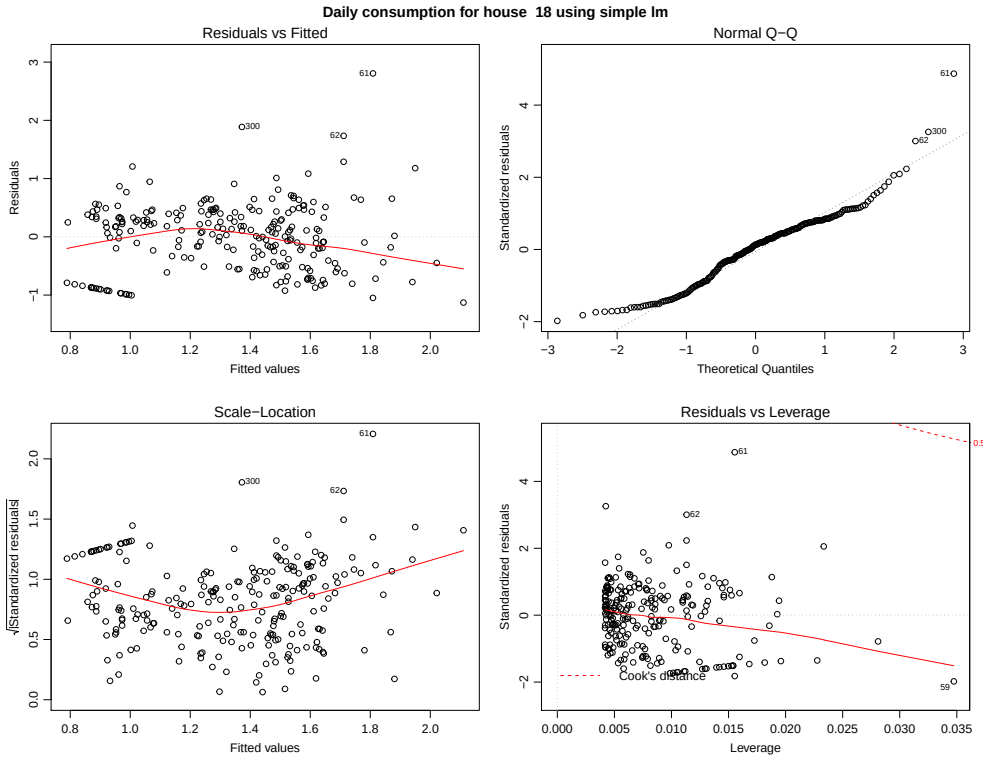


Figure 4.1: Residual plots of house 18 based on the simple linear regression model given in (4.3). The model assumptions of a linear regression model are not fulfilled for this specific house.

approximately six months. Thus, the two distinct lengths of observations are modeled slighty different. There do not exist observations for winter break and spring break in the data containing the short houses. This lead to the following two multiple linear regression models:

$$Y_{m,L} = \mu + \beta_{\mathrm{Q}} \cdot x_{\mathrm{T}} + \ldots + \varepsilon \tag{4.4}$$

$$Y_{m,S} = \mu + \beta_{\mathrm{Q}} \cdot x_{\mathrm{T}} + \ldots + \varepsilon \tag{4.5}$$

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), the

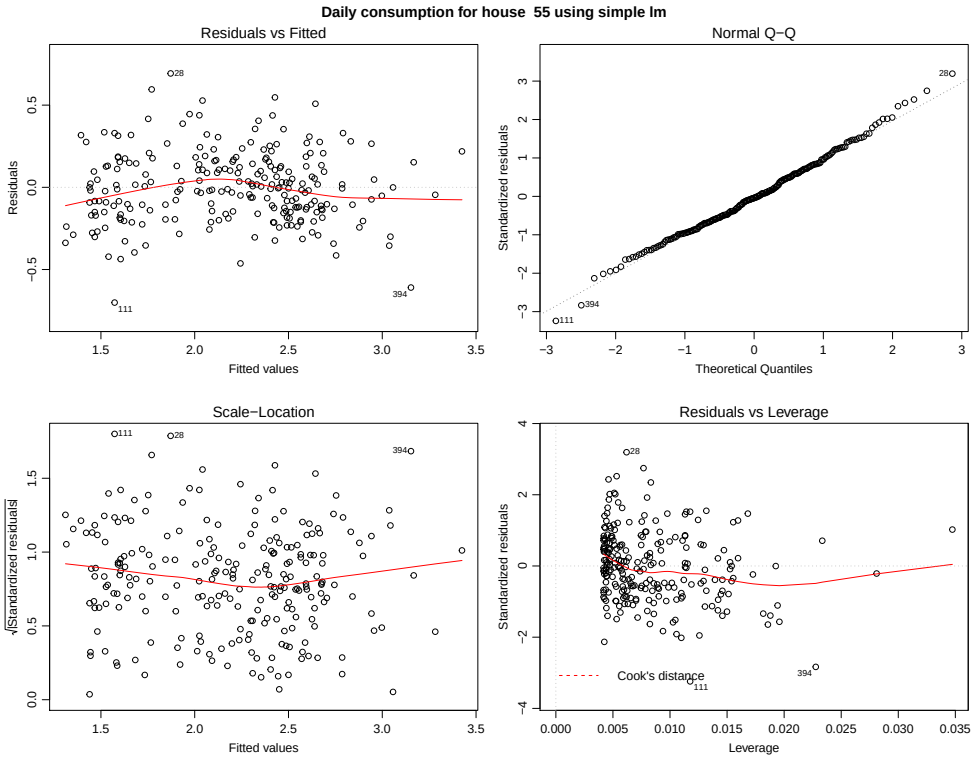


Figure 4.2: Residual plots of house 55 based on the simple linear regression model given in (4.3). The model assumptions of a linear regression model are overall fulfilled.

interaction between the temperature and the different wind directions (T:N, T:E, T:S, T:W). Mangler en lille overgang her, synes jeg.

4.3.1 Splines

In the multiple regression model, splines will be used to model the wind direction. It does not make much sence to include the wind direction as it is in the model. It is not useful to know how significant the wind direction is, if it is not connected to the wind speed and if it is not known which directions are important. By modeling the wind direction with splines, each spline will represent a specific general direction.

Modellere wind direction Lave en parameter om til flere vind retninger. 2. degree splines Knots Mellem retningerne Giver mere mening for brugeren

Vi vil gerne vægte vores vind i forskellige retninger i vores model, så derfor bruger

vi splines til at modellere de forskellige retninger.

Tilføj billede af splines

4.3.2 Validation

As for the simple linear regression model, the multiple models has to fulfill several assumptions.

4.3.3 Results

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 each parameter for the different holidays occurs in less than half of the cases.

Hvis bare én af retningerne har over halvdelen signifikant skal alle retningerne med.

	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 signifance 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	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 signifance is in relation to the number of long houses.

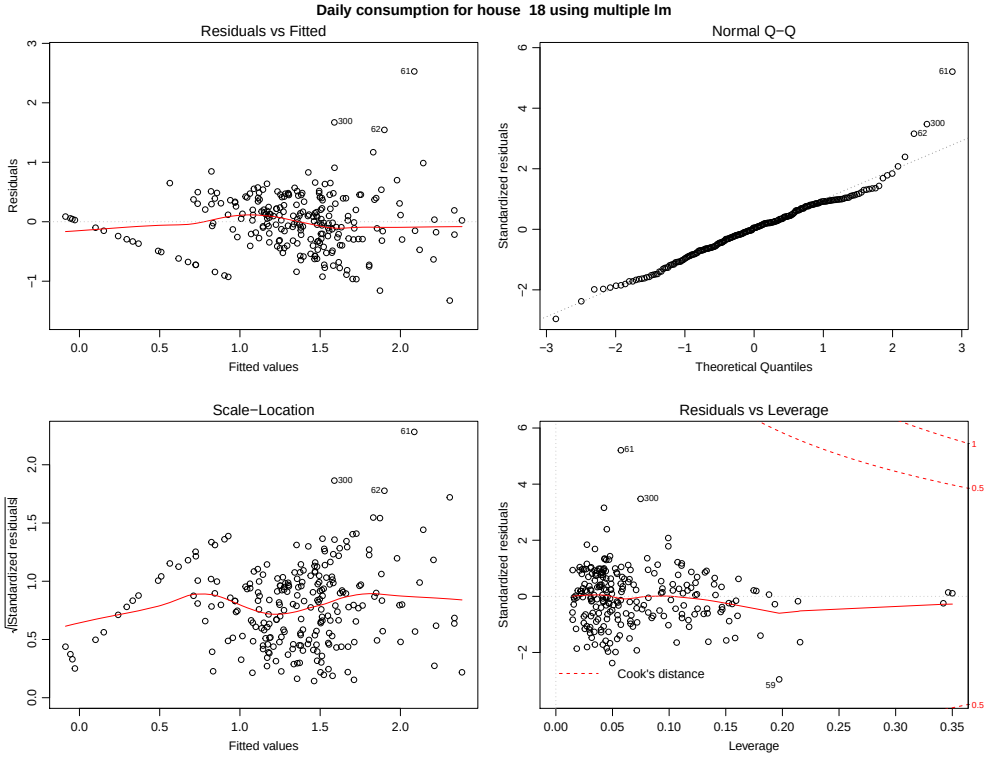


Figure 4.3: Diagnostic plot of the multiple linear regression model for long house 18.

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.

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. Punktum betyder, at det er mellem 5 og 10%.

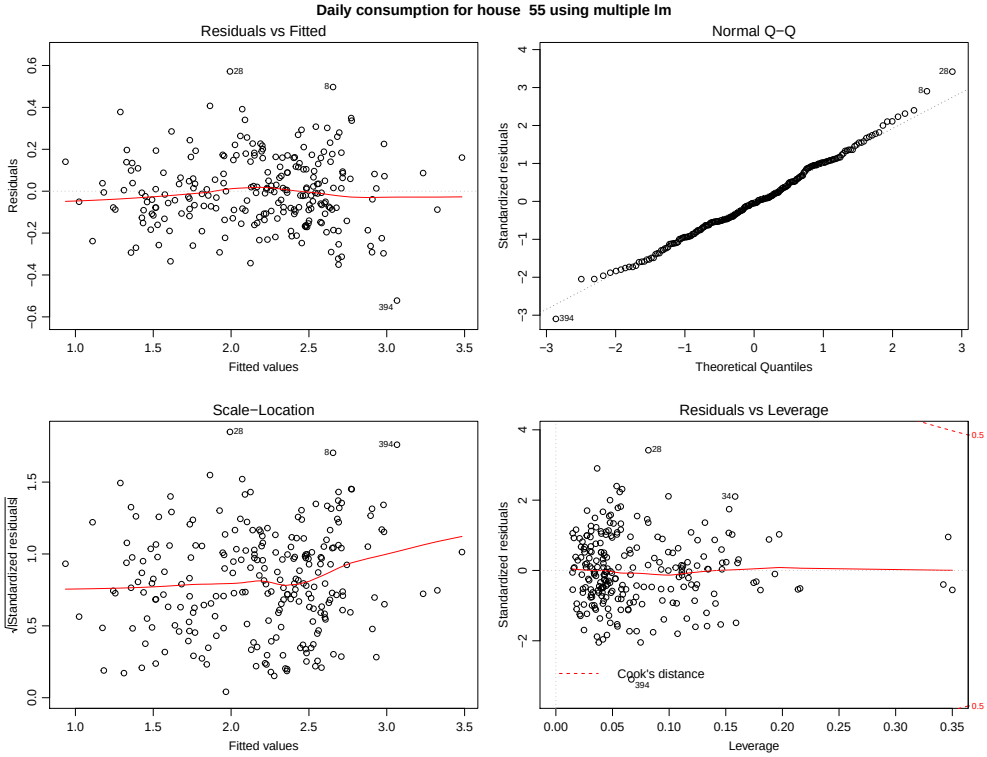


Figure 4.4: .

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	+***	−***	−*	+**		+***	−***	+**		+**	−***
62	+***	−***				+*	−*				

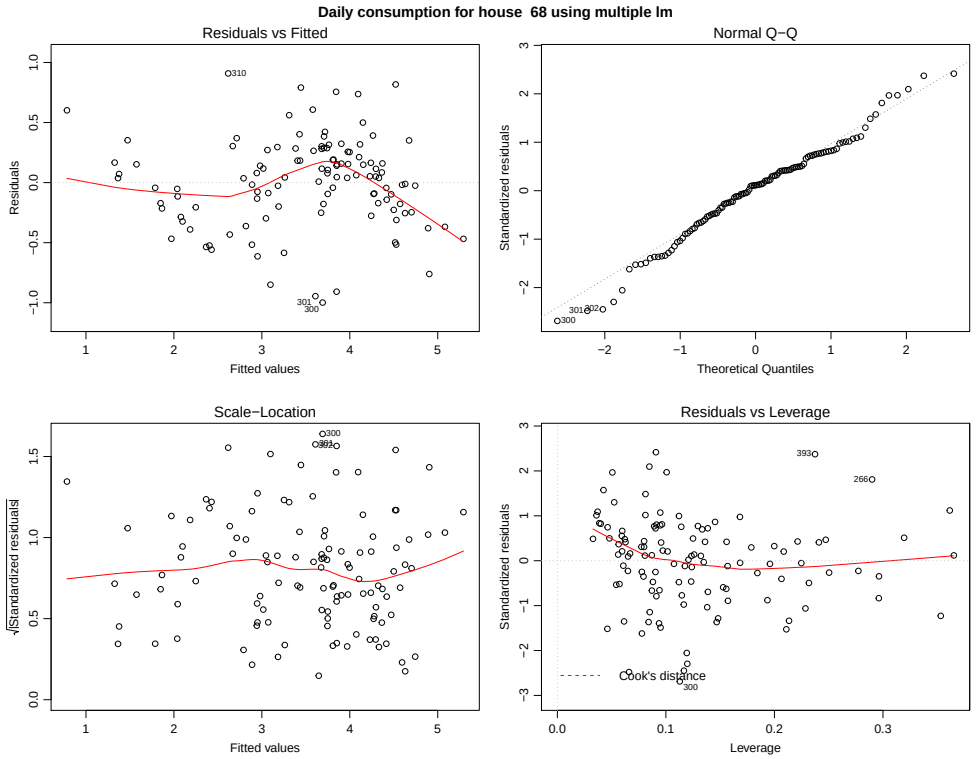


Figure 4.5: .

4.4.1 Validation

4.4.2 Results

4.5 Comparison

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 generel 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. These trends make sence. 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 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 con-

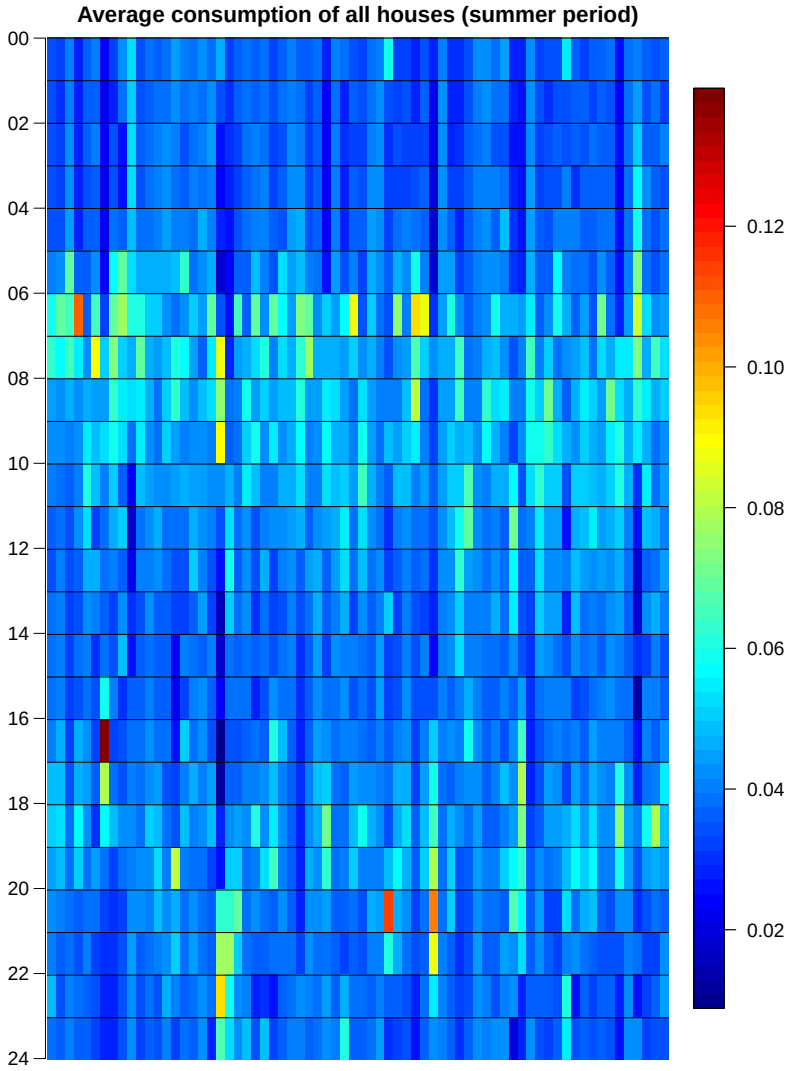


Figure 5.1: The normalized average consumption of every house during the day in the summer period. This is characterized by the days where the average outside temperature is above 15 degrees. The horizontal lines indicate the hours and each vertical strip is a house. The scale indicates the fraction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

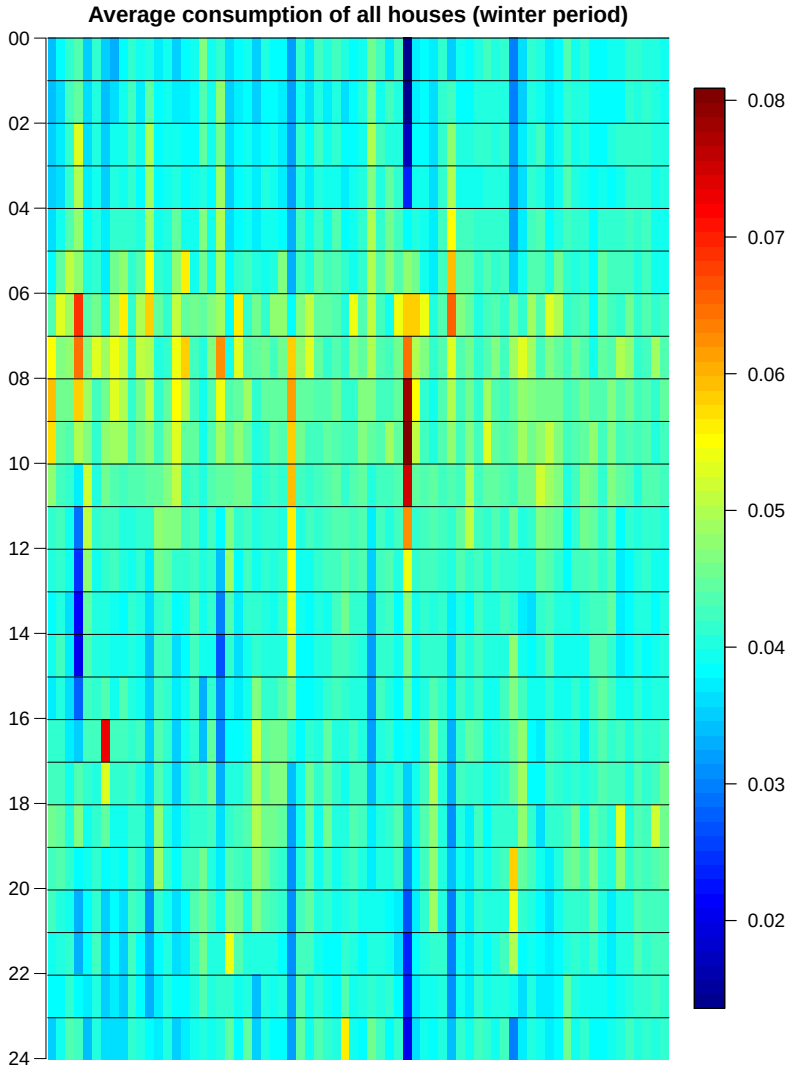


Figure 5.2: This figure shows the same as Figure 3.1, but only in the winter period, characterized by an outside temperature below 12 degrees.

sumption 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 $138m^2$. 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 surpricing that the consumption in the night hours is lower than the average. Neither is it surpricing 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 surpricing 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. ehh, den sidste sætning er måske lidt cheesy?

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 [3]. The ARMA model fits the data to a linear stochastic process, with an autoregression part (AR) and a moving average part (MA). A linear process $\{Y_t\}$ is a process that can be written as

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

where μ is the mean of the process, $\{\epsilon_i\}$ is white noise and $\{\psi_i\}$ is the weights. For now, the mean μ is assumed to be zero. To define the ARMA model, the backwards

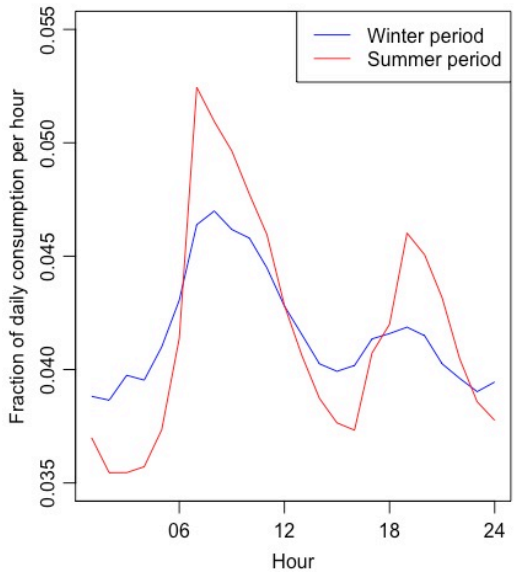


Figure 5.3: The average distribution of the heat consumption during the day for the winter period and the summer period respectively. The winter period is more smoothed out, but they are very similar.

shift operator B is first introduced as $B(Y_t) = Y_{t-1}$. An ARMA process has the form

$$\phi(B)Y_t = \theta(B)\epsilon_t, \tag{5.2}$$

where ϕ and θ 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 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

$$\phi(B)\Delta^d Y_t = \theta(B)\epsilon_t, \tag{5.3}$$

where $d \in \mathbb{N}$ is the differencing factor.

An ARIMAX model is constructed for the hourly data of every house. The outside temperature is used as the exogenous variates. The exploratory analysis has shown that there is a high covariance between the temperature and the consumption, making this the obvious choice. The model is expected to have a seasonal component with season 24, because the consumption in certain time periods are likely to be close to each other from day to day.
In the following a general modelling approach will be used where a large model is applied at first. It is then reduced by looking at the standard deviation of the estimates on all houses. The initial model is an ARIMAX(2,2,2) model with a seasonal (0,1,1) component. This is a very extensive model, and differencing is used to make it stationary. Figure 5.5 shows an example of how the model performs on house 55, while Figure 5.4 shows the same model for house 18. For house 18 both the ACF and the PACF ressemble white noise fairly well, with a few exceptions. The same goes for house 55. Both have a few significant lags, and they both have a significant lag around lag 61, even though it is barely so. It is supricing that neither house have lags in the seasons that are even close to being significant. This indicates that the dependencies from day to day might not be as important as it was initially assumed. But as mentioned, this model has some parameters that might not be necessary to include.

Parameters	AR1	AR2	MA1	MA2	SMA1	Intercept	Temperature
Insignificance	24%	81%	15%	13%	0%	0%	3%

Table 5.1: For each parameter in the first ARIMAX model, this table shows how many of the houses have an estimate that is less than two standard deviations. The average Log Likelihood of the model is -77.

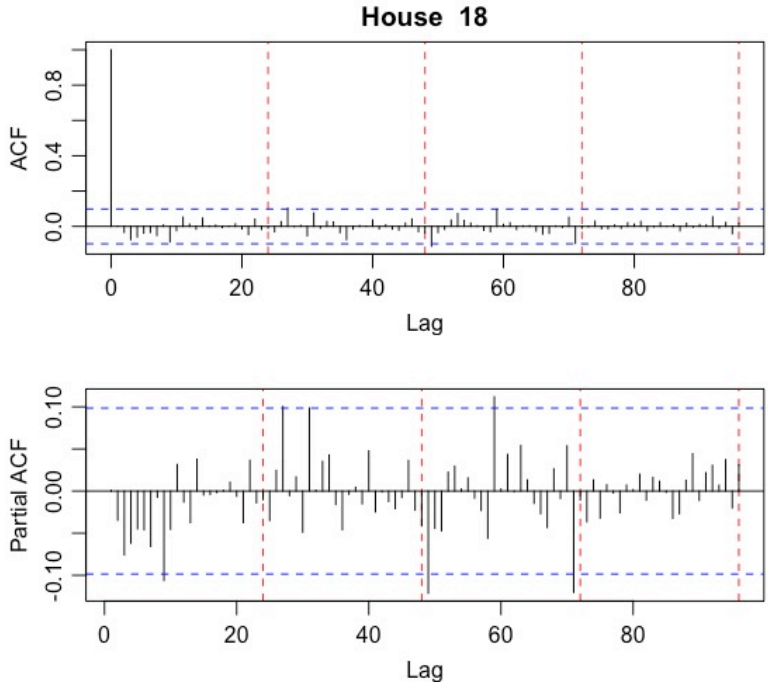


Figure 5.4: The autocorrelation function for the ARIMAX(2,2,2)×(0,1,1) model, based on the data from house 18.

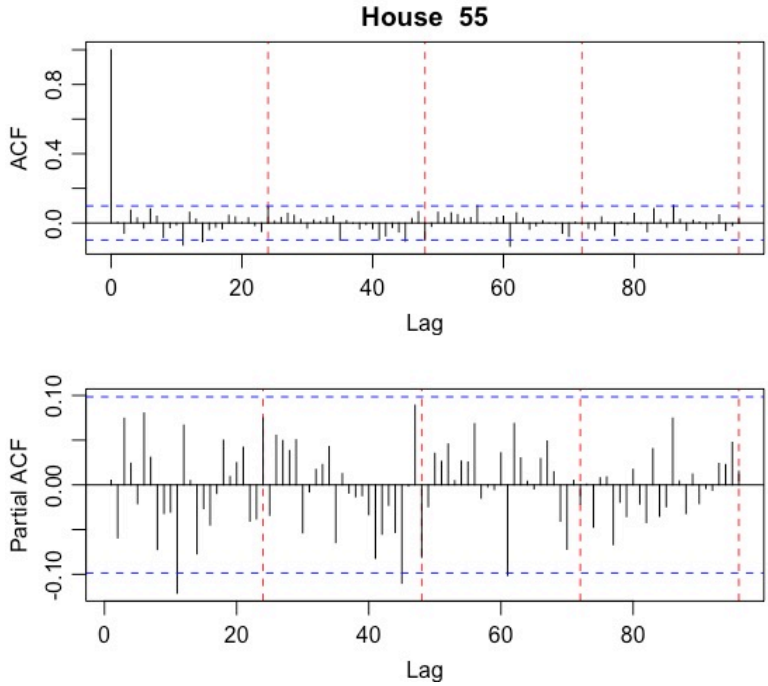


Figure 5.5: The partial autocorrelation function for the ARIMAX(2,2,2)×(0,1,1) model, based on the data from house 18.

APPENDIX A

Tables

A.1 Estimates of the simple linear regression model

A.2 Significance of parameters from multiple linear regression 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	P-value	House	P-value
1	1.269662e-07	36	0.0318435
2	0.2120532	37	0.5588932
3	0.06289541	38	0.002469713
4	0.1693807	39	0.1319804
5	0.9020079	40	0.2151643
6	0.511858	41	0.7165372
7	2.400099e-10	42	0.5484956
8	0.3662207	43	0.8539095
9	0.01292524	44	0.09057111
10	0.5039661	45	0.8307638
11	0.2445798	46	0.7925841
12	5.838099e-06	47	1.763434e-05
13	0.01173622	48	4.364622e-05
14	0.2712269	49	0.09749062
15	0.9682545	50	0.6649854
16	0.07409811	51	0.6341959
17	0.9973268	52	4.159959e-05
18	1.017533e-06	53	0.4897047
19	0.2113022	54	0.008756327
20	0.6677595	55	0.5201494
21	0.155591	56	0.4827258
22	0.003089513	57	0.005584645
23	0.09877236	58	0.4809962
24	0.461353	59	0.6763879
25	0.9855876	60	0.6583508
26	0.01226589	61	2.841694e-05
27	0.5094042	62	0.009033716
28	0.5752663	63	0.6070834
29	0.1056845	64	0.002157001
30	0.4512082	65	0.09290001
31	0.4481276	66	0.002398244
32	4.134475e-08	67	0.735418
33	0.4464175	68	0.2093689
34	5.011367e-06	69	0.0001833683
35	0.8539095	70	0.1393319

Table A.2: P-value from Shapiro-Wilk test for normality.

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.Punktum betyder, at det er mellem 5 og 10%.

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. Punktum betyder, at det er mellem 5 og 10%.

APPENDIX B

Figures

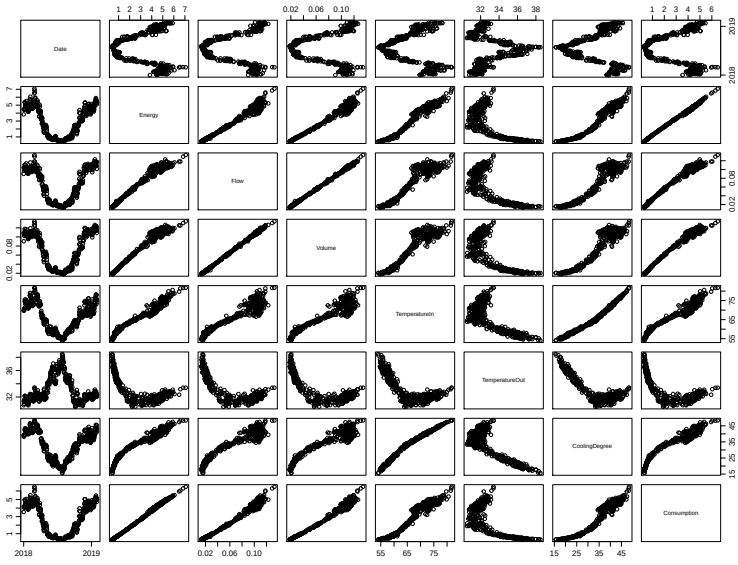


Figure B.1: Scatterplot showing the average of relevant attributes from house data.

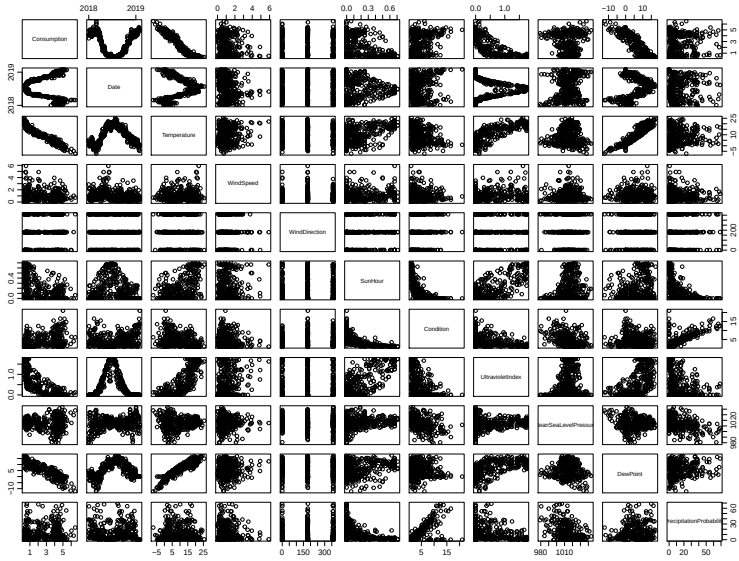


Figure B.2: Scatterplot showing the average of relevant attributes from weather data.

Bibliography

[1] TV 2. *Se billederne - Danmark igen ramt af vintervejr*. URL: `http://vejr.tv2.dk/2018-04-03-se-billederne-danmark-igen-ramt-af-vintervejr`.

[2] Michael J. Crawley. *Statistics: An introduction using R*.

[3] Henrik Madsen. *Time Series Analysis*.

[4] Simon J. Sheather. *A Modern Approach to Regression with R*.

{2.7182818284}ϑφερτυθιοπσδφγηξκλ