

 **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, May 10, 2019

A handwritten signature in black ink, appearing to read 'Anton Stockmarr', with a large, stylized 'S' at the end.

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

Abstract	i
Preface	iii
Acknowledgements	v
Contents	vii
Todo list	ix
1 Introduction	1
1.1 Motivation	1
2 Data	3
2.1 Original data	3
2.2 Cleaning and preparation	4
2.2.1 The sun and the wind	5
3 Exploratory Analysis	7
3.1 Examination of heat consumption	7
3.1.1 BBR data	8
3.2 Data segmentation	10
3.2.1 Segmentation by piece-wise optimization	11
3.2.2 Segmentation by significant deviations	11
3.3 Multicollinearity	13
4 Statistical models	17
4.1 Linear regression	17
4.1.1 Model assumptions	17
4.2 Simple linear regression model	18
4.2.1 Results	18
4.3 Multiple linear regression model	18
4.3.1 Splines	19
4.3.2 Results	20
4.4 Regression model for comparing houses	20

4.5	Comparison	20
5	Vejledningsmøder	21
5.1	19. februar	21
5.1.1	Spørgsmål	21
5.1.2	Noter	21
5.1.3	Hvad skal vi?	22
5.2	26. februar	23
5.2.1	Spørgsmål	23
5.2.2	Noter	23
5.2.3	Hvad skal vi lave?	24
5.3	5. marts	25
5.3.1	Spørgsmål	25
5.3.2	Hvad skal vi have lavet?	25
5.3.3	Noter	25
5.3.3.1	Til projektplan:	25
5.3.3.2	Andet:	26
5.4	12. marts	27
5.5	19. marts	28
5.5.1	Spørgsmål	28
5.5.2	Noter	28
5.5.3	Hvad skal vi lave?	29
5.6	26. marts	30
5.6.1	Noter	30
5.6.2	Hvad skal vi lave?	30
5.7	5. april	31
5.7.1	Spørgsmål	31
5.7.2	Noter	31
5.8	8. april	33
5.8.1	Noter	33
5.9	23. april	35
5.9.1	Spørgsmål	35
5.9.2	Noter	35
5.9.3	Hvad skal vi lave?	36
5.10	30. april	37
5.10.1	Inden møde	37
5.10.2	Spørgsmål	37
5.10.3	Noter	37
5.10.4	Hvad skal vi lave?	38
5.11	7. maj	38

Todo list

	5.2 (1) Daily averages of consumption versus temperature differences	23
	5.2 (2) Læse artikler fra Peder	23
	5.3 (3) få styr på lorte parskip-pakken	25
	5.3 (4) Få aksefis af Grønning eller Maika	25

CHAPTER 1

Introduction

1.1 Motivation

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 **antal** data points in all. The second data set includes weather data containing 11,845 observations with 11 attributes. Furthermore, the third data set is from Byggnings- og Boligregistret (BBR) and contain details for each of the houses e.g. total area, year of construction and type of house. **Mangler muligvis lidt mere her.**

The main focus of this section will be how 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 to 29-01-2019. 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. Hourly values.
EndDateTime	End time and date for measurements.
Energy	Electricity consumption in <i>kWh</i> .
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 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	
WindDirection	
SunHour	
Condition	
UltravioletIndex	
MeanSeaLevelPressure	
DewPoint	
Humidity	
PrecipitationProbability	
IsHistoricalEstimated	

Table 2.2: Attributes from the original weather data..

StartDateTime and EndDateTime are always one hour apart. When there are missing observations the following the next StartDateTime is simply delayed. Energy is the measured energy consumption on the meter in the houses.

Noget med at vi også har 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. [Synes der mangler et eller andet her](#).

Both weather data and the house data are aggregated in order to convert hourly values into daily values since there are of interest when modelling i chapter 3. [Loader en temporary data ind, som vi modificerer indtil vi putter den ind i vores endelige data](#). Data from 2017 in the house data are removed since data for the same period is missing in the weather data. The format for the attributes `StartDateTime` and `EndDateTime` is changed to d-m-Y H:min:sec. Likewise, the attribute `StartDateTime` in the weather data is converted to the same format as in the house data in order to merge the two data sets.

For nogle huse er der nogle hourly measurements der ikke er der. Der er huller i målingerne. Disse udfyldes med null, hvilket er bedre/lettere at arbejde med.

Attributen `IsHistoricalEstimated` ændres til logical, så vi kan compute med den.

Vi laver så temp. weather data så vi kan merge det med house data. Vi merger ikke al data, da mængden vil være en del større. Vi merger tmp weather data på house data i model processen.

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.

Vi tilføjer en binær attribute for hver ferie, og endnu en for weekender. De forskellige ferier vi tager med er christmas break, winter break, spring break, autumn break

2.2.1 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 heat consumption

To get an overview of the heat consumption for each house, the daily average consumption for each house has been calculated and can be seen as a function of the time in figure 3.1.

Figure 3.1 shows the daily average consumption for all the houses and the daily consumption of two houses - one that follows the trend at 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 it was snowing in Denmark at that time [Tilføj reference på det her](#).

The average of the attributes from the house data is examined through a scatterplot in order to find possible correlations. Figure 3.2 clearly shows that the consumption is close to 0 in the summer period. [Pairs af gennemsnitlig house data - vi ser en masse sammenhænge mellem de forskellige attributer. Vi kan se at CoolingDegree skal være over 25, før at varmekonsumet stiger. CoolingDegree begynder at stige et stykke tid før flowet stiger, hvilket hænger godt sammen med at når man fx tænder en radiator så stiger CoolingDegree. De efterfølgende radiatorer man tænder øger volumnet.](#)

The figure 3.3 shows the dependencies between the average consumption of the houses and the weather attributes.

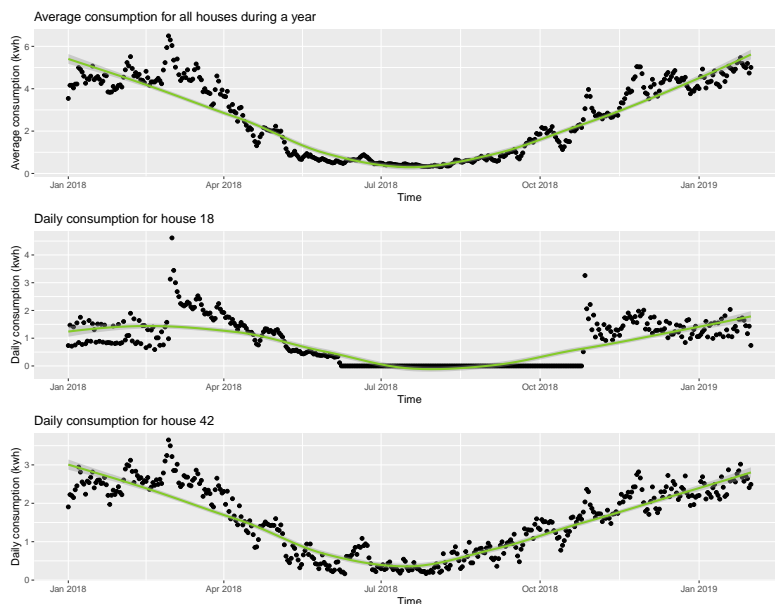


Figure 3.1: Daily consumption during a year (2018). The top plot shows the average consumption for all the houses. The plot in the middle shows an example of a house that follows the trend and the last plot shows a house that deviates from the trend.

It is already known that there is a dependency between the heat consumption and the time of year. During the summer period there is almost no consumption. The 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. *Anton nævnte noget med SunHour og Ultravioletindex.*

Figures 3.2 and 3.3 are used to investigate linear relationships which is desired when modeling. If a linear relation is not *obtained* this could give rise to a transformation on either the dependent or the independent variable. *Jeg synes der mangler lidt her.*

3.1.1 BBR data

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

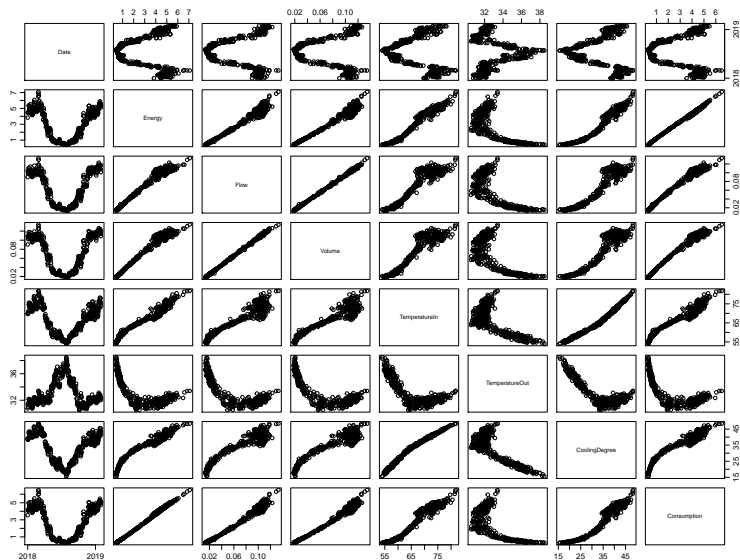


Figure 3.2: .

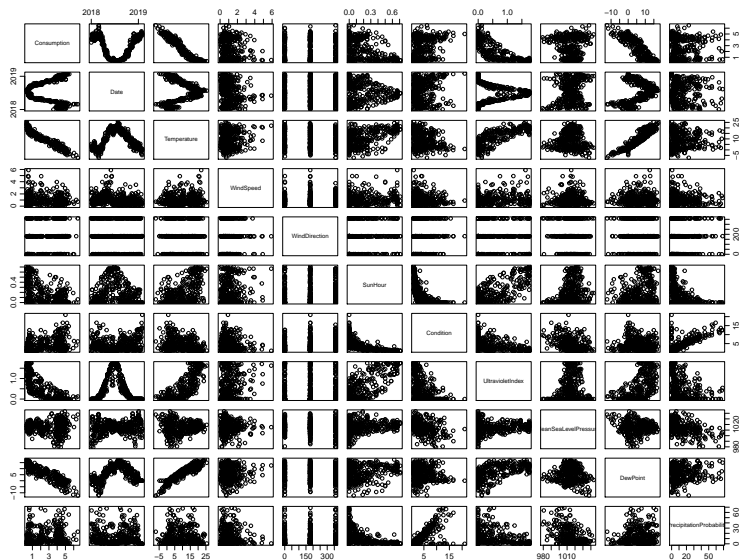


Figure 3.3: .

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 that the outlier is an apartment of 61 m^2 build in 1920.

3.2 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

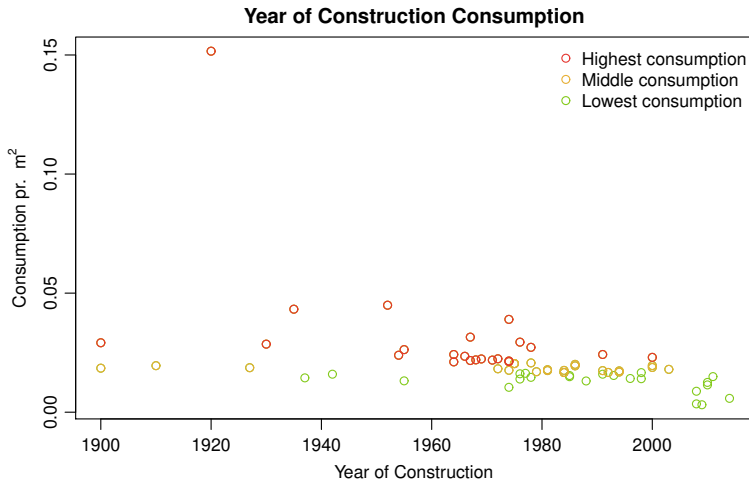


Figure 3.4: Plot showing the year of construction and the average consumption pr. m^2 for each house. It is clearly seen that there is a tendency that the later a house is built or reconstructed, the better is the insulation of the house.

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.

3.2.1 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 3.5 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.

3.2.2 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 3.6. On the left the data points are plotted with standard deviations on the y-axis. The red line highlights the two standard deviations. 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.

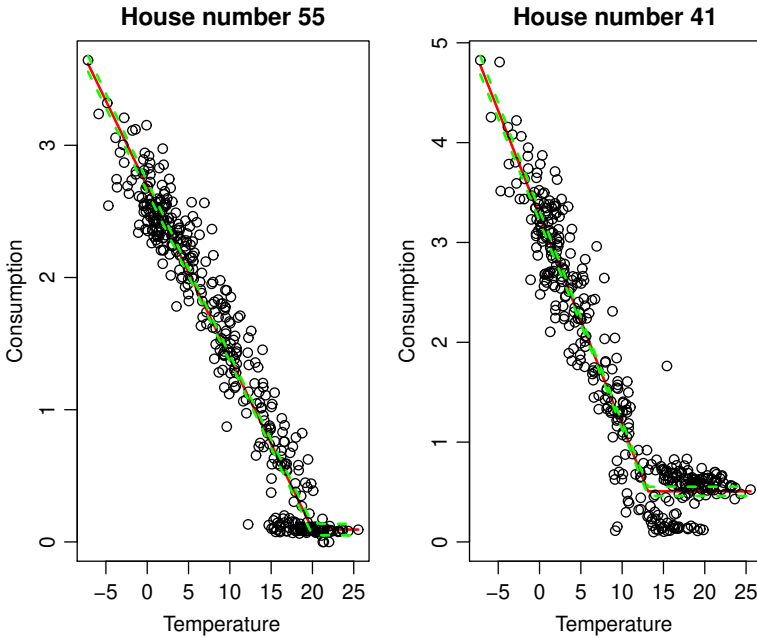


Figure 3.5: Piece-wise optimization of the consumption. The red line is the regression line and the green line is the confidence interval..

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 3.6 to figure 3.5, 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 3.7 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.

3.3 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 3.3 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 3.8 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.

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.

Breakpoint for house number 55

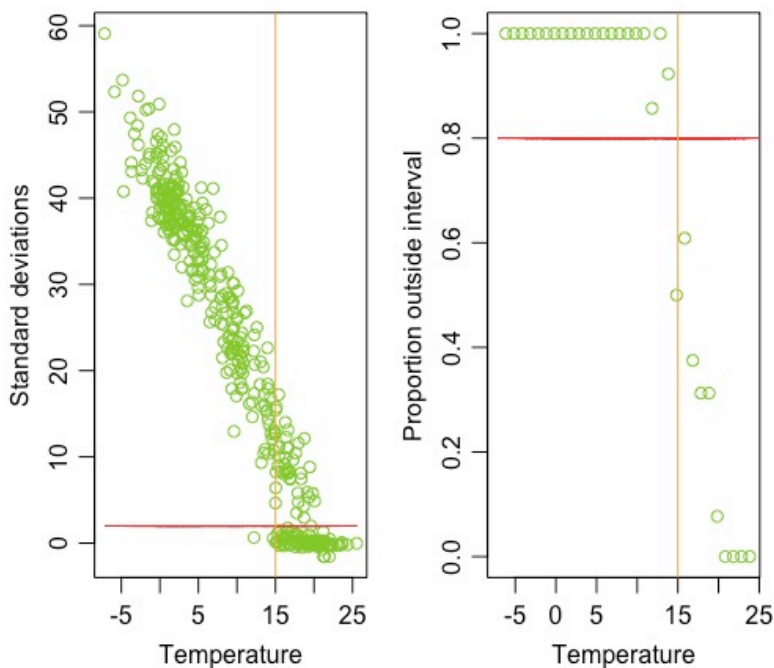


Figure 3.6: An illustration of how the breakpoint is found using segmentation by significant deviations. On the left figure the line illustrates two standard deviations from the high temperature distribution. The right figure shows how many points are outside the two standard deviations. The last point below 80% is the chosen breakpoint.

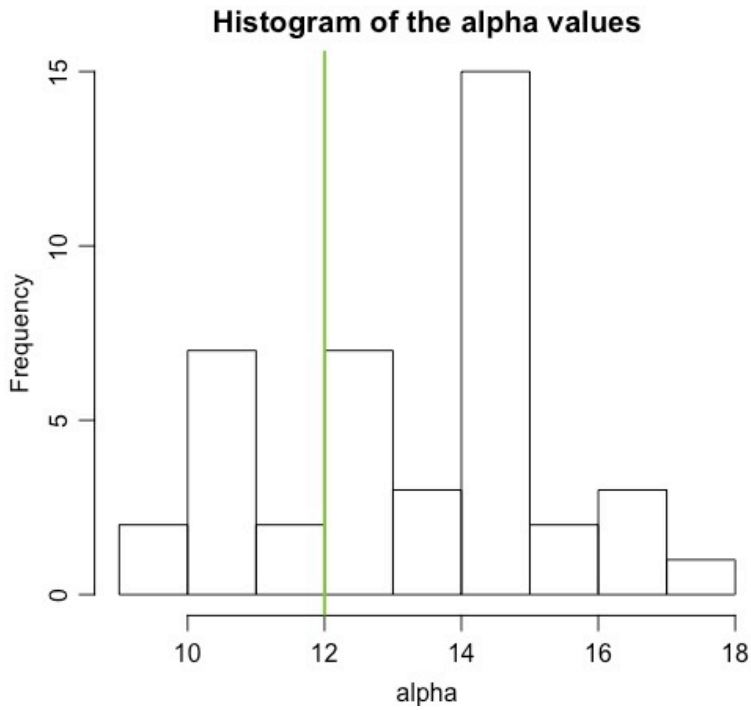


Figure 3.7: A histogram of the alpha values for every house in the third segmentation method. The first quantile is chosen as the overall breakpoint. It is 12 degrees, illustrated by the green line.

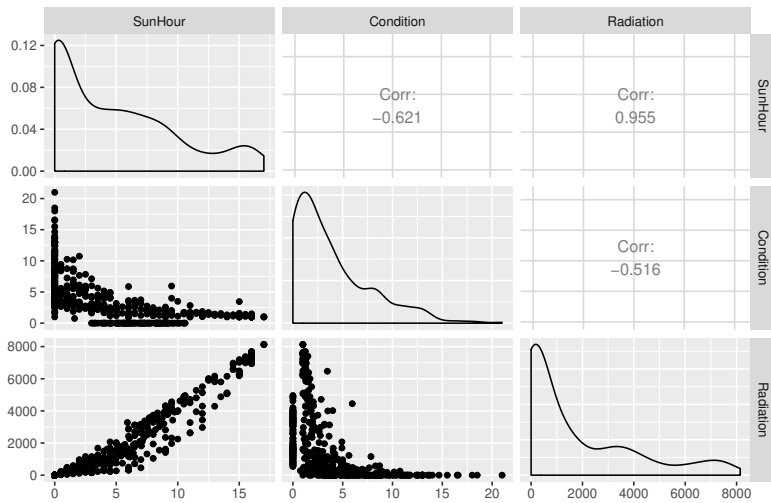


Figure 3.8: Scatterplot showing the correlations between the three attributes Condition, Radiation and SunHour. It is clearly seen that the radiation and the sun hour are highly correlated.

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, \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 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

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

is chosen for estimating the model. The parameters 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 homogeneity

3 Variance should be independent of location

4 Linear relationship between x_j and Y

If these assumptions are not met ...

The fitting of the regression models is carried out by using the method stepwise regression **Bruger vi adjusted R-squared?** 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.

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**. 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.

Opskriv hvilken simpel lineær regressionsmodel, vi bruger.

4.2.1 Results

Overordnet kan den simple lineær 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.

4.3 Multiple linear regression model

Da alle attributter er gennemsnitlige for at få dagsværdier, giver condition ikke rigtig mening at have med i modellen. Derfor undlades den for nu.

Vi medtager ikke interactions mellem holiday attributen og de andre attributter, da dette ikke er main focus.

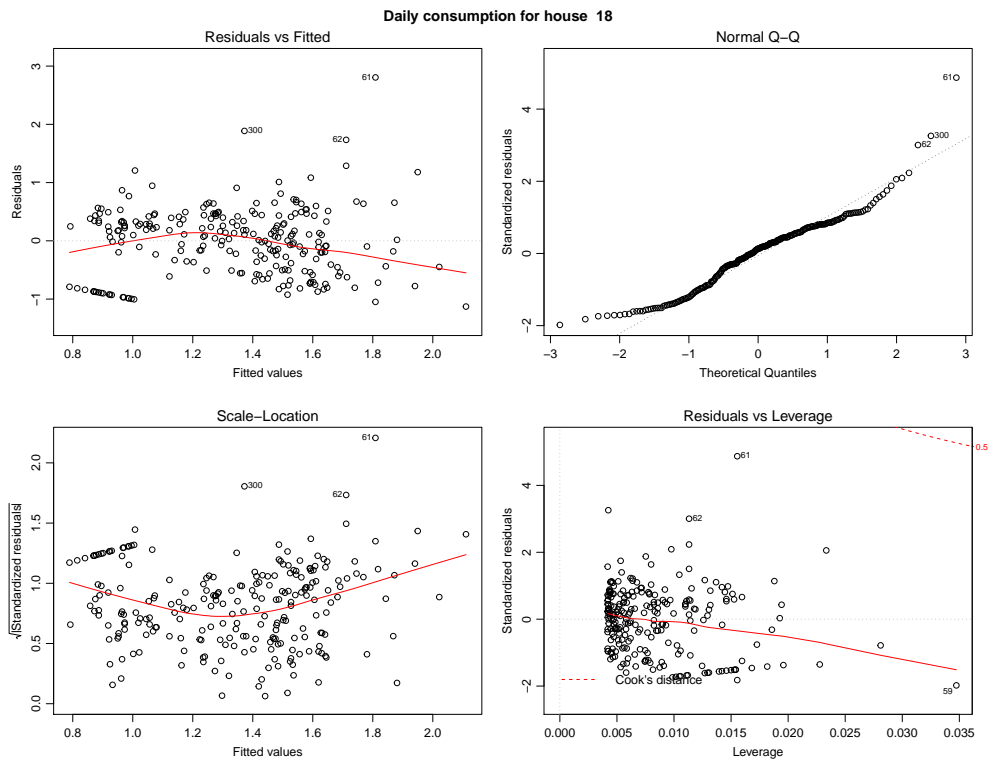


Figure 4.1: Residual plots of house 18 based on the simple linear regression model given in equation (??). The model assumptions of a linear regression model are not fulfilled for this specific house..

4.3.1 Splines

In the multiple regression model, splines will be used to model the wind direction. It does not make much sense 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 Results

4.4 Regression model for comparing houses

Baseret på tabeller over signifikante parametre for både korte og lange huse, kan vi lave en general model for alle huse, der inkluderer temperaturen, splines og radiation.

4.5 Comparison

CHAPTER 5

Vejledningsmøder

5.1 19. februar

5.1.1 Spørgsmål

1. Hvorfor er der nogle af husene, som kun har omkring 3600 observationer, mens andre har 9400? Hvad vil det betyde for os? Hvad kan vi gøre? **Vi skal i sidste ende lave noget der virker på tilgængeligt data. Realistisk problem hæhæ. Vi må godt sige, at vi skal have nok data. En delopgave: hvor mange data skal der til for at kunne sige noget konstruktivt. Ændrer det på konklusionerne? Få denne perspektivering ind på et eller andet tidspunkt.**
2. Må vi fjerne hus 5? Den giver os problemer... **Vi skal bare ændre på datoerne for hus 5 inde i en text editor.**

5.1.2 Noter

- Hvornår er der informationer nok, hvornår er der ikke?
- Når vi laver vores modeller, skal vi lave dem således at mængden af data kan variere. Man laver noget for hvert hus, så man så kan sammenligne et eller andet. Hvad er ens, og hvad er forskelligt for hvert hus?
- Lasse forventer ikke, at vi ender med perfekte modeller. Thank God!
- Brug `as.POSIX` til at lave tiden. Kig på input- og outputtype.
- Der er to måder at lave varmt brugsvarme på - enten varmeveksler eller varmegvandsbeholder. Beholder: hvis temp. i bunden bliver for lav - opvarmningen bliver dermed mere jævn. Pladevarmeveksler: ligesom radiator, fjernvarme igennem radiatoren og brugsvarme i midten eller sådan noget.
- Vi har også sommerdata - kig på varmeforbruget der til at få en idé om hvordan huset opfører sig. Er der et hårdt forbrug mellem kl. 7-8? Maj eller september måned kan vise hvordan deres varmegvandsforbrug er. Er der peaks, eller er det jævnt fordelt?

- Man skal ikke kaste for meget væk.
- Brugsvand er støj, men det ikke tilfældig støj. Det er positivt, så det påvirker estimerne. Noget af det kan vi fjerne, men vi skal se på data hvor der ikke er varme - er der nogle mønstre?
- Hvilken ugedag er bedst til at repræsentere en weekend? Måske lørdage?
- Skal vi kigge på hvordan huset performer, eller skal vi kigge på hvordan huset performer her og nu?
- Hvor stopper vi? Det vigtigste er, at vi laver nogle ting, som vi ved kommer til at virke.
- Teoridelen: det er vigtigere at vi får tydeliggjort hvad den her metode kan.

5.1.3 Hvad skal vi?

- Tjek forskel på ugedage, weekender, helligdage, ferier - hvad gør vi med disse forskelle?
- Få lavet plots.
- Markér underlig opførsel i data i plots.
- Find de normale perioder og så gør noget dér. Alt det andet kigger vi på senere.

5.2 26. februar

5.2 (1) Daily averages of consumption versus temperature differences

5.2 (2) Læse artikler fra Peder

5.2.1 Spørgsmål

1. abline på Q-plot - kan vi optimere den på nogen måde, eller er det okay vi bare vælger en temperatur? Det er meget realistisk, at folk tænder for varmen, når der er under 13 grader udenfor. Vi har brug for en smart måde at optimere på. Vi kan sagtens optimere denne. Vi skal dog lave plottet på døgnværdier i stedet.
2. Hvordan sorterer man rækkerne i et data.frame ud fra en bestemt søjle? Den her er vist fikset.
3. Idéen var at udfylde de punkter vi mangler og så fylde dem ud med NA værdier. Så rækkerne mangler ikke, men de er tomme. Er det en korrekt måde at håndtere dette problem på? Peder siger det giver mening og så tage højde for det derfra. Det giver mening fordi det er samplet meget skarpt. Lav en vektor med de tidspunkter vi gerne vil have og så merge data.frame med vektoren og så keep left, så fylder den ind. Husk én detalje: sommertid og vintertid.
4. Vise plots - er det godt eller skidt?

5.2.2 Noter

- Al data er højst sandsynligt målt i samme tidszone.
- Peders strategi: fortæl den at det er "GMT" eller "UTC" tid.
- Vi laver en model for hvert hus, fordi det skalerer til mange huse. 69 forskellige sæt parametre men det kan godt være samme model. Det er en af de diskussioner vi kommer til at skulle lave.
- Hvad effekten af at bruge forskellige modeller? Der kommer forskellige ting ind, vi kan sammenligne huse, hvor mange data har man? Hvilken betydning har det?
- Vi tager ét hus - hvad kan vi gøre med en månedsdata og så laver vi et rulende vindue. Hvilke estimerer et eller andet. Er det faktisk robust det vi har gang i? Plot parameter estimererne gør nok noget henover året. Hvad gør konfidensintervallerne?

- Brug subset af data til at estimere med, forskellige længder, overlap osv. Det er en god måde at lave robuste modeller på. Kan man fx overhovedet se at folks juleferier har betydning?
- I første omgang er det at kigge på hvordan husene opfører sig. Vi starter med at bygge ting op, som vi ved virker. Forudsigelse og undersøgelse af robusthed.
- Tag en eller to dages gennemsnit på varmesæsonen og så tage parametrene og plot dem for den model eller så noget.
- Normaliseret pr. kvadratmeter i huset.
- Når vi ikke har indetemperaturen, er vi nødt til at have mu med. Hvis man bruger en masse el, så påvirker det også estimatet af indetemperaturen.
- Plot af hele data, pairs plot, vinterperioder - plot for alle sammen. Fx et hus der opfører sig helt gakket.
- Det plot med knækket vi har - vi skal tage det over hele dagen og ikke baseret på timerne. Man kan også lave en model, hvor man tager autokorrelationen med og så bruger weighted least squares.
- **aggregate** fra Peder.
- Hvis man laver modelreduktion - hvad er altid med? Brug **step**-funktionen til at reducere. Er weekdays signifikant?
- Helsingørdata: Nogenlunde samme modeller som for Aalborg. Vi har el og vand og vil lave dagsværdier, hvad kan vi bruge det til? Hvad hvis vi ikke bruger el og vand, hvad hvis vi gør? Får vi merværdi.

5.2.3 Hvad skal vi lave?

- Lave vektor og merge med data.frame
- Lave projektplan: kursusbeskrivelse og læringsmål ligesom for et kursus. Brug teksten fra mda'en eller sådan noget. 10 linjer eller noget. Hvad er læringsmål, som vi skal måles på?
- Hvad er egentlig det nye vi laver/undersøger?

5.3 5. marts

5.3 (3) få styr på lorte parskip-pakken

5.3 (4) Få aksefis af Grønning eller Maika

5.3.1 Spørgsmål

- Vi vil gerne aflevere den 20. juni, så vi kan fremlægge senest den 27. juni.
- Hvad er det helt præcist volume er? Umiddelbart ville vi mene det var det samme som flow, men værdierne er forskellige og flow er pr. time mens volume ikke er.
- Vil det have nogen betydning senere hen, hvis vi har fjernet EndDateTime nu?
- Hvad skal vi lægge i korrelationerne? Fortæl os det.

5.3.2 Hvad skal vi have lavet?

- Læse notefis grundigt.
- Kigge på fejl i optim-funktion (Anton).
- Få styr på ggplot.

5.3.3 Noter

5.3.3.1 Til projektplan:

- SEAS-NVE vil gerne vide hvad for nogle forskellige ting man kan lave med de data.
- Sammenligne huse - hvad kan vi sammenligne, hvad kan vi ikke sammenligne?
- Hvad er det de godt vil kommunikere til beboerne på den lange bane? Hvor godt performer beboerens hus.
- Relativt sammenlignelige huse - hvordan er deres temperaturafhængighed?
- Hvad der er signifikant ligger bagved.
- Forecasts: hvad er der af døgnvariationer? Er der specifikke mønstre? Der er nogen der har en brændeovn - kan vi se om den er tændt? varmegvandsforbrug - hvordan er det fordelt på døgnet? Har man natsænkning/dagssænkning?
- Til tidsrækkedelen: det skal være en dynamisk model. Der er en overførsels-funktion, der kan være svær at identificere.

- Døgnvariation som ikke kobler til dynamikken og heller ikke temperaturen, DET er det spændende, siger Lasse!
- Kør to ting parallelt, når vi er tre.
- IKKE START MED ET HUS MED BRÆNDEOVN!
- Vi skal nok lege manuelt med et par forskellige huse og så tage den derfra.
- Fix punkt 2 og så kommentarer omkring hvordan det skal kommunikeres.

5.3.3.2 Andet:

- Optimeringen af α ligger udenfor - i optim.
- Se Lasses tegning - lav en funktion som hedder piecewise
- Hvis der er huller i data: lave rå-gennemsnit og så bagefter se hvornår et eller andet.
- Kig på residualer fra en model. Lav plot på dagsværdier og håb på ting ser mere robust ud.
- Varians inhomogenitet????? Plot residualerne mod de forskellige variable.
- Variablen flow er det flow der er her og nu, når målingen laves. Volumen er det flow der er løbet igennem siden sidste måling. Lasse forventer, at volumen er det robuste tal.
- Flow og temperature kommer fra EndDateTime, så det er det vi skal bruge.
- Til Anders: Noget med volumen og de temperaturer vi har her er de fra øjeblikket eller er de for den forgående time.
- Energidata: kan vi gange volumen og temp.forskellen sammen og få noget der ligner energidata. FØR VI SKRIVER TIL ANDERS.
- Energi er vist ikke electricity consumption.
- Vi skal sige til Anders, at alle huses data ser sådan her ud. Men inden skal vi lige tjekke forholdet mellem volumen og coolingdegree.
- Energi burde være $4.186 \text{ blabla} * \text{temp forskel} * \text{flow}$
- Kig på dagsværdier nu.

5.4 12. marts

Bestemmelse af max temp. Hvor stabil er hældningen? Jo flere data der er med i modellen, jo bedre er estimatet. Men når de dårlige værdier inkluderes bliver det dårligere igen. Kig på diagnostic plot af de fittede huse. Lav en linear model med backward selection. Fuld regressionsmodel. Se på hvilke variable som er vigtige. Sammenlign hældninger. Kig på hvad de forskellige variable gør. Hvilke variable skal med i modellen og hvilke skal bruges til at estimere parametre. Man tager tit varmekonsum pr. kvadratmeter. Enten ved at dividere forbruget eller hved at scalere parametre. Når vi snakker tidsrække modeller skal vi se på ACF.

5.5 19. marts

5.5.1 Spørgsmål

- Fortælle om vores bud på at bestemme overgangsperioden for fjernvarme.

5.5.2 Noter

- Lav heat maps til exploratory
- I forhold til at bestemme α , så se billede af Mikkels figur. Når den går over for good i standard afvigelser, så er det det punkt et eller andet. De har valgt 3. Lav noget qsum eller sådan noget. Når standard afvigelserne skal plottes skal de transformeres med $\log + 1$. Bruger +20 grader som træningsdata. Vælger et robust estimat af middelværdi og standard afvigelse.
- Denne måde går oftest rigtig godt, men det kan gå dårligt, hvis det er 5 grader, og de så skruer ned for varmen. Også offentlige bygninger med weekendsænkning.
- Vælge et andet sigma niveau og så gøre det i 1-grads intervaller og se hvor mange der ligger under. Finn og Anton har styr på det.
- Der er en prior på 13 grader (det er for at være robuste), og hvis vi så har mere information, så gør vi noget andet.
- Det en fordeling fra 10 til 16 grader, hvor de fleste ligger på 13-14 grader.
- Lave et kriterie (threshold) for hvor mange observationer der skal ligge i 20+.
- Man burde måske bruge første kvartil for de huse hvor vi ikke har nok data, da man begår færre fejl ved at vælge en for lav værdi.
- Hvornår ved man at man kan stole på metoden? Lasse plejer at sige 12 grader.
- Lave et afsnit om hvordan man finder breakpointet - hvilket kapitel?
- Man får et større estimat af variansen, når der er ferier, weekend osv. Det gode ville være at lave en multiple linear model og tilføje parametre som påvirker forbruget et eller andet. Hvor mange af husene har sådanne effekter, og hvornår har de ikke de effekter.
- <http://skoleferie-dk.dk/skoleferie-aalborg/>
- Dag til dag variationen er mindre, fordi en af energikilderne forsvinder.

- Vi laver en lineær regressionsmodel, og vi kigger på variansen som funktion af periode. Hvis modellen er stort set perfekt til at forudsige, når der ikke bliver brugt varmt vand, så bør man pille disse perioder ud af modellen. De skal ikke smides ud i første omgang, men farv dem og se om de ligger anderledes end de andre. Tjek perioderne signifikans først.
- Kig på autokorrelationen i lag 1 for nogle huse. Er der væsentlige korrelation - ja eller nej? Sikre sig at man kigger på de rigtige lags. Det er stadig dagsværdier. Skal vi korregere for det? Lave weighted least squares i stedet for bare least squares.
- cuesum senere hen.

5.5.3 Hvad skal vi lave?

- Lav et plot for alle autokorrelationer for husene.
- Få lavet en standard måde vi plotter data for alle husene på. Split husene op alt efter hvor mange observationer der er.

5.6 26. marts

5.6.1 Noter

- På time niveau når du går længere ind i appen, har de en simpel døgnkurve model. Noget med en døgnmodel med simple gennemsnit.
- akkumuleret, cumulativt plot, hvor man summerer op i energi. På dagsværdier gør det ikke så stor forskel.
- Kig på hvor mange af de interpolerede værdier der rent faktisk er. Hvis der er mere end 2 decimaler, så er det interpoleret. Bare lav modulus.
- Hvordan er hældningerne i forhold til areal, bygningsår, hvilken type hus.
- Vi kan sagtens smide temp. variabel ind i ggplots. Loess laver en trekant.
- Smid temperaturen ind også, fordi det er den vi forventer der har mest indflydelse. Lav den kun som funktion af temperaturen og så plot residualer over datoerne.
- Hvis man skal være pragmatisk, så skal vi vælge et fast tal. For at få variansen ned skal vi holde os skarpt under 15 og over 10. 12-13 grader.
- Det der kendetegner at der er varme i forhold til brugsvarme, er at der er flow hele vejen, så vi skal kigge på flow.
- Hvordan kan vi blande solen ind? Sunhour og condition. Sunhour kan kun have en værdi forskellig fra nul, når condition er 0,1,2.
- **Multiple linear regression model:**

5.6.2 Hvad skal vi lave?

- LAV SIMPLE LINEAR REGRESSION MODELS

5.7 5. april

5.7.1 Spørgsmål

- Hvordan skal vi tolke det, når temperaturen ikke er signifikant?
- Skal vi forsøge at trække vores parametre ud af vores modeller?
- Vise vores model - er det rigtigt? HJÆLP TIL MODEL! Lasse fikser
- Skal vi overveje at transformere vores data, når nu vi har vores QQ plots?
- Anders' mail. Hvad synes du?

5.7.2 Noter

Vindhastighed. Plot punkterne med vinkel * hastighed. Lav et gennemsnit i et rektangulært koordinatsystem. Tag gennemsnittet. Brug den retning og den vinkel som den nye vindhastighed og retning (i dagsdata).

Splines: pbs(rad). En spline som function af vinklen. Hvis man brug deg 2 splines, har man først 3 parametre + 1 pr. knudepunkt man rammer. For 4 punkter er der 6 degrees of freedom.

- Parametre i LM: sol, wind, temperatur, wind direction.
- Wind direction regnes om til rektangulært form og se hvordan de ligger polært.
- Man kan bruge splines til at gøre wind direction lineært - cyklusspline.
- Lav parentes og plus attributer og så i anden eller tredje i multiple.
- Vind: Den energi der skal bruges bliver højere jo mere udskiftning der er. $WindSpeed * Temperaturforskellen * alpha$. Meget naturligt at der er interaction mellem WindSpeed og temp. forskel. Der kan være et lineært led på vindhastigheden.
- Når vinden kommer fra en bestemt retning, så er der mere varme-/energitab, hvis der er vinduer der.
- Transformerig af WindDirection: plot punkterne i et rektangulært koordinatssystem, hvor der er ganget med vindhastigheden. Man skal måske lave gennemsnittet af vindhastigheden for sig. Man skal vægte med vindretningen - det er mere robust. Find retningen vægtet med hastighed.
- Splines: periodiske splines (pbs) i R. Hvilken orden skal splinen have? Lave dem som et antal veldefinerede knudepunkter. Koster det mere energi, når det blæser for en bestemt retning?

- Definere basisfunktioner rundt. Det største bidrag er, når vinden kommer fra den her retning.
- Kigge på om der er nogle huse der har nogle sider hvor de er meget følsomme.
- Andenordens spline med 4 knudepunkter, tænker Lasse. 6 frihedsgrader og 6 parametre.
- Citat Lasse: "Prøv at lave et matplot af Lasse"

5.8 8. april

5.8.1 Noter

- Pakken der ikke kommer ind, bliver ikke burgt. Wtf.
- Hint fra Lasse: et eksempel hvor der bliver lavet en måling lidt før klokken hel - der er en time, hvor der ikke er målinger i. Interpolationen til det sidste punkt, er der vist en fejl. **Tjek det her!**
- I stedet for at flytte punkter fremad, så se hvad det rigtige data er og så brug det eller noget.
- Vi vil gerne komme med et forslag til en lille ændring.
- Vores splines ligger forskudt med 5 grader, og så går de i 0 +/- 90.
- Hver basisspline er kun forskellig fra 0 i et interval på 180. 270 grader support, men 180 passer måske bedre.
- Lasse foreslår, at vi laver de 6/8 WindDirection i en parentes og så ganger med temperaturen. $\bar{T} * (w_1 + w_2 + \dots + w_8) * WindSpeed + .$
- Designmatricen skal være singulær blabla...
- Der er nogle fysiske rammer der gør, at vi ikke skal transformere. Men man kan diskutere om støjen skal have samme varians.
- adjusted.r.squared
- Vi skal måske tilføje hvilken dag på ugen det er??
- Test af varians homogenitet som funktion af fitted values - levene.test
- Vi er kede af det, når vi ser at vi burde lave weighted least squares i stedet for. Men vi skal gøre det alligevel..
- Hvis korrelationen er for stor, kommer det til at gøre alting lidt mere usikkert.
- For at residualerne giver mening, skal man hen på de rigtige dage.
- Brændeovn, brændeovn, brændeovn.
- Hvis der sker væsentlige forbrugsændringer, hvad sker der så med data? Kan vi så bruge data?
- Når man skal sammenligne dagsværdier, så er det en god idé at pille temperaturafhængigheden for et døgn ud.

- For hvert klokkeslag der kan man tegne hvordan tætheden ser ud ved bare at finde nogle quantiles.
- $Vreg = (t_t, t_{t-1}, t_{t-2}, \dots, S_t, S_{t-1})$. Bruge ARIMAX, bruge ARIMA til at fitte blabla.
- Argument i lm der hedder weight - hvor meget skiller en dag sig ud og så give den mindre vægt eller smide den ud.

5.9 23. april

5.9.1 Spørgsmål

- Hvorfor sidste spline altid bliver NA? $Temp * I(WindSpeed * Splinebasis) = Temp + I(\dots) + T : I(\dots)$. Ved at gange vindhastighed på fås noget med ortogonal.
- Vi vil gerne have retningerne skal være toppene i spline plot, så ikke knuderne. Ved at lægge 45 grader til passer det ikke med boundary knot, så den tager ikke højde for dem der ligger udenfor de 360 grader. **Løsning 1: Man kan flytte det hele. Løsning 2: Lasse har fikset :-).**
- Timeværdier:

5.9.2 Noter

- $\beta(WindDirection)$ bestemmes ved at tage koefficienten estimeret for spline 1 ganget med basisspline 1 + den næste osv.
Så får vi en kurve for $\beta(WindDirection)$. Så kan man kigge på, hvordan den ser ud ved 90 grader agtigt.
- Når vi skal sammenligne to huse, skal vi så bruge den samme model på de to huse eller skal vi et eller andet. Det er den fulde model, som IKKE er reduceret vi skal kigge på.
- Når vi så reducerer modellen ser vi hvad der er vigtigt for lige netop ét hus.
- Køre store model, reducer den og se hvilke parametre vi er nødt til at have med generelt set. Den nye model vi så laver ud fra de vigtige ting, bruger vi så til at sammenligne huse, som er bygget i samme år og samme areal osv.
- For kunden kunne det være interessant at se i hvilket interval af vinkler der påvirker huset, og hvor skal man så isolere bedre.
- En stor udfordring ved timedata: Kan vi bruge timedata til at sige noget om dagsværdi, hvornår er der slukket for varmen?
- Hvis der kun er varmt vandsforbrug, så vil der være nogle andre mønstre, som er on and off. Kig hen over en sommerperiode for at undersøge karakteristikken for varmt vandsforbruget - kan vi se det samme, når vi er inde i opvarmingsperioden.

5.9.3 Hvad skal vi lave?

- Vi skal tage hver `WindDirection*Windspeed` med som én variabel, for alle de forskellige retninger vi vælger. For så fjerner den de retninger der er ikke-signifikante.
- Se Bros besked på Facebook for resten.
- Tilføj farver og rigtige koefficienter i den der 'vinkelcirkel'.
- Se hvad der sker, hvis vi har 4 eller 8 splines med.
- Fikse spline plot, hvor x-aksen er de rigtige retninger.

5.10 30. april

5.10.1 Inden møde

- Vi skal have vores data delt op, så vi for et hus har alle tidspunkter på samme række.

5.10.2 Spørgsmål

- Hvor finder vi konstanterne til at vægte de forskellige vindretninger? **Vi skal gange estimerne fra modellen på, men det giver problemer når fortegnet af estimerne er negativ.**
- Hvilken måde skal vi modellere over data der har forskellige længder? Der er nogle der ikke har winter og spring med? **Tror vi fikser denne selv.**

5.10.3 Noter

- Hvilken model skal man bruge som den reduceret model? Lave en tabel med signifikans af reduceret model (inkluder fortegn på estimerne foran stjerne(r)). Alt der har mere end 1 stjerne er signifikant uanset hvad. Så man vil tage alle med, hvor der indgår minimum en stjerne mere end 1 gang.
- Vi vil gerne påvise, at huse er følsomme overfor forskellige retninger, så der skal hele splinen med i den reduceret model.
- Er der kollinearitet mellem ferierne? De burde være uafhængige. Solen og temperatur i forhold til weekender? Nope det er der ikke.
- Det betyder noget for usikkerheder: ingen af ferierne er signifikante, og vi tilføjer de fem parametre, så er forskellen next to nothing. Hvis man de er signifikante, og vi ikke tager dem med, så gør det usikkerheden for de medtagne parametre større.
- Hvor meget ændrer estimatet sig for sigma for hele modellen?
- For hver enkelt retning skal man lave en forudsigelse for de parametre man gerne vil forudsige med. Brug **predict** og lav et prediction interval eller et confidence interval. Vælg median for SolarRadiation.
- Lav en dataframe med alle retninger hvor du har fast temperatur og SolarRadiation hvor man predicer med det. Ændre WindDirection til Splinebasis. Vælg vinden til en realistisk vindhastighed.
- Det er interessant at kigge på forskellige vindhastigheder og forskellige temperaturer.

- Polygon funktion, brug origo
- Sammenligning af huse: 70 huse for et halvt år - er der forskel på at tage det samme hus og modellere det på et helt eller et halvt år. Er estimatorne robuste?
- Lasse: "Naturlig forklaring på negative estimator: en usikkerhed kan være at et vindue ikke er helt tæt, så presser vinden vinduet tættere." Hvis der er en utæthed der bliver mindre, som betyder mere end afkølingen, så et eller andet..

5.10.4 Hvad skal vi lave?

- Vi skal have vores data delt op, så vi for et hus har alle tidspunkter på samme række, så vi kan sammenligne dem.
- Vi skal nok dele data op i korte og lange huse.

5.11 7. maj

funktion med prediction af middelvarmen. Input temp og vind forhold + sol. Predict. Confidence interval eller prediction interval.

vind = 1 giver varme pr. vindhastighed. Men confidence interval vil du gerne kun have usikkerheden på det estimat.

I en specifik retning er der også en effekt af vindretning og temperatur.

Predict giver et godt samlet overblik, men hvis man vil se på specifikke afhængigheder skal man sætte input fx temp=12 grader.

$Y = \sigma x$

$V[y] = x^T V[\sigma] x$

Plot: Hvordan afhænger hver spline af temperaturen for w=1.

simulering pakke: mess

