

 **DTU Compute**  
Department of Applied Mathematics and Computer Science

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

In cooperation with SEAS-NVE

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

Kongens Lyngby 2019



**DTU Compute**

**Department of Applied Mathematics and Computer Science**

**Technical University of Denmark**

Matematiktorvet

Building 303B

2800 Kongens Lyngby, Denmark

Phone +45 4525 3031

[compute@compute.dtu.dk](mailto:compute@compute.dtu.dk)

[www.compute.dtu.dk](http://www.compute.dtu.dk)

# 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, April 12, 2019

A handwritten signature in black ink, appearing to read 'Anton Stockmarr', with a long vertical stroke on the left and several loops on the right.

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

---







<b>Abstract</b>	<b>i</b>
<b>Preface</b>	<b>iii</b>
<b>Acknowledgements</b>	<b>v</b>
<b>Contents</b>	<b>vii</b>
<b>Todo list</b>	<b>ix</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Torquent Arcu . . . . .	2
1.2 Luctus . . . . .	3
1.3 Sollicitudin vestibulum . . . . .	6
<b>2 Data</b>	<b>7</b>
2.1 Original data . . . . .	7
2.2 Cleaning and preparation . . . . .	8
<b>3 Exploratory Analysis</b>	<b>11</b>
3.1 Data segmentation . . . . .	13
3.2 BBR data . . . . .	15
3.3 Multicollinearity . . . . .	15
<b>4 Statistical models</b>	<b>17</b>
4.1 Linear regression . . . . .	17
4.2 Simple linear regression model . . . . .	18
4.3 Multiple linear regression model . . . . .	18
<b>5 Vejledningsmøder</b>	<b>19</b>
5.1 19. februar . . . . .	19
5.2 26. februar . . . . .	21
5.3 5. marts . . . . .	23
5.4 12. marts . . . . .	25
5.5 19. marts . . . . .	26

---

5.6	26. marts . . . . .	28
5.7	5. april . . . . .	29
5.8	8. april . . . . .	31

# Todo list

---

	1.0 (1) Make a cake . . . . .	1
	1.0 (2) Do it now . . . . .	1
	Figure: 1.0 (3) This is some text that is with the todo and in the figure . . . .	2
	Figure: 1.1 (4) This is some text that is with the todo and in the figure . . . .	4
	Figure: 1.1 (5) This is some text that is with the todo and in the figure . . . .	4
	Figure: 1.2 (6) This is some text that is with the todo and in the figure . . . .	5
	5.2 (7) Daily averages of consumption versus temperature differences . . . .	21
	5.2 (8) Læse artikler fra Peder . . . . .	21
	5.3 (9) få styr på lorte parskip-pakken . . . . .	23
	5.3 (10) Få aksefis af Grønning eller Maika . . . . .	23



# CHAPTER 1

# Introduction

- Upright shape
- *Italic shape*
- *Slanted shape*
- SMALL CAPS SHAPE
- Medium series
- **Bold sereies**
- Roman family
- Sans serif family
- Typewriter family

I love to write special characters like øæå indside my TeX document. Also á, à, ü, û, ë, ê, î, ï could be nice. So waht about the “ ” chracter. What about ° é ® † ¥ ù | œ ‘ @ ö ä ¬ ‹ « © f ß º Ω ... § √ ñ μ , · ¡ ¢ £ ∞ ™ [ ] ± ’.

Some dashes - - —, and the latex form - - —

$$x = \mathbf{x}, \mathbf{x}, x, x_{1_{23_4}^{2^3^4}} \cdot \textit{hello} * \textit{hello worldmy world\grave{u}third worldt}$$

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 [adams1980hitchhiker].

Mauris id quam non magna fermentum malesuada id mattis lorem. In a dapibus neque. Etiam lacus dui, malesuada ac eleifend imperdiet, imperdiet ut ipsum. Vestibulum id ultricies est. Phasellus augue mauris, semper a luctus vel, faucibus in risus. Fusce commodo augue quis elit sagittis non viverra turpis bibendum. Nunc placerat sem non sapien malesuada malesuada ullamcorper orci luctus [adams1980hitchhiker]. Morbi pharetra ligula integer mollis mi nec neque ultrices vitae volutpat leo ullamcorper. In at tellus magna. Curabitur quis posuere purus. Cum sociis natoque penatibus

1.0 (1) Make a cake

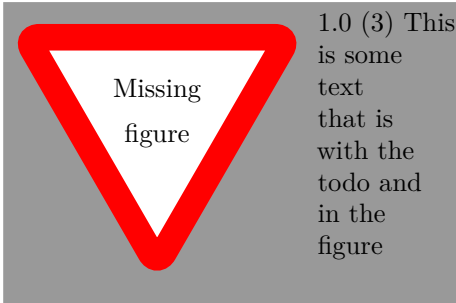
1.0 (2) Do  
it now

et magnis dis parturient montes, nascetur ridiculus mus. Suspendisse tristique placerat feugiat. Aliquam vitae est at enim auctor ultrices eleifend a urna. Donec non tincidunt felis. Maecenas at suscipit orci. See Figure 11.

Fusce id suscipit sem. Aliquam venenatis nibh nec nisl luctus vel consectetur neque dapibus. Nulla feugiat egestas turpis, ac viverra eros cursus sit amet. Cras tincidunt felis vel tellus ultrices condimentum. Quisque vehicula, arcu vitae interdum dignissim, purus tortor cursus libero, sit amet accumsan quam magna in neque. Phasellus luctus leo odio. Aliquam ultricies, arcu quis tempor rhoncus, tellus nisl tempus justo, condimentum tempor erat odio ac purus. Integer quis ipsum felis. Aliquam volutpat, leo ac consequat egestas, lectus lacus adipiscing quam, id iaculis dolor quam in erat. Phasellus tempor interdum arcu quis vestibulum. Pellentesque sit amet augue purus. See Table 11.

## 1.1 Torquent Arcu

Curabitur condimentum suscipit arcu, sit amet convallis urna pellentesque ac. Quisque fringilla tincidunt risus nec accumsan. Curabitur vel sagittis ante. Integer eget placerat leo. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Vestibulum quis risus in nulla fermentum pellentesque dictum et erat. Nulla vel pretium nunc. Integer tortor lorem, suscipit sit amet ultricies non,



**Figure 11:** This is my special figure. Aliquam ultricies, arcu quis tempor rhoncus, tellus nisl tempus justo, condimentum tempor erat odio ac purus. Integer quis ipsum felis. Aliquam volutpat, leo ac consequat egestas, lectus lacus adipiscing quam, id iaculis dolor quam in erat. Phasellus tempor interdum arcu quis vestibulum.

h		h	h
e		e	e

**Table 11:** This is a caption to the table.

porta at metus. Sed pharetra, ante facilis interdum porta, mi dolor fringilla quam, ac porttitor urna dolor quis massa. Proin viverra semper tincidunt. Vivamus pulvinar pharetra condimentum. Pellentesque rutrum mollis tellus ac scelerisque.

### 1.1.1 Vestibulum

Mauris luctus sollicitudin vestibulum. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos Figure 12(b). Duis eu nisl nec turpis porttitor bibendum eget sed orci. Aliquam consequat lorem a dui viverra porta facilis augue rutrum. Cras luctus tellus in lectus egestas eu consequat magna cursus. Aenean aliquam neque a nibh elementum ornare. Integer eleifend imperdiet commodo. Morbi auctor, dui vel laoreet congue, purus est accumsan augue, sit amet feugiat neque nisl vel lorem. Curabitur ante sem, lacinia id adipiscing quis, viverra tristique nulla. Pellentesque ullamcorper pellentesque metus varius facilis. Cras ac dui id odio tempor scelerisque. Curabitur a egestas risus. Pellentesque quis velit in sapien accumsan auctor. Phasellus aliquam, sapien eget lobortis volutpat, libero metus porttitor nisl, sed hendrerit urna dolor nec mi. See Listing 1.1.

```
1 # This is a comment
2 import easy
3 str = "I am a string"
4 str2 = "Now i have an awesome string with ` ' ' ` which are not TeX'ed"
5 str3 = "What about awesome unicode characters? Like ", , ", Ω, ç. \" This"
6 def fib(n):
7     if n == 0:
8         return 0
9     elif n == 1:
10        return 1
11    else:
12        return fib(n-1) + fib(n-2)
13 str4 = "Yes it is possible with 80 charactes. Which this string proves. Wiiii."
14 str5 = "It adjusts according to the spine"
```

Listing 1.1: Fibonacci.

## 1.2 Luctus

Praesent et pellentesque arcu. Phasellus venenatis mi eu lorem convallis et iaculis ante aliquet. Aenean rhoncus placerat metus, vel convallis leo suscipit eu. Integer dapibus venenatis commodo. Cras laoreet faucibus sem nec luctus. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Cras consectetur lacinia dolor at gravida. Phasellus ipsum arcu, vulputate fermentum ultricies eget, tempor eu odio. Aenean accumsan vestibulum risus a mattis. See it on Algorithm 1.

Fusce id suscipit sem. Aliquam venenatis nibh nec nisl luctus vel consectetur neque dapibus. Nulla feugiat egestas turpis, ac viverra eros cursus sit amet. Cras tincidunt



(a) 1 pass



(b) 5 passes

**Figure 12:** loop performance comparison.

felis vel tellus ultrices condimentum. Quisque vehicula, arcu vitae interdum dignissim, purus tortor cursus libero, sit amet accumsan quam magna in neque. Phasellus luctus leo odio. Aliquam ultricies, arcu quis tempor rhoncus, tellus nisl tempus justo, condimentum tempor erat odio ac purus. Integer quis ipsum felis. Aliquam volutpat, leo ac consequat egestas, lectus lacus adipiscing quam, id iaculis dolor quam in erat. Phasellus tempor interdum arcu quis vestibulum. Pellentesque sit amet augue purus. Curabitur condimentum suscipit arcu, sit amet convallis urna pellentesque ac. Quisque fringilla tincidunt risus nec accumsan. Curabitur vel sagittis ante. Integer eget placerat leo. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Vestibulum quis risus in nulla fermentum pellentesque dictum et erat. Nulla vel pretium nunc. Integer tortor lorem, suscipit sit amet ultricies non, porta at metus. Sed pharetra, ante facilisis interdum porta, mi dolor fringilla quam, ac porttitor urna dolor quis massa. Proin viverra semper tincidunt. Vivamus pulvinar pharetra condimentum. Pellentesque rutrum mollis tellus ac scelerisque.



**Algorithm 1** Modified mini-batch  $K$ -means

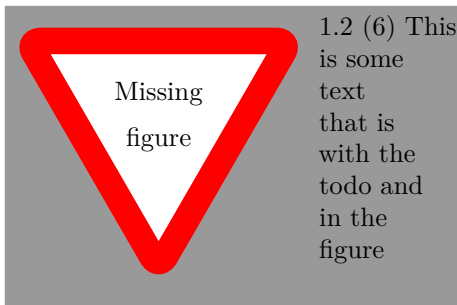
---

```

1: Given:  $K$ , mini-batch size  $B$ , iterations  $T$ , dataset  $X$ , correlation matrix  $P$ .
2: Initialize  $C = \{\mathbf{c}^{(1)}, \mathbf{c}^{(2)}, \dots, \mathbf{c}^{(K)}\}$  with random  $\mathbf{x}$ 'es picked from  $X$ .
3:  $A \leftarrow B \cdot T$  sorted random indexes to  $X$ , denoted  $a_1, a_2, \dots, a_{B \cdot T}$ .
4:  $X' \leftarrow \{\mathbf{x}^{(a_1)}, \mathbf{x}^{(a_2)}, \dots, \mathbf{x}^{(a_{B \cdot T})}\}$  ▷ Cache all points
5: size  $\leftarrow 0$ 
6: for  $i = 1$  to  $T$  do
7:    $M \leftarrow B$  examples picked randomly from  $X'$ 
8:   for  $\mathbf{x} \in M$  do ▷ Assignment step
9:      $\mathbf{d}[\mathbf{x}] \leftarrow f(C, \mathbf{x}, P)$  ▷ Cache closest center
10:  end for
11:  for  $\mathbf{x} \in M$  do ▷ Update step
12:     $\mathbf{c} \leftarrow \mathbf{d}[\mathbf{x}]$  ▷ Get cached center for current  $\mathbf{x}$ 
13:     $\text{size}[\mathbf{c}] \leftarrow \text{size}[\mathbf{c}] + 1$  ▷ Update cluster size
14:     $\eta \leftarrow \frac{1}{\text{size}[\mathbf{c}]}$  ▷ Get learning rate
15:     $\mathbf{c} \leftarrow (1 - \eta)\mathbf{c} + \eta\mathbf{x}$  ▷ Take gradient step
16:  end for
17: end for
18: return  $C$ , size

```

---

**Figure 13:** This is the caption I wrote.

## 1.3 Sollicitudin vestibulum

Mauris luctus sollicitudin vestibulum. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Duis eu nisl nec turpis porttitor bibendum eget sed orci. Aliquam consequat lorem a dui viverra porta facilisis augue rutrum. Cras luctus tellus in lectus egestas eu consequat magna cursus. Aenean aliquam neque a nibh elementum ornare. Integer eleifend imperdiet commodo. Morbi auctor, dui vel laoreet congue, purus est accumsan augue, sit amet feugiat neque nisl vel lorem. Curabitur ante sem, lacinia id adipiscing quis, viverra tristique nulla. Pellentesque ullamcorper pellentesque metus varius facilisis. Cras ac dui id odio tempor scelerisque. Curabitur a egestas risus. Pellentesque quis velit in sapien accumsan auctor. Phasellus aliquam, sapien eget lobortis volutpat, libero metus porttitor nisl, sed hendrerit urna dolor nec mi.

Praesent et pellentesque arcu. Phasellus venenatis mi eu lorem convallis et iaculis ante aliquet. Aenean rhoncus placerat metus, vel convallis leo suscipit eu. Integer dapibus venenatis commodo. Cras laoreet faucibus sem nec luctus. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Cras consecetur lacinia dolor at gravida. Phasellus ipsum arcu, vulputate fermentum ultricies eget, tempor eu odio. Aenean accumsan vestibulum risus a mattis.

```
1 # This is a comment
2 import easy
3 str = "I am a string"
4 str2 = "Now i have an awesome string with ` ' ' ` which are not TeX'ed"
5 str3 = "What about awesome unicode characters? Like ", , ", Ω, ç. \" This"
6 def fib(n):
7     if n == 0:
8         return 0
9     elif n == 1:
10        return 1
11    else:
12        return fib(n-1) + fib(n-2)
13 str4 = "Yes it is possible with 80 charactes. Which this string proves. Wiiii."
14 str5 = "It adjusts according to the spine"
```

Listing 1.2: Fibonacci2.

# CHAPTER 2

## Data

The data is provided by SEAS-NVE in two data sets. The house data consists of 69 .csv-files containing 8 attributes for each house which is 499,499 data points in all. The second data set includes weather data containing 11,845 observations with 11 attributes. **Noget med hvordan data er blevet målt - hvilket udstyr, af hvilken virksomhed osv.** 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 21 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 21:** Attributes from the original house data..

The weather data set consists of the attributes seen in Table 22.

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 22:** 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.

## 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 attribute Holiday, som er defineret som working days, christmas break, winter break, spring break, autumn break

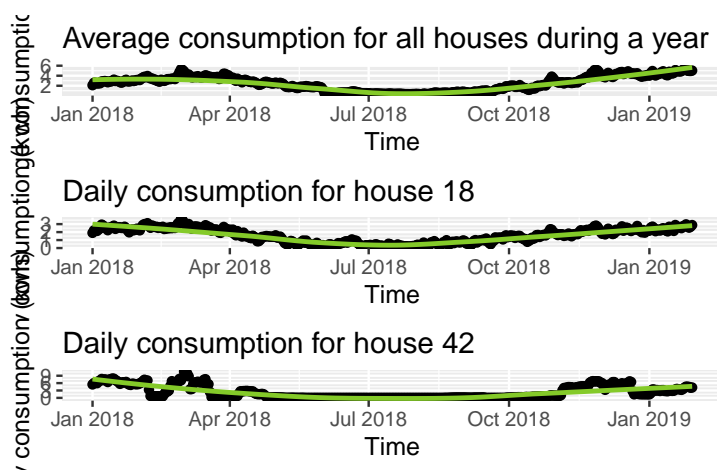


## CHAPTER 3

# 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. **Mangler nok noget lidt mere her.**

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 the following figure.



**Figure 31:** 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.

Figure 31 shows the daily average consumption for all the house 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. **Noget med at vi kun er interesserede i perioden, hvor der er tændt for varmen, så derfor fjerner vi perioden hvor consumption er tæt på 0.** All three plots show some unusual high

data points around April 2018. This can be due to the fact that it was snowing **blabla**.

The average of the attributes from the house data is examined through a scatter-plot in order to find possible correlations. Figure ?? 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 varmemeforbruget 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.**

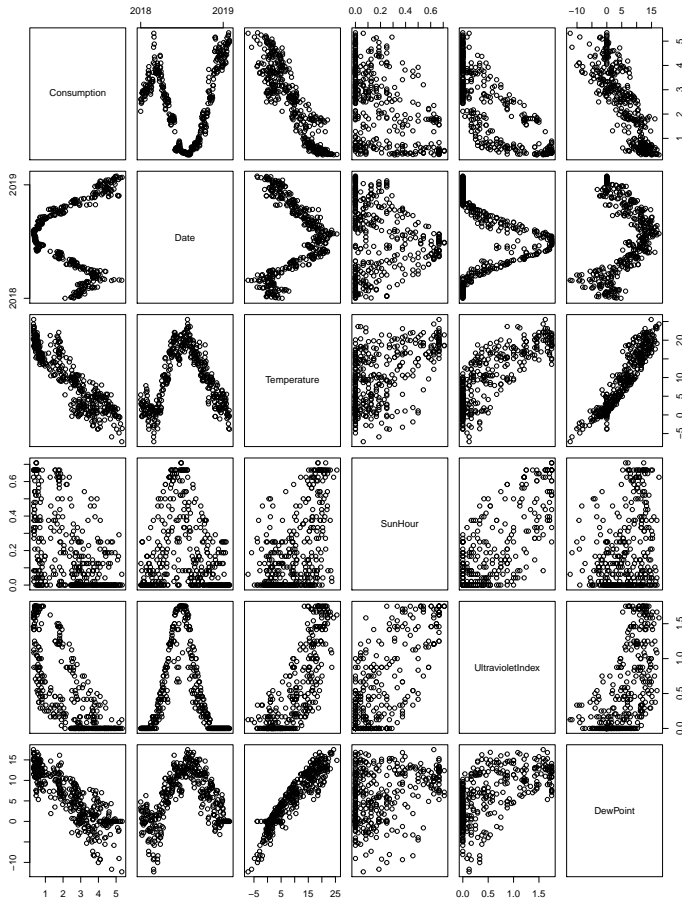


Figure 32: .



The figure (de udvalgte weather pairs) shows the dependencies between the average consumption of the houses and the weather attributes. We already know that there is a dependency between the 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 ?? and 32 are used to investigate linear relationships which is desired when modelling. If a linear relation is not obtained this could give rise to a transformation on either the dependent or the independent variable.

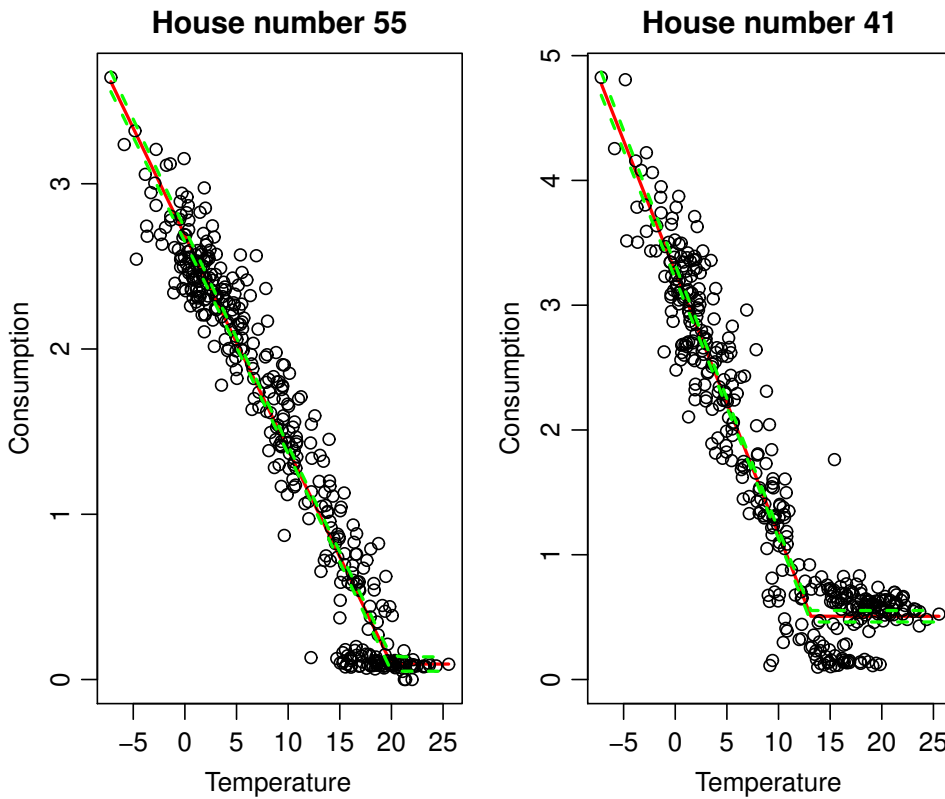
## 3.1 Data segmentation

When looking at the consumption data, it is useful to distinguish between periods where the heat is turned off and periods where it is turned on. The models that will be applied in the following sections should only consider periods with heating turned on, if possible. Because of the tap-water consumption, this difference is not always clear cut. 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. Three approaches will be described below, together with their pros and cons.

### Segmentation by piece-wise optimization

The first approach is to make a linear regression on the data with two segments. A breakpoint  $\alpha$  is found, such that the SSE is as small as possible. The second segment is restricted to being constant. This way the breakpoint illustrates when the consumption goes from being linearly dependent on the temperature, to having a constant value. If that assumption holds, all data points below the breakpoint can be used in the statistical model for analysing heat consumption, and the rest can be discarded as tap water consumption.

Figure 33 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.



**Figure 33:** Piece wise optimization of the consumption. The red line is the regression line and the green line is the confidence interval..

### 3.1.1 Segmentation by investigation slopes

### 3.1.2 Segmentation by significant deviations

### 3.1.3 Bestemmelse af temperatur breakpoint:

- Vi kigger på de huse der har mindst et års data (så vi er sikker på at hele sommeren er med)
- Vi antager at dagene med over 20 grader udenfor, der er der slukket for varmen, og der er dermed kun varmvandsforbruget med, som vi antager er hvid støj.
- For hver grad under 20 ser vi hvor mange procent af consumptionen der ligger indenfor  $\pm 2$  standardafvigelser fra 20+ sættet.

- Den første grad hvor mere end 20% af datapunkterne ligger indenfor intervallet gemmes som det hus' alpha.
- vi tager til sidst 15% "percentile" af alphaerne, og skærer alt data fra vores datasæt som er over 12 grader.

## 3.2 BBR data

## 3.3 Multicollinearity

Multicollinearity occurs when two or more explanatory variables are highly correlated.

In linear regression, multicollinearity

The complete data set used for modelling in chapter 4 can be seen in table 31

Variable	Description
Date	End time and date for measurements. Hourly values.
Temperature	Temperature outside in Degrees/C.
WindSpeed	
WindDirection	
SunHour	
Condition	
UltravioletIndex	
MeanSeaLevelPressure	
PrecipitationProbability	
Observation	
Consumption	CoolingDegree times Volume from House data
Holiday	

**Table 31:** Attributes used for modelling..



# 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/object 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  $\varepsilon_i$  are errors which are assumed to be white noise while also being i.i.d (independent and identically distributed). Equation (4.1) shows a multiple linear regression model as it contains more than one explanatory variable. In this section both a simple linear model and a multiple linear model has been fitted to data given in table 31.

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

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

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

#### 4.1.1 Model assumptions

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

- 1 Normality of residuals
- 2 Variance homogeneity

**3** Variance should be independent of location

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

4.2 Simple linear regression model

4.3 Multiple linear regression model

# 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 estimerterne. 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 (7) Daily averages of consumption versus temperature differences

5.2 (8) 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 (9) få styr på lorte parskip-pakken

5.3 (10) 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  $\alpha$  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 ved at skalere 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  $\alpha$ , 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 er 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.



